Data Engineering

Big data refers to data that is so large, fast and complex that it’s difficult or impossible to process using traditional tools like a data warehouse.

Big Data

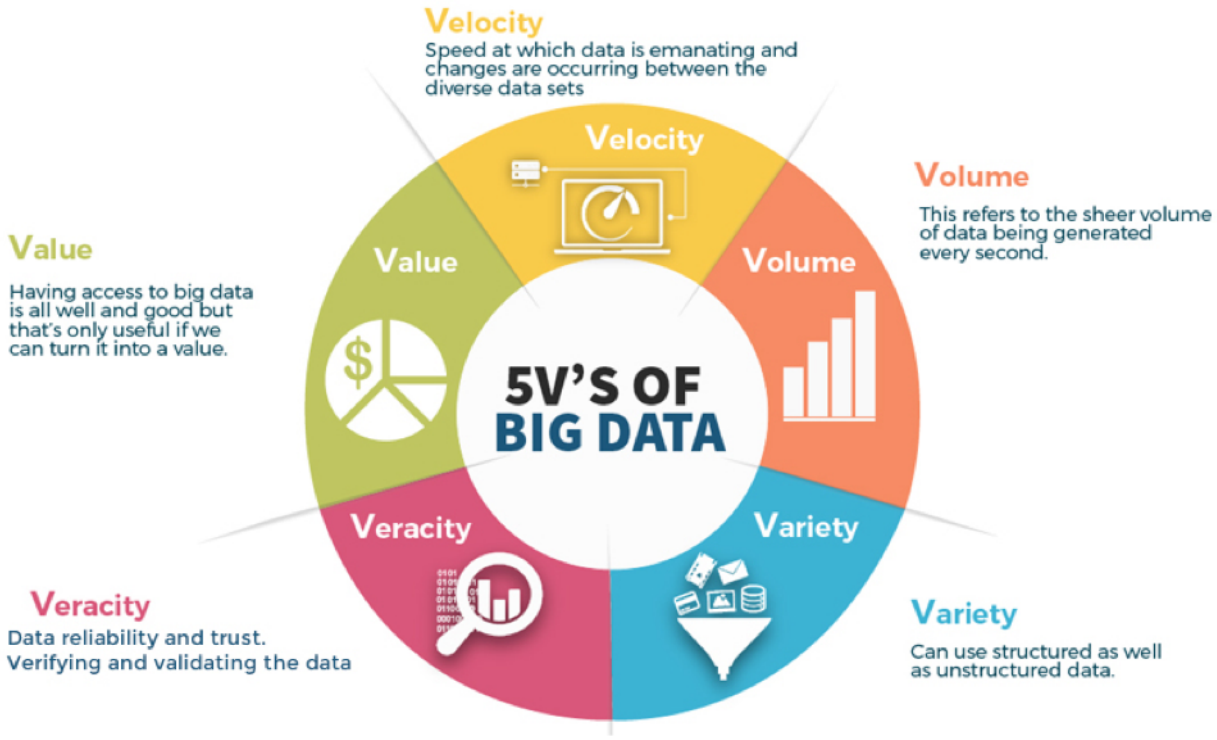
We might have heard the term Big Data before, but what does it really mean? Below is a high-level definition:

> Big data refers to data that is so large, fast and complex that it’s difficult or impossible to process using traditional tools like a data warehouse.

Traditional methods of data storage included storing data locally on a computer, on a server, mainframe system or in a database. Over time, due to the rapid evolution in data production, these traditional approaches are no longer able to cope with the new reality.

Accordingly, a new group of tools and technologies were introduced starting mid to late 2000's. The logic behind these new tools was to switch the mentality from storing and analyzing small quantities of high-value data in structured format using expensive systems such as a data warehouse (DWH) to being able to capture and store **all** raw data (both structured and unstructured) in a central repository, and then running various operations and transformations on this comprehensive dataset using tools that were open-source and inexpensive.

Hence, these new big data tools were developed to address the below challenges called the "V's" of big data:



- **Volume**:
  - The quantity of data is growing, and so is the number of data production sources (such as from Internet of Things devices and machines)
- **Variety**:
  - The format of produced data is also evolving. Now we have:
    - Structured: Table-like data organized into rows and columns. This used to be the standard format for data storage for decades.
    - Unstructured: More modern data types including text documents, emails, videos, images, and audio.
- **Velocity**:
  - The speed at which the data is arriving (batch initially and real-time more recently)
- **Veracity**:
  - The quality of the data (inconsistencies, uncertainties, empty data records etc.)
- **Value**:

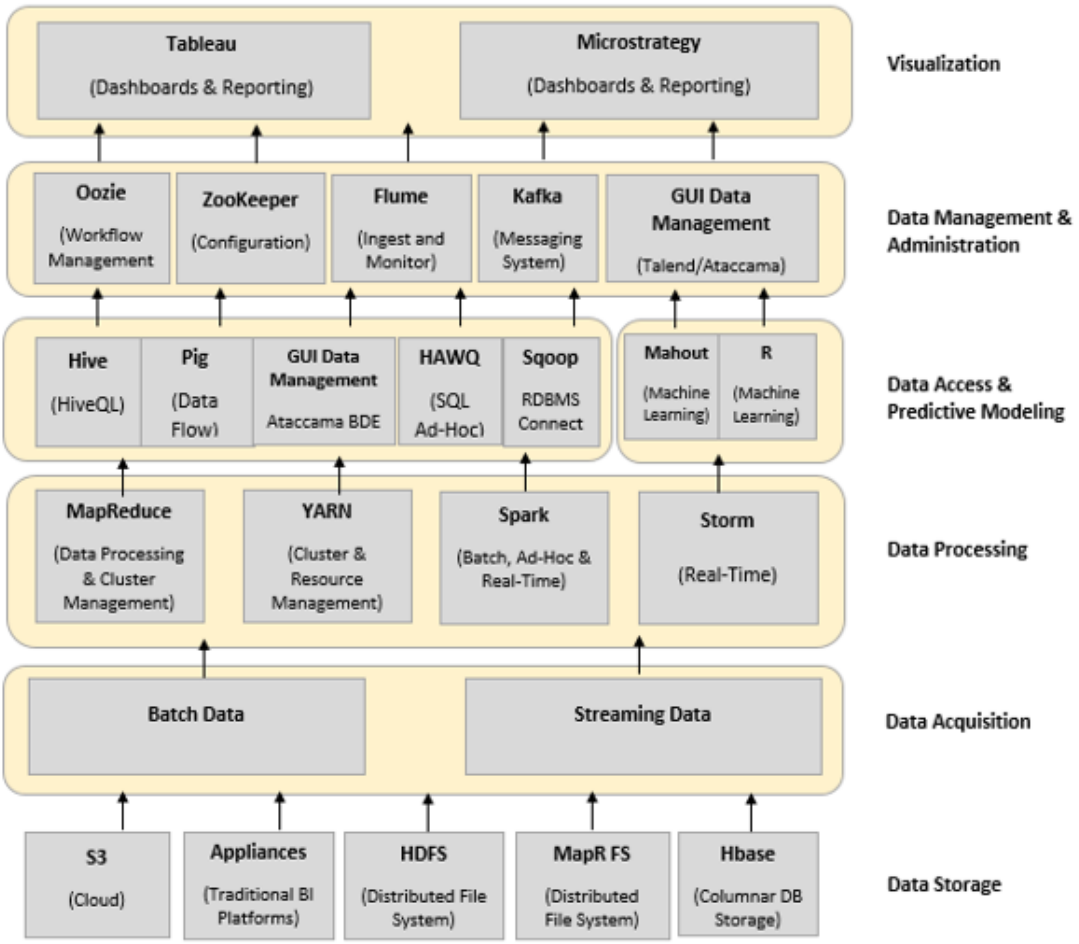
- How valuable is the captured data? How to increase its business value?

The first true Big data technology was the introduction of Apache Hadoop in 2006. This was a revolutionary step, as it finally enabled the capture, storage and analysis of structured and unstructured data in one centralised location called the data lake. Hadoop also provided the capabilities to manipulation, transform and explore the data using tools such as: - Java MapReduce - HiveQL (which is a SQL-like language) - Pig Latin (which is a scripting language similar to Bash)

Data Ecosystem

- The modern data infrastructure consists of various concepts, tools and components to be able to handle the various types of data being generated.
- These tools and components are arranged in groupings, with each layer performing a specific activity on the data before passing it on to the next layer
- The main layers in the data ecosystem include:
  - Data Storage
  - Data Acquisition
  - Data Processing
  - Data Access
  - Data Management
  - Data Visualisation

    Here is an example diagram showing how these layers and their corresponding tools fit in the big picture:



Data Storage

> Data storage, especially within the context of big data, is a compute-and-storage architecture that can be used to collect and manage huge-scale datasets and perform real-time data analysis.

In this layer, data that is generated from sources is captured and persisted. There are several alternatives available for this approach, each one tailored for specific types of data and its related use cases. The most common types of enterprise data storage components include:

Data Lakes

A data lake is a centralised repository that allows you to store all your structured and unstructured data in one location. This type of storage approach enables easy scaling to cope with the increasing quantity of data.

Data Warehouses

A data warehouse (DWH), also known as an enterprise data warehouse (EDW), is a system used for reporting and data analysis and is considered a core component of the traditional business intelligence architecture. DWHs are central repositories of integrated and structured data from one or more disparate sources. They store current and historical data in one single place that are used for creating analytical reports. These types of systems are expensive to maintain and usually store high-value data.

Cloud Storage

Cloud storage is a cloud computing model that stores data on the Internet through a distributed computing provider who manages and operates data storage as a service. It’s delivered on demand with just-in-time capacity and costs, and eliminates buying and managing data storage infrastructure. This gives users the agility, global scale and durability, with “anytime, anywhere” data access.

Data Aquisition

> Data acquisition is the process of collecting the data from various sources and moving it from point of origin to the target destination.

Depending on the type of data being produced, and the frequency of its production, we can classify this layer into two main types:

Batch Data:

This is the acquisition of a vast amount of data at-rest, in its entirety and at the same time. This approach is usually implemented at regular intervals (such as

once per day or once a week) on a full dataset, which includes the historical data plus any new incremental data. An example of batch data is moving a dataset of the entire list of customers for a certain company.

Real-time Data:

This is the acquisition of continuous data while it's in-motion in very short, near instantaneous intervals as the data arrives from its source. An example of real-time data is mobile application data that is constantly turned on, such as GPS locations.

Data Processing

> Data processing is, generally, the automated manipulation of raw data and transforming it to produce meaningful information. This is the layer that handles the ETL/ELT operations on the data.

Data processing is handled by specialized tools, with the vast majority of these tools being open-source and oftentimes free to download and use.

The most popular industry-ready frameworks are Hadoop (including Yarn), Spark, Flink and Storm. The choice on which framework to use, and even which component within that tool to deploy (as they all come with various modules) is determined by an enterprise system Architect. There are several criteria that support in determining the tool of choice. Some of these criteria include:

- Is the incoming data batch or streaming?
- What is the type of data being captured and stored (structured, unstructured?)
- Where will the data be stored?
- What transformations are required on the data?
- What is the quality of the data?
- What are we expected to do with the data after cleaning and transforming?

Once the data passes through the required processing steps, it's ready to be accessed via the required stakeholders as it's now in a cleaned and integrated state.

Data Access

> Data access refers to the capability to interact with the data once it's been through the data processing steps.

In the data access layer, the aim is to expose the cleaned, integrated and prepared data to the downstream systems and various stakeholders. This can be done via several tools, depending on our objective. For instance, if we are querying the data for analytics purposes, then we'll use tools such as HiveQL, Spark SQL or Python. If the goal is to create predictive models, we can use a tool like R, Pandas, and Numpy to create and run these algorithms.

Data Management

> Data management is the process of managing and synchronizing the data based on requirements.

In the data management layer, the objective is to determine what to do next with the data after access is provided. For instance, it may be required that certain jobs need to run hourly or daily. In this case, we'll use tools like Oozie or Airflow to create and schedule such jobs and to provide the required parameters.

If the requirement is to move or copy the data into another system, we can use Kafka to transport the data to the target system. If the requirement is to store the data into an enterprise data warehouse (EDW), we can use a tool like Sqoop to place the data there directly.

Finally, if the need is to create various dashboards and graphs, we can feed the data to the next (and final) layer, which is the data visualisation layer.

Data Visualisation

> Data visualisation is the process of displaying data in charts, graphs, maps, and other visual forms. It is used to help people easily understand and interpret their data at a glance, and to clearly show trends and patterns that arise from this data.

After the raw data has been through the previous 5 steps, it's now ready to be presented to business leaders and executives. It's quite common for non-technical professionals to refer to dashboards and graphs to analyze how the business is performing, track metrics and key performance indicators (KPI's) and to monitor the day-to-day operations of the company. Actually, in reality most data related projects are initiated to at least partially support dashboards and visualisation tools for the top-level business leaders.

In this layer, a tool like Tableau can be leveraged to represent the information we have in an easy to understand graphical format. Some common types of charts include:

- Pie charts
- Bar charts
- Time series charts
- Graphs

Data Fabric

> A data fabric is a novel data management architecture that serves as an integrated layer (fabric) of data and connecting processes. It can optimise access to distributed data and intelligently curate and orchestrate it for self-service delivery to data consumers.

Data fabric is a very modern approach to managing the data within large organizations. With a data fabric, you can elevate the value of enterprise data

by providing users access to the right data just in time, regardless of where it is stored. A data fabric architecture is agnostic to data environments, data processes, data use and geography, while integrating core data management capabilities. It automates data discovery, data governance and consumption, delivering business-ready data for analytics and AI.

Top performing enterprises are data driven. However, several challenges block them from fully exploiting all data:

- Lack of data access.
- Numerous data sources and data types.
- Data integration complexities.

Research shows that up to 74% of data is not analyzed in most organizations and up to 82% of enterprises are inhibited by data silos.

With a data fabric, business users and data scientists can access trusted data faster for their applications, analytics, AI and machine learning models, and business process automation, helping to improve decision-making and drive digital transformation. Technical teams can use a data fabric to radically simplify data management and governance in complex hybrid and multi-cloud data landscapes while significantly reducing costs and risk.

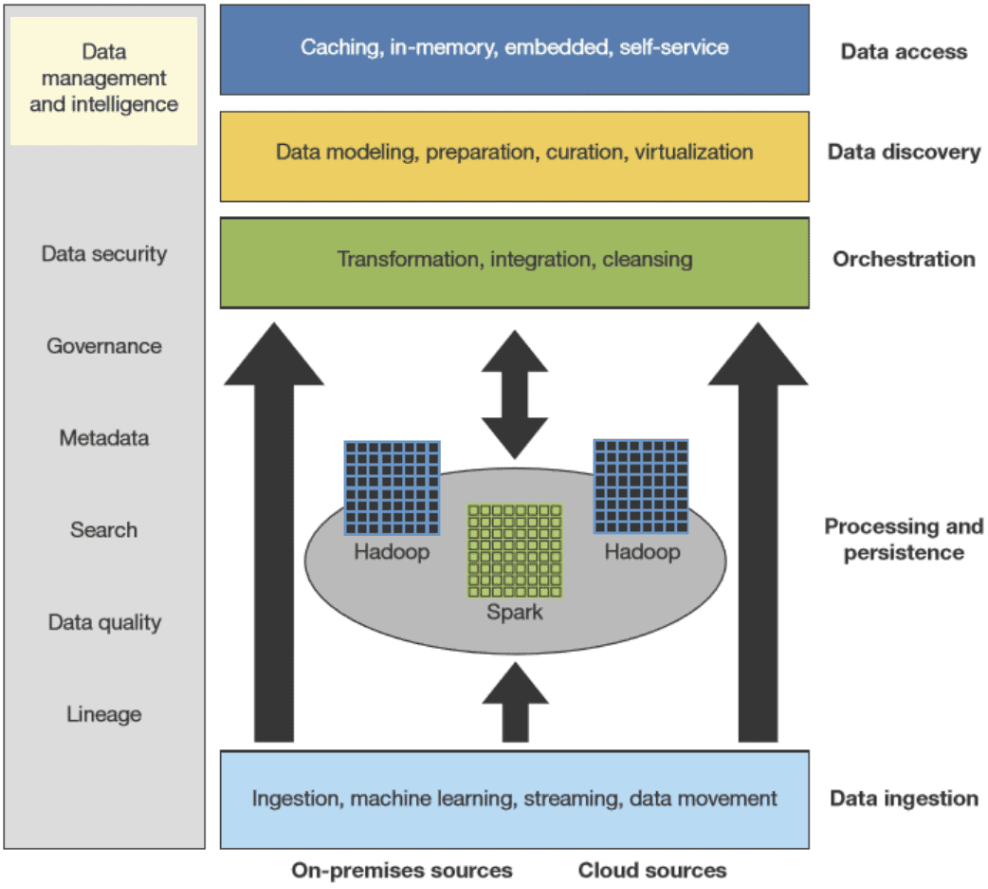
Data fabric enables a permanent and scalable mechanism for business to consolidate all its data under the umbrella of one unified platform. It leverages storage and processing power from multiple heterogeneous nodes to enable enterprise-wide access to all data assets of an enterprise. According to Forrester, a big data fabric assists enterprises to “…quickly ingest, transform, curate, and prepare streaming and batch data to support a real-time trusted view of the customer and the business.” *

Furthermore, big data fabric enables companies to:

- Effectively consolidate data assets with on-premises and Cloud data sources, for a complete view of enterprise-wide information.
- Gain access to the latest data in real-time.
- Easily onboard new big data systems and retire legacy systems, while keeping business systems running continuously without disruption.

- From a problem-solving perspective, data fabric overcomes the challenges of insufficient data availability, unreliability of data storage and security, siloed data, poor scalability, and reliance on underperforming legacy systems.

Below is what a typical data fabric ecosystem would look like in a global company:



- Big data can be defined using the 5 V's: volume, velocity, variety, veracity and value
- Big data is different from traditional data, as it comes in both structured and unstructured formats and the data arrives in various velocities in accelerating quantities

- Traditional relational database and data warehouse models were well suited for structured data with low volumes, however such systems are expensive to maintain, and can't handle unstructured data. This is why they are currently being augmented with more modern big data tools.
- The modern data ecosystem is composed of 6 layers, namely: storage, acquisition, processing, modelling, management and administration. Data must flow sequentially from the first layer (data storage) till the last layer (visualisation) and each layer performs a specific function on the data.
- Each of the 6 layers of the modern data ecosystem has specific tools that have become the industry standard
- In the data storage layer, cloud data storage is quickly becoming the dominant trend in industry