الفصل

الثامن

البيانات الضخمة

Big Data

ويشتمل هذا الفصل على النقاط التالية:

- ✓ مفهوم البيانات الضخمة
- ✓ خصائص البيانات الضخمة
- ✓ البيانات الضخمة والذكاء الاصطناعي

الفصل الثامن

البيانات الضخمة Big Data

مفهوم البيانات الضخمة:

توجد ثمة علاقة طردية بين التقدم التقني في عالمنا الذي نعيشه الآن وتطور وانتشار الانترنت وبين البيانات التي يتم إنتاجها باستمرار بسرعات وأحجام وأشكال كبيرة، ومما يوضح ويؤكد هذا إنترنت الأشياء Internet of Things (IOT)[1]، والانتشار الهائل للأجهزة المتصلة بالإنترنت، فقد ارتفع عدد الأجهزة المؤهلة عبر الإنترنت بنسبة 31٪ من عام 2016 إلى 8.4 مليار في عام 2017، ويقدر الخبراء أن إنترنت الأشياء سيشمل حوالي 30 مليار قطعة بحلول عام 2020، ومن المقدر أيضاً أن تبلغ القيمة السوقية العالمية لإنترنت الأشياء 7.1 تريليون دولار بحلول عام 2020.

وقد أجبر هذا الانفجار الكمي للبيانات الرقمية كبار مشغلي الانترنت (Yahoo،

(1) **إنترنت الأشياء**: Internet Of Things هو مفهوم متطور لشبكة بالإنترنت بحيث تمتلك كل الأشياء في حياتنا قابلية الاتصال بالإنترنت أو ببعضها البعض لإرسال واستقبال البيانات لأداء وظائف محددة من خلال الشبكة.

Google، Amazon، Facebook، وما إلى ذلك)، والتي تعد البيانات هي "المواد الخام" لهذه الشركات، على البحث عن طرق جديدة لمعالجة وتحليل هذه البيانات الضخمة (باستخدام صناعة الكمبيوتر) مع عنصر السرعة (القدرة على تنفيذ تحليلات في الوقت الحقيقي القريب). وهكذا، ولد مفهوم البيانات الضخمة.

في نهاية المطاف، تعتبر **البيانات الضخمة**: مجموعة أو مجموعات من البيانات الكبيرة والمعقدة لها خصائصها الفريدة (مثل الحجم، السرعة، التنوع، التباين، صحة البيانات)، والتي لا يمكن مُعالجتها بكفاءة باستخدام أدوات معالجة البيانات التقليدية لتحقيق الاستفادة منها. وتكمن التحديات التي ترافق هذا النوع من البيانات في توفيرها، معالجتها، تخزينها، تحليلها، البحث فيها، ومشاركتها ونقلها، تصويرها، تحديثها بالإضافة إلى المحافظة على الخصوصيات التي ترافقها.

وعلى الرغم من تزايد حجم البيانات المتاحة بتنسيقات مختلفة فإن تكلفة تخزين هذه البيانات آخذة في التناقص، مما يجعل تخزين كميات كبيرة من البيانات أمراً سهلاً للغاية، ومع ذلك لا تزال مشكلة معالجة هذه البيانات (من حيث الحجم، التنسيق، والتنوع) مستمرة؛ حيث تركز البيانات الضخمة على البُعد التقني، بينما تهتم البيانات الذكية Smart Data أكثر بالبعد التحليلي، وقيمة وتكامل البيانات في عمليات صنع القرار في المؤسسات المختلفة.

ينبغي النظر إلى البيانات الضخمة على أنها مصدر بيانات جديد يجب على الشركات والمؤسسات دمجها وربطها ببياناتها الحالية، وليس كمفهوم يمكن أن يحل محل "ذكاء الأعمال" الحالي، تضيف البيانات الضخمة نفسها إلى مجموعة الحلول التي وضعتها الشركات من أجل معالجة البيانات الكبيرة واستغلالها ونشرها بهدف مساعدتها في اتخاذ قرارات مستنيرة، سواء لأغراض استراتيجية أو تشغيلية.

خصائص البيانات الضخمة:

يتم تعريف البيانات الضخمة بشكل رئيس بخصائص (عناصر) أربعة: الحجم، التنوع، السرعة، الصدق بالإضافة إلى القيمة المرتبطة بالبيانات، تشير جوانب الحجم والسرعة إلى عملية إنشاء البيانات (كيفية التقاط البيانات وتخزينها)، بينما تتعامل الصدق والقيمة مع جودة فائدة من البيانات.

1- الحجم Volume:

هو حجم البيانات المستخرجة من مصدر ما، وهو ما يحدد قيمة وحجم البيانات لكي تُصنف ضمن البيانات الضخمة، وبحلول عام 2020 سيحتوي الفضاء الإلكتروني على ما يقرب من 40.000 ميتابايت من البيانات الجاهزة للتحليل واستخلاص المعلومات.

في عام 2014، كان هناك حوالي 3 مليارات مستخدم للإنترنت متصلا عبر أكثر من 6 مليارات جهاز (الخوادم، أجهزة الكمبيوتر الشخصية، الأجهزة اللوحية، والهواتف الذكية)، ومع ظهور الأجهزة الأخرى في حياتنا اليومية التي يتم توصيلها بالإنترنت مثل (أجهزة التلفزيون، الأجهزة المنزلية، كاميرات المراقبة، وما إلى ذلك) سيكون لدينا عشرات المليارات من هذه الأجهزة (حوالي 50 مليار جهاز) والتي يمكن أن تولد أكثر من (40،000 مليار بايت) من البيانات في السنة.

البيانات الضخمة المولدة عبر مليارات الأجهزة قد يكون بعضها ذا قيمة وصلة بالنسبة لشركة ما، والبعض الآخر أقل أهمية، ولكن لكي تتمكن من تحديد ذلك، من الضروري أن تكون قادراً على قراءتها، فرزها، تنظيمها، تحليلها ثم الانتقال نحو التخزين.

2- التنوع Variety:

يُقصد به تنوع البيانات المستخرجة، والتي تساعد المستخدمين سواء كانوا باحثين أو مُحللين على اختيار البيانات المناسبة لمجال بحثهم، وتتضمن بيانات مُهيكلة (Structured Data) في قواعد بيانات، وبيانات غير مُهيكلة (Un Structured Data)

مثل: الصور، مقاطع وتسجيلات الصوت، الفيديو، الرسائل القصيرة، سجلات المكالمات، وبيانات الخرائط (GPS)، وتتطلب وقتاً وجهداً لتهيئتها في شكل مناسب للتجهيز والتحليل.

3- السرعة Velocity:

يُقصد بها سرعة إنتاج واستخراج البيانات لتغطية الطلب عليها حيث تعتبر السرعة عنصراً حاسماً في اتخاذ القرار بناء على هذه البيانات، وهو الوقت الذي نستغرقه من لحظة وصول هذه البيانات إلى لحظة الخروج بالقرار بناءً عليها.

4- الموثوقية والصحة Veracity:

ما هي موثوقية مصدر البيانات، ومدى دقتها وصحتها وحداثة تلك البيانات، حيث أن هناك مدير تنفيذي من بين كل ثلاثة مدراء لا يثقون في البيانات التي تعرض عليه لاتخاذ القرار، كما أن هناك دراسات تقدر أن حجم ضرر البيانات الغير جيدة على الاقتصاد الامريكي يقدر 3.1 ترليون دولار سنويا.

5- القيمة Value:

ما هي القيمة التي يمكن الحصول عليها من البيانات الضخمة؟ هذا هو جوهر الموضوع، ما ينطبق على البيانات الضخمة ينطبق على جميع البيانات، يمكننا أن نقول: أن البيانات بدون قيمة (التي لا يتم استغلالها) أقل تكلفة (معالجتها، التخزين، إلخ) من البيانات التي يتم استغلالها وبالتالي، فإن قيمة البيانات تكمن في استخدامها.

تدرك الشركات تماماً أنها بعيدة كل البعد عن استغلال جميع البيانات المتاحة لها، وذلك بسبب أنها تركز بشكل أساسي على البيانات عالية التنظيم من أنظمة المعاملات. لكن في ظل العولمة بالإضافة إلى التضخم في رقمنة عالمنا الذي نعيش فيه أدى إلى زيادة الوعي، وأصبحت المنافسة أكثر صرامة، والفرص أكبر والقدرة على "المعرفة" قبل التصرف هي ميزة حقيقية.

البيانات الضخمة يجب أن يُنظر إليها على أنها مصدر إضافي للمعلومات (منظم وغير منظم) من شأنها إثراء عمليات صنع القرار في الشركات والمؤسسات (التقنية والبشرية)، حيث تبدأ بتحويل البيانات الضخمة إلى بيانات ذكية.

البيانات الضخمة والذكاء الاصطناعي:

تعد كمية البيانات التي يتم إنشاؤها كل عام ضخمة وتستمر في النمو بشكل كبير من حوالي 150 Exabytes إلى ما يقرب من 1200 Exabytes خلال خمس سنوات، نقوم بإنشاء 2.5 تريليون بايت من البيانات كل يوم، يقوم مستخدمو Twitter بتوليد أكثر من 500 مليون تغريدة كل يوم، ويتم تحميل كمية مماثلة من الصور على Facebook، في عام 2016 يتميز الرسم البياني على Facebook والذي يعكس علاقة الصداقة بين مستخدميه بأكثر من مليار عقدة وأكثر من مئات مليارات الصداقات، ويبلغ حجم شبكة الويب العالمية المفهرسة (المقدرة عبر حجم مؤشر Google) أكثر من 45 مليار صفحة ويب، وتقوم Google وحدها بإجراء عدة مليارات من عمليات البحث عليها يومياً.

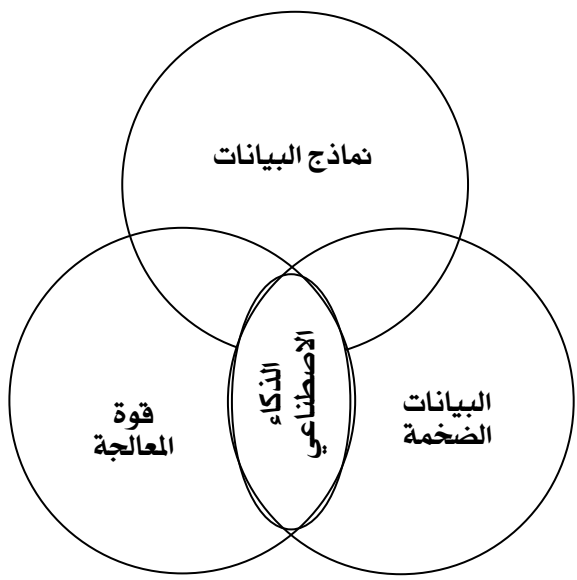
بالإضافة إلى القفزة الكبيرة في ظهور المعالجات المتوازية على نطاق واسع، ولا سيما المعالجات الرسومية GPU، وهي وحدات معالجة متوازية ضخمة تضم الآلاف من النوى، مقابل العشرات في وحدة المعالجة المركزية CPU. وقد سارع هذا إلى حد كبير خوارزميات أنظمة الذكاء الاصطناعي القائمة وجعلها الآن قابلة للحياة.

البيانات الضخمة موجودة في كل مكان تقريباً، ومن المهم جداً الحفاظ على البيانات التي يتم توليدها بكمية هائلة، لذا لا يجب تفويت أي شيء. من الصعب للغاية تخزين هذا الكم الهائل من البيانات التي تولدها أي شركة. تقنيات الحوسبة التقليدية غير قادرة على التعامل مع مجموعات البيانات الكبيرة هذه. غالباً ما يستخدم الذكاء الاصطناعي لمعالجة هذا النوع من البيانات. أساساً، الذكاء الاصطناعي وعلومه الفرعية (على سبيل المثال آلة الميل، التعلم العميق، الشبكات المحايدة،...)، كلها خوارزميات تستخدم هذه الكميات الهائلة من البيانات (البيانات الضخمة) لتحقيق النتائج المرجوة وإيجاد الاتجاهات

والأنماط والتنبؤات، حيث يتم تنفيذ المهام التحليلية المعقدة بشكل أسرع من الخيال البشري على البيانات الضخمة بمساعدة أنظمة الذكاء الاصطناعي.

في العالم الحقيقي، تمتلك البيانات بعض الخصائص غير المرغوب فيها منها: حجمها الضخم بالشكل الذي لا يمكن تصوره، ليست منظمة تنظيماً جيداً أو منسقة بشكل جيد، أنها تتغير باستمرار بسرعات كبيرة، وغيرها، تقنية الذكاء الاصطناعي أحد الطرق الهامة لتنظيم المعرفة واستخدامها بكفاءة بحيث تكون هذه البيانات مفهومة من قبل الأشخاص الذين يقدمونها، وأن تكون قابلة للتعديل بسهولة لتصحيح الأخطاء، كما يجب أن تكون مفيدة في العديد من الحالات.

فالبيانات الضخمة هي العامل الرئيس والمدخلات الأولية التي تحتاج إلى تنظيف وهيكلة متكاملة قبل أن تصبح مفيدة، والتي يتم تنظيفها وإزالة البيانات المكررة وغير الضرورية منها باستخدام ماكينة (آلة) الذكاء الاصطناعي



العلاقة بين البيانات الضخمة والمعالجات والذكاء الاصطناعي

العلاقة التي شُكلت بين الذكاء الاصطناعي والبيانات الضخمة علاقة تبادلية طردية، لا يوجد ذكاء اصطناعي بدون بيانات كبيرة، لأن الذكاء الاصطناعي يحتاج إلى بيانات لبناء ذكائه ولإطعام هذه المعالجات، وكلما زادت مدخلات أنظمة الذكاء الاصطناعي (بيانات الضخمة)، زادت دقة النتائج التي يمكن أن يحققها.

في الماضي، لم يتم تطوير أنظمة الذكاء الاصطناعي بشكل جيد بسبب المعالجات البطيئة، ومجموعات البيانات الصغيرة، لم يكن هناك أجهزة استشعار مثل اليوم، حيث يمكن أن تحتوي السيارة على عشرات من أجهزة الاستشعار المدمجة فيها. ولم تكن هناك بيانات كبيرة كما في الوقت الحالي لأن الإنترنت لم تكن متوفرة على نطاق واسع.

لماذا تعمل البيانات الضخمة والذكاء الاصطناعي معاً بشكل جيد؟

تواجه حلول "ذكاء الأعمال"[1] التقليدية (تلك التي كانت قائمة منذ بدايات عام 2000) صعوبة كبيرة في دمج البيانات الضخمة، والتي تستمر في ملء عالمنا الرقمي كل يوم بشكل كبير. وقد أدى هذا الوضع إلى قيام العديد من الشركات بتحويل البيانات الضخمة إلى "بيانات غامضة" (بيانات ضعيفة أو غير مرئية، يتم تجاهلها في العمليات التحليلية والتي يتم التحدث عنها بشكل فضفاض ولكنها لا ترى في الواقع ضوء النهار)، الأسباب الرئيسية لهذا:

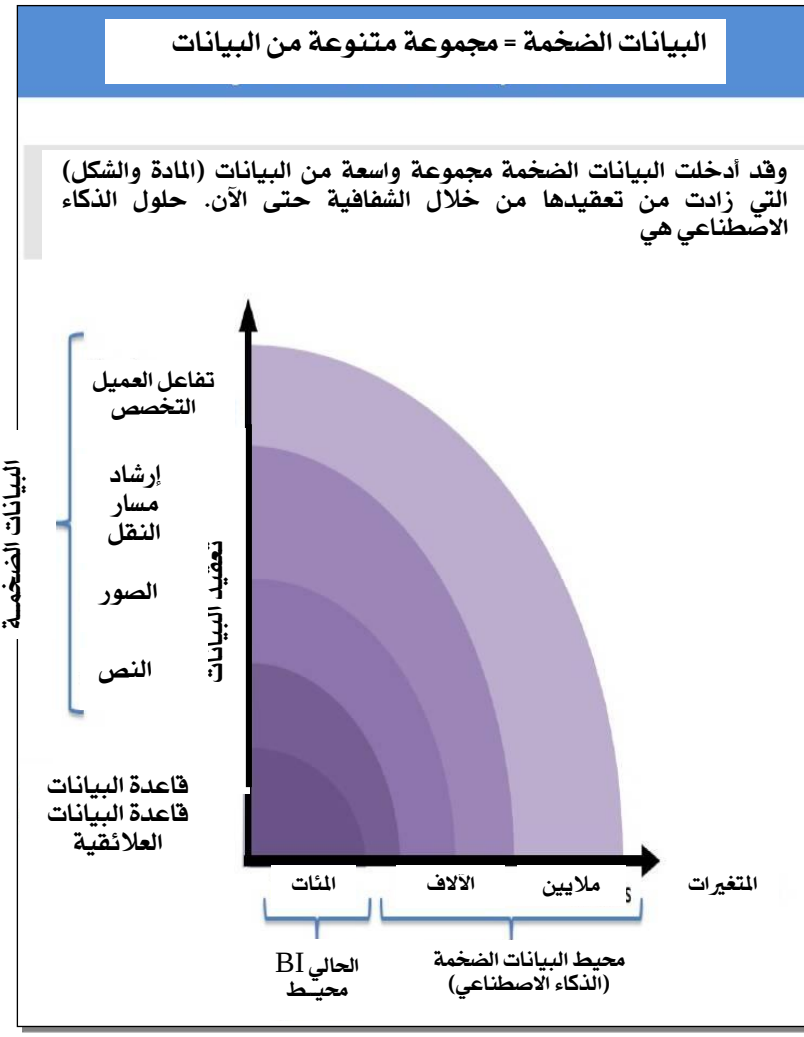
- نقص وقصور في البيانات المنظمة (الصور والأصوات والمدونات والنصوص، إلخ.)، والتي يصعب دمجها في حلول ذكاء الأعمال القائمة.
- تخزين البيانات في ملفات وقواعد بيانات متعددة وغير مترابطة مع بعضها البعض.

(1) **ذكاء الأعمال**: عبارة عن نظريات ومنهجيه وعمليات وخصائص وتقنيات تقوم على تحويل البيانات الأولية إلى معلومات مفيدة ذات معنى لأغراض الأعمال. له القدرة على التعامل مع أعداد هائلة من المعلومات لمساعدة المؤسسات في تحديد وتطوير الفرص التجارية الجديدة، والاستفادة منها.

- تقليدية العمليات التحليلية، والتي تم انسدادها بكميات كبيرة (الحجم والتنوع) من المتغيرات لتتكون مترابطة مع بعضها البعض.
- مؤقتية التحليل والقرار والعمل.

بالإضافة إلى أن حلول "ذكاء الأعمال" التقليدية ليست متصلة بشكل جيد بعالم المتغيرات، مما يؤدي إلى عدم وجود استجابة والتكيف مع التغيير. كما أن القيود التي تفرضها حلول "ذكاء الأعمال" التقليدية لا تسمح لنا أو تسمح بالكاد باستغلال EL Dorado الاتجاهات والمذاهب الجديدة من البيانات الضخمة.

الذكاء الاصطناعي ينقل التركيز والاهتمام على عمليات تنفيذ واتخاذ القرار، وكذلك الإجراءات الآلية والتعلم الذاتي، في إطار زمني لا يتجاوز بضعة أجزاء من الألف من الثانية. إن زيادة عدد الأجهزة المتصلة التي تدعمها رقمنة عالمنا الذي نعيش فيه الآن لن يؤدي إلا إلى زيادة حجم المعلومات المتاحة (في كل ثانية، ننتج 6000 تغريدة 40.000 بحث في Google، 2 مليون بريد إلكتروني)، بحلول عام 2019 حجم البيانات المتاحة سوف يتجاوز 4.4 زيتابايت (4.4 تريليون جيجابايت) سنوياً. ومن هنا يبدو التقارب بين البيانات الضخمة والذكاء الاصطناعي والذي سيجعل من الممكن استخراج المعلومات القيمة من هذه البيانات الهائلة والمتزايدة بسرعة متناهية. الشركات الكبيرة بدأوا بالفعل هذا التقارب.



البيانات الضخمة وتنوع البيانات