

تحليل البيانات الضخمة Big Data analytics :

أولاً تعريفات البيانات الضخمة:

يعود أصل مصطلح "البيانات الضخمة" إلى منتصف تسعينيات القرن العشرين، حيث استخدمه لأول مرة جون ماشي، كبير العلماء السابقين المتقاعدين في شركة سيليكون غرافيكس، للإشارة إلى التعامل مع مجموعات البيانات الضخمة وتحليلها،

1-البيانات الضخمة: هي كميات هائلة من البيانات التي تتجاوز القدرة على معالجتها باستخدام الأدوات والتقنيات التقليدية. هذا النوع من البيانات يتطلب حلولاً مبتكرة ومرنة لمعالجتها بسرعة وفعالية.

2- يتم تعريف البيانات الضخمة على أنها تدفق مستمر من المعلومات التي يتم مراقبتها وتحليلها بشكل مستمر، حيث تساهم مجموعة من العوامل مثل التفاعلات الاجتماعية، الأجهزة المحمولة، البحث والتطوير، والبنية التحتية المادية في تكوين هذا التدفق المستمر للبيانات.

3-البيانات الضخمة: تشير إلى كميات ضخمة ومعقدة من البيانات التي تحتاج إلى أدوات وتقنيات متطورة لمعالجتها وتحليلها، لتقديم رؤى ونتائج قيمة يمكن استخدامها في اتخاذ القرارات وتحقيق الكفاءة.

4-البيانات الضخمة: هي مجموعة واسعة من البيانات التي تتمتع بحجم كبير، وتنمو بسرعة عالية، وتأتي من مصادر متعددة. يتمثل التحدي في قدرتها على التوسع السريع واحتياجها إلى تقنيات متقدمة لتحليلها وتحويلها إلى معلومات قيمة تدعم اتخاذ القرارات وتحقيق الابتكار في مختلف المجالات.

ثانياً: أنواع البيانات الضخمة

1- البيانات المنظمة: (Structured Data)

البيانات المنظمة هي البيانات التي يتم تنظيمها في هيكل منظم، بحيث يمكن تخزينها في قواعد بيانات علائقية أو مستودعات منسقة مثل جداول البيانات. يتم تنظيم هذه البيانات في صفوف وأعمدة في الجداول، مما يسهل معالجتها وتحليلها باستخدام أدوات مثل نظم إدارة قواعد البيانات العلائقية (RDBMS) أو أنظمة إدارة قواعد البيانات (DBMS) على سبيل المثال، يمكن تخزين بيانات العملاء في جدول يحتوي على أعمدة مثل "الاسم"، "العنوان"، "رقم الهاتف"، و"البريد الإلكتروني". هذه البيانات تكون ذات هيكل ثابت، مما يسهل إجراء عمليات مثل الاستعلامات المعقدة والتحليل الإحصائي عليها. تعتبر البيانات المنظمة مثالية لاستخدام التقنيات التقليدية في تخزين البيانات وتحليلها.

2- البيانات غير المنظمة: (Unstructured Data)

البيانات غير المنظمة هي البيانات التي لا تتبع هيكلًا منظمًا أو قاعدة بيانات يمكن التعامل معها بسهولة باستخدام الأنظمة التقليدية مثل RDBMS. قد تتضمن هذه البيانات نصوصًا أو ملفات وسائط متعددة مثل الصور والفيديوهات والصوتيات، بالإضافة إلى محتوى منصات التواصل الاجتماعي مثل المنشورات والتعليقات والإعجابات. على سبيل المثال، يمكن أن تكون الصور ومقاطع الفيديو من Facebook أو YouTube أو ملفات PDF والرسائل النصية على منصات مثل WhatsApp أو Slack أمثلة على البيانات غير المنظمة. هذه البيانات يصعب تنظيمها وتحليلها باستخدام الأساليب التقليدية، ولذلك يتطلب التعامل معها تقنيات متقدمة مثل الذكاء الاصطناعي وتحليل النصوص أو أدوات متخصصة في تحليل الوسائط المتعددة.

3- البيانات شبه المنظمة: (Semi-structured Data)

البيانات شبه المنظمة هي بيانات تحتوي على بعض الهيكل التنظيمي ولكنها ليست منظمة تمامًا مثل البيانات المنظمة. هذه البيانات تتضمن بيانات وصفية تسمح بترتيبها وتنظيمها بطريقة مرنة، مما يسهل فهمها وتحليلها دون الحاجة إلى قاعدة بيانات صارمة. من الأمثلة الشائعة على البيانات شبه المنظمة هي ملفات HTML و XML و JSON. على سبيل المثال، تحتوي ملفات HTML على علامات مثل `

` و `

` التي توضح هيكل الصفحة، بينما تحتوي ملفات XML و JSON على عناصر وسمات يمكن استخدامها لتخزين البيانات بطريقة مرنة لكنها منظمة جزئيًا. بالرغم من أن البيانات شبه المنظمة قد لا تكون منظمة تمامًا مثل البيانات المنظمة، إلا أن هيكلها الأساسي يساعد في التعامل معها باستخدام تقنيات مثل قواعد البيانات NoSQL أو أدوات تحليل البيانات المهيكلة جزئيًا.

ثالثًا: خصائص البيانات الضخمة:

1- الحجم : يحدد حجم البيانات ما إذا كانت تعتبر بيانات ضخمة أم لا. تجمع معظم المنظمات البيانات من مصادر متعددة مثل وسائل التواصل الاجتماعي وأجهزة الآلات المختلفة. يتم جمع هذه البيانات بتردد عالٍ مما يؤدي إلى تكوين كميات كبيرة في فترة زمنية قصيرة.

2- السرعة : تشير إلى السرعة التي يتم بها جمع البيانات. تولد أجهزة إنترنت الأشياء والأجهزة المحمولة وأجهزة الاستشعار الأخرى كميات ضخمة من البيانات خلال ساعات قليلة.

3- التنوع : تأتي البيانات من أجهزة متعددة وتكون في تنسيقات مختلفة مثل النصوص، الصور، مقاطع الفيديو، وغيرها من التنسيقات المتنوعة.

4- الشمولية : البيانات الضخمة تلتقط النظام بالكامل بدلاً من أخذ عينات منه. يعني أنه يتم جمع جميع البيانات المتاحة بدلاً من العمل مع عينة فقط.

5- الدقة: البيانات الضخمة تتمتع بمستوى عالٍ من الدقة، مما يتيح استخدامها بشكل فعال في التحليل والتنبؤ.

6- العلاقة : تحتوي البيانات الضخمة على روابط بين حقول مشتركة تمكن من ربط مجموعات بيانات مختلفة مع بعضها البعض.

7- الامتدادية : البيانات الضخمة يمكن إضافة أو تعديل حقول جديدة فيها بسهولة، مما يساهم في تطوير وتحسين البيانات بمرور الوقت.

8- القدرة على التوسع: البيانات الضخمة يمكن أن تتوسع بسرعة كبيرة من حيث الحجم، مما يجعلها قابلة للتكيف مع الزيادة السريعة في حجم البيانات.

9- التباين: البيانات قد تتغير معانيها بشكل مستمر بناءً على السياق الذي يتم إنشاء البيانات فيه.

10- التعدد: البيانات تأتي من مصادر وأنواع متعددة، مثل المستشعرات، الإنترنت، والمزيد. من الصعب على الأنظمة التقليدية معالجة تنسيقات البيانات المتعددة في نفس الوقت.

11- الصدق : البيانات الضخمة تتنوع في جودتها. قد تكون بعض البيانات ذات مغزى، بينما تكون بيانات أخرى غير ذات معنى وتعتبر "ضوضاء".

رابعا تحليل البيانات الضخمة:

تحليل البيانات الضخمة هو عملية فحص مجموعات كبيرة من البيانات تحتوي على مجموعة متنوعة من أنواع البيانات لاكتشاف بعض المعرفة في قواعد البيانات، وتحديد الأنماط المثيرة للاهتمام وإقامة العلاقات لحل المشكلات واتجاهات السوق وتفضيلات العملاء وغيرها من المعلومات المفيدة. غالبًا ما تجني الشركات

والأعمال التي تنفذ تحليلات البيانات الضخمة العديد من الفوائد التجارية. تنفذ الشركات تحليلات البيانات الضخمة لأنها تريد اتخاذ قرارات تجارية أكثر استنارة.

أدوات وتقنيات تحليل البيانات الضخمة: يشمل تحليل البيانات الضخمة العديد من الأدوات والتقنيات المتقدمة، من أبرزها:

Hadoop: هو إطار عمل مفتوح المصدر يسمح بتخزين ومعالجة كميات كبيرة من البيانات على مجموعة من الخوادم.

Spark: هو إطار عمل لمعالجة البيانات الضخمة بسرعة عالية مقارنة بـ Hadoop.

NoSQL Databases: مثل MongoDB و Cassandra، وهي قواعد بيانات غير علائقية مصممة للتعامل مع البيانات غير المنظمة.

الذكاء الاصطناعي (AI) والتعلم الآلي (Machine Learning): يتم استخدام تقنيات الذكاء الاصطناعي لتحليل البيانات واستخراج الأنماط القيمة.

تصنيف تحليل البيانات الضخمة

التحليلات الوصفية: هي طريقة إحصائية تستخدم للبحث عن البيانات التاريخية وتلخيصها من أجل تحديد الأنماط أو المعنى.

تعتبر عملية تجميع البيانات واستخراج البيانات تقنيتين مستخدمتين في التحليلات الوصفية لاكتشاف البيانات التاريخية. يتم أولاً جمع البيانات وفرزها حسب عملية تجميع البيانات من أجل جعل مجموعات البيانات أكثر قابلية للإدارة من قبل المحللين.

يصف استخراج البيانات الخطوة التالية من التحليل وينطوي على البحث في البيانات لتحديد الأنماط والمعنى. يتم تحليل الأنماط المحددة لاكتشاف الطرق المحددة التي تفاعل بها المتعلمون مع محتوى التعلم وداخل بيئة التعلم.

يعتبر هذا النوع من التحليل أساسياً لفهم الماضي وتحديد ما حدث بالفعل. يمكن استخدامه في مجالات مثل تقارير الأداء، وتحليل سلوك العملاء في الماضي، وفحص الاتجاهات في البيانات القديمة. من خلال

تجميع البيانات واستخراج الأنماط منها، يمكن للشركات والمؤسسات تحديد ما إذا كانت استراتيجيات معينة قد حققت النجاح في الماضي أو لا.

المزايا:

يساعد في تحديد الأنماط المتكررة والسلوكيات السابقة.
مهم لتحديد المشكلات المحتملة أو الفرص في الماضي.
يوفر أساسًا متينًا لبناء استراتيجيات مستقبلية.

القيود:

لا يمكن استخدامه للتنبؤ بما سيحدث في المستقبل.
يقتصر على تحليل البيانات التاريخية فقط.

التحليلات التنبؤية

التحليلات التنبؤية هي طريقة إحصائية تستخدم الخوارزميات والتعلم الآلي لتحديد الاتجاهات في البيانات والتنبؤ بالسلوكيات المستقبلية بناءً على البيانات الحالية أو السابقة. باستخدام هذه الأدوات، يمكن توقع ما قد يحدث في المستقبل استنادًا إلى الأنماط والسلوكيات التي تم التعرف عليها.

يمكن استخدامه في العديد من المجالات مثل التنبؤ بمبيعات المنتجات، وتحليل المخاطر في الأعمال، وتوقع احتياجات العملاء المستقبلية، وغيرها.

المزايا:

يساعد في اتخاذ قرارات مستنيرة بناءً على التوقعات المستقبلية.
يمكن أن يوفر ميزة تنافسية من خلال التنبؤ بالاتجاهات قبل أن تحدث.
يستخدم تقنيات متقدمة مثل الخوارزميات والتعلم الآلي، مما يعزز دقة التنبؤات.

القيود:

يتطلب بيانات دقيقة وكبيرة لبناء نماذج التنبؤ.
لا يمكن ضمان 100% دقة التنبؤ، خاصة عندما تكون البيانات غير كافية أو مليئة بالضوضاء.