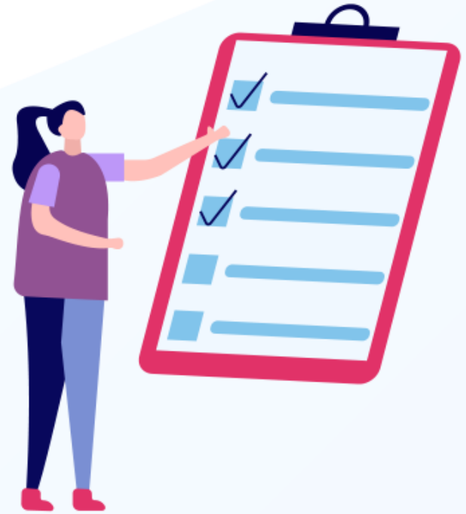
Big Data Hadoop and Spark Developer

Introduction to Big Data and Hadoop

Learning Objectives

By the end of this lesson, you will be able to:

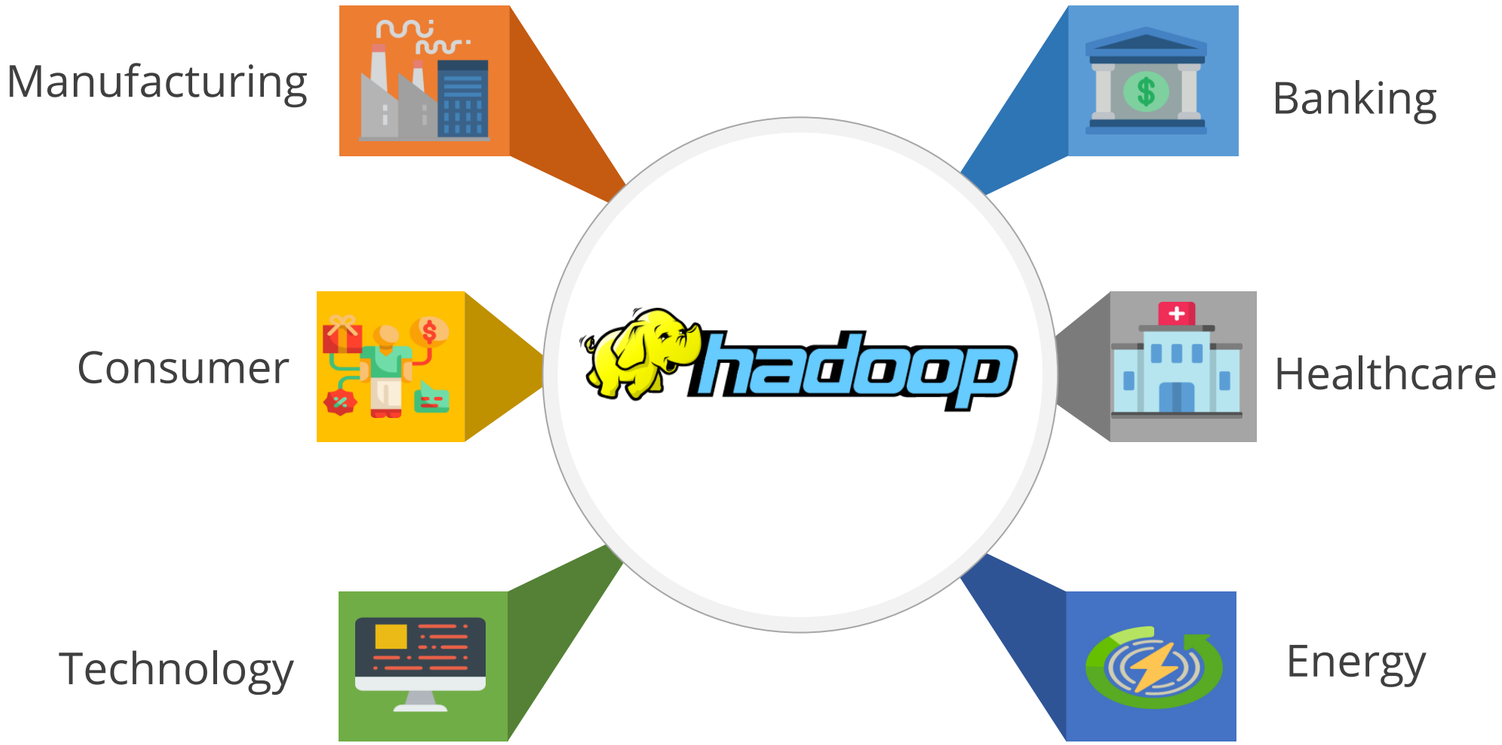
- Define the concepts of Big Data
- Explain Hadoop and how it addresses Big Data challenges
- Illustrate the components of the Hadoop Ecosystem



Introduction to Big Data and Hadoop

Big Data Overview

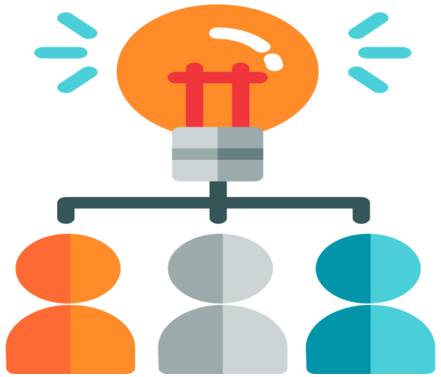
Big Data is data that has high volume, variety, velocity, veracity, and value.



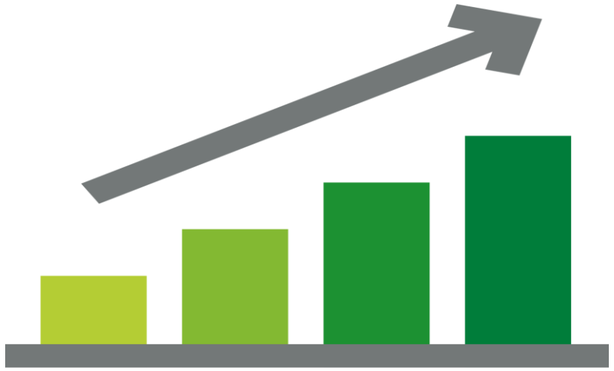
According to the US Bureau of Labor Statistics, Big Data alone will fetch 11.5 million jobs by 2026.

Traditional Decision-Making

Different ways of traditional decision-making are:



What We Think



Experience and Intuition

Rule of Thumb

Challenges of Traditional Decision-Making

Takes a long time to arrive at a decision, therefore losing the competitive advantage

Requires human intervention at various stages

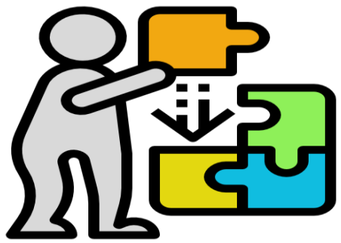
Lacks systematic linkage among strategy, planning, execution, and reporting

Provides limited scope of data analytics, that is, it provides only a bird's eye view

Obstructs company's ability to make fully informed decisions

Big Data Analytics

The Solution: Big Data Analytics



The decision-making is based on what users know which in turn is based on data analytics.

It provides a comprehensive view of the overall picture which is a result of analyzing data from various sources.

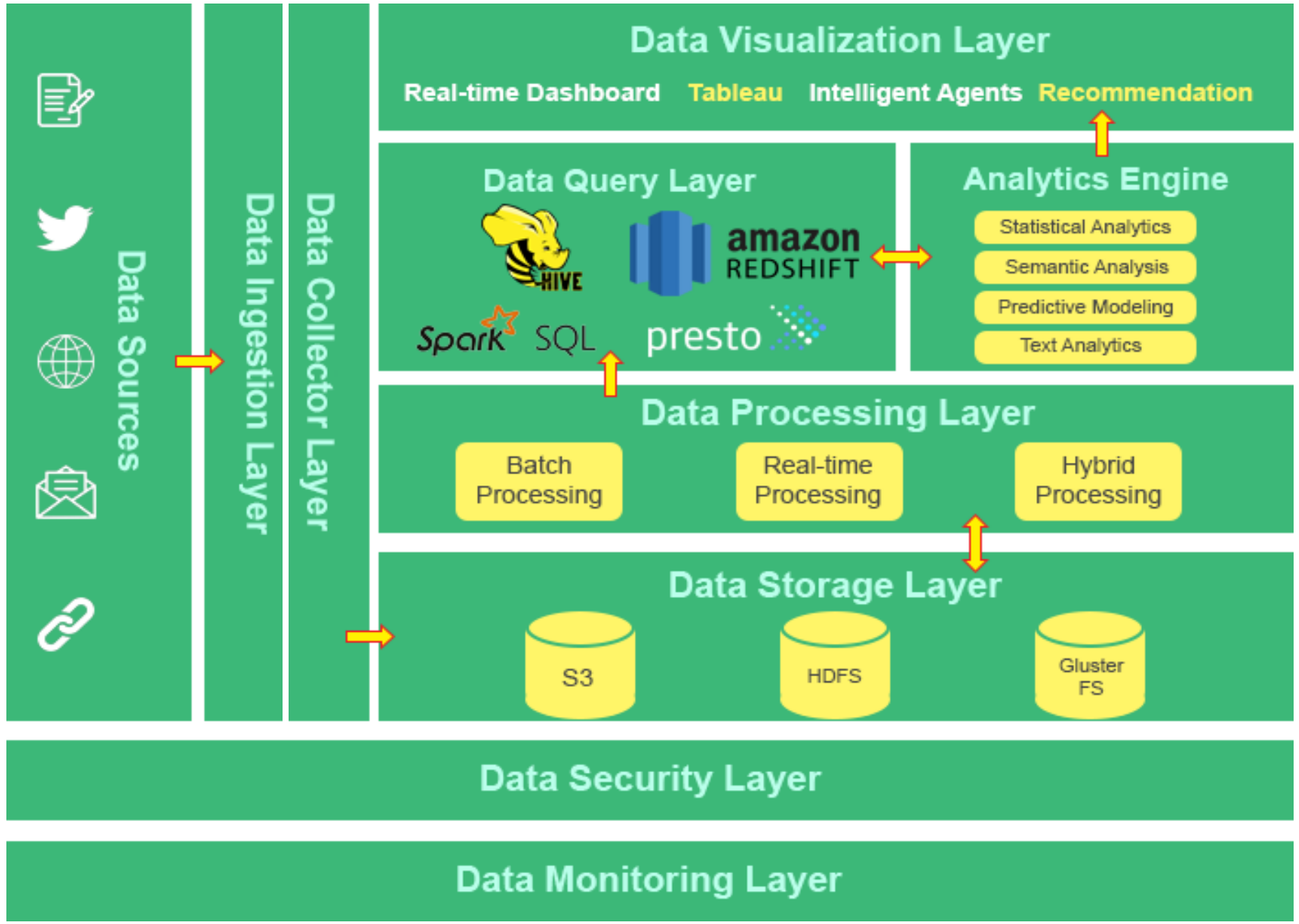
It provides streamlined decision-making from top to bottom.

Big data analytics helps in analyzing unstructured data.

It helps in faster decision-making thus improving the competitive advantage and saving time and energy.

Big Data Analytics Pipeline

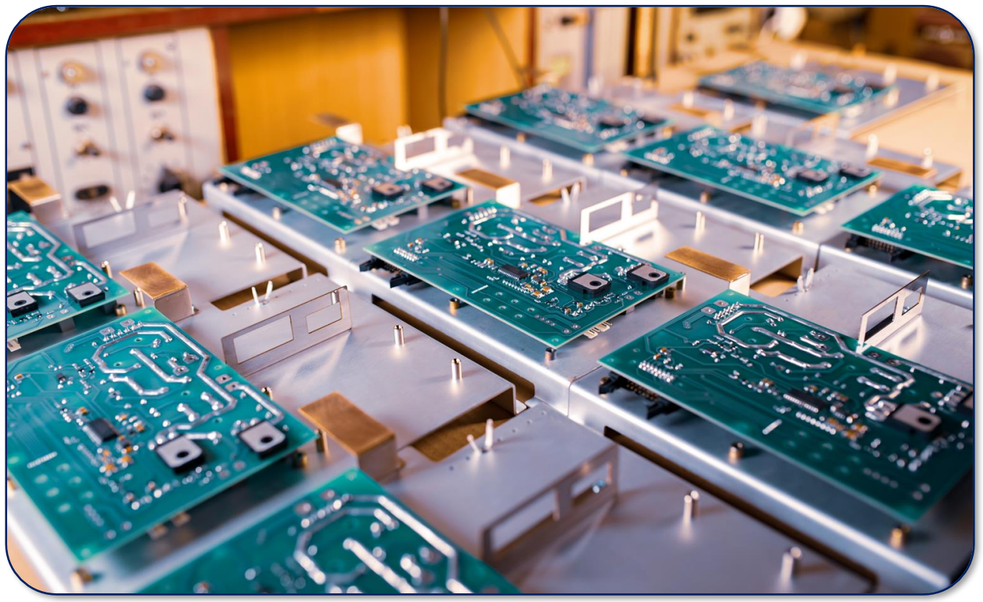
A data pipeline is a series of steps that ingest raw data from multiple sources and transfer the data to a destination for storage and analysis.



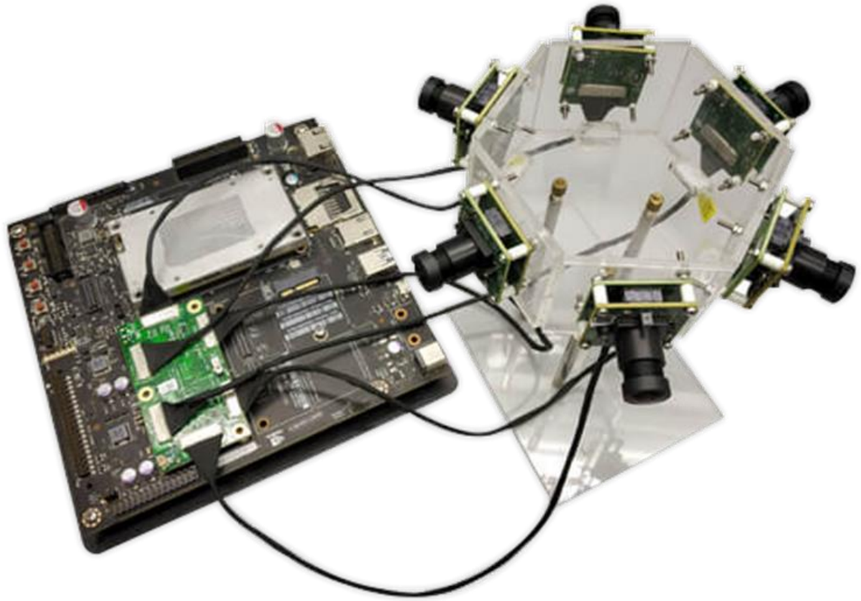
Case Study: Big Data Using NVIDIA Jetson Camera

Big Data Using NVIDIA Jetson Camera

A semiconductor manufacturing firm is testing chips in traditional ways. A total of 1.3 million chips are made, and a person is required to test and validate if the chips are ready to be dispatched to dealers.
It takes nearly 8 minutes to test a chip, and the company wants to reduce this time to increase production.

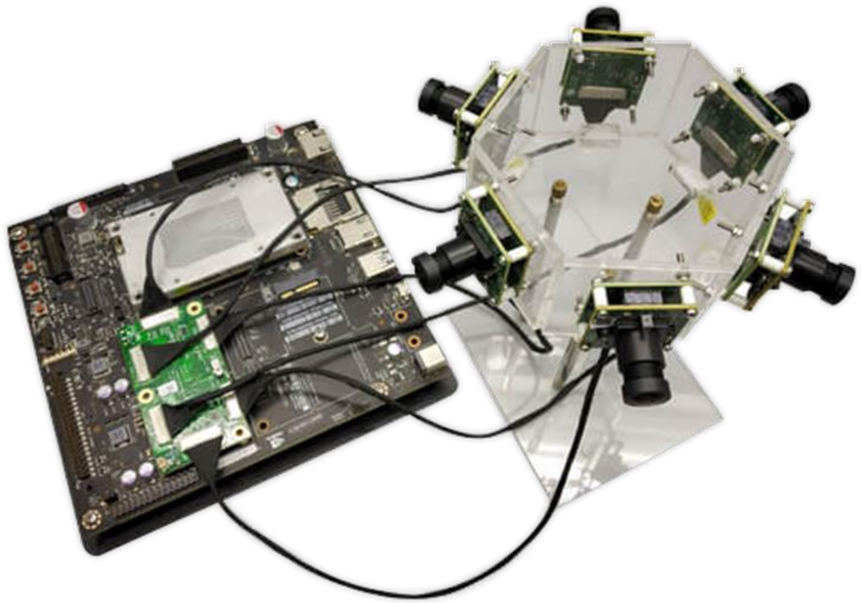


Big Data Using NVIDIA Jetson Camera



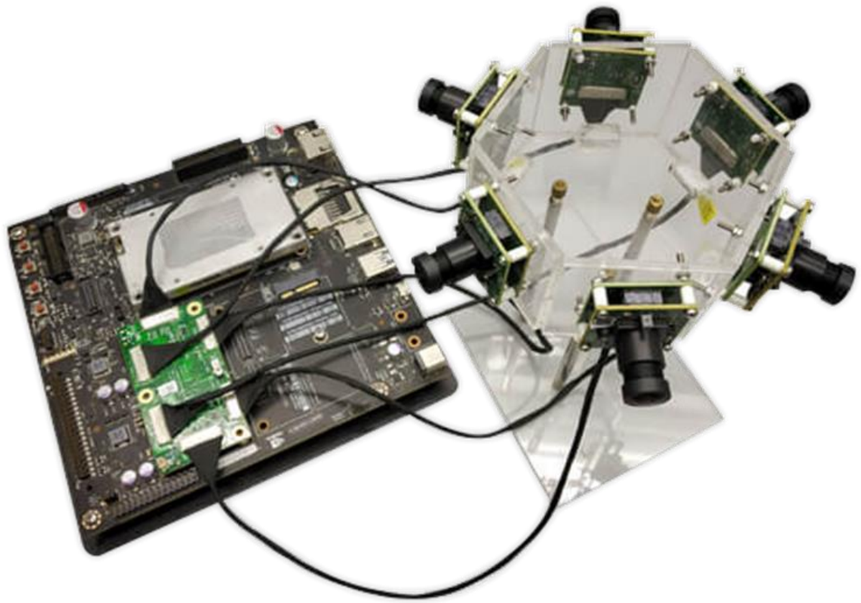
- The company installed six NVIDIA Jetson cameras in the testing utility to create a fully automated system for fast processing.
- It captures real-time photos of chips and sends them to the Hadoop Big Data pipeline for processing.

Big Data Using NVIDIA Jetson Camera



- When images are received into the system, ETL is performed on them to remove noise and transform them into fixed pixel length images.
- The images are then dumped into the Hadoop system for further processing.

Big Data Using NVIDIA Jetson Camera



- If the Hadoop system rejects some images, the batch of chips is manually checked by humans to remove any defects.
- It is estimated that Hadoop system testing was nearly 95% accurate, with 5% requiring manual intervention, which is a significant advancement in chip manufacturing.

What Is Big Data?

What Is Big Data?

Big data refers to large data sets that may be analyzed computationally to reveal patterns, trends, and associations, especially relating to human behavior and interactions.

Big Data at a Glance

28,000
Subscribers watching

695,000
Stories shared

9,132
Connection made

69 million
Messages sent

5,000
Downloads

2 million
Views

1.6 million USD
Spent online

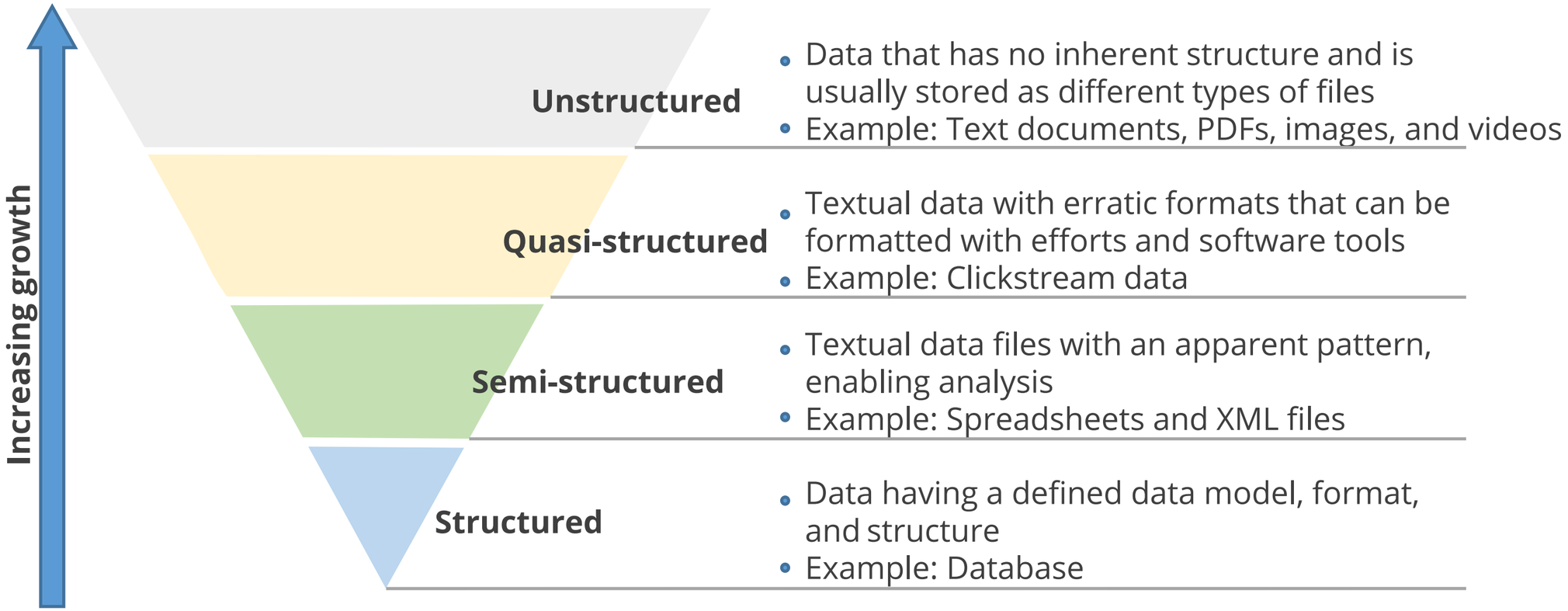
2 million
Swipes

197.6 million
Emails sent

500 hours
Content uploaded

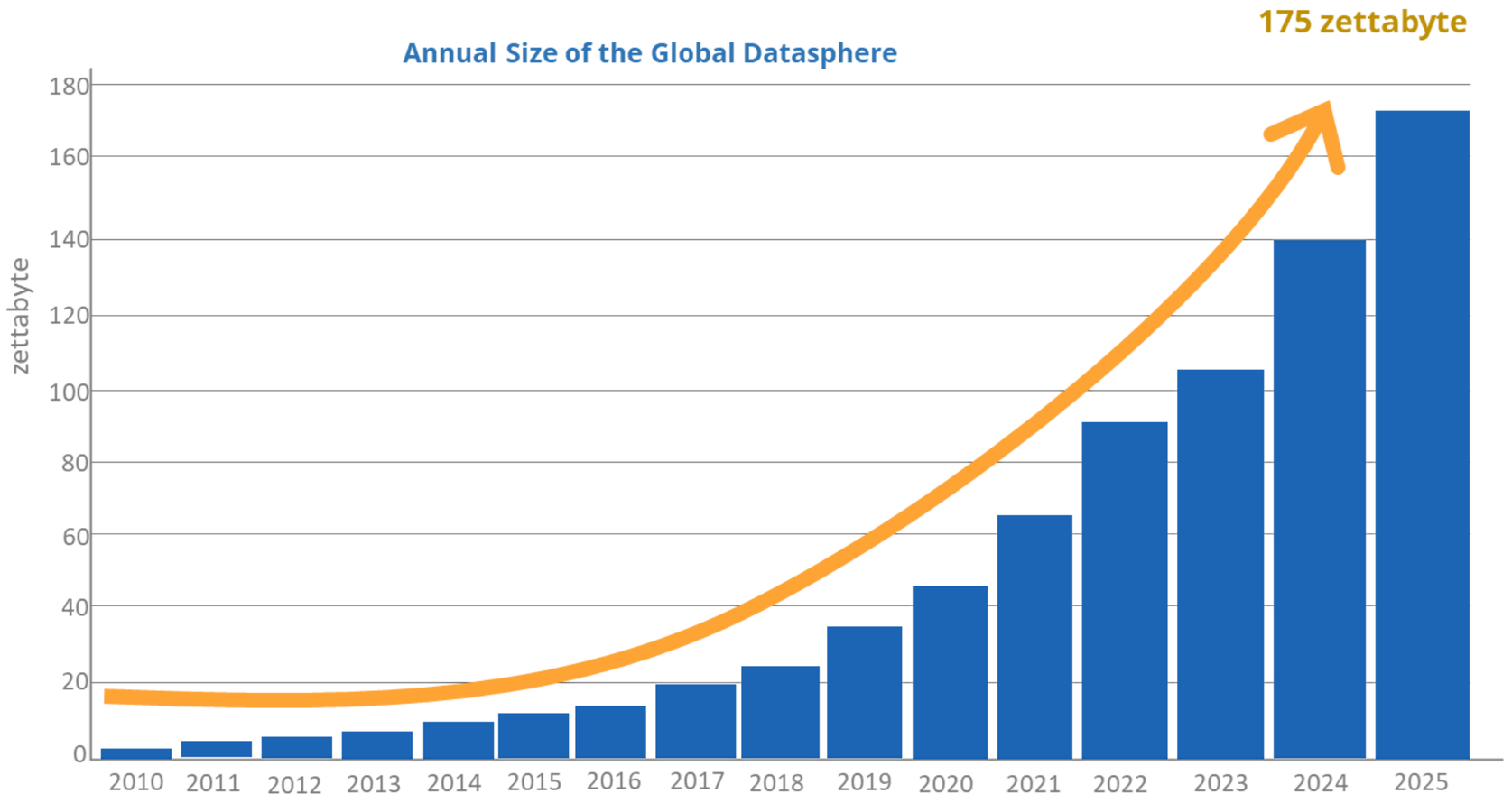
Different Types of Data

Following are the types of data:



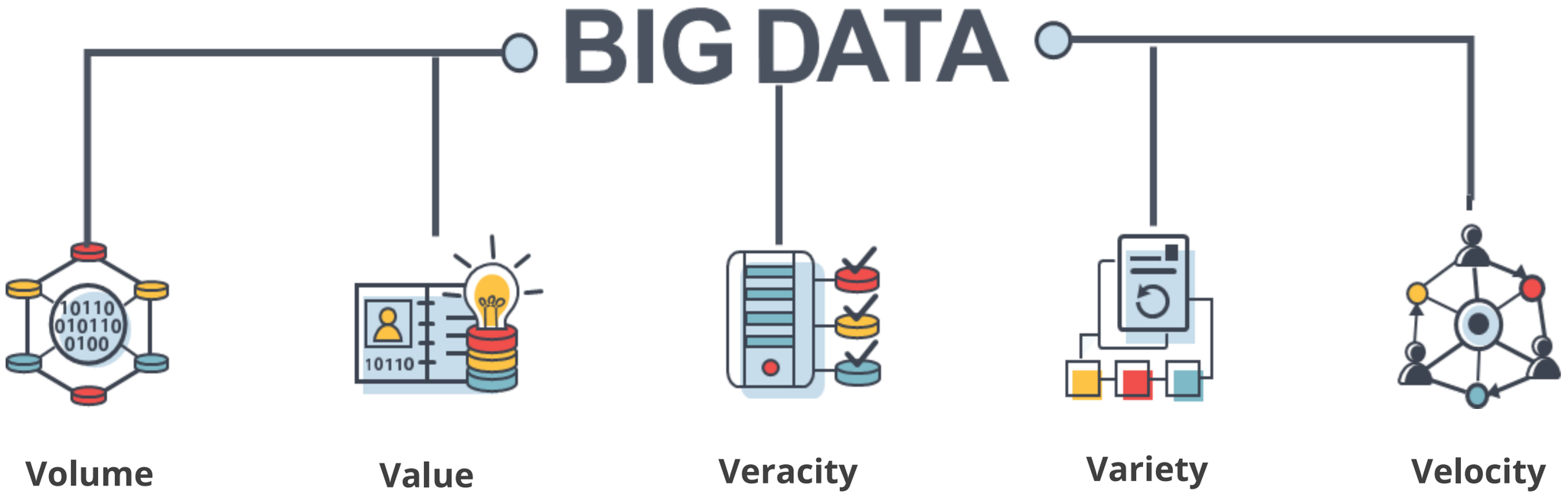
Growth in Data

It is estimated that the data will rise exponentially by 2025.

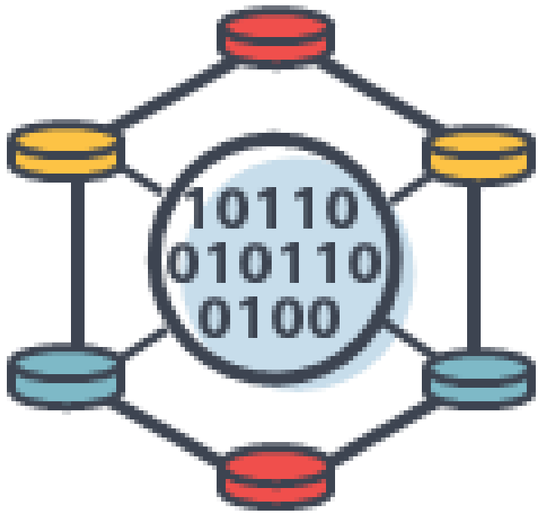


Five V’s of Big Data

Five V’s


BIG DATA
10110
010110
0100
10110
Volume
Value
Veracity
Variety
Velocity

Volume



- Volume refers to the massive amounts of data generated by social media every second.
- Facebook generates about a billion messages which can only be handled by Big Data Technologies.

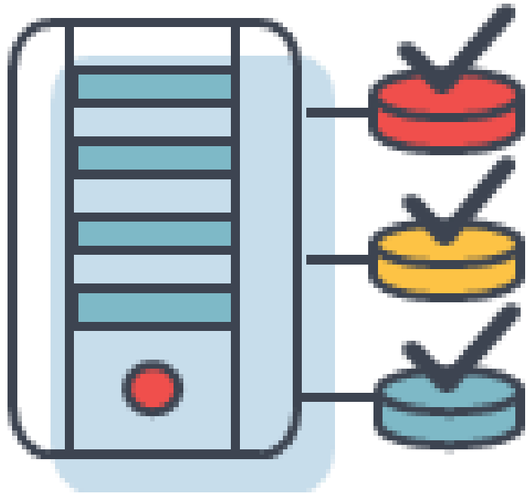
Velocity

- Velocity is the rate at which data is generated and transferred.
- Big Data provides data on demand and at a faster pace.
- Amazon records every click made when customers browse the website.

Variety

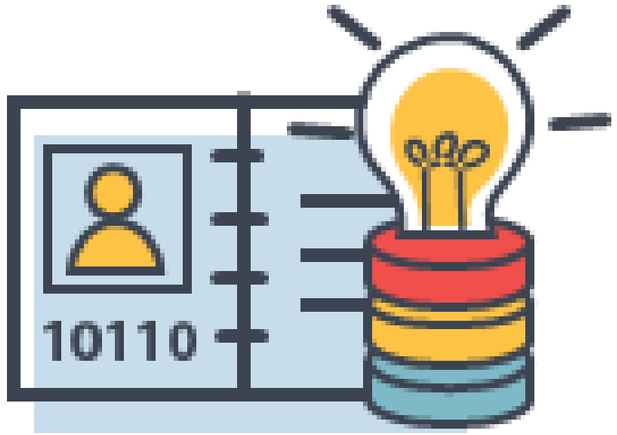
- Variety refers to all structured and unstructured data that can be generated by humans or machines.
- Social media, CRM systems, e-mails, audio, and video forms produce varied data.
- Analytics tools are used to segregate groups based on the type of data generated.

Veracity



- Veracity refers to the consistency, accuracy, and trustworthiness of data.
- It is required to filter and process unstructured and irrelevant data using analytics and algorithms to reveal meaningful information.

Value



- Value refers to the quality and accuracy of data.
- Data could be incomplete, erroneous, or incapable of providing genuine or useful information.
- Due to inherent discrepancies in the data collected, inaccurate predictions are made.

Unstructured Data Conundrum

The advent of big data seemed to rise to a new way of dealing with unstructured data. It allows to centralize the data and apply big data analytics to the problem.

Unstructured Data

Web Logs

Multimedia

Social Media

Semi-structured Data

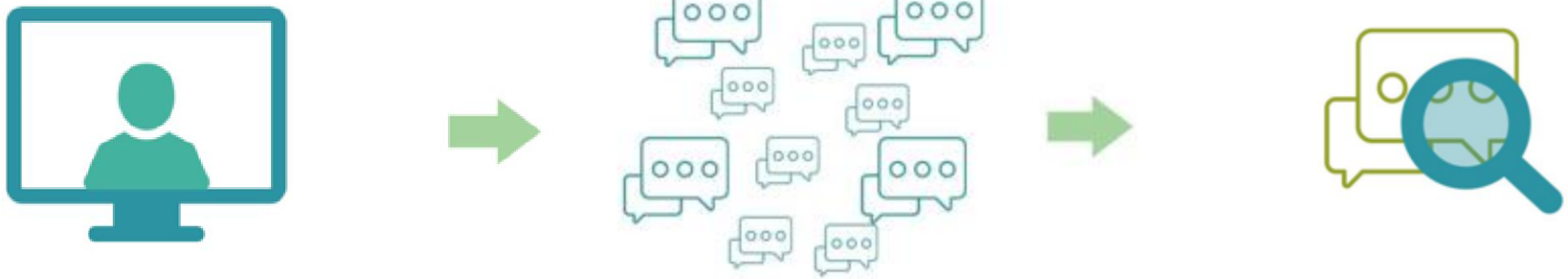
JSON

XML

Case Study: Royal Bank of Scotland

Case Study: Royal Bank of Scotland

100% of this data could be processed, whereas only 3% could be processed earlier with traditional systems.



250,000
Customer chats per month launched for multiple banking needs

Large volumes of unstructured web chat data

200
Chats manually extracted per month and analyzed for quality assurance

Previous Web Chat Analysis Approach

Challenges of Traditional System

Challenges of Traditional Systems (RDBMS And DWH)

There are three challenges in working with a Relational Databases and data warehousing systems:

Growth rate
RDBMS systems are designed for steady data retention rather than rapid growth.

Data size
Data ranges from terabytes (10^12 bytes) to exabytes (10^18 bytes).

Unstructured data
Relational databases can not categorize unstructured data.

Advantages of Big Data

Big Data provides users with the following advantages:

- Can run anywhere and hardware can be added
- Processes all types of data

- Has better decision-making
- Processes huge data quickly in real time

Companies Using Big Data

Some of the companies which use Big Data are as follows:

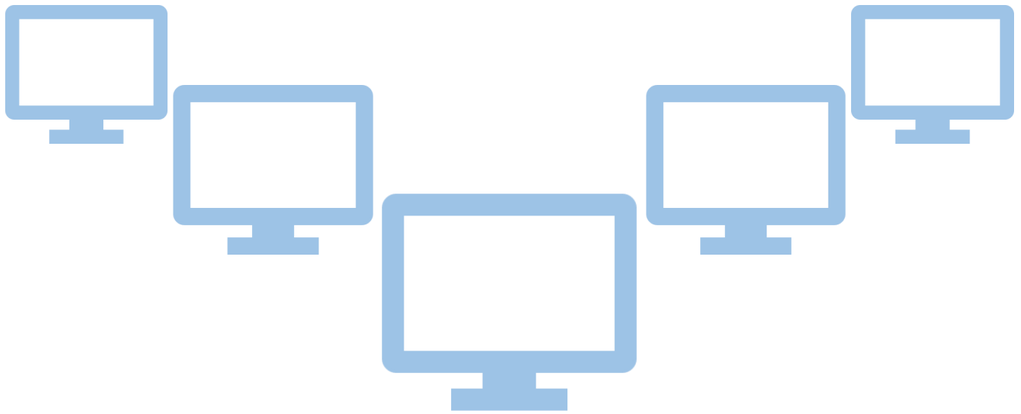
Case Study: Big Data in Netflix

Case Study: Big Data in Netflix

Netflix's enormous user base of over 148 million users offers it a big advantage when it comes to data collection. The following metrics are then focused on:

1. When do users watch a show?
2. Where do they watch it?
3. On which device do they watch the show?
4. How often do they pause a program?
5. How often do they re-watch a program?
6. Do they skip the credits?
7. What are the keywords searched?

Case Study: Big Data in Netflix



Multiple systems

Solution

- Traditionally, the analysis of such data was done using a computer algorithm that was designed to produce a correct solution for any given instance.
- As the data started to grow, a series of computers were employed to do the analysis.
- They are also known as distributed systems.

Distributed Systems

Distributed Systems

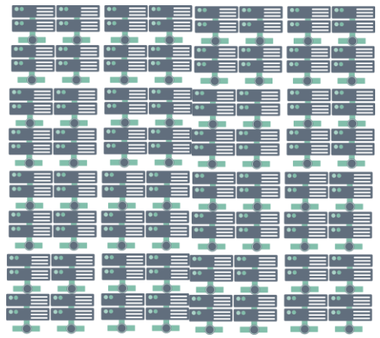
A distributed system is a model in which components located on networked computers communicate and coordinate their actions by passing messages.

How Does a Distributed System Work?

In recent times, distributed systems have been replaced by Hadoop.

1 Machine 4 I/O Channels Each Channel – 100 MB/s

Data =1 Terabyte

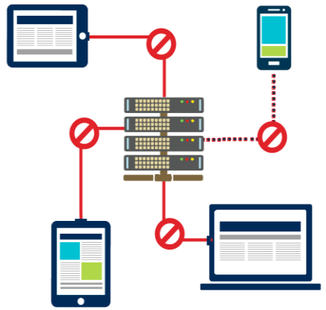


Data =1 Terabyte

100 Machine 4 I/O Channels Each Channel – 100 MB/s

Challenges of Distributed Systems

Since multiple computers are used in a distributed system, there are high chances of:



System failure

High programming complexity

Limited bandwidth

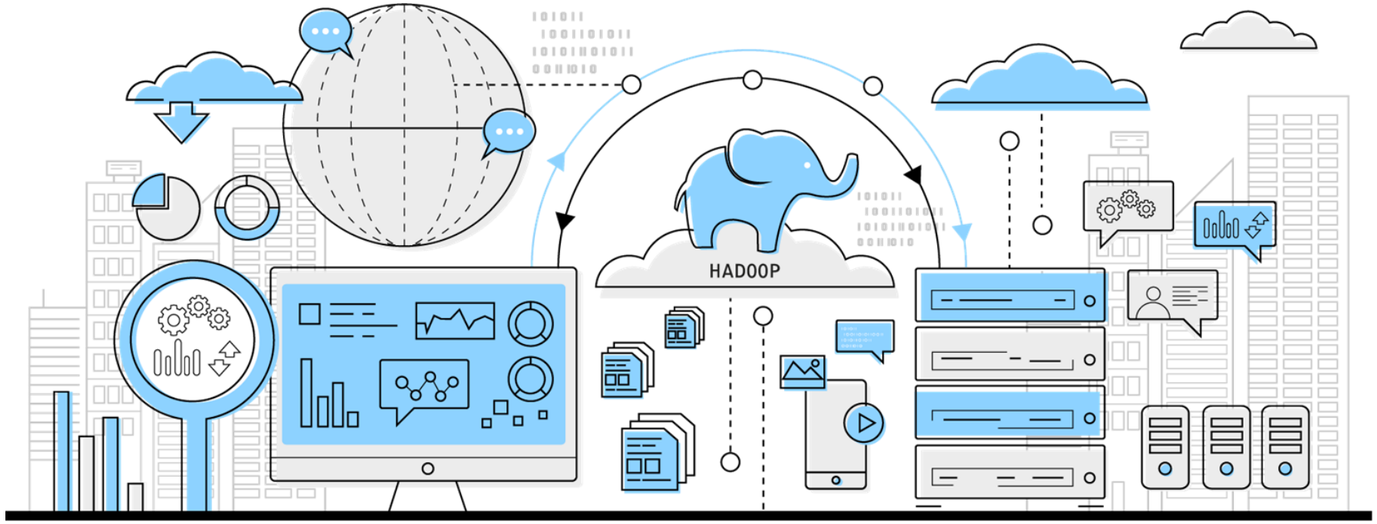
Solution

Hadoop overcomes these challenges.

Introduction to Hadoop

What Is Hadoop?

Hadoop is a framework that allows distributed processing of large datasets across clusters of commodity computers using simple programming models.



Characteristics of Hadoop

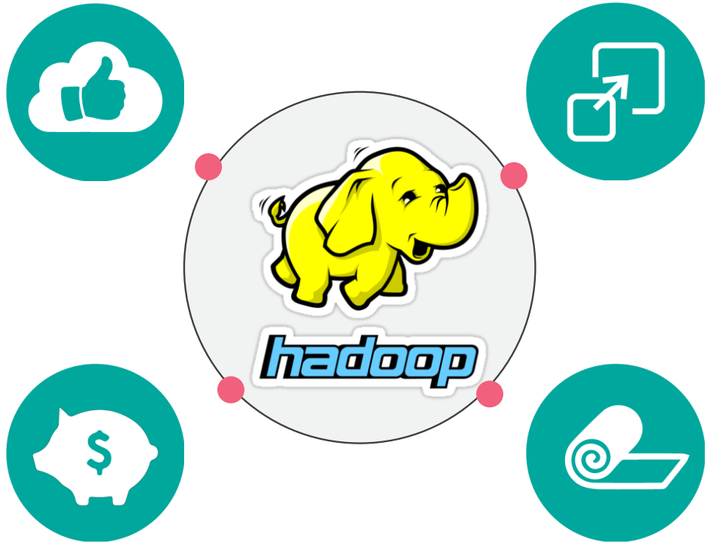
The four key characteristics of Hadoop are:

Reliable

Stores copies of the data on different machines and is resistant to hardware failure

Scalable

Can follow both horizontal and vertical scaling



Economical

Can use ordinary computers for data processing

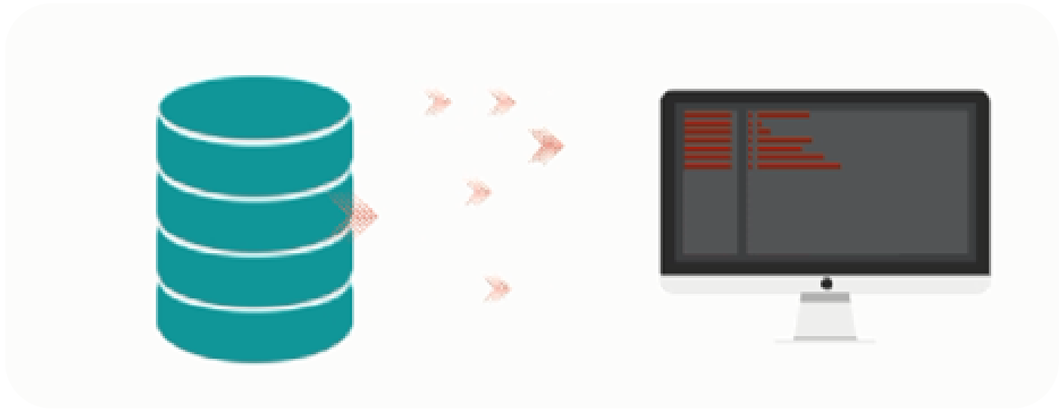
Flexible

Can store huge data and decide to use it later

Traditional Database Systems vs. Hadoop

The basic difference in functionality between traditional database systems and Hadoop are:

Traditional System



Data sent to the program

VS.

Hadoop

Program sent to the data

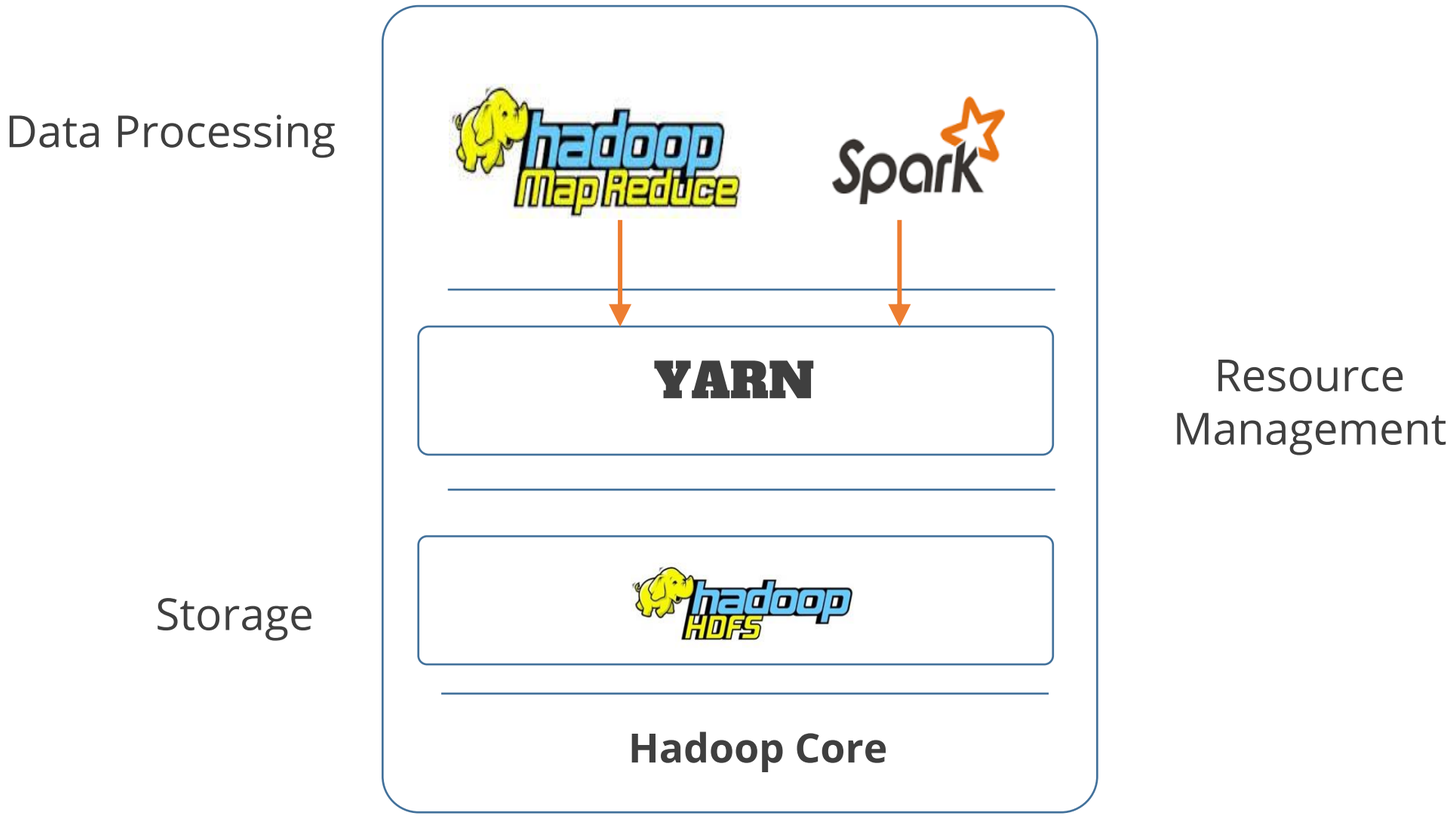
Traditional Database Systems vs. Hadoop

	RDMS	Hadoop
Data Types	Structured	Multi and unstructured
Processing	Limited, No data processing	Processing data in a distributed manner
Governance	Standards and structured	Loosely structured
Schema	Required on write	Required on read
Speed	Reads are fast	Writes are fast
Cost	Software license	Supports only
Resources	Known entity	Growing, Complexities, and wide
Best fit use	OLTP, Complex ACID Transactions, and operational data store	Data discovery, processing unstructured data, and massive storage or processing

Components of Hadoop Ecosystem

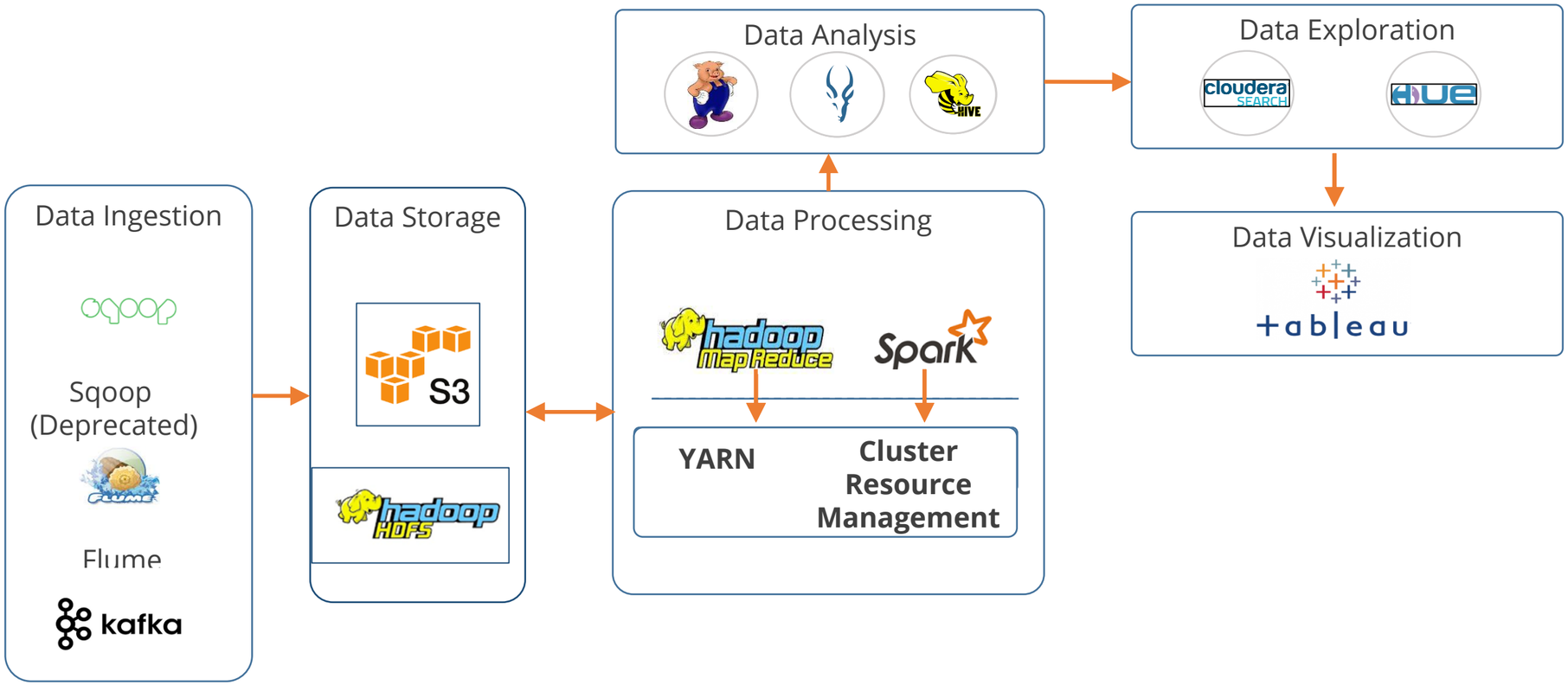
Hadoop Core Components

Hadoop is composed of three major components: HDFS, MapReduce, and YARN.



Components of Hadoop Ecosystem

These are the various components of the Hadoop ecosystem, which works together to provide data absorption, analysis, storage, and maintenance.

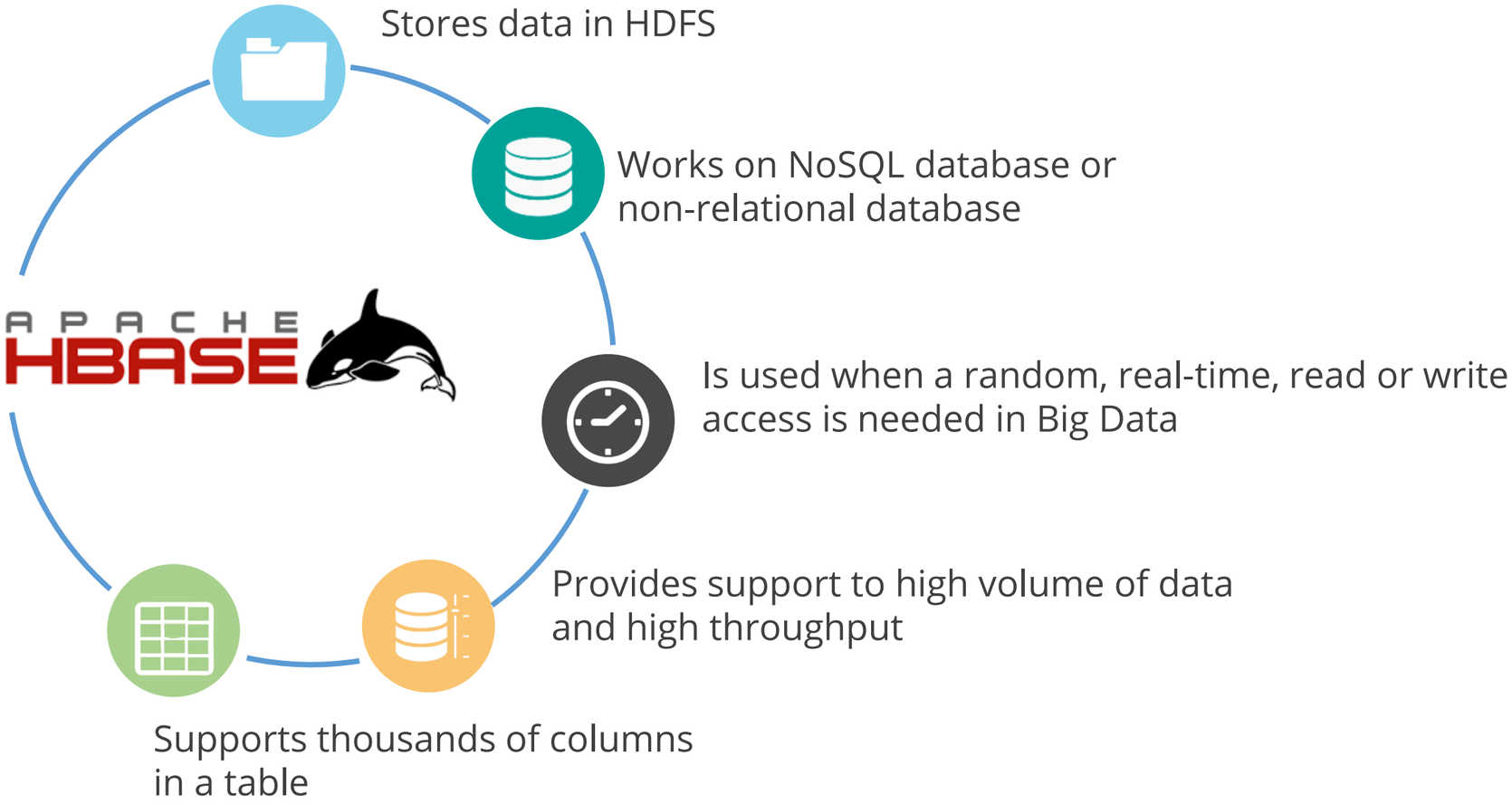


HDFS (Hadoop Distributed File System)

HDFS is a distributed file system that runs on commodity hardware and can handle massive data collections. Salient features of HDFS are as follows:

HBase

HBase is a column-oriented non-relational database management system that runs on top of the HDFS. Some of the features of Hbase are as follows:



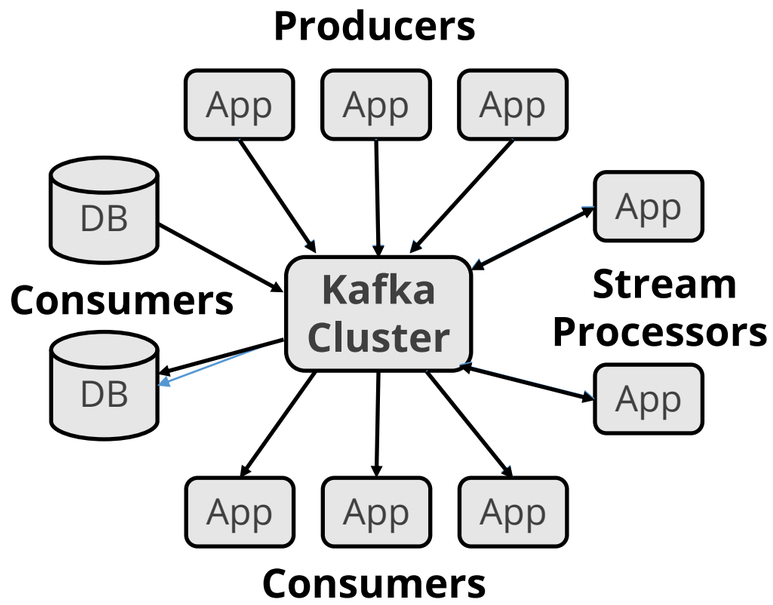
SQOOP

- Sqoop is a tool designed to transfer data between Hadoop and relational database servers, such as Oracle and MySQL to HDFS, and export data from HDFS to relational database.

- The order of columns in both MySQL and Hive should be the same while importing or exporting.

KAFKA

- Kafka is an open-source distributed event stream platform that provides high-performance data pipelines, streaming analytics, and data integration for critical applications.
- Kafka has a modern, cluster-centric design that offers strong durability and fault-tolerance guarantees.



Flume

Flume is used to ingest event data, such as streaming data, sensor data, and log files.

Distributed service for ingesting streaming data

Ideally suited for event data from multiple systems

Spark

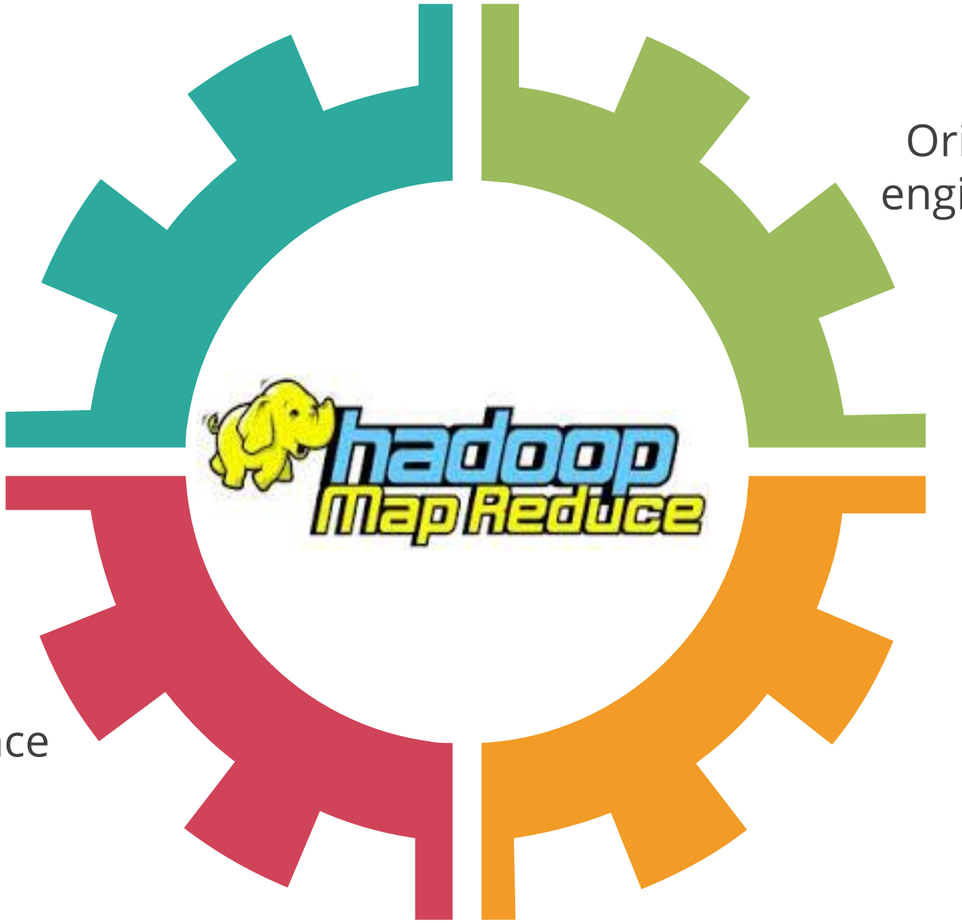
- It is an open-source cluster computing framework.
- Spark provides 100 times faster performance than MapReduce.
- Supports machine learning, business intelligence, streaming, and batch processing

Hadoop MapReduce

Some of Hadoop MapReduce features are:

Commonly used

Original Hadoop processing engine which is primarily Java-based

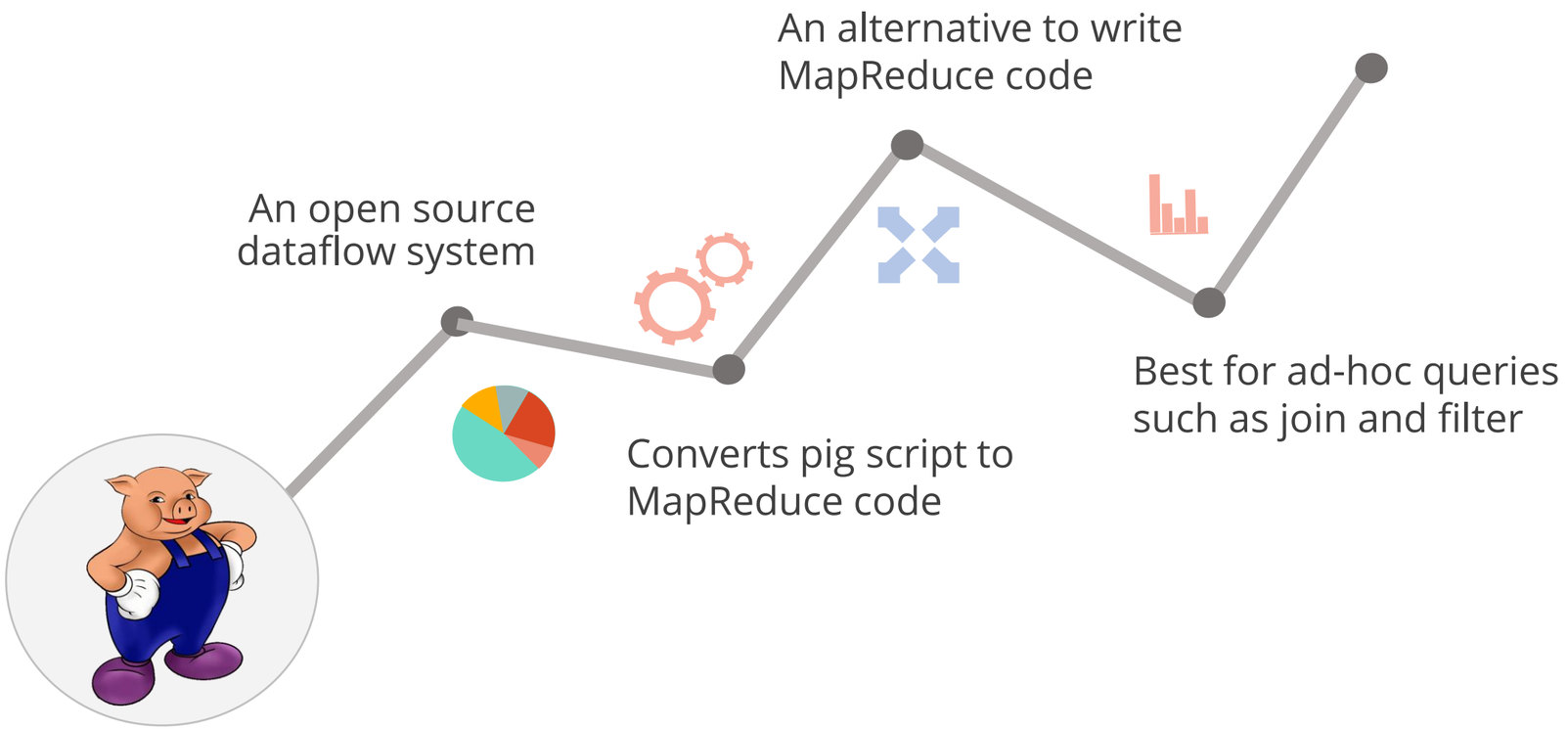


An extensive and mature fault tolerance framework

Based on the map and reduce programming model

Pig

Pig is a high-level data flow framework for executing Hadoop MapReduce programs.
Pig Latin is the language used in Pig.



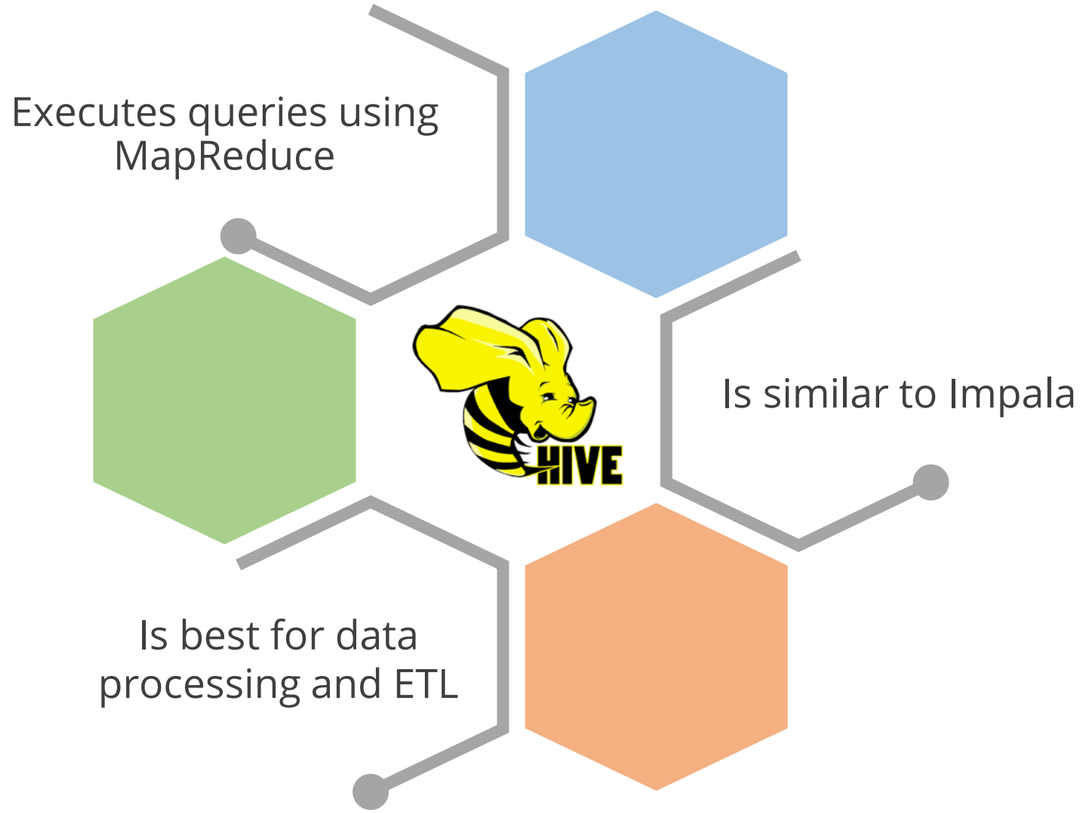
Impala

Impala is a SQL query engine that uses MPP (Massive Parallel Processing) to process large amounts of data stored in a Hadoop cluster.

- Is a high-performance SQL engine that runs on the Hadoop cluster
- Is ideal for interactive analysis
- Has very low latency –measured in milliseconds
- Supports a dialect of SQL (Impala SQL)

Hive

Hive is a data warehouse system that is used to analyze structured data. It is built on the top of Hadoop.



Cloudera Search

It is an Apache Solr that is fully integrated into the Cloudera platform, taking advantage of the flexible, scalable, and robust storage system and data processing frameworks included in the Cloudera Data Platform (CDP).

Is one of Cloudera's near-real-time access products

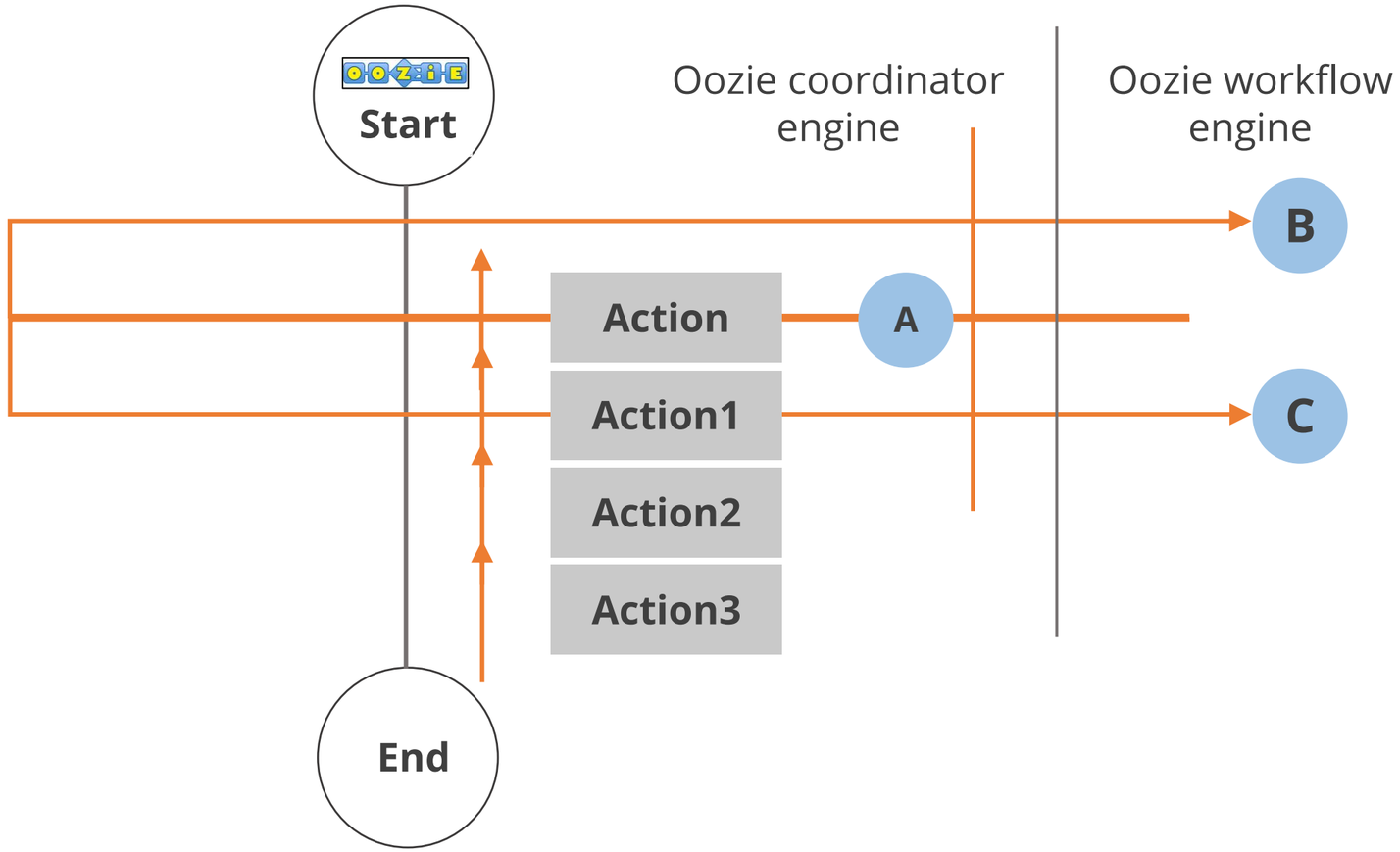
Eliminates the need to move large datasets across infrastructures to address business tasks

Enables nontechnical users to search and explore data stored in or ingested into Hadoop and HBase

Provides a fully integrated data processing platform

Oozie

Oozie is a workflow or coordination system that manages the Hadoop jobs.



Hadoop User Experience (HUE)

Hue is an acronym for Hadoop User Experience.

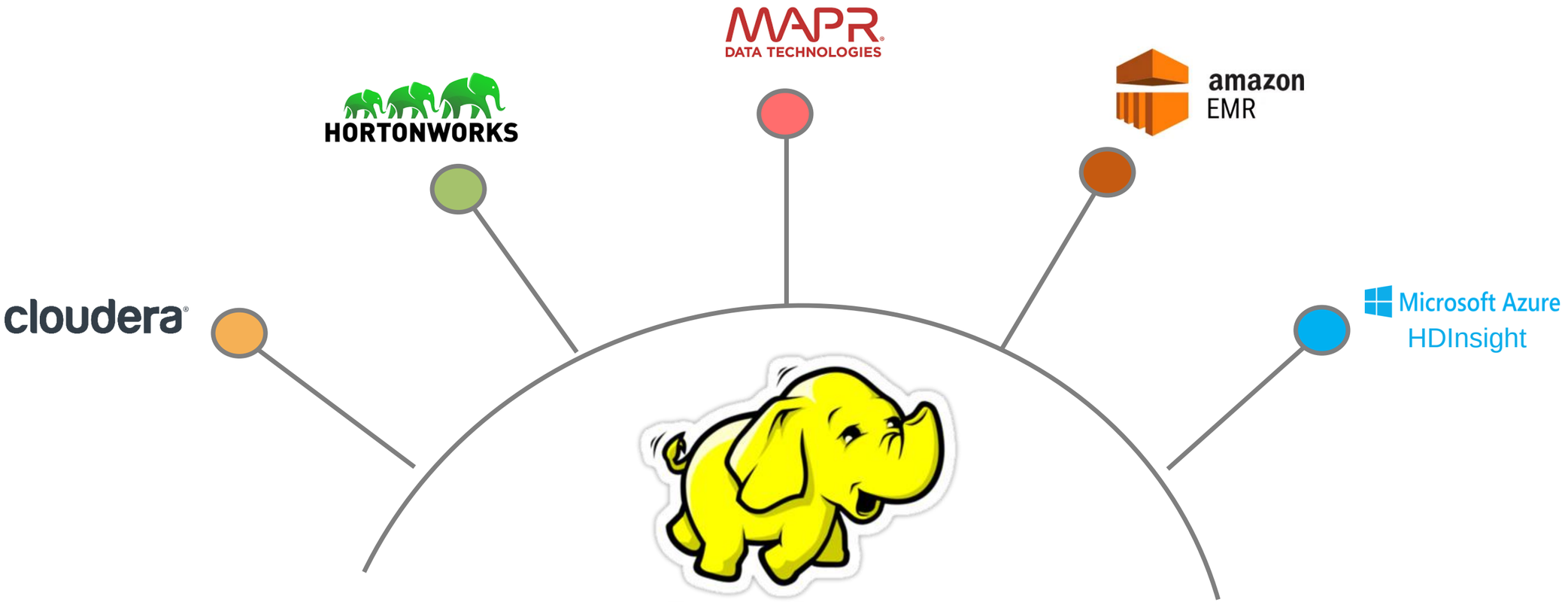
It provides SQL editors for Hive, Impala, MySQL, Oracle, PostgreSQL, Spark SQL, and Solr SQL.

Hue is an open-source Web interface for analyzing data with Hadoop.

Commercial Hadoop Distributions

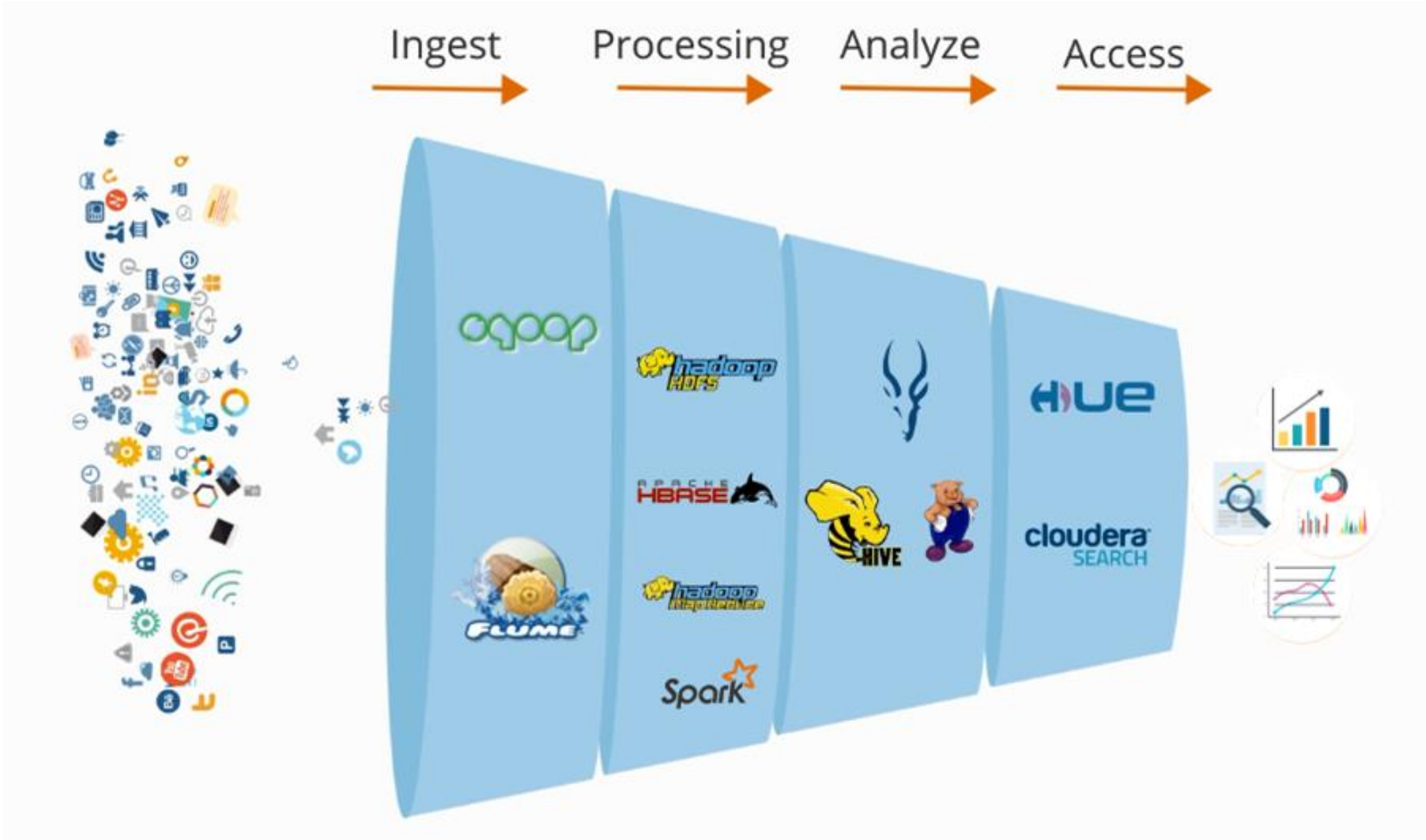
Various Commercial Hadoop Distributions

Following are some of the commercial Hadoop Distributions:



Big Data Processing

Components of Hadoop ecosystem work together to process big data. There are four stages of big data processing:



Assisted Practice: Walk-Through of the Simplilearn Cloud Lab

Duration: 10 Minutes

ASSISTED PRACTICE

Problem Statement: In this demonstration, you will walk-through through the Simplilearn cloud lab.

Objective: In this Assisted Practice, You will learn how to log in to the Simplilearn lab.

Tasks to Perform:

Step 1: Click on the **Practice Labs** tab on the left side panel of the LMS

Step 2: Copy the username and password that are generated

Step 3: Click on the **Launch Lab** button

Step 4: Enter the username and password in the respective fields on the page and click **Login**

Key Takeaways

- Big Data has high volume, variety, velocity, veracity, and value.
- Hadoop is a framework that allows the distributed processing of large datasets across clusters of commodity computers using simple programming models.
- The Hadoop Ecosystem is a platform that offers a variety of services to address big data issues. It consists of Apache projects and a variety of commercial tools and solutions.

Knowledge Check

Knowledge Check 1

Which Hadoop tool is an alternative to writing MapReduce code in a high-level language?

A. Pig

B. Hive

C. Spark

D. Impala

Knowledge Check 1

Which Hadoop tool is an alternative to writing MapReduce code in a high-level language?

A. Pig

B. Hive

C. Spark

D. Impala

The correct answer is **A**

Pig Latin is a high-level language for writing data analysis programs that are converted into MapReduce jobs through Pig.

Knowledge Check 2

Which tool can perform random read and write operations on datasets of petabyte size?

A. Hive

B. Impala

C. HBase

D. HDFS

Knowledge Check 2

Which tool can perform random read and write operations on datasets of petabyte size?

A. Hive

B. Impala

C. HBase

D. HDFS

The correct answer is **C**

HBase is a database that leverages milliseconds of delay to read and write data even at a petabyte scale.

Knowledge Check

3

Which of the following is a source of unstructured data?

A. Data from social media websites

B. Transactional data in Amazon’s database

C. Web and server logs

D. All of the above

Knowledge Check

3

Which of the following is a source of unstructured data?

A. Data from social media websites

B. Transactional data in Amazon's database

C. Web and server logs

D. All of the above

The correct answer is **A**

Unstructured data comprises data that is usually not easily searchable, including formats like audio, video, and social media postings.

Knowledge Check

4

What is Hadoop?

A. It is an in-memory tool used in Mahout algorithm computing.

B. It is a computing framework used for resource management.

C. It is a framework that allows distributed processing of large datasets across clusters of commodity computers using a simple programming model.

D. It is a search and analytics tool that provides access to analyze data.

Knowledge Check

4

What is Hadoop?

A. It is an in-memory tool used in Mahout algorithm computing.

B. It is a computing framework used for resource management.

C. It is a framework that allows distributed processing of large datasets across clusters of commodity computers using a simple programming model.

D. It is a search and analytics tool that provides access to analyze data.

The correct answer is **C**

Hadoop is a framework that allows distributed processing of large datasets across clusters of commodity computers using a simple programming model.

Thank You