جامعة عبد الحميد بن باديس مستغانم

كلية العلوم الاقتصادية، التجارية وعلوم التسيير

قسم العلوم الاقتصادية

مذكرة تخرج مقدمة ضمن متطلبات نيل شهادة ماستر أكاديمي

الشعبة: علوم اقتصادية التخصص: اقتصاد كمي

البيانات الضخمة واستخداماتها في العلوم الاقتصادية:

دراسة برنامج Apache Hadoop

مقدمة من طرف الطالبين:

بلعربي حمزة

عزيز محمد الأمين

تحت اشراف الأستاذ:

يخلف عبد الله

أعضاء لجنة المناقشة:

الصفة	الاسم واللقب	الرتبة	عن الجامعة
رئيسا	د/ مصطفى بن عامر	أستاذ محاضر " أ "	جامعة مستغانم
مقررا	أ/ يخلف عبد الله	أستاذ محاضر " أ "	جامعة مستغانم
مناقشا	أ/ القري عمار	أستاذ مساعد " أ "	جامعة مستغانم

السنة الجامعية: 2022/2021

جامعة عبد الحميد بن باديس مستغانم

كلية العلوم الاقتصادية، التجارية وعلوم التسيير

قسم العلوم الاقتصادية

مذكرة تخرج مقدمة ضمن متطلبات نيل شهادة ماستر أكاديمي

الشعبة: علوم اقتصادية التخصص: اقتصاد كمي

البيانات الضخمة واستخداماتها في العلوم الاقتصادية:

دراسة برنامج Apache Hadoop

مقدمة من طرف الطلبة:

بلعربي حمزة

عزيز محمد الأمين

تحت اشراف الأستاذ:

يخلف عبد الله

أعضاء لجنة المناقشة:

الصفة	الاسم واللقب	الرتبة	عن الجامعة
رئيسا	د/ مصطفى بن عامر	أستاذ محاضر " أ "	جامعة مستغانم
مقررا	أ/ يخلف عبد الله	أستاذ محاضر " أ "	جامعة مستغانم
مناقشا	أ/ القري عمار	أستاذ مساعد " أ "	جامعة مستغانم

السنة الجامعية: 2022/2021

الشكر

و

الاهداء

الإهداء

الى من ربيانا منذ الصغر الى الآن وغرسا فينا طاعة الله ورسوله
وحب الدراسة والعمل باجتهاد
الى الوالدين الغاليين اللذان نأسأل الله أن يطيل في عمرهما.
الى من يعمل كل يوم ويتعب للعائلة، الى من علم الدين وأن الأخلاق تأتي أولا
الى من حرص على اتباع الصراط المستقيم والاجتهاد في الدراسة
الى من أحمل اسمه بكل افتخار وعزة "أبي الغالي".
الى من سهرت الليالي في تربيتي، الى من صبرت في تعليمي ما لم يكن لي به علم
الى منبع العطاء والحنان من أنارة لي الحياة عند ظلمتها
الى بسمة الدنيا ومفتاح الآخرة "أمي العزيزة".

الى كل أفراد العائلة كبيرا وصغيرا.

الى الأستاذ المحترم الدكتور يخلف عبد الله الذي ساعدنا في انجاز هذا البحث
حفضه الله وحفظنا جميعا.
الى جميع أساتذتي المحترمين.

الى كل أصدقائي ورفقائي في الدرب.

الى كل من ذكرتهم أهدي هذا العمل البسيط.

الشكر

قال رسول الله صلى الله عليه وسلم: "من لم يشكر الناس لم يشكر الله".
الحمد لله على احسانه والشكر له على توفيقه وامتنانه وأشهد ألا إله الا الله وحده لا شريك له تعظيما لشأنه، وأشهد أن محمد عبده ورسوله الداعي الى رضوانه صلى الله عليه وسلم، أما بعد:

نتقدم بجزيل الشكر للوالدين العزيزين اللذان دائما حرصا على التشجيع بالاستمرار في مسيرة العلم والنجاح أدامهما الله.

كما نتوجه بالشكر الجزيل الى من شرفنا بإشرافه على هذا البحث، الأستاذ الدكتور "يخلف عبد الله" الذي نقدره لصبره علينا في اعداد البحث، ولتوجيهاتها الثمينة التي ساعدتنا في اتمام هذا العمل بأفضل شكل ممكن.

نتوجه بالشكر الى جميع أساتذتنا المحترمين.

نوجه خالص الشكر والتقدير الى كل من ساعدنا من قريب وبعيد على اتمام هذا العم

الفهرس:

قائمة الأشكال والجداول:

الأشكال:

الجداول:

مقدمة

عامة

1. مقدمة

من السهل رؤية ثمار مجتمع المعلومات، مع وجود هاتف محمول في كل جيب، وجهاز كمبيوتر في كل منزل تقريبا، وأنظمة تكنولوجيا كبيرة ومعقدة في كل مؤسسة وفي كل مكان. لكن المعلومات نفسها أقل وضوحًا. فبعد نصف قرن من دخول أجهزة الكمبيوتر إلى المجتمع السائد، بدأت البيانات تتراكم إلى الحد الذي يحدث فيه شيء جديد ومميز. لا يقتصر الأمر على أن العالم غارق أكثر بمزيد من المعلومات من أي وقت مضى، ولكن هذه المعلومات تنمو بشكل أسرع، حيث أدى تغيير الحجم إلى تغيير الحالة والتغيير الكمي إلى تغيير نوعي. صاغت علوم مثل علم الفلك وعلم الجينوم، التي شهدت الانفجار لأول مرة في العقد الأول من القرن الحادي والعشرين، مصطلح "البيانات الضخمة". المفهوم يهاجر الآن إلى جميع مجالات النشاط البشري، ولا يوجد تعريف دقيق للبيانات الضخمة. في البداية كانت الفكرة هي أن حجم المعلومات قد نما بشكل كبير لدرجة أن الكمية التي يتم فحصها لم تعد تتناسب مع الذاكرة التي تستخدمها أجهزة الكمبيوتر للمعالجة، لذلك احتاج المهندسون إلى تجديد الأدوات التي استخدموها لتحليلها بالكامل. هذا هو أصل تقنيات المعالجة الجديدة مثل MapReduce من Google ومكافئها مفتوح المصدر، Hadoop، الذي انبثق منYahoo .يتيح ذلك للمرء إدارة كميات أكبر بكثير من البيانات عن ذي قبل، ولا يلزم وضع البيانات في صفوف مرتبة أو جداول قاعدة بيانات كلاسيكية. كما تلوح في الأفق تقنيات أخرى لمعالجة البيانات التي تستغني عن التسلسل الهرمي الصارم والتجانس القديم. في الوقت نفسه، نظرًا لأن شركات الإنترنت يمكن أن تجمع مجموعة كبيرة من البيانات ولديها حافز مالي كبير لفهمها، فقد أصبحوا المستخدمين الرئيسيين لأحدث تقنيات المعالجة، ليحلوا محل الشركات غير المتصلة بالإنترنت التي لديها، في بعض الحالات، عقود من الخبرة.

تتمثل إحدى طرق التفكير في المشكلة اليوم فيما يلي: تشير البيانات الضخمة إلى أشياء يمكن للمرء القيام بها على نطاق واسع ولا يمكن القيام بها على نطاق أصغر، لاستخراج رؤى جديدة أو إنشاء أشكال جديدة من القيمة، بطرق تغير الأسواق والمؤسسات والعلاقة بين المواطنين والحكومات والمزيد غيرها. من العلوم إلى الرعاية الصحية، من الخدمات المصرفية إلى الإنترنت، قد تكون القطاعات متنوعة ولكنها معًا تروي قصة مماثلة: كمية البيانات في العالم تنمو بسرعة، وتتفوق ليس فقط على أجهزتنا بل على خيالنا. حاول العديد من الأشخاص وضع رقم فعلي لكمية المعلومات التي تحيط بنا وحساب مدى سرعة نموها. لقد حققوا درجات متفاوتة من النجاح لأنهم قاموا بقياس أشياء مختلفة. أجرى مارتن هيلبرت إحدى الدراسات الأكثر شمولاً من كلية أنينبيرج للتواصل والصحافة بجامعة جنوب كاليفورنيا. لقد سعى جاهداً لوضع رقم على كل ما تم إنتاجه وتخزينه ونقله. لن يشمل ذلك الكتب واللوحات ورسائل البريد الإلكتروني والصور والموسيقى والفيديو (التناظرية والرقمية) فحسب، بل يشمل ألعاب الفيديو والمكالمات الهاتفية وحتى أنظمة الملاحة في السيارات والرسائل المرسلة عبر البريد. كما قام بتضمين وسائل البث مثل التلفزيون والراديو، بناءً على مدى وصول الجمهور، ووفقًا لتقدير هيلبرت، كان هناك أكثر من 300 اكسابايت من البيانات المخزنة في عام 2007، وفقط لتوضيح كم هو هائل هذا الحجم من المعلومات،

ففيلم واحد يمكن تخزينه في ملف حجمه 1 جيجابايت، و1 اكسابايت تمثل بليون جيجابايت، لذا يمكننا ان نعلم حقا أنه حجم كبير جدا. العديد من الخبراء والمختصين ينظرون إلى البيانات بأنها "نفط المستقبل" ذلك مع ظهور مجموعة من العلوم الجديدة، مثل التنقيب في البيانات التي تهتم بالصورة الخام للبيانات ومعالجتها وتحويلها في شكل يمكن الإستفادة منها في مجالات المعرفة والذكاء الاصطناعي وفروعها المختلفة، من تعلم الآلة والتعلم العميق. حيث يمكن القول أن البيانات هي الأساس أو الوقود لتلك العلوم التي لم تكن تحظى بكثير من النجاح في تلك الفترة، مثل ما هو الحال في الوقت الحاضر، ذلك لأن حجم البيانات ومصادرها التي كانت موجوده آنذاك لم تكن بنفس الكمية والعدد الموجود اليوم، نتيجة للتحول الرقمي والاعتماد على الخدمات الإلكترونية والأنظمة الذكية واستخدام مواقع التواصل الاجتماعي بشكل أساسي ويومي، والتوجه نحو الشراء الإلكتروني، علما وأن حجم البيانات تضخم بشكل غير مسبوق كما تبرزه البيانات التالية حيث تشير الإحصاءات إلى أن 90 % من البيانات المتوفرة اليوم تم إنتاجها خلال السنوات القليلة الماضية وهي في زيادة مستمرة.

ولكنها فقط البداية. يتحدى عصر البيانات الضخمة الطريقة التي نعيش بها ونتفاعل مع العالم. الأمر الأكثر لفتًا للنظر هو أن المجتمع سيحتاج إلى التخلص من بعض هوسه بالسببية في مقابل ارتباطات بسيطة: عدم معرفة السبب ولكن ماذا فقط. هذا يقلب قرونًا من الممارسات الراسخة ويتحدى فهمنا الأساسي لكيفية اتخاذ القرارات وفهم الواقع. تمثل البيانات الضخمة بداية تحول كبير. مثل العديد من التقنيات الجديدة، ستصبح البيانات الضخمة بالتأكيد ضحية لدورة الضجيج سيئة السمعة في وادي السيليكون: بعد أن تم تكريمها على غلاف المجلات وفي مؤتمرات الصناعة، سيتم رفض هذا الاتجاه وسوف تتعثر العديد من الشركات الناشئة المهتمة بالبيانات. لكن كل من الافتتان واللعنة يسيئون فهمًا عميقًا لأهمية ما يحدث. تمامًا كما مكننا التلسكوب من فهم الكون وسمح لنا المجهر بفهم الجراثيم، فإن التقنيات الجديدة لجمع وتحليل مجموعات ضخمة من البيانات ستساعدنا في فهم عالمنا بطرق بدأنا للتو في تقديرها. فالثورة الحقيقية ليست في الآلات التي تحسب البيانات ولكن في البيانات نفسها وكيف نستخدمها.

2. الإشكالية

في ضوء ما تقدم، تتمثل إشكالية هذا البحث في السؤال الرئيسي التالي:

كيف يتم معالجة البيانات الضخمة واستخدامها في مجال العلوم الاقتصادية؟ وعليه نطرح الأسئلة الفرعية التالية:

- ماهي البيانات والبيانات الضخمة؟
- ما هي آليات تحليل ومعالجة البيانات الضخمة؟

- ما هي تطبيقات البيانات الضخمة في العلوم الاقتصادية؟
- ماهي التحديات التي تواجه استخدام البيانات الضخمة؟

3. فرضيات الدراسة

من خلال الإشكالية والأسئلة الفرعية السابقة، قمنا بطرح الفرضيات التالية:

- البيانات الضخمة هي تلك البيانات التي تفوق قدرة قواعد البيانات العادية على معالجتها.
- يعتبر حجم البيانات الضخمة المولدة آليا من التحديات الكبيرة خاصة في كيفية التعامل معها إضافة إلى تنوع البيانات.
- توجد العديد من الأدوات والتقنيات التي تستخدم لتحليل البيانات الكبيرة مثل: HPCC،GridGain ، Storm،Cassandra، Hadoop، MapReduce .
- توفر البيانات الضخمة لمؤسسات الأعمال المعلومات اللازمة لمعرفة رغبات الزبائن وميولهم ونفسياتهم، هذا ما يتيح للشركات الربحية توفير منتجات وخدمات بناءا على ذلك، وعليه تضمن تلك الشركات رضا عملاءها مما يؤدي الى زيادة مبيعاتها.

4. أسباب اختيار الدراسة

الداعي لاختيار هذا الموضوع يكمن أساسا فيما يلي:

يعتبر موضوع الدراسة من أحد المواضيع العصرية والمتجددة في البوابة الاقتصادية والمعلوماتية.

وفرت المعلومات النظرية والقياسية نظرا لتعدد الباحثين في هذا المجال مما يأكد أهمية موضوع الدراسة.

الدور البارز الذي تلعبه البيانات الضخمة في جميع المجالات خاصة المجال الاقتصادي، وهذا ما نلاحظه في الدول المتقدمة.

ميل الباحثين الى مجال المعلوماتية خصوصا جانب تحليل البيانات (Data Analysis)، لذا فاختيار هذا الموضوع يساهم أكثر في تغذية المعرفة بذلك الجانب.

5. **أهمية الدراسة**

انطلاقا من مشكلة البحث، يمكن ابراز أهميته من خلال الدور الكبير الذي تلعبه البيانات الضخمة في وقتناالراهن، فهي تحدث تغيرات كبيرة داخل المجتمعات إذا ما تم استغلالها بعناية ودقة، فالبيانات الضخمة إذا تمت ادارتها بشكل صحيح فمن شأنها أن تساهم بشكل فعال في التنمية الاقتصادية والاجتماعية للدول والمجتمعات.

6. أهدف الدراسة

تهدف هذه الدراسة إلى الإجابة عن التساؤلات السابقة بالإضافة إلى:

اعداد إطار نظري لأهم المرتكزات الفكرية لمفهوم البيانات الضخمة بالنسبة لتخصص العلوم الاقتصادية.

شرح أهم الأساليب والبرامج المستخدمة في تحليل ومعالجة البيانات الضخمة، حيث نلاحظ أن العديد من الدراسات العربية لا تقوم بإعطاء القدر الكافي لهذه البرامج عكس الدراسات الأجنبية، فمعرفة كيفية التعامل مع البيانات الضخمة وتحليلها يساهم في إدراك القارء للمهارة البرمجية والدقة التي تحتاجها هذه العملية.

ابراز الصعوبات التي تواجه استخدام البيانات الضخمة من مشاكل الخصوصية، الحجم الهائل وسرعة تدفق المعلومات، عدم توفر الآليات اللازمة لتخزين البيانات مع مرور الوقت وغيرها.

7. المنهج المستخدم

اعتمدت مجموعة الدراسة على المنهج الوصفي التحليلي الذي يتماشى مع طبيعة هذه الدراسة التي تعتبر استكشافية بالدرجة الأولى، فالمنهج الوصفي التحليلي يقوم بدارسة الظاهرة، ويصفها وصفا دقيقا، كما تم الاعتماد على المنهج القياسي ليعطيها وصفا رقميا يوضح مقدار هذه الظاهرة وحجمها ودرجة ارتباطها مع الظواهر الأخرى، كما تم الاعتماد على هذا المنهج الذي يقوم على تحليل المحتوى والكشف عن الاتجاهات والميول، ويهدف للوصف الموضوعي والمنظم للمحتوى ووسيلة أنجع للقيام باستنتاجات، عن طريق التحديد المنظم والموضوعي.

الفصل الأول:

ماهية

البيانات

الضخمة

تمهيد:

تعتبر البيانات الضخمة نقطة تحول في عصرنا الحالي، إذ تمكننا من تحويل الحياة الاجتماعية والتي هي ثمرة الاستخدام المكثف للتكنولوجيا إلى بيانات قابلة للقياس، بحيث يمكننا مراقبة الأنشطة البشرية والتفاعلات مع البيئة مما ينتج عنه بصمة رقمية هائلة.

ومع زيادة الاستخدام للتكنولوجيا هناك زيادة موازية في تدفقات المعلومات وجمع البيانات التي تنشأ يوميا، والسبب الأهم لزيادة حجم البيانات، لأنها تستمر بالتولد بشكل أكبر بكثير من السابق من خلال عدة أجهزة ومصادر، والأهم أن معظم تلك البيانات ليست مهيكلة، كغريدات تويتر والفيديوهات على يوتيوب وتحديثات الحالة على فيس بوك وغيرها، ما يعني أنه لا يمكن استخدام أدوات إدارة قواعد البيانات وتحليلها التقليدية مع هذه البيانات لأنها ببساطة ليست وفق الهيكل المعتاد الذي يتم التعامل معه كالجداول بالكسل فرضا. وفي هذا الفصل، سنحاول إعطاء صورة واضحة لماهية البيانات الضخمة كونها موضوع مهم في عصرنا هذا.

المبحث الأول: البيانات

تتزايد ظاهرة تحليلات البيانات الضخمة باستمرار حيث تعيد المؤسسات تصميم عملياتها التشغيلية للاعتماد على البيانات الحية على أمل دفع تقنيات التسويق الفعالة، وتحسين مشاركة العملاء، وربما تقديم منتجات وخدمات جديدة. لكن قبل التطرق لكل هذا يجب أن نعلم ما هي البيانات أصلا، أنواعها وكيفية قياسها.

المطلب الأول: مفهوم البيانات

في الحوسبة، البيانات هي المعلومات التي تمت ترجمتها إلى شكل فعال للحركة أو المعالجة. بالنسبة إلى أجهزة الكمبيوتر ووسائط النقل الحالية، فإن البيانات هي معلومات يتم تحويلها إلى شكل رقمي ثنائي. من المقبول استخدام البيانات كموضوع مفرد أو موضوع جمع. البيانات الأولية هي مصطلح يستخدم لوصف البيانات بأبسط تنسيق رقمي لها.

يمكن تعريفها بأنها مجموعة من الحروف، أو الكلمات، أو الأرقام، أو الرموز، أو الصور المتعلقة بموضوع ما . والبيانات في حد ذاتها ليس لها معنى أو قيمة، وهي الصورة الخام للمعلومة مثال بيانات الموظفين وصورهم.[1]

هي صفات وأرقام مشوشة وغير مرئية ومزدحمة بحيث لا يمكن استخراج أي حكمة أو قاعدة منها قبل أن يتم معالجتها[2].

وتعتبر هي أيضا المادة الخام التي سيتم تجهيزها للحصول على معلومات أو لجمع مزيد من التفاصيل ولا تكون لها فائدة إلا في حال معالجتها وتحليلها، فبيانات الطالب على سبيل المثال المكونة من الاسم واللقب، واسم الأب، عنوان الطالب ... الخ، تعتبر بيانات ولا تعتبر معلومات، بيانات المواطنين أيضا التي يتم جمعها أثناء التعداد مثلا، والبيانات التي يتم جمعها عن طريق استطلاعات الرأي حول منتج معين.[3]

تعود جذور مفهوم البيانات في سياق الحوسبة إلى أعمال كلود شانون، عالم الرياضيات الأمريكي المعروف باسم أبو نظرية المعلومات. لقد بشر بالمفاهيم الرقمية الثنائية القائمة على تطبيق منطق منطقي ثنائي القيمة على الدوائر الإلكترونية. تشكل تنسيقات الأرقام الثنائية أساس وحدات المعالجة المركزية وذاكرة أشباه الموصلات ومحركات الأقراص، بالإضافة إلى العديد من الأجهزة الطرفية الشائعة في الحوسبة اليوم. اتخذت المدخلات الحاسوبية المبكرة لكل من التحكم والبيانات شكل بطاقات مثقبة، متبوعة بشريط مغناطيسي والقرص الصلب[4].

---

[1] Sanders, John : Defining Terms: Data, Information and Knowledge, 2016, pp.1-3

[2] هناء قيراطي: توظيف البيانات الضخمة في الشركات التقنية وخصوصية المستخدم، مذكرة تخرج شهادة ماستر، جامعة 8 ماي 1945 –قالمة-, 2016-2017، ص18.

[3] https://differencebtw.com/data-vs-information/ , viewed at, 17:07, 12/05/2022.

[4] https://www.techtarget.com/searchdatamanagement/definition/data , viewed at, 17:09, 12/05/2022

في وقت مبكر، أصبحت أهمية البيانات في حوسبة الأعمال واضحة من خلال شعبية مصطلحات "معالجة البيانات" و "معالجة البيانات الإلكترونية"، والتي أصبحت، لبعض الوقت، تشمل النطاق الكامل لما يعرف الآن باسم تكنولوجيا المعلومات. على مدار تاريخ حوسبة الشركات، حدث التخصص وظهرت مهنة بيانات متميزة جنبًا إلى جنب مع نمو معالجة بيانات الشركة.

على الرغم من أن المصطلحين "بيانات" و "معلومات" غالبًا ما يتم استخدامهما بالتبادل، إلا أن هذين المصطلحين لهما معاني مختلفة عن بعضهما البعض. في بعض المنشورات الشائعة، يُقال أحيانًا أن البيانات تتحول إلى معلومات عندما يتم عرضها في السياق أو في التحليل اللاحق. ومع ذلك، في المعالجات الأكاديمية، تكون بيانات الموضوع مجرد وحدات معلومات. تُستخدم البيانات في البحث العلمي، وإدارة الأعمال (على سبيل المثال، بيانات المبيعات، والإيرادات، والأرباح، وأسعار الأسهم، والتمويل، والحوكمة كمثال نجد، معدلات الجريمة، ومعدلات البطالة، ومعدلات معرفة القراءة والكتابة، وفي كل شكل آخر من أشكال النشاط التنظيمي البشري تقريبًا على سبيل المثال، تعدادات عدد الأشخاص الذين لا مأوى لهم من قبل المنظمات غير الهادفة للربح.

بشكل عام، البيانات هي ذرات صنع القرار: إنها أصغر وحدات المعلومات الواقعية التي يمكن استخدامها كأساس للتفكير أو المناقشة أو الحساب. يمكن أن تتراوح البيانات من الأفكار المجردة إلى القياسات الملموسة، وحتى الإحصائيات. يتم قياس البيانات وجمعها والإبلاغ عنها وتحليلها واستخدامها لإنشاء تصورات للبيانات مثل الرسوم البيانية أو الجداول أو الصور. تشير البيانات كمفهوم عام إلى حقيقة أن بعض المعلومات أو المعرفة الموجودة يتم تمثيلها أو ترميزها في شكل ما مناسب للاستخدام أو المعالجة بشكل أفضل. البيانات الأولية ("البيانات غير المعالجة") هي مجموعة من الأرقام أو الأحرف قبل "تنظيفها" وتصحيحها بواسطة الباحثين. يجب تصحيح البيانات الأولية لإزالة القيم المتطرفة أو أخطاء إدخال البيانات أو الأداة الواضحة على سبيل المثال، قراءة مقياس الحرارة من موقع خارجي في القطب الشمالي يسجل درجة حرارة استوائية). تحدث معالجة البيانات عادة على مراحل، ويمكن اعتبار "البيانات المعالجة" من مرحلة واحدة "البيانات الأولية" للمرحلة التالية. البيانات الميدانية هي بيانات أولية يتم جمعها في بيئة "في الموقع" غير خاضعة للرقابة. البيانات التجريبية هي البيانات التي يتم إنشاؤها في سياق التحقيق العلمي عن طريق الملاحظة والتسجيل.

المطلب الثاني: قياس البيانات وأنواعها

قد يكون من السهل أن تضيع في عالم مصطلحات تخزين البيانات، لا سيما عند مناقشة وحدات قياس تخزين البيانات. ما هو الفرق بين البايت والبايت؟ ميغا بايت وجيجابايت؟ تيرابايت وكي بايت؟ سيساعد هذا المطلب في تقسيم هذه المفاهيم إلى أجزاء يسهل إدارتها بحجم البايت.

الفرع الأول: قياس البيانات

عندما يتعلق الأمر بتخزين البيانات، من المهم أن تكون واقعيًا بشأن احتياجاتك. هل أنت شركة صغيرة تستخدم حلًا بسيطًا ولكنه متعدد الاستخدامات مثلDAS ؟ أم أن مؤسستك تدرس مزايا SAN وNAS؟ بغض النظر عن الحل الذي ينتهي بك العمل معه، كل هذا يتوقف على مقدار البيانات التي تحتاجها بالفعل لتخزينها والوصول إليها. سواء كنت تصل إلى السحابة أو محرك أقراص ثابت محلي، فإن كمية البيانات التي يتفاعل معها نشاطك التجاري ستحدد في النهاية نوع التكنولوجيا التي ستحتاج إليها.

تمثل أجهزة الكمبيوتر البيانات، بما في ذلك الفيديو والصور والأصوات والنصوص، كقيم ثنائية باستخدام أنماط مكونة من رقمين فقط: 1 و0. البتة هي أصغر وحدة بيانات، وتمثل قيمة واحدة فقط. البايت يتكون من ثمانية أرقام ثنائية. يتم قياس التخزين والذاكرة بالميجابايت والجيجابايت. تستمر وحدات قياس البيانات في النمو مع زيادة كمية البيانات التي يتم جمعها وتخزينها. المصطلح الجديد نسبيًا "brontobyte"على سبيل المثال، هو تخزين البيانات الذي يساوي 10 أس 27 بايت[1].

يمكن تخزين البيانات في تنسيقات ملفات، كما هو الحال في أنظمة الحواسيب المركزية باستخدام ISAM وVSAM. تتضمن تنسيقات الملفات الأخرى لتخزين البيانات وتحويلها ومعالجتها قيمًا مفصولة بفواصل. استمرت هذه التنسيقات في العثور على استخدامات عبر مجموعة متنوعة من أنواع الماكينات، حتى مع اكتساب المزيد من الأساليب الموجهة نحو البيانات المنظمة مكانتها في حوسبة الشركات. تم تطوير تخصص أكبر مثل نظام إدارة قواعد البيانات وقواعد البيانات ثم نشأت تقنية قواعد البيانات العلائقية لتنظيم المعلومات.

يمكن توضيح وحدات القياس الخاصة بالبيانات في الجدول التالي:

[1] https://www.techtarget.com/searchdatamanagement/definition/data , viewed at, 17:41, 14/05/2022

الجدول أ: جدول وحدات قياس البيانات.

وحدة القياس	الحجم	الاختصار
البت	0 أو 1	b
بايت	8 بت	B
كيلوبايت Kilobyte	1024 بايت	Kb
ميجابايت Megabyte	1024 كيلوبايت	Mb
جيجابايت Gigabyte	1024 ميجابايت	Gb
تيرابايت Terabyte	1024 جيجابايت	Tb
بيتابايت Petabyte	1024 تيرابايت	Pb
اكسابايت Exabyte	1024 بيتابايت	Eb
زيتابايت Zettabyte	1024 اكسابايت	Zb
يوتابايت Yottabyte	1024 زيتابايت	Yb
زينوتابايت Xenottabyte	1024 يوتابايت	Xb
شايلينوبايت Shilentnobyte	1024 زينوتابايت	Sb
دوميجمقروبايتDomegemegrottebyte	1024 شايلينو بايت	Db

المصدر: من اعداد الطالبين اعتمادا على المكتسبات القبلية.

الفرع الثاني: أنواع البيانات

أدى نمو الويب والهواتف الذكية خلال العقد الماضي إلى زيادة في إنشاء البيانات الرقمية. تتضمن البيانات الآن معلومات نصية وصوتية وفيديو، بالإضافة إلى سجلات نشاط الويب والسجل. الكثير من ذلك هو بيانات غير منظمة.

تم استخدام مصطلح البيانات الضخمة لوصف البيانات الموجودة في نطاق بيتابايت أو أكبر. تصور لقطة الاختزال البيانات الضخمة ذات 3-Vsالحجم والتنوع والسرعة. مع انتشار التجارة الإلكترونية المستندة إلى الويب، تطورت نماذج الأعمال القائمة على البيانات الضخمة والتي تتعامل مع البيانات كأصل في حد ذاتها. لقد ولدت هذه الاتجاهات أيضًا انشغالًا أكبر بالاستخدامات الاجتماعية للبيانات وخصوصية البيانات.

البيانات لها معنى يتجاوز استخدامها في تطبيقات الحوسبة الموجهة نحو معالجة البيانات. على سبيل المثال، في التوصيل البيني للمكونات الإلكترونية واتصالات الشبكة، غالبًا ما يتم تمييز مصطلح البيانات عن "معلومات التحكم" و "بتات التحكم" والمصطلحات المماثلة لتحديد المحتوى الرئيسي لوحدة الإرسال. علاوة على ذلك، في العلم، يستخدم مصطلح البيانات لوصف مجموعة من الحقائق المجمعة. هذا هو الحال أيضًا في مجالات مثل التمويل والتسويق والتركيبة السكانية والصحة.

عموما عند الفصل في تقسيم البيانات نجد قسمين، الأول يقسم البيانات وفق شكلها الخام الى نوعين (بيانات مهيكلة، بيانات غير مهيكلة)، أما الثاني فيقسمها الى وفق التحليل الإحصائي الى أربعة أنواع، وهذا ما سنتطرق اليه في هذا الفرع:

1) **التصنيف وفق أشكال البيانات:** البيانات المنظمة مقابل البيانات غير المهيكلة حيث تتكون البيانات المنظمة من أنواع بيانات محددة بوضوح مع أنماط تجعلها سهلة البحث؛ بينما تتكون البيانات غير المهيكلة من بيانات لا يمكن البحث عنها بسهولة، بما في ذلك تنسيقات مثل الصوت والفيديو ومنشورات الوسائط الاجتماعية.

- **البيانات المهيكلة:** عادة ما توجد البيانات المهيكلة في قواعد البيانات العلائقية (RDBMS). تخزن الحقول بيانات محددة الطول مثل أرقام الهواتف أو أرقام الضمان الاجتماعي أو الرموز البريدية. تحتوي السجلات حتى على سلاسل نصية ذات أطوال متغيرة مثل الأسماء، مما يجعل البحث فيها أمرًا بسيطًا. قد تكون البيانات من صنع الإنسان أو الآلة، طالما يتم إنشاء البيانات داخل بنية RDBMS. هذا التنسيق قابل للبحث بشكل بارز، سواء من خلال الاستعلامات التي ينشئها الإنسان وعبر الخوارزميات باستخدام أنواع البيانات وأسماء الحقول، مثل الأبجدية أو الرقمية أو العملة أو التاريخ.[1]

تشمل تطبيقات قواعد البيانات العلائقية الشائعة مع البيانات المنظمة أنظمة حجز شركات الطيران، ومراقبة المخزون، ومعاملات المبيعات، ونشاط أجهزة الصراف الآلي. تتيح لغة الاستعلام الهيكلية (SQL) الاستعلامات حول هذا النوع من البيانات المنظمة داخل قواعد البيانات العلائقية.

تقوم بعض قواعد البيانات العلائقية بتخزين أو الإشارة إلى بيانات غير منظمة، مثل تطبيقات إدارة علاقات العملاء (CRM). يمكن أن يكون التكامل محرجًا في أحسن الأحوال لأن حقول المذكرات لا تصلح لاستعلامات قاعدة البيانات التقليدية. ومع ذلك، فإن معظم بيانات CRM منظمة.

[1] https://www.datamation.com/big-data/structured-vs-unstructureddata/#:~:text=unstructured%20data%3A%20structured%20data%20is,video%2C%20and%20social%20media%20postings , viewed at, 17:22, 15/05/2022.

- **البيانات غير المهيكلة**: البيانات غير المهيكلة هي في الأساس كل شيء آخر. البيانات غير المهيكلة لها بنية داخلية ولكنها غير منظمة عبر نماذج بيانات محددة مسبقًا أو مخطط. قد تكون نصية أو غير نصية، ومن صنع الإنسان أو الآلة. يمكن أيضًا تخزينها في قاعدة بيانات غير علائقية مثل NoSQL.

تشمل البيانات النموذجية غير المهيكلة التي ينتجها الإنسان ما يلي:[1]

-الملفات النصية: معالجة الكلمات وجداول البيانات والعروض التقديمية ورسائل البريد الإلكتروني والسجلات.

-البريد الإلكتروني: يحتوي البريد الإلكتروني على بعض الهياكل الداخلية بفضل البيانات الوصفية الخاصة به، ونشير إليه أحيانًا على أنه شبه منظم. ومع ذلك، فإن حقل رسالته غير منظم ولا يمكن لأدوات التحليل التقليدية تحليله.

-وسائل التواصل الاجتماعي: بيانات من Facebook وTwitter وLinkedIn.

-الموقع الإلكتروني: يوتيوب، إنستغرام، مواقع مشاركة الصور.

-بيانات الجوال: الرسائل النصية والمواقع.

-الاتصالات: الدردشة، والمراسلة الفورية، وتسجيلات الهاتف، وبرامج التعاون.

-الوسائط: ملفات MP3 والصور الرقمية وملفات الصوت والفيديو.

-تطبيقات الأعمال: مستندات MS Office، تطبيقات الإنتاجية.

تتضمن البيانات النموذجية غير المهيكلة المُنشأة آليًا ما يلي:

-صور القمر الصناعي: بيانات الطقس، التضاريس، التحركات العسكرية.

-البيانات العلمية: استكشاف النفط والغاز، استكشاف الفضاء، الصور الزلزالية، بيانات الغلاف الجوي.

-المراقبة الرقمية: صور المراقبة والفيديو.

-بيانات أجهزة الاستشعار: حركة المرور والطقس وأجهزة الاستشعار الأوقيانوغرافية.

إلى جانب الاختلاف الواضح بين التخزين في قاعدة بيانات علائقية والتخزين خارج واحدة، فإن الاختلاف الأكبر بين البيانات المهيكلة وغير المهيكلة هو سهولة التحليل. توجد أدوات تحليلات ناضجة للبيانات المهيكلة، لكن أدوات التحليل لتعدين البيانات غير المهيكلة لا تزال وليدة وتتطور.

يمكن للمستخدمين إجراء عمليات بحث بسيطة عن المحتوى عبر بيانات نصية غير منظمة. لكن افتقارها إلى البنية الداخلية المنظمة يهزم الغرض من أدوات التنقيب عن البيانات التقليدية، ولا تحصل

[1] https://www.ibm.com/cloud/blog/structured-vs-unstructured-data , viewed at 18:53, 15/05/2022.

المؤسسة على قيمة تذكر من مصادر البيانات ذات القيمة المحتملة مثل الوسائط الغنية، والشبكات أو مدونات الويب، وتفاعلات العملاء، وبيانات الوسائط الاجتماعية.

علاوة على ذلك، هناك ببساطة بيانات غير منظمة أكثر بكثير من البيانات المهيكلة. تشكل البيانات غير المهيكلة 80٪ وأكثر من بيانات المؤسسة، وتنمو بمعدل 55٪ و65٪ سنويًا. وبدون الأدوات اللازمة لتحليل فئة البيانات الضخمة هذه، تترك المؤسسات كميات هائلة من البيانات القيمة في جدول ذكاء الأعمال.

كما يوجد هناك نوع آخر يجدر بنا ذكره وهو البيانات شبه المهيكلة, وتشير الى البيانات التي لم يتم التقاطها أو تنسيقها بطرق تقليدية. لا تتبع البيانات شبه المنظمة تنسيق نموذج البيانات المجدول أو قواعد البيانات العلائقية لأنها لا تحتوي على مخطط ثابت. ومع ذلك، فإن البيانات ليست أولية أو غير منظمة تمامًا، وتحتوي على بعض العناصر الهيكلية مثل العلامات والبيانات الوصفية التنظيمية التي تسهل التحليل. تتمثل مزايا البيانات شبه المنظمة في أنها أكثر مرونة وأسهل في القياس مقارنةً بها.

كود HTML والرسوم البيانية والجداول ورسائل البريد الإلكتروني ووثائق XML هي أمثلة على البيانات شبه المنظمة، والتي توجد غالبًا في قواعد البيانات الموجهة للكائنات.[1]

2) **التصنيف وفق التحليل الإحصائي:** في الإحصاء، هناك أربعة أنواع للبيانات وفق مقاييس قياس البيانات: الاسمية والترتيبية والفاصل الزمني والنسبة. هذه مجرد طرق لتصنيف أنواع مختلفة من البيانات (فيما يلي نظرة عامة على أنواع البيانات الإحصائية). عادة ما تتم مناقشة هذا الموضوع في سياق التدريس الأكاديمي إذا كنت تقوم بصقل هذا المفهوم لإجراء اختبار إحصائي، اشكر عالمًا نفسيًا باحثًا يدعى ستانلي ستيفنز على ابتكار هذه المصطلحات. من الأفضل فهم هذه المقاييس الأربعة لقياس البيانات (الاسمي والترتيبي والفاصل الزمني والنسبة) مع المثال، كما سنرى أدناه: [2]

- **البيانات الاسمية:** فلنبدأ بأسهل طريقة يمكن فهمها. تستخدم المقاييس الاسمية لوصف المتغيرات، دون أي قيمة كمية. يمكن ببساطة تسمية المقاييس "الاسمية" "بالعلامات". علما أن كل هذه المقاييس متنافية (لا يوجد تداخل) وليس لأي منها أي أهمية عددية. هناك طريقة جيدة لتذكر كل هذا وهي أن الأصوات "الاسمية" تشبه إلى حد كبير "الاسم" وأن المقاييس الاسمية تشبه نوعًا ما "الأسماء" أو الملصقات.
- **البيانات الترتيبية:** باستخدام المقاييس الترتيبية، يكون ترتيب القيم هو المهم والمهم، لكن الاختلافات بين كل منها غير معروفة حقًا. مثلا. في كل حالة، نعلم أن رقم 4 أفضل من رقم 3 أو رقم 2، لكننا لا نعرف -

[1] https://www.teradata.com/Glossary/What-is-Semi-Structured-Data#:~:text=Semi%2Dstructured%20data%20refers%20to,not%20have%20a%20fixed%20schema , viewed at 19:06, 16/05/2022

[2] https://www.mymarketresearchmethods.com/types-of-data-nominal-ordinal-interval-ratio/, viewed at 19:38, 16/05/2022.

ولا يمكننا تحديد مقدار ذلك -إلى أي مدى هو أفضل. على سبيل المثال، هل الفرق بين "حسنًا" و "غير سعيد" هو نفسه الفرق بين "سعيد جدًا" و "سعيد" لا نستطيع ان نقول.

المقاييس الترتيبية هي مقاييس لمفاهيم غير رقمية مثل الرضا والسعادة وعدم الراحة وما إلى ذلك. من السهل تذكر كلمة "ترتيبي" لأنها تبدو مثل "ترتيب" وهذا هو المفتاح الذي يجب تذكره باستخدام "المقاييس الترتيبية". الترتيب هو المهم، ولكن هذا كل ما تحصل عليه حقًا من هذه.

- **البيانات الزمنية:** مقاييس الفاصل الزمني هي مقاييس رقمية نعرف من خلالها كلاً من الترتيب والاختلافات الدقيقة بين القيم. المثال الكلاسيكي لمقياس الفاصل الزمني هو درجة الحرارة المئوية لأن الفرق بين كل قيمة هو نفسه. على سبيل المثال، الفرق بين 60 و50 درجة هو 10 درجات قابلة للقياس، كما هو الحال بين 80 و70 درجة.

المقاييس الفاصلة جيدة لأن مجال التحليل الإحصائي على مجموعات البيانات هذه ينفتح. على سبيل المثال، يمكن قياس الاتجاه المركزي بالنمط أو الوسيط أو المتوسط؛ يمكن أيضًا حساب الانحراف المعياري.

مثل الآخرين، يمكنك تذكر النقاط الرئيسية للمقياس الفاصل بسهولة تامة. الفاصل نفسه يعني المسافة بين، وهو الشيء المهم الذي يجب تذكره لا تخبرنا المقاييس الفاصلة عن الترتيب فحسب، بل تخبرنا أيضًا بالقيمة بين كل عنصر.

لكن المشكلة في مقاييس الفترات أنه ليس لديهم "صفر حقيقي". على سبيل المثال، لا يوجد شيء مثل "لا درجة حرارة"، على الأقل ليس بالدرجة المئوية. في حالة مقاييس الفترات، لا يعني الصفر غياب القيمة، ولكنه في الواقع رقم آخر مستخدم على المقياس، مثل 0 درجة مئوية. الأرقام السالبة لها معنى أيضًا. بدون صفر حقيقي، من المستحيل حساب النسب. باستخدام بيانات الفاصل الزمني، يمكننا الجمع والطرح، لكن لا يمكننا الضرب أو القسمة.

مثلا: 10 درجات م+ 10 درجات م = 20 درجة م. لا توجد مشكلة هناك. 20 درجة مئوية ليست ضعف درجة حرارة 10 درجات مئوية، لأنه لا يوجد شيء مثل "لا درجة حرارة" عندما يتعلق الأمر بمقياس سيليزي. عند التحويل إلى فهرنهايت، يكون واضحًا: 10 درجة مئوية = 50 درجة فهرنهايت و20 درجة مئوية = 68 درجة فهرنهايت، والتي من الواضح أنها ليست ضعف درجة الحرارة. وخلاصة القول، المقاييس الفاصلة رائعة، لكن لا يمكننا حساب النسب، وهو ما يقودنا إلى آخر مقياس للقياس.

- **البيانات النسبية:** مقاييس النسبة هي النيرفانا النهائية عندما يتعلق الأمر بمقاييس قياس البيانات لأنها تخبرنا عن الترتيب، وتخبرنا بالقيمة الدقيقة بين الوحدات، ولديها أيضًا صفر مطلق مما يسمح بمجموعة واسعة من الإحصائيات الوصفية والاستنتاجية ليتم تطبيقها. مع المخاطرة بتكرار نفسي، فإن كل ما ورد

أعلاه حول بيانات الفاصل الزمني ينطبق على مقاييس النسبة، بالإضافة إلى مقاييس النسبة لها تعريف واضح للصفر. تشمل الأمثلة الجيدة لمتغيرات النسبة الطول والوزن والمدة.

توفر مقاييس النسبة ثروة من الاحتمالات عندما يتعلق الأمر بالتحليل الإحصائي. يمكن إضافة هذه المتغيرات بشكل مفيد وطرحها ومضاعفتها وتقسيمها (النسب). يمكن قياس الاتجاه المركزي بالوسط أو الوسيط أو المتوسط؛ يمكن أيضًا حساب مقاييس التشتت، مثل الانحراف المعياري ومعامل الاختلاف من مقاييس النسبة.

المطلب الثالث: مدخل في قاعدة البيانات

مع انتشار البيانات في المنظمات، تم التركيز بشكل إضافي على ضمان جودة البيانات عن طريق تقليل الازدواجية وضمان استخدام السجلات الحالية الأكثر دقة. تتضمن الخطوات العديدة المتضمنة في إدارة البيانات الحديثة تنقية البيانات، وكذلك عمليات الاستخراج والتحويل والتحميل (ETL) لدمج البيانات. أصبحت بيانات المعالجة تُستكمل ببيانات وصفية، يشار إليها أحيانًا باسم "بيانات حول البيانات"، والتي تساعد المسؤولين والمستخدمين على فهم قاعدة البيانات والبيانات الأخرى.

أصبحت التحليلات التي تجمع بين البيانات المنظمة وغير المهيكلة مفيدة، حيث تسعى المؤسسات إلى الاستفادة من هذه المعلومات. تسعى أنظمة مثل هذه التحليلات بشكل متزايد إلى الأداء في الوقت الفعلي، لذا فهي مصممة للتعامل مع البيانات الواردة المستهلكة بمعدلات استيعاب عالية، ومعالجة تدفقات البيانات للاستخدام الفوري في العمليات.

بمرور الوقت، تم توسيع فكرة قاعدة البيانات للعمليات والمعاملات لتشمل قاعدة البيانات لإعداد التقارير وتحليلات البيانات التنبؤية. والمثال الرئيسي هو مستودع البيانات، والذي تم تحسينه لمعالجة الأسئلة المتعلقة بالعمليات لمحللي الأعمال وكبار رجال الأعمال. أدت زيادة التركيز على إيجاد الأنماط والتنبؤ بنتائج الأعمال إلى تطوير تقنيات التنقيب عن البيانات. مهنة مسؤول قاعدة البيانات هي فرع من تكنولوجيا المعلومات. يعمل خبراء قواعد البيانات هؤلاء على تصميم قاعدة البيانات وضبطها وصيانتها. وقد اكتسبت مهنة البيانات جذورًا راسخة حيث اكتسب نظام إدارة قواعد البيانات العلائقية (RDBMS) استخدامًا واسعًا في الشركات، بدءًا من الثمانينيات. تم تمكين صعود قاعدة البيانات العلائقية جزئيًا من خلال لغة الاستعلام الهيكلية .(SQL) في وقت لاحق، ظهرت قواعد البيانات غيرSQL ، والمعروفة باسم قواعد بياناتNoSQL ، كبديل لقواعدRDBMS .[1]

[1] عمار محمد هلال: قواعد البيانات باستخدام SQL, 2017-2018، ص ص11-13.

اليوم، توظف الشركات متخصصين في إدارة البيانات أو تكلف العمال بدور الإشراف على البيانات، والذي يتضمن تنفيذ استخدام البيانات وسياسات الأمان على النحو المبين في مبادرات إدارة البيانات.

ظهر عنوان مميز -عالم البيانات -لوصف المهنيين الذين يركزون على التنقيب عن البيانات وتحليلها. أدت فائدة تقديم علم البيانات بطريقة مثيرة للعواطف إلى ظهور فنان البيانات؛ أي فرد بارع في رسم البيانات وتصورها بطرق إبداعية.

المبحث الثاني: البيانات الضخمة

ظهور موجة جديدة من البيانات من المصادر، مثل إنترنت الأشياء، وشبكات الاستشعار، والبيانات المفتوحة على الويب، والبيانات من تطبيقات الهاتف المحمول، وبيانات الشبكات الاجتماعية، جنبًا إلى جنب مع النمو الطبيعي لمجموعات البيانات داخل المنظمات، يخلق طلبًا على استراتيجيات إدارة البيانات الجديدة التي يمكنها التعامل مع هذه المقاييس الجديدة لبيئات البيانات. البيانات الضخمة هي مجال ناشئ حيث تقدم التكنولوجيا المبتكرة طرقًا جديدة لإعادة استخدام المعلومات واستخلاص القيمة منها. يُنظر الآن إلى القدرة على إدارة المعلومات بشكل فعال واستخراج المعرفة على أنها ميزة تنافسية رئيسية، وتقوم العديد من المنظمات ببناء أعمالها الأساسية بناءً على قدرتها على جمع المعلومات وتحليلها لاستخراج المعرفة التجارية والبصيرة. إن اعتماد تكنولوجيا البيانات الضخمة داخل القطاعات الصناعية ليس رفاهية ولكنه حاجة ملحة لمعظم المنظمات لاكتساب ميزة تنافسية. ولهذا سنتطرق في هذا الجزء من البحث نحو إعطاء صورة واضحة عن ماهية البيانات الضخمة والبحث عن التعريفات والمبادئ المتعلقة بالبيانات الضخمة.

المطلب الأول: مفهوم البيانات الضخمة

لنفترض أن شركة اتصالات تريد تقليل مخاطر فقدان العملاء، لذا فهي تحلل المليارات من سجلات تفاصيل المكالمات لمعرفة العملاء الأكثر اتصالاً (أي إجراء أو استقبال معظم المكالمات من مجموعة متنوعة من أرقام الهواتف). تركز الشركة بعد ذلك على العروض الترويجية على هؤلاء الأفراد لإبقائهم عملاء سعداء، لأنهم إذا غادروا، فقد يسحبون الكثير من الأصدقاء معهم إلى شركة اتصال جديدة. هذا النوع من الرؤى الخفية يوضح كيف توسع البيانات الضخمة نطاق المعلومات المستخدمة في صنع القرار. يمكن للشركات الآن إنشاء قيمة عمل جديدة من خلال الاستفادة من مصادر البيانات التي كان من الصعب في السابق الحصول عليها والوصول إليها وتحليلها بسبب التحديات المتعلقة بحجمها وسرعتها وهيكلها.

فالعديد من الخبراء والمختصين ينظرون إلى البيانات بأنها "نفط المستقبل" ذلك مع ظهور مجموعة من العلوم الجديدة، مثل التنقيب في البيانات التي تهتم بالصورة الخام للبيانات ومعالجتها وتحويلها في شكل يمكن الاستفادة منه في مجالات المعرفة والذكاء الاصطناعي وفروعه المختلفة، من تعلم الآلة والتعلم العميق. حيث يمكن القول أن البيانات هي الأساس أو الوقود لتلك العلوم التي لم تكن تحظى بكثير من النجاح في تلك الفترة، مثل ما هو الحال في الوقت الحاضر، ذلك لأن حجم البيانات ومصادرها التي كانت موجوده آنذاك لم تكن بنفس الكمية والعدد الموجود اليوم، نتيجة للتحول الرقمي والاعتماد على الخدمات الإلكترونية والأنظمة الذكية واستخدام مواقع التواصل الاجتماعي بشكل أساسي ويومي، والتوجه نحو الشراء الإلكتروني، علما وأن حجم البيانات تضخم بشكل غير مسبوق كما تبرزه البيانات التالية حيث تشير الإحصاءات إلى أن 90 % من البيانات المتوفرة اليوم تم إنتاجها خلال السنوات القليلة الماضية وهي في زيادة مستمرة.

وقد تختلف وتتنوع التعريفات والمفاهيم لهذا المجال ما بين الخبراء والشركات والمنظمات المتخصصة، حيث عرف معهد ماكنزي العالمي البيانات الضخمة، أنها مجموعة من البيانات التي يفوق حجمها القدرة على معالجتها باستخدام أدوات قواعد البيانات التقليدية، من التقاط ومشاركة ونقل وتخزين وإدارة وتحليل، في غضون فترة زمنية مقبولة.[1]

كما تعرف شركة جارتنر المتخصصة في أبحاث واستشارات تقنية المعلومات البيانات الضخمة على أنها الأصول المعلوماتية كبيرة الحجم وسريعة التدفق وكثيرة التنوع التي تتطلب طرق معالجة مجدية اقتصاديا ومبتكرة من أجل تطوير البصائر والمساعدة على اتخاذ القرارات.[2]

ويشير هذا المصطلح إلى تضخُّم حجم البيانات من ناحية عددها وسرعتها والتنوع في إنتاجها ولهذا أصبح البحث عن حلول جديدة لإدارة هذا الحجم الكبير سواءً في القدرة على التخزين أو القدرة على التحليل للاستفادة من هذه البيانات[3].

فعلى مدى السنوات الماضية، تم استخدام مصطلح البيانات الضخمة من قبل لاعبين رئيسيين مختلفين لتسمية البيانات بسمات مختلفة. تم اقتراح العديد من تعريفات البيانات الضخمة على مدار العقد الماضي، يُعرِّف Loukides البيانات الضخمة على أنها "عندما يصبح حجم البيانات نفسها جزءًا من المشكلة والتقنيات التقليدية للعمل مع البيانات تنفد من القوة".[4]

[1] مركز الإحصاء: مفاهيم عامة حول البيانات الكبيرة، أدلة المنهجية والجودة، دليل رقم 13، أبو ظبي، ص4.

[2] Douglas Laney: Big data means big business, 2013, Gartner, Inc. p1.

[3] هناء قيراطي: توظيف البيانات الضخمة في الشركات التقنية وخصوصية المستخدم، مصدر سابق، ص25.

[4] Laney, D: 3D data management: Controlling data volume, velocity, and variety. Technical Report, META Group, 2001.

ويصف جاكوبس البيانات الضخمة بأنها "البيانات التي يجبرنا حجمها على النظر إلى ما وراء الأساليب المجربة والحقيقية السائدة في ذلك الوقت". تجمع البيانات الضخمة مجموعة من تحديات إدارة البيانات للعمل مع البيانات ضمن مقاييس جديدة من الحجم والتعقيد.[1]

تشير البيانات الضخمة إلى مجموعات البيانات التي يتجاوز حجمها قدرة أدوات برامج قواعد البيانات النموذجية على الالتقاط والتخزين والإدارة والتحليل. لا يوجد تعريف صريح لحجم مجموعة البيانات التي يجب أن يتم اعتبارها بيانات ضخمة. يجب أن تكون هناك تقنية جديدة لإدارة ظاهرة البيانات الضخمة هذه. تُعرّف IDC تقنيات البيانات الضخمة بأنها جيل جديد من التقنيات والبنى المصممة لاستخراج القيمة اقتصاديًا من كميات كبيرة جدًا من مجموعة متنوعة من البيانات من خلال تمكين الالتقاط والاكتشاف والتحليل بسرعة عالية.[2] البيانات الكبيرة هي البيانات التي تتجاوز قدرة معالجة قاعدة البيانات التقليدية أنظمة. البيانات كبيرة جدًا، أو تتحرك بسرعة كبيرة، أو لا تتناسب مع هياكل هياكل قواعد البيانات الحالية. للحصول على قيمة من هذه البيانات، يجب أن تكون هناك طريقة بديلة لمعالجتها.

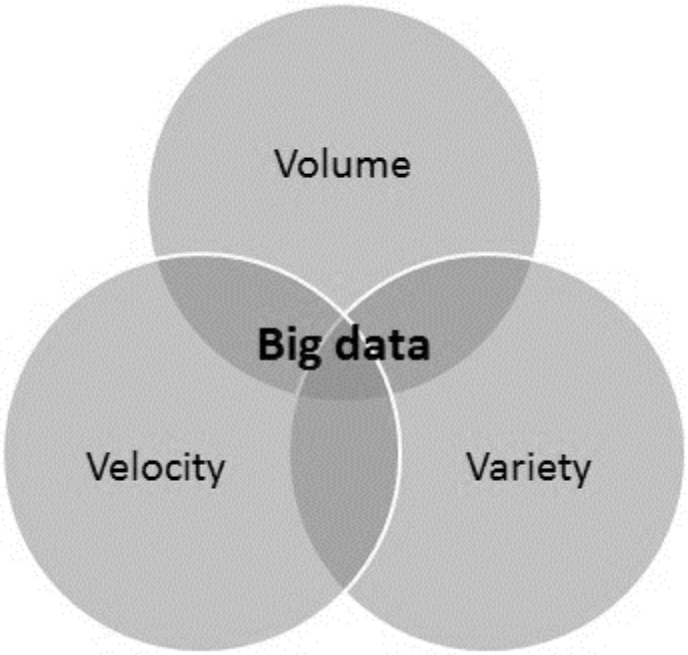
في الوقت الحاضر، أصبحت البيانات عامل إنتاج مهم يمكن مقارنته بالأصول المادية ورأس المال البشري. مع تطور الوسائط المتعددة والوسائط الاجتماعية وإنترنت الأشياء، ستجمع الشركات المزيد من المعلومات، مما يؤدي إلى ترقية الحجم الضخم وعدم تجانس البيانات الضخمة. اقترح مجتمع البحث بعض الحلول من وجهات نظر مختلفة. على سبيل المثال، تُستخدم الحوسبة السحابية لتلبية متطلبات البنية التحتية للبيانات الضخمة، ولتحسين كفاءة التكلفة والمرونة والتخفيض السلس.

المطلب الثاني: خصائص البيانات الضخمة

قال ستونبراكر أن "البيانات الضخمة يمكن أن تعني حجمًا كبيرًا أو سرعة كبيرة أو تنوعًا كبيرًا"[3]، أي أن للبيانات الضخمة ثلاث خصائص تحددها عن باقي أنواع البيانات في عالم الحوسبة، والملقبة ب 3Vs، وهي ما سنتطرق اليه في هذا الجزء من البحث:

[1] Jacobs. A: The pathologies of big data, Communications of the ACM, 2009.
[2] M.H.Padgavankar: Big Data Storage and Challenges, International Journal of Computer Science and Information Technologies, Amravati, Maharashtra, India, Vol. 5 (2), 2014. P 2219.
[3] Stonebraker, M: What does 'big data' mean? Communications of the ACM, BLOG, 2012.

الشكل أ: الخصائص الثلاث للبيانات الضخمة 3Vs.



المصدر: من اعداد الطالبين اعتمادا على مكتسبات قبلية.

الفرع الأول: الحجم

يرتبط اسم البيانات الضخمة بحد ذاته بحجم هائل. يلعب حجم البيانات دورًا مهمًا للغاية في تحديد قيمة البيانات. أيضًا، ما إذا كان يمكن اعتبار بيانات معينة بالفعل بيانات كبيرة أم لا، فهذا يعتمد على حجم البيانات. ومن ثم، فإن "الحجم" هو إحدى السمات التي يجب مراعاتها أثناء التعامل مع حلول البيانات الضخمة.

فالبيانات الضخمة هي "كميات" هائلة من البيانات التي يتم إنشاؤها من العديد من المصادر يوميًا، مثل العمليات التجارية، والآلات، ومنصات الوسائط الاجتماعية، والشبكات، والتفاعلات البشرية، وغير ذلك الكثير.

يمكن لـ Facebook إنشاء ما يقرب من مليار رسالة، 4.5 مليار مرة يتم تسجيل زر "أعجبني"، ويتم تحميل أكثر من 350 مليون مشاركة جديدة كل يوم. يمكن لتقنيات البيانات الضخمة التعامل مع كميات كبيرة من البيانات.[1]

[1] https://www.javatpoint.com/big-data-characteristics , viewed at 20:16, 18/05/2022.

الفرع الثاني: السرعة

يشير مصطلح "السرعة" إلى سرعة توليد البيانات. مدى سرعة إنشاء البيانات ومعالجتها لتلبية الطلبات، يحدد الإمكانات الحقيقية في البيانات. فتلعب السرعة دورًا رئيسيًا مقارنة بالآخرين، فلا جدوى من استثمار الكثير حتى ينتهي بك الأمر في انتظار البيانات. لذا، فإن الجانب الرئيسي للبيانات الضخمة هو توفير البيانات عند الطلب وبوتيرة أسرع.

تتعامل سرعة البيانات الكبيرة مع السرعة التي تتدفق بها البيانات من مصادر مثل العمليات التجارية وسجلات التطبيقات والشبكات ومواقع الوسائط الاجتماعية وأجهزة الاستشعار والأجهزة المحمولة وما إلى ذلك. تدفق البيانات هائل ومستمر.

تلعب السرعة دورًا مهمًا مقارنة بالآخرين. فتُنشئ السرعة التي يتم بها إنشاء البيانات في الوقت الفعلي. يحتوي على سرعات مجموعات البيانات الواردة ومعدل التغيير ودفعات النشاط. يتمثل الجانب الأساسي للبيانات الضخمة في توفير البيانات المطلوبة بسرعة.

الفرع الثالث: التنوع

يشير التنوع إلى مصادر غير متجانسة وطبيعة البيانات، سواء كانت منظمة أو غير منظمة. خلال الأيام السابقة، كانت جداول البيانات وقواعد البيانات هي المصادر الوحيدة للبيانات التي نظرت فيها معظم التطبيقات. في الوقت الحاضر، يتم أيضًا مراعاة البيانات في شكل رسائل بريد إلكتروني وصور ومقاطع فيديو وأجهزة مراقبة وملفات PDF والصوت وما إلى ذلك في تطبيقات التحليل. يطرح هذا التنوع من البيانات غير المهيكلة بعض المشكلات المتعلقة بالتخزين والتعدين وتحليل البيانات.

يمكن أن تكون البيانات الضخمة منظمة وغير منظمة وشبه منظمة والتي يتم جمعها من مصادر مختلفة. سيتم جمع البيانات فقط من قواعد البيانات والأوراق في الماضي، لكن قد تعدد مصادرها في وقتنا الحالي:

- **البيانات المنظمة:** في المخطط المهيكل، جنبًا إلى جنب مع جميع الأعمدة المطلوبة. إنه في شكل جدول. يتم تخزين البيانات المهيكلة في نظام إدارة قواعد البيانات العلائقية.
أي أنها بيانات يمكن تخزينها والوصول إليها ومعالجتها في شكل ثابت، حقق التقدم في العلوم التكنولوجية نجاحاً أكبر في تطوير تقنيات للعمل مع هذا النوع من البيانات حيث يكون الشكل معروفاً مقدما قابلا لاستخلاص القيمة منه.

- **شبه منظم:** في شبه منظم، لم يتم تعريف المخطط بشكل مناسب، على سبيل المثال، JSON وXML وCSV وTSV والبريد الإلكتروني. تم تصميم أنظمة OLTP (معالجة المعاملات عبر الإنترنت) للعمل مع البيانات شبه المنظمة. يتم تخزينها في العلاقات، أي الجداول.

  يمكن أن تحتوي البيانات شبه المنظمة على كلا النموذجين من البيانات،حيث يمكننا أن نرى البيانات شبه المنظمة على شكل منظم أو عشوائي[1]
- **البيانات غير المنظمة:** يتم تضمين جميع الملفات غير المهيكلة وملفات السجل وملفات الصوت وملفات الصور في البيانات غير المهيكلة. بعض المنظمات لديها الكثير من البيانات المتاحة، لكنهم لم يعرفوا كيفية اشتقاق قيمة البيانات لأن البيانات أولية.

  يتم تصنيف أي بيانات ذات شكل أو بنية غير معروفة على أنها بيانات غير منظمة بالإضافة إلى كونها ذات حجم ضخم، والبيانات غير المنظمة تطرح تحديات متعددة من حيث معالجتها والخروج بقيمة ذات فائد منها، مثال على ذلك مصدر بيانات غير متجانس يحوي على نصوص وصور ومقاطع فيديو حيث يبرز التحدي هنا بربط مختلف أنواع هذه البيانات ببعضها بصورة متجانسة تؤدي إلى استخلاص قيمة منها[2].

تتحدى متغيرات البيانات الضخمة أساسيات الأساليب التقنية الحالية وتتطلب أشكالًا جديدة من معالجة البيانات لتمكين اتخاذ القرار المحسن واكتشاف البصيرة وتحسين العملية. مع نضوج حقل البيانات الضخمة، أصبح صنع القرار والعمل التنظيمي يعتمد على عملية صنع المعنى وخلق المعرفة. وفيما يلي بعض الخصائص الاضافية الجديدة للبيانات الضخمة:

- **التقلب :Variability** يشير هذا إلى عدم الاتساق الذي يمكن أن تظهره البيانات في بعض الأحيان، مما يعيق عملية القدرة على التعامل مع البيانات وإدارتها بشكل فعال.
- **القيمة:Value** القيمة هي خاصية أساسية للبيانات الضخمة. ليست البيانات التي نعالجها أو نخزنها. نقوم بتخزين ومعالجة وتحليل البيانات بحيث لا يقتصر الأمر على كمية البيانات التي نخزنها أو نعالجها. إنه في الواقع مقدار البيانات القيمة والموثوقة والجديرة بالثقة التي يجب تخزينها ومعالجتها وتحليلها للعثور على رؤى.[3]
- **الصدق: Veracity** يقصد بالصدق مقدار موثوقية البيانات. له طرق عديدة لتصفية البيانات أو ترجمتها. الحقيقة هي عملية القدرة على التعامل مع البيانات وإدارتها بكفاءة. البيانات الضخمة ضرورية أيضًا في تطوير الأعمال. على سبيل المثال، منشورات Facebook مع علامات التصنيف.

[1]. مركز الإحصاء والتنافسية: مفاهيم عامة عن البيانات الضخمة 2021، مرجع سابق، ص9

[2]نفس المرجع، ص9.

[3] https://www.edureka.co/blog/big-data-characteristics/#volume viewed at 23:23, 18/05/2022.

الخلاصة:

في هذا الفصل تعرفنا عن ماهية البيانات والبيانات الضخمة التي أصبحت المجال الناشئ حيث تقدم التكنولوجيا المبتكرة طرقًا جديدة لاستخراج القيمة من تسونامي للمعلومات المتاحة. كما هو الحال مع أي منطقة ناشئة، يمكن أن تكون المصطلحات والمفاهيم مفتوحة لتفسيرات مختلفة. توضح التعريفات المختلفة "للبيانات الضخمة" التي ظهرت تنوع واستخدام المصطلح لتسمية البيانات بسمات مختلفة. كما تعرفنا عن المميزات الأساسية للبيانات الضخمة التي من خلالها يمكن استخدام أداتين من مجتمع الأعمال، سلاسل القيمة والأنظمة البيئية للأعمال، لنمذجة أنظمة البيانات الضخمة وبيئات أعمال البيانات الضخمة. يمكن لسلاسل قيمة البيانات الضخمة أن تصف تدفق المعلومات داخل نظام البيانات الضخمة كسلسلة من الخطوات اللازمة لتوليد القيمة والرؤى والمساعدة في اتخاذ القرار.

الفصل الثاني:

تحليل

البيانات

الضخمة

تمهيد:

يسعى الإنسان دائماً إلى تسهيل المهام عليه وتيسير كل عقبات الحياة، فلو نظرنا إلى أول إصدارات الحواسيب لوجدنا مساحات التخزين لديها صغيرة جداً، ناهيك عن سرعة الوصول البطيئة للبيانات بسبب ضعف أداء الحواسيب من جهة، وبسبب رداءة نظام التشغيل من جانب آخر، ولكن الإنسان بسبب ملكته الإبداعية فإنه طور ومازال يطور اداء الحاسوب آلياً وبرمجياً، حتى حصلنا على حواسيب بكفاءات عالية وبطرق فعالة وسريعة لحفظ البيانات ولاستغلالها. وفي زمننا هذا، أصبحت البيانات أكبر بكثير مما كانت عليه، يهتم تخزين البيانات الضخمة بتخزين البيانات وإدارتها بطريقة قابلة للتطوير، مما يلبي احتياجات التطبيقات التي تتطلب الوصول إلى البيانات. سيسمح نظام تخزين البيانات الكبيرة المثالي بتخزين كمية غير محدودة تقريبًا من البيانات، والتعامل مع المعدلات العالية للوصول العشوائي للكتابة والقراءة، بمرونة والتعامل بكفاءة مع مجموعة من نماذج البيانات المختلفة، ودعم كل من البيانات المنظمة وغير المنظمة، وفي هذا الفصل سنتطرق الى أبرز الطرق، التقنيات والتطبيقات التي تتم بها معالجة البيانات الضخمة وتخزينها.

المبحث الأول: تخزين البيانات الضخمة

في الوقت الحاضر، تعد البيانات الضخمة الموضوع الأهم بالنسبة للباحثين، لأنه يشير إلى كميات متزايدة بسرعة من البيانات التي تم جمعها من أجهزة غير متجانسة. تنتج شبكات الاستشعار والتجارب العلمية والمواقع الإلكترونية والعديد من التطبيقات الأخرى البيانات بتنسيقات مختلفة. الميل إلى التحول من البيانات المهيكلة إلى البيانات غير المهيكلة يجعل قواعد البيانات العلائقية التقليدية غير مناسبة للتخزين. هذا النقص في قواعد البيانات العلائقية يحفز تطوير آليات التخزين الموزعة الفعالة.

إن توفير تخزين عالي الكفاءة وموثوق وقابل للتطوير للبيانات المتزايدة ديناميكيًا هو الهدف الرئيسي في نشر أداة لتخزين البيانات الضخمة وبالتالي، فإن التطوير المبتكر لأنظمة التخزين مع أداء الوصول المحسن والتسامح مع الأخطاء مطلوب.

المطلب الأول: تطور طرق تخزين البيانات

لطالما كان تخزين البيانات مجالًا مثيرًا للقلق في مجال إدارة المعرفة. بعد وقت قصير من وضع تصور لانفجار المعلومات في الثلاثينيات، فكان التركيز مدفوعًا نحو فهم كيفية إدارة هذا النمو الدائم للبيانات والمعلومات. وبالإشارة إلى المكتبات، التي تعتبر المصدر الأول لتنظيم البيانات وتخزينها، فقد تم الإبلاغ عن علامة على زيادة سعة تخزين البيانات في عام 1944 عندما قدر فريمونت ريدر، أمين مكتبة من جامعة ويسليان، أن "مكتبات الجامعة الأمريكية تتضاعف في الحجم كل 16 عامًا".[1]

مع هذا التقدير، كان من الضروري تغيير طرق تخزين واسترجاع البيانات، ليس فقط فيما يتعلق بالمكتبات ولكن فيما يتعلق بجميع القطاعات المعنية بإدارة المعرفة. قبل تحليل أهمية التخزين الفعال للبيانات، تم توثيق فهم أفضل لتطور تخزين البيانات بإيجاز فيما يلي:

- في عام 1928، قدمت شركة IBM نسخة جديدة من البطاقة المثقبة، وتبين أنها واحدة من أهم الابتكارات التكنولوجية لشركة IBM، مما دفع الشركة إلى الصدارة في معالجة البيانات. لما يقرب من أربعة عقود، كانت الوسيلة الرئيسية لتخزين وفرز وإعداد التقارير عن البيانات التي تمت معالجتها أولاً من خلال معدات البطاقات المثقبة وأجهزة الكمبيوتر اللاحقة. في أواخر منتصف الخمسينيات من القرن الماضي، شكلت مبيعات البطاقات المثقوبة 20 في المائة من عائدات شركة IBM و 30% من أرباحها النهائية.[2]

[1] Alexander Iadarola: Lars TCF Holdhus, 2015, http://dismagazine.com/discussion/73314/tcf-data-awareness ,viewed at 18:49, 06/03/2022.

[2] IBM 100: the IBM Punched Card, https://www.ibm.com/ibm/history/ibm100/us/en/icons/punchcard/ , viewed at 19:03, 06/03/2022.

- في عام 1952، أعلنت شركة IBM عن أول وحدة تخزين شريط مغناطيسي، وأصبح الشريط المغناطيسي هو تقنية تخزين البيانات القياسية في الخمسينيات من القرن الماضي ولا يزال يستخدم بكثافة لتخزين أرشفة المحتوى في صناعة الترفيه اليوم، وكذلك لتطبيقات التخزين الرقمية العامة الأخرى حيث التكلفة المنخفضة أهم من الوصول السريع.[1]
- عام 1956 ابتكرت شركة IBM أول محرك قرص صلب ثابت قادر على استيعاب ما يصل إلى 5 ميجا بايت، ورفعها إلى 1 جيجا بايت في عام 1982، وعدد قليل من تيرابايت حاليًا.
- اخترع Alan Shugart محرك الأقراص المرنة (FDD) في شركة IBM عام 1967. استخدمت محركات الأقراص المرنة الأولى قرصًا مقاس 8 بوصات (أطلق عليه لاحقًا اسم "القرص المرن" حيث أصبح أصغر حجمًا)، والذي تطور إلى قرص مقاس 5.25 بوصة كان تم استخدامه على أول كمبيوتر شخصي من IBM في اوت 1981. وكان القرص 5.25 بوصة يحتوي على 360 كيلو بايت مقارنة بسعة 1.44 ميجا بايت للقرص المرن مقاس 3.5 بوصة الحالي.[2]
- عام 1982 تم اختراع مفهوم القرص المضغوط (CD) في اليابان، وتم تطوير القرص المضغوط في وقت لاحق بسعة تخزين من 650 ميجابايت إلى 700 ميجابايت؛ ما يعادل 450 قرصًا مرنًا.[3]
- عام 2000 ظهور محركات أقراص فلاش USB؛ على غرار الأقراص المرنة، فتحسنت سعة تخزين البيانات بمرور الوقت وتتحسن باستمرار.[4]
- عام 2010 "يُقدر أن السحابة تساهم بأكثر من 1 اكسابايت من البيانات"[5].

في أوائل الخمسينيات من القرن الماضي، طور فريتز-رودولف غونتش مفهوم الذاكرة الافتراضية، مما سمح بالتخزين المحدود ليتم التعامل معه على أنه لانهائي. سمح مفهوم Güntsch بمعالجة البيانات بدون قيود ذاكرة الأجهزة التي فرضت سابقًا تقسيم المشكلة بعبارة أخرى، كان التركيز على بنية الأجهزة وليس البيانات نفسها. اتفق ديريك برايس، عالم المعلومات المعروف باسم والد علم القياس مع بيان رايدر بشأن الحمل الزائد للتخزين من خلال الملاحظة والتعبير عن أن الكم الهائل من البحث العلمي كان أكثر من أن يواكبه البشر. أظهر تطور تخزين

[1] The Storage Engine 1951: Tape unit developed for data storage, https://www.computerhistory.org/storageengine/tape-unit-developed-for-data-storage/ , viewed at 16:39, 07/03/2022.
[2] Gray brown: History of the Floppy Disk Drive, https://computer.howstuffworks.com/floppy-disk-drive1.htm , viewed at 16:51, 07/03/2022.
[3] Philips Research: History of the CD - The CD Family, https://www.philips.com/a-w/about/innovation/research.html ,viewed at 16:58, 07/03/2022.
[4] Backupify Bit & Bytes: A History of Data Storage, https://www.backupify.com/history-of-data-storage/ ,viewed at 17 :04, 07/03/2022.
[5] Same as the previous reference.

البيانات أنه يعمل باستمرار على تحسين تقنيات التخزين، والتخزين الفعال "المزيد على القليل"، والتعامل الآن مع معظم الأجهزة الداخلية والتركيز على التخزين السحابي. ومع ذلك، قد تكون طريقة تخزين البيانات الضخمة هذه مشكلة إذا لم يتم إعدادها وفقًا لذلك.

ولهذا فان أنظمة تخزين البيانات الفعالة هي تلك التي تصنف المعلومات المستردة بتنسيق مناسب لاستخراج القيمة وتحليلها. يعتقد الباحثون في الصين أنه لكي يحدث هذا، يجب أن يوفر النظام الفرعي ما يلي:[1]

- يجب أن تستوعب البنية التحتية للتخزين المعلومات بشكل مستمر وموثوق.
- واجهة وصول قابلة للتطوير للاستعلام عن كمية كبيرة من البيانات وتحليلها.

المطلب الثاني: أساسيات تخزين البيانات الضخمة

لقد أثرت البيانات الضخمة على منظور البحث والإدارة والأعمال، وقد استحوذت على اهتمام مزودي حلول البيانات نحو نشر تقنيات مرضية لتخزين البيانات الضخمة كانت قواعد البيانات العلائقية فعالة للغاية بالنسبة لكميات مكثفة من البيانات من حيث عمليات التخزين والاسترجاع للعديد عقود. ومع ذلك، مع ظهور الإنترنت وإمكانية الوصول إليها، حولت التكنولوجيا للجمهور هيكل البيانات نحو مخطط أقل، ومترابط، وسريع النمو.[2] بصرف النظر عن ذلك، فإن تعقيد البيانات التي تم إنشاؤها بواسطة موارد الويب لا يسمح باستخدام تقنيات قواعد البيانات العلائقية لتحليل بيانات الصورة، والنمو الأسي، ونقص البنية، والتنوع في الأنواع يجلب تحديات تخزين البيانات وتحليلها لأنظمة إدارة البيانات التقليدية. يعتبر تحويل هياكل البيانات الضخمة إلى نماذج البيانات العلائقية، والمخططات العلائقية المحددة بدقة، وتعقيد الإجراءات للمهام البسيطة من السمات الصارمة لقواعد البيانات العلائقية، والتي لا تقبل البيانات الضخمة.

نتيجة لتحليل تقنيات تخزين البيانات الحالية والمستقبلية، تم الحصول على عدد من الأفكار المتعلقة بتقنيات تخزين البيانات. فأصبح من الواضح أن تخزين البيانات الضخمة أصبح عملاً تجاريًا للسلع الأساسية وأن تقنيات التخزين القابلة للتطوير قد وصلت إلى مستوى المؤسسة بحيث يمكنها إدارة أحجام غير محدودة تقريبًا من البيانات. والدليل على ذلك واضح من خلال الاستخدام الواسع النطاق للحلول المستندة إلى Hadoop التي يقدمها بائعون مثل (2014) Cloudera و (2014) Hortonworksو (2014) MapRبالإضافة إلى العديد من بائعي قواعد بياناتNoSQL2 ، لا سيما أولئك الذين يستخدمون الذاكرة المضمنة والعمودية في التخزين مقارنة بأنظمة إدارة قواعد البيانات العلائقية التقليدية التي تعتمد على التخزين المستند إلى الصفوف واستراتيجيات التخزين المؤقت

[1] Hu .h: Toward Scalable Systems for Big Data Analytics: A Technology Tutorial. Toward Scalable Systems for Big Data Analytics: A Technology Tutorial, 2014.

[2] Abdullah Gani: Big data storage technologies: a survey, Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur 50603, Malaysia, August 8, 2017, pages 1-2.

باهظة الثمن، فان تقنيات تخزين البيانات الضخمة الجديدة هذه لها قابلية تطوير أفضل مع تعقيد تشغيلي وتكاليف أقل.

على الرغم من هذه التطورات التي تعمل على تحسين الأداء وقابلية التوسع وسهولة استخدام تقنيات التخزين، لا تزال هناك إمكانات كبيرة غير مستغلة لتقنيات تخزين البيانات الضخمة، لاستخدام وتطوير التقنيات على حد سواء:[1]

- **إمكانية تغيير المجتمع والشركات عبر القطاعات:** تعد تقنيات تخزين البيانات الضخمة عاملاً تمكينيًا رئيسيًا للتحليلات المتقدمة التي لديها القدرة على تحويل المجتمع والطريقة التي يتم بها اتخاذ القرارات التجارية الرئيسية. هذا له أهمية خاصة في القطاعات التي لا تعتمد تقليديا على تكنولوجيا المعلومات مثل الطاقة. بينما تواجه هذه القطاعات مشاكل غير فنية مثل نقص المهارة وخبراء البيانات الضخمة والحواجز التنظيمية، تقنيات تخزين البيانات الجديدة لديها القدرة على تمكين تحليلات جديدة لتوليد القيمة في وعبر القطاعات الصناعية المختلفة.
- **يعد الافتقار إلى المعايير عائقًا رئيسيًا:** يعتمد تاريخ NoSQL على حل تحديات تكنولوجية محددة تؤدي إلى مجموعة من تقنيات التخزين المختلفة. مجموعة الخيارات الكبيرة إلى جانب عدم وجود معايير للاستعلام عن البيانات تجعل من الصعب تبادل مخازن البيانات لأنها قد تربط شفرة تطبيق معين بحل تخزين معين.
- **تحديات قابلية التوسع المفتوحة في مخازن البيانات المستندة إلى الرسم البياني:** تعد معالجة البيانات استنادًا إلى هياكل بيانات الرسم البياني مفيدة في عدد متزايد من التطبيقات الذي يسمح بالتقاط الدلالات والعلاقات المعقدة بشكل أفضل مع أجزاء أخرى من المعلومات القادمة من مجموعة كبيرة ومتنوعة من مصادر البيانات المختلفة، ولديه القدرة على تحسين القيمة الإجمالية التي يمكن توليدها من خلال تحليل البيانات. بينما يتم استخدام قواعد بيانات الرسم البياني بشكل متزايد لهذا الغرض، لا يزال من الصعب توزيع بنية البيانات القائمة على الرسم البياني بكفاءة عبر عقد الحوسبة.
- **تخلف نظم الخصوصية والأمان:** على الرغم من وجود العديد من المشاريع والحلول التي تتناول الخصوصية والأمان، فإن حماية الأفراد وتأمين بياناتهم تبقى دائما غير مواكبة للتقدم التكنولوجي السريع لأنظمة تخزين البيانات. فالمطلوب بحث كبير لفهم أفضل لكيفية إساءة استخدام البيانات، وكيف يجب حمايتها ودمجها في حلول تخزين البيانات الضخمة.

[1] José María Cavanillas: New Horizons for a Data-Driven Economy a Roadmap for Usage and Exploitation of Big Data in Europe, Springer Open, Springer International Publishing AG Switzerland, 2016, pages 121-122.

كما تظهر تقنيات البيانات الضخمة الناشئة واستخدامها في القطاعات المختلفة، القدرة على تخزين وإدارة وتحليل كميات كبيرة من تلميحات البيانات غير المتجانسة نحو ظهور مجتمع واقتصاد يحركهما البيانات مع إمكانات تحويلية هائلة. يمكن للمؤسسات الآن تخزينها وتحليلها للمزيد من البيانات بتكلفة أقل مع تحسين قدراتها التحليلية في نفس الوقت. في حين أن شركات مثل Google وTwitter وFacebook هي جهات فاعلة تشكل البيانات الأصل الرئيسي لها، تميل القطاعات الأخرى أيضًا إلى أن تصبح أكثر اعتمادًا على البيانات. على سبيل المثال، يعد قطاع الصحة مثالًا ممتازًا يوضح كيف يمكن للمجتمع أن يتوقع خدمات صحية أفضل من خلال تكامل وتحليل أفضل للبيانات المتعلقة بالصحة.[1]

المطلب الثالث: نظم تخزين البيانات الضخمة والتقنيات

خلال العقد الماضي، أدت الحاجة إلى التعامل مع انفجار البيانات وتحول الأجهزة من نهج التوسع إلى توسيع النطاق إلى انفجار أنظمة تخزين البيانات الضخمة الجديدة التي ابتعدت عن نماذج قواعد البيانات العلائقية التقليدية. النمو الهائل للبيانات له متطلبات أكثر صرامة على التخزين والإدارة. في هذا القسم، نركز على نظم تخزين البيانات الضخمة والتقنيات المستعملة في ذلك.

الفرع الأول: أنظمة تخزين البيانات الضخمة

يشير تخزين البيانات الضخمة إلى تخزين وإدارة مجموعات البيانات واسعة النطاق أثناء تحقيق الموثوقية وتوافر الوصول إلى البيانات. سنراجع القضايا المهمة بما في ذلك أنظمة التخزين الضخمة وأنظمة التخزين الموزعة وآليات تخزين البيانات الضخمة. من ناحية أخرى، تحتاج البنية التحتية للتخزين إلى توفير خدمة تخزين المعلومات مع مساحة تخزين موثوقة؛ من ناحية أخرى، يجب أن يوفر واجهة وصول قوية للاستعلام عن كمية كبيرة من البيانات وتحليلها. تقليديًا، كمعدات مساعدة للخادم، يتم استخدام جهاز تخزين البيانات لتخزين البيانات وإدارتها والبحث عنها وتحليلها باستخدام أنظمة RDBMS المنظمة. مع النمو الحاد للبيانات، أصبحت أجهزة تخزين البيانات أكثر أهمية بشكل متزايد، وتسعى العديد من شركات الإنترنت إلى تحقيق سعة تخزين كبيرة لتكون قادرة على المنافسة. لذلك، هناك حاجة ملحة للبحث في تخزين البيانات.

نظام تخزين للبيانات الضخمة. تظهر أنظمة تخزين مختلفة لتلبية متطلبات البيانات الضخمة. يمكن تصنيف تقنيات التخزين الضخمة الحالية على أنها تخزين مرفق مباشر (DAS) وتخزين شبكة، بينما يمكن تصنيف تخزين الشبكة إلى التخزين المتصل بالشبكة (NAS) وشبكة منطقة التخزين (SAN). في DAS، ترتبط العديد من الأقراص الصلبة بشكل مباشر بالخوادم، وتكون إدارة البيانات تتمحور حول الخادم، مثل أن أجهزة التخزين عبارة

[1] M.H.Padgavankar: Big Data Storage and Challenges, previous reference, P 2221.

عن معدات طرفية، يأخذ كل منها قدرًا معينًا من موارد الإدخال / الإخراج ويتم إدارتها بواسطة برنامج تطبيق فردي. لهذا السبب، فإن DAS مناسب فقط لربط الخوادم على نطاق صغير. ومع ذلك، نظرًا لقابلية التوسع المنخفضة، ستظهر DAS كفاءة غير مرغوب فيها عند زيادة سعة التخزين، أي أن قابلية الترقية وقابلية التوسع محدودة للغاية. وبالتالي، يتم استخدام DAS بشكل أساسي في أجهزة الكمبيوتر الشخصية والخوادم صغيرة الحجم.[1]

يستخدم التخزين الشبكي الشبكة لتزويد المستخدمين بواجهة موحدة للوصول إلى البيانات ومشاركتها. تشتمل معدات تخزين الشبكة على معدات تبادل البيانات الخاصة، ومجموعة الأقراص، ومكتبة النقر، ووسائط التخزين الأخرى، فضلاً عن برامج التخزين الخاصة. يتميز بقابلية توسعة قوية.

NAS هي في الواقع معدات تخزين مساعدة للشبكة. إنه متصل مباشرة بشبكة من خلال محور أو تبديل عن طريق بروتوكولات TCP / IP. حيث في NAS، يتم نقل البيانات في شكل ملفات. مقارنةً بـ DAS، يتم تقليل عبء الإدخال / الإخراج في خادم NAS بشكل كبير نظرًا لأن الخادم يصل إلى جهاز تخزين بشكل غير مباشر عبر الشبكة.

في حين أن NAS موجهة نحو الشبكة، فإن SAN مصممة خصيصًا لتخزين البيانات بشبكة مكثفة ذات عرض نطاق ترددي وقابلة للتطوير، على سبيل المثال، شبكة عالية السرعة مع اتصالات الألياف الضوئية. حيث في SAN، تكون إدارة تخزين البيانات مستقلة نسبيًا داخل شبكة منطقة تخزين محلية، حيث يتم استخدام تبديل البيانات متعدد المسارات بين أي عقد داخلية لتحقيق أقصى درجة من مشاركة البيانات وإدارة البيانات.

من أساسيات نظام تخزين البيانات، يمكن تقسيم كل من DAS وNAS وSAN إلى ثلاثة أجزاء: (1) مجموعة الأقراص: إنها أساس نظام التخزين والضمان الأساسي لتخزين البيانات؛ (2) أنظمة الاتصال والشبكة الفرعية: والتي توفر الاتصال بين مصفوفات الأقراص والخوادم؛ (3) برنامج إدارة التخزين: الذي يتعامل مع مشاركة البيانات والتعافي من الكوارث ومهام إدارة التخزين الأخرى لخوادم متعددة.

الفرع الثاني: تقنيات تخزين البيانات الضخمة

عادةً ما تضحي هذه الأساليب بخصائص مثل تناسق البيانات من أجل الحفاظ على استجابات استعلام سريعة مع زيادة كميات البيانات. تُستخدم مخازن البيانات الضخمة بطرق مماثلة لأنظمة إدارة قواعد البيانات العلائقية التقليدية، على سبيل المثال لحلول معالجة المعاملات عبر الإنترنت (OLTP) ومستودعات البيانات عبر البيانات المنظمة أو شبه المنظمة. توجد نقاط قوة خاصة في التعامل مع البيانات غير المهيكلة وشبه المنظمة على نطاق واسع.

[1] M.H.Padgavankar: Big Data Storage and Challenges, Previous reference, p 2220.

يقيِم هذا الفرع أحدث ما توصلت إليه تقنيات تخزين البيانات القادرة على التعامل مع كميات كبيرة من البيانات، ويحدد الاتجاهات ذات الصلة بمخزن البيانات. فيما يلي أنواع مختلفة من أنظمة التخزين:

1) قواعد بيانات NoSQL:

ربما تكون أهم عائلة من تقنيات تخزين البيانات الضخمة هي أنظمة إدارة قواعد بيانات NoSQL. تستخدم قواعد بيانات NoSQL نماذج بيانات من خارج العالم العلائقي لا تلتزم بالضرورة بخصائص المعاملات الخاصة بالذرية والاتساق والعزلة والمتانة (ACID).

تم تصميم قواعد بيانات NoSQL من أجل قابلية التوسع، غالبًا عن طريق التضحية بالاتساق. بالمقارنة مع قواعد البيانات العلائقية، فإنها غالبًا ما تستخدم واجهات استعلام منخفضة المستوى وغير قياسية، مما يزيد من صعوبة دمجها في التطبيقات الحالية التي تتوقع واجهة SQL. يؤدي عدم وجود واجهات قياسية إلى صعوبة التبديل بين البائعين. يمكن تمييز قواعد بيانات NoSQL بنماذج البيانات التي تستخدمها:

- **مخازن القيمة الرئيسية:** تسمح مخازن القيمة الرئيسية بتخزين البيانات بطريقة بدون مخطط. يمكن أن تكون كائنات البيانات غير منظمة أو منظمة تمامًا ويمكن الوصول إليها عن طريق مفتاح واحد. نظرًا لعدم استخدام مخطط، فليس من الضروري حتى أن تشترك كائنات البيانات في نفس البنية.
- **المخازن العمودية:** نظام DBMS الموجه نحو الأعمدة هو نظام إدارة قواعد البيانات (DBMS) الذي يخزن جداول البيانات كأقسام من أعمدة البيانات بدلاً من صفوف من البيانات، مثل معظم أنظمة DBMS العلائقية". قواعد البيانات هذه عادةً ما تكون خرائط متفرقة وموزعة ومستمرة ومتعددة الأبعاد يتم فرزها حيث يتم فهرسة البيانات بواسطة ثلاثة أضعاف مفتاح الصف ومفتاح العمود والطابع الزمني. يتم تمثيل القيمة كنوع بيانات سلسلة غير متقطع. يتم الوصول إلى البيانات بواسطة مجموعات الأعمدة، أي مجموعة من مفاتيح الأعمدة ذات الصلة التي تضغط بشكل فعال البيانات المتفرقة في الأعمدة. يتم إنشاء مجموعات الأعمدة قبل تخزين البيانات ومن المتوقع أن يكون عددها صغيرًا. في المقابل، فإن عدد الأعمدة غير محدود. من حيث المبدأ، تكون المخازن العمودية أقل ملاءمة عند الحاجة إلى الوصول إلى جميع الأعمدة. ومع ذلك، نادرًا ما يكون هذا هو الحال في الممارسة العملية، مما يؤدي إلى أداء متفوق للمخازن العمودية.[1]
- **قواعد بيانات المستندات:** على عكس القيم الموجودة في مخزن القيمة الرئيسية، يتم تنظيم المستندات. ومع ذلك، لا توجد متطلبات لمخطط مشترك يجب أن تلتزم به جميع المستندات كما في حالة السجلات في

[1] http://virtuoso.openlinksw.com/dataspace/doc/dav/wiki/Main/VOSArticleVirtuosoAHybridRDBMSGraphColumnStore ,viewed at 1:00, 13/03/2022.

قواعد البيانات العلائقية. وبالتالي، يشار إلى قواعد بيانات المستندات على أنها تخزين البيانات شبه المنظمة. على غرار مخازن القيمة الرئيسية، يمكن الاستعلام عن المستندات باستخدام مفتاح فريد. ومع ذلك، من الممكن الوصول إلى المستندات عن طريق الاستعلام عن هيكلها الداخلي، مثل طلب جميع المستندات التي تحتوي على حقل بقيمة محددة. تعتمد قدرة واجهة الاستعلام عادةً على تنسيق الترميز المستخدم بواسطة قواعد البيانات. تتضمن الترميزات الشائعة XML أو JSON.

- **قواعد بيانات الرسم البياني:** قواعد بيانات الرسم البياني، مثل (Neo4J 2015)، تخزن البيانات في هياكل الرسوم البيانية مما يجعلها مناسبة لتخزين البيانات الترابطية للغاية مثل الرسوم البيانية للشبكات الاجتماعية. هناك نكهة خاصة لقواعد بيانات الرسم البياني هي المتاجر الثلاثية مثل AllegroGraph وVirtuoso المصممة خصيصًا مصممة لتخزين ثلاث مرات RDF. ومع ذلك، فإن تقنيات المتاجر الثلاثية الحالية ليست مناسبة حتى الآن لتخزين مجموعات البيانات الكبيرة حقًا بكفاءة.[1]

وبشكل عام، فإن حجم مخازن بيانات NoSQL أفضل من قواعد البيانات العلائقية، تقل قابلية التوسع مع زيادة تعقيد نموذج البيانات المستخدم بواسطة مخزن البيانات. ينطبق هذا بشكل خاص على قواعد بيانات الرسم البياني التي تدعم التطبيقات التي تقوم بالكتابة والقراءة بشكل مكثف. تتمثل إحدى طرق تحسين الوصول للقراءة في تقسيم الرسم البياني إلى رسوم بيانية فرعية متصلة بالحد الأدنى بين بعضها البعض وتوزيع هذه الرسوم البيانية الفرعية بين العقد الحسابية. ومع ذلك، مع إضافة حواف جديدة إلى الرسم البياني، قد يزداد الاتصال بين الرسوم البيانية الفرعية بشكل كبير. قد يؤدي هذا إلى زيادة زمن انتقال الاستعلام بسبب زيادة حركة مرور الشبكات والحسابات غير المحلية. لذلك يجب أن تراعي مخططات التجزئة الفعالة مقدار الحمل المطلوب لإعادة التوزيع الديناميكي لبيانات الرسم البياني.

2) قواعد بيانات NewSQL:

قواعد بيانات NewSQL هي شكل حديث من قواعد البيانات العلائقية التي تهدف إلى قابلية قابلة للمقارنة مع قواعد بيانات NoSQL مع الحفاظ على ضمانات المعاملات التي تقدمها أنظمة قواعد البيانات التقليدية ,تم استخدام المصطلح لأول مرة من قبل محلل مجموعة 451 ماثيو أسليت في ورقة بحثية عام 2011 تناقش ظهور جيل جديد من أنظمة إدارة قواعد البيانات. كان نظام قاعدة البيانات المتوازية H-Store من أوائل أنظمة NewSQL. لديها الخصائص التالية:

- SQLهي الآلية الأساسية لتفاعل التطبيق

[1] http://franz.com/agraph/allegrograph/, viewed at 1:15 13/03/2022.

- دعم ACID للمعاملات
- آلية تحكم التزامن غير مقفل
- بنية توفر أداءً أعلى بكثير لكل عقدة
- بنية قابلة للتوسيع، لا تشارك في أي شيء، قادرة على العمل على عدد كبير من العقد دون المعاناة من الاختناقات

من المتوقع أن تكون أنظمة NewSQL أسرع بحوالي 50 مرة من OLTP RDBMS التقليدية. على سبيل المثال، يتوسع (VoltDB 2014) بشكل خطي في حالة الاستعلامات غير المعقدة (قسم واحد) ويوفر دعم ACID. يتسع لعشرات العقد حيث يتم تقييد كل عقدة بحجم الذاكرة الرئيسية.

فيما يلي عرض لأفضل قواعد بيانات NewSQL الموجودة حاليًا في السوق. القائمة ليست شاملة، لذا ابحث أكثر إذا كنت تخطط لاستخدام إحدى قواعد البيانات:[1]

- **VoltDB:** يعمل VoltDB بشكل جيد مع تطبيقات المعاملات عالية السرعة. تقوم قاعدة البيانات بمعالجة الذاكرة على بنية موزعة. البرنامج متاح كمصدر مفتوح وملك. من أهم خصائصه (اتخاذ القرار في الوقت الحقيقي. دعم استيراد وتصدير كافكا. التعافي من الكوارث من خلال نسخ قاعدة البيانات. تكامل تصدير Hadoop وOLAP.)
- **CockroachDB:** هي قاعدة بيانات قوية وقابلة للتطوير. توفر قاعدة البيانات تناسقًا قويًا للبيانات وتعمل بشكل جيد مع الموارد ذات زمن الانتقال المنخفض. من أهم خصاءصه (نظام قوي للتعافي من الكوارث. عرض البيانات التاريخية، والتسجيل، وخيارات التخزين. عمليات تنظيف مدمجة للأقراص وأجهزة التخزين. يعمل CockroachDB في ظروف غير مواتية.)
- **NuoDB:** هي قاعدة بيانات موزعة جغرافيًا مع تحجيم مرن لمواقع جغرافية مختلفة. تقوم قاعدة البيانات بتعيين البيانات عبر نقاط مختلفة مع الحفاظ على توافقها مع ACID. من مميزاته (تحويلات البيانات عالية الجودة. متوفر دائمًا مع تطورات المخطط عبر الإنترنت والترقيات المتجددة. ميزات مخصصة لتخزين البيانات والتحكم فيها. دعم كامل لمعاملات ACID.)
- **ClustrixDB:** هي قاعدة بيانات NewSQL ذاتية الإدارة. يقوم البرنامج بأتمتة عمليات القياس ويدعم التوافر العالي وتصنيف البيانات الفعال. من مميزاته (خيارات ترحيل كود SQL. المقاييس الصحية المضمنة في واجهة المتصفح. مساعدة DevOps وذاكرة التخزين المؤقت للاستعلام).

[1] https://phoenixnap.com/kb/newsql, viewed at 11:15, 15/03/2022

- **Altibase:** Altibase هي قاعدة بيانات في الذاكرة ذات بنية هجينة. تعمل قاعدة البيانات على تقليل تكاليف الأجهزة والبرامج من خلال الجمع بين معالجة البيانات في الذاكرة ونظام إدارة قواعد البيانات على القرص بترخيص واحد. من مميزاته (Altibase يأتي في إصدارات المجتمع والملكية. محرك ذاكرة مُحسّن لسرعات أعلى. الثبات المخصص ومستويات توازن الأداء. خيارات نشر مرنة. الوصول في الوقت الحقيقي إلى البيانات الحيوية.)

3) منصات استعلام البيانات الضخمة

توفر الأنظمة الأساسية لاستعلام البيانات الضخمة واجهات استعلام أعلى مخازن البيانات الضخمة الأساسية التي تبسط الاستعلام عن مخازن البيانات الأساسية. إنها تقدم عادةً واجهة استعلام تشبه SQL للوصول إلى البيانات، ولكنها تختلف في نهجها وأدائها. يوفر Hive فكرة تجريدية أعلى نظام الملفات الموزعة (Hadoop HDFS) الذي يسمح بالاستعلام عن الملفات المهيكلة بلغة استعلام تشبه SQL. تنفذ الخلية الاستعلامات من خلال ترجمة الاستعلامات في وظائف MapReduce. نتيجة لذلك، تتمتع استعلامات Hive بزمن انتقال عالٍ حتى بالنسبة لمجموعات البيانات الصغيرة.[1] تشمل مزايا Hive واجهة استعلام تشبه SQL والمرونة لتطوير المخططات بسهولة. هذا ممكن حيث يتم تخزين المخطط بشكل مستقل عن البيانات ويتم التحقق من صحة البيانات فقط في وقت الاستعلام. يشار إلى هذا الأسلوب باسم قراءة المخطط مقارنة بنهج المخطط عند الكتابة لقواعد بيانات SQL. لذلك فإن تغيير المخطط هو عملية رخيصة نسبيًا. متجر Hadoop العمودي HBase مدعوم أيضًا بواسطة Hive.

على عكس Hive، تم تصميم Impala لتنفيذ الاستعلامات مع زمن انتقال منخفض. يعيد استخدام نفس البيانات الوصفية وواجهة المستخدم الشبيهة بـ SQL مثل Hive ولكنه يستخدم محرك الاستعلام الموزع الخاص به والذي يمكنه تحقيق زمن انتقال أقل. كما أنه يدعم HDFS وHBase كمخازن بيانات أساسية.

Spark SQL هي واجهة استعلام أخرى ذات زمن انتقال منخفض تدعم واجهة Hive. يدعي المشروع أنه "يمكنه تنفيذ استعلامات Hive QL حتى 100 مرة أسرع من Hive دون أي تعديل على البيانات أو الاستفسارات الموجودة". يتم تحقيق ذلك من خلال تنفيذ الاستعلامات باستخدام إطار عمل Spark بدلاً من إطار عمل MapReduce Hadoop.[2]

أخيرًا، يعد Drill تطبيقًا مفتوح المصدر لبرنامج Dremel الخاص بـ Google والذي تم تصميمه على غرار Impala كنظام استعلام مخصص وتفاعلي وقابل للتطوير للبيانات المتداخلة. يوفر Drill لغة استعلام تشبه SQL

[1] Thusoo, A Sarma: Hive – a warehousing solution over a map-reduce framework. Statistics and Operations Research Transactions, 2,2009, pp.1626–1629

[2] Shenker, S Stoicam: SQL and rich analytics at scale. In Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data, 2013, pp. 13–24.

الخاصة به وهو DrQL متوافق مع Dremel، ولكنه مصمم لدعم لغات الاستعلام الأخرى مثل Mongo Query Language. على عكس Hive وImpala، فإنه يدعم مجموعة من مصادر البيانات التي لا تحتوي على مخطط، مثل قواعد بيانات HDFS وHBase وCassandra وMongoDB وSQL.[1]

4) سحابة التخزين (Cloud Storage):

مع تزايد شعبية الحوسبة السحابية، ينمو تأثيرها على البيانات الضخمة أيضًا. بينما تقوم Amazon وMicrosoft وGoogle بالبناء على الأنظمة الأساسية السحابية الخاصة بها، فإن الشركات الأخرى بما في ذلك IBM وHP وDell وCisco وRackspace وما إلى ذلك، تبني اقتراحها حول OpenStack، وهو نظام أساسي مفتوح المصدر لبناء أنظمة السحابة. وفقًا لـ IDC بحلول عام 2020، "ستتأثر" 40٪ من الكون الرقمي بالحوسبة السحابية "، و" ربما يتم الاحتفاظ بنسبة تصل إلى 15٪ في السحابة "[2].

يمكن للمؤسسات والمستخدمين النهائيين استخدام السحابة بشكل عام، وخاصة التخزين السحابي. بالنسبة للمستخدمين النهائيين، يتيح تخزين بياناتهم في السحابة الوصول إليها من كل مكان ومن كل جهاز بطريقة موثوقة. بالإضافة إلى ذلك، يمكن للمستخدمين النهائيين استخدام التخزين السحابي كحل بسيط للنسخ الاحتياطي عبر الإنترنت لبيانات سطح المكتب الخاصة بهم. وبالمثل بالنسبة للمؤسسات، يوفر التخزين السحابي وصولاً مرنًا من مواقع متعددة وسعة نطاق سريعة وسهلة بالإضافة إلى أسعار تخزين أرخص ودعم أفضل استنادًا إلى وفورات الحجم (CloudDrive 2013) مع فعالية التكلفة المرتفعة بشكل خاص في بيئة تكون فيها المؤسسة تتغير احتياجات التخزين بمرور الوقت صعودًا وهبوطًا. يمكن التمييز بين حلول التخزين السحابي تقنيًا بين تخزين الكائنات والكتل. تخزين الكائنات "هو مصطلح عام يصف أسلوب معالجة وحدات التخزين المنفصلة التي تسمى". في المقابل، يتم تخزين بيانات التخزين الكتلي في وحدات تخزين يشار إليها أيضًا باسم الكتل...تعمل كل كتلة كمحرك أقراص ثابت فردي"[3] وتتيح الوصول العشوائي إلى أجزاء وأجزاء من البيانات، وبالتالي تعمل بشكل جيد مع تطبيقات مثل قواعد البيانات. بالإضافة إلى تخزين العناصر والكتل، توفر الأنظمة الأساسية الرئيسية دعمًا للتخزين المستند إلى قاعدة البيانات العلائقية وغير العلائقية بالإضافة إلى التخزين في الذاكرة وتخزين قائمة الانتظار. في التخزين السحابي، هناك اختلافات كبيرة يجب أخذها في الاعتبار في مرحلة تخطيط التطبيق:

[1] Melnik, S Garcia-Molina: Similarity flooding: A versatile graph matching algorithm and its application to schema matching. In Proceedings of the 18th International Conference Data Engineering. IEEE Computer Society, 2012, pp. 117–128.

[2] http://www.1cloudroad.com/is-enterprise-cloud-storage-a-good-fit-for-your-business; viewed at 10 :31, 17/03/2022.

[3] http://searchstorage.techtarget.com/definition/block-storage , viewed at 17:54, 19/03/2022.

- نظرًا لأن التخزين السحابي عبارة عن خدمة، فإن التطبيقات التي تستخدم هذا التخزين تتمتع بقدر أقل من التحكم وقد تواجه أداءً منخفضًا نتيجة للشبكات. يجب أن تؤخذ هذه الفروق في الأداء في الاعتبار أثناء مراحل التصميم والتنفيذ.
- الأمن هو أحد الاهتمامات الرئيسية المتعلقة بالسحب العامة. ونتيجة لذلك، يتوقع كبير مسؤولي التكنولوجيا في أمازون أنه في غضون خمس سنوات، سيتم تشفير جميع البيانات الموجودة في السحابة افتراضيًا.
- تدعم السحب الغنية بالميزات مثل AWS معايرة الكمون والتكرار ومستويات الإنتاجية للوصول إلى البيانات، مما يسمح للمستخدمين بالعثور على المفاضلة الصحيحة بين التكلفة والجودة.

هناك مشكلة أخرى مهمة عند التفكير في التخزين السحابي وهي نموذج الاتساق المدعوم (وما يرتبط به من قابلية التوسع والتوافر وتحمل القسم وزمن الانتقال). بينما تدعم خدمة التخزين البسيط (S3) من Amazon الاتساق النهائي، تدعم تخزين Microsoft Azure blob الاتساق القوي وفي نفس الوقت التوافر العالي وتحمل القسم. تستخدم Microsoft طبقتين: (1) طبقة تيار "توفر توفرًا عاليًا في مواجهة تقسيم الشبكة وغيرها من حالات الفشل"، و (2) طبقة التقسيم التي "توفر ضمانات تناسق قوية"[1].

المبحث الثاني: تحليل البيانات الضخمة

تأتي البيانات في أشكال عديدة، وأحد الأبعاد التي يجب مراعاتها ومقارنة تنسيقات البيانات المختلفة هو مقدار البنية الواردة فيها. كلما زادت بنية مجموعة البيانات، زادت قابليتها للمعالجة الآلية. في أقصى الحدود، ستمكّن التمثيلات الدلالية من التفكير الآلي. تحليل البيانات الضخمة هو المجال الفرعي للبيانات الضخمة المعني بإضافة هيكل إلى البيانات لدعم اتخاذ القرار وكذلك دعم سيناريوهات الاستخدام الخاصة بالمجال. يلخص هذا المبحث الرؤى الرئيسية، وأحدث ما توصلت إليه التكنولوجيا، والاتجاهات الناشئة، والمتطلبات المستقبلية، ودراسات الحالة القطاعية لتحليل البيانات.

المطلب الأول: أساسيات تحليل البيانات الضخمة وطرق الاستدلال والتعلم الآلي

يتمثل التحدي الأكبر لمعظم الصناعات الآن في دمج تقنيات البيانات الضخمة في عملياتها وبنيتها التحتية. تطبق الصناعة اليوم التعلم الآلي واسع النطاق وخوارزميات أخرى لتحليل مجموعات البيانات الضخمة، جنبًا إلى جنب

[1] Calder, B: Windows azure storage: a highly available cloud storage service with strong consistency. In: Proceedings of the Twenty-Third ACM Symposium on Operating Systems Principles, 2013, pp.143–157.

مع معالجة الأحداث المعقدة ومعالجة الدفق للتحليلات في الوقت الفعلي.

الفرع الأول: رؤى أساسية في تحليل البيانات الضخمة

تحدد العديد من الشركات الحاجة إلى إجراء تحليل للبيانات الضخمة، ولكن ليس لديها الموارد اللازمة لإنشاء بنية تحتية لتحليل خط أنابيب التحليلات وصيانته ستساعد زيادة بساطة التكنولوجيا في معدل التبني. علاوة على ذلك، يجب بناء مجموعة كبيرة من المعرفة بالمجال داخل كل صناعة حول كيفية استخدام البيانات: ما هي القيمة التي يجب استخراجها وما هي المخرجات التي يمكن استخدامها في العمليات اليومية.

- **البساطة في التحليل:** تشير بساطة تقنيات البيانات الضخمة إلى مدى سهولة قدرة المطورين على اكتساب التكنولوجيا واستخدامها في بيئتهم الخاصة. البساطة مهمة لأنها تؤدي إلى زيادة قابلية تبني التكنولوجيا. حدد العديد من الخبراء الدور الحاسم للبساطة في تقنيات البيانات الضخمة الحالية والمستقبلية. يرجع نجاح Hadoop وMapReduce أساسًا إلى بساطتهما تتوفر منصات البيانات الضخمة الأخرى التي يمكن اعتبارها أكثر قوة، ولكن لديها مجتمع أصغر من المستخدمين لأن اعتمادها يصعب إدارته. وبالمثل، تم الإبلاغ عن تقنيات البيانات المرتبطة، على سبيل المثال، RDF SPARQL، على أنها معقدة للغاية وتحتوي على منحنى تعليمي شديد الانحدار. يبدو أن مثل هذه التقنيات مفرطة التصميم ومعقدة للغاية ولا تصلح إلا للمتخصصين. بشكل عام، توجد بعض التقنيات الناضجة جدًا لتحليلات البيانات الضخمة، ولكن يجب تصنيع هذه التقنيات وجعلها في متناول الجميع. ستساعد منصة البيانات الضخمة سهلة الاستخدام في تبني الصناعات غير التقنية لتقنيات البيانات الضخمة[1].

- **تنقيب بيانات البث:** هذا مطلوب للتعامل مع كميات كبيرة من بيانات التدفق التي ستأتي من شبكات الاستشعار أو الأنشطة عبر الإنترنت من أعداد كبيرة من المستخدمين. ستسمح هذه القدرة للمؤسسات بتوفير تخصيص دقيق وقابل للتكيف بدرجة عالية.
- **اكتشاف البيانات**: الأسئلة المتكررة التي يطرحها المستخدمون والمطورون هي: أين يمكننا الحصول على البيانات حول X وأين يمكننا الحصول على معلومات عن Y. فمن الصعب العثور على البيانات وغالباً ما تكون البيانات التي تم العثور عليها قديمة وليست بالتنسيق الصحيح. هناك حاجة إلى برامج الزحف للعثور على مجموعات البيانات الكبيرة، والبيانات الوصفية للبيانات الضخمة، والروابط ذات المغزى بين مجموعات البيانات ذات الصلة، وآلية تصنيف مجموعة البيانات التي تعمل بالإضافة إلى تصنيف الصفحة لمستندات الويب.

[1] http://hadoop.apache.org/, viewed at 15:12, 20/03/2022.

- **التقنيات القديمة المطبقة في سياق جديد:** يتم تطبيق تقنيات فردية ومجموعات من التقنيات القديمة في سياق البيانات الضخمة. الفرق هو المقياس (الحجم) ومقدار عدم التجانس المصادف (التنوع). على وجه التحديد، في سياق الويب، يتم التركيز على مجموعات البيانات الكبيرة القائمة على المعنى مثل Freebase وعلى استخراج بيانات عالية الجودة من الويب. إلى جانب الحجم، هناك حداثة في حقيقة أن هذه التقنيات تأتي معًا في نفس الوقت.

- **التعامل مع كل من البيانات واسعة النطاق والمحددة للغاية:** الميزة القريبة لاستخراج المعلومات من الويب هي أن الويب يدور حول كل شيء، لذا فإن التغطية واسعة النطاق. قبل الويب، كان التركيز على مجالات محددة عند إنشاء قواعد البيانات وقواعد المعرفة. لم يعد من الممكن القيام بذلك في سياق الويب. تم تغيير المفهوم الكامل لـ "تصور المجال": الآن المجال هو كل شيء في العالم. على الجانب الإيجابي، تتمثل الفائدة في أنك تحصل على قدر كبير من الاتساع، والتحدي البحثي هو كيف يمكن للمرء أن يتعمق في مجال ما مع الحفاظ على السياق الواسع.
- **الأنظمة البيئية المبنية حول مجموعات الأدوات:** لها تأثير كبير غالبًا ما تكون مدفوعة من قبل الشركات الكبيرة حيث يتم إنشاء تقنية لحل مشكلة داخلية ثم يتم التخلي عنها. Apache Cassandra3 هو مثال على ذلك تم تطويره في البداية بواسطة Facebook لتشغيل ميزة البحث في البريد الوارد حتى عام 2010. وربما يكون النظام البيئي حول Hadoop هو الأكثر شهرة.

ستشارك المجتمعات مع البيانات الضخمة في جميع مراحل سلسلة القيمة وبطرق متنوعة. على وجه الخصوص، ستشارك المجتمعات بشكل وثيق في جمع البيانات، وتحسين دقة البيانات واستخدام البيانات. ستعمل البيانات الضخمة أيضًا على تعزيز مشاركة المجتمع في المجتمع بشكل عام. يصف قسم البيع بالتجزئة للمتطلبات المستقبلية والاتجاهات الناشئة مثالاً على ذلك. أطلقت O2 UK بالاشتراك مع Telef onica Digital مؤخرًا خدمة تقوم بتعيين بيانات الهاتف المحمول وإعادة توظيفها في صناعة البيع بالتجزئة. تتيح هذه الخدمة لتجار التجزئة التخطيط لمكان منافذ البيع بالتجزئة بناءً على الحركة اليومية للعملاء المحتملين. تسلط هذه الخدمة الضوء على أهمية البيانات الضخمة الداخلية (في هذه الحالة سجلات المحمول) التي يتم دمجها لاحقًا مع مصادر البيانات الخارجية (البيانات الجغرافية وبيانات التفضيل) لإنشاء أنواع جديدة من الأعمال. بشكل عام، سيؤدي تجميع البيانات عبر المؤسسات وعبر القطاعات إلى تعزيز القدرة التنافسية للصناعة الأوروبية.

الفرع الثاني: الاستدلال في البيانات الضخمة

تم العثور أيضًا على أن الاتجاهات الحالية بشأن البيانات المرتبطة والتقنيات الدلالية والتفكير على نطاق واسع هي بعض الموضوعات التي أبرزها الخبراء الذين تمت مقابلتهم فيما يتعلق بتحديات البحث الرئيسية

والمتطلبات التكنولوجية الرئيسية للبيانات الضخمة. يقدم هذا الفرع مراجعة حديثة فيما يتعلق بتحليل البيانات الضخمة والأدبيات المنشورة، ويحدد مجموعة متنوعة من الموضوعات التي تتراوح من العمل بكفاءة مع البيانات إلى إدارة البيانات على نطاق واسع. ومن بين هذه المواضيع أنه قد يحول حجم الويب وعدم تجانسه من إجراء التفكير الكامل ويتطلب حلولًا تقنية جديدة لتلبية إمكانات الاستدلال المطلوبة. تم توسيع هذه الميزة المطلوبة أيضًا لتشمل تقنيات التعلم الآلي وهذه التقنيات مطلوبة لاستخراج معلومات مفيدة من كميات هائلة من البيانات. على وجه التحديد، ذكر فرانسوا بانسيلهون مدى أهمية التعلم الآلي لاكتشاف الموضوعات وتصنيف الوثائق في داتا بابليكا. بعد ذلك، سلط ريكاردو بايزا-ياتس الضوء على الحاجة إلى معايير في حساب البيانات الضخمة من أجل السماح لمزودي البيانات الضخمة بمقارنة أنظمتهم. فالوعد المنطقي كما تم الترويج له في سياق الويب الدلالي لا يتوافق حاليًا مع متطلبات البيانات الضخمة بسبب مشكلات قابلية التوسع. يتم تعريف الاستدلال بمبادئ معينة، مثل السلامة والاكتمال، وهي بعيدة كل البعد عن العالم العملي وخصائص الويب، حيث تكون البيانات غالبًا متناقضة وغير كاملة وذات حجم ساحق. علاوة على ذلك، توجد فجوة بين الاستدلال على نطاق الويب والتفكير الأكثر تفصيلاً على مجموعات فرعية مبسطة من منطق الدرجة الأولى، نظرًا لحقيقة أنه يتم افتراض العديد من الجوانب، والتي تختلف عن الواقع (على سبيل المثال، مجموعة صغيرة من البديهيات والحقائق، والكمال وصحة قواعد الاستدلال).

تعد المقارنة المعيارية وليدة في مجال معالجة البيانات الدلالية على نطاق واسع، وفي الواقع يتم إنتاجها حاليًا فقط. على وجه الخصوص، يهدف مشروع مجلس معايير البيانات المرتبطة (LDBC) إلى "إنشاء مجموعة من المعايير لإدارة بيانات الرسم البياني واسع النطاق وRDF (إطار وصف الموارد) وكذلك إنشاء هيئة مستقلة لتطوير المعايير". جزء من مجموعة المعايير التي تم إنشاؤها في LDBC هو قياس الأداء واختبار تكامل البيانات ووظائف التفكير كما تدعمها أنظمة RDF. تركز هذه المعايير على اختبار: (1) مطابقة المثيل واستخراجها وتحويلها وتحميلها التي تلعب دورًا مهمًا في تكامل البيانات؛ و (2) ا قدرات منطق محركات RDF الحالية. كلا الموضوعين مهمان للغاية في الممارسة العملية، وقد تم تجاهلهما إلى حد كبير من خلال المعايير الحالية لمعالجة البيانات المرتبطة. عند إنشاء مثل هذه المعايير، يحلل LDBC مختلف السيناريوهات المتاحة لتحديد تلك التي يمكن أن تعرض بشكل أفضل تكامل البيانات ووظائف التفكير لمحركات RDF. بناءً على هذه السيناريوهات، يتم تحديد قيود أنظمة RDF الحالية من أجل جمع مجموعة من المتطلبات لتكامل بيانات RDF ومعايير الاستدلال. على سبيل المثال، من المعروف أن الأنظمة الحالية لا تعمل بشكل جيد في ظل وجود قواعد تفكير غير قياسية (مثل التفكير المتقدم الذي يأخذ في الاعتبار النفي والتجميع). علاوة على ذلك، يقوم المنطقيون الحاليون بإجراء الاستدلال من خلال تجسيد إغلاق

مجموعة البيانات (باستخدام التسلسل الخلفي أو الأمامي).[1] ومع ذلك، قد لا يكون هذا النهج قابلاً للتطبيق عندما يتم توفير قواعد الاستدلال الخاصة بالتطبيق، ومن ثم فمن المحتمل أن يؤدي تحسين حالة الفن إلى دعم استراتيجيات التفكير المختلط التي تتضمن كلاً من التسلسل الخلفي والأمامي، وإعادة كتابة الاستعلام (أي دمج مجموعة القواعد في الاستعلام).

الفرع الثالث: التعلم الآلي في البيانات الضخمة

تستخدم خوارزميات التعلم الآلي البيانات لتتعلم تلقائيًا كيفية أداء المهام مثل التنبؤ والتصنيف واكتشاف الانحراف. صُممت معظم خوارزميات التعلم الآلي لتعمل بكفاءة على معالج أو نواة واحدة. أدت التطورات في البنى متعددة النواة والحوسبة الشبكية إلى زيادة الحاجة إلى التعلم الآلي للاستفادة من توفر وحدات معالجة متعددة. توجد العديد من واجهات البرمجة واللغات المخصصة للبرمجة المتوازية مثل Orca MPI أو OpenACC، وهي مفيدة للبرمجة المتوازية للأغراض العامة. ومع ذلك، ليس من الواضح دائمًا كيف يمكن تنفيذ خوارزميات التعلم الآلي الحالية بطريقة متوازية. هناك مجموعة كبيرة من الأبحاث حول التعلم الموزع واستخراج البيانات التي تشمل خوارزميات التعلم الآلي التي تم تصميمها خصيصًا لأغراض الحوسبة الموزعة.[2]

بدلاً من إنشاء إصدارات موازية محددة من الخوارزميات، تتضمن الأساليب الأكثر عمومية أطر عمل لبرمجة التعلم الآلي على وحدات معالجة متعددة. تتمثل إحدى الطرق في استخدام تجريد عالي المستوى يبسط بشكل كبير تصميم وتنفيذ فئة محدودة من الخوارزميات المتوازية. على وجه الخصوص، تم تطبيق تجريد MapReduce بنجاح على مجموعة واسعة من تطبيقات التعلم الآلي. أظهر أن أي خوارزمية تناسب نموذج الاستعلام الإحصائي يمكن كتابتها في نموذج تجميع معين، والذي يمكن تنفيذه بسهولة بطريقة MapReduce ويحقق تسريعًا شبه خطي مع عدد وحدات المعالجة المستخدمة. لقد أظهروا أن هذا ينطبق على مجموعة متنوعة من خوارزميات التعلم.

أدت التطبيقات الموضحة في الورقة إلى الإصدار الأول من مكتبة التعلم الآلي MapReduce Mahout التي تشرح كيف يقيد نموذج MapReduce المستخدمين باستخدام افتراضات نمذجة بسيطة للغاية لضمان عدم وجود تبعيات حسابية في معالجة البيانات. يقترحون تجريد Graphlab الذي يعزل مستخدمين من تعقيدات البرمجة المتوازية (مثل سباقات البيانات، والمآزق)، مع الحفاظ على القدرة على التعبير عن التبعيات الحسابية المعقدة باستخدام رسم بياني للبيانات. تسمح لغات البرمجة ومجموعات الأدوات والأطر التي تمت مناقشتها بالعديد من التكوينات المختلفة لتنفيذ التعلم الآلي على نطاق واسع. يعتمد التكوين المثالي للاستخدام على التطبيق، حيث

[1] Fensel, D van Harmelen: Towards LarKC: A platform for web-scale reasoning. Los Alamitos, CA: IEEE Computer Society Press, 2007, pp. 94-96.
[2] http://www.cs.uinbc.edu/-hillol/DDM-BIB, viewed at 20:47, 23/03/2022.

سيكون للتطبيقات المختلفة مجموعات مختلفة من المتطلبات. ومع ذلك، فإن أحد أكثر الأطر شيوعًا المستخدمة في السنوات الأخيرة هو أباتشي هادوب Apache Hadoop، وهو تطبيق مفتوح المصدر ومجاني لنموذج MapReduce الذي تمت مناقشته أعلاه. يحدد أحد الخبراء Andrazˇ Tori بساطة Hadoop وMapReduce باعتباره المحرك الرئيسي لنجاحه. يوضح أن تنفيذ Hadoop يمكن أن يتفوق عليه من حيث وقت الحساب، على سبيل المثال، تنفيذ باستخدام Open MP، لكن Hadoop فاز من حيث الشعبية لأنه كان سهل الاستخدام.[1]

تتيح جهود الحساب المتوازية الموضحة أعلاه معالجة كميات كبيرة من البيانات. إلى جانب التطبيق الواضح لتطبيق الأساليب الحالية على مجموعات البيانات الكبيرة بشكل متزايد، تؤدي الزيادة في قوة الحساب أيضًا إلى مناهج جديدة للتعلم الآلي واسعة النطاق. أحد الأمثلة على ذلك هو العمل الأخير الذي تم فيه استخدام مجموعة بيانات من عشرة ملايين صورة لتعليم كاشف الوجه باستخدام البيانات غير الموسومة فقط. أدى استخدام الميزات الناتجة في مهمة التعرف على الكائن إلى زيادة في الأداء بنسبة 70٪ عن أحدث التقنيات. يمكن أن يصبح استخدام كميات كبيرة من البيانات للتغلب على الحاجة إلى بيانات التدريب المسمى اتجاهًا مهمًا. باستخدام البيانات غير الموسومة فقط، يتم تجاوز واحدة من أكبر الاختناقات أمام التبني الواسع للتعلم الآلي. إن استخدام أساليب التعلم غير الخاضعة للإشراف له حدوده على الرغم من أنه يبقى أن نرى ما إذا كان يمكن أيضًا تطبيق تقنيات مماثلة في مجالات التطبيق الأخرى.

المطلب الثاني: تقنيات تحليل البيانات الضخمة

تحليلات البيانات الكبيرة هي استخدام تقنيات تحليلية متقدمة مقابل مجموعات كبيرة ومتنوعة من البيانات الضخمة التي تتضمن بيانات منظمة وشبه منظمة وغير منظمة، من مصادر مختلفة، وبأحجام مختلفة من تيرابايت إلى زيتابايت.

الفرع الأول: تطور تحليل البيانات الضخمة

قد يكون من الصعب استخراج رؤى ذات مغزى حول الاتجاهات والارتباطات والأنماط الموجودة في البيانات الضخمة بدون قوة حوسبية هائلة. لكن التقنيات والتقنيات المستخدمة في تحليلات البيانات الضخمة تجعل من الممكن تعلم المزيد من مجموعات البيانات الكبيرة. وهذا يشمل البيانات من أي مصدر وحجم وبنية. باستخدام تحليلات البيانات الضخمة، يمكنك في النهاية تعزيز عملية صنع القرار والنمذجة والتنبؤ بالنتائج المستقبلية وتحسين ذكاء الأعمال بشكل أفضل وأسرع. أثناء قيامك ببناء حل البيانات الضخمة، ضع في اعتبارك البرامج

[1] Chu, C Kim: Map-reduce for machine learning on multicore. Advances in Neural Information Processing Systems, 2007.

مفتوحة المصدر مثل Apache Hadoop وApache Spark والنظام البيئي Hadoop بأكمله كأدوات فعالة من حيث التكلفة لمعالجة البيانات وتخزينها مصممة للتعامل مع حجم البيانات التي يتم إنشاؤها اليوم.[1]

لا تستطيع مستودعات البيانات التقليدية وقواعد البيانات العلائقية التعامل مع المهمة. كانت هناك حاجة إلى الابتكار. في عام 2006، تم إنشاء Hadoop بواسطة مهندسين في Yahoo وتم إطلاقه كمشروع Apache مفتوح المصدر. أتاح إطار المعالجة الموزع تشغيل تطبيقات البيانات الضخمة على نظام أساسي مجمع. هذا هو الفرق الرئيسي بين تحليلات البيانات الكبيرة التقليدية مقابل البيانات الكبيرة.

في البداية، استفادت الشركات الكبيرة فقط مثل Google وFacebook من تحليل البيانات الضخمة. بحلول عام 2010، بدأ تجار التجزئة والبنوك والشركات المصنعة وشركات الرعاية الصحية في رؤية قيمة كونها أيضًا شركات تحليلات البيانات الضخمة.

كانت المؤسسات الكبيرة ذات أنظمة البيانات المحلية هي الأنسب في البداية لجمع مجموعات البيانات الضخمة وتحليلها. لكن AWS وموردي الأنظمة الأساسية السحابية الآخرين سهّلوا على أي شركة استخدام منصة تحليلات البيانات الضخمة. أعطت القدرة على إعداد مجموعات Hadoop في السحابة شركة من أي حجم حرية التدوير وتشغيل ما يحتاجون إليه فقط عند الطلب.

يُعد النظام البيئي لتحليلات البيانات الضخمة مكونًا رئيسيًا للرشاقة، وهو أمر ضروري لشركات اليوم لتحقيق النجاح. يمكن اكتشاف الرؤى بشكل أسرع وأكثر كفاءة، مما يُترجم إلى قرارات أعمال فورية يمكن أن تحدد الفوز.

الفرع الثاني: أدوات تحليل البيانات الضخمة

تُستخدم قواعد بيانات NoSQL، (ليس فقط SQL) أو غير العلائقية، في الغالب لجمع وتحليل البيانات الضخمة. وذلك لأن البيانات الموجودة في قاعدة بيانات NoSQL تسمح بالتنظيم الديناميكي للبيانات غير المهيكلة مقابل التصميم المنظم والجداول لقواعد البيانات العلائقية.

تتطلب تحليلات البيانات الضخمة إطار عمل برمجيًا للتخزين الموزع ومعالجة البيانات الضخمة. تعتبر الأدوات التالية حلول برمجية لتحليل البيانات الضخمة:

[1] https://www.heavy.ai/learn/big-data-analytics, viewed at 00:07, 24/03/2022.

1) HBase:

HBase هو نظام إدارة قواعد بيانات غير علائقية موجه نحو العمود يعمل أعلى نظام الملفات الموزعة Hadoop(HDFS). يوفر HBase طريقة تتسامح مع الأخطاء لتخزين مجموعات البيانات المتفرقة، وهو أمر شائع في العديد من حالات استخدام البيانات الضخمة. إنه مناسب تمامًا لمعالجة البيانات في الوقت الفعلي أو الوصول العشوائي للقراءة / الكتابة إلى كميات كبيرة من البيانات.

على عكس أنظمة قواعد البيانات العلائقية، لا يدعم HBase لغة الاستعلام المهيكلة مثل SQL؛ في الواقع، HBase ليس مخزن بيانات علائقي على الإطلاق. تتم كتابة تطبيقات HBase بلغة Java ™ بشكل يشبه إلى حد كبير تطبيق Apache MapReduce النموذجي. يدعم HBase كتابة التطبيقات في Apache Avro وREST وThrift. تم تصميم نظام HBase للقياس الخطي. وهو يتألف من مجموعة من الجداول القياسية ذات الصفوف والأعمدة، مثل قاعدة البيانات التقليدية. يجب أن يحتوي كل جدول على عنصر محدد كمفتاح أساسي، ويجب أن تستخدم جميع محاولات الوصول إلى جداول HBase هذا المفتاح الأساسي.[1]

يدعم Avro، كمكون، مجموعة غنية من أنواع البيانات الأولية بما في ذلك: البيانات الرقمية والثنائية والسلاسل؛ وعدد من الأنواع المعقدة بما في ذلك المصفوفات والخرائط والتعداد والسجلات. يمكن أيضًا تحديد ترتيب الفرز للبيانات. ويعتمد HBase على ZooKeeper للتنسيق عالي الأداء. تم دمج ZooKeeper في HBase، ولكن إذا كنت تدير مجموعة إنتاج، فمن المقترح أن يكون لديك مجموعة ZooKeeper مخصصة مدمجة مع مجموعة HBase الخاصة بك. كما يعمل HBase بشكل جيد مع Hive، وهو محرك استعلام للمعالجة المجمعة للبيانات الضخمة، لتمكين تطبيقات البيانات الضخمة المتسامحة مع الأخطاء.

2) HIVE:

Hive، الذي تم تطويره في الأصل بواسطة Facebook وامتلكه لاحقًا Apache ، هو نظام تخزين بيانات تم تطويره بغرض تحليل البيانات المنظمة. يعمل Apache Hive في إطار منصة بيانات مفتوحة المصدر تسمى Hadoop، وهو نظام تطبيق تم إصداره في عام 2010 (أكتوبر).

تم تقديم Hive لتسهيل التحليل المتسامح (قليل التأثر بالأخطاء) للبيانات الضخمة على أساس منتظم، وقد تم استخدامه في تحليلات البيانات الضخمة وكان شائعًا في العالم لأكثر من عقد حتى الآن. على الرغم من أنه

[1] https://www.ibm.com/topics/hbase#:~:text=HBase%20is%20a%20column%2Doriented,many%20big%20data%20use%20cases., viewed at 18:56, 26/03/2022

يحتوي على العديد من المنافسين مثل Impala، إلا أن Apache Hive يقف بعيدًا عن بقية الأنظمة نظرًا لطبيعته المتسامحة مع الأخطاء في عملية تحليل البيانات وتفسيرها.

تعد Apache Hive أداة فعالة بشكل خاص عندما يتعلق الأمر بالبيانات الضخمة (البيانات الأسية التي سيتم تحليلها). برنامج بيانات المستودعات الذي يدعم عملية تحليل البيانات الخاصة بالبيانات الضخمة على أساس منتظم، يعتبر مفهوم البيانات الضخمة للخلية شائعًا جدًا في المجال التكنولوجي. نظرًا لأنه يتم تخزين البيانات في نظام الملفات الموزعة (Apache Hadoop HDFS) حيث يتم تنظيم البيانات وتنظيمها، كما تساعد Apache Hive في معالجة هذه البيانات وتحليلها لإنتاج أنماط واتجاهات تعتمد على البيانات. يصلح للاستخدام من قبل المنظمات أو المؤسسات، Apache Hive مفيد للغاية في البيانات الضخمة ونموها المتغير باستمرار.[1]

يشارك مفهوم لغة الاستعلام الهيكلية أو برنامج SQL في العملية التي تتواصل مع العديد من قواعد البيانات وتجمع البيانات المطلوبة. يمكن أن يساعدنا فهم بيانات Hive الضخمة من خلال عدسة تحليلات البيانات في الحصول على مزيد من الأفكار حول عمل Apache Hive. باستخدام تسلسل معالجة الدُفعات، يُنشئ Hive تحليلات البيانات بشكل أسهل بكثير ومنظم ويتطلب أيضًا وقتًا أقل مقارنة بالأدوات التقليدية. HiveQL هي لغة مشابهة لـ SQL تتفاعل مع قاعدة بيانات Hive عبر مؤسسات مختلفة وتحلل البيانات الضرورية بتنسيق منظم.

HIVE في البيانات الضخمة هي ابتكار بارز أدى في النهاية إلى تحليل البيانات على نطاق واسع. تحتاج المؤسسات الكبيرة إلى بيانات ضخمة لتسجيل المعلومات التي يتم جمعها بمرور الوقت. لإنتاج تحليل يعتمد على البيانات، تقوم المؤسسات بجمع البيانات واستخدام مثل هذه التطبيقات البرمجية لتحليل بياناتها. يمكن استخدام هذه البيانات، مع Apache Hive، لقراءة وكتابة وإدارة المعلومات التي تم تخزينها في نموذج منظم. منذ ظهور تحليلات البيانات، كان تخزين البيانات موضوعًا شائعًا.

نظرًا لأن جمع البيانات أصبح مهمة يومية وتوسعت المنظمات في جميع الجوانب، أصبح جمع البيانات أساسيًا وواسعًا. علاوة على ذلك، بدأ التعامل مع البيانات بالبيتابايت الذي يحدد تخزين البيانات الضخمة لهذا، احتاجت المؤسسات إلى معدات ضخمة وربما كان هذا هو السبب في ضرورة إصدار برنامج مثل Apache Hive. وبالتالي، تم إصدار Apache Hive بهدف تحليل البيانات الضخمة وإنتاج قياسات تعتمد على البيانات.

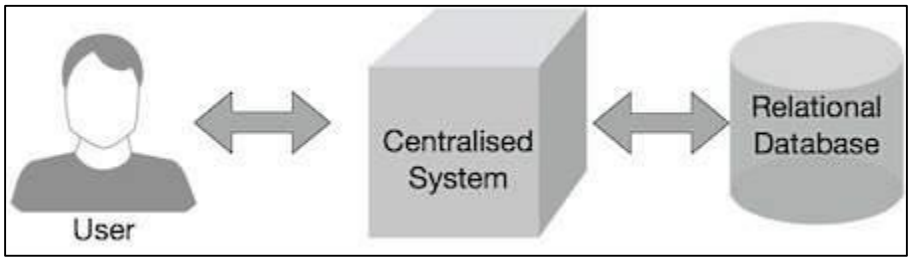
3) MapReduce:

عادةً ما تحتوي أنظمة المؤسسات التقليدية على خادم مركزي (Centralized System) لتخزين البيانات ومعالجتها. يوضح الرسم التوضيحي التالي عرضًا تخطيطيًا لنظام مؤسسة تقليدي. من المؤكد أن النموذج التقليدي

[1] https://www.analyticssteps.com/blogs/what-hive-big-data-and-its-benefits, viewed at 16:23, 27/03/2022.

ليس مناسبًا لمعالجة كميات ضخمة من البيانات القابلة للتطوير ولا يمكن أن تستوعبه خوادم قواعد البيانات القياسية(Relational Database) . علاوة على ذلك، فإن النظام المركزي يخلق الكثير من الاختناق للمستخدم (User) أثناء معالجة ملفات متعددة في وقت واحد.

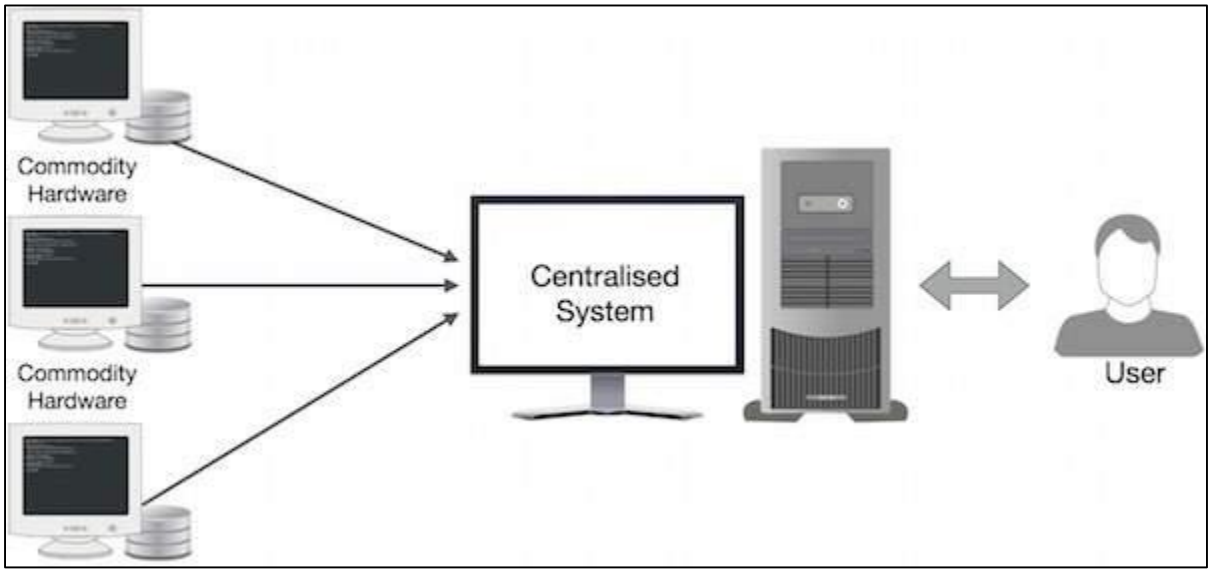
الشكل ب: عرض تخطيطي لنظام مؤسسة تقليدي.



المصدر: www.tutorialspoint.com , viewed at 17:45, 28/03/2022.

قامت Google بحل مشكلة الاختناق هذه باستخدام خوارزمية تسمى MapReduce. يقسم MapReduce المهمة إلى أجزاء صغيرة ويعينها للعديد من أجهزة الكمبيوتر. في وقت لاحق، يتم جمع النتائج في مكان واحد ودمجها لتشكيل مجموعة البيانات الناتجة.

الشكل ج: عرض تخطيطي لنظام MapReduce.



المصدر: www.tutorialspoint.com , viewed at 17:45, 28/03/2022.

تحتوي خوارزمية MapReduce على مهمتين، وهما Map وReduce:

- تأخذ مهمة Map مجموعة من البيانات وتحولها إلى مجموعة أخرى من البيانات، حيث يتم تقسيم العناصر الفردية إلى مجموعات.
- تأخذ مهمة Reduce الإخراج من الخريطة كمدخلات وتجمع مجموعات البيانات هذه (أزواج القيمة الرئيسية) في مجموعة أصغر من المجموعات.

يتم تنفيذ مهمة Reduce دائمًا بعد وظيفة Map.

نقوم الآن بإلقاء نظرة المراحل التي تمر علها هذه الخوارزمية في التحليل:[1]

- **طور الإدخال:** لدينا هنا قارئ سجل يقوم بترجمة كل سجل في ملف إدخال ويرسل البيانات التي تم تحليلها إلى مصمم الخرائط في شكل أزواج قيمة رئيسية.
- **الخريطة:** الخريطة (Map)هي وظيفة يحددها المستخدم، والتي تأخذ سلسلة من أزواج القيمة الرئيسية وتعالج كل واحد منهم لتوليد صفر أو أكثر من أزواج القيمة الرئيسية.
- **المفاتيح الوسيطة:** تُعرف أزواج مفاتيح القيمة الرئيسية التي تم إنشاؤها بواسطة مصمم الخرائط على أنها المفاتيح الوسيطة.
- **الموحد أو المدمج (Combiner):** هو نوع من المخفض المحلي الذي يقوم بتجميع البيانات المتشابهة من مرحلة الخريطة في مجموعات قابلة للتحديد. يأخذ المفاتيح الوسيطة من معين الخرائط كمدخلات ويطبق كود معرف من قبل المستخدم لتجميع القيم في نطاق صغير من مخطط واحد. إنه ليس جزءًا من خوارزمية MapReduce الرئيسية؛ إنه اختياري.
- **التبديل والفرز:** تبدأ مهمة Reducer بخطوة Shuffle and Sort. يقوم بتنزيل أزواج القيمة الرئيسية المجمعة على الجهاز المحلي، حيث يعمل Reducer. يتم فرز أزواج القيمة الرئيسية الفردية حسب المفتاح في قائمة بيانات أكبر. تجمع قائمة البيانات المفاتيح المكافئة معًا بحيث يمكن تكرار قيمها بسهولة في مهمة Reducer.
- **المخفض:** يأخذ Reducer البيانات المجمعة ذات القيمة الرئيسية المقترنة كإدخال ويقوم بتشغيل وظيفة Reducer على كل منها. هنا، يمكن تجميع البيانات وتصفيتها ودمجها بعدة طرق، وتتطلب نطاقًا واسعًا من المعالجة. بمجرد انتهاء التنفيذ، فإنه يعطي صفرًا أو أكثر من أزواج القيمة الرئيسية للخطوة النهائية.

[1] https://www.tutorialspoint.com/map_reduce/map_reduce_introduction.htm#:~:text=MapReduce%20is%20a%20programming%20model,huge%20volumes%20of%20complex%20data., viewed at 19:17, 02/04/2022.

- **مرحلة الإخراج:** في مرحلة الإخراج، لدينا مُنسق إخراج يقوم بترجمة أزواج القيمة الرئيسية النهائية من وظيفة Reducer وكتابتها في ملف باستخدام كاتب سجل.

كملخص، يعزز إطار عمل MapReduce جدولة المهام ومراقبتها. يتم إعادة تنفيذ المهام الفاشلة بواسطة إطار العمل. يمكن استخدام إطار العمل هذا بسهولة، حتى من قبل المبرمجين ذوي الخبرة القليلة في المعالجة الموزعة. يمكن تنفيذ MapReduce باستخدام لغات برمجة مختلفة مثل Java وHive وPig وScala وPython.

4) Pig:

في وقت سابق من عام 2006، تم تطوير Apache Pig بواسطة باحثي Yahoo. في ذلك الوقت، كانت الفكرة الرئيسية لتطوير Pig هي تنفيذ وظائف MapReduce على مجموعات بيانات كبيرة للغاية. في عام 2007، انتقلت إلى (Apache Software Foundation. ASF) مما يجعلها مشروعًا مفتوح المصدر. الإصدار الأول (0.1) من Pig جاء في عام 2008. أحدث إصدار من Apache Pig هو 0.18 والذي جاء في عام 2017.

Pig هي منصة أو أداة عالية المستوى تُستخدم لمعالجة مجموعات البيانات الكبيرة. يوفر مستوى عاليًا من التجريد للمعالجة عبر MapReduce. يوفر لغة برمجة عالية المستوى، تُعرف باسم Pig Latin والتي تُستخدم لتطوير أكواد تحليل البيانات لمعالجة البيانات المخزنة في HDFS، سيكتب المبرمجون البرامج النصية باستخدام Pig Latin Language. قام Internally Pig Engine (أحد مكونات Apache Pig) بتحويل كل هذه البرامج النصية إلى خريطة محددة وتقليل المهمة. لكن هذه ليست مرئية للمبرمجين من أجل توفير مستوى عالٍ من التجريد. يعد Pig LatinوEngine Pig المكونين الرئيسيين لأداة Apache Pig. يتم تخزين نتيجة Pig دائمًا في HDFS.[1]

أحد قيود MapReduce هو أن دورة التطوير طويلة جدًا. تعد كتابة المخفض والمخطط، وتجميع حزم الشفرة، وإرسال المهمة واسترداد المخرجات مهمة تستغرق وقتًا طويلاً. يقلل Apache Pig من وقت التطوير باستخدام أسلوب الاستعلام المتعدد. أيضًا، يعد Pig مفيدًا للمبرمجين الذين ليسوا من خلفية Java. يمكن كتابة 200 سطر من كود Java في 10 أسطر فقط باستخدام لغة Pig Latin. يحتاج المبرمجون الذين لديهم معرفة بـ SQL إلى جهد أقل لتعلم Pig Latin.

5) Spark:

بدأ Apache Spark في عام 2009 كمشروع بحثي في جامعة كاليفورنيا في بيركلي AMPLab، وهو تعاون

[1] https://www.geeksforgeeks.org/introduction-to-apache-pig/#:~:text=Pig%20is%20a%20high%2Dlevel,develop%20the%20data%20analysis%20codes, viewed at 22:35, 03/04/2022.

يضم الطلاب والباحثين وأعضاء هيئة التدريس، ويركز على مجالات التطبيق كثيفة البيانات. كان الهدف من Spark هو إنشاء إطار عمل جديد، مُحسَّن للمعالجة التكرارية السريعة مثل التعلم الآلي، وتحليل البيانات التفاعلي، مع الاحتفاظ بقابلية التوسع، والتسامح مع أخطاء Hadoop MapReduce. نُشرت الورقة الأولى بعنوان "Spark: Cluster Computing with Working Sets" في يونيو 2010، وكان Spark مفتوح المصدر بموجب ترخيص BSD. في يونيو 2013، دخلت Spark حالة الحضانة في Apache Software Foundation (ASF)، وتم تأسيسها كمشروع Apache Top-Level في فبراير 2014. يمكن تشغيل Spark بشكل مستقل، على Apache Mesos، أو في أغلب الأحيان على Apache Hadoop.

Apache Spark هو نظام معالجة موزع مفتوح المصدر يستخدم لأحمال عمل البيانات الضخمة. يستخدم التخزين المؤقت في الذاكرة وتنفيذ الاستعلام المحسن للاستعلامات التحليلية السريعة مقابل البيانات من أي حجم. يوفر تطوير واجهات برمجة التطبيقات في Java وScala وPython وR، ويدعم إعادة استخدام الكود عبر أحمال عمل متعددة -معالجة الدُفعات والاستعلامات التفاعلية والتحليلات في الوقت الفعلي والتعلم الآلي ومعالجة الرسم البياني. ستجده مستخدمًا من قِبل مؤسسات من أي مجال، بما في ذلك FINRA وYelp وZillow وDataXu وInstitute Urban وCrowdStrike. أصبح Apache Spark واحدًا من أشهر أطر معالجة البيانات الضخمة الموزعة مع 365000 عضو في لقاء عام 2017.[1]

تم إنشاء Spark لمعالجة القيود المفروضة على MapReduce، من خلال إجراء المعالجة في الذاكرة، وتقليل عدد الخطوات في الوظيفة، وإعادة استخدام البيانات عبر عمليات متعددة متوازية. باستخدام Spark، هناك حاجة إلى خطوة واحدة فقط حيث تتم قراءة البيانات في الذاكرة، وتنفيذ العمليات، وإعادة كتابة النتائج مما يؤدي إلى تنفيذ أسرع بكثير. يعيد Spark أيضًا استخدام البيانات باستخدام ذاكرة التخزين المؤقت في الذاكرة لتسريع خوارزميات التعلم الآلي التي تستدعي بشكل متكرر وظيفة في نفس مجموعة البيانات. تتم إعادة استخدام البيانات من خلال إنشاء إطارات البيانات، وهي عبارة عن تجريد عبر مجموعة البيانات الموزعة المرنة (RDD)، وهي مجموعة من الكائنات المخزنة مؤقتًا في الذاكرة، وإعادة استخدامها في عمليات شرارة متعددة. هذا يقلل بشكل كبير من زمن الوصول مما يجعل Spark أسرع عدة مرات من MapReduce، خاصة عند القيام بالتعلم الآلي والتحليلات التفاعلية.

6) YARN:

YARN (Yet Another Resource Negotiator) تعني "مع ذلك مفاوض موارد آخر". تم تقديمه في Hadoop 2.0 لإزالة الاختناق في Job Tracker الذي كان موجودًا في Hadoop 1.0. تم وصف YARN بأنه "مدير موارد معاد

[1] https://aws.amazon.com/big-data/what-is-spark/, viewed at 20:43, 05/04/2022

تصميمه" وقت إطلاقه، ولكنه تطور الآن ليصبح معروفًا بنظام التشغيل الموزع واسع النطاق المستخدم في معالجة البيانات الضخمة.

يسمح YARN أيضًا بمحركات معالجة البيانات المختلفة مثل معالجة الرسم البياني والمعالجة التفاعلية ومعالجة الدفق بالإضافة إلى معالجة الدُفعات لتشغيل ومعالجة البيانات المخزنة في HDFS (نظام الملفات الموزعة Hadoop) مما يجعل النظام أكثر كفاءة. من خلال مكوناته المختلفة، يمكنه تخصيص موارد مختلفة ديناميكيًا وجدولة معالجة الطلب. لمعالجة البيانات ذات الحجم الكبير، من الضروري تمامًا إدارة الموارد المتاحة بشكل صحيح بحيث يمكن لكل تطبيق الاستفادة منها.

تشمل المكونات الرئيسية لهندسة YARN ما يلي:[1]

- **العميل:** يقدم وظائفmap-reduce.
- **مدير الموارد:** هو البرنامج الخفي الرئيسي لـ YARN وهو مسؤول عن تخصيص الموارد وإدارتها بين جميع التطبيقات. عندما يتلقى طلب معالجة، فإنه يعيد توجيهه إلى مدير العقدة المقابل ويخصص الموارد لإكمال الطلب وفقًا لذلك. يتكون من مكونين رئيسيين: -**المجدول:** ينفذ الجدولة بناءً على التطبيق المخصص والموارد المتاحة. إنه برنامج جدولة خالص، مما يعني أنه لا يؤدي مهام أخرى مثل المراقبة أو التتبع ولا يضمن إعادة التشغيل في حالة فشل المهمة. يدعم جدولة YARN المكونات الإضافية مثل جدولة السعة و Fair Scheduler لتقسيم موارد الكتلة. و-**مدير التطبيق:** وهو مسؤول عن قبول التطبيق والتفاوض بشأن الحاوية الأولى من مدير الموارد. يقوم أيضًا بإعادة تشغيل حاوية التطبيق الرئيسية في حالة فشل المهمة.
- **مدير العقدة:** يعتني بالعقدة الفردية على كتلة Hadoop ويدير التطبيق وسير العمل وتلك العقدة المعينة. وظيفتها الأساسية هي مواكبة مدير الموارد. يسجل مع مدير الموارد ويرسل نبضات مع الحالة الصحية للعقدة. إنه يراقب استخدام الموارد، ويؤدي إدارة السجل، كما أنه يقتل حاوية بناءً على توجيهات مدير الموارد. كما أنها مسؤولة عن إنشاء عملية الحاوية وبدء تشغيلها بناءً على طلب مدير التطبيق.
- **التطبيق الرئيسي:** التطبيق هو وظيفة واحدة يتم تقديمها إلى إطار عمل. مدير التطبيق مسؤول عن التفاوض على الموارد مع مدير الموارد، وتتبع الحالة ومراقبة التقدم في تطبيق واحد. يطلب مدير التطبيق الحاوية من مدير العقدة عن طريق إرسال سياق تشغيل الحاوية (CLC) والذي يتضمن كل ما يحتاجه التطبيق للتشغيل. بمجرد بدء التطبيق، يرسل التقرير الصحي إلى مدير الموارد من وقت لآخر.
- **الحاوية:** هي عبارة عن مجموعة من الموارد المادية مثل ذاكرة الوصول العشوائي (RAM)، وأنوية وحدة المعالجة المركزية (CPU) والقرص الموجود على عقدة واحدة. يتم استدعاء الحاويات بواسطة Container

[1] https://www.geeksforgeeks.org/hadoop-yarn-architecture, viewed at 19:39, 06/04/2022.

Launch Context (CLC) وهو سجل يحتوي على معلومات مثل متغيرات البيئة ورموز الأمان والتبعيات وما إلى ذلك.

أدى استخدام Apache Hadoop YARN لفصل HDFS عن MapReduce إلى جعل بيئة Hadoop أكثر ملاءمة لاستخدامات المعالجة في الوقت الفعلي والتطبيقات الأخرى التي لا يمكنها الانتظار حتى تنتهي المهام المجمعة. الآن، MapReduce هو مجرد واحد من العديد من محركات المعالجة التي يمكنها تشغيل تطبيقات Hadoop. لم يعد لديه قفل على معالجة الدُفعات في Hadoop بعد الآن: في كثير من الحالات، يقوم المستخدمون باستبداله بـ Spark للحصول على أداء أسرع في التطبيقات المجمعة، مثل وظائف الاستخراج والتحويل وتحميل.[1]

[1] https://www.techtarget.com/searchdatamanagement/definition/Apache-Hadoop-YARN-Yet-Another-Resource-Negotiator, viewed at 21:28, 06/04/2022

الخلاصة:

للتلخيص، يعد تحليل البيانات الضخمة جزءًا أساسيًا من سلسلة القيمة الإجمالية للبيانات الضخمة التي تعد بأن يكون لها تأثير اقتصادي واجتماعي كبير في الاتحاد الأوروبي على المدى القريب إلى المتوسط. بدون تحليل البيانات الضخمة، لا تعمل بقية السلسلة. ومع مرور الوقت ازداد حجم البيانات مما فرض ابتكار وسائل تخزين جديد. وقد قمنا بتوضيح ذلك من خلال هذا الفصل، كما أشرنا الى أساسيات مهمة في تخزين البيانات الضخمة كالخصوصية، والنمو الحاد للبيانات، وأهم وسائل وتقنيات تخزين البيانات الضخمة (NewSQL،.NoSQL)، وهي التي تسمح بتخزين هذا الكم الهائل من المعلومات في شكل بسيط ويسهل استعمالها في التحليل، الذي يعد جزءًا أساسيًا من سلسلة قيمة البيانات الضخمة. يمكننا رسم هذه العملية كاريكاتوريًا باستخدام قول إنكليزي قديم أن ما يحققه هذا المكون هو "تحويل الرصاص إلى ذهب". يتم تحويل كميات كبيرة من البيانات التي قد تكون غير متجانسة فيما يتعلق بآلية التشفير، والتنسيق، والبنية، والدلالات الأساسية، والمصدر، والموثوقية، والجودة إلى بيانات قابلة للاستخدام من خلال عدة آليات وتقنيات (MapReduce،HIVE، Hbase، Spark).

الفصل الثالث:

تطبيقات

البيانات

الضخمة

تمهيد:

إحدى مهام العمل الأساسية لاستخدام البيانات الضخمة هي المساعدة في اتخاذ قرارات العمل. حيث تتضمن عملية اتخاذ القرار إعداد التقارير واستكشاف البيانات (التصفح والبحث) والبحث الاستكشافي (إيجاد الارتباطات والمقارنات وسيناريوهات ماذا لو وما إلى ذلك). إن القيمة التجارية لهذه الخدمات اللوجستية للمعلومات ذات شقين: (1) التحكم في سلسلة القيمة و (2) شفافية سلسلة القيمة. السابق بشكل عام مستقل عن البيانات الضخمة؛ ومع ذلك، يوفر الأخير الفرص والمتطلبات لأسواق البيانات والخدمات فاستخدام البيانات هو مجال واسع يتم تناوله في هذا الفصل من خلال عرض تطبيقات البيانات من وجهات نظر مختلفة، بما في ذلك مكدسات التكنولوجيا الأساسية، والاتجاهات في مختلف القطاعات، والتأثير على نماذج الأعمال، والمتطلبات على التفاعل بين الإنسان والحاسوب.

المبحث الأول: فرص استخدام البيانات الضخمة

تعتبر البيانات الضخمة حدثا ضخما وواقعا جديدا أفرزته التطورات المتواصلة في عالم التكنولوجيا، وتعتبر البيانات الضخمة كحل للتساؤلات المطروحة حيث أصبح الاعتماد عليها امرا حتميا نظرا لإمكانياتها الكبيرة لدفع الابتكار والرقي في جميع المجالات خاصة الاقتصادية والاجتماعية، وما تقدمه من فرص حقيقية لحل المشاكل من خلال الاعتماد على عدة أدوات والتي سنتطرق اليها في هذا المبحث.

المطلب الأول: أساسيات في استخدام البيانات الضخمة

فيما يلي النقاط الرئيسية لاستخدام البيانات الضخمة والتي سنقوم بشرحها أكثر فيما بعد:[1]

- **التحليلات التنبؤية:** من الأمثلة الرئيسية لتطبيق التحليلات التنبؤية في الصيانة التنبؤية استنادًا إلى بيانات المستشعر والسياق للتنبؤ بالانحرافات عن فترات الصيانة القياسية. عندما تشير البيانات إلى نظام مستقر، يمكن تمديد فترات الصيانة، مما يؤدي إلى انخفاض تكاليف الصيانة. عندما تشير البيانات إلى وجود مشاكل قبل الوصول إلى الصيانة المجدولة، يمكن أن تكون الوفورات أعلى إذا أمكن تجنب الأعطال وتكلفة الإصلاح وأوقات التعطل. تتجاوز مصادر المعلومات بيانات المستشعر وتميل إلى تضمين البيانات البيئية والسياقية، بما في ذلك معلومات الاستخدام (مثل الحمل العالي) للآلة. نظرًا لأن التحليل التنبئي يعتمد على أجهزة استشعار جديدة وبنية تحتية لمعالجة البيانات، فإن كبار المصنّعين يغيرون نموذج أعمالهم ويستثمرون في البنية التحتية الجديدة بأنفسهم (مع تحقيق تأثيرات الحجم على الطريق) وتأجير الآلات لعملائهم.

- **الصناعة 4.0:** الاتجاه المتنامي في التصنيع هو استخدام الأنظمة الإلكترونية الفيزيائية. إنه يؤدي إلى تطور عمليات التصنيع القديمة، من ناحية توفير كمية هائلة من المعلومات الحساسة والبيانات الأخرى ومن ناحية أخرى جلب الحاجة إلى ربط جميع البيانات المتاحة من خلال شبكات الاتصال وسيناريوهات الاستخدام التي تجني الفوائد المحتملة. الصناعة 4.0 تعني دخول تكنولوجيا المعلومات إلى الصناعة التحويلية وتجلب معها عددًا من التحديات لدعم تكنولوجيا المعلومات. يتضمن ذلك خدمات لمهام متنوعة مثل التخطيط والمحاكاة، والرصد والتحكم، والاستخدام التفاعلي للآلات، واللوجستيات وتخطيط موارد المؤسسة (ERP)، والتحليل التنبئي، والتحليل الإرشادي في النهاية حيث يمكن التحكم في عمليات اتخاذ القرار تلقائيًا عن طريق تحليل البيانات.

[1] José María Cavanillas: New Horizons for a Data-Driven Economy, previous refrence, p145.

- **تكامل البيانات والخدمات الذكية:** عند تطوير سيناريو الصناعة 4.0 أعلاه، يتم التركيز على الخدمات التي تحل المهام المطروحة. لتمكين تطبيق الخدمات الذكية من التعامل مع مشاكل استخدام البيانات الضخمة، هناك أمور فنية وتنظيمية. يجب معالجة قضايا حماية البيانات والخصوصية، والقضايا التنظيمية، والتحديات القانونية الجديدة (على سبيل المثال، فيما يتعلق بقضايا الملكية للبيانات المشتقة). على المستوى التقني، هناك أبعاد متعددة يجب على طولها تمكين تفاعل الخدمات: على مستوى الأجهزة من الأجهزة الفردية، إلى المرافق، إلى الشبكات؛ على المستوى المفاهيمي من الأجهزة الذكية إلى الأنظمة الذكية والقرارات؛ على مستوى البنية التحتية من IaaS إلى PaaS و SaaSإلى الخدمات الجديدة لاستخدام البيانات الضخمة وحتى العمليات التجارية والمعرفة كخدمة.

- **الاستكشاف التفاعلي:** عند العمل مع كميات كبيرة من البيانات بتنوع كبير، غالبًا ما تكون النماذج الأساسية للعلاقات الوظيفية مفقودة. هذا يعني أن محللي البيانات لديهم حاجة أكبر لاستكشاف مجموعات البيانات والتحليلات. تتم معالجة ذلك من خلال التحليلات المرئية والطرق الجديدة والديناميكية لتصور البيانات، ولكن هناك حاجة إلى واجهات مستخدم جديدة ذات قدرات جديدة لاستكشاف البيانات. توفر بيئات استخدام البيانات المتكاملة الدعم، على سبيل المثال، من خلال آليات التاريخ والقدرة على مقارنة التحليلات المختلفة وإعدادات المعلمات المختلفة والنماذج المتنافسة.

تؤثر البيانات الضخمة على صحة اتخاذ القرار المستند إلى البيانات في المستقبل. العوامل المؤثرة هي أولا النطاق الزمني للقرارات والتوصيات، من المدى القصير إلى المدى الطويل وثانيا قواعد البيانات المختلفة (بالمعنى غير التقني) من البيانات السابقة والتاريخية إلى البيانات الحالية والحديثة. ستؤثر التطبيقات الجديدة التي تعتمد على البيانات بقوة على تطوير أسواق جديدة. من العوائق المحتملة لمثل هذه التطورات دائمًا الحاجة إلى شبكات شركاء جديدة (مجموعة من القدرات المنفصلة حاليًا) وعمليات الأعمال والأسواق.

مجال خاص لحالات الاستخدام للبيانات الضخمة هو التصنيع والنقل وقطاع الخدمات اللوجستية. تخضع هذه القطاعات لتغير تحولي كجزء من اتجاه على مستوى الصناعة، يسمى "الصناعة 4.0"، والذي ينشأ في رقمنة وربط المنتجات، ومرافق الإنتاج، والبنية التحتية للنقل كجزء من تطوير "إنترنت الأشياء". استخدام البيانات له تأثير عميق في هذه القطاعات، على سبيل المثال تؤدي تطبيقات التحليل التنبئي في الصيانة إلى نماذج أعمال جديدة حيث أن مصنعي الآلات هم في أفضل وضع لتقديم الصيانة القائمة على البيانات الضخمة. يجلب ظهور الأنظمة السيبرانية الفيزيائية (CPS) للإنتاج والنقل والخدمات اللوجستية والقطاعات الأخرى تحديات جديدة للمحاكاة والتخطيط، للمراقبة والتحكم والتفاعل (من قبل الخبراء وغير الخبراء) مع تطبيقات استخدام الآلات أو البيانات.

على نطاق أوسع، هناك حاجة إلى خدمات جديدة وبنية تحتية جديدة للخدمات. تحت عنوان "البيانات الذكية" وخدمات البيانات الذكية، تمت صياغة متطلبات البيانات وأسواق الخدمة أيضًا. إلى جانب البنية التحتية التقنية للتفاعل والتعاون في الخدمات من مصادر متعددة، هناك قضايا قانونية وتنظيمية تحتاج إلى معالجة. تعد البنية التحتية المناسبة للخدمة أيضًا فرصة للشركات الصغيرة والمتوسطة للمشاركة في سيناريوهات استخدام البيانات الضخمة من خلال تقديم خدمات محددة، على سبيل المثال، من خلال أسواق خدمات استخدام البيانات.

المطلب الثاني: أثر استخدام البيانات الضخمة في العلوم الاقتصادية

أحد أهم تأثيرات سيناريوهات استخدام البيانات الضخمة هو اكتشاف علاقات وتبعيات جديدة في البيانات تؤدي، على السطح، إلى فرص اقتصادية ومزيد من الكفاءة. على مستوى أعمق، يمكن أن يوفر استخدام البيانات الضخمة فهمًا أفضل لهذه التبعيات، مما يجعل النظام أكثر شفافية ويدعم عمليات صنع القرار الاقتصادي والاجتماعي. حيثما تكون البيانات متاحة للجمهور، يتم دعم اتخاذ القرار الاجتماعي؛ حيثما تتوفر البيانات ذات الصلة على المستوى الفردي، يتم دعم اتخاذ القرارات الشخصية. تأتي إمكانية الشفافية من خلال استخدام البيانات الضخمة مع عدد من المتطلبات:

- اللوائح والاتفاقيات بشأن الوصول إلى البيانات والملكية والحماية والخصوصية.
- المطالب المتعلقة بجودة البيانات (على سبيل المثال بشأن اكتمال ودقة وتوقيت).
- الوصول إلى البيانات الأولية وكذلك الوصول إلى الأدوات أو الخدمات المناسبة لاستخدام البيانات الضخمة.

وبالتالي فإن الشفافية لها بعد اقتصادي واجتماعي وشخصي. عندما يمكن تلبية المتطلبات المذكورة أعلاه، تصبح القرارات شفافة ويمكن اتخاذها بطريقة أكثر موضوعية وقابلة للتكرار، حيث تكون عمليات اتخاذ القرار مفتوحة لإشراك لاعبين آخرين.

الدوافع الاقتصادية الحالية لاستخدام البيانات الضخمة هي الشركات الكبيرة التي تتمتع بإمكانية الوصول إلى البنى التحتية الكاملة. وتشمل هذه قطاعات مثل الإعلان في شركات الإنترنت وبيانات الاستشعار من البنى التحتية الكبيرة (مثل الشبكات الذكية أو المدن الذكية) أو للآلات المعقدة (مثل محركات الطائرات). في الأمثلة الأخيرة، هناك اتجاه نحو تكامل أوثق لاستخدام البيانات في الشركات الكبيرة حيث تظل قدرات البيانات الضخمة مع الشركات المصنعة (وليس العملاء)، على سبيل المثال عندما يتم تأجير المحركات فقط وتكون البنية التحتية للبيانات الضخمة مملوكة للمصنعين وتديرها.

هناك متطلبات متزايدة للمعايير والأسواق التي يمكن الوصول إليها للبيانات وكذلك للخدمات لإدارة وتحليل واستغلال المزيد من استخدامات البيانات. عند استيفاء هذه المتطلبات، يتم إنشاء الفرص للشركات الصغيرة والمتوسطة للمشاركة في حالات استخدام أكثر تعقيدًا لاستخدام البيانات الضخمة. في الفروع التالية سنناقش أهم ما يقدمه مجال استخدام البيانات الضخمة في العلوم الاقتصادية وهو موضوع دراستنا:

الفرع الأول: دعم القرار

نظام دعم القرار (DSS) هو تطبيق برنامج كمبيوتر يستخدم لتحسين قدرات اتخاذ القرار في الشركة. يحلل كميات كبيرة من البيانات ويقدم للمؤسسة أفضل الخيارات الممكنة المتاحة. تجمع أنظمة دعم القرار البيانات والمعرفة من مختلف المجالات والمصادر لتزويد المستخدمين بمعلومات تتجاوز التقارير والملخصات المعتادة. هذا يهدف إلى مساعدة الناس على اتخاذ قرارات مستنيرة.[1]

نظام دعم القرار هو تطبيق إعلامي وليس تطبيق تشغيلي. تزود التطبيقات المعلوماتية المستخدمين بالمعلومات ذات الصلة بناءً على مجموعة متنوعة من مصادر البيانات لدعم اتخاذ قرارات مستنيرة بشكل أفضل. على النقيض من ذلك، تسجل التطبيقات التشغيلية تفاصيل المعاملات التجارية، بما في ذلك البيانات المطلوبة لاحتياجات دعم القرار للأعمال.

تستخدم أنظمة دعم القرار الحالية بقدر ما تعتمد على التقارير الثابتة هذه التقنيات ولكنها لا تسمح بالاستخدام الديناميكي الكافي لجني الإمكانات الكاملة للبحث الاستكشافي. ومع ذلك، في ترتيب متزايد من التعقيد، تشمل هذه المجموعات أهداف العمل التالية:

- **البحث:** في أدنى مستوى من التعقيد، يتم استرداد البيانات فقط لأغراض مختلفة. وتشمل هذه استرجاع الحقائق والبحث عن العناصر المعروفة، على سبيل المثال لأغراض التحقق. تشمل الوظائف الإضافية التنقل عبر مجموعات البيانات والمعاملات.
- **التعلم:** في المستوى التالي، يمكن لهذه الوظائف أن تدعم اكتساب المعرفة وتفسير البيانات، مما يتيح الفهم. تشمل الوظائف الداعمة للمقارنة والتجميع وتكامل البيانات. قد تدعم المكونات الإضافية الوظائف الاجتماعية لتبادل البيانات. تتضمن أمثلة التعلم عمليات بحث بسيطة عن عنصر معين (اكتساب المعرفة)، على سبيل المثال أحد المشاهير واستخدامهم في الإعلان (البيع بالتجزئة). من المتوقع أن يقوم تطبيق البحث عن البيانات الضخمة بالعثور على جميع البيانات ذات الصلة وتقديم عرض متكامل.

[1] https://www.techtarget.com/searchcio/definition/decision-support-system#:~:text=A%20decision%20support%20system%20(DSS)%20is%20a%20computer%20program%20application,the%20best%20possible%20options%20available, viewed at 19:44, 10/04/2022.

- **التحقيق:** على أعلى مستوى من أنظمة دعم القرار، يمكن تحليل البيانات وتجميعها وتوليفها. يتضمن ذلك دعم الأداة للاستبعاد والنفي والتقييم. في هذا المستوى من التحليل، يتم دعم الاكتشافات الحقيقية وتؤثر الأدوات على التخطيط والتنبؤ. ستحاول المستويات الأعلى من البحث (الاكتشاف) إيجاد ارتباطات مهمة، مثل تأثير المواسم و / أو الطقس على مبيعات منتجات معينة في أحداث معينة، لا سيما استخدام البيانات الضخمة لقرارات الأعمال الاستراتيجية عالية المستوى.

على مستوى أعلى، قد تكون هذه الوظائف آلية (جزئيًا) لتقديم تحليلات تنبؤية وحتى معيارية. يشير الأخير إلى القرارات المشتقة والمنفذة تلقائيًا بناءً على نتائج التحليل التلقائي (أو اليدوي). ومع ذلك، فإن مثل هذه الوظائف هي خارج نطاق أنظمة دعم القرار النموذجية ومن المرجح أن يتم تضمينها في بيئات معالجة الأحداث المعقدة (CEP) حيث يتم وزن الكمون المنخفض للقرار الآلي أعلى من الأمان الإضافي للإنسان في الحلقة التي يتم توفيرها بواسطة أنظمة دعم القرار. حيث يمكن للمستخدمين تطبيق الذكاء الاصطناعي (AI) في أنظمة دعم القرار. يطلق عليها اسم أنظمة دعم القرار الذكية(IDSS) ، حيث يقوم الذكاء الاصطناعي باستخراج كميات كبيرة من البيانات ومعالجتها للحصول على رؤى وتقديم توصيات لاتخاذ قرارات أفضل. يقوم بذلك عن طريق تحليل مصادر متعددة للبيانات وتحديد الأنماط والاتجاهات والجمعيات لمحاكاة قدرات صنع القرار البشري[1].

تم تصميم IDSS للعمل بشكل مشابه للمستشار البشري، حيث يجمع البيانات ويحللها لدعم صانعي القرار من خلال تحديد المشكلات وحلها، وتقديم الحلول الممكنة وتقييمها. يحاكي مكون الذكاء الاصطناعي في DSS القدرات البشرية بأكبر قدر ممكن، مع معالجة وتحليل المعلومات بشكل أكثر كفاءة كنظام كمبيوتر.

قد يشتمل IDSS على إمكانات متقدمة مثل قاعدة المعرفة والتعلم الآلي واستخراج البيانات وواجهة المستخدم. تتضمن أمثلة تطبيقات IDSS أنظمة التصنيع المرنة أو الذكية وأنظمة دعم اتخاذ القرارات التسويقية الذكية وأنظمة التشخيص الطبي.

الفرع الثاني: التحليل التنبئي

التحليلات التنبؤية هي فرع من التحليلات المتقدمة التي تقوم بعمل تنبؤات حول النتائج المستقبلية باستخدام البيانات التاريخية جنبًا إلى جنب مع النمذجة الإحصائية وتقنيات استخراج البيانات والتعلم الآلي.

تستخدم الشركات التحليلات التنبؤية للعثور على أنماط في هذه البيانات لتحديد المخاطر والفرص. وغالبًا ما ترتبط التحليلات التنبؤية بالبيانات الضخمة وعلوم البيانات. تسبح الشركات اليوم في البيانات الموجودة عبر

[1] https://www.techtarget.com/searchcio/definition/decision-support-system#:~:text=A%20decision%20support%20system%20(DSS)%20is%20a%20computer%20program%20application,the%20best%20possible%20options%20available , viewed at 17: 13, 13/04/2022.

قواعد بيانات المعاملات أو ملفات سجل المعدات أو الصور أو الفيديو أو أجهزة الاستشعار أو مصادر البيانات الأخرى. لاكتساب رؤى من هذه البيانات، يستخدم علماء البيانات خوارزميات التعلم العميق والتعلم الآلي للعثور على الأنماط والتنبؤ بالأحداث المستقبلية. وتشمل هذه الانحدار الخطي وغير الخطي والشبكات العصبية وآلات ناقلات الدعم وأشجار القرار. [1]يمكن بعد ذلك استخدام التعلم الذي تم الحصول عليه من خلال التحليلات التنبؤية في التحليلات الوصفية لدفع الإجراءات بناءً على الرؤى التنبؤية.

أحد الأمثلة الرئيسية للتحليل التنبئي هو الصيانة التنبؤية القائمة على استخدام البيانات الضخمة. يتم تحديد فترات الصيانة عادةً على أنها توازن بين تكرار الصيانة المكلف والعالي وخطر الفشل المكلف بنفس القدر قبل الصيانة. اعتمادًا على سيناريو التطبيق، غالبًا ما تتطلب مشكلات السلامة صيانة متكررة، على سبيل المثال، في صناعة الطيران. ومع ذلك، في حالات أخرى، لا تكون تكلفة أعطال الماكينة كارثية ويصبح تحديد فترات الصيانة عملية اقتصادية بحتة.

الافتراض الأساسي للتحليل التنبئي هو أنه بالنظر إلى معلومات المستشعر الكافية من جهاز معين وقاعدة بيانات كبيرة بما فيه الكفاية من أجهزة الاستشعار وبيانات الفشل من هذا الجهاز أو نوع الجهاز العام، يمكن التنبؤ بالوقت المحدد لفشل الجهاز بشكل أكثر دقة. يعد هذا النهج بخفض التكاليف بسبب:

- يمكن تجنب فترات الصيانة الأطول باعتبارها انقطاعات "غير ضرورية" للإنتاج (أو العمالة) عند الوصول إلى الوقت العادي للصيانة. يسمح النموذج التنبئي بتمديد فترة الصيانة، بناءً على بيانات المستشعر الحالية.
- عدد أقل من حالات الفشل حيث يمكن تقليل عدد حالات الفشل التي تحدث قبل الصيانة المجدولة بناءً على بيانات المستشعر والصيانة التنبؤية التي تتطلب أعمال صيانة سابقة.
- انخفاض تكاليف حالات الفشل حيث يمكن التنبؤ بالفشل المحتمل من خلال الصيانة التنبؤية مع وقت تحذير مسبق معين، مما يسمح بجدولة أعمال الصيانة / التبادل، وتقليل أوقات الانقطاع.

يتطلب تطبيق التحليلات التنبؤية توافر بيانات المستشعر لجهاز معين (حيث يتم استخدام "الجهاز" كمصطلح عام إلى حد ما) بالإضافة إلى مجموعة بيانات شاملة لبيانات المستشعر جنبًا إلى جنب مع بيانات الأعطال. قد يكون تجهيز الآلات الحالية بأجهزة استشعار إضافية، وإضافة مسارات اتصال من أجهزة الاستشعار إلى خدمات الصيانة التنبؤية، وما إلى ذلك، اقتراحًا مكلفًا. بناءً على تجربة إحجام عملائها عن مثل هذه الاستثمارات، طور عدد من الشركات (خاصة مصنعي الآلات) نماذج أعمال جديدة لمعالجة هذه المشكلات. ومن الأمثلة البارزة على ذلك

[1]https://www.ibm.com/analytics/predictiveanalytics#:~:text=Predictive%20analytics%20is%20a%20branch,to%20identify%20risks%20and%20opportunities , viewed at 19:40, 19/04/2022.

توربينات الرياح من جنرال إلكتريك ومحركات الطائرات من رولز رويس. يتم تقديم محركات Rolls Royce بشكل متزايد للإيجار، مع عقود خدمة كاملة بما في ذلك الصيانة، مما يسمح للشركة المصنعة برفع الفوائد من تطبيق الصيانة التنبؤية. من خلال ربط السياق التشغيلي ببيانات مستشعر المحرك، يمكن توقع الأعطال مبكرًا، مما يقلل (تكاليف) الاستبدالات، مما يسمح بالصيانة المخطط لها بدلاً من الصيانة المجدولة فقط. تقدم حلول GE OnPoint حزم خدمة مماثلة تُباع جنبًا إلى جنب مع محركات GE.

الفرع الثالث: الاستكشاف

استكشاف البيانات هو الخطوة الأولى في تحليل البيانات المستخدمة لاستكشاف البيانات وتصورها لكشف الرؤى من البداية أو تحديد المناطق أو الأنماط للبحث في المزيد. باستخدام لوحات المعلومات التفاعلية واستكشاف البيانات عن طريق التأشير والنقر، يمكن للمستخدمين فهم الصورة الأكبر بشكل أفضل والحصول على الرؤى بشكل أسرع. فيساعد البدء باستكشاف البيانات المستخدمين على اتخاذ قرارات أفضل بشأن مكان التعمق في البيانات والحصول على فهم واسع النطاق للأعمال عند طرح أسئلة أكثر تفصيلاً لاحقًا. من خلال واجهة سهلة الاستخدام، يمكن لأي شخص عبر المؤسسة التعرف على البيانات واكتشاف الأنماط وتوليد أسئلة مدروسة قد تحفز على تحليل أعمق وقيِّم.[1]

تعمل أدوات استكشاف البيانات والتحليلات المرئية على بناء الفهم، وتمكين المستخدمين من استكشاف البيانات في أي تصور. يعمل هذا النهج على تسريع وقت الإجابات وتعميق فهم المستخدمين من خلال تغطية المزيد من التفاصيل في وقت أقل. يعد استكشاف البيانات مهمًا لهذا السبب لأنه يضفي الطابع الديمقراطي على الوصول إلى البيانات ويوفر تحليلات الخدمة الذاتية المحكومة. علاوة على ذلك، يمكن للشركات تسريع استكشاف البيانات من خلال توفير البيانات وتسليمها من خلال مجموعات البيانات المرئية التي يسهل استكشافها واستخدامها.

يمكن أن يساعد استكشاف البيانات الشركات في استكشاف كميات كبيرة من البيانات بسرعة لفهم الخطوات التالية بشكل أفضل من حيث التحليل الإضافي. يمنح هذا العمل نقطة انطلاق أكثر قابلية للإدارة وطريقة لاستهداف مجالات الاهتمام. في معظم الحالات، يتضمن استكشاف البيانات استخدام تصورات البيانات لفحص البيانات على مستوى عالٍ. من خلال اتباع هذا النهج عالي المستوى، يمكن للشركات تحديد البيانات الأكثر أهمية والتي قد تشوه التحليل وبالتالي يجب إزالتها. يمكن أن يكون استكشاف البيانات مفيدًا أيضًا في تقليل الوقت المستغرق في التحليل الأقل قيمة عن طريق تحديد المسار الصحيح للأمام من البداية.

[1] https://www.tibco.com/reference-center/what-is-data-exploration#:~:text=Data%20exploration%20is%20the%20first,and%20get%20to%20insights%20faster, viewed at 20:51, 25/04/2022.

يمكن أن يضيف دعم نهج التجربة والخطأ البشري قيمة من خلال توفير طرق ذكية لاستخراج المعلومات تلقائيًا وتجميعها للإجابة على الأسئلة المعقدة. يمكن لهذه الأساليب تحويل عملية تحليل البيانات لتصبح استكشافية ومتكررة. في المرحلة الأولى، يتم تحديد البيانات ذات الصلة ثم يتم إضافة سياق مرحلة التعلم الثانية لهذه البيانات. تتيح مرحلة الاستكشاف الثالثة عمليات مختلفة لاشتقاق القرارات من البيانات أو تحويل البيانات وإثرائها.

الفرع الرابع: التحليل التكراري

في كثير من الأحيان، يكون تدفق العمليات للعديد من مشاريع البيانات الضخمة متكررًا، مما يعني الكثير من اختبار الأفكار الجديدة ذهابًا وإيابًا، وميزات جديدة لتضمينها، وتعديل العديد من المعلمات المفرطة، وما إلى ذلك، مع موقف سريع الفشل. عادة ما تكون النتيجة النهائية لهذه المشاريع نموذجًا يمكنه الإجابة على السؤال المطروح. لاحظ أننا لم نقول بدقة الإجابة على السؤال المطروح! تتمثل إحدى المآزق التي يواجهها العديد من علماء البيانات هذه الأيام في عدم قدرتهم على تعميم نموذج للبيانات الجديدة، مما يعني أنهم قد تجاوزوا بياناتهم بحيث يقدم النموذج نتائج سيئة عند إعطائهم بيانات جديدة. تعتمد الدقة بشكل كبير على المهمة وعادة ما تملها احتياجات العمل مع بعض تحليل الحساسية الذي يتم إجراؤه لموازنة التكلفة والفوائد لنتائج النموذج. ومع ذلك، هناك عدد قليل من مقاييس الدقة القياسية التي سنستعرضها خلال هذا الكتاب حتى تتمكن من مقارنة النماذج المختلفة لمعرفة كيف تؤثر التغييرات التي تم إجراؤها على النموذج على النتيجة.

تجلب المعالجة المتوازية والفعالة لتدفقات البيانات التكرارية عددًا من التحديات التقنية. عادةً ما تحسب عمليات تحليل البيانات التكرارية نتائج التحليل في سلسلة من الخطوات. في كل خطوة، يتم حساب نتيجة أو حالة وسيطة جديدة وتحديثها. نظرًا للأحجام الكبيرة في تطبيقات البيانات الضخمة، يتم تنفيذ العمليات الحسابية بالتوازي، وتوزيع وتخزين وإدارة الحالة بكفاءة عبر أجهزة متعددة. تحتاج العديد من الخوارزميات إلى عدد كبير من التكرارات لحساب النتائج النهائية، مما يتطلب تكرارات بزمن انتقال منخفض لتقليل أوقات الاستجابة الإجمالية. ومع ذلك، في بعض التطبيقات، يتم تقليل الجهد الحسابي بشكل كبير بين التكرار الأول والأخير. الأنظمة القائمة على الدُفعات مثل MapReduce وApache Spark، تكرر جميع العمليات الحسابية في كل تكرار حتى عندما لا تتغير النتائج (الجزئية). تستغل أنظمة تدفق البيانات التكرارية حقًا مثل ستراتوسفير لأنظمة الرسوم البيانية المتخصصة مثل GraphLab وPregel Google هذه الخصائص وتقليل التكلفة الحسابية في كل تكرار.[1]

الفرع الخامس: التصور للبيانات الضخمة

[1] Dean. J: MapReduce: Simplified data processing on large clusters. Communications of the ACM, 51(1), 2008, pp. 107–113.

يشير تصور البيانات الضخمة إلى تنفيذ تقنيات التصور الأكثر حداثة لتوضيح العلاقات داخل البيانات. تتضمن أساليب التصور التطبيقات التي يمكنها عرض التغييرات في الوقت الفعلي والمزيد من الرسومات التوضيحية، وبالتالي تتجاوز الرسوم البيانية الدائرية والشريطية والمخططات الأخرى. تنحرف هذه الرسوم التوضيحية عن استخدام مئات الصفوف والأعمدة والسمات نحو تمثيل بصري فني أكثر للبيانات.

يعد تصور نتائج التحليل بما في ذلك عرض الاتجاهات والتنبؤات الأخرى بواسطة أدوات التصور المناسبة جانبًا مهمًا من استخدام البيانات الضخمة. يعد اختيار المعلمات والمجموعات الفرعية والميزات ذات الصلة عنصرًا حاسمًا في استخراج البيانات والتعلم الآلي مع العديد من الدورات اللازمة لاختبار الإعدادات المختلفة. نظرًا لتقييم الإعدادات على أساس نتائج التحليل المقدمة، يسمح التصور عالي الجودة بإجراء تقييم سريع ودقيق لجودة النتائج، على سبيل المثال، في التحقق من الجودة التنبؤية لنموذج من خلال مقارنة النتائج مع مجموعة بيانات الاختبار. بدون التصور الداعم، يمكن أن تكون هذه عملية مكلفة وبطيئة، مما يجعل التصور عاملاً مهمًا في تحليل البيانات.

لاستخدام نتائج تحليلات البيانات في الخطوات اللاحقة لسيناريو استخدام البيانات، على سبيل المثال، السماح لعلماء البيانات وصناع القرار في مجال الأعمال باستخلاص استنتاجات من التحليل، يمكن أن يكون العرض التقديمي المرئي المختار جيدًا أمرًا حاسمًا لجعل مجموعات النتائج الكبيرة قابلة للإدارة وفعال. اعتمادًا على مدى تعقيد التصورات، يمكن أن تكون مكلفة من الناحية الحسابية وتعيق الاستخدام التفاعلي للتصور.ومع ذلك، فإن البحث الاستكشافي في نتائج التحليلات ضروري للعديد من حالات استخدام البيانات الضخمة. في بعض الحالات، سيتم تطبيق نتائج تحليل البيانات الضخمة على حالة واحدة فقط، مثل محرك الطائرة. ومع ذلك، في كثير من الحالات، ستكون مجموعة بيانات التحليل معقدة مثل البيانات الأساسية، حيث تصل إلى حدود تقنيات التصور الإحصائي الكلاسيكية وتتطلب استكشافًا وتحليلاً تفاعليًا. [1]

في عمل شنايدرمان الأساسي حول التصور، حدد سبعة أنواع من المهام: نظرة عامة، والتكبير / التصغير، التصفية، التفاصيل عند الطلب، الربط، التاريخ والاستخراج.[2]

مع ذلك، هناك مجال آخر من التصور ينطبق على نماذج البيانات المستخدمة في العديد من خوارزميات التعلم الآلي وتختلف عن التنقيب عن البيانات التقليدية وتطبيقات إعداد التقارير. عند استخدام نماذج البيانات هذه للتصنيف والتجميع والتوصيات والتنبؤات، يتم اختبار جودتها باستخدام مجموعات بيانات مفهومة جيدًا. يدعم التصور هذا التحقق من الصحة وتكوين النماذج ومعلماتها.

[1] Spence. R: Information visualization – design for interaction, 2nd edition, Upper Saddle River, NJ: Prentice Hall, 2006.
[2] Shneiderman. B: The eyes have it: A task by data type taxonomy for information visualizations, In Proceedings of Visual Languages, 2002.

أخيرًا، يمثل الحجم الهائل لمجموعات البيانات تحديًا مستمرًا لأدوات التصور مدفوعة بالتقدم التكنولوجي في وحدات معالجة الرسومات والشاشات والاعتماد البطيء لبيئات التصور الغامرة مثل الكهوف والواقع الافتراضي والواقع المعزز. يتم تغطية هذه الجوانب في مجالات التصور العلمي والمعلوماتي.

المطلب الثالث: أهمية استخدام البيانات الضخمة في العلوم الاقتصادية

تقوم العديد من المؤسسات المتقدمة مثل غوغل بمعالجة ملايين البيانات الرقمية وتحويلها إلى معلومات تستفيد منها للتوصل إلى أخذ قرارات سليمة بالإضافة إلى التنبؤ بالمستقبل فيما يخص عملهم، فالبيانات الضخمة تعتبر مثل العديد من التكنولوجيات الجديدة، حيث تتمثل أهميتها في امكانية تخفيض التكاليف وإدخال تحسينات كبيرة وعروض منتجات وخدمات جديدة بناءًا على رأي الجمهور أو المستهلك وبإمكانها أيضا دعم القرارات التجارية الداخلية فهي تسمح للمنظمات بتحقيق مجموعة متنوعة من الأهداف. وفي هذا الجزء من البحث، نحاول ابراز مدى أهمية استخدام البيانات الضخمة في مجال العلوم الاقتصادية من خلال النقاط التالية:

الفرع الأول: السوق

يعد سوق تكنولوجيا البيانات الضخمة في المجالات المالية والتأمينية من أكثر الأسواق الواعدة. وفقًا لتوقعات TechNavio، سينمو سوق البيانات الضخمة العالمية في قطاع الخدمات المالية بمعدل نمو سنوي مركب يبلغ 56.7٪ خلال الفترة 2012-2016. أحد العوامل الرئيسية التي تساهم في نمو هذا السوق هو الحاجة إلى تلبية اللوائح المالية، لكن الافتقار إلى الموارد الماهرة لإدارة البيانات الضخمة يمكن أن يشكل تحديًا.

البائعون الرئيسيون المهيمنون على هذه المساحة هم Hewlett-Packard وIBM وMicrosoft وOracle وهم لاعبون عالميون راسخون يتمتعون بملف تعريف عام. ومع ذلك، فإن جاذبية السوق ستكون عامل جذب للوافدين الجدد في السنوات القادمة.

نظرًا لأن البيانات هي أهم الأصول، فإن هذه التكنولوجيا مواتية ومميزة بشكل خاص لمؤسسات الخدمات المالية، كما ذكر تقرير معهد IBM لقيمة الأعمال "التحليلات: الاستخدام الفعلي للبيانات الضخمة في الخدمات المالية". من خلال الاستفادة من هذا الأصل، يمكن للبنوك وشركات الأسواق المالية اكتساب فهم شامل للأسواق والعملاء والقنوات والمنتجات واللوائح والمنافسين والموردين والموظفين مما يتيح لهم المنافسة بشكل أفضل. لذلك، يعد هذا اتجاهًا إيجابيًا في السوق ومن المتوقع أن يقود نمو سوق البيانات الضخمة العالمية في قطاع الخدمات المالية.

فيما يتعلق باستراتيجية البيانات، تتبع مؤسسات الخدمات المالية نهجًا مدفوعًا بالأعمال تجاه البيانات الضخمة،

يتم تحديد متطلبات العمل في المقام الأول ثم يتم مواءمة الموارد والقدرات الداخلية الحالية لدعم فرص العمل، قبل الاستثمار في مصادر البيانات والبنى التحتية. ومع ذلك، لا تحافظ جميع المؤسسات المالية على نفس الوتيرة. وفقًا لتقرير IBM، بينما يركز 26٪ على فهم المفاهيم الأساسية (مقارنة بـ 24٪ من المؤسسات العالمية)، فإن الغالبية إما تحدد خارطة طريق تتعلق بالبيانات الضخمة (47٪) أو تجري بالفعل عمليات تجريبية وتنفيذية للبيانات الضخمة (27٪). حيث يتخلفون عن أقرانهم عبر الصناعة في استخدام أنواع بيانات أكثر تنوعًا ضمن تطبيقات البيانات الضخمة الخاصة بهم. تقوم أكثر من 21٪ بقليل من هذه الشركات بتحليل البيانات الصوتية (غالبًا ما يتم إنتاجها بكثرة في مراكز الاتصال التابعة لبنوك التجزئة)، بينما أفاد أكثر من 27٪ بقليل بتحليل البيانات الاجتماعية (مقارنة بـ 38٪ و43٪ على التوالي، من بين أقرانهم في الصناعة). يُعزى هذا النقص في التركيز على البيانات غير المهيكلة إلى الكفاح المستمر لدمج البيانات المنظمة الضخمة للمؤسسات.[1]

الفرع الثاني: المالية

يمكن أن يجلب ظهور البيانات الضخمة في الخدمات المالية مزايا عديدة للمؤسسات المالية. يتم إبراز الفوائد التي تأتي مع أكبر تأثير تجاري على النحو التالي:

- **مستويات محسّنة من رؤية العملاء والمشاركة والخبرة:** مع رقمنة المنتجات والخدمات المالية والاتجاه المتزايد للعملاء الذين يتفاعلون مع العلامات التجارية أو المؤسسات في الفضاء الرقمي، هناك فرصة لمؤسسات الخدمات المالية لتعزيز مستوى مشاركة العملاء وتحسين تجربة العملاء بشكل آني. يجادل الكثيرون بأن هذا هو المجال الأكثر أهمية بالنسبة للمؤسسات المالية للبدء في الاستفادة من تكنولوجيا البيانات الضخمة للبقاء في المقدمة، أو حتى لمواكبة المنافسة. للمساعدة في تحقيق ذلك، يمكن لتقنيات البيانات الضخمة والتقنيات التحليلية أن تساعد في استنباط نظرة ثاقبة من مصادر غير منظمة حديثة مثل وسائل التواصل الاجتماعي.
- **قدرات معززة للكشف عن الاحتيال والوقاية منه:** لطالما كانت مؤسسات الخدمات المالية عرضة للاحتيال. هناك أفراد ومنظمات إجرامية تعمل على الاحتيال على المؤسسات المالية ويتطور تعقيد هذه المخططات مع مرور الوقت. في الماضي، كانت البنوك تحلل عينة صغيرة من المعاملات في محاولة لاكتشاف الاحتيال. قد يؤدي ذلك إلى تسلل بعض الأنشطة الاحتيالية عبر الشبكة وإبراز "الإيجابيات الكاذبة" الأخرى. يعني استخدام البيانات الضخمة أن هذه المؤسسات قادرة الآن على استخدام مجموعات بيانات أكبر لتحديد الاتجاهات التي تشير إلى الاحتيال للمساعدة في تقليل التعرض لمثل هذه المخاطر.

[1] José María Cavanillas: New Horizons for a Data-Driven Economy, previous reference, p210.

- **تحليل محسن للتداول في السوق:** بدأ التداول في الأسواق المالية في التحول إلى مساحة رقمية منذ سنوات عديدة، مدفوعًا بالطلب المتزايد على تنفيذ أسرع للصفقات. تعتبر استراتيجيات التداول التي تستخدم الخوارزميات المتطورة للتداول السريع في الأسواق المالية من أهم المستفيدين من البيانات الضخمة.

يمكن اعتبار بيانات السوق نفسها بيانات ضخمة. إنه حجم كبير، ويتولد من مجموعة متنوعة من المصادر، ويتولد بسرعة هائلة. ومع ذلك، لا تُترجم هذه البيانات الضخمة بالضرورة إلى معلومات قابلة للتنفيذ. تكمن الفائدة الحقيقية من البيانات الضخمة في الاستخراج الفعال للمعلومات القابلة للتنفيذ ودمج هذه المعلومات مع المصادر الأخرى. يمكن دمج بيانات السوق من أسواق ومناطق جغرافية متعددة بالإضافة إلى مجموعة متنوعة من فئات الأصول مع مصادر أخرى منظمة وغير منظمة لإنشاء مجموعات بيانات غنية ومختلطة (مجموعة من البيانات المهيكلة وغير المهيكلة). يوفر هذا رؤية شاملة ومتكاملة لحالة السوق ويمكن استخدامه في مجموعة متنوعة من الأنشطة مثل إنشاء الإشارات وتنفيذ التجارة وإعداد تقارير الأرباح والخسائر وقياس المخاطر، كل ذلك في الوقت الفعلي وبالتالي تمكين تداول أكثر فعالية.

الفرع الثالث: القطاع العام

يمكن تصنيف أهمية البيانات الضخمة في القطاع العام في ثلاثة مجالات رئيسية، بناءً على تصنيف أنواع الفوائد:[1]

- **تحليلات البيانات الضخمة:** يغطي هذا المجال التطبيقات التي لا يمكن إجراؤها إلا من خلال الخوارزميات الآلية للتحليلات المتقدمة لتحليل مجموعات البيانات الكبيرة لحل المشكلات التي يمكن أن تكشف عن رؤى تعتمد على البيانات. يمكن استخدام هذه القدرات لاكتشاف الأنماط والتعرف عليها أو لإنتاج تنبؤات. تشمل التطبيقات في هذا المجال كشف الاحتيال؛ الإشراف على الأنشطة المنظمة للقطاع الخاص؛ تحليل المشاعر لمحتوى الإنترنت لتحديد أولويات الخدمات العامة؛ الكشف عن التهديدات من مصادر البيانات الخارجية والداخلية لمنع الجريمة والاستخبارات والأمن؛ والتنبؤ لأغراض التخطيط للخدمات العامة.
- **تحسينات في الفعالية:** تغطي تطبيق البيانات الضخمة لتوفير قدر أكبر من الشفافية الداخلية. يمكن للمواطنين والشركات اتخاذ قرارات أفضل وأن يكونوا أكثر فعالية، وحتى إنشاء منتجات وخدمات جديدة بفضل المعلومات المقدمة. تتضمن بعض أمثلة التطبيقات في هذا المجال توافر البيانات عبر الصوامع التنظيمية؛ تبادل المعلومات من خلال منظمات القطاع العام على سبيل المثال تجنب المشاكل الناجمة عن عدم وجود قاعدة بيانات واحدة للهوية؛ الحكومة المفتوحة والبيانات المفتوحة تسهل التدفق الحر

[1] McKinsey Global Institute: Big Data: The next frontier for innovation, competition, and productivity. McKinsey & Company, June 2011.

للمعلومات من المؤسسات العامة إلى المواطنين والشركات، وإعادة استخدام البيانات لتقديم خدمات جديدة ومبتكرة للمواطنين.

- **تحسينات في الكفاءة:** يغطي هذا المجال التطبيقات التي تقدم خدمات أفضل وتحسينًا مستمرًا على أساس تخصيص الخدمات والتعلم من أداء هذه الخدمات. بعض الأمثلة على التطبيقات في هذا المجال هي تخصيص الخدمات العامة للتكيف مع احتياجات المواطنين وتحسين الخدمات العامة من خلال التحليلات الداخلية القائمة على تحليل مؤشرات الأداء.

الفرع الرابع: الصناعة

المتطلبات الأساسية في قطاع التصنيع هي تخصيص المنتجات والإنتاج -" حجم القطعة واحد" -تكامل الإنتاج في سلسلة قيمة المنتج الأكبر، وتطوير المنتجات الذكية. حيث تشهد الصناعة التحويلية تغيرات جذرية مع إدخال تكنولوجيا تكنولوجيا المعلومات على نطاق واسع. تشمل التطورات في إطار "الصناعة 4.0" عددًا متزايدًا من أجهزة الاستشعار والاتصال في جميع جوانب عملية الإنتاج. وبالتالي، فإن الحصول على البيانات يهتم بجعل البيانات المتاحة بالفعل قابلة للإدارة، أي أن التوحيد القياسي وتكامل البيانات هما أكبر المتطلبات. تم تطبيق تحليل البيانات بالفعل في التطبيقات داخل اللوحات الجدارية وسيكون مطلوبًا لمزيد من التطبيقات المتكاملة التي تغطي سلاسل لوجستية كاملة عبر المصانع في سلسلة الإنتاج وحتى في استخدام ما بعد البيع للمنتجات (الذكية). يحتاج تخطيط الإنتاج إلى أن يكون مدعومًا بالمحاكاة المستندة إلى البيانات لهذه البيئات الكاملة.

يمكن للآلات المعقدة والذكية، مثل محركات الطائرات، الاستفادة من الصيانة التنبؤية الكبيرة القائمة على قاعدة البيانات حيث يتم استخدام معلومات المستشعر والسياق مع خوارزميات التعلم الآلي لتجنب الصيانة غير الضرورية وجدولة الإصلاحات الوقائية عند توقع حدوث أعطال. نظرًا لتكاليف البنية التحتية الإضافية، يستخدم المصنعون نماذج أعمال جديدة حيث يتم تأجير الآلات وعدم بيعها؛ وبالتالي، فإن بيانات وخدمات أجهزة الاستشعار مملوكة ومُنفذة من قبل الشركة المصنعة وليس مستخدم الآلات. هذا يؤدي إلى تحديات في اللوائح والعقود المتعلقة بملكية البيانات.

يمكن أن يكون قطاع التصنيع الأوروبي رائدًا في السوق باستخدام البيانات الضخمة في سياق الصناعة 4.0، وسوقًا رائدًا، حيث يتم دمج تصنيع البيانات الضخمة في سلسلة قيمة المنتج الأكبر ويمكن استخدام المنتجات الذكية.

الفرع الخامس: الابتكار

الابتكار هو عملية تكرارية تهدف إلى إنشاء منتجات أو عمليات أو معارف أو خدمات جديدة باستخدام

المعرفة الجديدة أو حتى الموجودة، يستلزم الابتكار القائم على البيانات استغلال أي نوع من البيانات في عملية الابتكار لخلق قيمة، يقود الاتجاه الناشئ للابتكار القائم على البيانات الضخمة إلى تطوير السلع والخدمات التي تعتمد على البيانات ويمكن أن تمكن من التخطيط المستند إلى البيانات والتسويق المستند إلى البيانات والعمليات القائمة على البيانات في جميع القطاعات والمجالات الصناعية. من المنظور الاقتصادي، فإن البيانات باعتبارها سلعة أو مشاعات غير متنافسة مثل النفط تخدم موردًا للبنية التحتية (من منظور وظيفي) يمكن استغلاله في وقت واحد من قبل العديد من المستخدمين أو الجهات الفاعلة لغايات منافسة أو تكميلية مختلفة. الطلب على البيانات بهذا المعنى وفقًا لمنظمة التعاون الاقتصادي والتنمية مدفوع بشكل أساسي بالأنشطة الإنتاجية النهائية التي تتطلب البيانات كمدخل، وفي الواقع، رأس مال غير تافه. بالإضافة إلى ذلك، يؤكد المؤلفون أنفسهم أنه يمكن استخدام موارد البيانات كمدخلات في مجموعة متنوعة من السلع، بما في ذلك السلع الخاصة والعامة والاجتماعية. بعبارة أخرى، من المحتمل أن توفر البيانات الضخمة عوائد كبيرة على الحجم والنطاق.

ترتبط الابتكارات القائمة على البيانات الضخمة ضمنيًا بنموذج سلسلة القيمة أو بشكل أكثر دقة "سلسلة القيمة الافتراضية" التي تحدد كيفية جمع البيانات محل الاهتمام وتنظيمها واختيارها وتحويلها إلى منتجات أو خدمات وتوزيعها.

على المستوى التنظيمي، يمكن أن تنجم فئتان على الأقل من المبادرات الإستراتيجية عن الابتكار القائم على البيانات الضخمة وسلسلة القيمة الأساسية للبيانات الضخمة. تهدف الفئة الأولى من المبادرات إلى إتاحة المعلومات حول جوانب العمليات والخدمات التنظيمية لتمكين التحسينات. بشكل عام، من خلال أدوات العمليات التنظيمية، يتم إنشاء كميات كبيرة من البيانات (أي البيانات الضخمة) التي تُعلِم أو تقود التغييرات المطلوبة. المجموعة الثانية من المبادرات هي مواجهة خارجية وتنطوي على استغلال بيانات العملاء مثل البحث وسجلات المستخدم وسجلات المعاملات والمحتويات الأخرى التي ينشئها العميل لدفع التسويق طويل الأمد والتوصية المستهدفة والشخصية وزيادة المبيعات ورضا العملاء. ومن الأمثلة الشائعة على ذلك خوارزمية التصفية التعاونية لـ Netflix للتنبؤ بتصنيفات أفلام المستخدم. مثال آخر هو استخدام Google لسلوك بحث المستخدمين لاستهداف الإعلانات.

في الولايات المتحدة، تستخدم مئات الشركات البيانات المفتوحة والكبيرة (مثل بيانات الطقس وبيانات نظام تحديد المواقع العالمي) كمصادر رئيسية لتوليد القيمة عبر مختلف القطاعات بما في ذلك التمويل والاستثمار والتعليم والبيئة والطقس والإسكان والعقارات والغذاء والزراعة. يوضح القسم التالي بالتفصيل عددًا من التحولات القائمة على البيانات عبر مختلف القطاعات بما في ذلك الاتصالات والرعاية الصحية والقطاع العام والتمويل والتأمين والإعلام والترفيه والطاقة والنقل.

المبحث الثاني: تحديات استخدام البيانات الضخمة

يمكن لاستراتيجية البيانات الضخمة التي يتم تنفيذها بشكل جيد أن تعمل على تبسيط التكاليف التشغيلية وتقليل الوقت اللازم للتسويق وتمكين المنتجات الجديدة. لكن الشركات تواجه مجموعة متنوعة من تحديات البيانات الضخمة في نقل المبادرات من مناقشات مجلس الإدارة إلى الممارسات الناجحة.

قالBill Szybillo : "ربما الأهم أن الشركات تحتاج إلى معرفة كيف ولماذا تعتبر البيانات الضخمة مهمة لأعمالها في المقام الأول". مدير ذكاء الأعمال في مزود برمجياتERP VAI . فيتمثل أحد أكبر التحديات المتعلقة بمشاريع البيانات الضخمة في التطبيق الناجح للرؤى التي تم الحصول عليها.[1]كما أوضح أن العديد من التطبيقات والأنظمة تلتقط البيانات، لكن المنظمات غالبًا ما تكافح لفهم ما هو ذي قيمة، ومن هناك، لتطبيق تلك الأفكار بطريقة مؤثرة.

المطلب الأول: المشاكل التحليلية والفنية والمعيقات التقنية

في كثير من الأحيان، تفشل الشركات في معرفة حتى الأساسيات: ما هي البيانات الضخمة في الواقع، وما هي فوائدها، وما هي البنية التحتية المطلوبة، وما إلى ذلك. بدون فهم واضح، فإن مشروع اعتماد البيانات الضخمة قد يكون محكومًا عليه بالفشل. قد تضيع الشركات الكثير من الوقت والموارد في أشياء لا يعرفون حتى كيفية استخدامها.

الفرع الأول: المشاكل التحليلية والفنية

يفتح الافتقار إلى فهم كيفية التعامل مع البيانات الضخمة قائمة تحديات البيانات لدينا. عندما تبدأ الشركات في الانتقال إلى المنتجات الرقمية التي تستخدم البيانات الضخمة، فقد لا يكون موظفوها مستعدين للعمل مع مثل هذه الحلول المتقدمة. نتيجة لذلك، يمكن أن يتسبب التنفيذ مع موظفين غير مدربين في حدوث تباطؤ كبير في عمليات العمل، واضطراب في تدفقات العمل المألوفة، والعديد من الأخطاء. حتى يدرك موظفوك الفوائد الكاملة للابتكار ويتعلمون كيفية استخدامها، قد يكون هناك انخفاض في الإنتاجية.

للتغلب على تحديات البيانات هذه، من المهم جدًا ربط المتخصصين المؤهلين أو تدريب المختصين الحاليين بسير العمل الحالي جنبًا إلى جنب مع إنشاء واعتماد حلول رقمية متقدمة جديدة. كما تبين الممارسة، فإن الخيار البديل ليس دائمًا فعالاً، لأن الموظفين سيحتاجون إلى بعض الوقت ليتم تدريبهم. علاوة على ذلك، ستجلب الحلول الرقمية الجديدة أعباء عمل إضافية لقسم تكنولوجيا المعلومات لديك. لذلك، من الأفضل بكثير إما الجمع بين

[1] https://www.techtarget.com/searcherp/The-ultimate-guide-to-ERP , viewed at 21:12, 01/05/2022.

التدريب وتوظيف متخصصين جدد أو العثور على فريق كامل الموظفين ومخصص مقدمًا من شركات تطوير البرمجيات والذي سيتولى مسؤولية دعم البرامج الجديدة.

وإذا لم يفهم الموظفون قيمة البيانات الضخمة و / أو لا يريدون تغيير العمليات الحالية من أجل اعتمادها، فيمكنهم مقاومتها وإعاقة تقدم الشركة.

قد يكون من السهل أن تضيع في مجموعة متنوعة من تقنيات البيانات الضخمة المتوفرة الآن في السوق. هل تحتاج إلى Spark أم أن سرعات Hadoop MapReduce ستكون كافية؟ هل من الأفضل تخزين البيانات في Cassandra أو HBase؟ يمكن أن يكون العثور على الإجابات خادعًا. ومن الأسهل أيضًا الاختيار بشكل سيئ، إذا كنت تستكشف محيطًا من الفرص التكنولوجية دون رؤية واضحة لما تحتاجه.

تمتلك فرق إدارة البيانات مجموعة واسعة من تقنيات البيانات الضخمة للاختيار من بينها، وغالبًا ما تتداخل الأدوات المختلفة من حيث قدراتها.

يوصيLenley Hensarling ، كبير مسؤولي الإستراتيجيات في شركة Aerospike لقاعدة بياناتNoSQL ، بأن تبدأ الفرق من خلال النظر في الاحتياجات الحالية والمستقبلية للبيانات من الدفق ومصادر الدُفعات، مثل الحواسيب المركزية والتطبيقات السحابية وخدمات البيانات التابعة لجهات خارجية. على سبيل المثال، منصات البث على مستوى المؤسسات التي يجب مراعاتها تشمل Apache Kafka و Apache Pulsarو AWS Kinesisو Google Pub/Subوكلها توفر حركة سلسة للبيانات بين أنظمة السحابة السحابية المحلية وأنظمة السحابة المختلطة، على حد قوله. [1]بعد ذلك، يجب أن تبدأ الفرق في تقييم قدرات إعداد البيانات المعقدة المطلوبة لتغذية الذكاء الاصطناعي والتعلم الآلي وأنظمة التحليلات المتقدمة الأخرى. من المهم أيضًا التخطيط للمكان الذي قد تتم فيه معالجة البيانات. في الحالات التي يكون فيها وقت الاستجابة مشكلة، تحتاج الفرق إلى التفكير في كيفية تشغيل التحليلات ونماذج الذكاء الاصطناعي على خوادم الحافة، وكيفية تسهيل تحديث النماذج. يجب موازنة هذه الإمكانات مقابل تكلفة نشر وإدارة المعدات والتطبيقات التي يتم تشغيلها في أماكن العمل أو في السحابة أو على الحافة.

ومع تقدم التكنولوجيا الرقمية، تتغير أيضًا أهداف أعمال الشركات واحتياجات عملائها. من وجهة نظر التحديات في تحليلات البيانات الضخمة، يشير هذا إلى أنها يجب أن تكون محدثة، مما يعني أن بعضًا منها، والتي كانت ذات صلة بالأمس، قد تكون قديمة بالفعل. بالإضافة إلى ذلك، فإن جائحةCOVID-19 ، الذي غيّر بشكل

[1] https://www.techtarget.com/searchdatamanagement/tip/10-big-data-challenges-and-how-to-address-them , viewed at 16:49, 03/05/2022.

كبير الأنماط المعتادة للمستخدمين، يؤدي إلى تفاقم مشكلة الصلة بالموضوع. هذا يعني أنه لم يعد بإمكانك الاعتماد على تحليلات البيانات التاريخية للتسويق وتحليل المستهلك[1].

من وجهة نظر فنية، تكمن تحديات البيانات هذه في الحاجة إلى أداة توفر ترشيحًا محدثًا للبيانات غير ذات الصلة وتقصير دورة المعالجة للبيانات الجديدة بحيث يتم تقديم الابتكارات في أسرع وقت ممكن.

على وجه الخصوص، سيتعين التفكير في طريقة لتحديد أولويات البيانات الضخمة وتقسيمها بحيث تستغرق معالجتها حدًا أدنى من الوقت، ويؤدي كل تكرار إلى نتيجة مهمة للشركة. هذا هو المكان الذي تكون فيه المنهجية الرشيقة في متناول اليد، والتي، بالمناسبة، قابلة للتطبيق ليس فقط على تطوير البرامج.

أيضًا، يجب توفير البرمجة الآلية حيثما أمكن ذلك. قد يلعب الذكاء الاصطناعي دورًا، والذي سيكون قادرًا على تحمل مسؤولية معالجة وتحليل التدفقات الجديدة غير المنظمة للمعلومات. أيضًا، لا ننسى إجراء تحليل متعمق للبيانات الموجودة بالفعل، للتخلص من البيانات الغير ذات الصلة.

كما يمكننا حصر المشاكل التحليلية والفنية التي تواجهها البيانات الضخمة في النقاط الأساسية التالية:[2]

- **الحدود الإدارية وحماية البيانات:** يشكل تجميع البيانات عبر الحدود الإدارية بطريقة غير قائمة على الطلب تحديًا حقيقيًا، نظرًا لأن هذه المعلومات قد تكشف عن معلومات شخصية وأمنية شديدة الحساسية عند دمجها مع مصادر بيانات أخرى مختلفة، مما يؤدي ليس فقط إلى المساس بالخصوصية الفردية ولكن أيضًا الأمن المدني. يجب تبرير حقوق الوصول إلى مجموعات البيانات المطلوبة لعملية ما والحصول عليها. عند إجراء عملية جديدة على بيانات موجودة، يجب الحصول على إشعار أو ترخيص من وكالة خصوصية البيانات. يجب الحفاظ على المجهولية في هذه الحالات، لذلك يلزم فصل البيانات. يجب معالجة مخاوف الخصوصية والأمن العام الفردية قبل إقناع الحكومات بمشاركة البيانات بشكل أكثر انفتاحًا، ليس فقط علنًا ولكن مشاركتها بطريقة مقيدة مع الحكومات الأخرى أو الكيانات الدولية. البعد الآخر هو تنظيم استخدام الحوسبة السحابية بطريقة يمكن للقطاع العام أن يثق فيها بمقدمي الخدمات السحابية. علاوة على ذلك، فإن الافتقار إلى موفري الحوسبة السحابية للبيانات الضخمة الأوروبية داخل السوق الأوروبية يمثل أيضًا عائقًا أمام التبني.

  على وجه الخصوص في الاتحاد الأوروبي، هناك العديد من قضايا حماية البيانات والخصوصية التي يجب مراعاتها عند إجراء تحليلات البيانات الضخمة. تملي المتطلبات التنظيمية أنه يجب معالجة البيانات الشخصية لأغراض محددة وقانونية وأن المعالجة يجب أن تكون كافية وذات صلة وليست مفرطة.

[1] https://nix-united.com/blog/12-big-data-issues-growing-companies-face/ , viewed at, 17:35, 03/05/2022.
[2] José María Cavanillas: New Horizons for a Data-Driven Economy, previous reference, p201.

إن تأثير هذه المبادئ على مؤسسات الخدمات المالية كبير، حيث يستطيع الأفراد مطالبة مؤسسات الخدمات المالية بإزالة أو الامتناع عن معالجة بياناتهم الشخصية في ظروف معينة. قد يؤدي هذا المطلب إلى زيادة تكاليف مؤسسات الخدمات المالية، لأنها تتعامل مع طلبات الأفراد. قد تؤدي إزالة البيانات هذه أيضًا إلى انحراف مجموعة البيانات، حيث ستكون مجموعات معينة من الأشخاص أكثر نشاطًا وإدراكًا لحقوقهم من الآخرين.

- **مهارات البيانات الضخمة:** قال مايك أومالي، نائب الرئيس الأول للاستراتيجية في شركة Seneca Global، وهي شركة لتطوير البرمجيات والاستعانة بمصادر خارجية لتكنولوجيا المعلومات، "إن أحد أكبر التحديات المتعلقة بتطوير برمجيات البيانات الضخمة هو العثور على العمال ذوي مهارات البيانات الضخمة والاحتفاظ بهم".[1] حيث أن هناك نقص في علماء البيانات المهرة والتقنيين الذين يمكنهم التقاط مصادر البيانات الجديدة هذه ومعالجتها. عندما يتم اعتماد تقنيات البيانات الضخمة بشكل متزايد في الأعمال التجارية، سيصبح من الصعب العثور على متخصصي البيانات الضخمة المهرة. يمكن لوكالات الهيئات العامة أن تقطع مسافة معقولة بالمهارات التي تمتلكها بالفعل، ولكن بعد ذلك سيحتاجون إلى التأكد من تقدم هذه المهارات. إلى جانب الأشخاص الموجهين تقنيًا، هناك نقص في المعرفة لدى الأشخاص الموجودين في مجال الأعمال والذين يدركون ما يمكن أن تفعله البيانات الضخمة لمساعدتهم على حل تحديات القطاع العام.
- **السرية والمتطلبات التنظيمية:** أي معلومات تتعلق بطرف ثالث يخضع لتحليلات البيانات الضخمة من المرجح أن تكون معلومات سرية. لذلك، ستحتاج مؤسسات الخدمات المالية إلى التأكد من امتثالها لالتزاماتها وأن أي استخدام لهذه البيانات لا يؤدي إلى انتهاك السرية أو الالتزامات التنظيمية.
- **قضايا المسؤولية:** فقط لأن البيانات الضخمة تحتوي على كمية هائلة من المعلومات، فهذا لا يعني أنها تعكس عينة تمثيلية من السكان. لذلك، هناك خطر سوء تفسير المعلومات المنتجة وقد تنشأ المسؤولية عندما يتم الاعتماد على تلك المعلومات. هذا عامل يجب على مؤسسات الخدمات المالية مراعاته عند النظر في استخدام البيانات الضخمة في النماذج التحليلية والتأكد من أن أي اعتماد على المخرجات يأتي مع إخلاء المسؤولية ذات الصلة.

الفرع الثاني: المعيقات التقنية

تستلزم مشاريع اعتماد البيانات الضخمة الكثير من النفقات. إذا تم اختيار حلًا محليًا، فسيتعين التفكير في تكاليف الأجهزة الجديدة والتعيينات الجديدة (المشرفون والمطورون) والكهرباء وما إلى ذلك. بالإضافة إلى ذلك:

[1] https://www.techtarget.com/searchdatamanagement/tip/10-big-data-challenges-and-how-to-address-them , viewed at 21:38, 05/05/2022.

على الرغم من أن الأطر المطلوبة مفتوحة المصدر، إلا أنك ستظل بحاجة إلى الدفع مقابل تطوير البرامج الجديدة وإعدادها وتكوينها وصيانتها. إذا قررت حل البيانات الضخمة المستند إلى السحابة، فستظل بحاجة إلى تعيين موظفين (كما هو مذكور أعلاه) والدفع مقابل الخدمات السحابية وتطوير حلول البيانات الضخمة بالإضافة إلى إعداد وصيانة الأطر المطلوبة. علاوة على ذلك، في كلتا الحالتين، ستحتاج إلى السماح بالتوسعات المستقبلية لتجنب خروج نمو البيانات الضخمة عن السيطرة وتكلفك ثروة.[1]

عاجلاً أم آجلاً، عند التعامل مع البيانات الضخمة ستواجه مشكلة تكامل البيانات، نظرًا لأن البيانات التي تحتاج إلى تحليلها تأتي من مصادر متنوعة في مجموعة متنوعة من التنسيقات المختلفة. على سبيل المثال، تحتاج شركات التجارة الإلكترونية إلى تحليل البيانات من سجلات مواقع الويب ومراكز الاتصال و "عمليات مسح" مواقع الويب الخاصة بالمنافسين ووسائل التواصل الاجتماعي. من الواضح أن تنسيقات البيانات ستختلف، وقد يكون مطابقتها مشكلة. على سبيل المثال، يجب أن يعرف الحل الخاص بك أن الزلاجات المسماة SALOMON QST 17/18 92 و Salomon QST 92 2017-18 و Salomon QST 92 Skis 2018 هي نفس الشيء، في حين أن شركات ScienceSoft و Scienceoft ليست كذلك.

لا أحد يخفي حقيقة أن البيانات الضخمة ليست دقيقة بنسبة 100٪. وبشكل عام، الأمر ليس بهذه الأهمية. لكن هذا لا يعني أنه لا يجب عليك التحكم على الإطلاق في مدى موثوقية بياناتك. لا يمكن أن تحتوي على معلومات خاطئة فحسب، بل يمكنها أيضًا تكرار نفسها، فضلاً عن احتوائها على تناقضات. ومن غير المحتمل أن توفر البيانات ذات الجودة المتدنية للغاية أي رؤى مفيدة أو فرص رائعة لمهام عملك التي تتطلب الدقة في العمل.

كما تعد التحديات الأمنية للبيانات الضخمة مشكلة كبيرة تستحق جزءا آخر كاملًا مخصصًا لهذا الموضوع. لكن دعونا نلقي نظرة على المشكلة على نطاق أوسع. في كثير من الأحيان، تؤخر مشاريع اعتماد البيانات الضخمة الأمان إلى مراحل لاحقة. وبصراحة، هذا ليس تحركًا ذكيًا كثيرًا. تتطور تقنيات البيانات الضخمة، لكن ميزات الأمان الخاصة بها لا تزال مهملة، حيث من المأمول أن يتم منح الأمان على مستوى التطبيق. وماذا حصلنا عليه؟ في المرتين (مع التقدم التكنولوجي وتنفيذ المشروع) يتم تجاهل أمان البيانات الضخمة.

الميزة الأكثر شيوعًا للبيانات الضخمة هي قدرتها الهائلة على النمو. وأحد أخطر تحديات البيانات الضخمة مرتبط بهذا بالضبط. يمكن التفكير في تصميم الحل الخاص بك وتعديله ليتناسب مع الارتقاء بالمستوى دون بذل جهود إضافية. لكن المشكلة الحقيقية ليست العملية الفعلية لإدخال قدرات معالجة وتخزين جديدة. إنه يكمن في تعقيد التوسع بحيث لا ينخفض أداء نظامك وتظل في حدود الميزانية.

[1] https://www.scnsoft.com/blog/big-data-challenges-and-their-solutions , viewed at 23:09, 05/05/2022.

يمكن أن تؤدي خوارزميات التحليلات وتطبيقات الذكاء الاصطناعي المبنية على البيانات الضخمة إلى نتائج سيئة عندما تتسلل مشكلات جودة البيانات إلى أنظمة البيانات الضخمة. يمكن أن تصبح هذه المشكلات أكثر أهمية وأصعب في المراجعة حيث تحاول فرق إدارة البيانات والتحليلات جذب المزيد من أنواع البيانات المختلفة. Bunddler ، سوق عبر الإنترنت للعثور على مساعدين للتسوق عبر الإنترنت يساعدون الناس على شراء المنتجات وترتيب الشحنات ، عانى من هذه المشكلات بشكل مباشر حيث وصل إلى 500000 عميل. كان المحرك الرئيسي لنمو الشركة هو استخدام البيانات الضخمة لتوفير تجربة شخصية للغاية، والكشف عن فرص زيادة المبيعات ومراقبة الاتجاهات الجديدة. كانت إدارة جودة البيانات الفعالة مصدر قلق رئيسي. خاصة عندما تأتي البيانات من مصادر مختلفة. لضمان جودة البيانات التي يجمعونها، أنشأ فريق كوفالينكو معرّف بيانات ذكيًا يطابق التكرارات مع تباينات بسيطة في البيانات ويبلغ عن أي أخطاء إملائية محتملة. أدى ذلك إلى تحسين دقة الرؤى التجارية الناتجة عن تحليل البيانات.

تستخدم بعض المؤسسات بحيرة البيانات كمستودع شامل لمجموعات البيانات الضخمة التي تم جمعها من مصادر متنوعة، دون التفكير في كيفية دمج البيانات المتباينة. تنتج مجالات الأعمال المختلفة، على سبيل المثال، بيانات مهمة للتحليل المشترك، ولكن هذه البيانات غالبًا ما تأتي مع دلالات أساسية مختلفة يجب توضيحها. ويجب الحذر من الدمج المخصص للمشاريع، والذي يمكن أن يتضمن الكثير من إعادة العمل. للحصول على عائد الاستثمار الأمثل في مشاريع البيانات الضخمة، من الأفضل بشكل عام تطوير نهج استراتيجي لتكامل البيانات.

قد تصبح معالجة مشكلات حوكمة البيانات أكثر صعوبة مع نمو تطبيقات البيانات الضخمة عبر المزيد من الأنظمة. تتفاقم هذه المشكلة لأن البنى السحابية الجديدة تمكن المؤسسات من التقاط وتخزين جميع البيانات التي تجمعها في شكلها غير المجمع. يمكن أن تتسلل حقول المعلومات المحمية بطريق الخطأ إلى مجموعة متنوعة من التطبيقات.

بدون إستراتيجية وضوابط حوكمة البيانات، يمكن فقد الكثير من فوائد الوصول الأوسع والأعمق إلى البيانات، ومن الممارسات الجيدة التعامل مع البيانات كمنتج، مع وضع قواعد الحوكمة المدمجة منذ البداية. إن استثمار المزيد من الوقت مقدمًا في تحديد وإدارة مشكلات حوكمة البيانات الضخمة سيجعل من السهل توفير وصول الخدمة الذاتية الذي لا يتطلب الإشراف على كل حالة استخدام جديدة.

فيما يلي وصف تفصيلي لكل من المتطلبات التقنية الثمانية التي تم استخلاصها من تطبيقات البيانات الضخمة التي تطرقنا اليها سابقا:[1]

[1] José María Cavanillas: New Horizons for a Data-Driven Economy, previous reference, pp.202-203.

- **اكتشاف الأنماط:** تحديد الأنماط وأوجه التشابه مثلا لاكتشاف السلوكيات الإجرامية أو غير القانونية المحددة في سيناريو التطبيق لرصد مشغلي الأسواق عبر الإنترنت والإشراف عليهم (وأيضًا لسيناريوهات المراقبة المماثلة داخل القطاع العام). هذا المطلب قابل للتطبيق أيضًا في السيناريو لتحسين الكفاءة التشغيلية في وكالة العمل، وفي سيناريو الشرطة التنبؤية.
- **مشاركة البيانات وتكامل البيانات:** مطلوب للتغلب على نقص توحيد مخططات البيانات وتجزئة ملكية البيانات. تكامل مصادر البيانات المتعددة والمتنوعة في منصة البيانات الضخمة.
- **رؤى في الوقت الفعلي:** تمكن من تحليل بيانات حديثة في الوقت الفعلي لاتخاذ قرارات فورية، للحصول على رؤى في الوقت الفعلي من البيانات التي تمت معالجتها.
- **أمن البيانات:** الإجراءات القانونية والوسائل التقنية التي تسمح بمشاركة البيانات بأمان وخصوصية. قد تفتح حلول هذا المطلب الاستخدام الواسع النطاق للبيانات الضخمة في القطاع العام. يعد التقدم في حماية وخصوصية البيانات أمرًا أساسيًا للقطاع العام، حيث قد يسمح بتحليل كميات هائلة من البيانات التي يمتلكها القطاع العام دون الكشف عن معلومات حساسة. تمنع مشكلات الخصوصية والأمان هذه استخدام البنى التحتية السحابية (المعالجة والتخزين) من قبل العديد من الوكالات العامة التي تتعامل مع البيانات الحساسة.
- **نقل البيانات في الوقت الفعلي:** نظرًا لتزايد القدرة على وضع أجهزة الاستشعار في سيناريوهات تطبيقات المدن الذكية، هناك طلب كبير على نقل البيانات في الوقت الفعلي. سيكون مطلوبًا توفير إمكانات معالجة وتنظيف موزعة لمستشعرات الصور حتى لا تنهار قنوات الاتصال وتوفر فقط المعلومات المطلوبة لتحليل الوقت الفعلي، والتي ستغذي أنظمة الوعي الظرفي لصانعي القرار.
- **تحليلات اللغة الطبيعية:** استخراج المعلومات من مصادر غير منظمة عبر الإنترنت (مثل وسائل التواصل الاجتماعي) لتمكين التنقيب عن المشاعر. التعرف على البيانات من مدخلات اللغة الطبيعية مثل النص والصوت والفيديو.
- **التحليلات التنبؤية:** كما هو موضح في سيناريو التطبيق للشرطة التنبؤية، حيث كمثال يكون الهدف هو توزيع قوات الأمن والموارد وفقًا لتوقع الحوادث، قم بتوفير تنبؤات تستند إلى التعلم من المواقف السابقة للتنبؤ بتخصيص الموارد الأمثل للخدمات العامة.
- **النمذجة والمحاكاة:** أدوات خاصة بالمجال لنمذجة ومحاكاة الأحداث وفقًا للبيانات من الأحداث الماضية لتوقع نتائج القرارات المتخذة للتأثير على الظروف الحالية في الوقت الفعلي، على سبيل المثال، في سيناريوهات السلامة العامة.

المطلب الثاني: الخصوصية والأمان

قد نكون أشرنا بشكل غير مباشر لمشكلة الخصوصية التي تواجهها البيانات بشكل عام، لكن في هذا الفرع سنقوم بالتعمق أكثر في هذا الموضوع للحصول على فهم شامل حوله نظرا لأهميته الحساسة.

تخلق البيانات الضخمة فرصة هائلة للاقتصاد العالمي ليس فقط في مجال الأمن القومي، ولكن أيضًا في مجالات تتراوح من التسويق وتحليل مخاطر الائتمان إلى البحث الطبي والتخطيط الحضري. في الوقت نفسه، يتم التخويف من استخدام البيانات الضخمة من خلال مشكل الخصوصية وحماية البيانات. يشعر المدافعون عن الخصوصية بالقلق من أن التقدم في النظام الإيكولوجي للبيانات سيقلب علاقات القوة بين الحكومة والشركات والأفراد، ويؤدي إلى التنميط العنصري أو غيره من أشكال التمييز والإفراط في التجريم والحريات المقيدة الأخرى.

فلا يقتصر أمان مشروعات البيانات الضخمة على إتاحة الوصول إلى المعلومات فحسب. البيانات التي تعمل كمصدر للتحليل، كقاعدة عامة، تحتوي على معلومات مهمة للأعمال: الأسرار التجارية، البيانات الشخصية، وما إلى ذلك. يمكن أن يتحول انتهاك سرية التعامل مع هذه البيانات إلى مشاكل خطيرة، بما في ذلك الغرامات المفروضة على المنظمين، خسارة العملاء، فقدان القيمة السوقية، إلخ.

الفرع الأول: مفهوم الخصوصية

هناك الكثير من التعاريف للخصوصية في الأدب والعلوم كونها محل دراسة من الكثير من المجالات والقضايا.

إن الرغبة في عدم استباق استفسارنا عن قيمة الخصوصية من خلال تبني مفهوم مليء بالقيمة في البداية كافية لتبرير النظر إلى الخصوصية كحالة للفرد مقابل الآخرين، أو كشرط من شروط الحياة. كما يتطلب أيضًا رفض محاولات وصف الخصوصية بأنها مطالبة أو حالة نفسية أو منطقة لا ينبغي غزوها. للأسباب نفسها، هناك وصف آخر يجب رفضه وهو الخصوصية كشكل من أشكال السيطرة.[1]

بشكل عام، الخصوصية هي الحق في عدم التدخل، أو التحرر من التدخل أو التطفل. خصوصية المعلومات هي الحق في الحصول على بعض التحكم في كيفية جمع معلوماتك الشخصية واستخدامها.

اسأل معظم الأشخاص هذه الأيام عما يفكرون فيه عندما يتعلق الأمر بالخصوصية ومن المحتمل أن تجري محادثة حول الانتهاكات الهائلة للبيانات والتكنولوجيا والشبكات الاجتماعية والأخطاء الإعلانية المستهدفة.

[1] Ruth Gavison: Privacy and the Limits of Law , The Yale Law Journal Company, Inc, Vol. 89, N. 3,January 1980, pp425.426.

أضف إلى ذلك أن الثقافات المختلفة لها وجهات نظر مختلفة على نطاق واسع حول ماهية حقوق الشخص عندما يتعلق الأمر بالخصوصية وكيف ينبغي تنظيمها.[1]

فالخصوصية في مفهومها البسيط هو عدم التدخل والسيطرة على المعلومات، والوصول المقيد إليها ومع ذلك، من الصعب ضبط مفهوم الخصوصية وليس من السهل تحديده لأن هذا المفهوم مرتبط بعدة أبعاد مثل الجسم الشخصي والسلوك الشخصي وحاليا الاتصالات الشخصية والمعلومات الشخصية وتخضع هذه المفاهيم لقوانين وسياسات معينة تشكل الأبعاد المبنية في هذه القوانين وتشريعات السياسة العامة.

الخصوصية حق أساسي، ضروري للاستقلالية وحماية كرامة الإنسان، وهي بمثابة الأساس الذي تُبنى عليه العديد من حقوق الإنسان الأخرى.

تمكننا الخصوصية من إنشاء حواجز وإدارة الحدود لحماية أنفسنا من التدخل غير المبرر في حياتنا، مما يسمح لنا بالتفاوض بشأن هويتنا وكيف نريد التفاعل مع العالم من حولنا. تساعدنا الخصوصية على وضع حدود للحد من من يمكنه الوصول إلى أجسادنا وأماكننا وأشياءنا، فضلاً عن اتصالاتنا ومعلوماتنا.

تمنحنا القواعد التي تحمي الخصوصية القدرة على تأكيد حقوقنا في مواجهة الاختلالات الكبيرة في توازن القوى.

ونتيجة لذلك، تعد الخصوصية طريقة أساسية نسعى إلى حماية أنفسنا والمجتمع من استخدامها التعسفي وغير المبرر للسلطة، من خلال تقليل ما يمكن معرفته عنا وفعله بنا، مع حمايتنا من الآخرين الذين قد يرغبون في ممارسة السيطرة.الخصوصية ضرورية لمن نحن كبشر، ونتخذ قرارات بشأنها كل يوم. إنه يمنحنا مساحة لنكون أنفسنا بدون حكم، ويسمح لنا بالتفكير بحرية دون تمييز، وهو عنصر مهم لمنحنا السيطرة على من يعرف ماذا عنا.[2]

الفرع الثاني: مشكل الخصوصية في البيانات الضخمة

منذ العصور القديمة، ناقش الناس في جميع المجتمعات تقريبًا قضايا الخصوصية، بدءًا من النميمة إلى التنصت إلى المراقبة. أدى تطوير التقنيات الجديدة إلى إبقاء القلق بشأن الخصوصية مشتعلًا لعدة قرون، ولكن الانتشار العميق لتقنيات المعلومات الجديدة خلال القرن العشرين -وخاصة ظهور الكمبيوتر -جعل الخصوصية تتفجر لتصبح قضية خط المواجهة في جميع أنحاء العالم. ابتداءً من الستينيات، تلقى موضوع الخصوصية اهتمامًا

[1] https://iapp.org/about/what-is-privacy/ , viewed at 17:55, 08/05/2022.
[2] https://privacyinternational.org/explainer/56/what-privacy , viewed at 20:18, 09/05/2022.

متزايدًا بشكل مطرد. وتراوح الخطاب من الكتاب المشهورين إلى الصحفيين إلى الخبراء في القانون والفلسفة وعلم النفس وعلم الاجتماع والأدب والاقتصاد ومجالات أخرى لا حصر لها.[1]

الخصوصية قضية ذات أهمية عميقة في جميع أنحاء العالم. في كل دولة تقريبًا، تسعى العديد من القوانين والحقوق الدستورية والقرارات القضائية إلى حماية الخصوصية. في القانون الدستوري للدول حول العالم، يتم تكريس الخصوصية كحق أساسي. على الرغم من أن دستور الولايات المتحدة لا يذكر صراحة كلمة "الخصوصية"، إلا أنه يحمي قدسية المنزل والسرية الاتصالات من تدخل الحكومة. خلصت المحكمة العليا إلى أن التعديل الرابع يحمي من عمليات التفتيش الحكومية عندما يكون لدى الشخص "توقع معقول للخصوصية". بالإضافة إلى ذلك، قررت المحكمة العليا أن الدستور يحافظ على "منطقة خصوصية" تشمل القرارات التي يتخذها الأشخاص بشأن الصحة، فضلاً عن حماية معلوماتهم الشخصية من الإفشاء غير المبرر من قبل الحكومة. تحمي العديد من الدول صراحة الخصوصية في دساتيرها.

قد يكون إيجاد التوازن الصحيح بين مخاطر الخصوصية ومكافآت البيانات الضخمة أكبر تحدٍ للسياسة العامة في عصرنا. إنه يدعو إلى اتخاذ خيارات بالغة الأهمية بين اهتمامات السياسة الكبيرة مثل البحث العلمي، والصحة العامة، والأمن القومي، وإنفاذ القانون والاستخدام الفعال للموارد من ناحية، وحقوق الأفراد في الخصوصية والإنصاف والمساواة وحرية التعبير من ناحية أخرى. وهو يتطلب تقرير ما إذا كانت الجهود المبذولة لعلاج مرض مميت أو نزع فتيل الإرهاب تستحق إخضاع الفردانية البشرية للمراقبة الشاملة واتخاذ القرارات باستخدام الخوارزميات.

لسوء الحظ، تتقدم المناقشة أزمة تلو الأخرى، مع التركيز غالبًا على الشكليات القانونية بينما يتم تجنب خيارات السياسة الأكبر. علاوة على ذلك، أصبح الجدل مستقطبًا بشكل متزايد، حيث تتجاهل كل مجموعة مخاوف الأخرى تمامًا. على سبيل المثال، في سياق الرقابة الحكومية، يصور المدافعون عن الحريات المدنية الحكومة على أنها تسعى إلى السلطة المطلقة. يبدو أنه بالنسبة لصقور الخصوصية، لا توجد فائدة مهما كانت كبيرة بما يكفي لتعويض تكاليف الخصوصية، بينما بالنسبة لعشاق البيانات، فإن مخاطر الخصوصية ليست أكثر من فكرة متأخرة في السعي وراء المعلومات الكاملة.

في بعض الحالات، يوفر تحليل البيانات الضخمة فائدة مباشرة للأفراد الذين يتم استخدام معلوماتهم. يوفر هذا دافعًا قويًا للمنظمات للمناقشة حول مزايا استخدامها بناءً على قيمتها العائدة للأفراد المتضررين، ناقشنا أنه في العديد من هذه الحالات، فإن الاعتماد على خيارات الأفراد لإضفاء الشرعية على استخدام البيانات هو أمر

[1] Daniel J. Solove: UNDERSTANDING PRIVACY, Harvard University Press, GWU Legal Studies Research Paper No. 420, May 2008, pp. 2-3.

فارغ نظرًا للتحيزات الموثقة جيدًا في عمليات صنع القرار الخاصة بهم. الفرصة التي يوليها الفرد اهتمامًا، بينما في الآخرين، قد يتراجع الأفراد على الرغم من مصلحتهم الفضلى. ومع ذلك، سيكون من المؤسف أن يؤدي الفشل في الحصول على موافقة ذات مغزى إلى تشويه سمعة ممارسة المعلومات التي تفيد الأفراد بشكل مباشر.[1]

ضع في اعتبارك الدرجة العالية من التخصيص التي تتبعها Netflix وAmazon، والتي توصي المستهلكين بالأفلام والمنتجات بناءً على تحليل تفاعلاتهم السابقة. يفيد تحليل البيانات هذا المستهلكين بشكل مباشر وقد تم تبريره حتى بدون التماس الموافقة الصريحة. وبالمثل، فإن قرار Comcast في عام 2010 بمراقبة أجهزة كمبيوتر عملائها بشكل استباقي للكشف عن البرامج الضارة، والقرارات الأحدث التي اتخذها مزودو خدمات الإنترنت، بما في ذلك Comcast وAT&T وVerizon للوصول إلى المستهلكين للإبلاغ عن إصابات البرامج الضارة المحتملة، كانت تهدف إلى إفادة المستهلكين بشكل مباشر.[2] تعتمد وظائف الإكمال التلقائي والترجمة من Google على جمع بيانات شامل وتحليل في الوقت الفعلي بضغطة مفتاح تلو الأخرى. عرض القيمة للمستهلكين واضح ومقنع.

[1] Omer Tene & Jules Polonetsky: To Track or 'Do Not Track': Advancing Transparency and Individual Control in Online Behavioral Advertising, 13 MINN. J.L. SCI. & TECH. 281, 2012, pp285-86.

[2] Roy Furchgott, *Comcast to Protect Customer's Computers from Malware*, N.Y. TIMES GADGETWISE (Sept. 30, 2010).

الخلاصة:

يمكن أن تكون التنمية الاقتصادية مهمة محفوفة بالمخاطر. بدون معلومات وإحصاءات دقيقة، من الصعب التنبؤ بالاتجاه والسرعة اللذين ينبغي تحقيق النمو بهما. من خلال تسخير البيانات الضخمة، تتاح الفرصة لمنظمي البيانات الإلكترونية للتخطيط بثقة، واكتساب فهم أعمق لاحتياجات المجتمعات، والمضي قدمًا بطريقة مستدامة، كما ناقشنا في هذا الفصل أن فوائد البيانات الضخمة هائلة. بالنسبة لفئة رجال الأعمال، تعتبر التكنولوجيا مجرد وسيلة لإبقاء الشركة قريبة من عملائها. شهدت الشركات التي شرعت في مشروع البيانات الضخمة (...Google، facebook) نموًا هائلاً في الأعمال. لقد ساعدت المنظمة بنجاح على تحقيق تخفيضات في التكاليف، وقرارات أسرع وأفضل، وحتى توفير خدمات جديدة للعميل. كما تطرقنا أيضا الى أكبر التحديات التي تقف دون استخدام البيانات الضخمة، حيث رأينا كيف أن الخصوصية هي قضية مهمة في تطبيق تقنيات البيانات الضخمة للتحليلات. نظرًا لأنه يتم جمع المزيد والمزيد من البيانات، فقد يؤدي تجميع البيانات هذا جنبًا إلى جنب مع تحليلات البيانات إلى انتهاك خصوصية المستخدم. إذا تم الاستعانة بمصادر خارجية لتحليلات البيانات، يمكن لموظف طرف ثالث غير موثوق به استنتاج المعلومات الشخصية للمستخدمين. ترغب المؤسسات في استخدام أدوات تحليلات البيانات الضخمة لتعزيز رضا العملاء، لكنها تحتاج إلى ضمان حماية خصوصية المستخدم أثناء القيام بذلك.

الفصل الرابع:

دراسة برنامج

Apache Hadoop

تمهيد:

أدت ثورة الحوسبة التي بدأت منذ أكثر من عقدين من الزمن إلى تراكم كميات كبيرة من البيانات الرقمية من قبل الشركات. التطورات في أجهزة الاستشعار الرقمية؛ انتشار أنظمة الاتصالات، وخاصة المنصات والأجهزة المحمولة؛ التسجيل على نطاق واسع لأحداث النظام؛ وقد أدى التحرك السريع نحو المنظمات غير الورقية إلى مجموعة ضخمة من موارد البيانات داخل المنظمات. ويضمن الاعتماد المتزايد للشركات على التكنولوجيا أن البيانات ستستمر في النمو بمعدل أسرع. قانون مور، الذي ينص على أن أداء أجهزة الكمبيوتر تضاعف تاريخيًا كل عامين تقريبًا، ساعد في البداية موارد الحوسبة لمواكبة نمو البيانات. ومع ذلك، بدأت وتيرة التحسن في موارد الحوسبة في التناقص في حوالي عام 2005. بدأت صناعة الحوسبة في البحث عن خيارات أخرى، وهي المعالجة المتوازية لتوفير حل أكثر اقتصادا. إذا تعذر على جهاز كمبيوتر واحد أن يصبح أسرع، كان الهدف هو استخدام العديد من موارد الحوسبة لمعالجة نفس المشكلة بشكل متوازٍ Hadoop .هو تطبيق لفكرة أجهزة الكمبيوتر المتعددة في الشبكة التي تطبق MapReduce نوع من التعليمات الفردية ، فئة البيانات المتعددةSIMD لتقنية الحوسبة لتوسيع نطاق معالجة البيانات.

أدى تطور الحوسبة السحابية من خلال بائعين مثل أمازون وجوجل ومايكروسوفت إلى تعزيز هذا المفهوم لأنه يمكننا الآن استئجار موارد الحوسبة مقابل جزء بسيط من التكلفة التي يتطلب شرائها.

هذا الفصل عبارة عن دراسة عملية لتطوير البرامج وتشغيلها باستخدام Hadoop، وهو مشروع تستضيفه مؤسسة Apache Software Foundation ويتم توسيعه ودعمه الآن من قبل العديد من البائعين مثل Cloudera وMapR وHortonworks.وسيناقش هذا الفصل الدافع وراء البيانات الضخمة بشكل عام وHadoop بشكل خاص.

لمحة تاريخية عن برنامج Hadoop:

بدأ Hadoop مع Doug Cutting و Mike Cafarellaفي عام 2002 عندما بدأ كلاهما العمل في مشروع .Apache Nutch كان مشروع Apache Nutch عبارة عن عملية بناء نظام محرك بحث يمكنه فهرسة مليار صفحة. بعد الكثير من البحث حولNutch ، خلصوا إلى أن مثل هذا النظام سيكلف حوالي نصف مليون دولار من الأجهزة، بالإضافة إلى تكلفة تشغيل شهرية تبلغ 30 ألف دولار تقريبًا، وهو مكلف للغاية. لذلك، أدركوا أن تصميم مشروعهم لن يكون قادرًا بشكل كافٍ على الحل البديل بمليارات الصفحات على الويب. لذلك كانوا يبحثون عن حل عملي يمكن أن يقلل من تكلفة التنفيذ بالإضافة إلى مشكلة تخزين مجموعات البيانات الكبيرة ومعالجتها.

في عام 2003، صادفوا ورقة تصف بنية نظام الملفات الموزعة منGoogle ، المسمى GFS نظام ملفات Googleالذي نشرتهGoogle ، لتخزين مجموعات البيانات الكبيرة. لقد أدركوا الآن أن هذه الورقة يمكن أن تحل مشكلة تخزين الملفات الكبيرة جدًا التي تم إنشاؤها بسبب عمليات الفهرسة على الويب. لكن هذه الورقة كانت مجرد نصف حل لمشكلتهم.

في عام 2004، نشرت Google ورقة أخرى حول تقنيةMapReduce ، والتي كانت حلاً لمعالجة مجموعات البيانات الكبيرة هذه. الآن هذه الورقة كانت نصف حل آخر لـDoug Cut و Mike Cafarellaلمشروع Nutch الخاص بهم. كانت هاتان التقنيتان GFSوMapReduce فقط على ورقة بيضاء في .Google لم تطبق Google هاتين الطريقتين. عرف دوج كاتنج من عمله على أباتشي لوسين (وهي مكتبة مجانية ومفتوحة المصدر لاسترجاع المعلومات، وقد كُتبت في الأصل بلغة جافا بواسطة دوج كاتنج في 1999) أن المصدر المفتوح هو وسيلة رائعة لنشر التكنولوجيا لعدد أكبر من الناس. لذلك، بدأ مع مايك كافاريلا في تطبيق تقنيات (Google (GFS & MapReduce كمصدر مفتوح في مشروع.Apache Nutch

في عام 2005، وجد كتنج أن Nutch يقتصر على 20 إلى 40 من مجموعات العقدة فقط. سرعان ما أدرك مشكلتين[1]:

- لن يحقق Nutch إمكاناته حتى يعمل بشكل موثوق على المجموعات الأكبر.
- وكان هذا يبدو مستحيلًا مع شخصين فقط دوغ كاتنج ومايك كافاريلا.

[1] Sameer Wadkar: Pro Apache Hadoop, The expert's voice in big data, Friendsof Apress, 2nd edition, 2014, p12.

كانت المهمة الهندسية في مشروع Nutch أكبر بكثير مما أدرك. لذلك بدأ في العثور على وظيفة في شركة مهتمة بالاستثمار في جهودهم. ووجد ياهو! كان لدى ياهو فريق كبير من المهندسين الذين كانوا حريصين على العمل في هذا المشروع هناك.

لذلك في عام 2006، انضم دوج كاتنج إلى Yahoo مع مشروع .Nutch لقد أراد أن يزود العالم بإطار عمل حوسبة مفتوح المصدر وموثوق وقابل للتطوير بمساعدة .Yahoo لذلك في موقع Yahoo أولاً، قام بفصل أجزاء الحوسبة الموزعة عن Nutch وشكل مشروعًا جديدًا أطلق على عليه اسم Hadoop اسمًا لفيل لعبة أصفر كان مملوكًا لابن دوج كاتنج. الآن أراد أن يصنع Hadoop بطريقة يمكن أن تعمل بشكل جيد على آلاف العقد. لذلك مع GFSوMapReduce، بدأ العمل على.Hadoop

في عام 2007، اختبرت Yahoo بنجاح Hadoop على مجموعة مكونة من 1000 عقدة وبدأت في استخدامها.

في جانفي من عام 2008، أصدرت Yahoo مكتوب Hadoop كمشروع مفتوح المصدر لـ ASF مؤسسة برامج .Apacheوفي جويلية من عام 2008، اختبرت Apache Software Foundation بنجاح مجموعة 4000 عقدة باستخدام.Hadoop في عام 2009، تم اختبار Hadoop بنجاح لفرز (PetaByte) PB من البيانات في أقل من 17 ساعة للتعامل مع المليارات من عمليات البحث وفهرسة الملايين من صفحات الويب. وترك دوج كاتنج شركة ياهو وانضم إلى كلوديرا لمواجهة التحدي المتمثل في نشر Hadoop في الصناعات الأخرى[1].

في ديسمبر 2011، أصدرت Apache Software Foundation الإصدار 1.0 من.Apache Hadoop

وفي وقت لاحق في أغسطس 2013، كان الإصدار 2.0.6 متاحًا.

وحاليًا، لدينا الإصدار 3.3.3 من Apache Hadoop والذي تم إصداره في ماي 2022.

ماهية برنامج Apache Hadoop:

هو عبارة عن مجموعة من أدوات البرامج مفتوحة المصدر التي تسهل استخدام شبكة من العديد من أجهزة الكمبيوتر لحل المشكلات التي تنطوي على كميات هائلة من البيانات والحسابات. يوفر إطارًا برمجيًا للتخزين الموزع ومعالجة البيانات الضخمة باستخدام نموذج برمجة.MapReduce

[1] https://www.geeksforgeeks.org/hadoop-history-or-evolution/ , viewed at 18:30, 30/05/2022.

Apache Hadoop هو إطار عمل برمجي مفتوح المصدر للتخزين والمعالجة على نطاق واسع لمجموعات البيانات على مجموعات من الأجهزة السلعية. Hadoop هو مشروع Apache عالي المستوى يتم بناؤه واستخدامه من قبل مجتمع عالمي من المساهمين والمستخدمين. تم ترخيصه بموجب ترخيص Apache 2.0.[1]

غالبًا ما يستخدم مصطلح Hadoop لكل من الوحدات الأساسية والوحدات الفرعية وكذلك النظام البيئي، أو مجموعة من حزم البرامج الإضافية التي يمكن تثبيتها فوق أو بجانبHadoop ، مثل Apache Pig و Apache Hiveو Apache HBaseوApache Phœnix، Apache Sparkو Apache ZooKeeperو Cloudera Impalaو Apache Flumeو Apache Sqoopو Apache Oozieو Apache Storm.التي سبق أن تطرقنا لبعض منها.

تم استلهام مكونات MapReduce و HDFSمن Apache Hadoop من أوراق Google على MapReduce و Google File System.

تتم كتابة إطار عمل Hadoop نفسه في الغالب بلغة برمجة Java، مع بعض التعليمات البرمجية الأصلية في C وأدوات مساعدة سطر الأوامر مكتوبة كنصوص برمجية shell. على الرغم من أن كود Java MapReduce شائع، يمكن استخدام أي لغة برمجة مع Hadoop Streaming لتنفيذ الخريطة وتقليل أجزاء من برنامج المستخدم. تعرض المشاريع الأخرى في نظام Hadoop البيئي واجهات مستخدم أكثر ثراءً.[2]

تم تصميم Hadoop في الأصل لمجموعات الكمبيوتر المبنية من الأجهزة السلعية، والتي لا تزال شائعة الاستخدام، ومنذ ذلك الحين وجد أيضًا استخدامها على مجموعات من الأجهزة المتطورة. تم تصميم جميع الوحدات في Hadoop بافتراض أساسي أن أعطال الأجهزة شائعة الحدوث ويجب أن يتم التعامل معها تلقائيًا بواسطة إطار العمل.

جوهر Apache Hadoop من جزء تخزين يُعرف باسم HDFS وجزء معالجة وهو نموذج برمجة MapReduce. يقوم Hadoop بتقسيم الملفات إلى كتل كبيرة وتوزيعها عبر العقد في مجموعة. ثم ينقل الكود المعبأ إلى عقد لمعالجة البيانات بالتوازي. يستفيد هذا الأسلوب من موقع البيانات، حيث تعالج العقد البيانات التي يمكنها الوصول إليها. يتيح ذلك معالجة مجموعة البيانات بشكل أسرع وأكثر كفاءة مما ستكون عليه في بنية الكمبيوتر العملاق الأكثر تقليدية التي تعتمد على نظام ملفات متوازي حيث يتم توزيع الحساب والبيانات عبر الشبكات عالية السرعة.

[1] https://opensource.com/life/14/8/intro-apache-hadoop-big-data#:~:text=Hadoop%20was%20created%20by%20Doug,after%20his%20son's%20toy%20elephant, viewed at 17:32, 01/06/2022.

[2] Sameer Wadkar: Pro Apache Hadoop, Previous reference, p11.

مكونات برنامج Apache Hadoop:

يتكون Hadoop من حزمة Hadoop المشتركة، والتي توفر نظام الملفات ومستويات نظام التشغيل، ومحرك MapReduceإما MR1 / MapReduce أوMR2 / YARN و .HDFSتحتوي حزمة Hadoop Common على ملفات Java Archive (JAR)والبرامج النصية اللازمة لبدء.Hadoop

لجدولة فعالة للعمل، يجب أن يوفر كل نظام ملفات متوافق مع Hadoop وعيًا بالموقع، وهو اسم الحامل، وتحديداً مفتاح الشبكة حيث توجد عقدة العامل. يمكن لتطبيقات Hadoop استخدام هذه المعلومات لتنفيذ التعليمات البرمجية على العقدة حيث توجد البيانات، وفشل ذلك، على نفس الرف / المفتاح لتقليل حركة المرور الأساسية. يستخدم HDFS هذه الطريقة عند نسخ البيانات لتكرار البيانات عبر رفوف متعددة. يقلل هذا النهج من تأثير انقطاع التيار الكهربائي على الرف أو فشل التبديل؛ في حالة حدوث أي من حالات فشل الأجهزة هذه، ستظل البيانات متاحة.

تتضمن مجموعة Hadoop الصغيرة عقدة رئيسية واحدة وعقد عمال متعددة. تتكون العقدة الرئيسية من Job Tracker وTracker Task وNameNode وDataNode. تعمل العقدة التابعة أو العاملة بمثابة DataNode وTaskTracker، على الرغم من أنه من الممكن أن يكون لديك عقد عاملة للبيانات فقط وعقد عاملة للحساب فقط. هذه تستخدم عادة فقط في التطبيقات غير القياسية. يتطلب Hadoop JRE 1.6 أو أعلى. تتطلب البرامج النصية القياسية لبدء التشغيل والإغلاق إعداد (Secure Shell (SSH بين العقد في المجموعة.

في مجموعة أكبر، تتم إدارة عقد HDFS من خلال خادم NameNode مخصص لاستضافة فهرس نظام الملفات، وNameNode ثانوي يمكنه إنشاء لقطات من هياكل ذاكرة الاسم، وبالتالي منع تلف نظام الملفات وفقدان البيانات. وبالمثل، يمكن لخادم JobTracker المستقل إدارة جدولة الوظائف عبر العقد. عند استخدام Hadoop MapReduce مع نظام ملفات بديل، يتم استبدال NameNode و NameNode الثانوي وهندسة DataNode لـ HDFS بالمكافئات الخاصة بنظام الملفات.[1]

في جوهر Hadoop أنه إطار عمل MapReduce يستند إلى .Java ومع ذلك، نظرًا للاعتماد السريع لمنصة Hadoop، كانت هناك حاجة لدعم مجتمع المستخدمين بخلاف .Java تطورت Hadoop لتصبح لديها التحسينات والمشاريع الفرعية التالية لدعم هذا المجتمع وتوسيع نطاق وصوله إلى المنظمة[2]:

[1] Druba Borthakur: Apache Hadoop goes realtime at facebook, Apache Hadoop Code Repository, 2006, p4.
[2] Sameer Wadkar: Pro Apache Hadoop, Previous reference, pp.11-13.

- **Hadoop Streaming ;** يسمح لك باستخدام MapReduce مع أي برنامج نصي لسطر الأوامر. هذا يجعل MapReduceقابلاً للاستخدام بواسطة مبرمجي UNIX ومبرمجي Python وما إلى ذلك لتطوير وظائف مخصصة.
- **Hadoop Hive :** أدرك المستخدمون بسرعة MapReduce أن تطوير برنامج MapReduce هو مهمة برمجة للغاية ، مما يجعله عرضة للخطأ ويصعب اختباره. كانت هناك حاجة إلى لغات أكثر تعبيرًا مثل SQL لتمكين المستخدمين من التركيز على المشكلة بدلاً من التطبيقات منخفضة المستوى لأدوات SQL النموذجية (على سبيل المثال جملة WHERE، جملة GROUP BY، جملة JOIN، إلخ). تم تطوير Apache Hive لتوفير إمكانية تخزين بيانات DW لمجموعات البيانات الكبيرة. يمكن للمستخدمين التعبير عن استفساراتهم في Hive Query، وهو مشابه جدًا لـ SQL. يحول محرك MySQL هذه الاستعلامات إلى وظائف MapReduce منخفضة المستوى. يمكن للمستخدمين الأكثر تقدمًا تطوير وظائف محددة بواسطة المستخدم (UDFs) في Java. يدعم Hive أيضًا برامج التشغيل القياسية مثل ODBC وHive .JDBC هو أيضًا نظام أساسي مناسب للاستخدام عند تطوير أنواع تطبيقات ذكاء الأعمال (BI) للبيانات المخزنة في Hadoop.
- **Hadoop Pig :** على الرغم من أن الدافع وراء Pig كان مشابهًا لـ Hive ، إلا أن Hive هي لغة شبيهة بلغة SQL ، وهي لغة توضيحية. من ناحية أخرى، فإن Pig هي لغة إجرائية تعمل بشكل جيد في سيناريوهات خط البيانات. سوف يروق Pig للمبرمجين الذين يطورون خطوط أنابيب معالجة البيانات (على سبيل المثال، مبرمجو SAS). وهي أيضًا منصة مناسبة لاستخدامها في استخراج أنواع التطبيقات وتحميلها وتحويلها (ELT).
- **Hadoop HBase :** جميع المشاريع السابقة ، بما في ذلك MapReduce ، هي عمليات مجمعة. ومع ذلك، هناك حاجة قوية للبحث عن البيانات في الوقت الفعلي في Hadoop. لم يكن لدى Hadoop متجر أصلي للمفتاح / القيمة. على سبيل المثال، فكر في أحد مواقع التواصل الاجتماعي مثل Facebook. إذا كنت تريد البحث عن ملف تعريف أحد الأصدقاء، فتوقع الحصول على إجابة على الفور (وليس بعد تشغيل مجموعة طويلة من الوظائف). كانت حالات الاستخدام هذه هي الدافع لتطوير منصة HBase.

نظام الملفات الموزعة Hadoop :HDFS

نظام الملفات الموزعة Hadoop HDFS هو نظام ملفات موزع وقابل للتطوير ومحمول مكتوب بلغة Java لإطار Hadoop. تحتوي كل عقدة في مثيل Hadoop عادةً على اسم واحد، وتشكل مجموعة من datanodes مجموعة HDFS. الوضع نموذجي لأن كل عقدة لا تتطلب وجود datanode. يقدم كل datanode مجموعات من

البيانات عبر الشبكة باستخدام بروتوكول كتلة خاص بـ HDFS. يستخدم نظام الملفات طبقة TCP / IP للاتصال. يستخدم العملاء استدعاء الإجراء البعيد (RPC) للتواصل بين بعضهم البعض.

يشتمل نظام ملفات HDFS على ما يسمى بـ namenode الثانوي، والذي يضلل بعض الأشخاص ليعتقدوا أنه عندما ينتقل namenodeالأساسي دون اتصال بالإنترنت، يتولى namenode الثانوي المهمة. في الواقع، يتصل namenodeالثانوي بانتظام برمز الاسم الأساسي ويبني لقطات من معلومات دليل namenode الأساسي، والتي يحفظها النظام بعد ذلك في الدلائل المحلية أو البعيدة. يمكن استخدام هذه الصور التي تم فحصها لإعادة تشغيل رمز اسم أولي فاشل دون الحاجة إلى إعادة تشغيل المجلة الكاملة لإجراءات نظام الملفات، ثم لتحرير السجل لإنشاء بنية دليل محدثة. نظرًا لأن namenode هو النقطة الوحيدة لتخزين وإدارة البيانات الوصفية، فقد يصبح عنق الزجاجة لدعم عدد كبير من الملفات، خاصة عدد كبير من الملفات الصغيرة. يهدف اتحادHDFS ، وهو إضافة جديدة، إلى معالجة هذه المشكلة إلى حد معين من خلال السماح بمسافات اسم متعددة تخدمها رموز أسماء منفصلة[1].

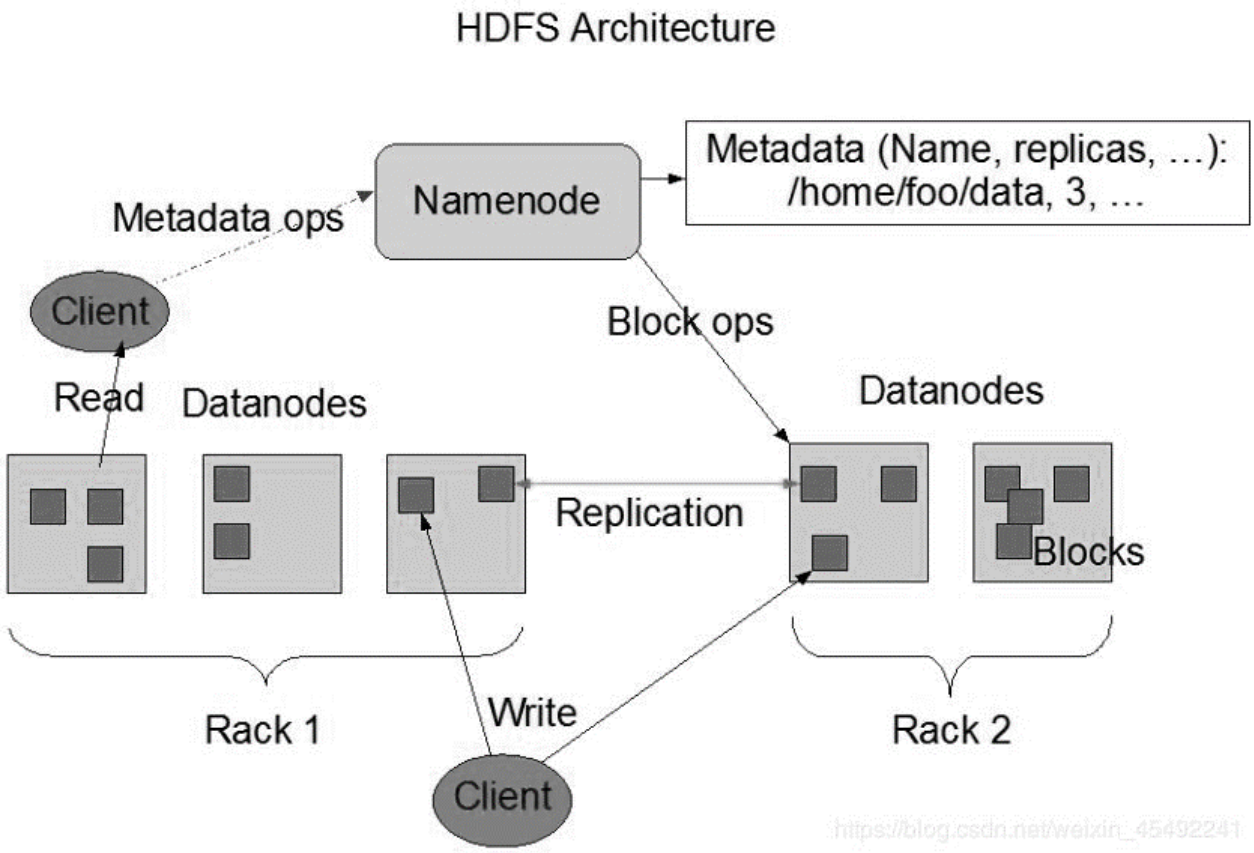
ميزة استخدام HDFS هي الوعي بالبيانات بين متتبع الوظائف ومتعقب المهام. يقوم متعقب الوظائف بجدولة الخريطة أو تقليل الوظائف إلى أدوات تعقب المهام مع إدراك موقع البيانات. على سبيل المثال، إذا كانت العقدة A تحتوي على بيانات x، y، z وكانت العقدة B تحتوي على بيانات (أ، ب، ج)، يقوم متتبع الوظائف بجدولة العقدة B لأداء الخريطة أو تقليل المهام على (أ، ب، ج) والعقدة ستتم جدولة A لأداء الخريطة أو تقليل المهام على x، y، z. هذا يقلل من كمية حركة المرور التي تمر عبر الشبكة ويمنع نقل البيانات غير الضروري. عند استخدام Hadoopمع أنظمة الملفات الأخرى، لا تتوفر هذه الميزة دائمًا. يمكن أن يكون لهذا تأثير كبير على أوقات إتمام العمل، وهو ما تم إثباته عند تشغيل وظائف كثيفة البيانات. تم تصميم HDFS للملفات الثابتة في الغالب وقد لا تكون مناسبة للأنظمة التي تتطلب عمليات كتابة متزامنة[2].

يتمثل أحد القيود الأخرى لـ HDFS في أنه لا يمكن تثبيته مباشرة بواسطة نظام تشغيل موجود. قد يكون إدخال البيانات وإخراجها من نظام ملفاتHDFS ، وهو إجراء غالبًا ما يتعين القيام به قبل وبعد تنفيذ مهمة ما، أمرًا غير مريح. تم تطوير نظام ملفات في نظام الملفات الافتراضي Userspace (FUSE) لمعالجة هذه المشكلة، على الأقل لنظام Linux وبعض أنظمة Unix الأخرى.

[1] Sameer Wadkar: Pro Apache Hadoop, Previous reference, pp.17-18.
[2] https://opensource.com/life/14/8/intro-apache-hadoop-big data#:~:text=Hadoop%20was%20created%20by%20Doug,after%20his%20son's%20toy%20elephant ,viewed at 23:47, 02/06/2022.

يمكن الوصول إلى الملفات من خلال Java API الأصلي، Thrift API، لإنشاء عميل باللغة التي يختارها المستخدمون (C ++، Java، Python، PHP، Ruby، Erlang، Perl، Haskell، C #، Cocoa، Smalltalk، أو OCaml)، أو واجهة سطر الأوامر، أو التصفح من خلال تطبيق الويب HDFS-UI عبر HTTP.

الشكل د: كيفية عمل HDFS.



المصدر: https://arabicprogrammer.com/article/7680515806/ , viewed at 00 :12, 03/06/2022.

الخلاصة:

من خلال هذا الفصل، قمنا بالتعرف على أحد أكثر الآليات المستخدمة في تحليل البيانات الضخمة وهو البرنامج Apache Hadoop. حيث مررنا على تاريخ تطور البرنامج، وقمنا بالتعرف عن ماهية هذا البرنامج كونه إطار عمل يسمح بالمعالجة الموزعة لمجموعات البيانات الكبيرة عبر مجموعات من أجهزة الكمبيوتر باستخدام نماذج برمجة بسيطة. إنه مصمم للارتقاء من الخوادم الفردية إلى آلاف الأجهزة، كل منها يقدم عمليات الحوسبة والتخزين المحلية. بدلاً من الاعتماد على الأجهزة لتقديم إتاحة عالية، تم تصميم المكتبة نفسها لاكتشاف حالات الفشل ومعالجتها في طبقة التطبيق، وبالتالي تقديم خدمة عالية التوفر أعلى مجموعة من أجهزة الكمبيوتر كما قدم هذا الفصل المفاهيم المختلفة لنظام Hadoop. لقد بدأ بمثال أساسي لعدد الكلمات وشرع في استكشاف العديد من الميزات الرئيسية في Hadoop. وتعرفنا على كيفية عمل هذا البرنامج من خلال نظام الملفات الموزعة Hadoop HDFSوشاهدنا كيفية إدارة الوظائف في Hadoopباستخدام برنامج JobTracker وTaskTracker. وأيضا من خلال خاصية namenode. مع ذكر بعض القيود وكيف تمت معالجتها مع مرور الوقت.

الخاتمة

1. الخاتمة

خلال فصول هذا البحث حاولنا الاجابة على الاشكالية المطروحة والمتمثلة في كيفية معالجة البيانات الضخمة واستخدامها في العلوم الاقتصادية، حيث تناولنا أربعة فصول كما يلي:

أشرنا في الفصل الأول إلى مفاهيم عامة حول البيانات، أنواعها (بيانات مهيكلةو بيانات غير مهيكلة والبيانات شبه المهيكلة) وكيفية قياسها، ومدى أهمية قواعد البيانات حيث اليوم، توظف الشركات متخصصين في إدارة البيانات أو تكلف العمال بدور الإشراف على البيانات، والذي يتضمن تنفيذ استخدام البيانات وسياسات الأمان على النحو المبين في مبادرات إدارة البيانات، كما تطرقنا أيضا الى مفهوم البيانات الضخمة كونها مصطلح ليس بالقديم والخصائص الثلاث التي تحددها وتميزها عن باقي البيانات الأخرى (السرعة، الحجم والتنوع)حيث أنها مجموعة من البيانات التي يفوق حجمها القدرة على معالجتها باستخدام أدوات قواعد البيانات التقليدية, وبهذا نكون قد أجبنا عن الفرضية الأولى والثانية.

وبذلك وجب اتباع وابتكار وسائل جديدة وهي موضوع الفصل الثاني. فتطرقنا الى معظم وسائل تخزين البيانات الضخمة وكيفية تطورها مع مرور الزمن، فكان من الضروري تغيير طرق تخزين واسترجاع البيانات، فانطلقنا من البطاقات المثقبة الى أحدث ما توصلت إليه تقنيات تخزين البيانات القادرة على التعامل مع كميات كبيرة من البيانات (NoSQL،NewSQL ، منصات الاستعلام)، ومنه تغطية الفرضية الثالثة.

وفي الفصل الثالث تطرقنا الى مدى أهمية استخدام تحاليل البيانات الضخمة في العلوم الاقتصادية التي تؤدي الى فرص اقتصادية ومزيد من الكفاءة. وعلى مستوى أعمق، يمكن أن توفر استخدام البيانات الضخمة فهمًا أفضل لهذه التبعيات، مما يجعل النظام أكثر شفافية ويدعم عمليات صنع القرار الاقتصادي والاجتماعي. حيثما تكون البيانات متاحة للجمهور، يتم دعم اتخاذ القرار الاجتماعي؛ حيثما تتوفر البيانات ذات الصلة على المستوى الفردي، يتم دعم اتخاذ القرارات الشخصية، وبالتالي تحقق الفرضية الرابعة.

لكن من جهة أخرى رأينا أن استخدام البيانات الضخمة يواجه عدة مشاكل تقنية وغير تقنية قد تطرقنا اليها ولعل أهمها هو موضوع الخصوصية الذي لطالما يطفو الى سطح كلما تحدثنا عن البيانات وتحليلها، كون الخصوصية قضية ذات أهمية عميقة في جميع أنحاء العالم. في كل دولة تقريبًا، تسعى العديد من القوانين والحقوق الدستورية والقرارات القضائية إلى حماية الخصوصية. في القانون الدستوري للدول حول العالم،

ثم قمنا بدراسة أحد أهم البرامج المتعلقة بالبيانات الضخمة وهو برنامج Apache Hadoop، أحد أكثر البرامج نجاحا واستعمالا في تحليل البيانات الضخمة، كونه يسمح باستعمال عدة حواسيب في أن واحد لمعالجة

الكم الهائل للبيانات الضخمة، الأمر الذي كان يبدو مستحيلا. فقمنا بالتعرف عن تاريخ هذا البرنامج، وكيفية عمله والأهم نظام الملفات الموزعة الخاص به الذي يتميز بالوعي بالبيانات بين متتبع الوظائف ومتعقب المهام.

فالأخير، وكإجابة عن إشكالية البحث نجد أن البيانات الضخمة تغير قواعد اللعبة. تستخدم العديد من المؤسسات المزيد من التحليلات لدفع الإجراءات الاستراتيجية وتقديم تجربة أفضل للعملاء. يمكن أن يؤدي التغيير الطفيف في الكفاءة أو المدخرات الصغيرة إلى ربح ضخم، وهذا هو السبب في أن معظم المؤسسات تتجه نحو البيانات الضخمة.

2. النتائج

من خلال بحثنا هذا توصلنا الى عدة نتاءج من أهمها نذكر النقاط التالية:

- تمثل البيانات الضخمة مرحلة هامة من مراحل تطور نظم المعلومات والإتصالات، وهي تعبر في مفهومها المبسط عن كمية هائلة من البيانات المعقدة التي يفوق حجمها قدرة البرمجيات والآليات الحاسوبية التقليدية على تخزينها ومعالجتها وتوزيعها، الأمر الذي أدى إلى وضع حلول بديلة متطورة تمكن من التحكم في تدفقها والسيطرة عليها.
- إن عملية اتخاذ القرارات تُعد محور العملية الإدارية وجوهرها وإن نجاح المؤسسة أوالقطاع الحكومي يتوقف إلى حد كبير على قدرة وكفاءة القيادة الإدارية على إتخاذ القرارات الإدارية المناسبة. إن عملية صنع القرار تبدأ بتجميع البيانات ومُعالجتها واستخلاص المعلومات التي بناء عليها يتم اتخاذ القرار حيث بدأت تعتمد العديد من الشركات الكبيرة والقطاعات الحكومية على سياسة تحليل البيانات الضخمة والمعقدة والتي تحتاج إلى البرمجيات المتخصصة في مجال إدارة البيانات والتحليلات.
- قد أصبح بإمكان الشركات والمؤسسات والهيئات اليوم على إختلاف أنواعها تحليل حركة العملاء من شراء وبيع ونحوه بدقة أكبر ليتمكنوا وفقاً لذلك من معرفة السلع الأكثر طلباً أو تلك الراكدة ويقترحوا على عملائهم سلع معينة وفقاً لعمليات الشراء التي تتم. كما أصبح لديهم القدرة على فهم سلوك العملاء بشكل أكثر دقة وتحديد المميزين منهم ومن هم بحاجة لمساعدة أو لتحديد توجهاتهم أو مُراقبة أدائهم. هذا الأمر ليس فقط لمراكز البيع التقليدية بل يشمل المتاجر الإلكترونية على شبكة الإنترنت وعلى نطاق أوسع.

3. توصيات

لاستغلال البيانات الضخمة بكفاءة يمكننا تقديم التوصيات التالية:

- عندما يتعلق الأمر بإدارة البيانات، أغلب المنظمات الحكومية تواجه مشكلة وجود كميات هائلة من البيانات في أنظمة الكمبيوتر، ومعظم هذة البيانات غير منظمة أو مُهيكلة (unstructured data) وهذا

يعني أنها لا تناسب أي نموذج بيانات معرّف مسبقاً. لفهم الأنماط الموجودة في هذه البيانات يجب أن تطبق المنظمات الحكومية نماذج إحصائية تسعى لإلتقاط ومعالجة كميات هائلة من البيانات غير المهيكلة.

- أدى إستخدام أدوات التعلم عبر الإنترنت والبرامج القائمة على التفاعل بصورة متزايدة في مجال التعليم إلى زيادة حجم البيانات، واختلاف نوعية البيانات الكبيرة التي يُمكن جمعها من بيئات التعلم، فهنا نجد بيانات كبيرة عن المتعلمين، وخبرات التعلم لدى المتعلمين، كما تختلف هذه البيانات في نوعيتها وعمقها بنسب متفاوتة. لذا يجب توظيف أشخاص ذو تخصص وخبرة لتكوين موظفين يستطيعون التأقلم مع هذا التغير المستمر للبيانات الضخمة.
- لإدارة وتحليل جميع البيانات الخاصة بهم، ومع طبيعة البيانات المتغيرة وارتفاع حجمها أصبحت الإستعانة بأدوات البيانات الضخمة من خلال الحوسبة السحابية (Cloud Computing) أمراً ضرورياً. فأصبح بإمكان المختصين بتطوير الخدمات الحكومية رصد مدى رضا المواطنين عن الخدمات المقدَّمة لهم. وعلى ضوء النتائج المحللة يمكن استنتاج ما يلزم عمله للتطوير والتحسين، حيث أصبح مسح آراء الجمهور عن طريق الإستبيانات التقليدية مكلفاً وغير مجدٍ في كثير من الأحيان، وذلك نظراً لتنوع البيانات الديموغرافية وثقافات المتعاملين. إن من أكبر المصادر لتلك البيانات الضخمة هي البيانات المسجلة من خلال عمليات التعداد السكاني والتسجيل في قواعد البيانات الحكومية، حيث يمكن أن تستنتج الحكومات معلومات ثمينة جداً من خلال تحليل تلك البيانات المخزنة.
- على الدولة استغلال هذا المورد من خلال توفير البنى التحتية اللازمة من قواعد البيانات وأدوات التحليل الى خلق مشاريع جديدة لتكوين أشخاص قادرين على دراسة البيانات الضخمة، كما يجب اعداد سياسات واستراتيجيات للبيانات الضخمة على المستوى الاقليمي، لأن مستقبل البيانات الضخمة واضح وله وجهة واحدة فقط –النمو-.

4. آفاق الدراسة

موضوع البيانات الضخمة هو دراسة واسعة جدا دائمة التجدد والتغير مع مرور الوقت، فتعتبر هذه الورقة مجرد تمهيد لهذا الموضوع الثري بمحتواه المعقد ودائم النمو، وكطلبة نحن دائما نسعى للإثراء مكتسباتنا العلمية بالمزيد، ودراسة مثل البيانات الضخمة هي جديرة بالتعمق أكثر فيها، لذلك يمكننا طرح النقاط التالية في شكل آفاق يمكن دراستها في المستقبل بإذن الله:

- اعداد دراسات في مجال علوم البيانات
- محاولة تطبيق البيانات الضخمة على بعض المؤشرات الاقتصادية على المستوى الوطني
- رقمنة قواعد البيانات

المراجع

المراجع:

- **المراجع العربية**

1) عمار محمد هلال: قواعد البيانات باستخدام SQL, 2017-2018.
2) هناء قيراطي: توظيف البيانات الضخمة في الشركات التقنية وخصوصية المستخدم، مذكرة تخرج شهادة ماستر، جامعة 8 ماي 1945 –قالمة-, 2016-2017.
3) مركز الإحصاء والتنافسية: مفاهيم عامة عن البيانات الضخمة 2021، حكومة عجمان، الامارات العربية المتحدة، فيفري 2021.
4) مركز الإحصاء: مفاهيم عامة حول البيانات الكبيرة، أدلة المنهجية والجودة، دليل رقم 13، أبو ظبي.

- **المراجع الأجنبية**

1) Abdullah Gani: Big data storage technologies: a survey, Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur 50603, Malaysia, August 8, 2017.
2) Calder, B: Windows azure storage: a highly available cloud storage service with strong consistency. In: Proceedings of the Twenty-Third ACM Symposium on Operating Systems Principles, 2013.
3) Chu, C Kim: Map-reduce for machine learning on multicore. Advances in Neural Information Processing Systems, 2007.
4) Daniel J. Solove: UNDERSTANDING PRIVACY, Harvard University Press, GWU Legal Studies Research Paper No. 420, May 2008.
5) Dean. J: MapReduce: Simplified data processing on large clusters. Communications of the ACM, 51(1), 2008.
6) Douglas Laney: Big data means big business, 2013, Gartner, Inc.
7) Druba Borthakur: Apache Hadoop goes realtime at facebook, Apache Hadoop Code Repository, 2006.
8) Fensel, D van Harmelen: Towards LarKC: A platform for web-scale reasoning. Los Alamitos, CA: IEEE Computer Society Press, 2007.
9) Hu .h: Toward Scalable Systems for Big Data Analytics: A Technology Tutorial. Toward Scalable Systems for Big Data Analytics: A Technology Tutorial, 2014.
10) Jacobs. A: The pathologies of big data, Communications of the ACM, 2009.

11) José María Cavanillas: New Horizons for a Data-Driven Economy a Roadmap for Usage and Exploitation of Big Data in Europe, Springer Open, Springer International Publishing AG Switzerland, 2016.

12) Laney, D: 3D data management: Controlling data volume, velocity, and variety. Technical Report, META Group, 2001.

13) M.H.Padgavankar: Big Data Storage and Challenges, International Journal of Computer Science and Information Technologies, Amravati, Maharashtra, India, Vol. 5 (2).

14) McKinsey Global Institute: Big Data: The next frontier for innovation, competition, and productivity. McKinsey & Company, June 2011.

15) Melnik, S Garcia-Molina: Similarity flooding: A versatile graph matching algorithm and its application to schema matching. In Proceedings of the 18th International Conference Data Engineering. IEEE Computer Society, 2012.

16) Omer Tene & Jules Polonetsky: To Track or 'Do Not Track': Advancing Transparency and Individual Control in Online Behavioral Advertising, 13 MINN. J.L. SCI. & TECH. 281, 2012.

17) Roy Furchgott, Comcast to Protect Customer's Computers from Malware, N.Y. TIMES GADGETWISE Sept. 30, 2010.

18) Ruth Gavison: Privacy and the Limits of Law , The Yale Law Journal Company, Inc, Vol. 89, N. 3,January 1980.

19) Sameer Wadkar: Pro Apache Hadoop, The expert's voice in big data, Friendsof Apress, 2nd edition, 2014.

20) Sanders John: Defining Terms: Data, Information and Knowledge, 2016

21) Shenker, S Stoicam: SQL and rich analytics at scale. In Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data.

22) Shneiderman. B: The eyes have it: A task by data type taxonomy for information visualizations, In Proceedings of Visual Languages, 2002.

23) Spence. R: Information visualization – design for interaction, 2nd edition, Upper Saddle River, NJ: Prentice Hall, 2006.

24) Stonebraker, M: What does 'big data' mean? Communications of the ACM, BLOG, 2012.

25) Thusoo, A Sarma: Hive – a warehousing solution over a map-reduce framework. Statistics and Operations Research Transactions, 2, 2009.

- **المراجع الالكترونية**

1) http://dismagazine.com/discussion/73314/tcf-data-awareness
2) http://franz.com/agraph/allegrograph/
3) http://hadoop.apache.org
4) http://searchstorage.techtarget.com/definition/block-storage
5) http://virtuoso.openlinksw.com/dataspace/doc/dav/wiki/Main/VOSArticleVirtuosoAHybridRDBMSGraphColumnStore
6) http://www.1cloudroad.com/is-enterprise-cloud-storage-a-good-fit-for-your-business
7) http://www.cs.uinbc.edu/-hillol/DDM-BIB
8) https://arabicprogrammer.com/article/7680515806/
9) https://aws.amazon.com/big-data/what-is-spark/
10) https://computer.howstuffworks.com/floppy-disk-drive1.htm
11) https://differencebtw.com/data-vs-information/
12) https://iapp.org/about/what-is-privacy/
13) https://nix-united.com/blog/12-big-data-issues-growing-companies-face/
14) https://opensource.com/life/14/8/intro-apache-hadoop-big-data#:~:text=Hadoop%20was%20created%20by%20Doug,after%20his%20son's%20toy%20elephant
15) https://phoenixnap.com/kb/newsql
16) https://privacyinternational.org/explainer/56/what-privacy
17) https://www.analyticssteps.com/blogs/what-hive-big-data-and-its-benefits
18) https://www.backupify.com/history-of-data-storage/
19) https://www.computerhistory.org/storageengine/tape-unit-developed-for-data-storage
20) https://www.datamation.com/big-data/structured-vs unstructureddata/#:~:text=unstructured%20data%3A%20structured%20data%20is,video%2C%20and%20social%20media%20postings
21) https://www.edureka.co/blog/big-data-characteristics/#volume
22) https://www.geeksforgeeks.org/hadoop-history-or-evolution/
23) https://www.geeksforgeeks.org/hadoop-yarn-architecture
24) https://www.geeksforgeeks.org/introduction-to-apache-pig/#:~:text=Pig%20is%20a%20high%2Dlevel,develop%20the%20data%20analysis%20code
25) https://www.heavy.ai/learn/big-data-analytics

26) https://www.ibm.com/analytics/predictiveanalytics#:~:text=Predictive%20analytics%20is%20a%20branch,to%20identify%20risks%20and%20opportunities
27) https://www.ibm.com/cloud/blog/structured-vs-unstructured-data
28) https://www.ibm.com/ibm/history/ibm100/us/en/icons/punchcard/
29) https://www.ibm.com/topics/hbase#:~:text=HBase%20is%20a%20column%2Doriented,many%20big%20data%20use%20cases
30) https://www.javatpoint.com/big-data-characteristics
31) https://www.mymarketresearchmethods.com/types-of-data-nominal-ordinal-interval-ratio
32) https://www.philips.com/a-w/about/innovation/research.html
33) https://www.scnsoft.com/blog/big-data-challenges-and-their-solutions
34) https://www.techtarget.com/searchcio/definition/decision-support-system#:~:text=A%20decision%20support%20system%20(DSS)%20is%20a%20computer%20program%20application,the%20best%20possible%20options%20available
35) https://www.techtarget.com/searchcio/definition/decision-support-system#:~:text=A%20decision%20support%20system%20(DSS)%20is%20a%20computer%20program%20application,the%20best%20possible%20options%20available
36) https://www.techtarget.com/searchdatamanagement/definition/Apache-Hadoop-YARN-Yet-Another-Resource-Negotiator
37) https://www.techtarget.com/searchdatamanagement/definition/data
38) https://www.techtarget.com/searchdatamanagement/tip/10-big-data-challenges-and-how-to-address-them
39) https://www.techtarget.com/searcherp/The-ultimate-guide-to-ERP
40) https://www.teradata.com/Glossary/What-is-Semi-Structured-Data#:~:text=Semi%2Dstructured%20data%20refers%20to,not%20have%20a%20fixed%20schema
41) https://www.tibco.com/reference-center/what-is-data-exploration#:~:text=Data%20exploration%20is%20the%20first,and%20get%20to%20insights%20faster
42) https://www.tutorialspoint.com/map_reduce/map_reduce_introduction.htm#:~:text=MapReduce%20is%20a%20programming%20model,huge%20volumes%20of%20complex%20data
43) www.tutorialspoint.com

الملخص

تهدف الدراسة الى التعريف بموضوع البيانات الضخمة، التي أخذت بالتراكم بفضل تقنيات التخزين والتكنولوجيات الجديدة ، التي أصبحت منتجة للبيانات منظمة وغير منظمة، كالحواسيب والهواتف النقالة وصولا إلى الأجهزة المرتبطة بالأنترنت كأجهزة التلفاز وأدوات الملاحة وهناك العديد من المصادر الأخرى التي تولد هذا الكم من البيانات على غرار مواقع التواصل الاجتماعي، وتبيان أهمية وأثر استخدام البيانات الضخمة في مجال العلوم الاقتصادية من خلال استخراج القيمة من هذا الكم الهائل من البيانات، واستغلالها في مجال عملها، وذلك باستخدام برامج وأدوات تقنية جديدة قادرة على التعامل معها، والكشف عن التحديات والمعيقات التي تواجهها تطبيقات البيانات الضخمة في هذا المجال من مشكل نقص المهارات الى مشكل الخصوصية وأمن المعلومات، خاصة في ظل تسابق مختلف الشركات العالمية نحو امتلاك هذه البيانات الضخمة في شكلها الخام و تحليلها لاستخراج القيمة منها الى أن هذه العملية تشوبها الكثير من التجاوزات في حق أصحاب البيانات و من بينها انتهاك خصوصية مستخدمي الانترنت.

الكلمات المفتاحية: البيانات الضخمة، تخزين البيانات، تحليل البيانات، قاعدة البيانات، أباشي هادوب.

Abstract

The study aims to introduce the topic of big data, which is accumulating thanks to storage techniques And new technologies, which have become producers of organized and unorganized data, such as computers and mobile phones, and there are many other sources that generate GPS to Internet-related devices such as televisions and navigation tools, this amount of data such as social networking sites, and show the importance and impact of the use of big data in the field of Economic sciences by extracting value from this huge amount of data, and exploiting it in its field of work, using new technical programs and tools capable of dealing with it, and revealing the challenges and obstacles that big data applications face in this field from the problem of lack of skills to the problem of privacy and information security , especially in light of the various international companies racing towards owning this huge data in its raw form and analyzing it to extract value from it, that this process is marred by many abuses against data owners, including the violation of the privacy of internet users.

Key Words: Big Data, Data Storage, Data Analyses, Data Base, Apache Hadoop.