Modulo 1: Definición de la ciencia de datos y qué hacen los científicos de datos

Definición de la ciencia de datos

Definición de la ciencia de datos

Vídeo: ¿Qué es la ciencia de datos?

Ciencia de datos es un intento de trabajare con datos para encontrar respuesta a preguntas que se están explorando.

Se trata más de datos que de ciencia.

Si tienes datos y curiosidad, trabajas con datos, los manipulas, los exploras para obtener respuestas de ellos, eso es ciencia de datos.

La ciencia de datos es relevante hoy por:

- Abundancia de datos
- La existencia de algoritmos
- El software libre.
- El abaratamiento del almacenaje de datos

Fundamentos de la ciencia de datos

Basta disponibilidad de datos de muchas fuentes: log files, e-mails...

Potencia computacional para tratarlos.

Añadiremos conocimiento a la organización.

Primero hay que clarificar la pregunta que la organización quiere que le respondan los datos.

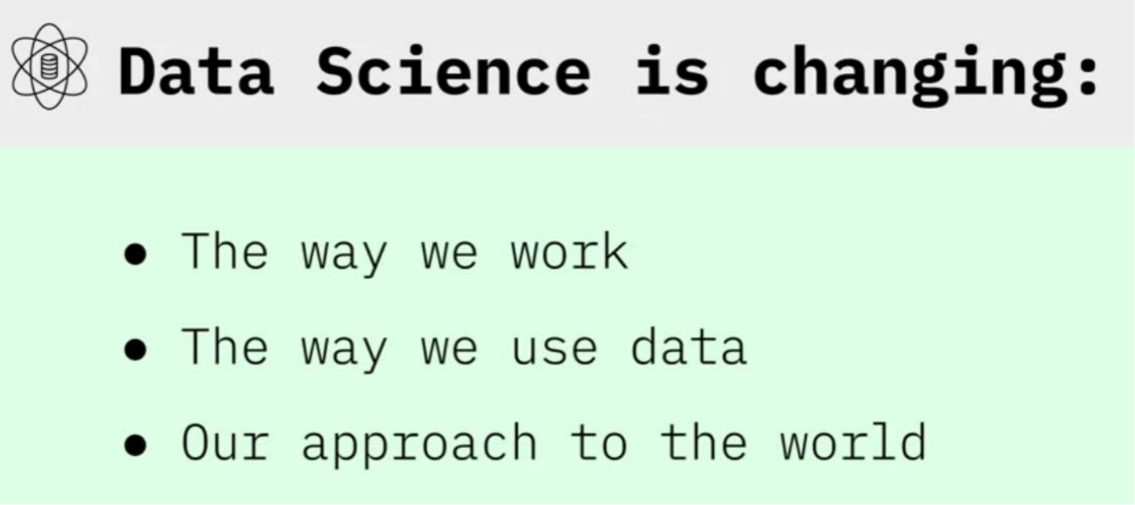
Segundo definir los datos necesarios para resolver el problema.

Definir el origen de lo datos, sean estos estructurado o no estructurados.

Elegir modelos de análisis para encontrar patrones o excentricidades.

Puede confirmar las sospechas o aportar nuevo conocimiento.

Cuando los datos hasn revelado sus entrañas el científico de datos se torna un contador de historias para comunicar los resultados.



Data Science is changing:

- The way we work
- The way we use data
- Our approach to the world

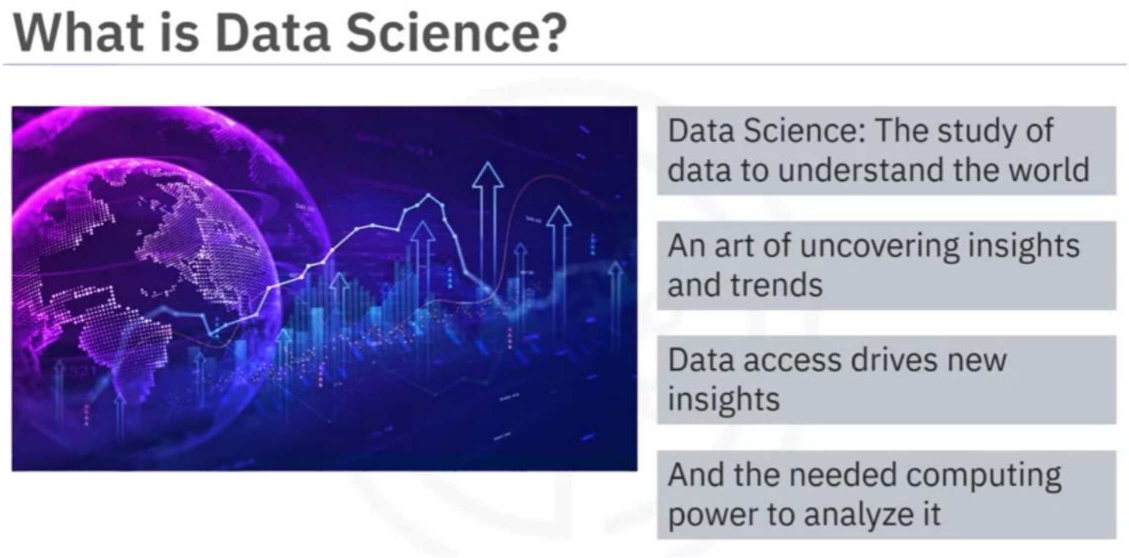
Los muchos caminos hacia la Ciencia de Datos

NO existió hasta 2010 cuando Andrew Gelman or DJ Patil cambiaron el nombre a la estadística

Ciencia de datos: El trabajo más sexy del siglo XXI

In fact, the firms are searching for well-rounded individuals who possess the subject matter expertise, some experience in software programming and analytics, and exceptional communication skills

Definición de la ciencia de datos



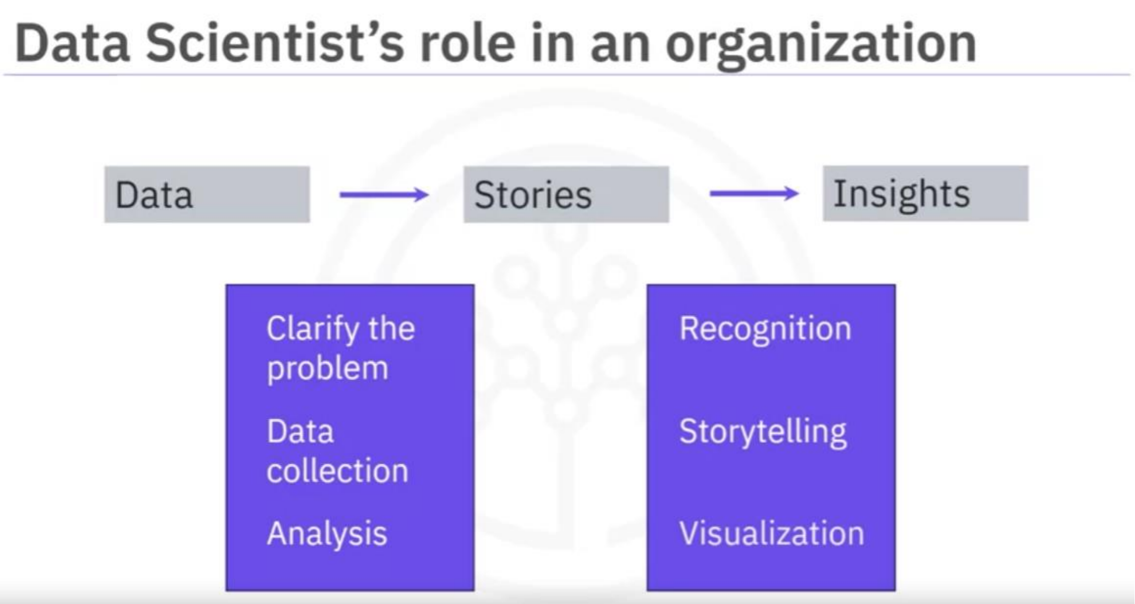
What is Data Science?

Data Science: The study of data to understand the world

An art of uncovering insights and trends

Data access drives new insights

And the needed computing power to analyze it

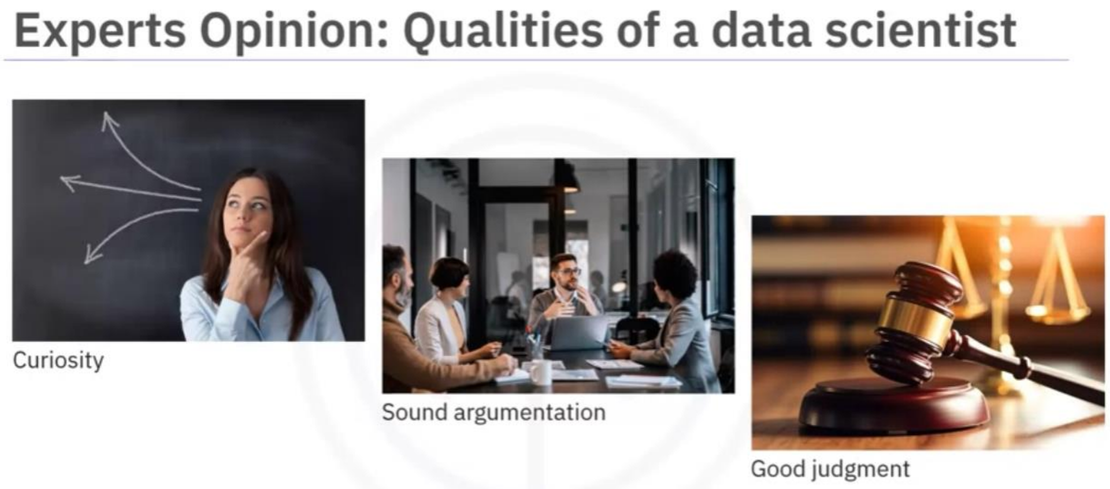


Data Scientist’s role in an organization

Data → Stories → Insights

- Clarify the problem
- Data collection
- Analysis

- Recognition
- Storytelling
- Visualization

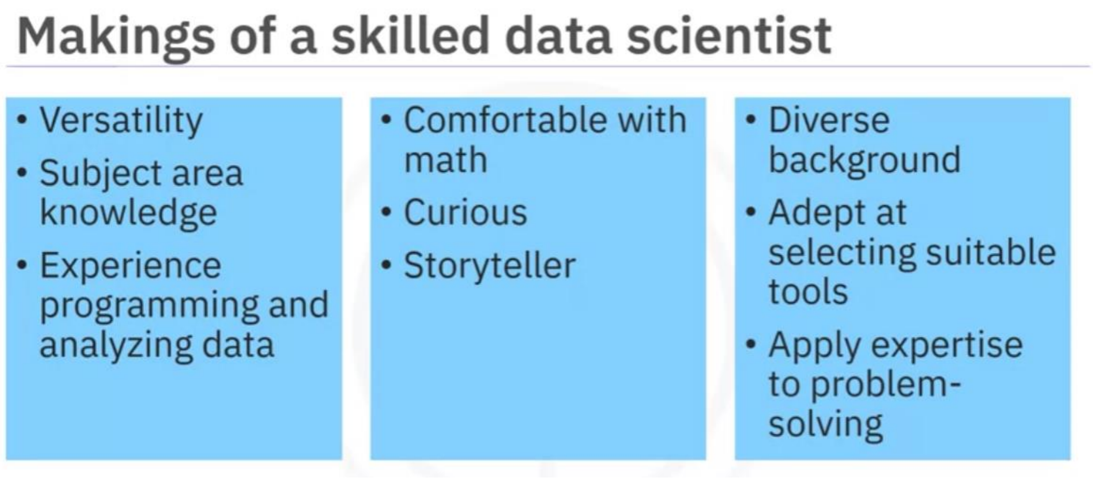


Experts Opinion: Qualities of a data scientist

Curiosity

Sound argumentation

Good judgment

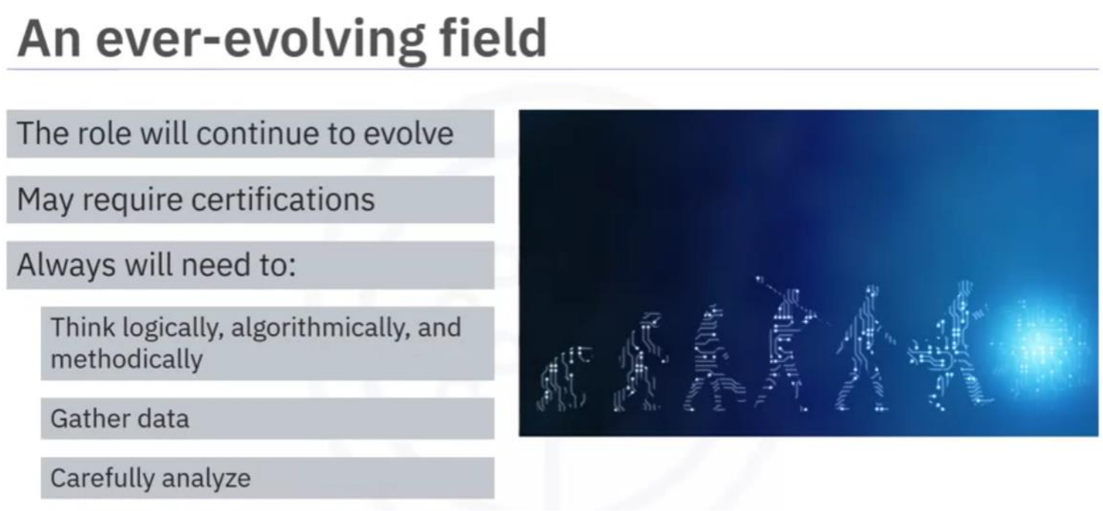


Makings of a skilled data scientist

- Versatility
- Subject area knowledge
- Experience programming and analyzing data

- Comfortable with math
- Curious
- Storyteller

- Diverse background
- Adept at selecting suitable tools
- Apply expertise to problem-solving



An ever-evolving field

The role will continue to evolve

May require certifications

Always will need to:

- Think logically, algorithmically, and methodically
- Gather data
- Carefully analyze

Consejos para los nuevos científicos de datos

Ser curioso. O no sabrás que hacer con los datos.

Ser Argumentative: often arguing or wanting to argue: Para defender una situación. Que pueda cambiar a medida que los datos te enseñan nuevas cosas.

Ser Judgmental: tending to form opinions too quickly: O no sabras por dónde empezar.

Habilidad para contar historias, o los descubrimientos no serán comunicados,

Cómodo y flexible con las plataformas de análisis de datos.

Tener una ventaja competitiva. Sectorial.

Term	**Definition**	**Video where the term is introduced**
Algorithms	A set of step-by-step instructions to solve a problem or complete a task.	What is Data Science?
Model	A representation of the relationships and patterns found in data to make predictions or analyze complex systems retaining essential elements needed for analysis.	What is Data Science?
Outliers	When a data point or points occur significantly outside of most of the other data in a data set, potentially indicating anomalies, errors, or unique phenomena that could impact statistical analysis or modeling.	What is Data Science?

Term	Definition	Video where the term is introduced
Quantitative analysis	A systematic approach using mathematical and statistical analysis is used to interpret numerical data.	Many Paths to Data Science
Structured data	Data is organized and formatted into a predictable schema, usually related tables with rows and columns.	What is Data Science?
Unstructured data	Unorganized data that lacks a predefined data model or organization makes it harder to analyze using traditional methods. This data type often includes text, images, videos, and other content that doesn’t fit neatly into rows and columns like structured data.	What is Data Science?

¿Qué hacen los científicos de datos?

Un día en la vida de un científico de datos

Habilidades de los científicos de datos y Big Data

Trabajar con distintos formatos de archivo

CSV/TSV: Delimiter is a sequence of one or more characters. (Comma, Tab, Colon, Vertical Bar, Space)

XLSX: an XML-based File Format created by MS that uses open file format. Saves all functions available in Excel. It is a secure file format as it cannot save malicious code.

XML: has set rules for encoding data. It is self-descriptive. Does not uses predefined tags like HTML does.

PDF: from Adobe, presents documents independently of SW, HW, and SO.

JSON: JavaScript Object Notation designed for transmitting structured data over the web. También audio y video

Temas y algoritmos de la ciencia de datos

Regression, visualización para los que no entienden muy bien la ciencia de datos. Neuronal networks. Nearest neighbor

Tema de debate: Preséntese

Lectura: ¿Qué hace que alguien sea un científico de datos?

Data scientist is someone who finds solutions to problems by analyzing Big or small data using appropriate tools and then tells stories to communicate her findings to the relevant stakeholders

Data science is what data scientists do.

Curiosity is equally important for data scientists as it is for journalists.

Raquel Schutt defined a data scientist as someone who is a part computer scientist, part software engineer, and part statistician (Miller, 2013). But that's the definition of an average data scientist. *"The best"*, she contended, *"tend to be really curious people, thinkers who ask good questions and are O.K. dealing with unstructured situations and trying to find structure in them."*

GLossario

Modulo 2: Temas de la ciencia de datos

Objetivos de aprendizaje

- Definir los macrodatos y sus características distintivas, como la velocidad, el volumen, la veracidad y el valor
- Describa cómo Hadoop y otras herramientas de big data, combinadas con la potencia informática distribuida, están desencadenando la transformación digital.
- Enumere algunas de las habilidades necesarias para ser un científico de datos y analizar big data.
- Describa las cinco características esenciales de la computación en nube
- Explique qué es la minería de datos.
- Resuma la importancia de establecer los objetivos, la selección de datos, el preprocesamiento, la transformación y el almacenamiento de los datos como preparación para la minería de datos.
- Explique la diferencia entre aprendizaje profundo y aprendizaje automático.

- Describa la regresión y cómo podría utilizarse para predecir el comportamiento del mercado y el análisis de tendencias.
- Describir la IA generativa

Big Data y minería de datos

Cómo impulsan los macrodatos la transformación digital

Digital Transformation affects business operations, updating existing processes and operations and creating new ones to harness the benefits of new technologies.

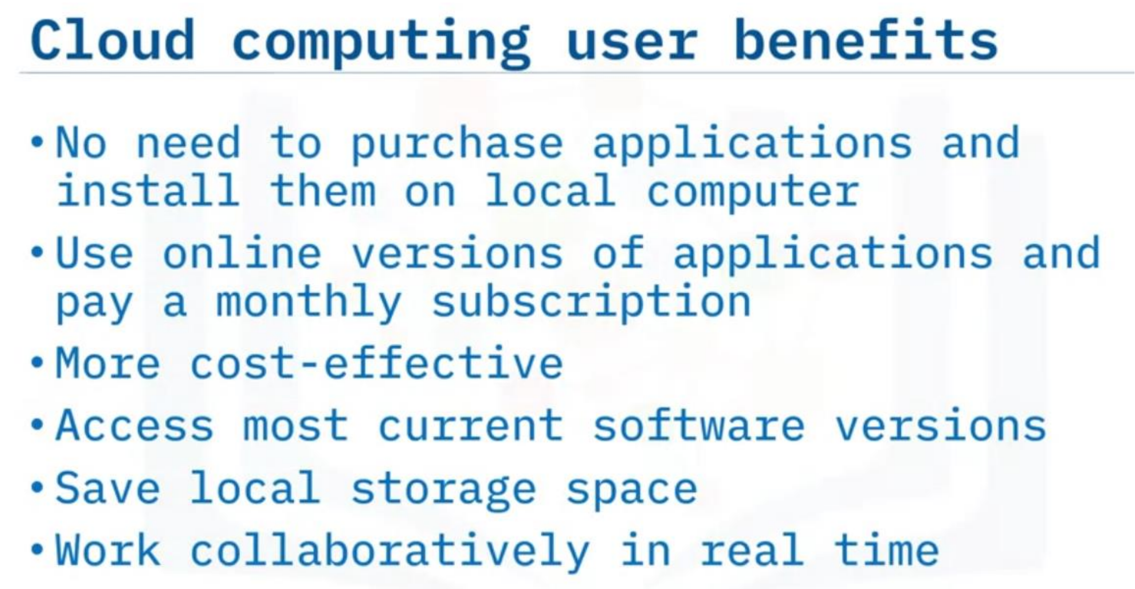
Digital Transformation is an organizational and cultural change driven by Data Science, and especially Big Data. It changes operations, and value delivery.

Digital transformation is not simply duplicating existing processes in digital form; the in-depth analysis of how the business operates helps organizations discover how to improve processes, and operations, and harness the benefits of integrating data science into their workflows.

Most organizations realize that digital transformation will require fundamental changes to their approach towards data, employees, and customers, and it will affect their organizational culture.

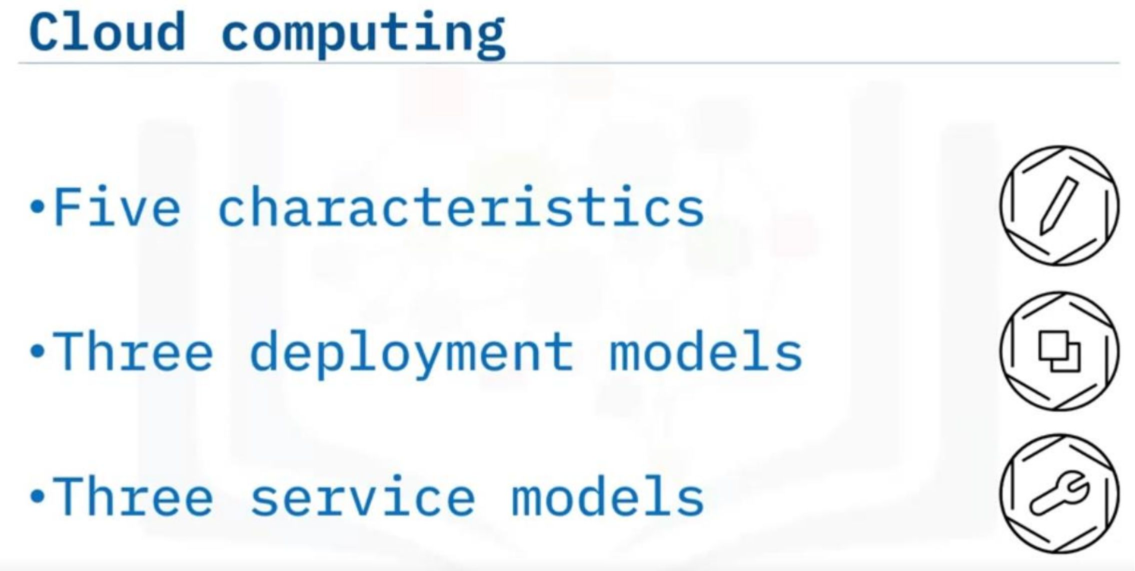
Introducción a la nube

Cloud computing, also referred to as the cloud, is the delivery of on-demand computing resources such as networks, servers, storage, applications, services, and data centers over the Internet on a pay-for-use basis.



Cloud computing user benefits

- No need to purchase applications and install them on local computer
- Use online versions of applications and pay a monthly subscription
- More cost-effective
- Access most current software versions
- Save local storage space
- Work collaboratively in real time



Cloud computing

- Five characteristics
- Three deployment models
- Three service models

Five Characteristics:

1.- On-demand self-service means that you get access to cloud resources you need, using a simple interface, without requiring human interaction with each service provider.

2.- Broad network access means that cloud computing resources can be accessed via the network through standard mechanisms and platforms such as mobile phones, tablets, laptops, and workstations.

3.- Resource pooling is what gives cloud providers **economies of scale,** which they pass on to their customers, making cloud cost-efficient. Using a multitenant **model**, computing resources are pooled to serve multiple consumers, and cloud resources are dynamically assigned and reassigned according to demand, without customers needing to know the physical location of these resources.

4.- Rapid elasticity implies that you can access more resources when you need them, and scale back when you don't, because **resources are elastically provisioned and released**.

5.- Measured service means that you only p**ay for what you use** or reserve as you go. If you're not using resources, you're not paying. Resource usage is monitored, measured, and reported transparently based on consumer utilization.

Cloud computing is using technology "as a service," leveraging remote systems on-demand over the open Internet, scaling up and scaling back, and paying for what you use. Cloud Computing has changed the way the world consumes computer services, by making them more cost-efficient while also making organizations more agile in response to changes in their markets.

Three Deployment models:

Indicate where the infrastructure resides, who owns and manages it, and how cloud resources and services are made available to users.

Public cloud means you leverage cloud services over the open internet on hardware owned by the cloud provider, but its usage is shared by other companies.

Private cloud means that the cloud infrastructure is provisioned for exclusive use by a single organization. It could run on premises, or it could be owned, managed, and operated by a service provider.

Hybrid cloud means you use a mix of both public and private clouds, working together seamlessly.

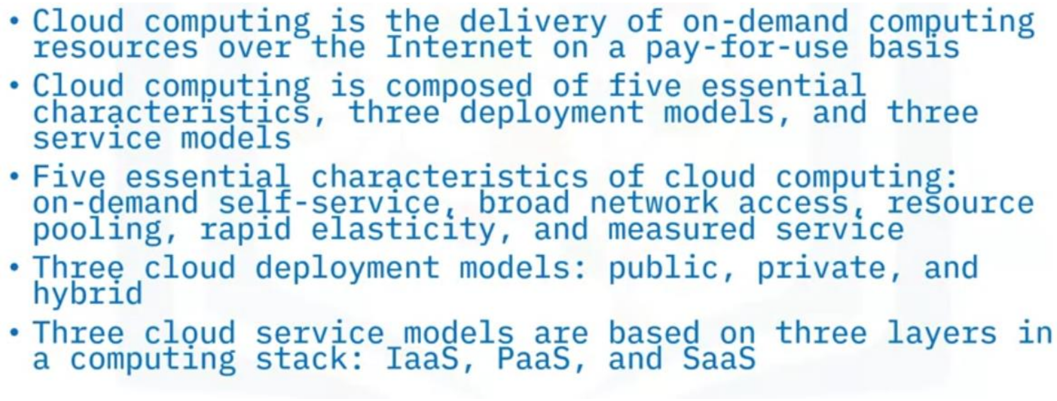
Three Service models

Based on the three layers in a computing stack: infrastructure, platform, and application.

In an **IaaS model,** you can access the infrastructure and physical computing resources such as servers, networking, storage, and data center space without the need to manage or operate them.

In a **PaaS model**, you can access the platform that comprises the **hardware** and **software** tools that are usually **needed to develop and deploy applications** to users over the Internet.

In a **SaaS model** is a software licensing and delivery model in which **software** and applications are centrally hosted and l**icensed on a subscription basis**. It is sometimes referred to as "on-demand software."



- Cloud computing is the delivery of on-demand computing resources over the Internet on a pay-for-use basis
- Cloud computing is composed of five essential characteristics, three deployment models, and three service models
- Five essential characteristics of cloud computing: on-demand self-service, broad network access, resource pooling, rapid elasticity, and measured service
- Three cloud deployment models: public, private, and hybrid
- Three cloud service models are based on three layers in a computing stack: IaaS, PaaS, and SaaS

La nube para la ciencia de datos.

Cloud is a godsend for data scientists. It allows you to:

- Bypass the physical limitations of the computers and the systems you're using.
- Deploy very advanced computing algorithms and the ability to do high-performance computing.
- Multiple entities to work with same data at the same time.
- Get instant access to open source technologies like Apache Spark
- Access to the most up-to-date tools and libraries.
- Cloud Platforms: IBM Cloud, Amazon Web Services (AWS), Google Cloud platform (GCP), Microsoft azure. IBM also provides Skills Network labs or SN labs to learners registered at any of the learning portals on the IBM Developer Skills Network,

Fundamentos del bigdata

Big Data refers to the dynamic, large and disparate volumes of data being created by people, tools, and machines. It requires new, innovative, and scalable technology to collect, host, and analytically process the vast amount of data gathered in order to derive real-time business insights that relate to consumers, risk, profit, performance, productivity management, and enhanced shareholder value.(Ernst and Young)

Velocity is the speed at which data accumulates.

Volume is the scale of the data, or the increase in the amount of data stored.

(Drivers:data sources, sensors resolution, and scalable infrastructure.

Variety is the **diversity** of the data (Structured and unstructured) coming from **different sources**.

Veracity is the quality and origin of data, and its conformity to facts and accuracy. Attributes include **consistency, completeness, integrity**, and ambiguity. Drivers include c**ost and the need for traceability.** With the large amount of data available, the debate rages on about the accuracy of data in the digital age. Is the information real, or is it false?

Value is our ability and need to turn data into value. Value isn't just profit also includes satisfaction.

The scale of the data being collected means that it's not feasible to use conventional data analysis tools. However, alternative tools that leverage distributed computing power can overcome this problem. Tools such as Apache Spark, Hadoop and its ecosystem provide ways to extract, load, analyze, and process the data across distributed compute resources, providing new insights and knowledge.

Data Science and Big Data

Data science and business analytics have become very hot subjects in the last four or five years.

Big data is data that is large enough and has enough volume and velocity that you cannot handle it with traditional database systems.

Big data, was started by Google, when (Larry Page and Sergey Brin) tried to figure out how to solve their page rank algorithm. There was nothing out there. They were trying to store all of the web pages in the world, and there was no technology, so they went out and developed the Hadoop Approach.

¿Qué es Hadoop?

They took the **data, and they sliced** it into pieces and they distributed each and they replicated each piece or triplicated each piece and they would send it the pieces of these files to hundreds of computers and then they would send the same program to all these computers in the cluster.

And each computer would run the program on its little piece of the file and send the results back.

The results would then be sorted, and those results would then be redistributed back to another process.

The first process is called a **map** or a mapper process and the second one was called a **reduce** process.

Fairly simple concepts but turned out that you could do lots and lots of different kinds of handle lots and lots of different kinds of problems and very, very, very large data sets.

So, the one thing that's nice about these **big data clusters is they scale linearly.**

How does Data science differ from traditional subjects like stadistics?

Most of the components of data science (probability, statistics, algebra, linear algebra programming, and databases) **have been around for many decades**. But instead of taking a sample and trying to test some hypothesis, we can take data sets and **look for patterns**. And so back off one level from hypothesis testing to **finding patterns that maybe will generate hypotheses**.

Now this can bother some very traditional statisticians and gets them really annoyed sometimes that you know **you're supposed to have a hypothesis that is independent of the data** and then you test it.

Do you recall a time when no one spoke about data science?

2016.

But it's **morphing and changing and growing**. I mean the last three years **deep learning** has just been added into the mix. **Neural networks** have been around for 20 or 30 years. 20 years ago, I would teach neural networks in a class and you really couldn't do very much with them. And now some researchers have come up with multi-layer neural networks in Toronto in particular the University of Toronto.

Herramientas de procesamiento de Big Data: Hadoop, HDFS, Hive y Spark.

Hadoop is a collection of tools that provides **distributed storage and processing** of big data.

Hive is a **data warehouse** for data query and analysis built on top of Hadoop.

Spark is a **distributed data analytics framework** designed to perform complex data analytics in real-time.

Hadoop

Hadoop, a java-based open-source framework, allows distributed storage and processing of large datasets across clusters of computers.

In Hadoop distributed system, **a node is a single computer**, and a collection of nodes forms a **cluster**.

Hadoop **can scale up** from a single node to any number of nodes, each offering local storage and computation.

Hadoop provides a **reliable**, **scalable**, and **cost-effective** solution for **storing data with no format requirements**.

Using Hadoop, you can: Incorporate emerging data formats, such as streaming audio, video, social media sentiment, and clickstream data, along with structured, semi-structured, and unstructured data not traditionally used in a data warehouse.

Provide **real-time, self-service** access for all stakeholders.

Optimize and streamline costs in your enterprise data warehouse by consolidating data across the organization and **moving "cold" data**, that is, data that is not in frequent use, to a Hadoop-based system.

HDFS

One of the four main components of Hadoop is Hadoop Distributed File System, or HDFS, which is a storage system for big data that runs on **multiple commodity hardware connected through a network**.

HDFS provides scalable and reliable big data storage by partitioning files over multiple nodes. It splits large files across multiple computers, allowing parallel access to them.

Computations can, therefore, **run in parallel on each node** where data is stored.

It also **replicates file blocks** on different nodes to prevent data loss, making it fault-tolerant.

Let's understand this through an example.

Consider a file that includes phone numbers for everyone in the United States; the numbers for people with last name starting with A might be stored on server 1, B on server 2, and so on. With Hadoop, pieces of this phonebook would be stored across the cluster. To reconstruct the entire phonebook, your program would need the blocks from every server in the cluster.

HDFS also **replicates** these smaller pieces **onto two additional servers by default**, ensuring availability when a server fails, In addition to **higher availability**, this offers multiple **benefits**.

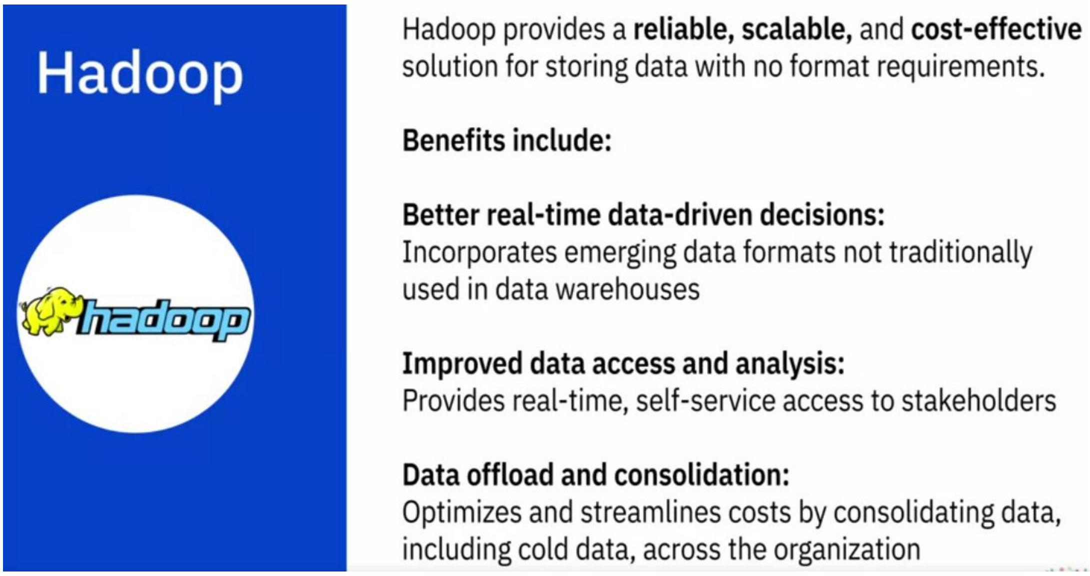
It allows the Hadoop cluster to break up work into smaller chunks and run those jobs on all servers in the cluster for **better scalability.**

Finally, you gain the benefit of **data locality**, which is the process of **moving the computation closer to the node on which the data resides**.

This is critical when working with large data sets because it **minimizes network congestion and increases throughput.**

Some of the other benefits that come from using HDFS include:

- **Fast recovery** from hardware failures, because HDFS is built to detect faults and automatically recover.
- Access to s**treaming data**, because HDFS supports high data throughput rates.
- Accommodation of **large data sets**, because HDFS can scale to hundreds of nodes, or computers, in a single cluster.
- **Portability**, because HDFS is portable across multiple hardware platforms and compatible with a variety of underlying operating systems.

- 

Hadoop

Hadoop provides a **reliable, scalable,** and **cost-effective** solution for storing data with no format requirements.

Benefits include:

Better real-time data-driven decisions:
Incorporates emerging data formats not traditionally used in data warehouses

Improved data access and analysis:
Provides real-time, self-service access to stakeholders

Data offload and consolidation:
Optimizes and streamlines costs by consolidating data, including cold data, across the organization

Hive

Hive is an open-source data warehouse software for reading, writing, and managing large data set files that are stored directly in either HDFS or other data storage systems such as Apache HBase.

Hadoop is intended for long sequential scans and, because Hive is based on Hadoop, queries have very high latency—which means **Hive is less appropriate for applications that need very fast response times**.

Also, **Hive is read-based**, and therefore **not suitable for transaction processing** that typically **involves a high percentage of write operations**. Hive is better suited for data warehousing tasks such as ETL, reporting, and data analysis and includes tools that enable easy access to data via SQL.

Spark

This brings us to Spark, a general-purpose data processing engine designed to extract and process large volumes of data for a **wide range of applications**, including Interactive Analytics, Streams Processing, Machine Learning, Data Integration, and ETL.

It takes advantage of i**n-memory processing to significantly increase the speed of computations** and spilling to disk only when memory is constrained. Spark has i**nterfaces for major programming languages**, including Java, Scala, Python, R, and SQL. It can run using i**ts standalone clustering technology** as well as on top of other infrastructures such as Hadoop.

And it can access data in a large variety of data sources, including HDFS and Hive, making it highly versatile.

The ability to **process streaming data fast and perform complex analytics in real-time is the key use case** for Apache Spark.

Lecturas: Minería de datos

Establishing Data Mining Goals

The first step in data mining requires you to **set up goals for the exercise**. Obviously, you must identify the key questions that need to be answered. However, going beyond identifying the key questions are the concerns about the costs and benefits of the exercise. Furthermore, you must determine, in advance, the **expected level of accuracy** and usefulness of the results obtained from data mining. If money were no object, you could throw as many funds as necessary to get the answers required. However, the cost-benefit trade-off is always instrumental in determining the goals and scope of the data mining exercise. The level of accuracy expected from the results also influences the costs. High levels of accuracy from data mining would cost more and vice versa. Furthermore, beyond a certain level of accuracy, you do not gain much from the exercise, given the diminishing returns. Thus, **the cost-benefit trade-offs for the desired level of accuracy are important considerations for data mining goals**.

Selecting Data

The **output** of a data-mining exercise largely **depends upon the quality of data** being used. At times, data are readily available for further processing. For instance, retailers often possess large databases of customer purchases and demographics. On the other hand, data may not be readily available for data mining. In such cases, you must identify other sources of data or even plan new data collection initiatives, including surveys. **The type of data, its size, and frequency of collection have a direct bearing on the cost of data** mining exercise. Therefore, identifying the right kind of data needed for data mining that could answer the questions at reasonable costs is critical.

Preprocessing Data

Preprocessing data is an important step in data mining. Often raw data are messy, containing erroneous or irrelevant data. In addition, even with relevant data, information is sometimes missing. In the preprocessing stage, you i**dentify the irrelevant attributes of data and expunge such attributes from further consideration**. At the same time, identifying the erroneous aspects of the data set and flagging them as such is necessary. For instance, human error might lead to inadvertent merging or incorrect parsing of information between columns. **Data should be subject to checks to ensure integrity.** Lastly, you must **develop a**

formal method of dealing with missing data and determine whether the data are missing randomly or systematically.

If the data were **missing randomly**, a simple set of solutions would suffice. However, when data are **missing in a systematic way,** you must determine the impact of missing data on the results. For instance, a particular subset of individuals in a large data set may have refused to disclose their income. Findings relying on an individual's income as input would exclude details of those individuals whose income was not reported. This would lead to systematic biases in the analysis. Therefore, you must **consider in advance if observations or variables containing missing data be excluded from the entire analysis or parts of it**.

Transforming Data

After the relevant attributes of data have been retained, the next step is to **determine the appropriate format in which data must be stored**. An important consideration in data mining is to r**educe the number of attributes needed to explain the phenomena**. This may require transforming data Data reduction algorithms, such as Principal Component Analysis (demonstrated and explained later in the chapter), can reduce the number of attributes without a significant loss in information. In addition, **variables may need to be transformed to help explain the phenomenon being studied**. For instance, an individual's income may be recorded in the data set as wage income; income from other sources, such as rental properties; support payments from the government, and the like. Aggregating income from all sources will develop a representative indicator for the individual income.

Often you need to **transform variables from one type to another.** It may be prudent to **transform the continuous variable for income into a categorical variable** where each record in the database is identified as low, medium, and high-income individual. This **could help capture the non-linearities in the underlying behaviors.**

Storing Data

The transformed data must be stored in a format that makes it conducive for data mining. The data must be stored in a format that gives unrestricted and immediate read/write privileges to the data scientist. During data mining, new variables are created, which are written back to the original database, which is why the **data storage scheme should facilitate efficiently reading from and writing to the database**. It is also important to store data on servers or storage media that keeps the data secure and also **prevents the data mining algorithm from**

unnecessarily searching for pieces of data scattered on different servers or storage media. Data safety and privacy should be a prime concern for storing data.

Mining Data

After data is appropriately processed, transformed, and stored, it is subject to data mining. This step covers **data analysis methods, including parametric and non-parametric** methods, and machine-learning algorithms. **A good starting point for data mining is data visualization**. Multidimensional views of the data using the advanced graphing capabilities of data mining software are very helpful in **developing a preliminary understanding of the trends hidden** in the data set.

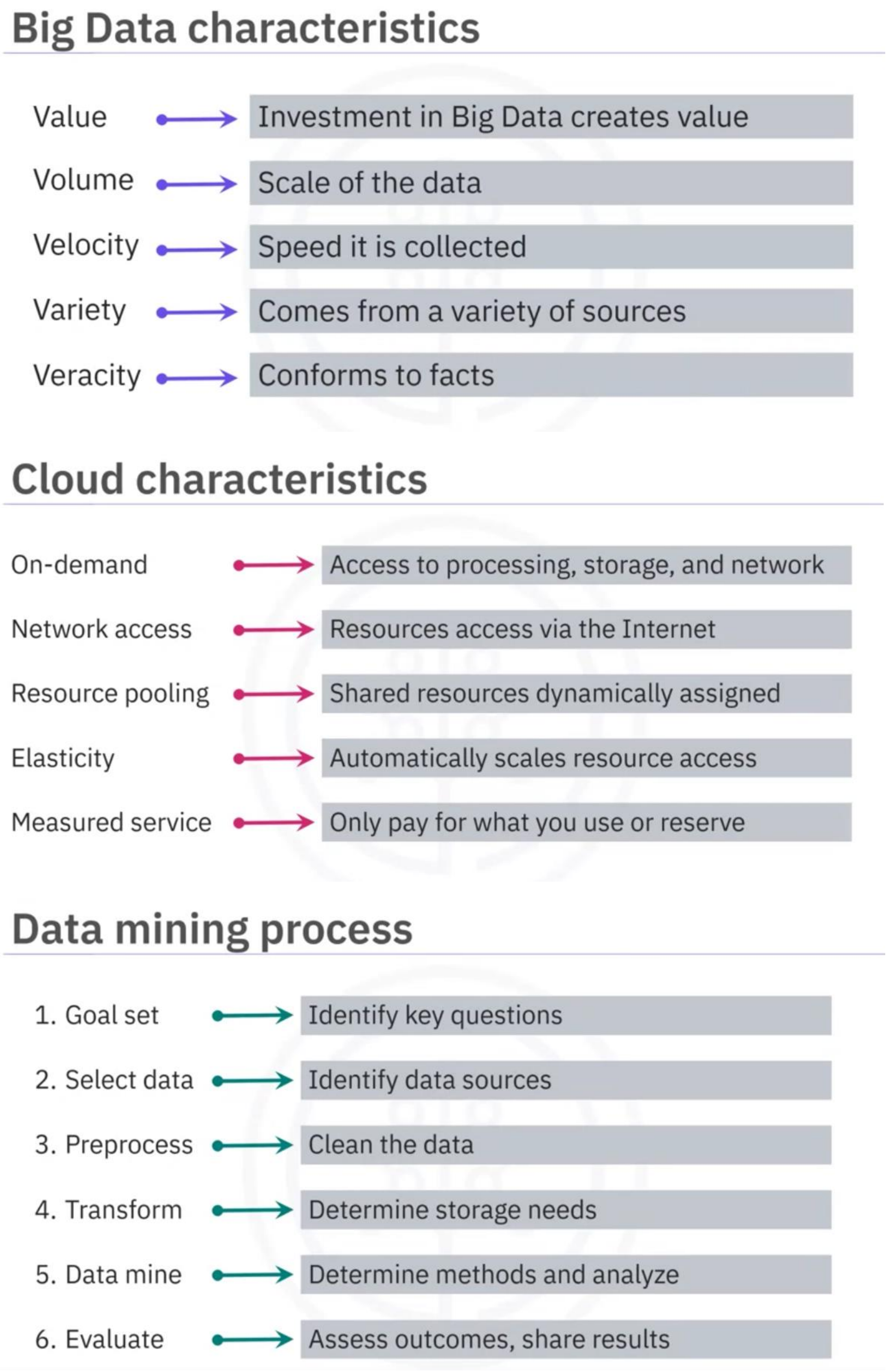
Later sections in this chapter detail data mining algorithms and methods.

Evaluating Mining Results

After results have been extracted from data mining, you do a f**ormal evaluation** of the results. Formal evaluation could include **testing the predictive capabilities of the models** on observed data to see how effective and efficient the algorithms have been in reproducing data. This is known as an "**in-sample forecast"**. In addition, the results are **shared with the key stakeholders for feedback**, which is then incorporated in the later iterations of data mining to improve the process.

Data mining and evaluating the results becomes an iterative process such that the analysts use better and improved algorithms to improve the quality of results generated in light of the feedback received from the key stakeholders.

Summary



Big Data characteristics

Value →	Investment in Big Data creates value
Volume →	Scale of the data
Velocity →	Speed it is collected
Variety →	Comes from a variety of sources
Veracity →	Conforms to facts

Cloud characteristics

On-demand →	Access to processing, storage, and network
Network access →	Resources access via the Internet
Resource pooling →	Shared resources dynamically assigned
Elasticity →	Automatically scales resource access
Measured service →	Only pay for what you use or reserve

Data mining process

1. Goal set →	Identify key questions
2. Select data →	Identify data sources
3. Preprocess →	Clean the data
4. Transform →	Determine storage needs
5. Data mine →	Determine methods and analyze
6. Evaluate →	Assess outcomes, share results

Aprendizaje profundo y aprendizaje automático

Inteligencia Artificial y Ciencia de Datos

In data science, there are many terms that are used interchangeably.

Big data refers to data sets that are so massive, so quickly built, and so varied that they defy traditional analysis methods such as you might perform with a relational database.

The concurrent development of enormous compute power in distributed networks and new tools and techniques for data analysis means that organizations now have the power to analyze these vast data sets.

A new knowledge and insights are becoming available to everyone. Big data is often described in terms of five V's; velocity, volume, variety, veracity, and value.

Data mining is the process of automatically **searching and analyzing data, discovering previously unrevealed patterns**. It involves preprocessing the data to prepare it and transforming it into an appropriate format.

Once this is done, insights and patterns are mined and extracted **using various tools and techniques** ranging from simple data visualization tools to machine learning and statistical models.

Machine learning is a subset of AI that uses computer algorithms to analyze data and **make intelligent decisions based on what it is learned** without being explicitly programmed. Machine learning algorithms are trained with large sets of data and they learn from examples. They do not follow rules-based algorithms.

Machine learning is what enables machines to solve problems on their own and make accurate predictions using the provided data.

Deep learning is a specialized subset of machine learning that uses layered neural networks to simulate human decision-making. Deep learning algorithms can **label and categorize information and identify patterns**. It is what enables AI systems to **continuously learn on the job and improve the quality and accuracy of results by determining whether decisions were correct**.

Artificial neural networks, often referred to simply as neural networks, take inspiration from biological neural networks, although they work quite a bit differently. A neural network in AI is a collection of **small computing units** called neurons that t**ake incoming data and learn to make decisions over time**. Neural networks are often layer-deep and are the reason **deep learning algorithms become more efficient as the data sets increase in volume**, as opposed to other machine learning algorithms that may plateau (to reach a particular level and then stay the same) as data increases.

Now that you have a broad understanding of the differences between some key AI concepts, there is one more differentiation that is important to

understand that between Artificial Intelligence and Data Science.

Data Science is the process and method for **extracting knowledge and insights from** large volumes of disparate **data**. It's an **interdisciplinary** field involving mathematics, statistical analysis, data visualization, machine learning, and more.

It's what makes it possible for us to appropriate information, see patterns, find meaning from large volumes of data and use it to make decisions that drive business.

Data Science **can use many of the AI techniques to derive insight from data**.

For example, it could use machine learning algorithms and even deep learning models to extract meaning and draw inferences from data. **There is some interaction between AI and Data Science, but one is not a subset of the other.** Rather, **Data Science is a broad term** that encompasses the entire data processing methodology while **AI includes everything that allows computers to learn how to solve problems and make intelligent decisions**.

Both AI and Data Science can involve the use of big data.

-
-

Inteligencia artificial generativa y ciencia de datos

Generative AI is a subset of artificial intelligence that **focuses on producing new data (**images, music, language, computer code, and more**)** mimicking creations by people.

How does generative AI operate, though?

Deep learning models like **Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) are at the foundation of this technique**.

These models create new instances that replicate the underlying distribution of the original data by learning patterns from enormous volumes of data.

Generative AI has found diverse applications across various industries:**NLP** revolutionizing content creation and chatbots. Medical **images**, aiding in the training of medical professionals. Stunning **artworks**, generating endless creative visual compositions. Realistic **environments**, characters, and game levels. Generative AI assists in **fashion styles** and creating personalized shopping.

Now, let's discuss how **data scientists use Generative AI**. Training and testing a model takes lots of data. Sometimes, the data you want to study doesn't have enough observations to build a model.

Interest in s**ynthetic data as a tool for analysis and model creation** has increased due to recent developments in generative AI.

Data scientists can **augment their data sets using generative AI** to create synthetic data. It creates this data with **similar** properties as the real data, such as its **distribution**, **clustering**, and many other factors the AI learned about the real data set.

Data scientists can then use the synthetic data along with the real data for model training and testing.

Data scientists, researchers, and analysts frequently find themselves **confined by time** when examining data patterns and insights. Due to this restriction, they can only conceive, develop, and **evaluate a small number of hypotheses**, leaving many other possibilities untested. With generative AI, data scientists can leverage generative AI to generate and test software code for constructing analytical models.

Coding automation has the potential to revolutionize the field of analytics, allowing the data scientist to focus on higher-level tasks such as identifying and clarifying the problem the models intend to solve and evaluating hypotheses from a wider range of data sources.

Generative AI can generate accurate business insights and comprehensive reports, making it possible to update these insights as the data evolves.

Furthermore, it can autonomously explore data to uncover hidden patterns and insights that might go unnoticed during manual analysis and enhance decision-making.

For instance, IBM's Cognos Analytics, which provides AI-powered automation, enables you to unlock the full potential of your data with the help of its **natural language AI assistant**. You describe the question you are looking to answer or the hypothesis you want to test, and it helps generate insights you need.

Redes neuronales y aprendizaje profundo

How does a neural network work?

Computer Sciences attempt to mimic how our real brain 's neurons actually functions.

A neural network would have some inputs that would come in. They would be fed into different processing nodes that would then do some transformation on them and aggregate them or something, and then maybe go to another level of nodes. And finally there would some output would come out.

Problem with neural networks was They were **computationally very intensive** and so they went out of favor.

Deep learning is Neural Networks on steroids.

Uses cases for deep learning

Speech Recognition. Face recognition.

How can one start with neural networks?

Matrix and linear algebra. Packages will do deep learning but **you should have some idea of what is happening underneath**. You have to have some special computational resources.

Lectura Regresión

Sir Frances Galton in 1886 studied landed upon a statistical technique we today know as regression models.

However, the real value added by egression models rested not just found the correlation between housing prices and the size of housing units, but I have also discovered the magnitude of those relationships.

Laboratorio Exploración de datos mediante IBM Cloud Gallery

- https://dataplatform.cloud.ibm.com/gallery

Lesson summary.

Artificial intelligence (AI) has boomed and become accessible to almost everyone. Data scientists use AI regularly when analyzing data.

AI is the branch of computer science that includes the development of systems that can **mimic** many of the **tasks associated with human intelligence**.

Machine learning is a subset of AI that uses computer algorithms to **learn about data and make predictions** with it **without** needing to explicitly **program** the analysis methods into the system.

Deep learning is a subset of machine learning that **uses layered neural networks to simulate human decision-making.**

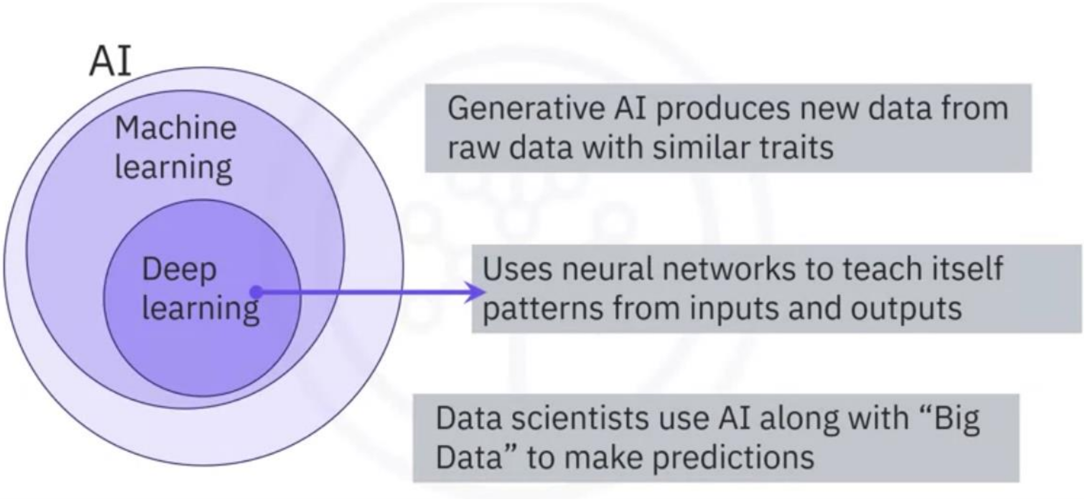
A neural network is a collection of small computing units, called neurons, that take incoming data and learn to make decisions over time, such as the difference between a dog and a cat.

Deep learning algorithms become **more efficient as the amount of data increases in volume,** unlike other machine learning algorithms, which tend to plateau.

Generative AI is a subset of AI, focuses on **producing new data (**images, music, languages, and computer code**)** rather than just analyzing existing data. Generative AI also can make data sets **with similar traits** to a raw data set.

Regression identifies the **strength and amount of the correlation** between one or more inputs and an output.

Using big data, data scientists use all of these areas of AI to make predictions.


AI
Machine learning
Deep learning
Generative AI produces new data from raw data with similar traits
Uses neural networks to teach itself patterns from inputs and outputs
Data scientists use AI along with "Big Data" to make predictions

At this point in the course, you know:

Big Data has five characteristics: velocity, volume, variety, veracity, and value.

The **five cloud computing characteristics** are on-demand self-service, broad network access, resource pooling, rapid elasticity, and measured service.

Data mining has a six-step process: goal setting, selecting data sources, preprocessing, transforming, mining, and evaluation.

The availability of so many disparate amounts of data created by people, tools, and machines requires new, innovative, and scalable technology to drive transformation.

Deep learning utilizes neural networks to teach itself patterns in inputs and outputs.

Machine learning is a subset of AI that uses computer algorithms to learn about data and **make predictions without explicitly programming** the analysis methods into the system.

Regression identifies the **strength and amount of the correlation between one or more inputs and an output**.

Skills involved in processing Big Data include the application of statistics, machine learning models, and some computer programming.

Generative AI, a subset of artificial intelligence, f**ocuses on producing new data rather than just analyzing existing data**. It allows machines to create content, including images, music, language, computer code, and more, mimicking creations by people.

Módulo 3: Aplicaciones y Carreras en Ciencia de Datos

Dominios de aplicación de la Ciencia de Datos

How should companies get started in Data Science?

So **the first thing a company has to do is to start recording information.** Once you have data, then you can a**pply algorithms and analytics to it.** Do not overwrite on your old data thinking you don't need it anymore. **Data never gets old**. **Data is always relevant**, even if it's 100 years old, 200 years old.

Make sure there's a **consistency**. So someone 20 years later trying to understand, that data should be able to do so, so **have proper documentation**. Do it now. Put the best practices for data archiving in place the moment you start a business. And if you're already in business and you haven't done it, do it now.

Data science inside a company is only going to be **as valuable as the data collected**. Garbage in, garbage out is a rule in any sort of analysis.

Look for a team who love to work as a data scientist. it's not one data scientist, but a team of them, that each of them have strengths in different areas of data science.

Old problems, new data science solutions

Uber uses data to put the right number of drivers in the right place, at the right time, for a cost the rider is willing to pay.

Toronto Transportation Commission **By focusing on peak hour clearances and identifying the most congested routes, monthly hours lost for commuters due to traffic congestion dropped from 4.75 hrs. in 2010 to 3 hrs. in mid-2014.**

The information collected will lead to better predictions of when and where cyanobacterial blooms take place, **enabling proactive approaches to protect public health in recreational lakes and in those that supply drinking water.**

Such interdisciplinary training **prepares the next generation of scientists** to address societal issues with the proper modernized data science tools.

You must: Identify the problem and establish a clear understanding of it. Gather the data for analysis. Identify the right tools to use, and develop a data strategy.

Case studies are also helpful in customizing a potential solution.

It will take time for an organization to refine best practices for data strategy using data science, but the benefits are worth it.

Aplicaciones de la Ciencia de Datos

Data science and big data are making an undeniable impact on businesses, **changing day to day operations, financial analytics, and especially interactions with customers**.

It's clear that **businesses can gain enormous value from the insights data science can provide**, but sometimes it's hard to see exactly how.

Recommendation engine.

Personal assistants.

Wearable devices.

In 2011, McKinsey and Company said that data science was going to become the **key basis of competition, supporting new waves of productivity, growth, and innovation.**

In 2013, UPS announced that it was using data from customers, drivers and vehicles in a new route guidance system aimed **to save time, money, and fuel**.

Thanks to data science. Netflix knows what people want before they do.

How data sciecne is saving lives?

Data science systems that use predictive analytics ensure that all physicians can also access **the latest information about the disease, tests, and treatment plans, tailored to their specific patient**.

By providing extra information through data science tools, **physicians can be made aware of the most helpful tests and treatments for a specific patient**.

There are many opportunities to explore other ways to mine data, such as from **electronic medical records** for different types of medical research.

Developing more sophisticated big data analytics capabilities helps **healthcare organizations move from basic descriptive analytics towards predictive insights, thanks to data science**.

Disaster Preparedness: Earthquakes, hurricanes & tornados, floods, and volcanic eruptions can be **predicted with the help of data science**.

When added to the information recorded by scientists and weather stations, this type of data can be used to **improve the predictions for localised weather events**. saving hundreds of lives.

Lectura: La entrega final

The Final Deliverable

The ultimate purpose of analytics is to communicate findings to the concerned who might use these insights to formulate policy or strategy. Analytics summarize findings in tables and plots. **The data scientist should then use the insights to build the narrative to communicate the findings.** In academia 1,000 to 7,000 words in length. In consulting and business, 1,500 words

Before the authors started their analysis, they must have **discussed the scope of the final deliverable**. They would have **deliberated the key message of the report** and then **looked for the data and analytics they needed to make their case.**

The initial **planning and conceptualizing** of the final deliverable **is therefore extremely importan**t for producing a compelling document.

Embarking on analytics, without due consideration to the final deliverable, is likely to result in a poor-quality document where the analytics and narrative would struggle to blend.

Carreras y contratación en la ciencia de datos

How Can Someone Become a Data Scientist?

A real data scientist, **the high-end data scientists, are mostly PhDs**.

They often come out of physics, out of statistics, they have to have a computer science background, they have to have a math background, they have to know about databases and statistics and probability and all that stuff.

However, if you're coming into a data science team, I think the **first skills** you need is you need to know how to program, at least have some **computational thinking**, so having taken a programing course, you need to know some **algebra**, at least up to analytics, **geometry**, and hopefully some **calculus**, some basic **probability**, some basic **statistics**, I mean really have to understand the difference and different statistical distributions, and **database**.

I mean, one of the **easiest places to start is relational databases**, You don't have to really have understand the whole MapReduce programming model.

But then, as you go further up in the field, then you have to know a lot of computer science theory and statistics, it's really, and probability, it's really the intersection of them that the high end data scientists, the PhD data scientists work with

What is the role of self-learning in data science?

I do a lot of self-learning. I'm a builder, I'm a tinkerer, so if I wanna figure out how to do something, I build it.

So I think one of t**he ways you learn things is you do them**, you have to do them, and these online learning platforms especially now that we have things like IPython and Jupyter Notebooks and I guess Zeppelin means that you can actually go in and take some of these courses and you can do things right then and you can see them and feel them and play with them and, at that point, you know, you'll start to get your head around what is actually happening.

Motivation is the key problem in all of these, is how to keep people motivated

The badge system is one of the ways you get people to keep going through.

Where should data Science fit in the org structure?

The place it can't sit is probably under the Chief Information Officer.

It has to come out of the research side. T**he demand out there for the PhD level data scientist is just unbelievable**.

Recruiting for Data science.

tendency would be to find the person who has all the skills, **you will realize you're looking for a unicorn**.

Given the pool of applicants you have, **who has the most resonance with your firm's DNA**. Because you can **teach analytics skills**,

But what really matters is **who's passionate about the kind of business that you do**.

And I would say if I'm looking for someone, if I have to put together a data science team, **I would first look for curiosity.** Is that person curious **about things not just for data science** The second thing is do they have a **sense of humor** because, you see, you have to have a lighthearted about it. If someone is too serious about it, they probably would take it too seriously, and would not be able to look at the lighter elements. The third thing I would look for are technical skills.I would go through the social skills, curiosity, and sense of humor. **The ability to tell a story.** The ability to know that there is a story there.

I find quite often nobody really cares about the r-square or the confidence interval. So you have to **be able to introduce those things and explain something in a compelling way**.

And they also have to find somebody who is **relatable**, because data science, it been typically new means that the person in that role has to **make relationships and they have to work across different departments**.

When a company is hiring anyone to work on a data science team, they need to think about **what role that person is going to take**. Before a company begins, they need to understand what they want out of their data science team.

From a skills point of view, let's focus on the technical skills and in that case, first thing would be **what kind of a technical platform would you like to adopt?** Let's say you want to work in a structured data environment and let's say you want to work in market research. **Then the type of skills you need are slightly different** than someone who would like to work in big data environments

It starts with where you would like to be, in what field, in what domain.In terms of platforms-
In addition to technical skills, the second aspect of the data science is to have the ability to communicate. **The communication skills** or presentation skills. I call them story telling skills

Careers in data science

Companies like LinkedIn, Glassdoor, Indeed, and Dice track employment trends which show a career in **data science moving up the list of most promising jobs** to become number one since 2016. It remains one of the top three career choices for 2020.

Dice noted that **job postings are from companies in a wide variety of industries**, not just tech. Global Industry Analysts Incorporated predicts that the data science platform market will grow by $314.8 billion US by 2025, driven by a compounded growth of 38.2%.

McKinsey Global Institute warned of **huge talent shortages for data and analytics** by 2018. Forrester Research Analyst Brandon Purcell said, in January of 2019, the demand for data scientists will only grow as organizations increasingly rely on data-driven insights.

Importancia de las Matemáticas y la Estadística para la Ciencia de Datos (sólo cambio de nombre)

La estructura del informe

Cover page: the title of the report, names of authors,their affiliations, and contacts, name of the institutional publisher (if any), and the date of publication

Table of contents; Never shy away from including a ToC, especially if your document, excluding cover page, table of contents, and references, is five or more pages in length.

Abstract/Executive summary: explaining the crux of your arguments in three paragraphs

Introduction: setting up the problem for the reader who might be new to the topic.

Literature review: highlight gaps in the existing knowledge, which your analysis will try to fill. This is where you formally introduce your research questions and hypothesis.

Methodology Section: research methods and data sources.

Results; Show empirical findings (Descriptive statistics/graphics, testing hypothesis, Statistical models,

Discussion: rely on the power of narrative to enable numbers to communicate your thesis to your readers.

Conclusion:

References, Acknowledgements, appendices

Summary

Recruiting

Desired skills:

Domain-specific knowledge

Analyzing both structured and unstructured data

Presenting and storytelling

Companies should develop teams with these skills

Excitement results in productivity

Excitement to work with data in their industry

Ask good questions

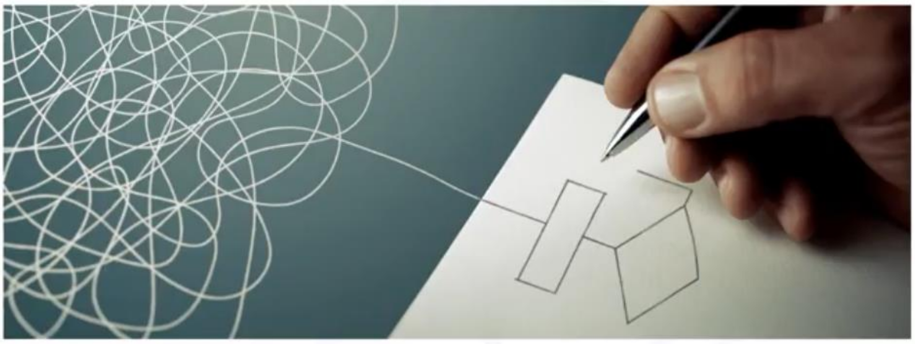
A data scientist in retail might not make a good data scientist in IT

Curious thinker

Curiosity leads to good questions

Curiosity encourages motivation

Logic minded

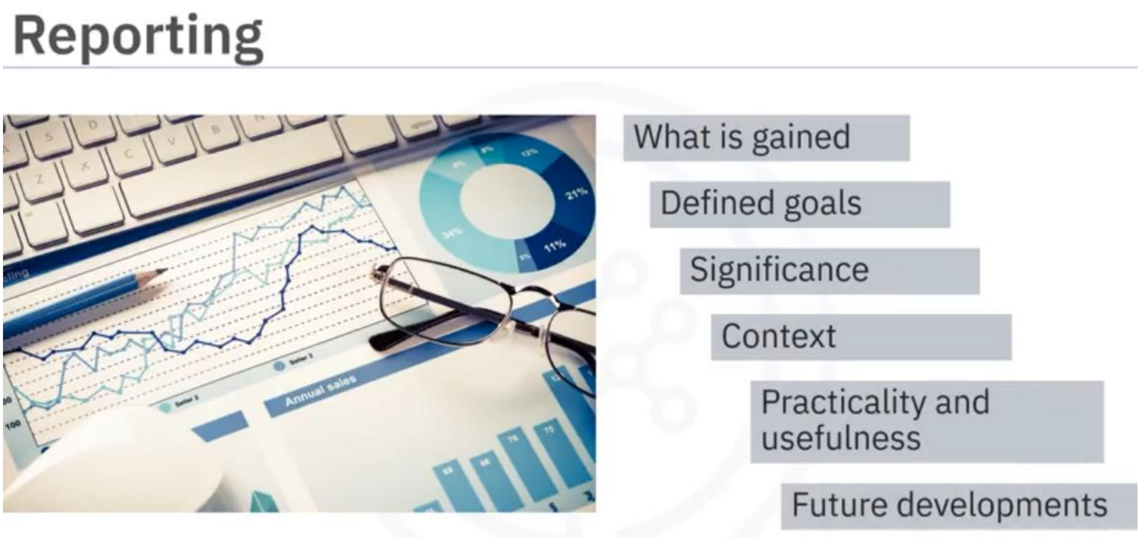


Think analytically and computationally

Background in mathematics, statistics, and probability

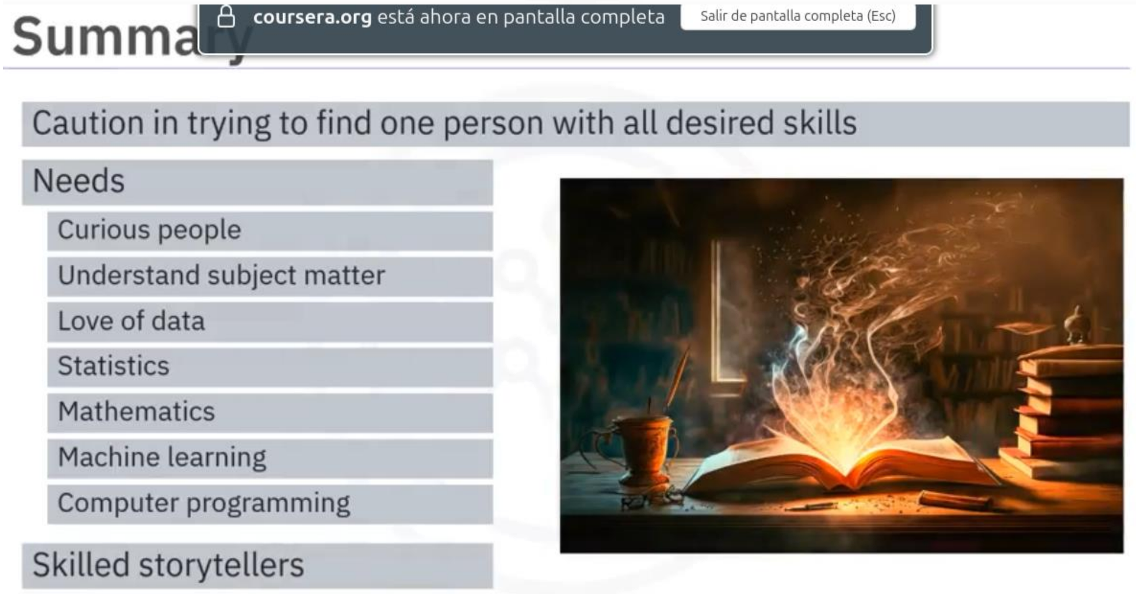
Creating the narrative

- Communicating
- Instructing
- Presenting
- Storytelling
- Synthesizing



Reporting

- What is gained
- Defined goals
- Significance
- Context
- Practicality and usefulness
- Future developments

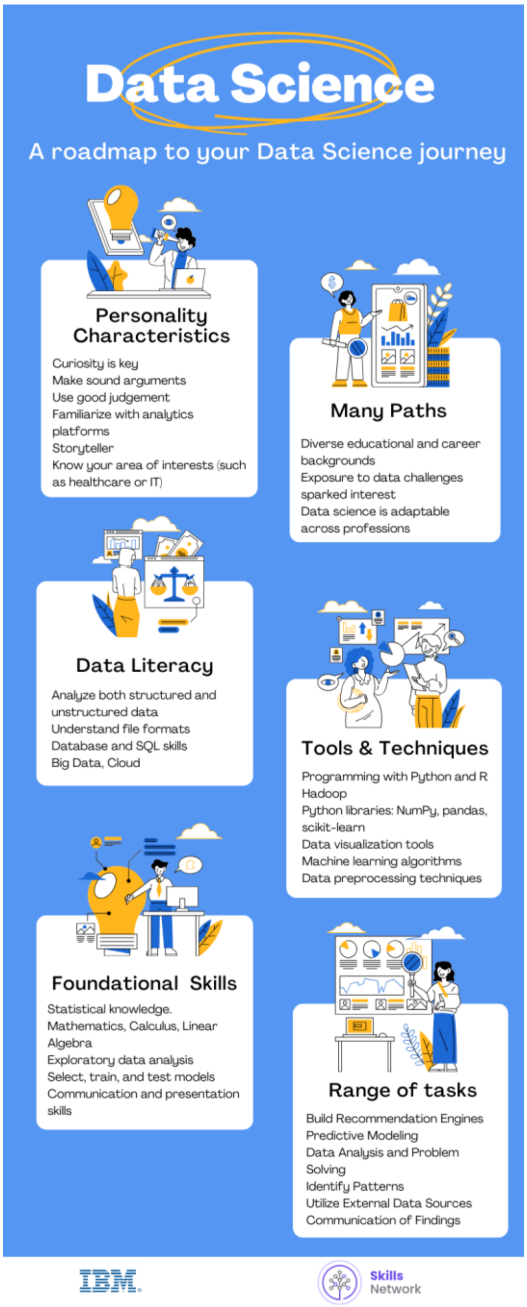


coursera.org está ahora en pantalla completa
Salir de pantalla completa (Esc)

Summary

- Caution in trying to find one person with all desired skills
- Needs
  - Curious people
  - Understand subject matter
  - Love of data
  - Statistics
  - Mathematics
  - Machine learning
  - Computer programming
- Skilled storytellers

Lectura: Infografía sobre la hoja de ruta



Data Science

A roadmap to your Data Science journey

Personality Characteristics

Curiosity is key
Make sound arguments
Use good judgement
Familiarize with analytics platforms
Storyteller
Know your area of interests (such as healthcare or IT)

Many Paths

Diverse educational and career backgrounds
Exposure to data challenges sparked interest
Data science is adaptable across professions

Data Literacy

Analyze both structured and unstructured data
Understand file formats
Database and SQL skills
Big Data, Cloud

Tools & Techniques

Programming with Python and R
Hadoop
Python libraries: NumPy, pandas, scikit-learn
Data visualization tools
Machine learning algorithms
Data preprocessing techniques

Foundational Skills

Statistical knowledge.
Mathematics, Calculus, Linear Algebra
Exploratory data analysis
Select, train, and test models
Communication and presentation skills

Range of tasks

Build Recommendation Engines
Predictive Modeling
Data Analysis and Problem Solving
Identify Patterns
Utilize External Data Sources
Communication of Findings

IBM Skills Network

Case Study: Lila's Journey to Becoming a Data Scientist: Her Working Approach on the First Task

This case study explores the data scientist's career path and key attributes, highlighting the skills, **education, and experiences required to excel** in this dynamic field. We'll follow the story of Lila, a fictional individual who aspires to become a successful data scientist.

There will be a quiz after this reading based on the contents of this case study.

Education and Skill Acquisition

With an economics undergraduate degree and a substantial data analysis background, Lila finds data science and its potential to drive meaningful change captivating. Inspired by her experiences, she makes a determined decision to transition her career and step into the role of a data scientist.

Lila realizes that to embark on her data science journey, she needs to enhance her skills and knowledge. She enrolled in the **IBM Data Science Professional Certificate online program** that covers key topics like statistics, machine learning, data analysis, and programming languages like Python and SQL. She diligently completes coursework and practices her coding skills on real datasets.

Building a Strong Foundation

As she progresses in her studies, Lila gains a deep understanding of data science fundamentals such as data manipulation and visualization with Python libraries like NumPy, Pandas, and Matplotlib. This strong foundation equips her with essential skills for data analysis.

Visualization for Storytelling

Lila knows she must communicate her findings effectively, so she learns which types of data visualizations will be most informative. She learns to create charts and graphs that visually represent data like sales trends, customer segmentation, and product popularity, allowing stakeholders to grasp the data's significance. These visualizations help in storytelling and decision-making.

Hands-On Experience

Lila understands that practical experience is invaluable in data science. She started participating in **Kaggle competitions** and working on personal data projects. These experiences expose her to real-world data problems and help her develop problem-solving skills. Furthermore, she created her **GitHub account** and uploaded her projects to build her profile.

Data Wrangling and Preprocessing

Lila learns that data scientists spend a significant portion of their time on data cleaning and preprocessing. She worked on various datasets, learned data preprocessing as she used sed NumPy and pandas Python libraries, and became skilled in **handling missing data, outlier detection, and feature engineering to improve model performance**.

Communication and Storytelling

Recognizing that data scientists must communicate their findings effectively, Lila honed her data storytelling skills. She learned various tools like matplotlib and plotly while she pursued her IBM Data Science Professional Certificate. She learned how to create compelling visualizations and present her insights in a clear and understandable manner.

Networking and Collaboration

Lila actively participates in **data science communities and attends meetups** and conferences. She collaborates on open-source projects, connects with fellow data scientists, and **gains exposure to various industries** when she attended the IBM TechXchange Conference.

Domain Expertise

Understanding that domain knowledge is crucial, Lila chooses a niche that aligns with her interests. She looks deeply into several domains, including e-commerce, healthcare, finance, and several other fields to which she could apply her data science skills effectively. Since her master's in economics, she chose e-commerce as her core domain to land herself a data science career.

Landing the First Job

After months of preparation, Lila started applying for data scientist positions. She tailors her **resume to highlight her relevant skills and projects**. Her **online portfolio** showcases her capabilities and demonstrates her commitment to the field.

Lila's Approach to Working on Her First Task as a Data Scientist

As a newly hired junior data scientist at a retail company, Lila uses data insights to improve customer service. Her first assignment involves diving into customer data to identify patterns and anomalies that could impact customer service. She uses data analysis to enhance the overall customer experience.

Dataset Selection and Sourcing

In the initial phase of her data science journey, Lila faced the challenge of selecting a suitable dataset and procuring it from different sources. Apart from the historical data available for the organizations for the past four years, she scoured **various repositories, websites, and databases** to find the right datasets for her project. Upon collecting data from diverse sources, Lila encountered another crucial decision point. She had to **decide how to harmonize and integrate these disparate datasets into a cohesive whole**. She reached out to product professionals, data engineers, and domain specialists, seeking their input and expertise in merging datasets.

Data Understanding and Cleaning

Lila begins by importing the dataset into her data analysis environment using Python and SQL. She loads the data and examines the first few rows to understand its structure and contents. Upon acquiring the dataset, Lila encounters her first challenge: data cleaning. Lila checks for missing values, duplicates, and outliers in the dataset. She addresses missing data by imputing or removing rows or columns with missing values. **Outliers are identified and treated appropriately** based on their impact on the analysis.

Exploratory Data Analysis (EDA)

As she delves into exploratory data analysis, Lila faces numerous choices. She must determine which summary statistics, visualizations, and distribution analyses will best reveal insights into customer behavior and sales trends. Each choice she makes during **EDA** influences the story the data will tell. Lila conducts EDA to gain insights into the dataset. She generates summary statistics and visualizations (histograms, scatter plots) and explores the distribution of variables. EDA helps her understand customer behavior, popular products, and sales trends.

Feature Engineering

Lila recognizes the potential for **feature engineering to enhance her analysis**. She assesses whether creating new features, such as calculating total purchase amounts, will improve the dataset's utility for her project.

Statistical Analysis, Machine Learning

Lila evaluates whether statistical tests or machine learning algorithms are necessary. She employs regression analysis to understand relationships between variables and explore machine learning models for demand forecasting or customer segmentation tasks. Lila also performs statistical tests to uncover patterns in the data. She uses regression analysis to understand the impact of unit price on sales.

Presentation and Reporting

At the culmination of her analysis, Lila faces the challenge of presenting her findings. **Lila compiles her analysis and findings using a Jupyter Notebook into a comprehensive report and presentation**. She highlights actionable insights and recommendations for the e-commerce platform's stakeholders.

Continuous Learning

After completing her first project, Lila continues to refine her skills, explores more complex datasets, and tackles increasingly challenging data science tasks.

Machine Learning Skills

Although Lila took an introductory course on Machine Learning as part of the IBM Data Science Professional Certificate, the field intrigues her, and she wants to develop her skills further by taking the **IBM Machine Learning Professional Certificate**. She identified Machine Learning Repository datasets in the course and experimented with various algorithms. Lila dives into machine learning to excel as a data scientist, wherein she studies various algorithms, such as linear regression, decision trees, and deep learning models. She continues to gain expertise in selecting and fine-tuning algorithms based on specific data problems.

Course Summary

Congratulations! You have completed this course. At this point, you know that:

- Data science is the practice of extracting valuable insights from vast datasets to guide strategic decision-making.
- Data science careers offer diverse paths, often involving mathematics, programming, and a curiosity for data exploration.
- Successful data scientists exhibit qualities like curiosity, critical judgment, and an aptitude for constructive argumentation.
- The data science field is characterized by high demand, resulting in attractive remuneration for skilled professionals.
- A Data Scientist's daily routine can vary significantly depending on the project's nature.
- A wide array of algorithms is available for extracting insights from data.
- Big Data plays a pivotal role in driving digital transformation across industries.
- Cloud computing is a fundamental technology in modern data science.
- Data mining techniques are essential for uncovering patterns and knowledge from data.
- Tools like Hadoop, HDFS, Hive, and Spark are employed for processing Big Data.
- Deep learning, machine learning, and regression are critical data science topics.
- Data science applications span diverse domains, solving complex problems.

- Companies can harness data science to address age-old challenges with innovative solutions.
- Data science contributes significantly to saving lives and improving various aspects of society.
- Careers in data science offer exciting opportunities, with mathematics and statistics being essential foundations.
- Reports in data science adhere to specific structures, and career roadmaps provide guidance.
- Case studies and projects offered practical application of the knowledge acquired during the course.

Modulo 4 Understanding data

Understanding data

Data is **unorganized information** (facts, observations, perceptions, numbers, characters, symbols, and images) that is processed to make it meaningful.

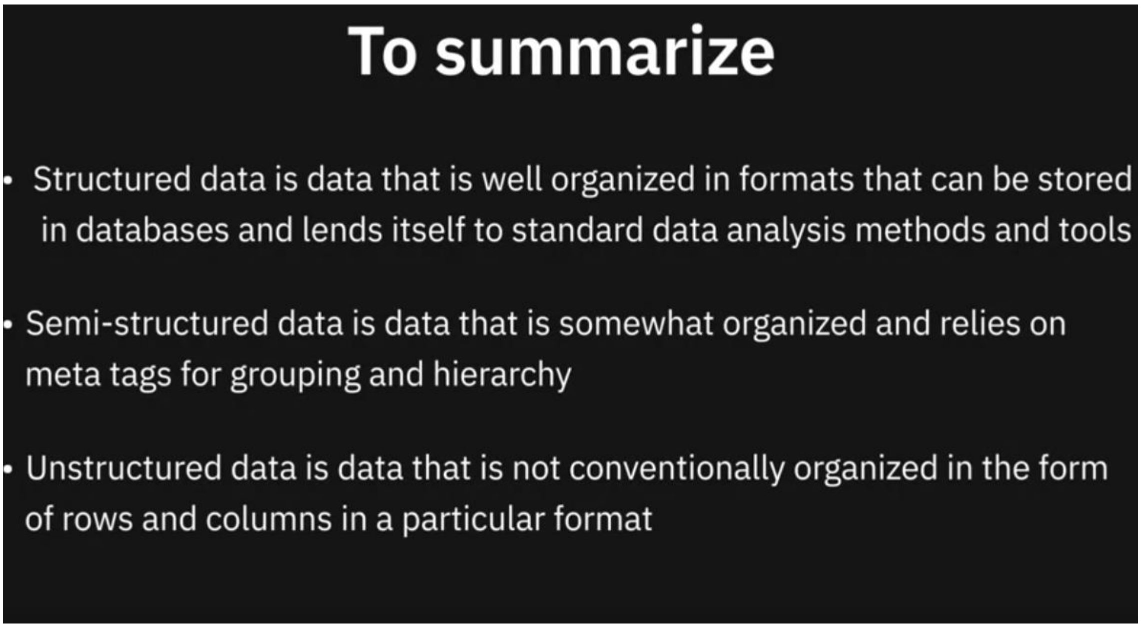
Structured data has a well-defined structure or adheres to a specified data model, can be stored in well-defined schemas such as databases, and in many cases can be represented in a tabular manner with rows and columns. Structured data is objective facts and numbers that can be collected, exported, stored, and organized in typical databases. **Sources could include**: SQL Databases and Online Transaction Processing (or OLTP), Spreadsheets, Online forms, **Sensors** (GPS or RFID) and Network and Web server **logs**. You can also easily examine structured data with standard data analysis methods and tools.

Semi-structured data is data that has some organizational properties but lacks a fixed or rigid schema. Semi-structured data **cannot be stored in the form of rows and columns as in databases**. It contains tags and elements, or metadata, which is used to group data and **organize it in a hierarchy**. **Sources could include**: E-mails, XML, JSON (facts, observations, perceptions, numbers, characters, symbols, and images), Binary executables, TCP/IP packets, Zipped files, Integration of data from different sources.

Unstructured data is data that does not have an easily identifiable structure and, therefore, **cannot be organized in a mainstream relational database** in the form of rows and columns. It does not follow any particular format, sequence, semantics, or rules.

Unstructured data can deal with the heterogeneity of sources and has a variety of business intelligence and analytics applications. **Sources could include**: Web pages, Social media feeds, Images video and audio files, documents and PDF files, PowerPoint presentations, media logs; and surveys. Unstructured **data can be**

stored in files and documents (such as a Word doc) for manual analysis or i**n NoSQL databases that have their own analysis tools** for examining this type of data.



To summarize

- Structured data is data that is well organized in formats that can be stored in databases and lends itself to standard data analysis methods and tools
- Semi-structured data is data that is somewhat organized and relies on meta tags for grouping and hierarchy
- Unstructured data is data that is not conventionally organized in the form of rows and columns in a particular format

Data sources

Relational Databases (QL Server, Oracle, MySQL, and IBM DB2,);

Organizations have internal applications that use databases to store data in a structured way and used as a source for analysis.

Flat files and XML Datasets

External to the organization, there are publicly and privately datasets. There are companies that sell specific data available as flat files, spreadsheet files, or XML documents.

Flat files (CSV), store data in plain text format, with one record or row per line, and each value separated by delimiters such as commas, semi-colons, or tabs. Data in a flat file maps to a single table.

Spreadsheet files (XLSX) are a special type of flat files that can be stored in custom formats and include additional information such as formatting, formulas.

XML files, contain data values that are identified or marked up using tags, support hierarchical data structures.

APIs and Web Services;

APIs and Web Services typically listen for incoming requests, which can be in the form of **web requests from users or network requests from application**s, and return data in plain text, XML, HTML, JSON, or media files.

Twitter and Facebook: **opinion mining or sentiment analys**is. Stock Marke APIs. Data Lookup and Validation APIs,

Web Scraping;

Web ScrapingWeb scraping is used to extract relevant data from unstructured sources. collecting product details, comparing prices.

BeautifulSoup, Scrapy, Pandas, and Selenium.

Data Streams and Feeds

Data streams are another widely used source for aggregating **constant streams of data** flowing from instruments, IoT devices, and applications, GPS data from cars, computer programs, websites, and social media posts.

This data is **generally timestamped and also geo-tagged** for geographical identification.

Some of the data streams and ways in which they can be leveraged include: stock and market tickers for **financial trading;** retail transaction streams for predicting demand and supply chain management; s**urveillance and video feeds for threat detection**; social media feeds for sentiment analysis;□sensor data feeds for monitoring industrial or **farming machinery**; web click feeds for monitoring web performanceand improving design; and real-time flight events for rebooking and rescheduling.

Some popular applications used to process data streams include **Apache Kafka, Apache Spark Streaming, and Apache Storm**.

RSS (or Really Simple Syndication) feeds, are another popular data source.

These are typically used for capturing updated data from online forums and news sites where data is refreshed on an ongoing basis.

Using a feed reader, which is an interface that converts RSS text files into a stream of updated data, updates are streamed to user devices.

Viewpoints: Working with varied data sources.

we're **often moving data from one relational database to another**. (Challenges: From **one vendor to another** or **versioning)**.

Working with multiple data sources is a**bout flexibility**.

Moving data one time is usually not all that hard as long as **we're sub-terabyte**.

Moving data consistently and continually, and in a performant way can cause us to evaluate a lot of different solutions. So we really need to **be open to new ideas and looking for new solutions** that meet the requirements that we have.

Unstructured data and data intensive applications started to look elsewhere from databases. The B-tree data structures that drive, or power, these relational databases slows down due to their nature of the random reads and random writes for the heavy write applications. (Google released a white paper back in 2006 called Google BigTable)

standard formats but also **proprietary formats.**

You will need to work with **data at rest, streaming data, or data in motion.**

When it comes to the data formats, log data, XML data, JSON, etc., each of them comes with their own challenges.

For example, **log data is extremely challenging** because it's unstructured and you may need to write **your own custom tools** to pass the data depending on what you want to look at.

XML is resource intensive, especially memory, because it has both the starting and ending tags.

JSON got it off the ending tags. looks like a key-value pairs. Saves some resources (Apache Avro).

A **lot of your challenge might come from the data itself**. **And in this particular case we had to use different separators for different tables, because every single special character that we could think of was in one of those tables**.

Metadata and Metadata Management

Objectives

After completing this reading, you will be able to:

- Define what metadata is
- Describe what metadata management is
- Explain the importance of metadata management
- List popular tools for metadata management

What is metadata?

Metadata is data that provides information about other data.

This is a very broad definintion. Here we will consider the concept of metadata within the context of databases, data warehousing, business intelligence systems, and all kinds of data repositories and platforms.

We'll consider the following **three main types of metadat**a:

- Technical metadata
- Process metadata, and
- Business metadata

Technical metadata

Defines the data structures from a technical perspective.For example:

- Tables that record information about the tables stored in a database, like:

  each table's name

  the number of columns and rows each table has

- A data catalog, which is an inventory of tables that contain information, like:

  the name of **each database** in the enteprise data warehouse

  the name of **each column** present in each database

  the **names of every table that each column is contained in**

  the **type of data** that each column contains

The technical metadata for relational databases is typically stored in specialized tables in the database called the System Catalog.

Process metadata

Process metadata **describes the processes from** enterprise systems. To be monitored for failures tracks things like:

- process start and end times
- disk usage
- where data was moved from and to, and
- how many users access the system at any given time

This sort of data is invaluable for t**roubleshooting and optimizing workflows** and ad hoc queries.

Business metadata

Users who want to explore and analyze data within and outside the enterprise are typically interested in data discovery. They need to be able to **find data which is meaningful and valuable to them and know where that data can be accessed from**. These business-minded users are thus interested in business metadata, which is information about the data described in readily interpretable ways, such as:

- how the data is acquired
- what the data is measuring or describing
- the connection between the data and other data sources

Business metadata also serves as documentation for the entire data warehouse system.

Managing metadata

Managing metadata includes developing and administering **policies and processes** to ensure information can be accessed and integrated from various sources and appropriately shared across the entire enterprise.

Creation of a reliable, user-friendly **data catalog** is a primary objective of a metadata management model. The data catalog is a core component of a modern metadata management system, serving as the main asset around which metadata management is administered. It serves as the basis by which companies **can inventory and efficiently organize their data systems**. A modern metadata managment model will include a **web-based user interface** that enables engineers and business users to easily search for and find information on key attributes such as CustomerName or ProductType. This kind of model is central to any Data Governance initiative.

Why is metadata management important?

Good metadata management has many valuable benefits. Having access to a well implemented data catalog greatly **enhances data discovery, repeatability, governance, and can also facilitate access to data**.

Well managed metadata helps you to understand both the business context associated with the enterprise data and the data lineage, which helps to improve data governance. **Data lineage provides information about the origin of the data and how it gets transformed and moved, and thus it facilitates tracing of data errors back to their root cause.** Data governance is a data management concept concerning the capability that enables an organization to **ensure that high data quality exists** throughout the complete lifecycle of the data, and data controls are implemented that support business objectives.

The **key focus areas of data governance include availability, usability, consistency, data integrity and data security** and includes establishing processes to ensure effective data management throughout the enterprise such as **accountability for the adverse effects of poor data quality** and ensuring that the data which an enterprise has can be used by the entire organization.

Popular tools for metadata management

Popular metadata management tools include:

- IBM InfoSphere Information Server
- CA Erwin Data Modeler
- Oracle Warehouse Builder
- SAS Data Integration Server
- Talend Data Fabric
- Alation Data Catalog
- SAP Information Steward
- Microsoft Azure Data Catalog
- IBM Watson Knowledge Catalog
- Oracle Enterprise Metadata Management (OEMM)
- Adaptive Metadata Manager
- Unifi Data Catalog
- data.world
- Informatica Enterprise Data Catalog

Summary

In this reading, you learned that:

- Metadata is data that provides information about other data, and includes three main types: technical, process, and business metadata
- The technical metadata for relational databases is typically stored in specialized tables in the database called the system catalog
- A primary objective of business metadata management modelling is the creation and maintenance of a reliable, user-friendly data catalog
- Having access to a well-implemented data catalog greatly enhances data discovery, repeatability, governance, and can also facilitate access to data
- Metadata management tools from IBM include InfoSphere Information Server and Watson Knowledge Catalog

Data literacy

Data Collection and Organization

Data repository: **data** that has been collected, organized, and isolated. It is **ready** to use, mine or analysis.

Types of repositories: **databases**, data **warehouses**, and **big data stores**.

Databases.

Collection of data designed for the input, storage, search, retrieval, and modification.

And a Database Management System (DBMS), is a set of programs that DO THAT

Database and DBMS mean different things.

Several factors influence the choice of database, such as the data type and structure, **querying mechanisms, latency requirements, transaction speeds**, and **intended use of the data**.

Relational databases

Or RDBMSes, build on the organizational principles of **tabular format** flat files, following a well-defined structure and schema.

However, unlike flat files, RDBMSes are **optimized for data operations and querying involving many tables** and much larger data volumes.

Structured Query Language, or SQL, is the standard querying language for relational databases.

non-relational databases.

Or NoSQL, or "Not Only SQL". emerged in response to the **volume, variety, and velocity** at which data is being generated today, made it possible to store data in a **schema-less** or free-form fashion.

NoSQL is widely used for processing big data.

Data warehouse

Works as a **central repository** that **merges information** coming from disparate sources and **consolidates it** (ETL), **into one comprehensive database** for analytics and business intelligence.

Related to Data Warehouses are the concepts of **Data Marts and Data Lakes**, have historically been relational, However, non-relational data repositories are also now being used for Data Warehousing.

Big Data Stores

Include **distributed computational and storage infrastructure to store, scale, and process** very large data sets.

Relational DataBase Management systems (RDMS)

A relational database is a collection of data organized into a table structure, where the **tables can be linked, or related, based on data common to eac**h. enables you to **retrieve** an **entirely new table from data in one or more tables** with a single query.

It also allows you to understand the relationships among all available data and gain new insights for making better decisions.

Relational databases use structured query language, (SQL), for querying **millions of records** and retrieving large amounts of data in a matter of seconds.

Relational databases, by design, are ideal for the **optimized storage, retrieval, and processing** of data for large volumes of data, unlike spreadsheets that have a limited number of rows and columns.

Relational databases's relationships **minimizes data redundancy**.

Relaional databases restrict data types and values to fields, which minimizes irregularities and leads to **greater consistency and data integrity**.

Relational databases's architecture controls access to data enforcing **data governance** rules.

Relational databases (small desktop systems to massive cloud-based systems) They can be either:

- open-source and internally supported,
- open-source with commercial support, or
- commercial closed-source systems.

IBM DB2
Microsoft SQL Server
MySQL
ORACLE DATABASE
PostgreSQL

Cloud-based relational databases, also referred to as Database-as-a-Service, are gaining wide use as they have access to the limitless compute and storage capabilities offered by the cloud.

Amazon RDS
Google SQL
IBM DB2 on Cloud
ORACLE CLOUD
SQL
Azure SQL

RDBMS is a mature and well-documented technology, making it **easy to learn and find qualified talent**.

RDBMS advantges:

Ability to create **meaningful information** by joining tables:

Flexibility: add/rename (columns tables relations) while running queries.

Reduced redundancy:

Ease of **backup and disaster recovery**: Exports can happen while the database is running. **Cloud-based relational databases do continuous mirroring**, which means the loss of data on restore can be measured in seconds or less.

ACID-compliance: ACID stands for **Atomicity, Consistency, Isolation, and Durability.** And ACID compliance implies that the data in the database remains accurate and consistent despite failures, and database transactions are processed reliably.

Now we'll look at some use cases for relational databases:

Online Transaction Processing: OLTP applications are focused on transaction-oriented tasks that run at high rates. Relational databases are well suited for OLTP applications because they can **accommodate a large number of users**; they support the ability to insert, update, or delete **small amounts of data**; and they also support f**requent queries** and updates as well as **fast response times**.

Data warehouses: In a data warehousing environment, relational databases can be **optimized for online analytical processing** (or OLAP), where historical data is analyzed for business intelligence.

IoT solutions: Internet of Things (IoT) solutions require speed as well as the ability to collect and process data from edge devices, which need a lightweight database solution.

RDBMS limitations:

RDBMS does **not work well with semi-structured and unstructured data** and is, therefore, not suitable for extensive analytics on such data.

For **migration** between two RDBMSs, s**chemas and type of data need to be identical** between the source and destination tables.

Relational databases have a l**imit on the length of data fields**, which means if you try to enter more information into a field than it can accommodate, the information will not be stored.

Despite the limitations and the evolution of data in these times of big data, cloud computing, IoT devices, and social media, RDBMS continues to be the **predominant technology** for working with structured data.

NoSQL

NoSQL, which stands for "not only SQL," provides **flexible schemas** for the storage and retrieval of **structured, semi-structured, or unstructured** data. **Existed for many years** but have only recently become popular for their attributes around s**cale, performance, and ease of use**.

NoSQL databases are built for specific data models and have flexible schemas that allow programmers to create and manage modern applications.

They do not use a traditional row/column/table database design with fixed schemas, and typically not use the structured query language (or SQL) to query data, although some may support SQL or SQL-like interfaces.

NoSQL allows data to be stored in a schema-less or free-form fashion.

Based on the model being used for storing data, there are **four common types of NoSQL databases**:

Key-value store, Document-based, Column-based, and graph-based.

Key-value store:

A collection of key-value pairs. The **key (attribute) is a unique identifier**.

Keys and values can be anything from simple integers or strings to **complex JSON documents**.

Are great for:

- User session data and user preferences.
- Real-time **recommendations**.
- Real-time **targeted advertising**.
- In-memory **data caching**.

Not be the best fit:

- To **query the data on specific data value.**
- need **relationships** between data values.
- need to have **multiple unique keys.**

Open Source   open source   AWS

Document-based:

Eeach **record** and its associated data within a single **document**. They enable flexible indexing, powerful ad hoc queries, and analytics over collections of documents.

Are preferable for:

- Platforms (**eCommerce, CRM, Analytics**),
- **Medical** records.

Not be the best fit.

- To run **complex search** queries-
- To run **multi-operation transactions**.

Thanks for your post!

When it comes to document-based databases, the difference between the two statements you mentioned can be confusing, but here's a breakdown:

1. **"Document-Based Databases Enable Powerful Ad Hoc Queries":**

- **Ad Hoc Queries**: These are queries created dynamically to answer specific questions. Document-based databases (like MongoDB) allow you to create flexible queries on the fly. This is powerful because you don't need to know the exact structure of the data beforehand, and you can search for specific fields, filter based on conditions, and even aggregate data.

- **Document Structure**: Document-based databases store data in collections of documents (usually in JSON-like formats). This means that the database can quickly find and return documents that match a certain pattern or condition.

Example: In a MongoDB collection of users, you could easily create a query to find all users from a specific city who are over 30 years old.

2. **"If You're Looking to Run Complex Search Queries ... Document-Based Database May Not Be the Best Option":**

- **Complex Search Queries**: These refer to more advanced queries that involve multiple joins, highly relational data, or complicated transactions. Document-based databases are often not as efficient for these types of queries compared to relational databases like MySQL or PostgreSQL.

- **Relational Data**: In traditional relational databases, data is often split across multiple tables and linked via foreign keys. This structure allows for complex joins and transactions but requires more rigid schema design and careful management.

Example: If your query needs to join multiple collections/tables to get the desired data (like finding all users and then filtering by specific actions in a separate collection), a document-based database may not be as optimized for these operations compared to a relational database.

Summary:

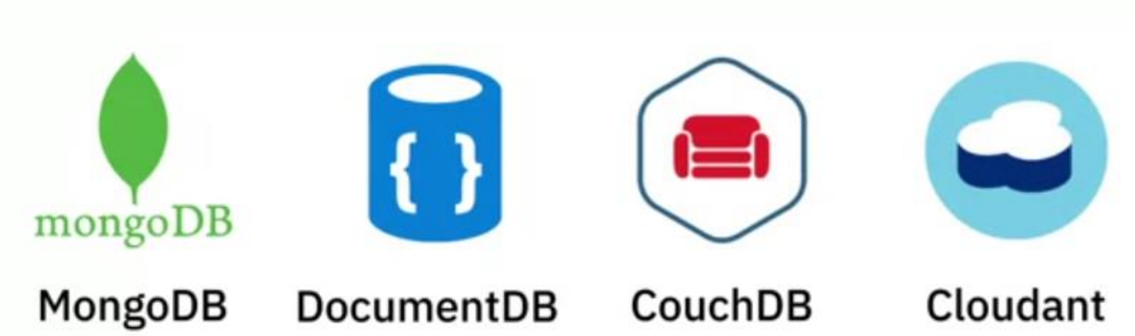
- **Ad Hoc Queries** in document-based databases are powerful because they allow flexibility and quick querying of documents based on fields and conditions.

- **Complex Queries** involving multiple collections, deep relationships, or extensive transactional operations might be better suited for **relational databases** rather than document-based ones.

Does this help clarify things? Let me know if you want to explore any specific examples!

Best Regards,

Geetika.



Open Source AWS Open Source IBM

Column-based:

Column-based models store data in cells grouped as columns of data instead of rows.

A logical grouping of columns, that is, columns that are usually accessed together, is called a column family.

For example, a customer's name and profile information will most likely be accessed together but not their purchase history.

So,customer name and profile information data can be grouped into a column family.

Since column databases **store all cells corresponding to a column as a continuous disk entry**, accessing and searching the data becomes very fast.

Are preferable for:

- **Heavy write** requests.
- **time-series** data, **weather** data, and **IoT** data.

Not be the best fit.

- To run **complex search** queries-
- Frequently change your querying patterns.

Graph-based:

Graph-based databases use a graphical model to represent and store data.

They are particularly useful for visualizing, analyzing, and finding connections between different pieces of data.

The circles are nodes, and they contain the data. The arrows represent relationships.

Are preferable for:

- Data **containing lots of interconnected relationships**.
- social **networks and diagram**, real-time product **recommendations**, **fraud detection**, and access management.

Not be the best fit.

- High **volumes** of transactions.
- Large-**volume** analytics queries

Advantages of NoSQL

NoSQL was created **in response to the limitations** of traditional relational database technology.

Handle **large volumes of structured, semi-structured, and unstructured data**.

The ability to run as distributed systems scaled across multiple data centers, which enables them to **take advantage of cloud computing infrastructure**;

An efficient and **cost-effective scale-out architecture** that provides additional capacity and performance with the addition of new nodes;

Simpler design, better control over **availability**, and improved **scalability** that enables you to be more **agile**, more **flexible**, and to iterate more quickly-

Differences between relational and non-relational databases:

RDBMS schemas rigidly define how all data inserted into the database must be typed and composed, whereas **NoSQL databases can be schema-agnostic**, allowing unstructured and semi-structured data to be stored and manipulated.

Maintaining high-end, commercial **RDBMS is expensive** whereas NoSQL databases are specifically designed for **low-cost commodity hardware.**

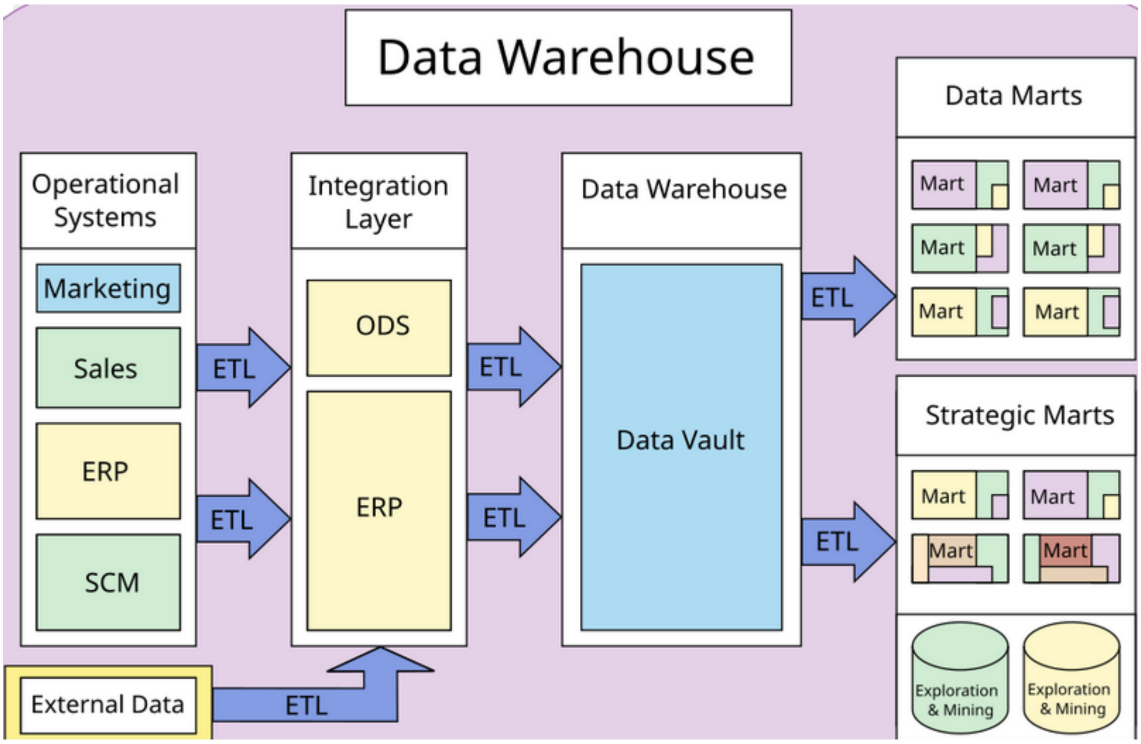
Unlike most NoSQL, RDBMS support ACID-compliance, which ensures reliability of transactions and crash recovery.

RDBMS is a mature and well-documented technology, which means the risks are more or less perceivable as compared to NoSQL, which is a relatively newer technology.

Nonetheless, NoSQL databases are here to stay, and are increasingly being used for mission critical applications.

Data Marts, Data Lakes, ETL y Data Pipelines

Data warehouse:



A multi-purpose storage for different use cases.

Data warehouses serve as the **single source of truth**—storing current and historical data that has been modeled, structured, cleansed, conformed, and categorized for a specific purpose, meaning it is **analysis ready**

A data mart

is a sub-section of the data warehouse, built specifically for a **particular business function, purpose, or community of users**.

Since a data mart offers analytical capabilities for a restricted area of the data warehouse, it **offers isolated security and isolated performance**.

The most important role of a data mart is **business-specific reporting and analytics**.

Data Lake

Is a storage repository that can store **large amounts of structured, semi-structured, and unstructured data** in their **native format, classified and tagged with metadata**.

So, while a data warehouse stores data processed for a specific need, a data lake is a pool of **raw data** where each data element is given a unique identifier and is tagged with metatags for further use.

You would opt for a data lake if you generate, or have access to, large volumes of data on an ongoing basis, but don't want to be restricted to specific or pre-defined use cases.

Unlike data warehouses, a data lake would **retain all source data**, without any exclusions. Data lakes are sometimes also used as a **staging area** of a data warehouse.

The most important role of a data lake is in **predictive and advanced analytic**s.

ETL

ETL converts **raw data into analysis-ready data**.

It is an **automated process** in which you gather raw data from identified sources, **extract the information that aligns with your reporting and analysis needs**, clean, standardize, and transform that data into a format that is usable in the context of your organization; and l**oad it into a data repository**.

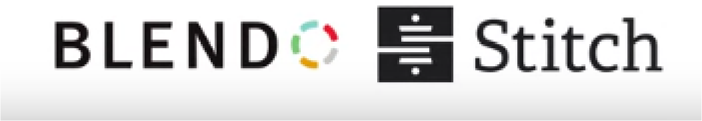
While ETL is a generic process, the actual job can be very different in usage, utility, and complexity.

Extract

Collects data from source locations for transformation.

Batch processing,

meaning source **data, is moved in large chunks** from the source to the target system at scheduled intervals.



Pentaho, Apache Airflow

Stream processing,

Meaning source **data is pulled in real-time** from the source and transformed while it is in transit and before it is loaded into the data repository.

Tools for stream processing include Apache Samza, Apache Storm, and Apache Kafka.

Transform

Rules and functions that converts raw data into data that can be used for analysis.

For example, making date formats and units of **measurement consistent across all sourced data, removing duplicate data, filtering out** data that you do not need, **enriching** data, for example, splitting full name to first, middle, and last names, establishing key relationships across tables, applying **business rules** and data **validations**.

Load

Data is transported to a destination system or data repository.

- **Initial** loading, that is, **populating** all the data in the repository,
- **Incremental** loading, that is, applying ongoing **updates and modifications** as needed periodically;
- **Full refresh**, that is, **erasing contents** of one or more tables and r**eloading with fresh data**.

Load verification, which includes data checks for missing or null values, server performance, and monitoring load failures, are important parts of this process step.

It is vital to keep an eye on load failures and **ensure the right recovery mechanisms are in place.**

ETL has historically been used for **batch** workloads on a large scale. However, with the emergence of streaming ETL tools, they are increasingly being used for **real-time** streaming event data as well.

Pipelines

Data pipeline is a **broader term** that encompasses the entire journey of moving data from one system to another, of **which ETL is a subset**.

Data pipelines can be architected for batch processing, for streaming data, and a combination of batch and streaming data.

In the case of **streaming data**, data processing or transformation, happens in a **continuous flow**. This is particularly **useful for data that needs constant updating**, such as data from a sensor monitoring traffic.

A data pipeline is a high performing system that supports both **long-running batch queries and smaller interactive queries**.

The destination for a data pipeline is typically a data lake, although the data may also be loaded to different target destinations, such as another application or a visualization tool.

Considerations for choice of a data repository.

Talk 2.

There's a number of factors to keep in mind while picking the right database for the job. You need to look at the use case. What is the data repository going to be **used for**? Is it going to be used for **storing structured information, semi- structured or unstructured** information. Or do you know beforehand what the **schema** of the data is?

Is there **performance** requirements? Are you working with **data at rest, or streaming data, or data in motion**?

Does the data need to be **encrypted**? Does there... is there, you know, what's the **volume** of data that you're working with? Do you need a big data system?

And what are the storage requirements? Does the data need to be **updated frequently and accessed frequently**, that it just needs to be stored and kept in a valt for a long time and is needed for backup purposes for example?

And then your organization might have certain **standards** that might have put in place of which databases or which data repositories you're allowed to use for different kinds of tasks.

Talk 2.

We look at what are the kind of capacities that this data repository is supposed to handle.

And then we also look at the **type of access** that we need this for. Do we access it in **short intervals** or do we run **long running queries** on it?

Am I **using** it more for **transaction processing** or am I using it for **analytics** or **archival purposes**, or for data **warehousing** purpose?

We also look for **compatibility**. How compatible this new data repository is with my existing ecosystem of programming languages, tools, and any processes that we have.

We also consider the **security** features this repository gives us.

And the most important thing is **scalability**. We may be happy with its performance today, but is it scalable enough? Can it scale along with the organization?

Talk 3.

I don't often get to choose the type of data repository that my organization uses, and **very few organizations use one data repository** these days.

On my team that I work on these days, we have a **set of preferred solutions**. We have a preferred **enterprise relational database**. We have a preferred **open-source** relational database for some of the smaller projects and for the microservices. And then we also have a **preferred unstructured data source**. So those are three main ones.

The important thing is to think about the skills **that you have within your organization** or that you want to foster within your organization.

And consider the **costs** of the various solutions.

In our case, we have some experts on Db2, so our enterprise database of choice is Db2. However, there are other projects that use different ones. For open source, we've changed that a couple of times. We've got a couple of different directions with where we really want to be there.

And all of these... the **hosting platform makes a difference as well**, because now it's not just do I want to use IBM Db2 or do I want to use some other vendors, Microsoft SQL Server or whatever. It's not between those two choices. It's when I do those, do I want to do them on AWS RDS? Maybe I should consider Amazon's Aurora. Maybe I should consider Googles relational offerings. There's so many different choices there that you have to consider.

There's the decision of **how should the data be stored**. There's the decision of how should the data **be retrieved**, and there's also the decision of **where**.

Talk 4.

I would say the **structur**e of the data, the **nature of the application**, and the **volume** at which the data is getting ingested into your database, all these factors determine the nature of the data source that you should pick.

In **most cases a relational database should be enough**, however, there will be edge cases where relational databases such as IBM Db2, Oracle or Postgres won't necessarily do the job for you.

In those cases, so depending on the use case, for example, if you are ingesting **gigabytes or terabytes of data per day**. Then document stores such as MongoDB, or wide column stores such as Cassandra might be a good fit for you.

At the same time, if you're trying to build a product recommendation engines or trying to show the network of relationships between different people on the social media, then graph data structures such as Neo4J or Apache TinkerPop would be an ideal fit for you.

At the same time, if you are **mining through terabytes or petabytes of data for analytics,** Hadoop engine with MapReduce would be a good fit for you.

So it really boils down to the **nature of the application** and the **volume** of the data, and the **structure of the data**, before you can pick the right database or data source whatever the use case.

Data integration Platforms

Gartner defines data i**ntegration as a discipline** comprising the practices, architectural techniques, and tools that allow organizations to i**ngest, transform, combine, and provision data across various data types**.

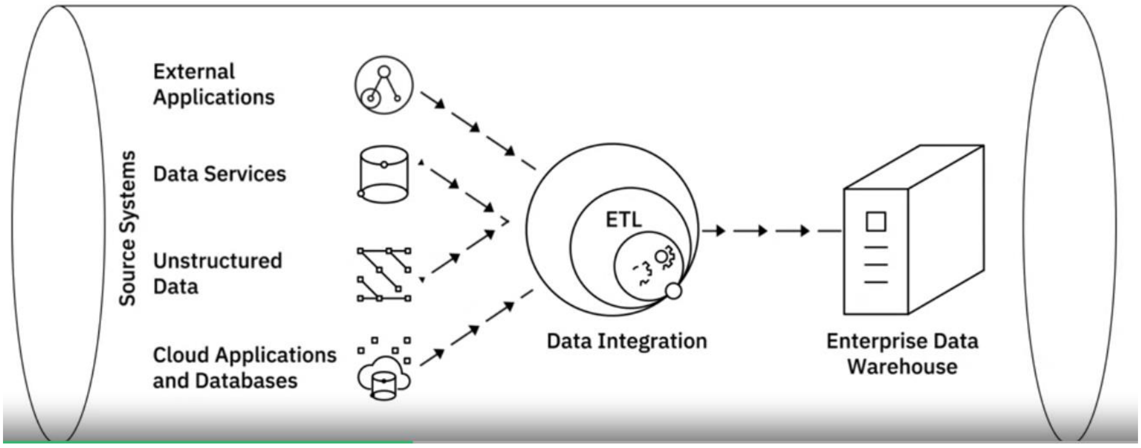
The report further explains that data integration has several **usage scenarios**, such as **data consistency** across applications, master **data management**, **data sharing** between enterprises, and **data migration** and **consolidation**.

In the **field of analytics and data science**, data integration includes **accessing**, **queueing**, or **extracting** data from operational systems **transforming** and **merging** extracted data either logically or physically **data quality and governance, and delivering** data through an integrated approach for analytics purposes

For example, to make customer data available for analytics, you would need to extract individual customers' information from operational systems such as sales, marketing, and finance. You would then need to provide a unified view of the combined data so that your users can access, query, and manipulate this data from a single interface to derive statistics, analytics, and visualizations.

How does a data integration platform relate to ETL and data pipelines?

While **data integration combines disparate data into a unified view of the data**, a **data pipeline covers the entire data movement journey from source to destination systems**.



In that sense, you use a data pipeline to perform data integration, while **ETL is a process within data integration.**

There is no one approach to data integration.

However, modern data integration solutions typically support the following capabilities:

An extensive **catalog of pre-built connectors and adopters** that help you connect and build integration flows with a wide variety of data sources such as databases, flat files, social media data, APIs, CRM and ERP applications.

Open-source architecture that provides greater flexibility and avoids vendor lock-in.

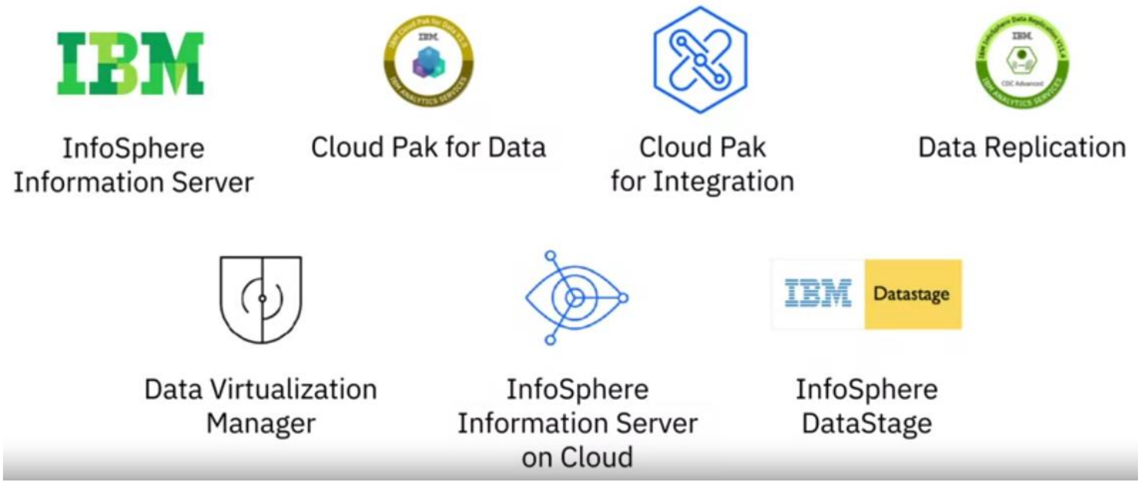
Optimization for both batch processing of large-scale data and continuous data streams, or both.

Integration with Big Data sources. Support for big data is increasingly driving the decision regarding choice of integration platforms.

Additional functionalities. For example, specific demands around **data quality** and **governance**, **compliance**, and **security**.

Portability, which ensures that as businesses move to cloud models, they should be able **to run their data integration platforms anywhere**. And data integration tools are able to work natively in a single cloud, multi-cloud, or hybrid cloud environment.

Data integration platforms and tools available (Privative).



InfoSphere Information Server

Cloud Pak for Data

Cloud Pak for Integration

Data Replication

Data Virtualization Manager

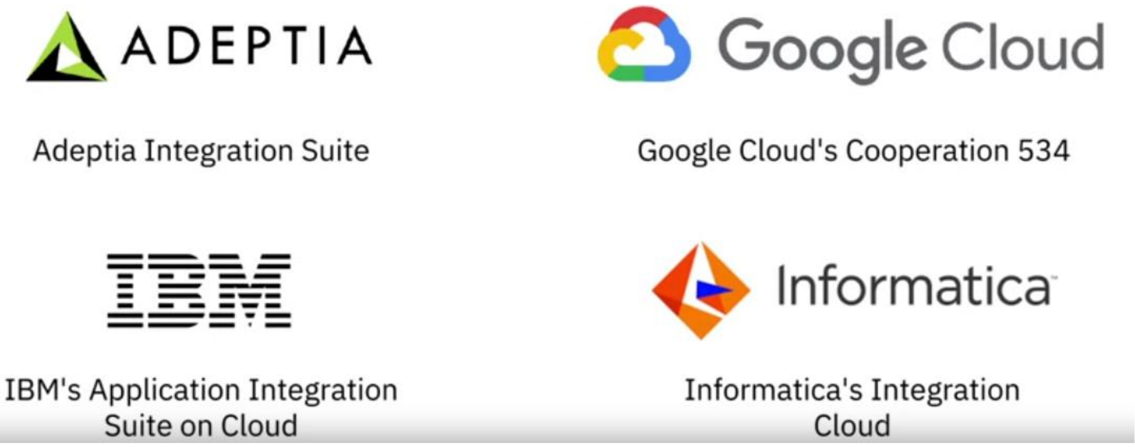
InfoSphere Information Server on Cloud

InfoSphere DataStage

- Talend Data Fabric
- Talend Cloud
- Talend Data Catalog
- Talend Data Management
- Talend Big Data
- Talend Data Services
- Talend Open Studio

open-source Data integration frameworks

cloud-based Integration Platform as a Service, (iPaaS)



Adeptia Integration Suite

Google Cloud's Cooperation 534

IBM's Application Integration Suite on Cloud

Informatica's Integration Cloud

The data integration space continues to evolve as businesses embrace newer technologies and as data grows, be it in the variety of sources or its use in business decision-making.

Summary

Welcome to the data literacy lesson summary.

As a data scientist, you need an awareness of data **storage** possibilities for its organization and **management**, and options for **retrieval**.

These systems enable you to find and analyze the data you need to make great discoveries hidden in that data.

In this video, we'll summarize what you learned in this lesson about the technologies and tools to handle large amounts of data.

Data repositories.

These repositories need the ability to find the data you want and return it to you in a usable format.

Your **data type helps determine the type of repository** you need.

You can store structured, semi structured or unstructured data.

Depending on the organization, you may need a relational or no SQL database.

For big data stores, your needs may call for a data warehouse, a data mart, or a data lake.

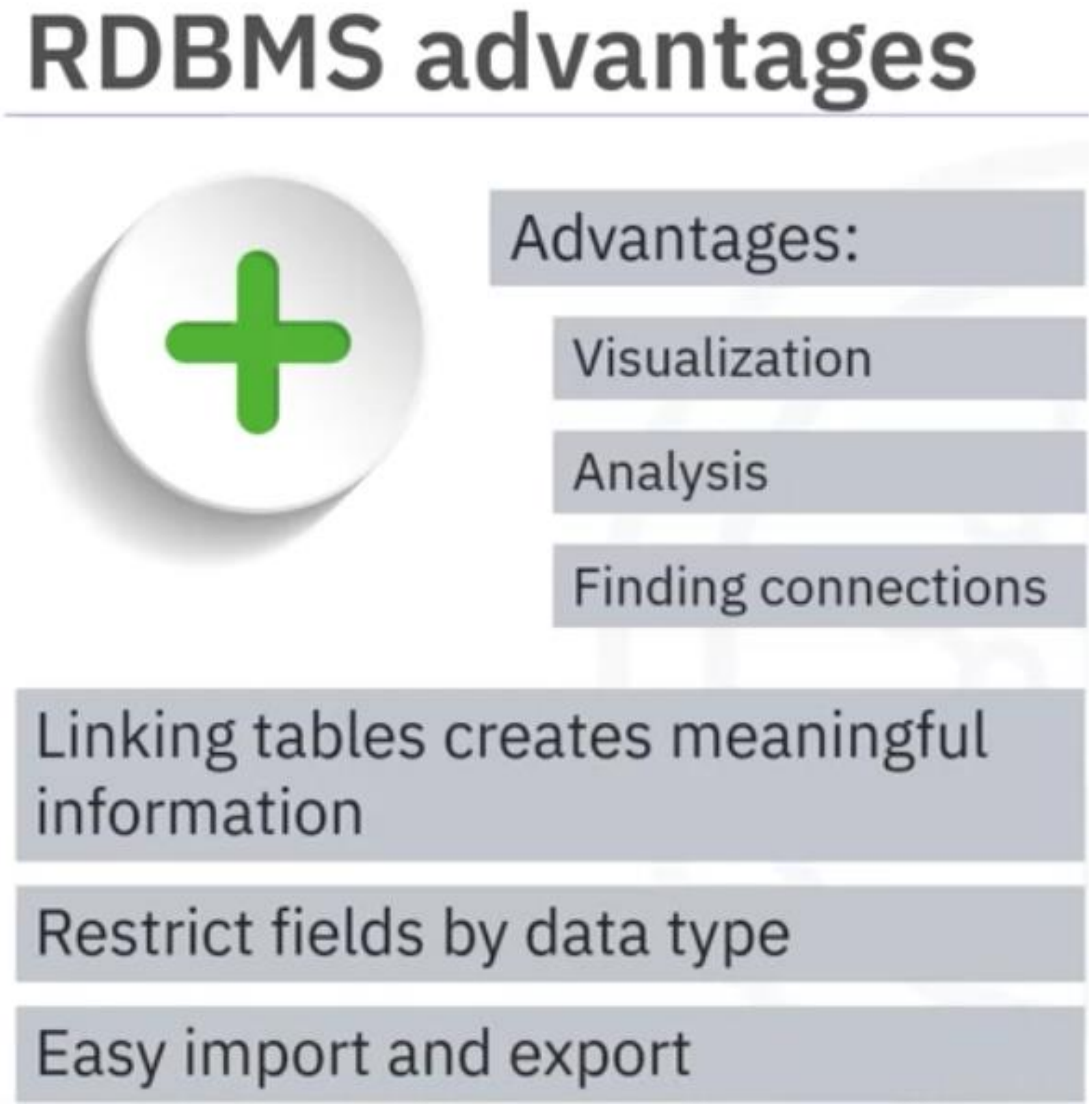
Relational databases store structured data. These are the oldest types of repositories.

The most conventional and frequently used relational database management systems, often abbreviated as RDBMSs, are based on the foundational concept

of structuring data in a tabular format with data arranged in rows and columns.

Each table usually relates to a topic, and the columns of data in the table contain a specific type of information related to that topic. Then the database contains a defined schema that describes the table to each other.

Relational databases usually rely on structured query language or SQL, to search for and retrieve the data you need. You use SQL to manipulate the data.



RDBMS advantages

Advantages:

- Visualization
- Analysis
- Finding connections

Linking tables creates meaningful information

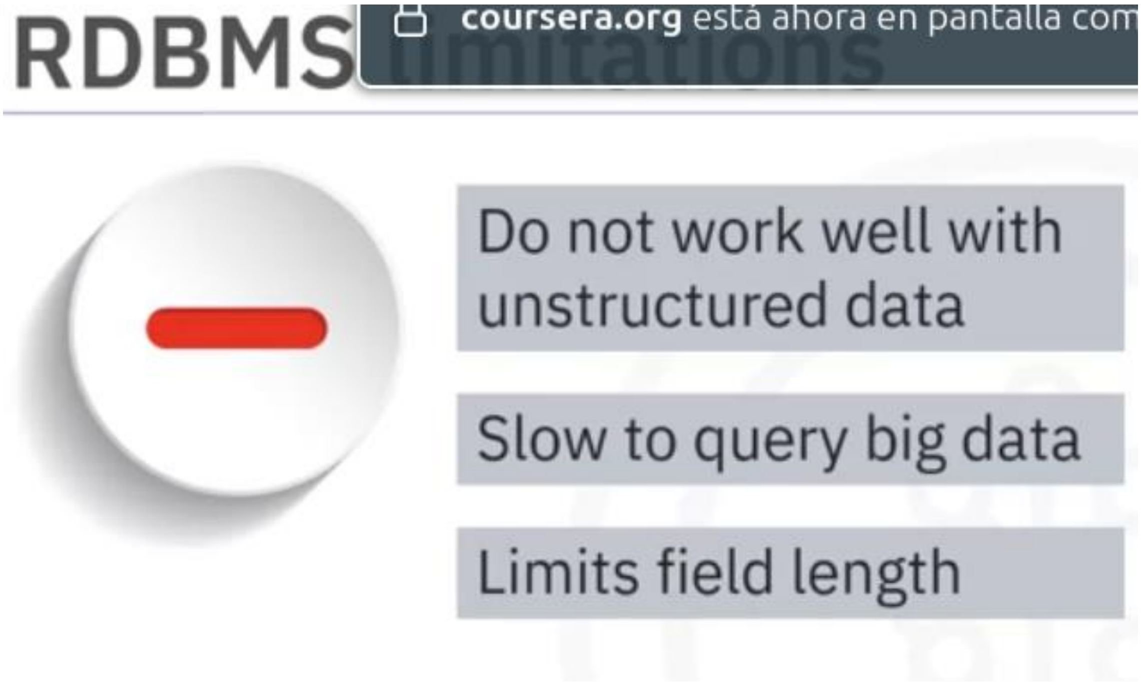
Restrict fields by data type

Easy import and export

Relational databases are beneficial for visualizing, analyzing, and finding connections between different pieces of data.

You link tables together by creating schemas, you can restrict database fields to specific data types and values which minimizes irregularities and leads to greater **consistency and data integrity.**

They offer easy export and import options, making back up and restoration easy.



RDBMS limitations

coursera.org está ahora en pantalla com

Do not work well with unstructured data

Slow to query big data

Limits field length

However, RBMS is do not work well with semi structured or unstructured data.

They are also slow to query with enormous datasets.

Since RDBMSs use predefined structures for data to reside in, it becomes problematic when the data evolves and no longer conforms to that structure.

Relational databases also limit field length, which means that sometimes they cannot accommodate the information you need.

Because of these limitations and the quantities and diversity of data collected, many organizations have turned to not only SQL databases, or no SQ L for short.

Built for speed, flexibility, and scale, non relational databases allow storing **data without stringent schemas**. They can house semi structured and unstructured data. No SQL databases include document based, key value, columnar and graph.

Document based databases store semi structured documents, such as Jason files. You group documents into collections, and each document has its structure.

Key value stores each piece of data as a key value pair, so you retrieve and update the data using the key.

Columnar databases store data and columns rather than rows, enabling storage of large volumes of data suitable for analytical workloads.

Graph databases store data in nodes. Nodes have relationships and properties, and can manage and query complex relationships between them.

You can use technologies such as data warehouses, data marts, and data lakes for high volumes of data. A **data warehouse** works like a multipurpose storage for different use cases. The data has already been modeled and structured for a specific purpose.

As an organization, you would opt for a data warehouse when you have a massive amount of data from your operational systems that must be readily available for reporting and analysis.

A **data mar**t is a subsection of the data warehouse built specifically for a **particular business function**, purpose, or community of users. A data mart offers analytical capabilities for restricted data warehouse area, offering isolated security and performance.

A **data lake** is a storage repository that can store large amounts of structured, semi structured, and unstructured data **in their native format**, classified and tagged with meta data.

Storage options.

Data pipelines address an organization's need to collect, transform, and move data.

Data pipelines have multiple steps, providing a systematic process to handle massive amounts of data as it is continually collected, processed, and made available.

ETL, which stands for extract, transform, and load, is a subset of a data pipeline, referring to **an automated process where an organization converts its raw data into data ready for analysis**.

Now as a future data scientist, you are aware of many technologies needed to handle big data before analysis can begin. These include data storage, organization and management and retrieval.

Data storage options depend on the type of data, its volume, and how you intend to organize it.

Using a data pipeline such as ETL, provides a process to manage and retrieve the data so you can analyze it as a data scientist.

GLOSSARY

Term	Definition	Video where the term is introduced
Comma-separated values (CSV) / Tab-separated values (TSV)	Commonly used format for storing tabular data as plain text where either the comma or the tab separates each value.	Working on Different File Formats
Data file types	A computer file configuration is designed to store data in a specific way.	Working on Different File Formats
Data format	How data is encoded so it can be stored within a data file type.	Working on Different File Formats

Data visualization	A visual way, such as a graph, of representing data in a readily understandable way makes it easier to see trends in the data.	Data Science Topics and Algorithms
Delimited text file	A plain text file where a specific character separates the data values.	Working on Different File Formats
Extensible Markup Language (XML)	A language designed to structure, store, and enable data exchange between various technologies.	Working on Different File Formats
Hadoop	An open-source framework designed to store and process large datasets across clusters of computers.	What Makes Someone a Data Scientist
JavaScript Object Notation (JSON)	A data format compatible with various programming languages for two applications to exchange structured data.	Working on Different File Formats
Jupyter notebooks	A computational environment that allows users to create and share documents containing code, equations, visualizations, and explanatory text. See Python notebooks.	Data Science Skills & Big Data
Nearest neighbor	A machine learning algorithm that predicts a target variable based on its similarity to other values in the dataset.	Working on Different File Formats
Neural networks	A computational model used in deep learning that mimics the structure and functioning of the human brain's neural pathways. It takes an input, processes it using previous learning, and produces an output.	A Day in the Life of a Data Scientist
Pandas	An open-source Python library that provides tools for working with structured data is often used for data manipulation and analysis.	Data Science Skills & Big Data
Python notebooks	Also known as a "Jupyter" notebook, this computational environment allows users to create and share documents containing code, equations,	Data Science Skills & Big Data

	visualizations, and explanatory text.	
R	An open-source programming language used for statistical computing, data analysis, and data visualization.	Data Science Skills & Big Data
Recommendation engine	A computer program that analyzes user input, such as behaviors or preferences, and makes personalized recommendations based on that analysis.	A Day in the Life of a Data Scientist
Regression	A statistical model that shows a relationship between one or more predictor variables with a response variable.	Data Science Topics and Algorithms
Tabular data	Data that is organized into rows and columns.	A Day in the Life of a Data Scientist
XLSX	The Microsoft Excel spreadsheet file format.	Working on Different File Formats

Term	Definition	Video where the term is introduced
Analytics	The process of examining data to draw conclusions and make informed decisions is a fundamental aspect of data science, involving statistical analysis and data-driven insights.	Data Scientists at New York University
Big Data	Vast amounts of structured, semi-structured, and unstructured data are characterized by its volume, velocity, variety, and value, which, when analyzed, can provide competitive advantages and drive digital transformations.	How Big Data is Driving Digital Transformation
Big Data Cluster	A distributed computing environment comprising thousands or tens of thousands of interconnected computers that collectively store and process large datasets.	What is Hadoop?
Broad Network Access	The ability to access cloud resources via standard mechanisms and platforms such as mobile devices, laptops, and workstations over networks.	Introduction to Cloud
Chief Data Officer (CDO)	An emerging role responsible for overseeing data-related initiatives, governance, and strategies, ensuring that data plays a central role in digital transformation efforts.	How Big Data is Driving Digital

		Transformation
Chief Information Officer (CIO)	An executive is responsible for managing an organization's information technology and computer systems, contributing to technology-related aspects of digital transformation.	How Big Data is Driving Digital Transformation
Cloud Computing	The delivery of on-demand computing resources, including networks, servers, storage, applications, services, and data centers, over the Internet on a pay-for-use basis.	Introduction to Cloud
Cloud Deployment Models	Categories that indicate where cloud infrastructure resides, who manages it, and how cloud resources and services are made available to users, including public, private, and hybrid models.	Introduction to Cloud
Cloud Service Models	Models based on the layers of a computing stack, including Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS), represent different cloud computing offerings.	Introduction to Cloud
Commodity Hardware	Standard, off-the-shelf hardware components are used in a big data cluster, offering cost-effective solutions for storage and processing without relying on specialized hardware.	What is Hadoop?
Data Algorithms	Computational procedures and mathematical models used to process and analyze data made accessible in the cloud for data scientists to deploy on large datasets efficiently.	Cloud for Data Science
Data Replication	A strategy in which data is duplicated across multiple nodes in a cluster to ensure data durability and availability, reducing the risk of data loss due to hardware failures.	What is Hadoop?
Data Science	An interdisciplinary field that involves extracting insights and knowledge from data using various techniques, including programming, statistics, and analytical tools.	Data Scientists at New York University
Deep Learning	A subset of machine learning that involves artificial neural networks inspired by the human brain, capable of learning and making complex decisions from data on their own.	Data Scientists at New York University
Digital Change	The integration of digital technology into business processes and operations leads to improvements and innovations in how organizations operate and deliver value to customers.	How Big Data is Driving Digital Transformation

Digital Transformation	A strategic and cultural organizational change driven by data science, especially Big Data, to integrate digital technology across all areas of the organization, resulting in fundamental operational and value delivery changes.	How Big Data is Driving Digital Transformation
Distributed Data	The practice of dividing data into smaller chunks and distributing them across multiple computers within a cluster enables parallel processing for data analysis.	What is Hadoop?
Hadoop	A distributed storage and processing framework used for handling and analyzing large datasets, particularly well-suited for big data analytics and data science applications.	Data Scientists at New York University
Hadoop Distributed File System (HDFS)	A storage system within the Hadoop framework that partitions and distributes files across multiple nodes, facilitating parallel data access and fault tolerance.	What is Hadoop?
Infrastructure as a Service (IaaS)	A cloud service model that provides access to computing infrastructure, including servers, storage, and networking, without the need for users to manage or operate them.	Introduction to Cloud
Java-Based Framework	Hadoop is implemented in Java, an open-source, high-level programming language, providing the foundation for building distributed storage and processing solutions.	Big Data Processing Tools: Hadoop, HDFS, Hive, and Spark
Map Process	The initial step in Hadoop’s MapReduce programming model, where data is processed in parallel on individual cluster nodes, often used for data transformation tasks.	What is Hadoop?
Measured Service	A characteristic where users are billed for cloud resources based on their actual usage, with resource utilization transparently monitored, measured, and reported.	Introduction to Cloud
On-Demand Self-Service	The capability for users to access and provision cloud resources such as processing power, storage, and networking using simple interfaces without human interaction with service providers.	Introduction to Cloud
Rapid Elasticity	The ability to quickly scale cloud resources up or down based on demand, allowing users to access more resources when needed and release them when not in use.	Introduction to Cloud
Reduce Process	The second step in Hadoop's MapReduce model is where results from the mapping process are	What is Hadoop?

	aggregated and processed further to produce the final output, typically used for analysis.	
Replication	The act of creating copies of data pieces within a big data cluster enhances fault tolerance and ensures data availability in case of hardware or node failures.	What is Hadoop?
Resource Pooling	A cloud characteristic where computing resources are shared and dynamically assigned to multiple consumers, promoting economies of scale and cost-efficiency.	Introduction to Cloud
Skills Network Labs (SN Labs)	Learning resources provided by IBM, including tools like Jupyter Notebooks and Spark clusters, are available to learners for cloud data science projects and skill development.	Cloud for Data Science
Spilling to Disk	A technique used in memory-constrained situations where data is temporarily written to disk storage when memory resources are exhausted, ensuring uninterrupted processing.	Big Data Processing Tools: Hadoop, HDFS, Hive, and Spark
STEM Classes	Science, Technology, Engineering, and Mathematics (STEM) courses typically taught in high schools prepare students for technical careers, including data science.	Data Scientists at New York University
Variety	The **diversity of data types**, including structured and unstructured data from various sources such as text, images, video, and more, posing data management challenges.	Foundations of Big Data
Velocity	The **speed at which data accumulates** and is generated, often in real-time or near-real-time, drives the need for rapid data processing and analytics.	Foundations of Big Data
Veracity	The quality and accuracy of data, ensuring that it conforms to facts and is consistent, complete, and free from ambiguity, impacts data reliability and trustworthiness.	Foundations of Big Data
Video Tracking System	A system used to capture and analyze video data from games, enabling in-depth analysis of player movements and game dynamics, contributing to data-driven decision-making in sports.	How Big Data is Driving Digital Transformation
Volume	The **scale of data generated and stored** is driven by increased data sources, higher-resolution sensors, and scalable infrastructure.	Foundations of Big Data
V's of Big Data	A set of characteristics common across Big Data definitions, including Velocity, Volume, Variety, Veracity, and Value, highlighting the rapid	Foundations of Big Data

	generation, scale, diversity, quality, and value of data.	

Term	Definition	Video where the term is introduced
Artificial Neural Networks	Collections of small computing units (neurons) that process data and learn to make decisions over time.	Artificial Intelligence and Data Science
Bayesian Analysis	A statistical technique that uses Bayes' theorem to update probabilities based on new evidence.	Applications of Machine Learning
Business Insights	Accurate insights and reports generated by generative AI can be updated as data evolves, enhancing decision-making and uncovering hidden patterns.	Generative AI and Data Science
Cluster Analysis	The process of grouping similar data points together based on certain features or attributes.	Neural Networks and Deep Learning
Coding Automation	Using generative AI to automatically generate and test software code for constructing analytical models, freeing data scientists to focus on higher-level tasks.	Generative AI and Data Science
Data Mining	The process of automatically searching and analyzing data to discover patterns and insights that were previously unknown.	Artificial Intelligence and Data Science
Decision Trees	A type of machine learning algorithm used for decision-making by creating a tree-like structure of decisions.	Applications of Machine Learning
Deep Learning Models	Includes Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) that create new data instances by learning patterns from large datasets.	Generative AI and Data Science
Five V's of Big Data	Characteristics used to describe big data: Velocity, volume, variety, veracity, and value.	Neural Networks and Deep Learning

Generative AI	A subset of AI that focuses on creating new data, such as images, music, text, or code, rather than just analyzing existing data.	Generative AI and Data Science
Market Basket Analysis	Analyzing which goods tend to be bought together is often used for marketing insights.	Neural Networks and Deep Learning
Naive Bayes	A simple probabilistic classification algorithm based on Bayes' theorem.	Applications of Machine Learning
Natural Language Processing (NLP)	A field of AI that enables machines to understand, generate, and interact with human language, revolutionizing content creation and chatbots.	Generative AI and Data Science
Precision vs. Recall	Metrics are used to evaluate the performance of classification models.	Applications of Machine Learning
Predictive Analytics	Using machine learning techniques to predict future outcomes or events.	Neural Networks and Deep Learning
Synthetic Data	Artificially generated data with properties similar to real data, used by data scientists to augment their datasets and improve model training.	Generative AI and Data Science

Term	Definition	Video where the term is introduced
Arithmetic Models	Data science often uses Mathematical models to analyze data and predict outcomes.	Old problems, new data science solutions
Case study	In-depth analysis of an instance of a chosen subject to draw insights that inform theory, practice, or decision-making.	Old problems, new data science solutions
Data mining	Extracting information from raw data, such as making decisions, predicting trends, or understanding phenomena.	How Data Science is Saving Lives
Data Science	The field involves collecting, analyzing, and interpreting data to extract valuable insights and make informed decisions.	Old problems, new data science solutions

Data Strategy	A plan that outlines how an organization will collect, manage, and use data to achieve its goals.	Old problems, new data science solutions
Predictive analytics	Using data, algorithms, models, and machine learning to make predictions.	How Data Science is Saving Lives

Term	**Definition**	**Video where the term is introduced**
Adobe Spark	A suite of software tools that allow users to create and share visual content such as graphics, web pages, and videos.	Recruiting for Data Science
Analytical skills	The ability to analyze information systematically, logically, and organized.	Recruiting for Data Science
Chief information officer (CIO)	A business executive is responsible for an organization's information technology systems and tech-related initiatives.	How Can Someone Become a Data Scientist
Computational thinking	Breaking problems into smaller parts and using algorithms, logic, and abstraction to develop solutions. Often used but not limited to computer science.	How Can Someone Become a Data Scientist
Data clusters	A group of similar, related data points distinct from other clusters.	How Can Someone Become a Data Scientist
Executive summary	Usually occurring at the beginning of a research paper, this section summarizes the important parts of the paper, including its key findings.	The Report Structure
High-performing computing (HPC) cluster	A computing technology that uses a system of networked computers designed to solve complex and computationally intensive problems in traditional environments.	How Can Someone Become a Data Scientist
Mathematical computing	The use of computers to calculate, simulate, and model mathematical problems.	Importance of Mathematics and Statistics for Data Science
Matrices	Plural for matric, matrices are a rectangular (tabular) array of numbers often used in mathematics, statistics, and computer science.	Recruiting for Data Science
Stata	A software package used for statistical analysis.	Recruiting for Data Science

Statistical distributions	A way of describing the likelihood of different outcomes based on a dataset. The “bell curve” is a common statistical distribution.	How Can Someone Become a Data Scientist
Structured Query Language (SQL)	A language used for managing data in a relational database.	Importance of Mathematics and Statistics for Data Science
TCP/IP network	A network that uses the TCP/IP protocol to communicate between connected devices on that network. The Internet uses TCP/IP.	How Can Someone Become a Data Scientist

Term	**Definition**	**Video where the term is introduced**
Comma-separated values (CSVs)	Delimited text files where the delimiter is a comma. Used to store structured data.	Understanding Different Types of File Formats
Delimited text file formats	Text files are used to store data where each line or row has values separated by a delimiter. A delimiter is a sequence of one or more characters specifying the boundary between values. Common delimiters include comma, tab, colon, vertical bar, and space.	Understanding Different Types of File Formats
NoSQL databases	Databases are designed to store and manage unstructured data and provide analysis tools for examining this type of data.	Types of Data
Online Transaction Processing (OLTP) Systems	Systems that focus on handling business transactions and storing structured data.	Types of Data
Relational databases	Databases are designed to store structured data with well-defined schemas and support standard data analysis methods and tools.	Types of Data
Sensors	Devices such as Global Positioning Systems (GPS) and Radio Frequency Identification (RFID) tags generate structured data.	Types of Data

Spreadsheets	Software applications like Excel and Google Spreadsheets are used for organizing and analyzing structured data.	Types of Data
SQL Databases	Databases that use Structured Query Language (SQL) for defining, manipulating, and querying data in structured formats.	Types of Data
Tab-separated values (TSVs)	Delimited text files where the delimiter is a tab. Used as an alternative to CSV when literal commas are present in text data.	Understanding Different Types of File Formats

Term	Definition	Video where the term is introduced
ACID-compliance	Ensuring data accuracy and consistency through Atomicity, Consistency, Isolation, and Durability (ACID) in database transactions.	Relational Database Management System
Cloud-based Integration Platform as a Service (iPaaS)	Cloud-hosted integration platforms that offer integration services through virtual private clouds or hybrid cloud models, providing scalability and flexibility.	Data Integration Platforms
Column-based Database	A type of NoSQL database that organizes data in cells grouped as columns, often used for systems requiring high write request volume and storage of time-series or IoT data.	NoSQL
Data at rest	Data that is stored and not actively in motion, typically residing in a database or storage system for various purposes, including backup.	Considerations for Choice of Data Repository
Data integration	A discipline involving practices, architectural techniques, and tools that enable organizations to ingest, transform, combine, and provision data across various data types, used for purposes such as data consistency, master data management, data sharing, and data migration.	Data Integration Platforms
Data Lake	A data repository for storing large volumes of structured, semi-structured, and unstructured data in	Data Marts,

	its native format, facilitating agile data exploration and analysis.	Data Lakes, ETL, and Data Pipelines
Data mart	A subset of a data warehouse designed for specific business functions or user communities, providing isolated security and performance for focused analytics.	Data Marts, Data Lakes, ETL, and Data Pipelines
Data pipeline	A comprehensive data movement process that covers the entire journey of data from source systems to destination systems, which includes data integration as a key component.	Data Integration Platforms
Data repository	A general term referring to data that has been collected, organized, and isolated for business operations or data analysis. It can include databases, data warehouses, and big data stores.	Data Collection and Organization
Data warehouse	A central repository that consolidates data from various sources through the Extract, Transform, and Load (ETL) process, making it accessible for analytics and business intelligence.	Data Collection and Organization
Document-based Database	A type of NoSQL database that stores each record and its associated data within a single document, allowing flexible indexing, ad hoc queries, and analytics over collections of documents.	NoSQL
ETL process	The Extract, Transform, and Load process for data integration involves extracting data from various sources, transforming it into a usable format, and loading it into a repository.	Data Marts, Data Lakes, ETL, and Data Pipelines
Graph-based Database	A type of NoSQL database that uses a graphical model to represent and store data, ideal for visualizing, analyzing, and discovering connections between interconnected data points.	NoSQL
Key-value store	A type of NoSQL database where data is stored as key-value pairs, with the key serving as a unique identifier and the value containing data, which can be simple or complex.	NoSQL

Portability	The capability of data integration tools to be used in various environments, including single-cloud, multi-cloud, or hybrid-cloud scenarios, provides flexibility in deployment options.	Data Integration Platforms
Pre-built connectors	Cataloged connectors and adapters that simplify connecting and building integration flows with diverse data sources like databases, flat files, social media, APIs, CRM, and ERP applications.	Data Integration Platforms
Relational databases (RDBMSes)	Databases that organize data into a tabular format with rows and columns, following a well-defined structure and schema.	Data Collection and Organization
Scalability	The ability of a data repository to grow and expand its capacity to handle increasing data volumes and workload demands over time.	Considerations for Choice of Data Repository
Schema	The predefined structure that describes the organization and format of data within a database, indicating the types of data allowed and their relationships.	Considerations for Choice of Data Repository
Streaming data	Data that is continuously generated and transmitted in real-time requires specialized handling and processing to capture and analyze.	Considerations for Choice of Data Repository
Use cases for relational databases	Applications such as Online Transaction Processing (OLTP), Data Warehouses (OLAP), and IoT solutions where relational databases excel.	Relational Database Management System
Vendor lock-in	A situation where a user becomes dependent on a specific vendor’s technologies and solutions, making it challenging to switch to other platforms.	Data Integration Platforms