What is Big Data?

Objectives

After watching this video, you will be able to:

- Explain Big Data
- Identify the characteristics of Big Data
- Explain the five V's of Big Data

What is Big Data?

"The basic idea behind the phrase 'Big Data' is that everything we do is increasingly leaving a digital trace (or data), which we can use and analyze to become smarter. The driving forces in this brave new world are access to ever-increasing volumes of data and our ever-increasing technological capability to mine that data for commercial insights."

—Bernard Marr

Big Data vs. Small Data

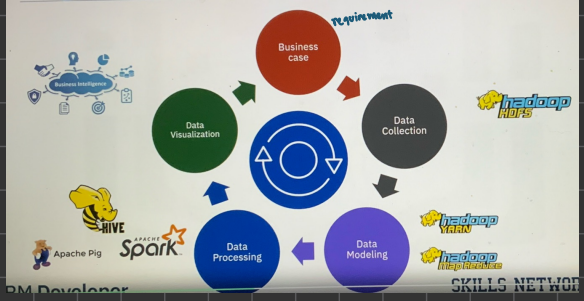
Small Data

- Small enough for human inference
- Accumulated slowly
- Relatively consistent and structured data usually stored in known forms such as JSON and XML
- Mostly located in storage systems within Enterprises or data centers

Big Data

- Data generated in huge volumes and could be structured, semi-structured, or unstructured
- Needs processing to generate insights for human consumption
- Arrives continuously at enormous speed from multiple sources
- Comprises any form of data including video, photos, and more
- Distributed on the cloud and server farms

Big Data life cycle



Characteristics of Big Data

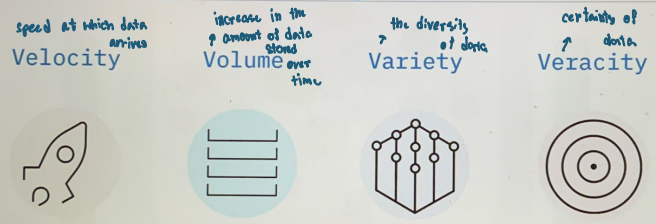
"Big Data is high-**volume**, high-**velocity** and/or high-**variety** information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation."

—Gartner

Bit, bytes and more

Bits = 0 or 1
8 Bits = 1 Byte
1024 Bytes= 1 Kilobyte(KB)
1024 KB = 1 Megabyte(MB)
1024 MB = 1 Gigabyte(GB)
1024 GB = 1 Terabyte(TB)
1024 TB = 1 Petabyte(PB)
1024 PB = 1 Exabyte(EB)
1024 EB = 1 Zettabyte(ZB)
1024 ZB = 1 Yottabyte(YB)

The four V's of Big Data



Velocity

Description

- Data that is generated fast
- Process that never stops

Attributes

- Batch
- Close to real time
- Streaming

Drivers

- Improved connectivity and hardware
- Rapid response times

Volume

Description

- Scale of data
- Increased amount of stored data

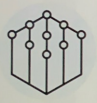
Attributes

- Petabytes
- Exa
- Zetta

Drivers

- Increase in data sources
- Higher resolution sensors
- Scalable infrastructure

Variety



Description

- Data that comes from machines, people, and processes
- Structured, semi-structured, and unstructured data

Attributes

- Structure, complexity, and origin

Drivers

- Mobile technologies
- Scalable infrastructure
- Resilience
- Fault recovery
- Efficient storage and retrieval

Veracity

Description

- Quality, origin, and conformity of facts
- Accuracy of data
- Data that comes from people and processes

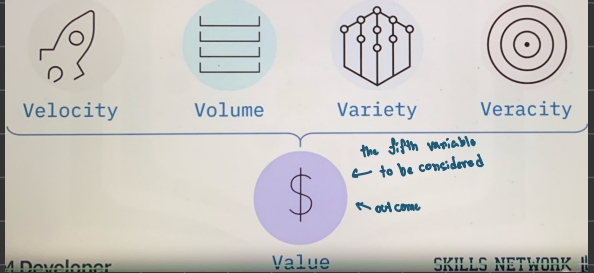
Attributes

- Consistency and completeness
- Integrity
- Ambiguity

Drivers

- Cost and traceability
- Robust ingestion
- ETL mechanisms

Adding the fifth V of Big Data



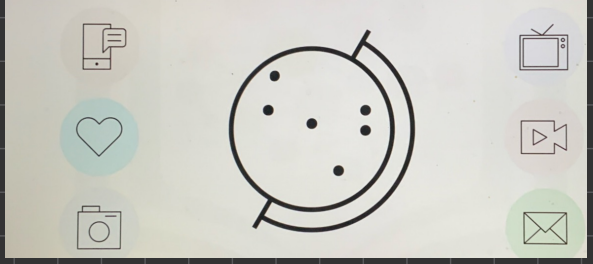
Impact of Big Data

Objectives

After watching this video, you will be able to:

- List examples of Big Data related technologies
- Explain the impact of Big Data on businesses and people
- Describe the Internet of Things (IoT) and its impact on Big Data

Generating and using Big Data



Big data in daily life

Recommendation engines use data from:

- Product searches
- Past orders
- Items in shopping carts
- Customer ratings and likes
- What other shoppers have looked at and bought

Amazon Netflix Spotify

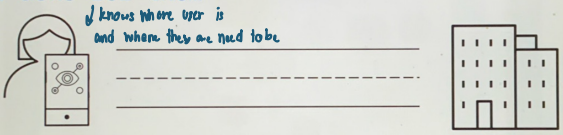
Big Data impacting people

Virtual personal assistants: (like Siri)

- Use Big Data to devise answers
- Use advanced neural networks
- Process speech into text
- Translate text into tasks

Big Data impacting people

- Google Now makes recommendations before the user asks for them (knows where user is and where they are need to be)



- Big Data forecasts future needs and behavior

The competitive advantage

"Data is the New Oil" —*Clive Humby (Chief Data Scientist, StartCount)*

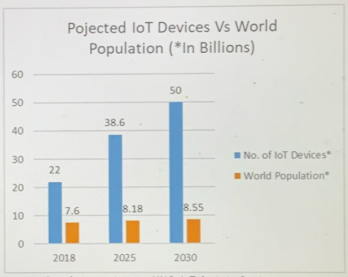
"Without big data, you are blind and deaf and in the middle of a freeway" — *Geoffrey Moore (Author and Consultant)*

- Powerful machine learning algorithms drive business decisions that increase efficiency
- Data scientists and Big Data engineers bring this value to companies

IoT – Internet of Things

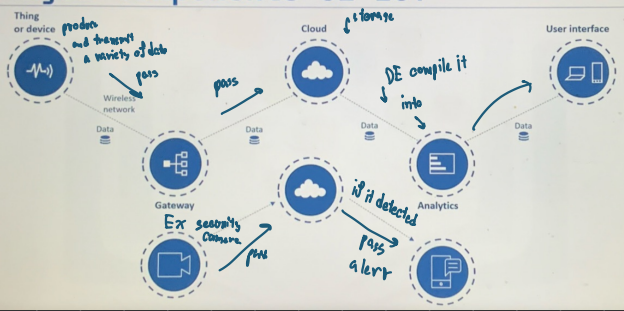
An Internet-enabled connected network of smart devices such as sensors, processors, embedded devices and communication hardware → These collect and transfer massive amounts of data over the Internet

Data collected, analyzed, and acted upon for benefits such as improving customer experience, enhanced productivity, and increased revenue



Population estimates: UNO, IoT devices: Statista.com

Major components of IoT



Parallel Processing, Scaling and Data Parallelism

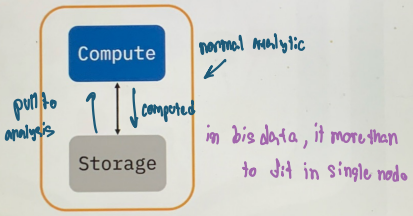
Objectives

After watching this video, you will be able to:

- Explain why Big Data requires parallel processing
- Identify the differences between linear and parallel processing
- Identify why parallel processing is apt for Big Data
- Explain parallel processing and scalability
- Describe the motivation for horizontal scaling
- Demonstrate "embarrassingly parallel"
- Explain fault tolerance in parallel computing

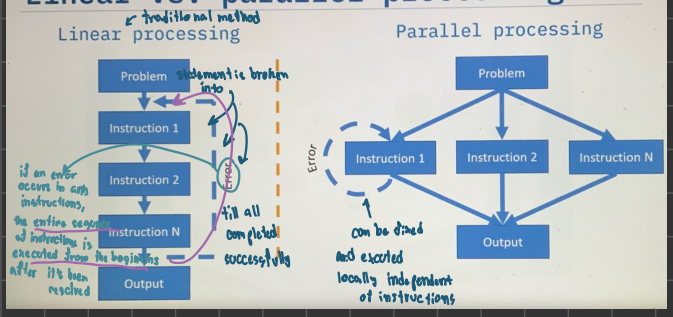
Big Data and parallel processing

Why parallel processing?



Single Node Capacity

Linear vs. parallel processing



Why parallel processing is apt for Big Data

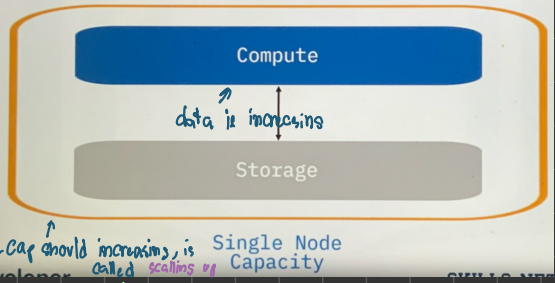
Parallel Processing advantages

- Parallel processing approach can process large datasets in a fraction of time → reduce processing time
- Less memory and compute requirements needed as set of instructions are distributed to smaller execution nodes
- More execution nodes can be added or removed from the processing network depending on complexity of the problem → flexibility

Data scaling in Big Data

to manage, store, and process overflow of data

What is data scaling?



Likewise, Cap should increasing, is called scaling up

Single Node Capacity

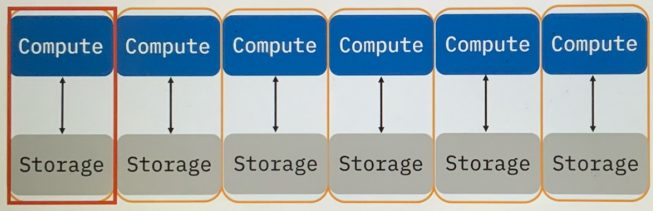
compute clusters can solve

Horizontal scaling

is adding additional nodes with the same cap until the problem is tractable

What people **really** mean when they say Big Data:

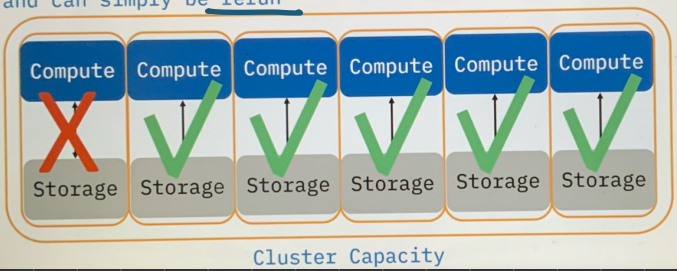
increase in this way are called a computer cluster



sorting a large data set adds significant complexity to the process

The "embarrassingly parallel"

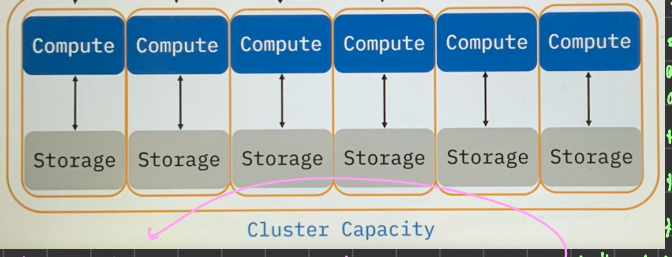
If any one process fails, it has no impact on the others and can simply be rerun



Cluster Capacity

The "not so easy" parallel

multiple computation must coordinate with one another. each process need to be aware of the state of its peer process



this requires sending messages across a network or writing them to a file system that is accessible to all processes

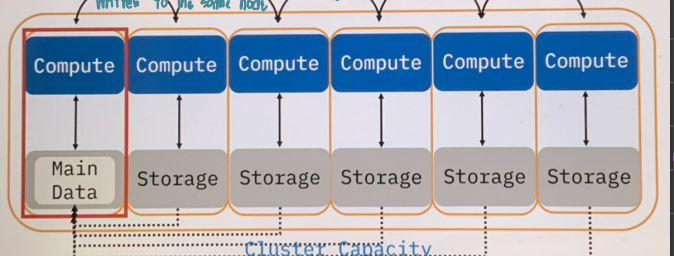
Cluster Capacity

the level of complexity increases significantly because you are asking on the cluster a cluster of computers to behave as a single computer

Data locality

Hadoop ecosystem the concept of "bringing compute to the data" is a central idea in the design of the cluster

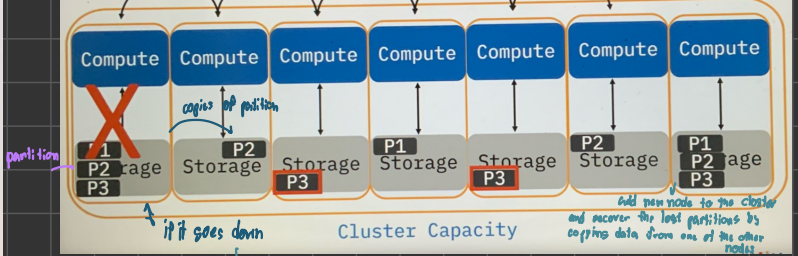
the resulting output will also be written to the same node



Cluster Capacity

Fault tolerance

refers ability of a system to continue operating without interruption when one or more of its components fail



Big Data Tools and Ecosystem

Objectives

After watching this video, you will be able to:

- Identify the key tooling categories within the Big Data ecosystem
- Describe the role of each tooling category in the Big Data life cycle
- List major tools and vendors within each Big Data tooling category

Categories of Big Data tooling

Data technologies

Data technologies allow enterprises to:

- Capture, process and share data at any scale and in any format
- Work with structured and unstructured data
- Leverage high-performance, parallel Big Data processing

Key technologies include Hadoop, HDFS, Spark, MapReduce, Cloudera, and Databricks

Analytics and Visualization

- Big Data analytics examines large amounts of data (uncover hidden patterns, correlations, and trends)
- Analyzed data is visualized using graphs, charts, and maps (to understand data insight)

Popular analytics tools available are Tableau, Palantir, SAS, Pentaho, and Teradata

Business Intelligence

- Business intelligence (BI) offers a range of tools that provide a quick and easy way to transform data into actionable insights
- Such insights inform an organization's strategic and tactical business decisions

Examples include Cognos, Oracle, PowerBI, Business Objects, and Hyperion

Cloud providers

(also provide "software as a service" models with point solutions to easily aggregate, process, and visualize data)

Offer fundamental infrastructure and support with shared resources including computing, storage, networking, and analytical software

AWS, GCP, IBM and ORACLE

NoSQL databases

NoSQL databases are best suited for Big Data processing:

- Store and process vast amounts of data at scale (built from the ground up)
- Store information in JSON documents instead of relational tables
- NoSQL database types include pure document databases, key-value stores, wide-column databases, and graph databases

Examples include MongoDB, CouchDB, Cassandra, Redis

Programming tools

Big Data programming tools:

- Perform large-scale analytical tasks and operationalize Big Data
- Provide all necessary functions for the Big Data life cycle (collection, cleaning, exploration, modeling, and visualization)

R, Python, SQL, Scala, and Julia are common programming tools

Open Source and Big Data

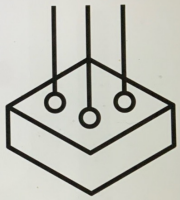
Objectives

After watching this video, you will be able to:

- Explain the role of open source in Big Data
- Describe platforms for coordinating open source
- Describe the most popular open source frameworks

What is an open source software?

- Free
- Source code is open, for review, to use, or re-use as needed in other projects
- Also allows any user to propose changes to the project



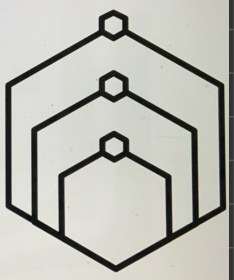
Open source license types:

- Public domain
- Permissive
- Copyleft
- Lesser General Public License

Open source for Big Data

Why is the open source model used for Big Data?

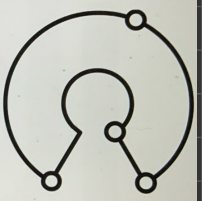
- Large and complex projects
- Projects serve the needs of many organizations
- Open source model has emerged as the best development strategy



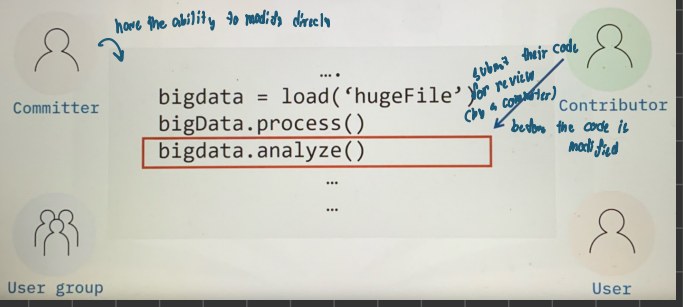
The lifeblood of Big Data

"I think, fundamentally, open source does tend to be a more stable software. It's the right way to do things." –Jim Whitehurst

"Open source isn't about saving money, it's about doing more stuff, and getting incremental innovation with the finite budget you have." –Jim Whitehurst



The community garden of code



Open source platforms

- The Apache Software Foundation
- Linux Foundation
- Cloud Native Computing Foundation

Open source in Big Data

Hadoop plays a major role in open source Big Data projects:

- Hadoop MapReduce
- Hadoop File System (HDFS)
- Yet Another Resource Negotiator (YARN)

MapReduce

HDFS

YARN

Open source tools in Big Data

- Hive
- Apache Spark
- Apache Hbase
- HDP

Beyond the Hype

Objectives

After watching this video, you will be able to:

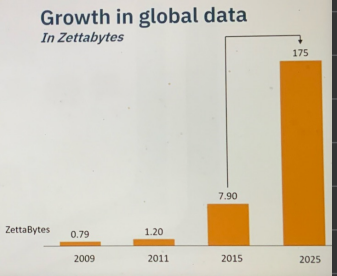
- Describe facts about Big Data
- List the key sources of Big Data
- Explain different types of Big Data
- Describe the contribution of cloud computing in Big Data

The hype about Big Data

Facts:

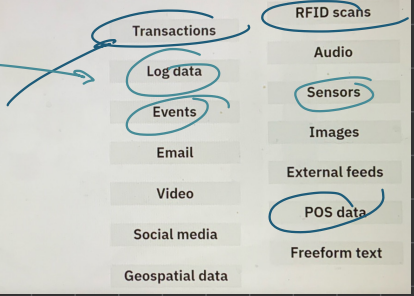
- More data has been created in the past two years than in the entire previous history of humankind
- 40% projected growth in global data generated per year
- The amount of digital data created over the next five years will be greater than twice the amount of data created since the advent of digital storage

Sources: McKinsey report – Digital era, data growth, estimates and, Gartner, IDC estimates



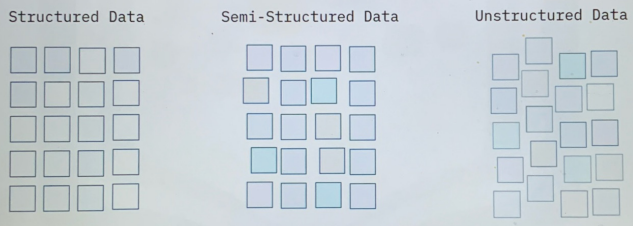
Where does Big Data come from?

- Social data
- Machine data
- Transactional data



Types of Big Data

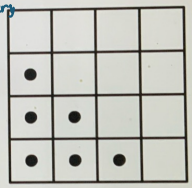
- Structured, unstructured, semi-structured



Sources: Structured

- Relational databases (SQL)
- Spreadsheets

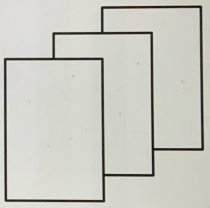
the knows its relationship with other data, what it means easy to query



Sources: Semi-structured

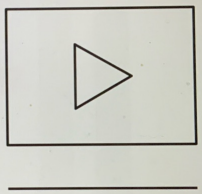
- XML
- JSON

use tags or other markers to enforce hierarchies of records and field within data.



Explosion in unstructured data

- Video production
- Social media
- Internet speeds



Advances in cloud computing

- Advances in cloud computing have contributed to the increasing potential of Big Data
- Cloud computing allows users to access highly scalable computing and storage resources through the Internet
- Organizations can expand server capacity to the large scale required to process Big Data

Big Data Use Cases

Objectives

After watching this video, you will be able to:

- Identify industries where Big Data is leveraged
- Discuss and outline the usage of Big Data in key industries

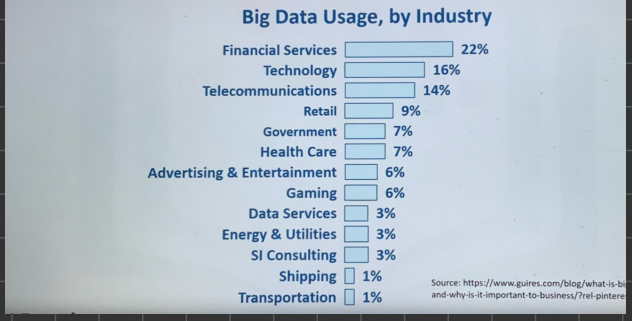
Industries leveraging Big Data

- Data driven companies have a competitive edge over peers
- Leverage data insights to improve decision making, enter new markets, and enhance customer experiences

Industries using Big Data

- Retail
- Finance
- Manufacturing
- Automobile
- Telecom
- Insurance

Industry-wise usage of Big Data



Big Data in retail

Price Analytics

- Understand market segmentation, identify best price points, and perform margin analysis

Sentiment Analysis

- Leverage social media to gauge customer perception and devise an effective marketing strategy → to connect with the customers

Big Data in insurance

Fraud Analytics

- Spot fraudulent claims and detect anomalies and prevent suspicious activities

Risk Assessment

- Use predictive modeling to identify high-risk customers of getting into accidents

Big Data in telecommunications

- Improved network security
  ↳ machine learning, identifies any threats and predictive maintenance resolve these threats
- Contextualized location-based promotions
- Real-time network analytics
- Optimized pricing ~ promotion

Big Data in manufacturing

Predictive Maintenance

- Analyze equipment usage patterns and predict equipment failure, maintenance requirements, and parts replacement

Production Optimization by AI

- Understand the production lines, analyze increased production time to recommend optimizations

Big Data in automotive industry

Predictive Support

- Predict malfunctions, pre-emptive ordering of parts and repair suggestions

Connected self-driven cars

- Real-time data analysis for autonomous driving real time adjustments and direction based on the data

Big Data in finance

Customer Segmentation

offer deeply personalized and targeted solutions to their customers

- Products offered based on customer demographics, transaction frequency, behavior patterns, social media posts and interactions with customer service systems

Algorithmic Trading

- Make machine learning-based decisions using instant analysis of large quantities of data

Term	Definition
Apache Spark	An open-source, in-memory application framework used for distributed data processing and iterative analysis of large data sets.
Apache HBase	A robust NoSQL datastore that efficiently manages storage and computation resources independently of the Hadoop ecosystem.
Business intelligence (BI)	Encompasses various tools and methodologies designed to convert data into actionable insights efficiently.
Big data	Data sets whose volume, velocity, or variety exceeds the capacity of conventional relational databases to effectively manage, capture, and process with minimal latency. Key characteristics of big data include substantial volume, high velocity, and diverse variety.
Big data analytics	Uses advanced analytic techniques against large, diverse big data sets that include structured, semi-structured, and unstructured data from different sources and sizes, from terabytes to zettabytes. It helps companies gain insights from the data collected by IoT devices.
Big data programming tools	Programming tools are the final component of big data commercial tools. These programming tools perform large-scale analytical tasks and operationalize big data. They also provide all necessary functions for data collection, cleaning, exploration, modeling, and visualization. Some popular tools you can use for programming include R, Python, SQL, Scala, and Julia.
Committer	Most open-source projects have formal processes for contributing code and include various levels of influence and obligation to the project: Committer, contributor, user, and user group. Typically, committers can modify the code directly.
Cloud computing	Allows customers to access infrastructure and applications over the internet without needing on-premises installation and maintenance. By leveraging cloud computing, companies can utilize server capacity on-demand and rapidly scale up to handle the extensive computational requirements of processing large data sets and executing complex mathematical models.
Cloud providers	Offer essential infrastructure and support, providing shared computing resources encompassing computing power, storage, networking, and analytical software. These providers also offer software as a service model featuring specific solutions, enabling enterprises to gather, process, and visualize data efficiently. Prominent examples of cloud service providers include AWS, IBM, GCP, and Oracle.
Extract, transform, and load (ETL) process	A systematic approach that involves extracting data from various sources, transforming it to meet specific requirements, and loading it into a data warehouse or another centralized data repository.
Hadoop	An open-source software framework that provides dependable distributed processing for large data sets through the utilization of simplified programming models.
Hadoop Distributed File System (HDFS)	A file system distributed on multiple file servers, allowing programmers to access or store files from any network or computer. It is the storage layer of Hadoop. It works by splitting the files into blocks, creating replicas of the blocks, and storing them on different machines. It is built to access streaming data seamlessly. It uses a command-line interface to interact with Hadoop.
Hive	A data warehouse infrastructure employed for data querying and analysis, featuring an SQL-like interface. It facilitates report generation and utilizes a declarative programming language, enabling users to specify the data they want to retrieve.
Internet of Things (IoT)	A system of physical objects connected through the internet. A thing or device can include a smart device in our homes or a personal communication device such as a smartphone or computer. These collect and transfer massive amounts of data over the internet without manual intervention by using embedded technologies.
Machine data	Refers to information generated by various sources, including the Internet of Things (IoT) sensors embedded in industrial equipment, as well as weblogs that capture user behavior and interactions.
Map	MapReduce converts a set of data into another set of data, and the elements are fragmented into tuples (key or value pairs).
MapReduce	A program model and processing technique used in distributed computing based on Java. It splits the data into smaller units and processes big data. It is the first method used to query data stored in HDFS. It allows massive scalability across hundreds or thousands of servers in a Hadoop cluster.
NoSQL databases	NoSQL databases are built from the ground up to store and process vast amounts of data at scale and support a growing number of modern businesses. NoSQL databases store data in documents rather than relational tables. Types of NoSQL databases include pure document databases, key-value stores, wide-column databases, and graph databases such as MongoDB, CouchDB, Cassandra, and Redis.
Open-source software	Not only is the runnable version of the code free, but the source code is also completely open, meaning that every line of code is available for people to view, use, and reuse as needed.
Price analytics	Helps understand market segmentation, identify the best price points for a product line, and perform margin analysis for maximum profitability.
Relational databases	Data is structured in the form of tables, with rows and columns, collectively forming a relational database. These tables are interconnected using primary and foreign keys to establish relationships across the data set.
Sentiment analysis	Utilizes social media conversations to gain insights into consumer opinions about a product. It is used to develop effective marketing strategies and establish customer connections based on their sentiments and preferences.
Social data	Comes from the likes, tweets and retweets, comments, video uploads, and general media that are uploaded and shared via the world's favorite social media platforms. Machine-generated data and business-generated data are data that organizations generate within their own operations.
Transactional data	Generated from all the daily transactions that take place both online and offline, such as invoices, payment orders, storage records, and delivery receipts.
Velocity	The speed at which data arrives. Velocity is one of the four main components used to describe the dimensions of big data.
Volume	The increase in the amount of data stored over time. Volume is one of the four main components used to describe the dimensions of big data.
Variety	The diversity of data or the various data forms that need to be stored. Variety is one of the four main components used to describe the dimensions of big data.
Veracity	The certainty of data, as with a large amount of data available, makes it difficult to determine if the data collected is accurate. Veracity is one of the four main components used to describe the dimensions of big data.
Yet Another Resource Negotiator (YARN)	Serves as the resource manager bundled with Hadoop and is typically the default resource manager for numerous big data applications, such as HIVE and Spark. While it remains a robust resource manager, it's important to note that more contemporary container-based resource managers, such as Kubernetes, are gradually emerging as the new standard practices in the field.