مبادئ علوم البيانات وتحليلات الأعمال

Data Science And Business Analytics

د. اسعد السعدني

أهداف التعلم

بنهاية هذا الفصل سيكون الطالب قادرا على أن:

١. يكتسب الطالب فهمًا لمصطلح علم البيانات ومفهومه.

٢. يميز بين المصطلحات: البيانات، المعلومات، والمعرفة.

٣. يفرق بين علم البيانات وذكاء الأعمال.

٤. يوضح التقارب بين علم البيانات والذكاء الاصطناعي.

٥. يعرف مراحل دورة حياة علم البيانات.

٦. يعرِّف مصطلح البيانات الضخمة، يحدد خصائصه، ويُصنِّف تقنياته.

٧. يعرف مفهوم إدارة البيانات ويُحدِّد مبادئ حوكمة البيانات.

٨. يناقش المهارات والأدوات الضرورية لممارسة علم البيانات.

٩. يُحدِّد مختلف المهن المرتبطة بمجال علم البيانات.

١٠. يعرف أهمية البيانات في تشكيل المجتمعات الرقمية وتطويرها

مقدمة في علم البيانات

Data Science

تكمن أهمية علم البيانات (Data Science) في أن البيانات أصبحت جزءا أساسيا في جميع الصناعات، فلقد أصبحت البيانات مطلبًا رئيسًا من قبل الشركات لكي تتوسع أعمالها وتتطور. حيث تمكن الشركات من اتخاذ القرارات المناسبة وذلك من خلال تحليل كميات كبيرة من البيانات لاستخراج رؤى وتوصيات قيمة لإدارة تلك الشركات.

علم البيانات هو مجال الدراسة الذي يتعامل مع كميات هائلة من البيانات باستخدام الأدوات والتقنيات الحديثة لإيجاد أنماط غير بديهية داخل تلك البيانات، وللوصول الى معلومات مهمّة يمكن أن تساهم في اتخاذ القرارات المتعلقة بكافة الأعمال.

مجالات تطبيق علم البيانات

- التطبيقات التجارية والصناعية
- الرعاية الصحية، والمعلوماتية الحيوية، والعلوم الطبيعية
- الاقتصاد الرقمي، وتحليل وسائل التواصل الاجتماعي والشبكات الاجتماعية
- المنازل الذكية، والمدن الذكية والمواصلات الذكية
- التعليم والتعلم الإلكتروني
- الطاقة، والاستدامة، والمناخ

البيانات والمعلومات Data and Information

تحيط بك البيانات بصورة يومية في كل مكان، فتتلقى المعلومات من التلفازومن الصحف والكتب وشبكة الإنترنت، ولكن هل فكرت في أن هناك فرقا بين البيانات والمعلومات؟ تعد البيانات تمثيلا للحقائق أو الأفكار بصورة شكلية، بحيث يمكن إيصالها أو معالجتها من خلال طريقة أو عملية ما.

- **البيانات**: تمثيل الحقائق أو الأفكار بتنسيق مناسب للتخزين أو المعالجة أو النقل.
- **المعلومات**: مجموعة من البيانات التي خضعت للمعالجة وأصبحت منظّمة ذات معنى وتقدّم في سياق محدد ومفيد وتُمكّن عمليات صنع القرار.

البيانات الأولية والمعلومات Raw Data and Information:

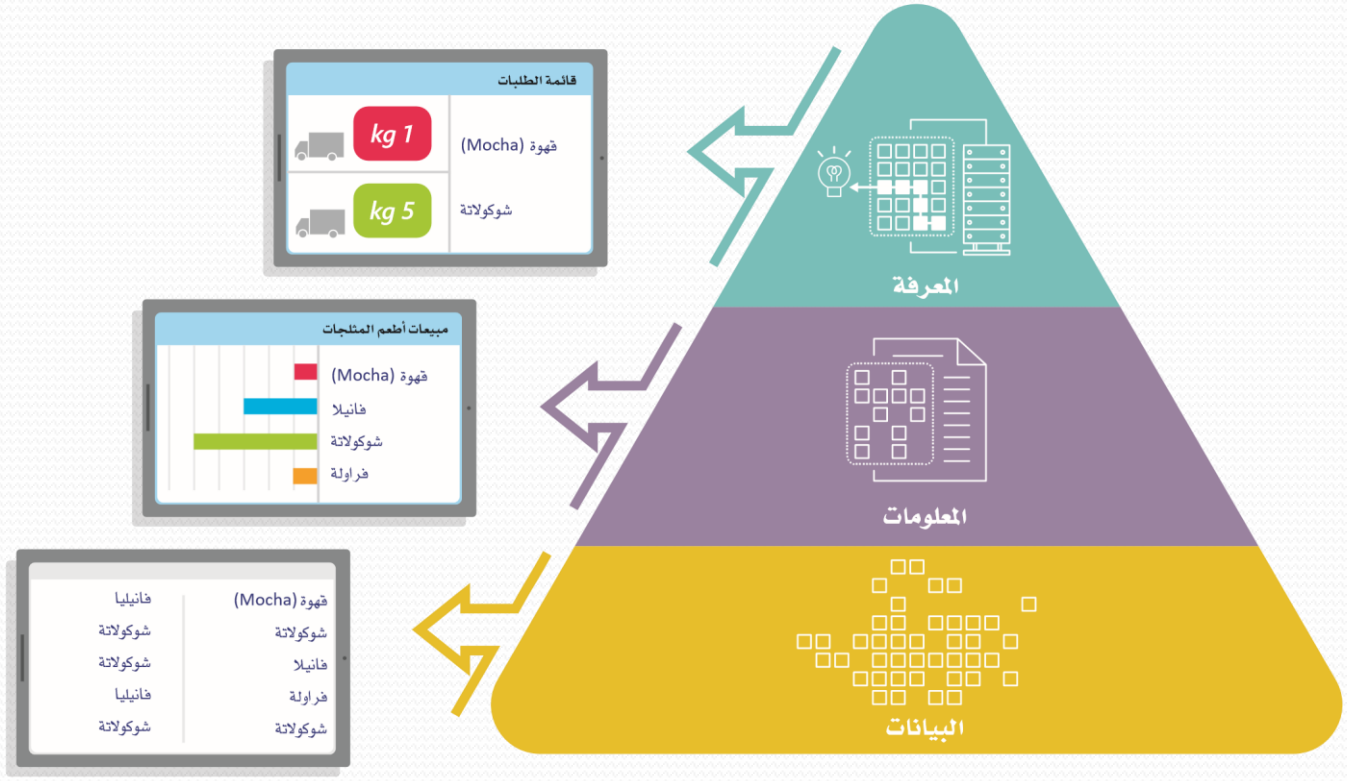
تطلق تسمية البيانات الأولية على البيانات التي تم جمعها حديثًا من مصادر مختلفة، ولكن لم يتم معالجتها أو تحليلها بعد بأي شكل من الأشكال، وعادة ما ترمز كلمة البيانات (Data) إلى البيانات الأولية، ولكن بمجرد تحليلها فإنها تتحول إلى معلومات.

جدول 1.1: أوجه الاختلاف بين البيانات والمعلومات

البيانات	المعلومات
غير منظَمة.	منظَمة منطقيًا.
يتم عرضها على صورة أرقام، رسومات، أو إحصائيات.	يتم تقديمها من خلال التقارير والرسوم البيانية أو المخططات.
مستقلة بذاتها.	تعتمد على البيانات.
يتم الحصول عليها من مدخلات المستخدمين أو من مدخلات محوسبة.	يتم الحصول عليها من عمليات معالجة البيانات.

Knowledge المعرفة

تعتبر المعرفة تمثيلا لفهمك للعالم، وهي بشكل أساسي مجموعة من المعلومات يتم استخدامها لتقديم فائدة أو تحقيق غرض معين. يمكنك القول إن فهم الشخص لبعض المعلومات حول شيء ما يوفر لديه معرفة به، فتصبح المعلومات معرفة عند تطبيق عمليات التفكير، أو التقييم، أو التخطيط، أو التنظيم.



شكل 1.3: هرم البيانات – المعلومات – المعرفة

جدول 1.2: أوجه الاختلاف بين المعلومات والمعرفة

	المعلومات	المعرفة
المعنى	بيانات تمت معالجتها لتصبح ذات سياق مفهوم.	استنتاجات مستقاة من المعلومات تساعد في اتخاذ القرارات.
القدرة على التنبؤ	لا تكفي وحدها للتوصل إلى استنتاجات أو قرارات.	تساهم في القدرة على التنبؤ واتخاذ القرارات.
إيصال النتائج للآخرين	يمكن إيصالها بسهولة من خلال الطرق الشفوية أو الورقية أو الإلكترونية.	تتطلب اطلاع بالموضوع المحدد.
النتائج	ينتج عنها الفهم.	ينتج عنها الإدراك.
الهدف	تجيب على أسئلة مثل من ومتى وماذا وأين.	تجيب على أسئلة مثل كيف ولماذا.

علم البيانات وذكاء الأعمال Business & Science Data Intelligence

- توجد البيانات في كل مكان من حولك، ويتم استخدامها ومعالجتها وتحليلها في جميع مجالات الحياة.
- تتطور نوعية البيانات واستخداماتها باستمرار، وتُستخدم بشكل خاص في العديد من التطبيقات المهمة مثل ذكاء الأعمال (Business Intelligence)، ولهذا يُعتبر ذكاء الأعمال عملية قائمة على التقنية لتحليل البيانات وتوفير معلومات مهمة تساعد المدراء التنفيذين وغيرهم من المسؤولين وصناع القرار على اتخاذ قرارات دقيقة خاصة بالأعمال.
- على الرغم من أن كلا من علم البيانات وذكاء الأعمال يتضمن العمل على البيانات، إلا أنهما يختلفان عن بعضهما.
- يُعد علم البيانات أكثر تعقيداً مقارنة بذكاء الأعمال، حيث يقتصر نطاق ذكاء الأعمال على مجال الأعمال، ويتم فيه تحليل البيانات السابقة من خلال تطوير لوحات المعلومات وعرض مستخلصات (رؤى) الأعمال، وكذلك ترتيب وتنظيم وتحليل البيانات وذلك لاستخراج المعلومات التي من شأنها مساعدة الشركات على النمو وتحقيق أهدافها بناء على فهم الاتجاهات الحالية للأعمال.
- يعتمد علم البيانات على استخدام البيانات المتوفرة للقيام بتنبؤات مستقبلية وعرض توقعات نمو الأعمال التجارية، وذلك بتوظيف مجموعة واسعة مما يسمى بالنماذج التنبؤية والخوارزميات الإحصائية المعقدة.

ذكاء الاعمال **Business Intelligence**: هو نظام مبني على البيانات ويشمل جمع وتخزين وتحليل وتمثيل البيانات لدعم عمليات اتخاذ القرارات.

يتمثل الدور الأساسي لأدوات ذكاء الأعمال في تحليل معلومات المؤسسات والشركات والمساهمة في إعداد استراتيجيات الأعمال، أما أدوات عالم البيانات فتشمل أدوات معالجة البيانات وأدوات البيانات الضخمة وكذلك نماذج خوارزمية معقدة لتحليل البيانات واستخلاص التوصيات.

جدول 1.3: أوجه الاختلاف بين علم البيانات وذكاء الأعمال

	علم البيانات	ذكاء الأعمال
المدى	تستخدم البيانات لعمل تنبؤات مستقبلية لتطوير الأعمال.	تحلل البيانات السابقة لاستنتاج الاتجاهات الحالية للأعمال.
الأدوات	يتضمن نماذج حسابية معقدة ومعالجة البيانات و أدوات البيانات الضخمة.	تقتصر الأدوات على تحليل المعلومات الإدارية والإشراف على استراتيجيات الأعمال.
أنواع البيانات	تتعامل بشكل أساسي مع البيانات غير المنظمة وشبه المنظمة، ويمكنها كذلك التعامل مع البيانات المنظمة.	تتعامل مع البيانات المنظمة التي يتم تخزينها عادة في مستودعات البيانات.
التعقيد	أكثر تعقيدا مقارنة بذكاء الأعمال.	أبسط بكثير مقارنة بعلم البيانات.
المرونة	أكثر مرونة حيث يمكن إضافة مصادر البيانات حسب الحاجة.	أقل مرونة حيث يجب تصميم مصادر البيانات مسبقا.

Data Science and علم البيانات والذكاء الاصطناعي Artificial Intelligence

كما تعرفت سابقًا على مفهوم علم البيانات، فإن مجال الذكاء الاصطناعي **Artificial Intelligence** يُعدُّ مجالً آخريتعامل مع كمٍ كبيرمن البيانات.

يمكن استخدام كل تقنية من هاتين التقنيتين بصوره منفصلة عن الأخرى للوصول لحلول لتحديات مختلفة، كذلك يمكن لكلٍ منهما إكمال بعضهما والتقارب معًا.

الذكاء الاصطناعي **Artificial Intelligence**:أحد مجالات علوم الحاسب ويهدف لبناء أنظمة قادرة على أداء المهام التي تتطلب عادةً ذكاءً بشريًا مثل القدرة على التعلم والاستدلال، وحل المشكلات ومعالجة اللغة الطبيعية والإدراك.

يختص علم البيانات بمعالجة البيانات التاريخية باستخدام أدوات حسابية للقيام بما يسمى بالتحليل الوصفي للبيانات **Descriptive Analysis** والذي يقدم وصفًا للمواقف المحددة، وكذلك للتنبؤ بالنتائج من خلال التحليل التنبؤي **Predictive Analysis**، ولتقديم الحلول والتوصيات للمشكلات من خلال التحليل التوجيهي **Prescriptive Analysis**. من أكثر الأدوات استخدامًا هي الأدوات الإحصائية والإدارية التي يمكن بواسطتها تحليل البيانات المؤرخة.

- يستخدم الذكاء الاصطناعي مجموعة متنوعة من التقنيات لمحاكاة الطريقة التي يفكر بها البشر والتي يقومون بناء عليها باتخاذ القرارات وتحليلها، فبدلً من التركيز على إجراء الحسابات الرياضية، يتم التركيز عند استخدام أدوات الذكاء الاصطناعي على عناصر المعرفة والذكاء كعناصر حاسمة لحل المشكلات. ويهتم الذكاء الاصطناعي كذلك بالحوسبة المعرفية.
- مشروعات علم البيانات المعقدة غالبًا ما تتضمن استخدام تقنيات تعلّم الآلة - أحد فروع الذكاء الاصطناعي - لتسهيل تحليل البيانات التنبؤية والتوجيهي.
- يستخدم الجيل القادم من أدوات علم البيانات ومنصات ذكاء الأعمال تعلّم الآلة للقيام ببعض الإجراءات مثل التعرّف على الأنماط في البيانات لاكتشاف الأنماط المخفية وتقديم التصورات والرؤى المهمة لاتخاذ القرارات
- عند الجمع بين علم البيانات والذكاء الاصطناعي، يمكن الحصول على طريقة فعالة جدًا في الحصول على نتائج دقيقة بشكل ملحوظ تساهم في اتخاذ قرارات أفضل وأسرع.

Data Science Life Cycle دورة حياة علم البيانات

يقوم علماء البيانات والمتخصصين في العمل على مشروعات علم البيانات بتوظيف خبراتهم من خلال خطوات محددة لتنفيذ كل مشروع جديد بكفاءة. يُطلق على هذه العملية اسم دورة حياة علم البيانات وتتضمن خمس مراحل تتميز كل مرحلة من المراحل المختلفة لهذه الدورة بخصائص معينة.

١. تعريف المشكلة وصياغتها Problem Definition and Formulation

فهم أهداف ومتطلبات العمل أو المشكلة العلمية وتحويل هذه المعرفة الى مسألة يمكن حلها بتحليل البيانات.

- من أجل تصميم وإيجاد حل لمشكلة بواسطة علم البيانات، فإنك تحتاج أولا إلى فهم ماهية المشكلة نفسها.
- يُعد التحليل الشامل للمشكلة وبيئتها والمتغيرات التي تؤثر عليها أمرًا ضروريًا لتطوير الحلول اللازمة لحل تلك المشكلة

جدول 1.4: أشهر أنواع تحليلات البيانات

تحليل الانحدار (Regression Analysis)	الحصول على الكميات أو الصفات الموجودة في مجموعة البيانات.
تحليل التصنيف (Classification Analysis)	تنظيم البيانات في فئات.
التحليل العنقودي (Clustering)	تنظيم البيانات في مجموعات.
تحليل انحراف البيانات (Anomaly Detection Analysis)	البحث عن انحراف أو شذوذ في البيانات.
نظم التوصية (Recommendation engines)	إعطاء توصية مستنيرة لمسألة محددة.

٢- جمع البيانات Data Collection

- عملية جمع القراءات أو الحقائق وتنسيقها، وتشمل الحصول عليها وتسميتها وتحسينها.

- بعد أن يتم تحديد الأهداف، يجب توفير مجموعة البيانات نفسها، ورغم انه قد يتم إدخال البيانات يدويًا أحيانًا، فمن المهم التنقيب وجمع البيانات، حيث يتعين في هذه المرحلة جمع بيانات كافية لمواصلة معالجتها.

- يمكن أن تأتي البيانات نفسها من مجموعة متنوعة من المصادر، فمثلًا تقوم أجهزة الاستشعار البيئية وتطبيقات الهاتف المحمول ومنصات الويب بتوليد البيانات بصورة مستمرة ليتم تخزينها تلقائيًا في قواعد البيانات.

جدول 1.5: تنسيقات تخزين البيانات الأكثر شيوعًا

الملفات المُنسقة (Formatted Files)	JSON و XML وCSV وجدول بيانات XLS.
قواعد البيانات العلائقية (Relational Databases)	خادم مايكروسوفت SQL وقاعدة بيانات أوراكل وأوراكل MySQL.
قواعد البيانات غير العلائقية (Non-Relational NoSQL Databases)	MongoDB و Azure Cosmos DB و AWS DynamoDB.
قواعد البيانات الرسومية (Graph Databases)	Neo4j و AWS Neptune و Dgraph.
قواعد بيانات السلاسل الزمنية (Time-series Databases)	InfluxDB وAWS Timescale.

٣- تجهيز البيانات وتنظيفها Data Preparation and Cleaning

- عملية متعددة المراحل لمراجعة البيانات وتصحيحها للتأكد من أنها في صيغة موحدة، ويتضمن ذلك معالجة القيم المفقودة والبيانات المشوشة، وحل التناقضات والتكرارات.
- تُعدُّ عملية "تنظيف" البيانات ومعالجتها أحد أهم المراحل في دورة حياة علم البيانات.
- يجب على عالم البيانات تصحيح وتجهيز البيانات التي تم جمعها في مرحلة التنقيب للتأكد من مناسبتها لمرحلة التحليل اللاحقة، وعند دمج البيانات من مصادر متعددة تزيد احتمالية تكرار البيانات أو تداخلها، الأمر الذي يتطلب عملية تصحيح وتصويب لتلك البيانات.
- كذلك هو الحال إذا وُجدت بيانات تالفة أو منسقة بشكل غير صحيح أو مكررة أو خاطئة أو حتى غير مكتملة.
- تكمن أهمية تصحيح تلك البيانات في أن الرؤى أو الاستنتاجات المستمدة في مرحلة التحليل من تلك البيانات ستكون خاطئة وسيصعب للغاية استنتاج ما إذا كانت المشكلة ناشئة من أخطاء في خطوات التحليل أو أن البيانات نفسها لم يتم تصحيحها، ولهذا السبب فإن عملية تنظيف البيانات والتحقق من صحتها جيدًا قبل تحليلها تُعدُّ أمرًا مهمًا للغاية للعملية بأكملها.

٤- التحليل الاستكشافي للبيانات Exploratory Data Analysis

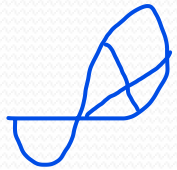
- هو نهج لتحليل مجموعات البيانات وتلخيص خصائصها الرئيسية، ويتم عادة باستخدام الأساليب المرئية.
- بعد أن جمعت البيانات وقمت بتصحيحها، يمكنك تحليل مجموعة البيانات واستنباط الإجابات المطلوبة لأسئلتك، ويتم إجراء تحليل البيانات باستخدام أدوات تحليل البيانات او الأكواد والمكتبات البرمجية المتخصصة، وقد يكون التحليل بسيطًا وذلك بدراسة متغيرواحد أو أكثر، وقد يتسع ليشمل عمليات أكثر تعقيدًا تتضمن عمليات إحصائية متقدمة.
- يُعدُّ تعلّم الآلة من أكثر الطرق شيوعا في الوقت الحالي لتحليل مجموعة البيانات، ويجب اتِّباع خطوات محددة لتحليل البيانات باستخدام تعلّم الآلة، ففي البداية يجب تحديد نموذج تعلّم الآلة بإيجاد قيم المدخلات والمخرجات يليها بناء خوارزمية التحليل نفسها.
- تعتبر هذه العملية معقدة، ولهذا فإن هناك متخصصين للقيام بها مثل علماء البيانات ومهندسي تعلّم الآلة. بعد الانتهاء من الخوارزمية، يتم بناء النموذج واختباره، وعند اكتمال هاتين المرحلتين يمكنك استخدام البيانات الناتجة منه للوصول للإجابات المرجوه الحصول عليها من عمليات التحليل.

٥- التمثيل الرسومي للبيانات Data Visualization

- يسلط التمثيل الرسومي للمعلومات الضوء على أنماط و اتجاهات البيانات، ويساعد القارئ على تطويررؤى وتوصيات بناءً على تلك البيانات.
- يتم تقديم البيانات التي يتم تحليلها عادة بصوره جداول بيانات، مما يتيح لمحللي البيانات ذوي الخبرة استخدامها، ويقدم التمثيل المرئي لتحليل البيانات إمكانية استخلاص رؤى وتوصيات ذات جودة أفضل، بينما توفر الرسوم البيانية والمخططات وحتى الخرائط، وكذلك التقارير المنسّقة طريقة فعالة لرؤية وفهم أنماط البيانات و اتجاهاتها أي ما توحي به تلك البيانات.
- يُعدُّ تمثيل النتائج أمرًا ضروريًا لاتخاذ قرارات مُستندة إلى البيانات عند التعامل مع كميات هائلة من المعلومات.

المجالات والمهارات الأساسية للدراسة في علم البيانات

- علم البيانات هو مصطلح واسع يتطلب الكفاءة في مختلف المجالات لإتقانها.
- بعض المجالات والجوانب الرئيسية اللازمة لإتقان علم البيانات:



❖ تعلم الآلة

بالنسبة لعالم البيانات، يعد تعلم الآلة مهارة أساسية. تتمثل الفكرة الأساسية للتعلم الآلي في السماح للآلات بالتعلم بشكل مستقل باستخدام كتلة البيانات التي يتم تغذيتها بالجهاز كمدخلات. مع تقدم التكنولوجيا، يتم تدريب الآلات على التصرف مثل البشر في القدرة على اتخاذ القرار.

❖ التعلم العميق

غالبًا ما يستخدم التعلم العميق في علم البيانات. لأنها تعمل بشكل أفضل بكثير من طرق التعلم الآلي التقليدية. حيث يستخرج التعلم العميق الميزات تلقائيًا من بنية البيانات.

❖ الرياضيات

لتحسين مهارات التعلم الآلي، يجب أن يكون لدى عالم البيانات معرفة عميقة بالرياضيات. موضوعان مهمان في الرياضيات من حيث التطبيق في علم البيانات هما الجبر والحساب. فإن الحسابات مطلوبة في مجالات مختلفة من التعلم الآلي، مثل تقنيات التحسين.

❖ الاحصاء والاحتمالية

العالم هو عالم احتمالي، لذلك نحن نعمل مع البيانات الاحتمالية؛ هذا يعني أنه وفقًا لمجموعة محددة من المتطلبات الأساسية، ستظهر لك البيانات جزءًا من الوقت فقط. لاستخدام علم البيانات بشكل صحيح، يجب أن تكون على دراية بالاحتمالات والإحصاءات. الإحصاء والاحتمالات من المتطلبات الأساسية في علم البيانات والمعرفة الجيدة في هذا المجال ضرورية.

❖ الخوارزميات

- نظرًا لأن جميع أنظمة التعلم الآلي تعتمد على الخوارزميات، فمن الضروري جدًا أن يكون لدى عالم البيانات فهم أساسي للخوارزميات وكيفية تصميمها.

❖ معالجة اللغة الطبيعية

في مجال علم البيانات، تعد معالجة اللغة الطبيعية مكونًا مهمًا للغاية مع تطبيقات واسعة في مختلف قطاعات الصناعة والشركات. من السهل على البشر فهم اللغة، ومع ذلك، فإن الآلات غير قادرة على التعرف عليها بشكل كافٍ. معالجة اللغة الطبيعية هي فرع من فروع الذكاء الاصطناعي يركز على سد الفجوة بين التواصل بين الإنسان والآلة لتمكين الآلة من التفسير والفهم.

❖ العرض المرئي للبيانات

يعد تصوير البيانات أحد أهم فروع علم البيانات. ببساطة، يتضمن الرسم التوضيحي عرض البيانات في شكل رسوم بيانية ومخططات بيانية.

❖ لغة البرمجة

يجب أن يتمتع عالم البيانات، بالإضافة إلى مهارات الكمبيوتر الأساسية مثل مهارات البرمجة حتى يتمكن من استخدامها للعمل مع البيانات والتمثيل المرئي واستخدام التعلم الآلي ومهارات التعلم العميق في تنفيذ المشروع.

تطبيقات علم البيانات

الآن بعد أن عرفت أهمية علم البيانات والمتطلبات الأساسية والمهارات اللازمة له، من المهم أن تعرف كيف يمكن استخدام علم البيانات في العالم الحقيقي، وسنرى كيف غيّرعلم البيانات العالم اليوم. لذلك، إليك قائمة بتطبيقات علوم البيانات لمعرفة المزيد عن تطبيقاتها:

- المواصلات

أهم تقدم أوتطويرحققه علم البيانات في مجال النقل هو إدخال السيارات ذاتية القيادة.

- لقد أسس علم البيانات موطئ قدم قوي في صناعة النقل من خلال التحليل المكثف لأنماط استهلاك الوقود.
- المر اقبة النشطة للمركبة وسلوك السائق.
- توفيربيئات قيادة أكثرأمانًا للسائقين.
- تحسين أداء السيارة، وإضافة الاستقلالية إلى السيارات.

باستخدام التعلم المعزز والاستقلالية، يمكن لشركات صناعة السيارات بناء سيارات أكثر ذكاءً وطرقًا منطقية أفضل.

- **كشف المخاطر والاحتيال**

تم استخدام علم البيانات لأول مرة في التمويل والمصارف. كانت العديد من المؤسسات المالية مثقلة بالديون في نهاية كل عام. لذلك، تم اعتبار اساليب علم البيانات كحل. لتحليل احتمالية المخاطر، تعلموا فصل البيانات بناءً على مواصفات العميل والتكاليف السابقة والمتغيرات الضرورية الأخرى. وبالتالي، يمكنهم القيام بالتسويق المستهدف بناءً على إيرادات كل عميل كل عام.

- **علم الوراثة والجينات الوراثية**

يساعد علم البيانات علماء الأحياء على تحليل استجابة الجينات للأدوية المختلفة. والغرض منه هو فهم ودراسة تأثير الحمض النووي على صحة الشخص، والذي يسعى إلى إيجاد روابط بيولوجية بين الأمراض والجينات والاستجابات للأدوية.

- **تطوير الادوية**

يتطلب اكتشاف دواء جديد سنوات من البحث والاختبار للوصول إلى مرحلة الإنتاج وفي النهاية يتم ترخيصه للمتاجر الطبية والمستشفيات للمرضى. يمكن استخدام خوارزميات التعلم الآلي وعلوم البيانات لتبسيط العملية وتقليل الوقت اللازم للفحص الأولي لمركبات الأدوية المستخدمة في إنتاج الأدوية. يمكن أن تتنبأ الخوارزميات وعلوم البيانات أيضًا بكيفية استجابة الجسم لمركبات دوائية معينة باستخدام نماذج ومحاكاة إحصائية ورياضية مختلفة. هذا أسرع بكثير من الاختبارات المختبرية التقليدية. يمكن للنماذج أيضًا توقع النتائج المستقبلية بشكل أكثر دقة.

التعامل مع البيانات

ما المقصود بالبيانات الضخمة؟ What is Big Data?

مجموعة بيانات كبيرة تتطلب تقنيات قابلة للتوسع لتخزينها ومعالجتها وإدارتها وتحليلها وذلك نظرًا لخصائص حجمها، وتنوعها وسرعتها وتباينها وبالطبع قيمتها.

يشير مصطلح البيانات الضخمة إلى البيانات الكبيرة جدًا أو المعقدة التي لا يمكن معالجتها بالطرق التقليدية، ونظرًا لأن كم هذه البيانات يُعدُّ كبيرًا جدًا لتتمّ معالجتها باستخدام أنظمة الحوسبة التقليدية، فإن تخزين مجموعاتها ومعالجتها يعتبر تحديًا كبيرًا، وكذلك قد تتطلب السرعة الهائلة لعملية جمع البيانات متطلبات تخزين عالية للغاية.

البيانات هي أساس علم البيانات: البيانات هي المكونات الرئيسية التي تستند إليها جميع التحليلات. في مجال علم البيانات، يمكن تقسيم هذه البيانات إلى مجموعتين: البيانات التقليدية والبيانات الضخمة.

تشير **البيانات التقليدية** إلى البيانات المخزنة في قواعد البيانات التي يمكن للمحللين إدارتها على جهاز الحاسوب. هذه البيانات في شكل جدول يحتوي على قيم عددية أو نصية. بالطبع، مصطلح "تقليدي" هو ما نستخدمه في أغلب الأحيان للتمييز بشكل أفضل بين البيانات الضخمة و أنواع البيانات الأخرى. البيانات الضخمة، من ناحية أخرى، هي بيانات أكبر من البيانات التقليدية وعادة ما يتم توزيعها عبر شبكة واسعة من أجهزة الحاسوب.

- **تعريف البيانات الضخمة**: تشير البيانات الضخمة إلى مجموعة بيانات هيكلية معقدة وغير منظمة وذات حجم كبير يتم إنشاؤها بسرعة من مجموعة متنوعة من المصادر، مما يساعد في اتخاذ القرار.

تشير البيانات الضخمة إلى مجموعة كبيرة من البيانات غير المتجانسة التي يتم الحصول عليها من مجموعة متنوعة من المصادروتتضمن أنواعًا مختلفة من البيانات على النحو التالي:

- **البيانات غير المهيكلة:** الشبكات الاجتماعية، ورسائل البريد الإلكتروني، والمدونات، والتغريدات، والصورالرقمية، وبيانات الجوال، وصفحات الويب، إلخ.
- **شبه المنظمة:** ملفات XML، ملفات نصية، إلخ.
- **البيانات المهيكلة:** قواعد البيانات والتنسيقات المهيكلة الأخرى.

البيانات الضخمة هي في الأساس تطبيق خاص لعلم البيانات حيث تكون مجموعة البيانات كبيرة جدًا وتحتاج إلى التغلب على التحديات المنطقية لمواجهتها.

علم البيانات هو نهج علمي يطبق الأفكار الخوارزمية والحاسوبية لمعالجة هذه البيانات الضخمة.

خصائص البيانات الضخمة

Characteristics of Big Data

هناك خمسة معايير أساسية تساعدنا في تصنيف أي بيانات تحت مصطلح "البيانات الضخمة" وهي: التنوع، والقيمة، والحجم، والموثوقية، والسرعة. وتعتبر البيانات "ضخمة" عندما تأتي بأحجام كبيرة، وبمعدل سريع جدًا، وبتنوع كبير، وبدقة عالية، وفائدة. ويجب أن تستوفي البيانات جميع هذه المعايير لكي يتم اعتبارها "بيانات ضخمة".

- **التنوع Variety**

يسير التنوع إلى العديد من أنواع البيانات المتو افرة ويتم هيكلة البيانات التقليدية المختلفة وتكييفها بدقة في قواعد البيانات العلائقية، ولكن مع ظهور البيانات الضخمة، أصبحت البيانات تتو افر في أنواع جديدة غير منظّمة. تتطلب أنواع البيانات غير المنظمة وشبه المنظمة) مثل النصوص والصوت والفيديو (معالجة اضافية مسبقة لاستخلاص المعاني ودعم معلومات البيانات الوصفية المتعلقة بتلك البيانات، وبدون هذه البيانات الوصفية يكون من المستحيل معرفة ما يتم تخزينه وكيف يمكن معالجته.

- **القيمةValue**

إن جمع الكثير من البيانات لا يعني أن تلك البيانات هي ذات قيمة، فقيمة البيانات تتمثل في امكانية الحصول على التوصيات والوصول إلى بعص الأفكار من خلالها .

يشير مصطلح القيمة الى مدى فائدة البيانات في اتخاذ القرارات، وبالطبع فإن إجراء التحليلات المناسبة هو وسيلة استخراج قيمة البيانات الضخمة.

- **الحجمVolume**

✓ نظرًا لأنه يجب معالجة كميات كبيرة من البيانات غير المنظمة، فإن كم البيانات يعد جانبًا مهمًا في البيانات الضخمة.

✓ يمكن أن تكون قيمة بعض هذه البيانات غير معروفة قبل القيام بتحليلها، مثل بيانات تصفح المستخدمين لأحد مو اقع الويب أو أحد تطبيقات الهاتف الذكي، أو تلك البيانات التي يتم الحصول عليها من أجهزة إنترنت الأشياء المدعمة بأجهزة الاستشعار.

✓ قد يصل حجم هذه البيانات إلى العشرات، بل المئات من التير ابايت من البيانات.

- **الموثوقيةVeracity**

✓ ترتبط صحة البيانات بمدى دقة مجموعة البيانات أو موثوقيتها.

✓ لا ترتبط الموثوقية بجودة البيانات نفسها فحسب، بل أيضا بمدى مصداقية مصدر البيانات ونوعها وكيفية معالجتها.

- **السرعةVelocity**

✓ يشير مصطلح السرعة إلى معدل التقاط البيانات وتخزينها.

✓ تنتج البيانات من معظم الأجهزة الذكية المتصلة بالإنترنت أجهزة إنترنت الأشياء والأجهزة المحمولة في الوقت الحقيقي أو قريبًا من الوقت الحقيقي، مما يتطلب الجمع الفوري لتلك البيانات وكذلك نقلها وتخزينها.

الفرق بين علم البيانات والبيانات الضخمة

الاختلافات بين علم البيانات والبيانات الضخمة:

- تحتاج المؤسسات إلى البيانات الضخمة لتحسين أدائها وزيادة نمو أعمالها وتقديم منتجات أفضل لعملائها. بينما يوفر علم البيانات أساليب وآليات لفهم واستغلال إمكانات البيانات الضخمة في الوقت المناسب.
- من الواضح أن علم البيانات يستخدم مناهج نظرية وعملية لاستكشاف معلومات البيانات الضخمة، والتي تلعب دوَرًاً مهمًا في استغلال إمكانات البيانات الضخمة. يمكن اعتبار البيانات الضخمة على أنها مجموعة من البيانات غير الصالحة، إلا إذا تم تحليلها بالاستدلال الاستنباطي والاستقرائي.
- يرتبط تحليل البيانات الضخمة بالتنقيب في البيانات. لكن علم البيانات يستخدم خوارزميات التعلم الآلي لتصميم وتطوير النماذج الإحصائية لتوليد المعرفة بكميات كبيرة من البيانات الضخمة.

تقنيات إدارة البيانات الضخمة

Technologies that Enable the Management of Big Data

- تستخدم الشركات أنظمة الحاسب وقواعد البيانات للاحتفاظ بالسجلات المختلفة مثل المعاملات المتعلقة بمعالجة الطلبات والمدفوعات وتتبع العملاء وإدارة التكلفة في الشركات تحتاج الشركات أيضًا إلى نظام لإعداد التقارير لتوفير المعلومات التي تساعدها على العمل بكفاءة والمساعدة المدراء التنفيذيين على اتخاذ القرارات المدروسة التي تضمن أداء أفضل للأعمال.
- يحتاج مديرو المتجر الإلكتروني إلى تحسين تجربة الشراء والتأكد من أن زوار الموقع الذين يتصفحون المنتجات سيصبحون زبائن للمتجر وذلك من خلال شراء المنتجات وكذلك العمل على عودة الزبائن للشراء مرات أخرى في المستقبل من خلال الموقع.
- ينتج عن هذه التفاصيل الدقيقة التي يتم جمعها كم هائل من البيانات التي يجب تحليلها لتقديم رؤية واضحة وقيمة للقائمين على أعمال الشركة. يتم استخدام نتائج تحليل تلك المعلومات لإحداث تغييرات في مخطط موقع الويب أو المتجر، ولتعديل أسعار المنتجات سواء بالزيادة أو بالخصم، ولتنظيم الحملات التسويقية للمنتجات على وسائل التواصل الاجتماعي للتأثير على سلوكيات الشراء لدى الزبائن.
- يتطلب القيام بهذا الأمر من الشركات توفير تقنيات وأدوات جديدة لإدارة وتحليل البيانات الضخمة لاستخراج قيمة الأعمال، ويجب جمع البيانات المطلوبة من المصادر الداخلية كدوائر المبيعات والتصنيع والمحاسبة، وكذلك من المصادر الخارجية كالبيانات الإحصائية عن النمو السكاني وطبيعة الزبائن وأعمارهم، وكذلك البيانات المتعلقة بالشركات المنافسة مثلا، وذلك لاستخراج معلومات موجزة وموثوقة حول الوضع الحالي والمستقبلي للشركة والتأثيرات المحتملة لمتغيرات السوق.

تقنيات إدارة البيانات الضخمة

مستودعات البيانات Data Warehouse

- ✓ قد تعتبر مستودعات البيانات الأداة الأقدم لتحليل بيانات الشركات.
- ✓ يسير مستودع البيانات إلى قاعدة البيانات التي تخزن البيانات الحالية والتاريخية التي نتجت عن العديد من أنظمة المعاملات التشغيلية الأساسية مثل أنظمة المبيعات، ودعم العملاء، والتصنيع، والتي تجعل البيانات متاحة لصانعي القرار في الشركة
- ✓ ويتم دمج هذه البيانات مع البيانات من المصادر الخارجية لتحويل البيانات غير المكتملة إلى بيانات منظّمة قبل تخزينها في مستودع البيانات.
- ✓ يوفر نظام مستودع البيانات أيضا مجموعة من الأدوات للتحليل والاستعلام وكذلك أدوات إعداد التقارير الرسومية.

تقنيات إدارة البيانات الضخمة

- **الحوسبة في الذاكرة In-Memory Computing**

هي طريقة لتسهيل عملية تحليل البيانات الضخمة لاعتمادها بصورة أساسية على ذاكرة الحاسب الرئيسة RAM لتخزين البيانات. يصل المستخدمون إلى البيانات المخزنة في الذاكرة الأساسية للنظام وبالتالي يتم تجاوز معوقات استرداد وقراءة البيانات الموجودة في قاعدة البيانات التقليدية المستندة إلى التخزين على الأقراص مما يعني تقليل وقت الاستعلام بشكل كبير. تتميز الخوادم السحابية بشكل خاص بوجود سعة كبيرة من ذاكرة الوصول العشوائي، مما يسهل استخدامها في عمليات الحوسبة في الذاكرة.

- **بحيرة البيانات Data Lake**

بحيرة البيانات هي مستودع بيانات عادةً ما يكون سحابيا يُستخدم لتخزين كميات هائلة من البيانات الأولية وغير المعالجة. في هذه الطريقة يتم استخدام عنوان URL ثابت لدعم كلٍ من البيانات المنظمة) مثل قواعد البيانات (والبيانات غير المنظمة) مثل رسائل البريد الإلكتروني والمستندات) .

التنقيب في البيانات الضخمة Mining Big Data

- ✓ عملية اكتشاف الأنماط في كمية كبيرة من البيانات واستخراج المعلومات المفيدة في توقع السلوك المستقبلي.
- ✓ يتم جمع البيانات الضخمة باستمرار بواسطة أجهزة الاستشعار والتطبيقات العامة والتطبيقات الشخصية.
- ✓ إن عملية جمع البيانات ليست سوى الخطوة الأولى في العملية المشار إليها باسم اكتشاف المعرفة.
- ✓ عملية اكتشاف الأنماط في كمية كبيرة يشير إلى العملية الشاملة للوصول إلى المعرفة المفيدة من البيانات، فالتنقيب عن البيانات هو تطبيق لخوارزميات في توقع السلوك المستقبلي.
- ✓ تحديد العلاقات المختلفة داخل هذه البيانات.
- ✓ تعتبر الخطوات الأخرى في عملية اكتشاف المعرفة مثل تنظيف البيانات، وتكامل البيانات، وتحويل صيغة البيانات، والتفسير الصحيح لنتائج التنقيب ضرورية لضمان اشتقاق المعرفة المفيدة من البيانات.

بعض المهام الرئيسة التي يتم إنجازها عن طريق التنقيب في البيانات:

تحليل البيانات لاكتشاف الأنماط والاتجاهات.
صياغة التنبؤات لمدخلات مجموعات البيانات المختلفة.
تصنيف أو تجميع أو توقع القيم المختلفة لمجموعة البيانات.
تسهيل عملية اتخاذ القرارات المدروسة.

الجدول 1.6: خطوات اكتشاف المعرفة

تصحيح البيانات:	تنظيف البيانات التالفة وغير المطابقة، وإزالة أنواع البيانات الخاطئة وما إلى ذلك.
تكامل البيانات:	يحدث التنقيب في البيانات من مصادر متعددة. يجب دمج مصادر البيانات هذه في مجموعة بيانات واحدة.
اختيار البيانات:	تحديد جزء مجموعة البيانات الذي يجب استخدامه لعملية استخراج البيانات. من المهم تحديد مجموعة البيانات الأكثر مواءمة لأهدافك لأن استخراج البيانات مهمة تستغرق وقتًا طويلًا.
تحويل صيغة البيانات:	يُعدُّ إعداد مجموعات البيانات الأولية وتنسيقها أمرًا ضروريًا لأن عمليات التنقيب عن البيانات تحتاج إلى أن يكون لمدخلاتها تنسيق محدد لتحليلها.
التنقيب في البيانات:	هي العملية الفعلية لتحليل البيانات واستخراج النتائج المرجوة من التحليل من خلال الأنماط.
تقييم النمط:	تقييم الأنماط التي تم إنشاؤها خلال خطوات التنقيب عن البيانات، وتحديد أيها مفيد لكل هدف محدد.
تمثيل المعرفة:	تمثيل النتائج التي تم الحصول عليها من خلال التقارير، والرسوم البيانية والمخططات الواضحة والمختصرة.

- **البيانات الضخمة والتخزين السحابي Big Data and Cloud Storage**

✓ هناك خياران معتمدان لتخزين البيانات الضخمة التخزين السحابي والتخزين الداخلي.

✓ ولقد كان تطوير تطبيقات البيانات الضخمة في الماضي يعتمد أساسا على حفظ البيانات في وسائط التخزين داخليا (على الخوادم داخل الشركات والمؤسسات)، مما تطلب توفر مستودعات بيانات محلية عالية التكلفة، وكذلك تثبيت برامج معقدة لإدارة تلك المستودعات.

✓ ساهمت التطورات الحديثة في علوم الحوسبة والبيانات في استبدال تلك الطريقة بالتخزين السحابي، والذي يعد بمثابة الحل الأمثل لتخزين البيانات الضخمة، وذلك لما يلي:

- تو افر النطاق العريض عالي السرعة على نطاق واسع يسهل حركة البيانات من مكان إلى آخر. ومع وجود بيانات منتجة محليا لم تعد هناك حاجة لتخزين البيانات داخليا، بل أصبح بالإمكان نقلها إلى التخزين السحابي لتحليلها.
- أصبحت غالبية التطبيقات تعتمد على التخزين السحابي، مما يعني أن عملية إنتاج المزيد من البيانات وتخزينها سحابيًا تزداد باستمرار، ولقد ساهم ذلك في قيام أعداد متزايدة من رواد الأعمال بعمل تحليلات جديدة للبيانات الضخمة لمساعدة الشركات على تحليل البيانات السحابية في كثير من المجالات مثل معاملات التجارة الإلكترونية وبيانات أداء تطبيقات الويب.

جدول 1.7: مزايا وعيوب تخزين البيانات الضخمة سحابيًا

المزايا	العيوب
تتطلب الكميات الكبيرة من البيانات المنظمة وغير المنظمة توفر شبكات ذات نطاق ارسال واسع وذلك لسرعة الإرسال والتخزين. يوفر التخزين السحابي بنية تحتية متاحة بسهولة مع القدرة على التوسع للتعامل مع أي مقدار من حركة مرور البيانات ومتطلبات التخزين.	تقدم إمكانيات تحكم مباشر أقل في أمن البيانات، وقد تتعرض لعمليات تؤدي إلى انتهاك البيانات، وبالتالي إلى عواقب خطيرة فيما يتعلق بلوائح خصوصية البيانات.
يؤدي تخزين البيانات الضخمة سحابيًا إلى التخلص من الحاجة إلى الاحتفاظ بأجهزة وبرامج وموظفين متخصصين عند الحاجة، ويُعدُّ نموذج الحوسبة السحابية المبني على الدفع حسب الحاجة إلى الخدمات أكثر فعالية من حيث التكلفة، مما يساهم في خفض التكلفة وزيادة الكفاءة والحد من هدر الموارد.	يمكن لمزود الخدمة السحابية رفع تكلفة الخدمات التي يقدمها في أي وقت، مما يعني ارتفاع التكلفة لأعمال الشركات المستخدمة لهذه الخدمات، والتي لا يمكنها الانتقال بسهولة إلى مقدم خدمات آخر يقدم أسعارًا تنافسية.
تركّز الشركة على عمليات تحليل البيانات بدلًا من إدارة البنية التحتية، مما ينعكس بشكل إيجابي على الأداء والميزة التنافسية.	يعني تخزين البيانات الضخمة سحابيًا أن توفر البيانات يعتمد على الاتصال بالشبكة. تؤثر المشاكل المتعلقة بالشبكات كتدني جودة الاتصال أو **تأخُّر الاستجابة** (latency)، والتي قد تظهر في البيئة السحابية على سرعة جمع البيانات ومعالجتها وتخزينها.

سياسة الشركات وحوكمة البيانات Data Governance and Policies

- ✓ تحدد الضوابط والهياكل التنظيمية للشركات والمؤسسات المسؤوليات وطرق اتخاذ القرارات المتعلقة بإدارة البيانات، والتي تتضمن تطوير السياسات والإجراءات الداخلية التي تتحكم بإدارة البيانات
- ✓ تساعد إدارة البيانات المؤسسات الخاصة أو المؤسسات الحكومية وغير الربحية في التعامل مع عمليات إدارة البيانات بجودة عالية خلال جميع مراحل دورة حياة البيانات، وتؤدي هذه السياسات والإجراءات الفعالة إلى تحسين الأعمال والنتائج، حيث تقوم الشركات والمؤسسات بجمع كميات هائلة من البيانات الداخلية والخارجية، وتعتبر إدارة البيانات ضرورية لاستخدام تلك البيانات بفعالية وإدارة المخاطروخفض التكاليف المختلفة
- ✓ أصبح واجبًا على المؤسسات أن تمتثل للتشريعات الجديدة الخاصة بخصوصية البيانات وحمايتها مثل اللائحة العامة لحماية البيانات في الاتحاد الأوروبي (GDPR) وقانون خصوصية المستهلك في كاليفورنيا (CCPA)، وذلك لأن حوكمة البيانات بصورة سيئة تجر المؤسسات إلى صعوبات وتجعلها تحت طائلة مواجهة العقوبات.

تضمن حوكمة البيانات أن البيانات (آمنة - موثوقة - موثقة – مدارة – مدققة)

معايير حوكمة البيانات Data Governance Standards

قامت منظمة المعايير الدولية ١٥٠ بتطوير معيار ISO/IEC ٣٨٥٠٥ لتطبيق مبادئ حوكمة تقنية المعلومات على متطلبات إدارة البيانات.

- المبادئ الستة لحوكمة البيانات

١. المسئولية: تحدد للأفراد.
٢. الاستراتيجية: تتوافق مع مهمة ورؤية المؤسسة.
٣. الحيازة: تتوافق مع المتطلبات التنظيمية
٤. التوافق: ضمان الامتثال للتشريعات والسياسات الداخلية وأخلاقيات العمل.
٥. الأداء: تلبية متطلبات المؤسسة.
٦. السلوك الإنساني: تشجيع الناس على المشاركة.

Data Governance versus Data Management حوكمة البيانات وإدارتها

- ✓ من الأهمية الإدراك أن حوكمة البيانات هي أحد مكونات إدارة البيانات الشاملة.
- ✓ إن وضع القواعد الإرشادية لحوكمة البيانات دون التنفيذ الفعلي لها يعتبر مضيعة للوقت والجهد دون معنى أو قيمة حقيقية، فحوكمة البيانات تحدد جميع الضوابط والسياسات والعمليات، والتي تنفذ بواسطة إدارة البيانات، والتي مهمتها هي جمع البيانات واستخدامها في صنع القرار من خلال اتباع أساسيات الحوكمة والتي تتمثل بالضوابط والسياسات والعمليات المتعلقة بالبيانات.
- ✓ تشبه حوكمة البيانات عملية تطوير التصميم لبناء منزل جديد، أما إدارة البيانات فهي عملية البناء نفسها.

- **إدارة البيانات: إدارة البيانات هي إنشاء وتنفيذ البنى والسياسات والإجراءات التي تدير احتياجات دورة حياة البيانات الكاملة للمؤسسة.**

Data Governance Challenges تحديات حوكمة البيانات

تعد التحديات المرتبطة بالبيانات السحابية والبيانات الضخمة من الأمور الشائعة التي تواجهها المؤسسات بخصوص حوكمة البيانات، فالخدمات السحابية و أنظمة البيانات الضخمة تستدعي متطلبات حوكمة جديدة. لقد كان تركيز برامج حوكمة البيانات حتى وقت قريب على البيانات المخزنة في مركز البيانات، أما الآن فأصبح من الضروري التعامل مع الكثير من البيانات المنظمة وغير المنظمة وشبه المنظمة التي قد تظهر معا في بيئات البيانات الضخمة، بالإضافة إلى تهديدات الخصوصية المرتبطة بأنظمة البيانات السحابية.

Who is Responsible? من المسؤول عن حوكمة البيانات؟

تضم عملية حوكمة البيانات مجموعة متنوعة من الأشخاص في معظم المؤسسات:

- ✓ المستخدمين النهائيين المطلعين على البيانات ذات العلاقة في أنظمة المؤسسة.
- ✓ مدراء الأعمال
- ✓ المتخصصين في إدارة البيانات
- ✓ موظفي تقنية المعلومات، ويتحمل المسئولية الرئيسة عن الحوكمة عادة رئيس قسم المعلومات كبير مسؤولي البيانات ومدير إدارة البيانات

يُعد رئيس قسم المعلومات أحد كبار المسؤولين التنفيذيين عن برنامج حوكمة البيانات وتشمل مسؤولياته الحصول على المو افقة والتمويل والتوظيف في البرنامج، وكذلك تقديم المبادرات، وتقييم تطور البرنامج، والترويج له بفاعلية.

فاعتمادًا على حجم المؤسسة، يتم تعيين مدير عام لإدارة البيانات ولقيادة وتنسيق مبادرة الحوكمة، يتولى عقد الاجتماعات، وتنفيذ الدورات التدريبية، وتتبع مؤشرات الأداء الرئيسة، وإدارة الاتصالات الداخلية للمبادرة. ويعمل مدير إدارة البيانات مع مالكي البيانات والمسؤولين الذين يضمنون تطبيق ضو ابط وقواعد حوكمة البيانات و اتباع المستخدمين النهائيين لها.