

Why Big Data and Where did it come from?,

Why Big Data?

1. **Volume:** The sheer amount of data generated daily is immense. Businesses, social media, IoT devices, and other sources produce terabytes and petabytes of data. Traditional data processing tools can't handle such vast amounts.
2. **Variety:** Data comes in various formats – structured (databases), semi-structured (XML, JSON), and unstructured (text, images, videos). Big Data technologies are designed to handle and analyze these diverse types of data.
3. **Velocity:** Data is generated at unprecedented speeds. For example, social media updates, sensor data, and financial transactions produce streams of data that need real-time processing.
4. **Veracity:** Ensuring the quality and accuracy of data is crucial. Big Data tools help in cleaning, processing, and analyzing data to ensure it's reliable.
5. **Value:** Analyzing Big Data helps businesses and organizations derive actionable insights, improve decision-making, enhance customer experiences, and create competitive advantages.

Where Did Big Data Come From?

1. **Technological Advancements:** The development of advanced computing technologies, such as distributed computing and parallel processing, enabled the handling of large datasets.
2. **Digital Transformation:** The digitalization of various sectors (e.g., finance, healthcare, retail) led to massive amounts of data generation. The

proliferation of internet use, social media, and e-commerce also contributed significantly.

3. **Mobile Devices and IoT:** The rise of smartphones, wearables, and IoT devices generated continuous streams of data from various sources, adding to the Big Data ecosystem.
4. **Cloud Computing:** The availability of scalable cloud storage and computing resources made it feasible to store and process large datasets without substantial investments in on-premises infrastructure.
5. **Open Source Tools:** The development and adoption of open-source Big Data tools and frameworks (e.g., Hadoop, Spark) made it easier for organizations to implement Big Data solutions.
6. **Data-Driven Culture:** As organizations recognized the importance of data-driven decision-making, they invested in technologies and strategies to harness the potential of Big Data.

Evolution of Big Data

1. **Early Days:** The term "Big Data" began to gain traction in the early 2000s. However, the concept of managing and analyzing large datasets has been around since the advent of computers and databases.
2. **2000s:** The development of Hadoop by Doug Cutting and Mike Cafarella in 2006 was a significant milestone. Hadoop's ability to process large datasets using a distributed computing approach revolutionized Big Data processing.
3. **2010s:** The rise of NoSQL databases, real-time data processing frameworks, and machine learning algorithms further propelled the Big Data movement. Companies like Google, Facebook, and Amazon leveraged Big Data to optimize their operations and services.

4. **Present and Future:** Today, Big Data continues to evolve with advancements in AI, machine learning, and edge computing. The focus is on making Big Data more accessible, scalable, and efficient, enabling organizations to unlock deeper insights and drive innovation.

Big Data has become a crucial component in various industries, driving significant advancements and transforming how businesses and organizations operate.

Big Data Characteristics

Big Data contains a large amount of data that is not being processed by traditional data storage or the processing unit. It is used by many **multinational companies** to **process** the data and business of many **organizations**. The data flow would exceed **150 exabytes** per day before replication.

There are five v's of Big Data that explains the characteristics.

5 V's of Big Data

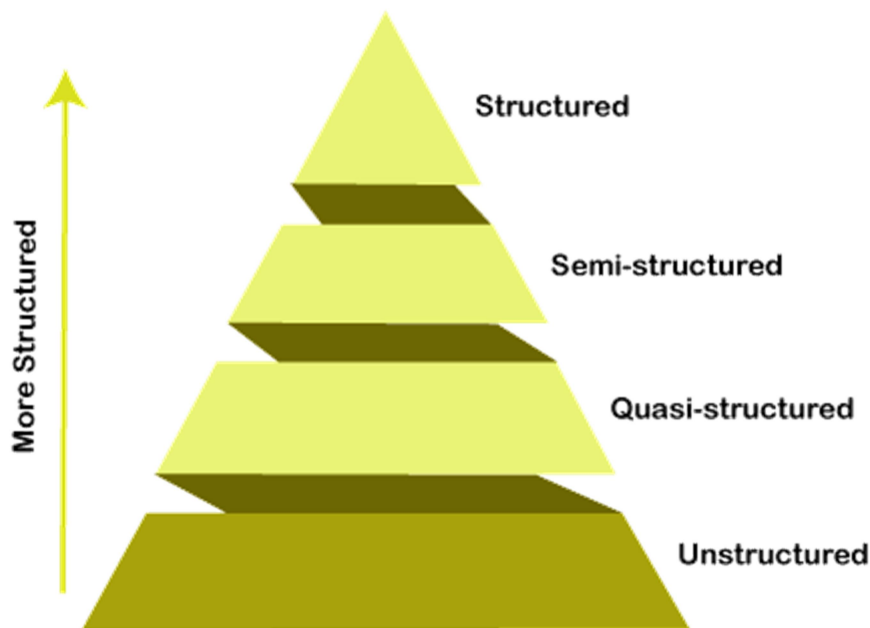
Volume

The name Big Data itself is related to an enormous size. Big Data is a vast 'volumes' of data generated from many sources daily, such as **business processes, machines, social media platforms, networks, human interactions**, and many more.

Facebook can generate approximately a **billion** messages, **4.5 billion** times that the "**Like**" button is recorded, and more than **350 million** new posts are uploaded each day. Big data technologies can handle large amounts of data.

Variety

Big Data can be **structured, unstructured, and semi-structured** that are being collected from different sources. Data will only be collected from **databases** and **sheets** in the past, But these days the data will comes in array forms, that are **PDFs, Emails, audios, SM posts, photos, videos, etc.**



The data is categorized as below:

- a. **Structured data:** In Structured schema, along with all the required columns. It is in a tabular form. Structured Data is stored in the relational database management system.
- b. **Semi-structured:** In Semi-structured, the schema is not appropriately defined, e.g., **JSON, XML, CSV, TSV**, and **email**. OLTP (**Online Transaction Processing**) systems are built to work with semi-structured data. It is stored in relations, i.e., **tables**.
- c. **Unstructured Data:** All the **unstructured files, log files, audio files**, and **image** files are included in the unstructured data. Some organizations

have much data available, but they did not know how to **derive** the value of data since the data is raw.

- d. **Quasi-structured Data:** The data format contains textual data with inconsistent data formats that are formatted with effort and time with some tools.

Example: Web server logs, i.e., the log file is created and maintained by some server that contains a list of **activities**.

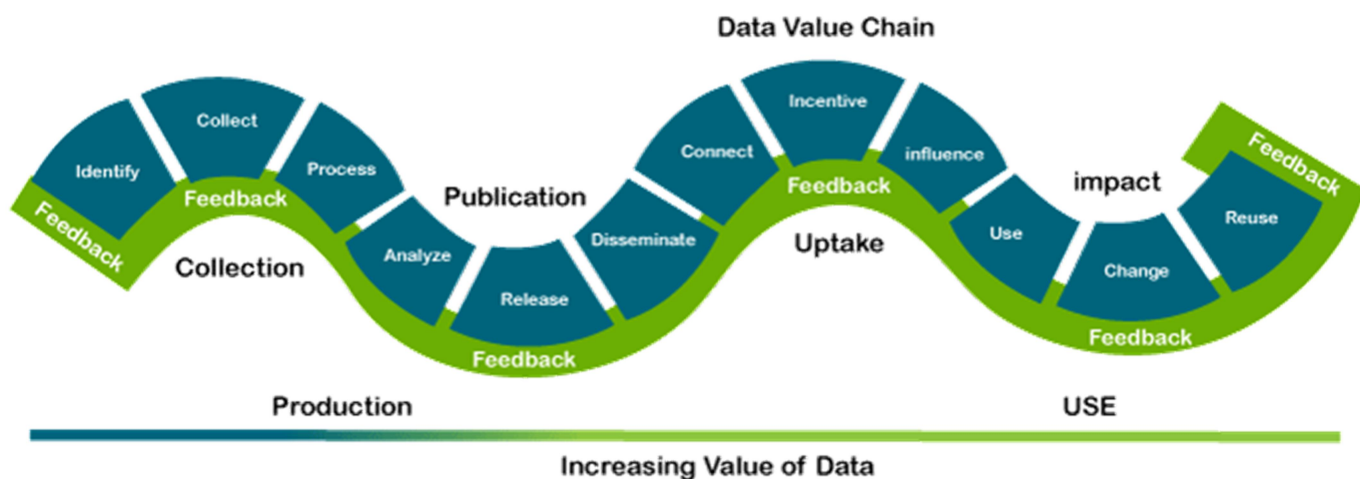
Veracity

Veracity means how much the data is reliable. It has many ways to filter or translate the data. Veracity is the process of being able to handle and manage data efficiently. Big Data is also essential in business development.

For example, **Facebook posts** with hashtags.

Value

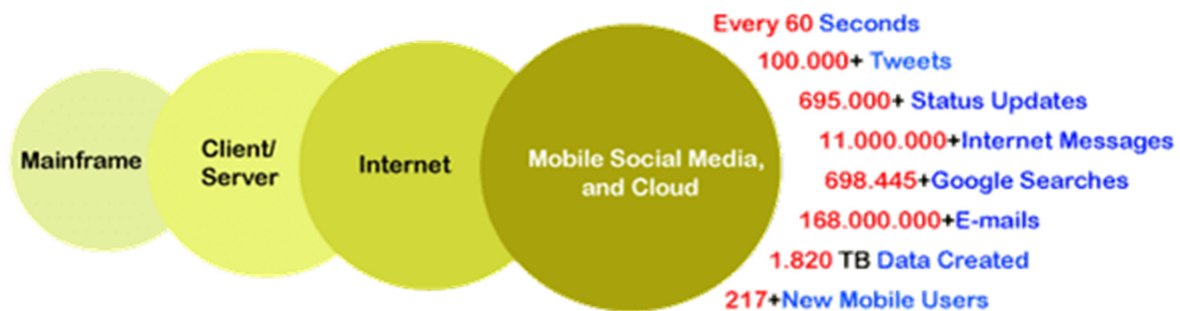
Value is an essential characteristic of big data. It is not the data that we process or store. It is **valuable** and **reliable** data that we **store, process,** and also **analyze**.



Velocity

Velocity plays an important role compared to others. Velocity creates the speed by which the data is created in **real-time**. It contains the linking of incoming **data sets speeds, rate of change, and activity bursts**. The primary aspect of Big Data is to provide demanding data rapidly.

Big data velocity deals with the speed at the data flows from sources like **application logs, business processes, networks, and social media sites, sensors, mobile devices, etc.**



Big Challenges with Big Data

Last Updated : 14 Jan, 2019

-
-
-

The challenges in **Big Data** are the real implementation hurdles. These require immediate attention and need to be handled because if not handled then the failure of the technology may take place which can also lead to some unpleasant result. Big data challenges include the storing, analyzing the extremely large and fast-growing data.

Some of the Big Data challenges are:

1. *Sharing and Accessing Data:*

- Perhaps the most frequent challenge in big data efforts is the inaccessibility of data sets from external sources.
- Sharing data can cause substantial challenges.
- It include the need for inter and intra- institutional legal documents.
- Accessing data from public repositories leads to multiple difficulties.
- It is necessary for the data to be available in an accurate, complete and timely manner because if data in the companies information system is to be used to make accurate decisions in time then it becomes necessary for data to be available in this manner.

2. *Privacy and Security:*

- It is another most important challenge with Big Data. This challenge includes sensitive, conceptual, technical as well as legal significance.
- Most of the organizations are unable to maintain regular checks due to large amounts of data generation. However, it should be necessary to perform security checks and observation in real time because it is most beneficial.
- There is some information of a person which when combined with external large data may lead to some facts of a person which may be secretive and he might not want the owner to know this information about that person.
- Some of the organization collects information of the people in order to add value to their business. This is done by making insights into their lives that they're unaware of.

3. *Analytical Challenges:*

- There are some huge analytical challenges in big data which arise some main challenges questions like how to deal with a problem if data volume gets too large?
- Or how to find out the important data points?

- Or how to use data to the best advantage?
- These large amount of data on which these type of analysis is to be done can be structured (organized data), semi-structured (Semi-organized data) or unstructured (unorganized data). There are two techniques through which decision making can be done:
 - Either incorporate massive data volumes in the analysis.
 - Or determine upfront which Big data is relevant.

4. *Technical challenges:*

- **Quality of data:**
 - When there is a collection of a large amount of data and storage of this data, it comes at a cost. Big companies, business leaders and IT leaders always want large data storage.
 - For better results and conclusions, Big data rather than having irrelevant data, focuses on quality data storage.
 - This further arise a question that how it can be ensured that data is relevant, how much data would be enough for decision making and whether the stored data is accurate or not.
- **Fault tolerance:**
 - Fault tolerance is another technical challenge and fault tolerance computing is extremely hard, involving intricate algorithms.
 - Nowadays some of the new technologies like cloud computing and big data always intended that whenever the failure occurs the damage done should be within the acceptable threshold that is the whole task should not begin from the scratch.
- **Scalability:**
 - Big data projects can grow and evolve rapidly. The scalability issue of Big Data has lead towards cloud computing.

- It leads to various challenges like how to run and execute various jobs so that goal of each workload can be achieved cost-effectively.
- It also requires dealing with the system failures in an efficient manner. This leads to a big question again that what kinds of storage devices are to be used.

Applications of Big Data

In today's world, there are a lot of data. Big companies utilize those data for their business growth. By analyzing this data, the useful decision can be made in various cases as discussed below:

1. Tracking Customer Spending Habit, Shopping Behavior: In big retail store (like Amazon, Walmart, Big Bazar etc.) management team has to keep data of customer's spending habit (in which product customer spent, in which brand they wish to spend, how frequently they spent), shopping behavior, customer's most liked product (so that they can keep those products in the store). Which product is being searched/sold most, based on that data, production/collection rate of that product get fixed.

Banking sector uses their customer's spending behavior-related data so that they can provide the offer to a particular customer to buy his particular liked product by using bank's credit or debit card with discount or cashback. By this way, they can send the right offer to the right person at the right time.

2. Recommendation: By tracking customer spending habit, shopping behavior, Big retail store provide a recommendation to the customer. E-commerce site like Amazon, Walmart, Flipkart does product recommendation. They track what product a customer is searching, based on that data they recommend that type of product to that customer.

As an example, suppose any customer searched bed cover on Amazon. So, Amazon got data that customer may be interested to buy bed cover. Next time when that customer will go to any google page, advertisement of various bed covers will be seen. Thus, advertisement of the right product to the right customer can be sent.

YouTube also shows recommend video based on user's previous liked, watched video type. Based on the content of a video, the user is watching, relevant advertisement is shown during video running. As an example suppose someone watching a tutorial video of Big data, then advertisement of some other big data course will be shown during that video.

3. Smart Traffic System: Data about the condition of the traffic of different road, collected through camera kept beside the road, at entry and exit point of the city, GPS device placed in the vehicle (Ola, Uber cab, etc.). All such data are analyzed and jam-free or less jam way, less time taking ways are recommended. Such a way smart traffic system can be built in the city by Big data analysis. One more profit is fuel consumption can be reduced.

4. Secure Air Traffic System: At various places of flight (like propeller etc) sensors present. These sensors capture data like the speed of flight, moisture, temperature, other environmental condition. Based on such data analysis, an environmental parameter within flight are set up and varied.

By analyzing flight's machine-generated data, it can be estimated how long the machine can operate flawlessly when it to be replaced/repared.

5. Auto Driving Car: Big data analysis helps drive a car without human interpretation. In the various spot of car camera, a sensor placed, that gather data like the size of the surrounding car, obstacle, distance from those, etc. These data are being analyzed, then various calculation like how many angles to rotate, what should be speed, when to stop, etc carried out. These calculations help to take action automatically.

6. Virtual Personal Assistant Tool: Big data analysis helps virtual personal assistant tool (like Siri in Apple Device, Cortana in Windows, Google Assistant in Android) to provide the answer of the various question asked by users. This tool tracks the location of the user, their local time, season, other data related to question asked, etc. Analyzing all such data, it provides an answer.

As an example, suppose one user asks “Do I need to take Umbrella?”, the tool collects data like location of the user, season and weather condition at that location, then analyze these data to conclude if there is a chance of raining, then provide the answer.

7. IoT:

- Manufacturing company install IOT sensor into machines to collect operational data. Analyzing such data, it can be predicted how long machine will work without any problem when it requires repairing so that company can take action before the situation when machine facing a lot of issues or gets totally down. Thus, the cost to replace the whole machine can be saved.
- In the Healthcare field, Big data is providing a significant contribution. Using big data tool, data regarding patient experience is collected and is used by doctors to give better treatment. IoT device can sense a symptom of probable coming disease in the human body and prevent it from giving advance treatment. IoT Sensor placed near-patient, new-born baby constantly keeps track of various health condition like heart bit rate, blood presser, etc. Whenever any parameter crosses the safe limit, an alarm sent to a doctor, so that they can take step remotely very soon.

8. Education Sector: Online educational course conducting organization utilize big data to search candidate, interested in that course. If someone searches for YouTube tutorial video on a subject, then online or offline course provider organization on that subject send ad online to that person about their course.

9. Energy Sector: Smart electric meter read consumed power every 15 minutes and sends this read data to the server, where data analyzed and it can be estimated what is the time in a day when the power load is less throughout the city. By this system manufacturing unit or housekeeper are suggested the time when they should drive their heavy machine in the night time when power load less to enjoy less electricity bill.

10. Media and Entertainment Sector: Media and entertainment service providing company like Netflix, Amazon Prime, Spotify do analysis on data collected from their users. Data like what type of video, music users are watching, listening most, how long users are spending on site, etc are collected and analyzed to set the next business strategy.

Why Big Data and Where did it come from?,
ChatGPT

Why Big Data?

1. **Volume:** The sheer amount of data generated daily is immense. Businesses, social media, IoT devices, and other sources produce terabytes and petabytes of data. Traditional data processing tools can't handle such vast amounts.
2. **Variety:** Data comes in various formats – structured (databases), semi-structured (XML, JSON), and unstructured (text, images, videos). Big Data technologies are designed to handle and analyze these diverse types of data.
3. **Velocity:** Data is generated at unprecedented speeds. For example, social media updates, sensor data, and financial transactions produce streams of data that need real-time processing.
4. **Veracity:** Ensuring the quality and accuracy of data is crucial. Big Data tools help in cleaning, processing, and analyzing data to ensure it's reliable.
5. **Value:** Analyzing Big Data helps businesses and organizations derive actionable insights, improve decision-making, enhance customer experiences, and create competitive advantages.

Where Did Big Data Come From?

1. **Technological Advancements:** The development of advanced computing technologies, such as distributed computing and parallel processing, enabled the handling of large datasets.
2. **Digital Transformation:** The digitalization of various sectors (e.g., finance, healthcare, retail) led to massive amounts of data generation. The proliferation of internet use, social media, and e-commerce also contributed significantly.

3. **Mobile Devices and IoT:** The rise of smartphones, wearables, and IoT devices generated continuous streams of data from various sources, adding to the Big Data ecosystem.
4. **Cloud Computing:** The availability of scalable cloud storage and computing resources made it feasible to store and process large datasets without substantial investments in on-premises infrastructure.
5. **Open Source Tools:** The development and adoption of open-source Big Data tools and frameworks (e.g., Hadoop, Spark) made it easier for organizations to implement Big Data solutions.
6. **Data-Driven Culture:** As organizations recognized the importance of data-driven decision-making, they invested in technologies and strategies to harness the potential of Big Data.

Evolution of Big Data

1. **Early Days:** The term "Big Data" began to gain traction in the early 2000s. However, the concept of managing and analyzing large datasets has been around since the advent of computers and databases.
2. **2000s:** The development of Hadoop by Doug Cutting and Mike Cafarella in 2006 was a significant milestone. Hadoop's ability to process large datasets using a distributed computing approach revolutionized Big Data processing.
3. **2010s:** The rise of NoSQL databases, real-time data processing frameworks, and machine learning algorithms further propelled the Big Data movement. Companies like Google, Facebook, and Amazon leveraged Big Data to optimize their operations and services.
4. **Present and Future:** Today, Big Data continues to evolve with advancements in AI, machine learning, and edge computing. The focus is on making Big Data more accessible, scalable, and efficient, enabling organizations to unlock deeper insights and drive innovation.

Big Data has become a crucial component in various industries, driving significant advancements and transforming how businesses and organizations operate.

Enabling Big Data Technologies

1. **Data Storage and Management**
 - **Hadoop Distributed File System (HDFS):** HDFS provides scalable and reliable data storage designed to handle large datasets by distributing data across multiple nodes.
 - **NoSQL Databases:** Databases like MongoDB, Cassandra, and HBase are designed for distributed data storage, capable of handling semi-structured and unstructured data.
 - **Data Lakes:** Platforms like AWS Lake Formation and Azure Data Lake store large amounts of raw data in its native format, enabling flexible analytics and processing.
2. **Data Processing and Analytics**

- **Apache Hadoop:** A framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models.
- **Apache Spark:** An open-source distributed computing system that provides an interface for programming entire clusters with implicit data parallelism and fault tolerance.
- **Apache Flink:** A stream processing framework that can process data in real-time, enabling low-latency analytics.
- **Dask:** A parallel computing library in Python that integrates with NumPy, pandas, and Scikit-learn, enabling Big Data processing in Python.

3. **Data Ingestion**

- **Apache Kafka:** A distributed streaming platform that can handle real-time data feeds, allowing for high-throughput, low-latency data ingestion.
- **Apache NiFi:** A data logistics platform for automating the movement of data between disparate systems.

4. **Data Integration and ETL (Extract, Transform, Load)**

- **Talend:** An open-source ETL tool that provides data integration, data management, and application integration solutions.
- **Apache Nifi:** A tool for automating the flow of data between systems, handling data ingestion, routing, and transformation.
- **Informatica:** Provides data integration and ETL capabilities for managing large volumes of data.

5. **Data Visualization and Business Intelligence**

- **Tableau:** A powerful data visualization tool that allows users to create a wide range of interactive and shareable dashboards.
- **Power BI:** Microsoft's business analytics tool for visualizing data and sharing insights across an organization.
- **Looker:** A business intelligence platform that helps in exploring, analyzing, and sharing real-time business analytics easily.

6. **Machine Learning and AI**

- **Apache Mahout:** A framework that allows building scalable machine learning algorithms.
- **TensorFlow:** An open-source machine learning framework developed by Google for building and training machine learning models.
- **Scikit-learn:** A Python library that provides simple and efficient tools for data mining and data analysis.

7. **Data Governance and Security**

- **Apache Ranger:** Provides centralized security framework to manage fine-grained access control across the Hadoop ecosystem.

- **Apache Atlas:** A data governance and metadata framework for managing data assets and ensuring data quality and compliance.
- **Collibra:** A data governance and catalog platform that helps organizations manage their data assets, ensuring compliance and data quality.

8. Cloud Platforms

- **Amazon Web Services (AWS):** Offers a suite of Big Data services, including Amazon S3 for storage, Amazon EMR for Hadoop processing, and Amazon Redshift for data warehousing.
- **Microsoft Azure:** Provides Big Data solutions like Azure HDInsight for Hadoop, Azure Databricks for Spark, and Azure Synapse Analytics for data warehousing.
- **Google Cloud Platform (GCP):** Offers Big Data services such as BigQuery for data warehousing, Google Cloud Storage for scalable storage, and Dataflow for stream and batch processing.

9. Edge Computing

- **Apache Edgent:** A programming model and micro-kernel style runtime for edge computing that enables the processing of data at the edge of the network.
- **AWS IoT Greengrass:** Extends AWS services to edge devices so they can act locally on the data they generate while still using the cloud for management, analytics, and durable storage.

10. Data Quality and Data Preparation

- **Trifacta:** A data preparation platform that helps in cleaning, structuring, and enriching raw data into refined formats for analysis.
- **DataRobot:** Provides automated machine learning, enabling users to prepare data, build and deploy models, and derive insights.

These technologies collectively enable organizations to harness the power of Big Data, turning vast amounts of data into valuable insights and actionable intelligence.

What is Big Data Stack?

Big data stack is a framework for the big data technologies that may satisfy the functional needs for big data projects is referred to as a technological stack. To understand big data, it helps to see how it stacks up that is, to set out the components of the architecture. Big data management architecture must incorporate a number of services that enable firms to make use of diverse data sources in a timely and effective manner.



Fig: Big Data Stack

- **Security infrastructure** - The information about your constituents must be protected in order to comply with regulatory requirements as well as to protect their privacy.
- **Operational data sources** - A relational database was used to store highly structured data that was handled by the line of business. Operational data sources were used to store highly-structured data.
- **Organizing Databases and tools** - structured database and tools used to organize the data and process this.
- **Analytical Data warehouse** - The addition of an analytical data warehouse simplifies the data for the development of reports.
- **Reporting and visualization** - Enable the processing of data while providing a user-friendly depiction of the results.

Big Data Distribution Packages

Big data distribution packages bundle various big data technologies into a single, integrated platform, making it easier for organizations to deploy and manage their big data infrastructure. Here are some of the most popular big data distribution packages:

1. Cloudera Data Platform (CDP)

- **Components:** Includes Apache Hadoop, Apache Spark, Apache HBase, Apache Hive, Apache Impala, Apache Kafka, and more.
- **Features:** Unified platform for data engineering, data warehousing, machine learning, and operational databases. It provides robust security, governance, and data management capabilities.
- **Deployment:** Available for on-premises, cloud, and hybrid environments.

2. Hortonworks Data Platform (HDP)

- **Components:** Includes Apache Hadoop, Apache Spark, Apache HBase, Apache Hive, Apache NiFi, Apache Kafka, and more.
- **Features:** Open-source platform designed for the enterprise, offering data governance, data security, and data management. HDP supports batch, interactive, and real-time analytics.
- **Deployment:** Available for on-premises and cloud environments.
- **Note:** Hortonworks merged with Cloudera, and their technologies are now part of Cloudera Data Platform.

3. MapR Data Platform

- **Components:** Includes Apache Hadoop, Apache Spark, Apache HBase, Apache Drill, Apache Kafka, and more, along with its own file system (MapR-FS) and database (MapR-DB).
- **Features:** Provides high performance, scalability, and reliability. MapR-FS is a distributed file system with POSIX compliance, and MapR-DB is a high-performance NoSQL database.
- **Deployment:** Available for on-premises, cloud, and hybrid environments.
- **Note:** MapR Technologies was acquired by HPE, and its products are integrated into the HPE Ezmeral Data Fabric.

4. Amazon EMR (Elastic MapReduce)

- **Components:** Supports Apache Hadoop, Apache Spark, Apache HBase, Apache Flink, Apache Hudi, and more.
- **Features:** Fully managed big data processing service that makes it easy to process large amounts of data using the Hadoop ecosystem. Provides scalability, flexibility, and integration with other AWS services.

- **Deployment:** Cloud-based, provided by Amazon Web Services (AWS).

5. **Google Cloud Dataproc**

- **Components:** Supports Apache Hadoop, Apache Spark, Apache Hive, Apache Pig, and more.
- **Features:** Fully managed service for running Apache Hadoop and Apache Spark clusters. It offers fast cluster startup times, autoscaling, and integration with other Google Cloud services.
- **Deployment:** Cloud-based, provided by Google Cloud Platform (GCP).

6. **Azure HDInsight**

- **Components:** Supports Apache Hadoop, Apache Spark, Apache HBase, Apache Storm, Apache Kafka, and more.
- **Features:** Fully managed cloud service that makes it easy to process big data using open-source frameworks. Provides enterprise-grade security, monitoring, and management capabilities.
- **Deployment:** Cloud-based, provided by Microsoft Azure.

7. **IBM Analytics Engine**

- **Components:** Supports Apache Hadoop, Apache Spark, Apache HBase, and more.
- **Features:** Managed service that provides a scalable and secure environment for running big data analytics. Integrates with other IBM Cloud services for comprehensive data management.
- **Deployment:** Cloud-based, provided by IBM Cloud.

8. **Databricks Unified Data Analytics Platform**

- **Components:** Built on Apache Spark, it also supports Delta Lake for reliable data lakes and MLflow for machine learning lifecycle management.
- **Features:** Provides a collaborative environment for data engineers, data scientists, and business analysts. It includes interactive notebooks, automated cluster management, and optimized runtime for Spark.
- **Deployment:** Available on AWS, Azure, and GCP.

9. **Qubole Data Service (QDS)**

- **Components:** Supports Apache Hadoop, Apache Spark, Presto, Apache Hive, and more.
- **Features:** Provides a self-service platform for big data analytics with auto-scaling, workload-aware autoscaling, and cost management features. Supports a variety of data processing frameworks.
- **Deployment:** Cloud-based, supporting AWS, Azure, and GCP.

These big data distribution packages offer comprehensive solutions for managing and processing large datasets, making it easier for organizations to implement and maintain their big data infrastructure.