WHAT IS BIG DATA ANALYTICS:

Big data analytics is the process of examining big data to uncover information -- such as hidden patterns, correlations, market trends and customer preferences -- that can help organizations make informed business decisions. This process allows organizations to leverage the exponentially growing data generated from diverse sources, including internet-of-things (IoT) sensors, social media, financial transactions and smart devices to derive actionable intelligence through advanced analytic techniques.

On a broad scale, data analytics technologies and techniques enable organizations to analyze data sets and gather new information. Big data analytics is a form of advanced analytics that involves more complex methods that include elements such as predictive models, statistical algorithms and what-if analysis powered by analytics systems.

An example of big data analytics can be found in the healthcare industry, where millions of patient records, medical claims, clinical results, care management records and other data must be collected, aggregated, processed and analyzed. Traditional data analysis methods can't support this level of complexity at scale, leading to the need for big data analytics systems.

Big data analytics is also commonly used for accounting, decision-making, predictive analytics and many other purposes. The data found in big data analytics varies greatly in type, quality and accessibility, presenting significant challenges but also offering tremendous benefits.

Benefits of big Data Analytics:

1. **More Complete Insights** – Big Data integrates vast amounts of structured and unstructured data, providing deeper and more accurate insights for informed decision-making.
2. **Enhanced Decision-Making** – Data-driven strategies improve confidence, enabling businesses to make proactive and well-informed choices rather than relying on intuition.
3. **Operational Efficiency** – Analytics help optimize workflows, reduce costs, and enhance productivity by identifying inefficiencies and improving resource utilization.
4. **Competitive Advantage** – Predictive analytics allow businesses to stay ahead of market trends, personalize customer experiences, and adapt to changing demands.
5. **Risk Management & Compliance** – Big Data helps detect fraud, assess risks, and ensure regulatory compliance, fostering security and transparency in operations.

TYPES OF DIGITAL DATA

Structured Data

Structured data is a type of data that is organized and easily managed using traditional data management tools such as spreadsheets, databases, or tables. Structured data is typically quantitative and numeric in nature, meaning that it consists of numbers, percentages, and other numerical values. Because of its organized nature, structured data is relatively easy to analyze using statistical methods such as regression analysis or correlation analysis.

Characteristics of Structured Data:

- Structured data is organized in a clear, predefined way, like rows and columns in a table. Think of it as a spreadsheet with organized boxes of information.
- It's easy to search for specific information because it follows a set pattern, and you can use tools like **SQL** (Structured Query Language) to ask questions or get results.
- The data is neatly arranged, making it quick to find what you need. You know exactly where to look for certain pieces of information.
- Structured data is usually stored in **databases** like **MySQL**, **PostgreSQL**, **Oracle**, or in **data warehouses** like **Amazon Redshift** or **Google BigQuery**.

Examples of Structured Data :

- **Financial Data:** Information like bank account balances, transaction histories, and stock prices. For example:
  - Stored in **CSV** (Comma-Separated Values) files or **SQL databases**.
  - Example: A list of transactions with dates, amounts, and account numbers in a table.
- **Sales Data:** Information on products sold, quantities, and prices. For example:
  - Stored in **Excel**, **CSV**, or **SQL databases**.
  - Example: A table listing products, the number sold, and the price per product..

Unstructured Data

Unstructured data is data that does not have a predefined format or organization, making it difficult to manage using traditional data management tools. Examples of unstructured data include social media posts, emails, images, and videos. Because of its unstructured nature, unstructured data is typically qualitative in nature, meaning that it is descriptive and narrative in nature. Analyzing unstructured data requires the use of advanced analytics techniques such as natural language processing (NLP) or sentiment analysis.

Characteristics of Unstructured Data (Simple Language)

- Unstructured data does not follow a set structure like tables or rows. It can be in the form of text, images, videos, or audio files.

- Since it is not organized in a clear way, it requires advanced tools like **Artificial Intelligence (AI)**, **Machine Learning (ML)**, and **Natural Language Processing (NLP)** to search, analyze, and process it.
- It can exist as **text**, **videos**, **audio files**, **images**, and more.
- Unlike structured data, it is stored in **NoSQL databases** (e.g., **MongoDB**, **Cassandra**) or in **data lakes** and **distributed file systems** (e.g., **Hadoop HDFS**).

Examples of Unstructured Data

- **Social Media Content:** Posts, tweets, comments, and videos shared on platforms like **Twitter, Facebook, YouTube**.
- **Customer Feedback:** Open-ended survey responses, live chat messages, and product reviews on shopping websites.

Semi-Structured Data

Semi-structured data is a type of data that has elements of both structured and unstructured data. This type of data includes information that is partially organized, but not to the extent that it can be classified as structured data. Examples of semi-structured data include XML and JSON files, which have some organization but also contain elements of unstructured data. Analyzing semi-structured data typically requires a combination of traditional data management tools and advanced analytics techniques.

Characteristics of Semi-Structured Data

- It has some structure, like labels or tags, but it is not arranged neatly in rows and columns like structured data.
- It is not as rigid as structured data but still has more organization than unstructured data.
- It is often stored in formats like **XML**, **JSON**, and **YAML**.
- Typically kept in **NoSQL databases** (e.g., **MongoDB, Cassandra**) or in **data lakes**.
- Can be analyzed using **Hadoop**, **Spark**, and other **Big Data tools** that handle semi-structured data.

Examples of Semi-Structured Data

- **Log Files:** Web server logs with **timestamps, IP addresses, and error messages** that provide some structure.
- **Geo-Tagged Data:** Photos or videos with extra details like **location, date, and camera settings** stored in the metadata.

FOUR MAIN TYPES OF DATA ANALYTICS

Before launching a data analytic effort, companies need to decide what they want to achieve: Do you have historical data to mine, to understand trends and patterns? Are you looking to make predictions, maybe even recommend actions to achieve desired results? Each type of data analytics serves a purpose and requires specific tools and techniques to succeed.

DESCRIPTIVE ANALYTICS

Descriptive analytics focuses on answering the question, **‘**What is happening?’ or ‘What has happened?’ by analyzing past data. Of all the types of data analytics, this is the most straightforward approach as it summarizes and simplifies the main features and characteristics of complex datasets through interactive visualizations.

How does descriptive analytics work?

To deploy descriptive analytics, analysts and data scientists typically follow five steps:

1. Define clear objectives or areas of interest you want to explore in the dataset.
2. Gather data relevant to your dataset.
3. Clean data for accurate results.
4. Apply advanced statistical techniques to condense large datasets into concise summaries.
5. Use cutting-edge BI tools like ThoughtSpot to run real-time data analysis, create interactive visualizations, and gain AI-assisted insights to make informed decisions.

Descriptive analytics examples

Descriptive analytics offers organizations a data-driven perspective on their operations. It’s no wonder that more businesses across every industry are starting to adopt this powerful approach to problem-solving. Let’s explore some popular descriptive analytics use cases:

1. **Customer relationship management:** With descriptive analytics, you can analyze customer data, including past surveys and feedback forms, to get insights into how customers interact with your products or services and what they want from your product. Armed with this information, you can adjust your offerings and optimize touchpoints across all channels.
2. **Financial reporting:** Finance teams often leverage descriptive analytics to summarize large volumes of transactional data, compare performance for optimal resource allocation, and assess the organization’s financial health.
3. **Supply chain management:** By monitoring supply chain metrics and assessing historical lead times, supply chain leaders can better understand the root causes for delays, minimize disruptions, and identify cost-saving opportunities.

PREDICTIVE ANALYTICS

Predictive analytics uses historical data to answer the question, ‘What may happen next?’ Businesses employ this model to predict future outcomes, find patterns, and identify risks or growth opportunities. While descriptive analytics serves as a reflective mirror, showing us a holistic picture of our past activities, predictive analytics acts as a crystal ball, providing a sneak peek into the future.

How does predictive analytics work?

Predictive analytics uses statistical modeling, data mining techniques, and machine learning to analyze large datasets and predict the likelihood of an event occurring. To apply this types of data analytics, you need to build a model and choose the correct analytical technique, depending on the problem to be solved and the nature of the dataset.

To deploy predictive analytics, here are the six steps you should follow:

1. Define the business problem.
2. Gather relevant data for analysis.
3. Pre-process the data to remove missing values.
4. Choose the right modeling predictive technique. Some of the most common types of analytics models include linear regression, decision trees, cluster models, and time series models.
5. Develop and train the model leveraging machine learning, AI, and advanced statistical techniques.
6. Test your machine learning model on a test dataset to assess how accurately the model analyzes new, unseen data.

Predictive analytics examples

Predictive analytics can help you boost operational efficiency, identify growth opportunities, and manage risks effectively across a wide range of industries. This includes sectors, such as banking, retail, utilities, public services, healthcare, and manufacturing. Here are some popular use cases:

1. **Inventory management:** Retailers and e-commerce businesses use this type of analytics to anticipate consumer demand for products and perform sales forecasts. By analyzing past sales data, critical retail KPIs, and market data, you can optimize inventory levels, manage resources, prevent stock-outs, and operate more efficiently.
2. **Credit scoring:** Banking and financial institutions leverage predictive analytics to make predictions about a customer’s ability to repay a loan from their past credit history. The

insights help decision-makers make data-driven decisions on who is likely to default on a loan and which customers pose a high credit risk.

3. **Sales forecasting:** Sales teams use predictive analytics to estimate future sales revenue and customer demand based on historical data, market analysis, and other relevant factors. Doing so helps businesses create personalized marketing campaigns, adjust product launches, and capitalize on opportunities that can drive sales.

PRESCRIPTIVE ANALYTICS

Unlike predictive analytics, which focuses on future outcomes, prescriptive analytics helps decision-makers identify the best course of action to help them achieve their business goals. The primary goal of this model is to answer the question: ‘What should we do?’.

How does prescriptive analytics work?

Prescriptive analytics relies on advanced techniques, such as machine learning, neural networks, recommendation engines, and mathematical algorithms, to offer actionable guidance on what to do next. Of all the types of types of data analytics, this model factors in past and *present* data to simulate scenarios and offer unbiased recommendations.

Here are six steps you should follow to deploy prescriptive analytics:

1. Define the business problem you want to address.
2. Gather relevant data for accurate results.
3. Develop the model using AI, machine learning, and advanced statistical techniques.
4. Test and train the model to check accuracy.
5. Map the outcomes and check whether they align with your business objective.
6. Monitor and adjust the model accordingly.

Prescriptive analytics examples

Advanced BI tools and prescriptive analytics have made it possible for businesses to analyze large datasets and make in-the-moment decisions. Here’s how different industries are using prescriptive analytics to get the most value out of their data:

1. **Handling multiple IT requests:** To keep up with the constant barrage of security and IT tickets, businesses need the right analytics capabilities to make the correct operational decisions. With an intuitive BI tool like ThoughtSpot, you can schedule alerts to quickly identify the problem. Consequently, you can apply prescriptive analytics models to get guidance about the next step.

This is exactly how data cloud leader Snowflake cleared its IT backlog by 20%—read the entire case study here.

1. **Employee training programs:** Prescriptive analytics models are helping HR leaders bridge skills gaps within their workforce. The model can recommend personalized learning paths for individual employees based on their roles and job responsibilities, ensuring everyone has the skills they need to excel.
2. **Campaign optimization:** As competition becomes fierce, it is critical to create campaigns that stand out. With prescriptive analytics, you can optimize your ongoing campaigns by getting insights into which channels are effective, what timing is the best, and which messaging will appeal to your target audience.

DIAGNOSTIC ANALYTICS

Diagnostic analytics examines past data to identify the root causes behind a particular outcome. This type of analytics aims to answer the question, ‘Why did this happen?’ It focuses on uncovering insights into historical data patterns, anomalies, and correlations to facilitate a deeper understanding of a particular business problem.

How does diagnostic analytics work?

Diagnostic analytics explores the relationships between variables and uses advanced statistical methods to pinpoint the root causes of specific events.

Here’s how this type of data analytics technique works:

1. Define the problem.
2. Collect relevant data for analysis.
3. Pre-process the data for accuracy.
4. Decide the right type of analytics technique. Some of the popular ones are: Hypothesis testing, anomaly detection, Root cause analysis, and regression analysis.
5. Apply statistical methods.
6. Leverage BI tools to visualize data.
7. Interpret data and share your findings.

Diagnostic analytics examples

Diagnostic analytics helps organizations dig deeper and find areas for improvement, implement changes, and optimize business processes. Here are some known use cases of diagnostic analytics across industries:

- **Optimizing patient treatments:** Diagnostic analytics can help healthcare providers analyze patient records and identify patterns leading to specific medical conditions or treatment outcomes. They can understand factors contributing to patient readmissions and identify areas for improvement in healthcare delivery.
- **Customer retention:** Diagnostic analytics can help you address customer churn by examining historical data to identify the reasons behind customer attrition. This includes identifying engagement patterns, analyzing channel performance, and pinpointing drop-off points.
- **Website traffic:** If you notice a sudden drop in website traffic, it is critical to pinpoint the exact cause so you don't lose prospects. Using this type of analytics, you can analyze data on user behavior on the website to identify and resolve the issue.

FEATURES OF BIG DATA ANALYTICS

5'V of Big Data Analytics

Volume

Volume refers to the colossal amount of data that inundates organizations. We're well past the days when companies resourced their data internally and stored it in local servers. Companies of 15 years ago handled terabytes of data.

Today, data has grown to petabytes if not exabytes of bytes (that's 1,000–1 million TB) that come from sources such as transaction processing systems, emails, social networks, customer databases, website lead captures, monitoring devices and mobile apps.

To handle all of this data, managers use data lakes and warehouses or data management systems. They store it on clouds or use service providers such as Google Cloud. And as global data grows from two zettabytes at the beginning of the decade to 181 zettabytes a day by 2025, even these may be insufficient.

An example of data volume

Walmart operates approximately 10,500 stores in 24 countries, handling more than 1 million customer transactions every hour. The result? Walmart imports more than 2.5 petabytes of data per hour, storing it internally on what happens to be the world's biggest private cloud.

Velocity

Big data grows fast. Consider that, according to Zettaspere, there are around 3,400,000 emails, 4,595 SMS, 740,741 WhatsApp messages, almost 69,000 Google searches, 55,000 Facebook posts, and 5,700 tweets made per minute.

Around five years ago, data scientists measured incoming data with computerized batch processing that read large files and generated reports. Today, batch processes are unable to handle the continuous rush of real-time data from a growing number of sources.

More critical still, data ages fast. As Walmart's former senior statistical analyst Naveen Peddamail said, "If you can't get insights until you've analyzed your sales for a week or a month, then you've lost sales within that time."

Competitive companies need some capable business intelligence (BI) tools to make timely decisions.

An example of data velocity

Using real-time alerting, Walmart sales analysts noted that a particular, rather popular, Halloween novelty cookie was not selling in two stores. A quick investigation showed that, due to a stocking oversight, those cookies hadn't been put on the shelves. By receiving automated alerts, Walmart was quickly able to rectify the situation and save its sales.

Variety

Variety refers to the different types of digitized data that inundate organizations and how to process and mine these various types of data for insights. At one time, organizations mostly gained their information from structured data that fit into internal databases like Excel.

Today, you also have unstructured information that evades management and comes in diverse forms such as emails, customer comments, SMS, social media posts, sensor data, audio, images, and video. Companies struggle with digesting, processing, and analyzing this type of data and doing so in real time.

An example of data variety

Walmart tracks each one of its 145 million American consumers individually, resulting in accrued data per hour that's equivalent to 167 times the books in America's Library of Congress. Most of that is unstructured data that comes from its videos, tweets, Facebook posts, call-center conversations, closed-circuit TV footage, mobile phone calls and texts, and website clicks.

How software can help: Walmart uses a 250-node Hadoop. For small to midsize companies, Tableau is ideal since it is also designed for non-technical users.

Veracity

Veracity is arguably the most important factor of all the five Vs because it serves as the premise for business success. You can only generate business profit and impact change with thorough and correct information.

Data can only help organizations if it's clean. That's if it's accurate, error-free, reliable, consistent, bias-free, and complete. Contaminating factors include:

- Statistical data that misrepresents the information of a particular market
- Meaningless information that creeps into and distorts the data
- Outliers in the dataset that make it deviate from the normal behavior
- Bugs in software that produce distorted information
- Software vulnerabilities that could cause bad actors to hack into and hijack data
- Human agents that make mistakes in reading, processing, or analyzing data, resulting in incorrect information

Example of data veracity

According to Jaya Kolhatkar, vice president of global data for Walmart labs, Walmart's priority is making sure its data is correct and of high quality. Clean data helps with privacy issues, ensuring sensitive details are encrypted while customer contact information is segregated.

How software can help: Multilingual and scalable Apache Spark is good for quick queries across data sizes. However, it's expensive and has latency issues.

Value

Big data is the new competitive advantage. But, that's only if you convert your information into useful insight.

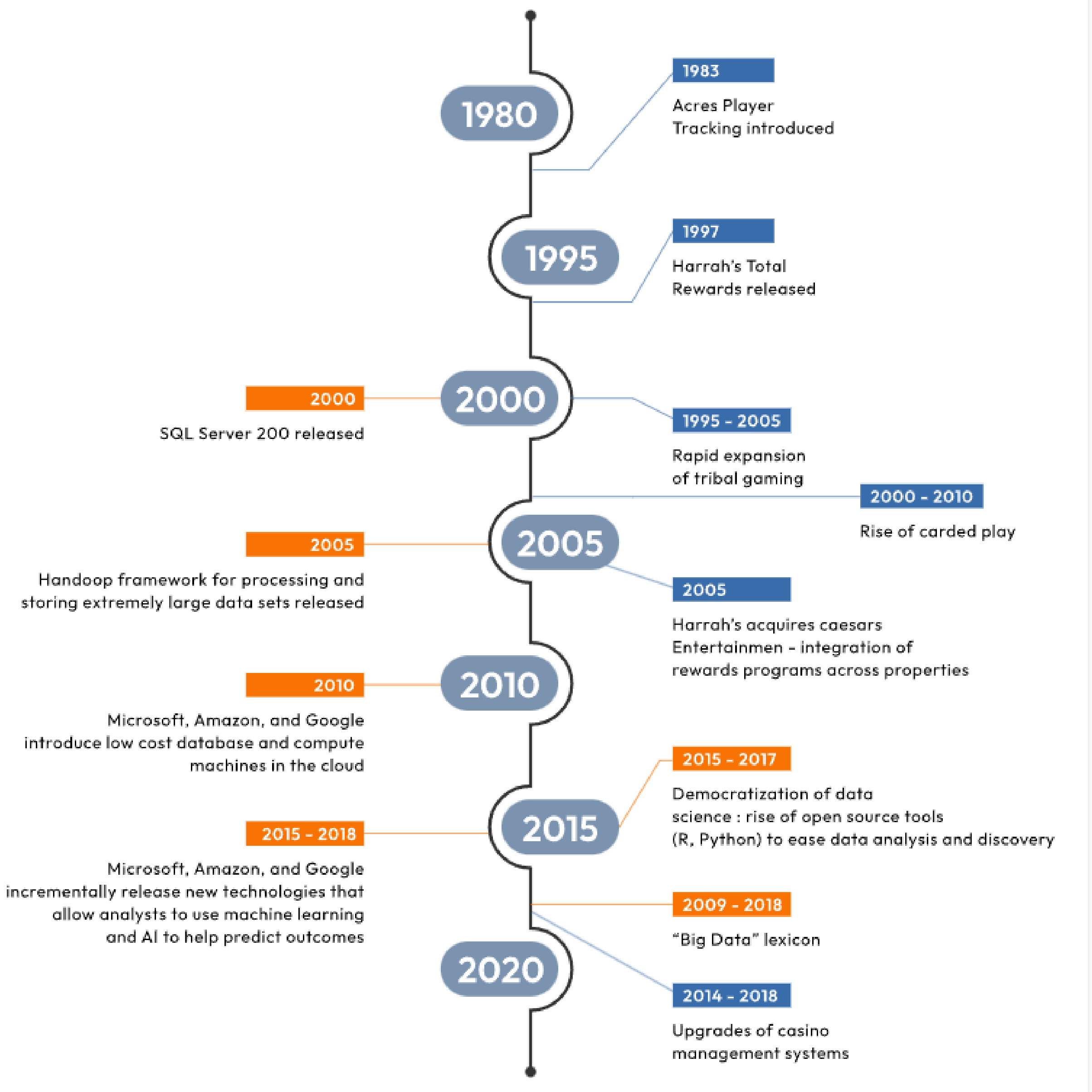
Users can capture value from that data through:

- Making their enterprise information transparent for trust
- Making better management decisions by collecting more accurate and detailed performance information across their business
- Fine-tuning their products or services to narrowly segmented customers
- Minimizing risks and unearthing hidden insights
- Developing the next generation of products and services

Example of data value

Walmart uses its big data to make its pharmacies more efficient, help it improve store checkout, personalize its shopping experience, manage its supply chain, and optimize product assortment among other ends

EVOLUTION OF BIG DATA


1980
1983
Acres Player
Tracking introduced
1995
1997
Harrah's Total
Rewards released
2000
2000
SQL Server 200 released
1995 - 2005
Rapid expansion
of tribal gaming
2000 - 2010
Rise of carded play
2005
2005
Handoop framework for processing and
storing extremely large data sets released
2005
Harrah's acquires caesars
Entertainmen - integration of
rewards programs across properties
2010
2010
Microsoft, Amazon, and Google
introduce low cost database and compute
machines in the cloud
2015
2015 - 2017
Democratization of data
science : rise of open source tools
(R, Python) to ease data analysis and discovery
2015 - 2018
Microsoft, Amazon, and Google
incrementally release new technologies that
allow analysts to use machine learning
and AI to help predict outcomes
2009 - 2018
"Big Data" lexicon
2020
2014 - 2018
Upgrades of casino
management systems

Big Data has transformed the way we analyze and interpret vast amounts of information. Emerging from the rise of the internet and digital technologies, Big Data represents the massive volumes of structured and unstructured data generated daily. This evolution began with the advent of digital storage and the development of sophisticated data analytics tools.

Over time, advancements in cloud computing, artificial intelligence, and machine learning have further enhanced our ability to process and analyze Big Data, leading to insights that drive innovation across various industries, from healthcare and finance to marketing and beyond.

The Advent of Digital Storage

The first step in the evolution of Big Data was the shift from analog to digital storage. As businesses and individuals started to store data digitally, the volume of available information began to grow exponentially. This transition laid the groundwork for the development of data analytics tools that could handle increasingly large datasets.

Emergence of Data Analytics Tools

As digital data grew, there was a pressing need for tools that could process and analyze this information efficiently. The development of data analytics tools, such as Hadoop and Spark, allowed businesses to harness the power of Big Data, uncovering trends and insights previously hidden within vast datasets.

Rise of Cloud Computing

Cloud computing has been a game-changer in the evolution of Big Data. By providing scalable storage and computing resources, cloud platforms have made it easier for businesses to store and process large datasets without the need for extensive physical infrastructure. This accessibility has democratized data analytics, enabling even small businesses to leverage Big Data for strategic decision-making.

Impact of Artificial Intelligence and Machine Learning

Artificial intelligence (AI) and machine learning (ML) have significantly advanced Big Data analytics. These technologies enable the automation of data analysis, uncovering complex patterns and predictions that were once the domain of human experts. AI and ML have expanded the possibilities of Big Data, driving innovation in areas such as personalized medicine, predictive maintenance, and targeted marketing.

Overview of SQL vs. NoSQL vs. NewSQL

SQL databases, also known as Relational Database Management Systems (RDBMS), follow a structured approach to storing and managing data using tables. They have been a reliable choice for handling historical data efficiently. However, as data grew in volume and complexity, traditional SQL databases started facing challenges in processing large amounts of unstructured information.

With the rise of Big Data, SQL databases struggled to provide fast performance and scalability. While improvements were made over time, they still had limitations in handling massive datasets that required quick responses and high scalability. This led to the introduction of NoSQL databases, which were designed to work efficiently with unstructured data.

NoSQL databases use different models such as key-value pairs, documents, graphs, or wide-column stores. Unlike SQL, they do not require a fixed schema, making them more flexible. They are also horizontally scalable, meaning they can distribute data across multiple servers, unlike SQL databases, which primarily scale by upgrading hardware (vertical scaling). NoSQL proved to be a great solution for Big Data applications but also had some drawbacks, such as weaker consistency guarantees and complex query handling.

To overcome these issues, a new approach called NewSQL was introduced. NewSQL combines the best of both SQL and NoSQL. It retains the structured, relational model of SQL while offering the scalability and flexibility of NoSQL. This makes it a powerful choice for modern applications that require both data integrity and high performance.

What is SQL?

SQL (Structured Query Language) databases, also known as Relational Database Management Systems (RDBMS), follow a structured approach using tables or relations. SQL has been the standard for managing and querying relational datasets since the mid-1980s, though its origins date back to the 1960s and 1970s. The primary motivation behind SQL's development was to separate application data from application code, allowing developers to focus on program development rather than data manipulation.

Features of SQL Databases:

- **ACID Compliance**: Ensures Atomicity, Consistency, Isolation, and Durability for reliable transactions.
- **Normalization**: The process of structuring data efficiently through First Normal Form (1NF), Second Normal Form (2NF), and Third Normal Form (3NF).
- **Scalability**: SQL databases primarily support vertical scaling (adding more power to a single server rather than distributing across multiple servers).

- **Data Integrity and Security**: Enforces data integrity rules and constraints, reducing redundancy and ensuring consistency.
- **JOIN Operations**: Allows complex queries to retrieve related data efficiently.

Drawbacks of SQL Databases:

- **Rigidity in Data Modeling**: Data must be structured in tables, which may not suit all data types (e.g., hierarchical or graph-based data).
- **Scalability Limitations**: Traditional SQL databases struggle with horizontal scaling, making them less suitable for large-scale distributed applications.
- **Storage Inefficiency**: Predefined schema may lead to unused storage space if data does not fully utilize the allocated size.
- **Schema Modifications**: Changing a database schema can be complex and may require extensive modifications.
- **Performance Bottlenecks**: Handling large-scale data transactions in cloud-based environments may not be optimal.

What is NoSQL?

NoSQL databases emerged in response to the limitations of SQL databases, particularly in handling Big Data applications. Unlike SQL, NoSQL databases do not rely on a structured table format. Instead, they store data in key-value pairs, document stores, wide-column stores, or graph formats.

Features of NoSQL Databases:

- **Schema Flexibility**: No predefined schema, allowing storage of structured, semi-structured, and unstructured data.
- **Horizontal Scalability**: Enables distributed data storage across multiple servers, ensuring high availability and fault tolerance.
- **Auto-Sharding**: Automatically distributes data across multiple nodes to balance the load.
- **High Performance**: Optimized for high-speed read and write operations.
- **BASE (Basically Available, Soft State, Eventually Consistent) Compliance**: Ensures high availability by allowing temporary inconsistencies that eventually resolve over time.

Drawbacks of NoSQL Databases:

- **Lack of ACID Compliance**: NoSQL databases prioritize availability and partition tolerance over consistency, making them unsuitable for financial transactions.
- **Limited Query Capabilities**: Unlike SQL, NoSQL lacks a standardized query language, making complex queries challenging.
- **Security Concerns**: Lacks built-in security features like SQL’s role-based access control.
- **Lack of Standardization**: Different NoSQL databases use different models, leading to compatibility challenges.

- **Inefficient for Analytical Processing**: NoSQL databases are optimized for real-time applications but lack robust analytical processing capabilities.

What is NewSQL?

NewSQL is an emerging database model that combines the best aspects of SQL and NoSQL. It provides the scalability and distributed nature of NoSQL while maintaining the ACID compliance of SQL. NewSQL databases are particularly designed for Online Transaction Processing (OLTP) workloads.

Features of NewSQL Databases:

- **Hybrid Scalability**: Supports both vertical and horizontal scaling, making it efficient for handling large workloads.
- **ACID Compliance**: Ensures transactional reliability, making it suitable for enterprise applications.
- **In-Memory Processing**: Enhances performance by utilizing memory for faster data retrieval.
- **Partitioning & Sharding**: Efficiently distributes data across multiple nodes.
- **Automatic Replication**: Provides high availability and fault tolerance.
- **Enhanced Query Optimization**: Improves query performance through indexing and query execution plans.

Drawbacks of NewSQL Databases:

- **Early Stage Technology**: NewSQL is still evolving, and adoption is relatively low compared to SQL and NoSQL.
- **Limited Vendor Support**: Fewer providers and solutions compared to well-established SQL and NoSQL databases.
- **Complex Implementation**: Requires specialized knowledge to configure and optimize efficiently.

Feature	SQL	NoSQL	NewSQL
Relational Property	Yes, it follows relational modelling to a large extent.	No, it doesn't follow a relational model. It was designed to be entirely different from that.	Yes, since the relational model is equally essential for real-time analytics.
ACID	Yes, ACID properties are fundamental to their application	No, rather provides for CAP support	Yes, Acid properties are taken care of.

SQL	Support for SQL	No support for old SQL	Yes, proper support and even enhanced functionalities for Old SQL
OLTP	Inefficient for OLTP databases.	It supports such databases, but it is not the best suited.	Fully functionally supports OLTP databases and is highly efficient
Scaling	Vertical scaling	Only Vertical scaling	Vertical + Horizontal scaling
Query Handling	Can handle simple queries with ease and fails when they get complex	Better than SQL for processing complex queries	Highly efficient in processing complex queries and smaller queries.
Distributed Databases	No	Yes	Yes

HADOOP

Hadoop is an open source framework based on Java that manages the storage and processing of large amounts of data for applications. Hadoop uses distributed storage and parallel processing to handle big data and analytics jobs, breaking workloads down into smaller workloads that can be run at the same time.

ADVANTAGES OF HADOOP:

1. Scalability: Hadoop is important as one of the primary tools to store and process huge amounts of data quickly. It does this by using a distributed computing model which enables the fast processing of data that can be rapidly scaled by adding computing nodes.

2. Low cost: As an open source framework that can run on commodity hardware and has a large ecosystem of tools, Hadoop is a low-cost option for the storage and management of big data.

3. Flexibility: Hadoop allows for flexibility in data storage as data does not require preprocessing before storing it which means that an organization can store as much data as they like and then utilize it later.

4. Resilience: As a distributed computing model, Hadoop allows for fault tolerance and system resilience, meaning if one of the hardware nodes fail, jobs are redirected to other nodes. Data stored on one Hadoop cluster is replicated across other nodes within the system to fortify against the possibility of hardware or software failure

5. Fast : Hadoop's unique storage method is based on a distributed file system that basically 'maps' data wherever it is located on a cluster. The tools for data processing are often on the same servers where the data is located, resulting in much faster data processing. If you're dealing with large volumes of unstructured data, Hadoop is able to efficiently process terabytes of data in just minutes, and petabytes in hours.

FEATURES OF HADOOP AND ITS ECOSYSTEM TOOLS:

1. **HDFS (Hadoop Distributed File System)**:

   - The core storage component of Hadoop, which stores large datasets across multiple machines.
   - Data is split into blocks (typically 128MB or 256MB) and stored across multiple nodes, providing redundancy and fault tolerance.

2. **MapReduce**:

   - A programming model for processing large datasets in parallel. The Map phase divides the input data into smaller chunks, and the Reduce phase aggregates the results.
   - It enables distributed data processing on a large scale, making it efficient for Big Data workloads.

3. **YARN (Yet Another Resource Negotiator)**:

   - YARN is the resource management layer of Hadoop that manages and schedules resources across the cluster.
   - It allows multiple applications (such as MapReduce, Spark, etc.) to run on the same cluster by allocating resources dynamically.

4. **HBase**:

- A NoSQL database built on top of HDFS, designed to store large amounts of sparse data in a distributed environment.
- It allows for real-time read/write access to data and is ideal for applications that require low-latency access to large datasets.

5. **Hive**:

- A data warehouse system built on top of Hadoop, which provides a high-level interface for querying and analyzing data using SQL-like language (HiveQL).
- It abstracts the complexities of writing MapReduce programs and makes it easier for users familiar with SQL to interact with Hadoop.

6. **Pig**:

- A platform that provides a high-level language (Pig Latin) for processing and analyzing large datasets.
- Pig is more procedural than Hive and is optimized for data transformations and loading tasks.

7. **Spark**:

- An in-memory data processing framework that provides faster processing than Hadoop's MapReduce by performing computations in memory.
- It supports batch processing and real-time streaming, and it can be used with Hadoop for faster analytics and data processing.

8. **Flume**:

- A tool for collecting, aggregating, and transporting large volumes of log data into Hadoop's HDFS.
- It is highly scalable and fault-tolerant, making it ideal for handling real-time data streams.

9. **Sqoop**:

- A tool used to import and export data between Hadoop and relational databases (e.g., MySQL, Oracle).
- It simplifies the process of transferring data from traditional databases to Hadoop, enabling seamless integration.

10. **Zookeeper**:

- A coordination service for distributed applications, providing synchronization and configuration management.

- It is used to maintain configuration information and coordinate distributed services in the Hadoop ecosystem.