Introduction to Big Data and Hadoop

Learning Objectives

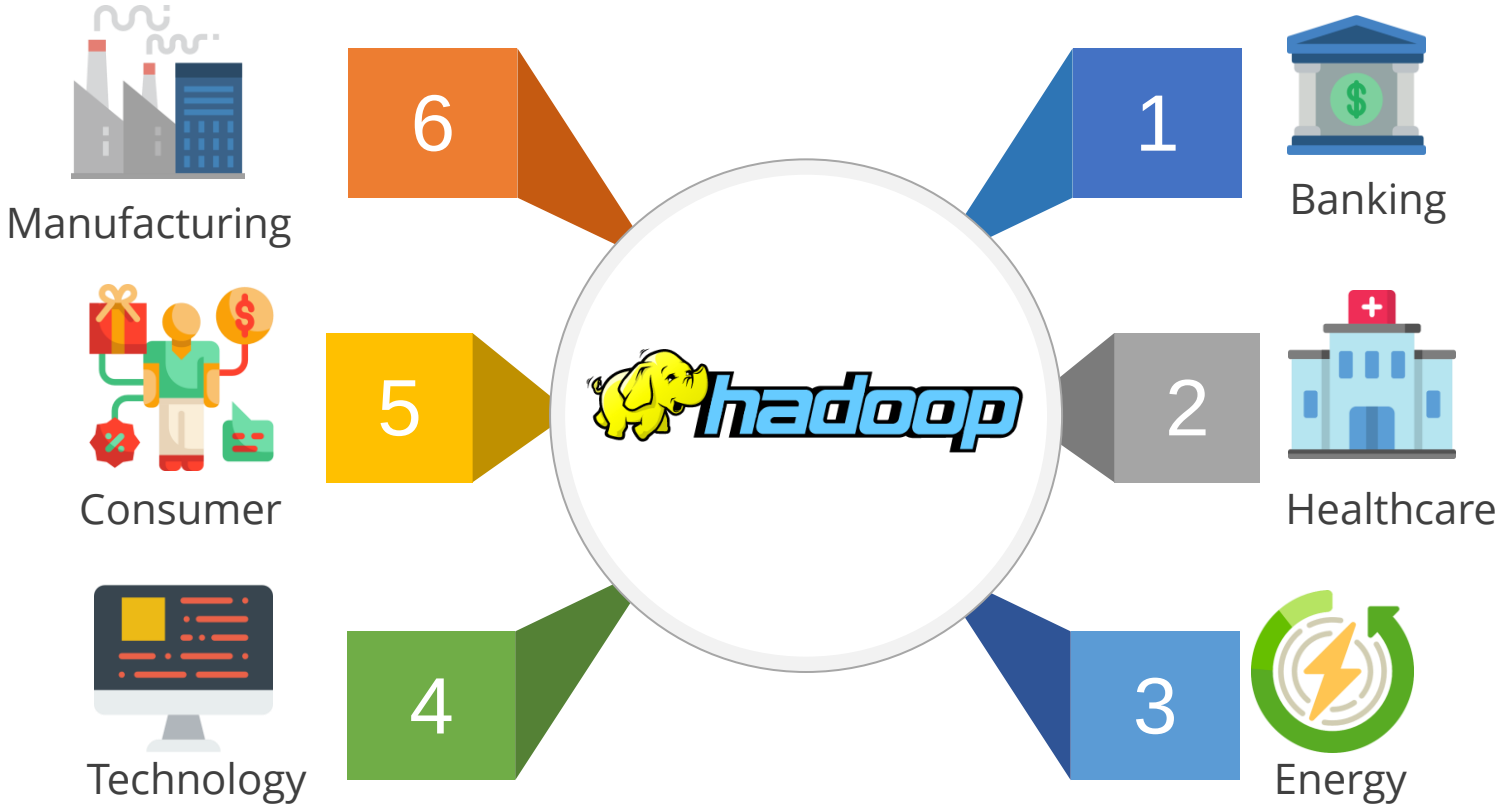
By the end of this lesson, you will be able to:

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

Introduction to Big Data

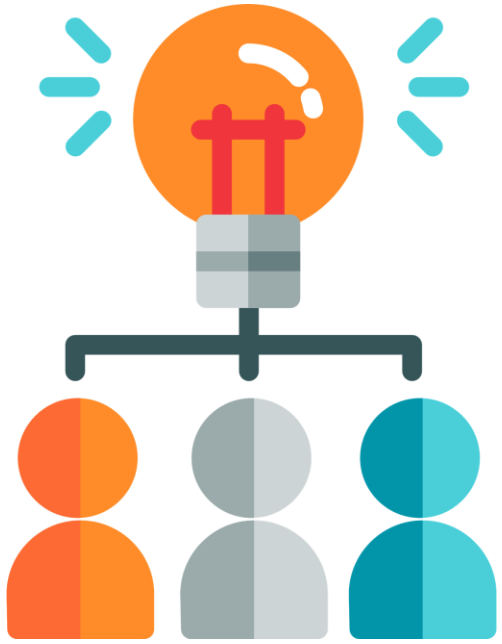
Big Data Overview

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

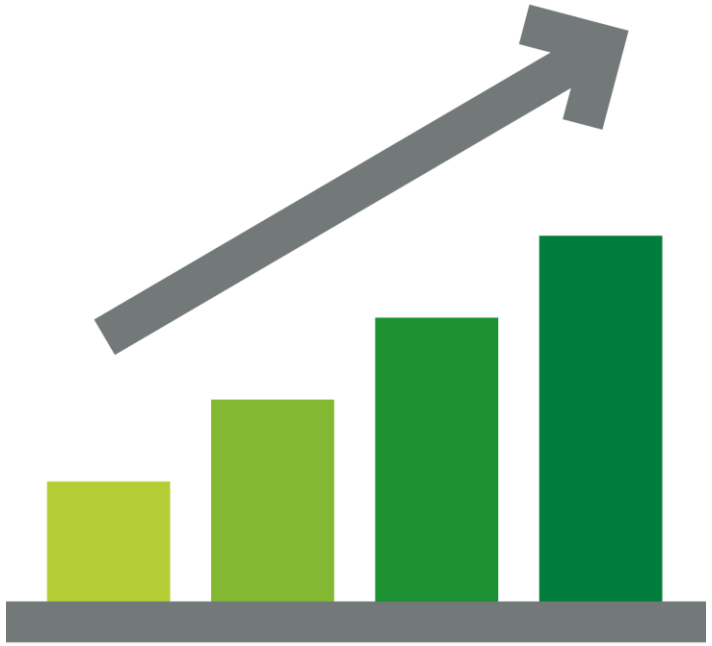


According to US Bureau of Labour Statistics, Big Data alone will fetch 11.5 million jobs by 2026.

Traditional Decision-Making



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

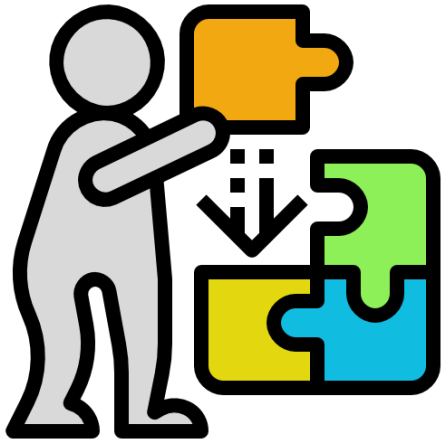
Lacks systematic linkage among strategy, planning, execution, and reporting

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



Solution

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

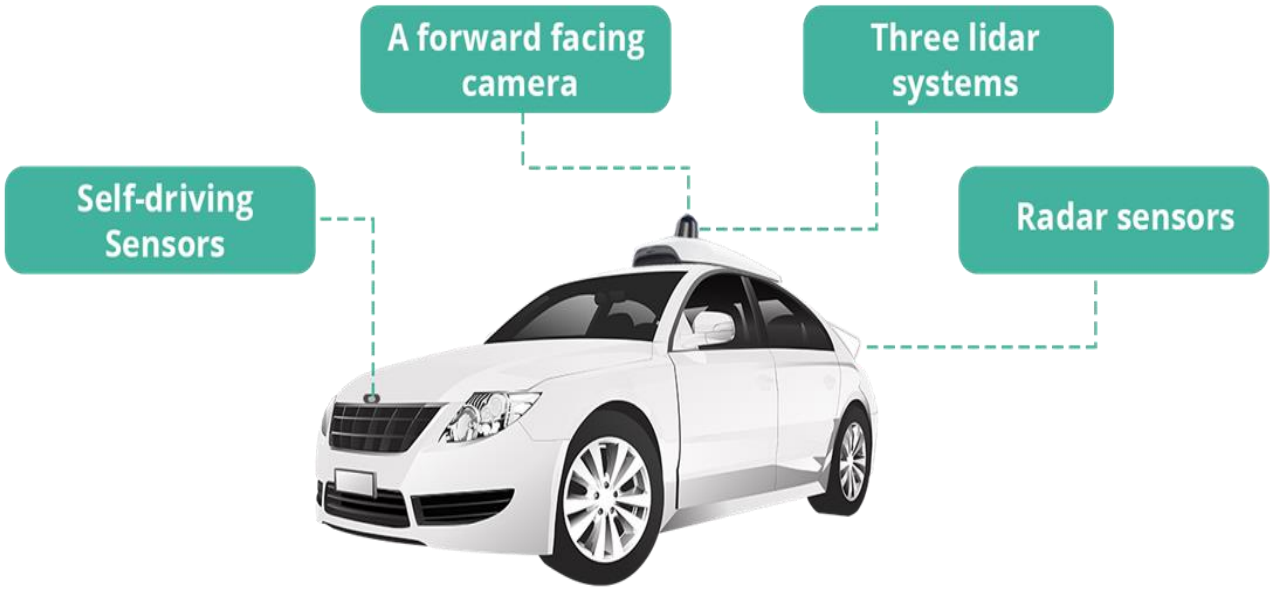
It provides a comprehensive view of the overall picture which is a result of analyzing data from various sources.

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.

Case Study: Google's Self-Driving Car

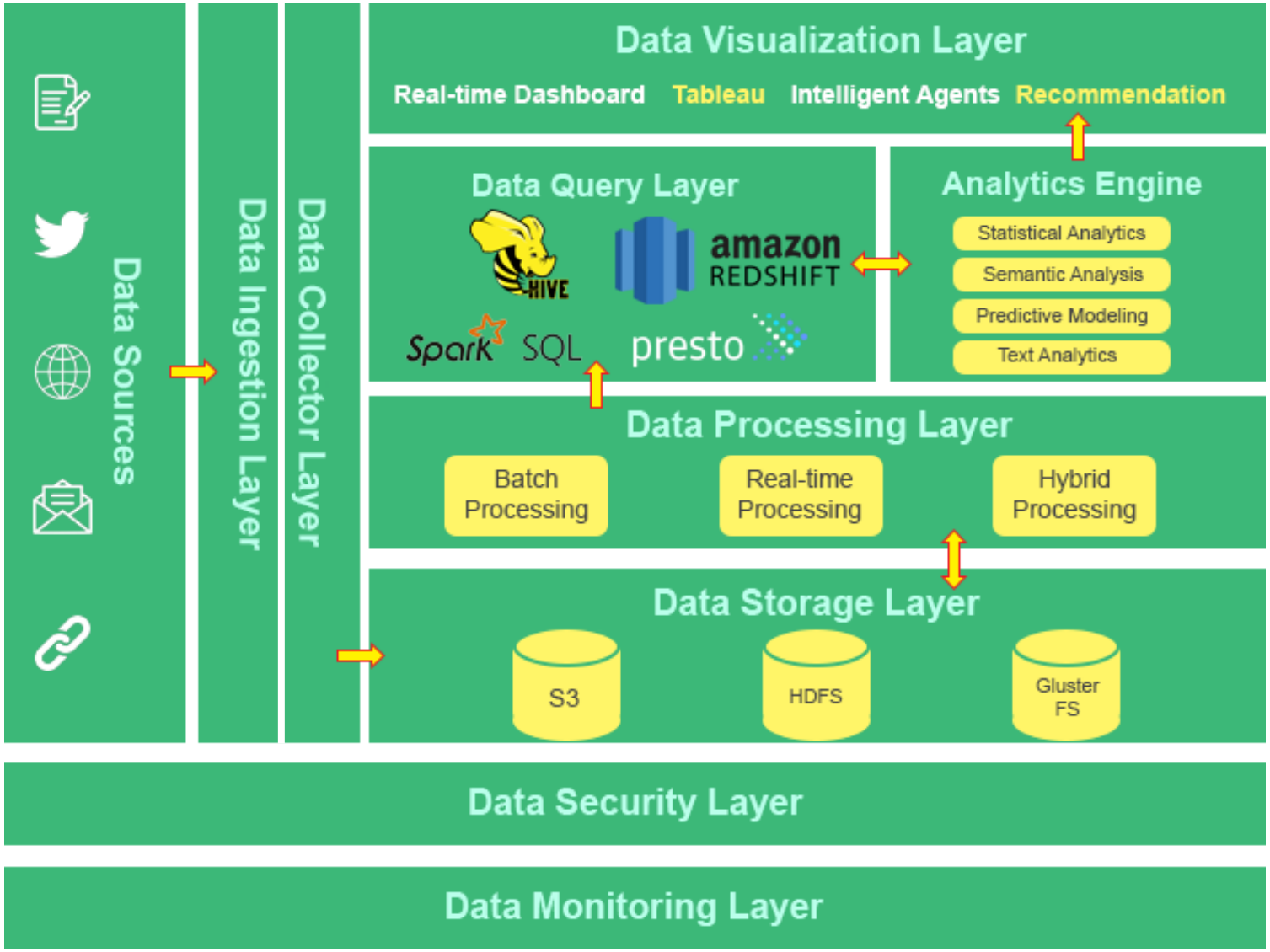


Technical Data

Community Data

Personal Data

Big Data Analytics Pipeline

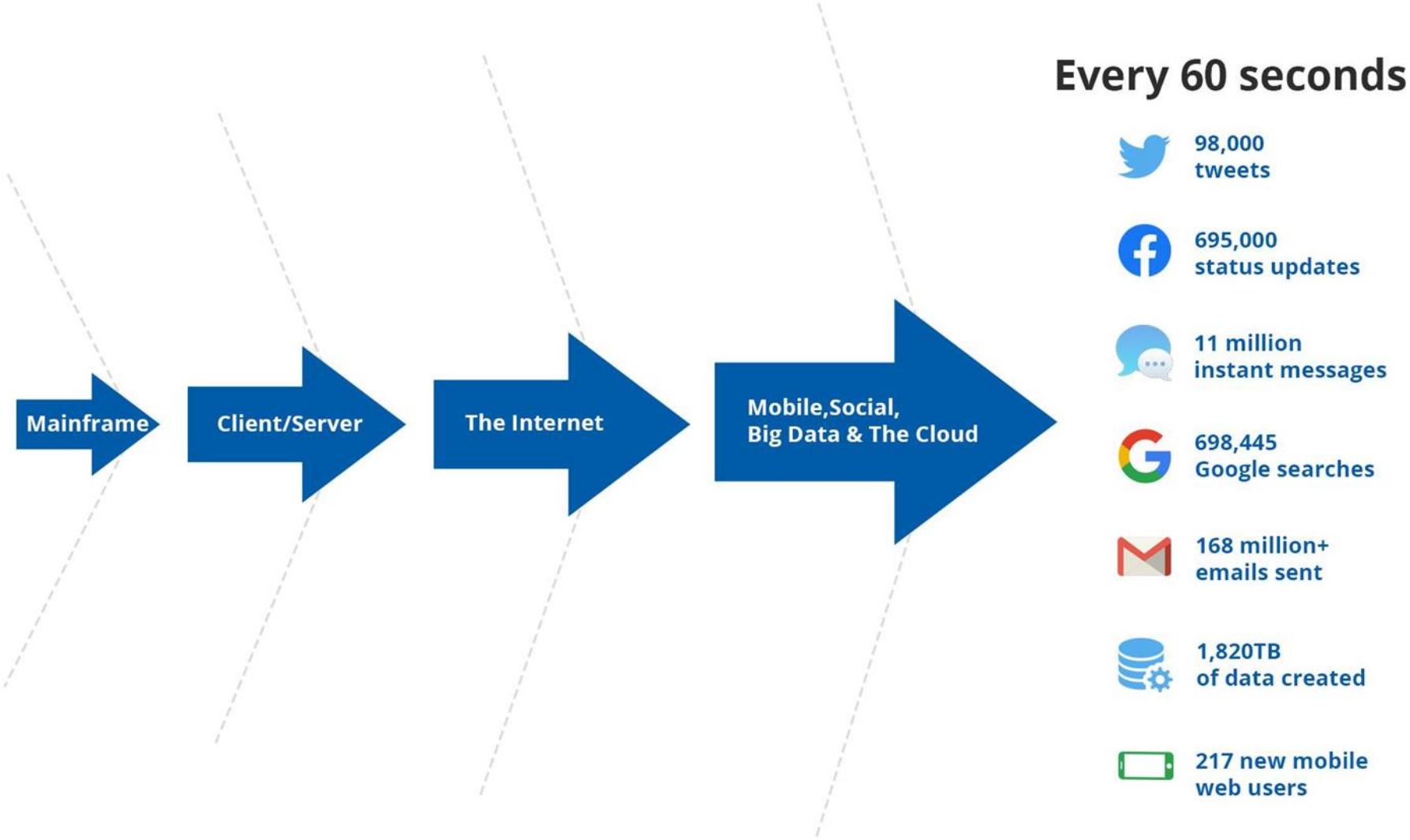


What Is Big Data?

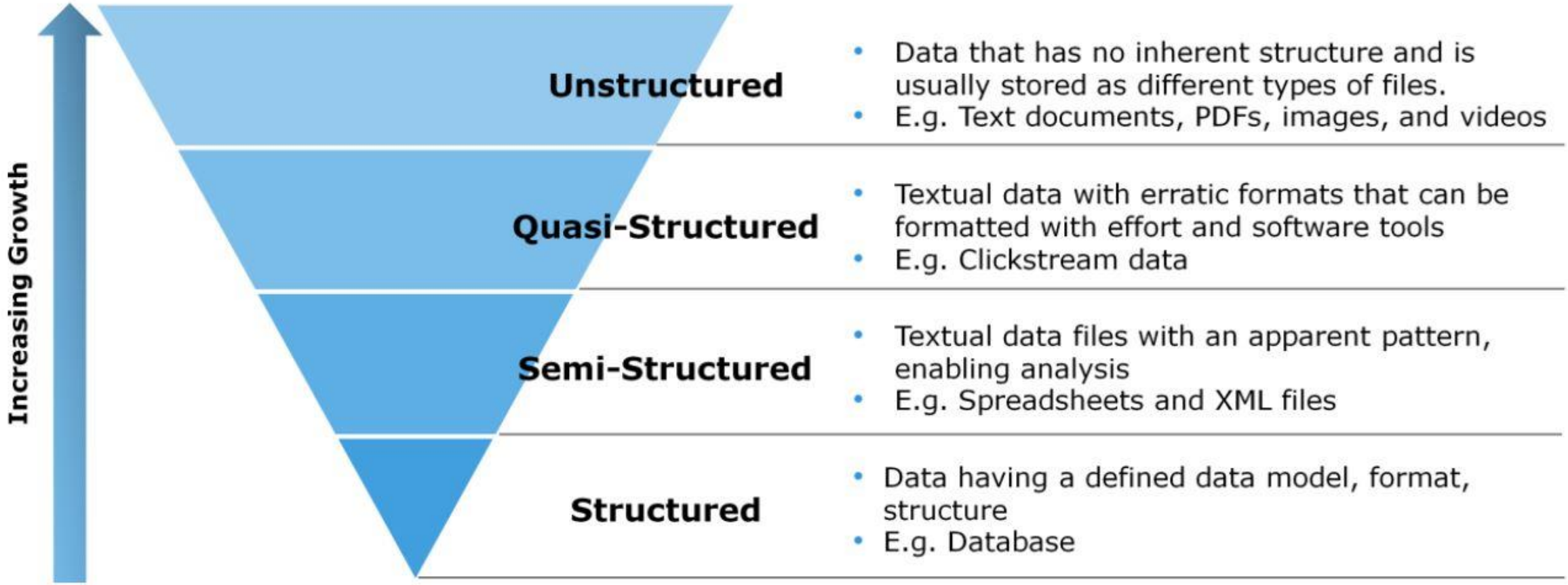
What Is Big Data?

> Big data refers to extremely 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

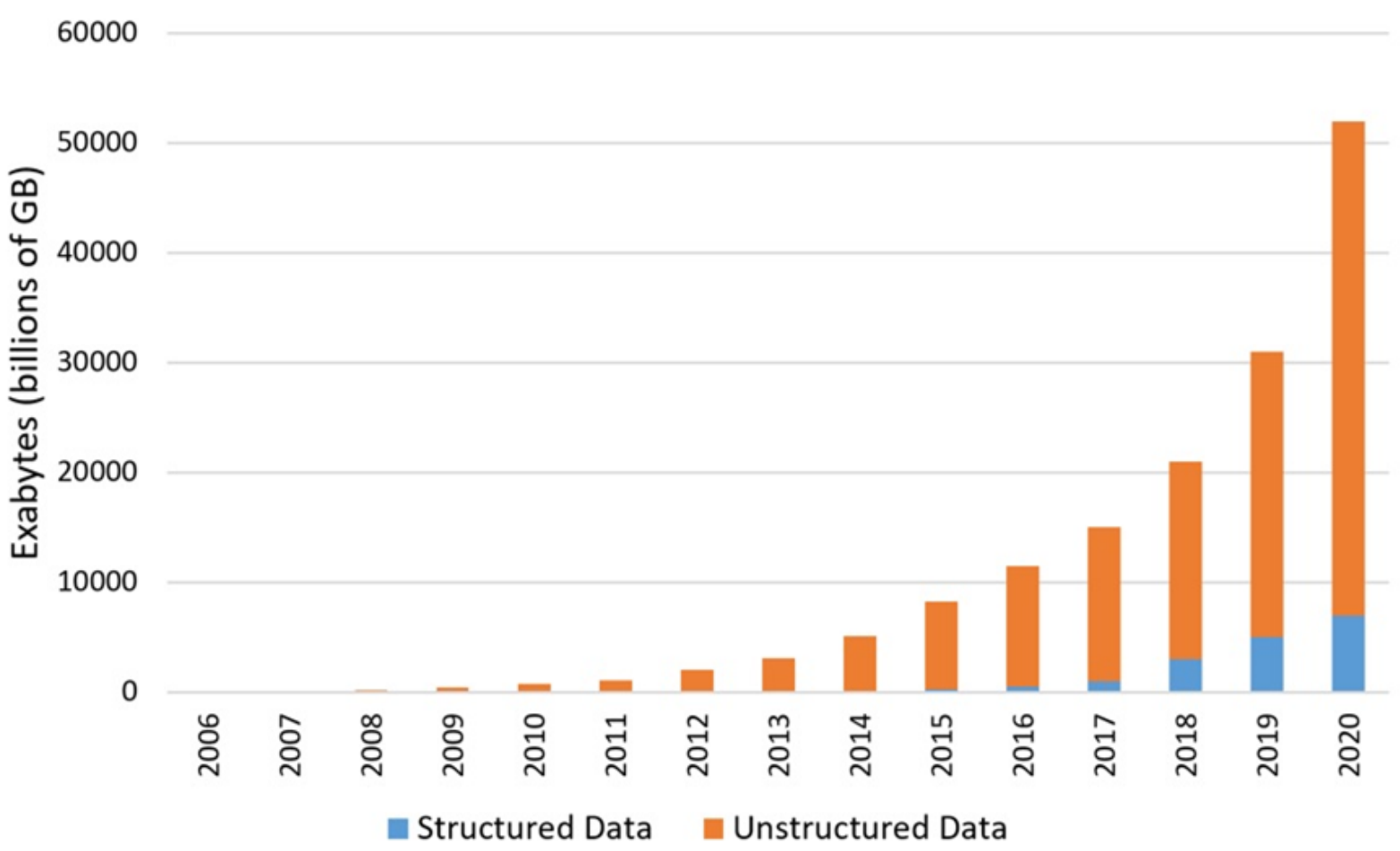


Different Types of Data



Growth in Data

By 2020, data will show an exponential rise!



Four Vs of Big Data

Four Vs of Big Data

Volume

- Overall amount of information produced every day is rising exponentially
- 2.3 trillion gigabytes of data is generated every day on the internet

Variety

- 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

Inherent discrepancies in the data collected results in inaccurate predictions

Velocity

- More than 50,000 Google searches are completed
- More than 125,000 YouTube videos are viewed
- 7,000 tweets are sent out
- More than 2 million e-mails are sent

Unstructured Data Conundrum

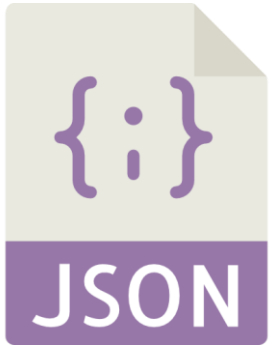
Unstructured Data

Web Logs

Multimedia

Social Media

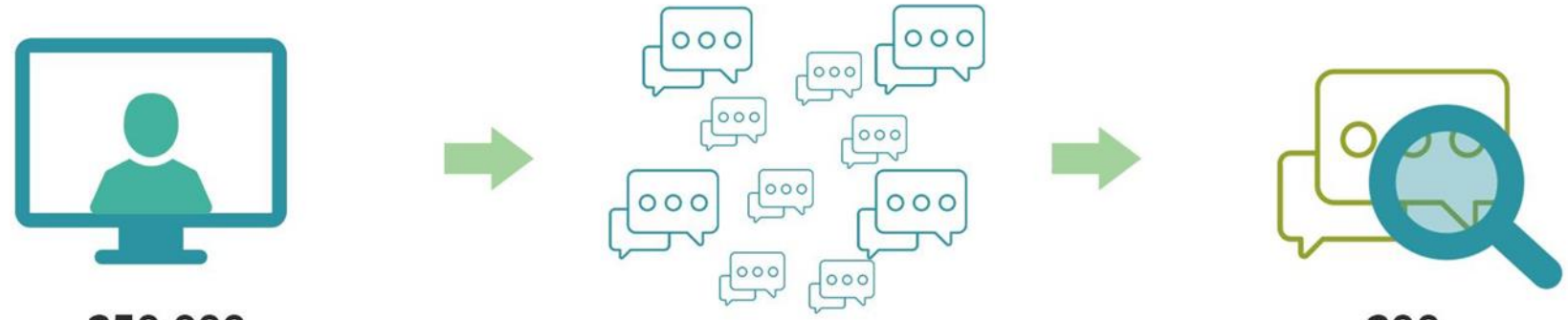
Semi-structured Data



Case Study: Royal Bank of Scotland

Case Study: Royal Bank of Scotland

Previous Web Chat Analysis Approach



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

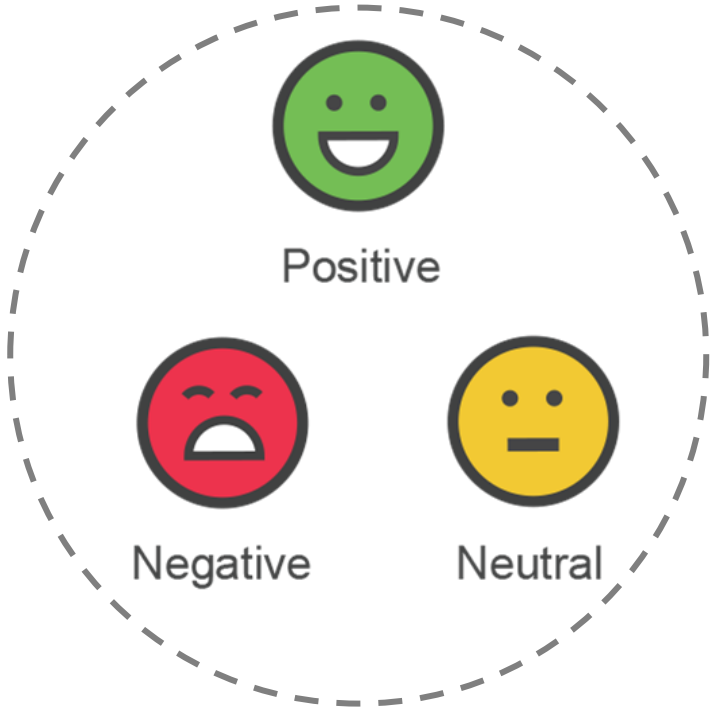
Large volumes of unstructured web chat data

200
chats manually extracted per month and analysed for quality assurance

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

Case Study: Royal Bank of Scotland

The case study of Royal Bank of Scotland gave the following three things:



Sentiment analysis

Reduced processing time

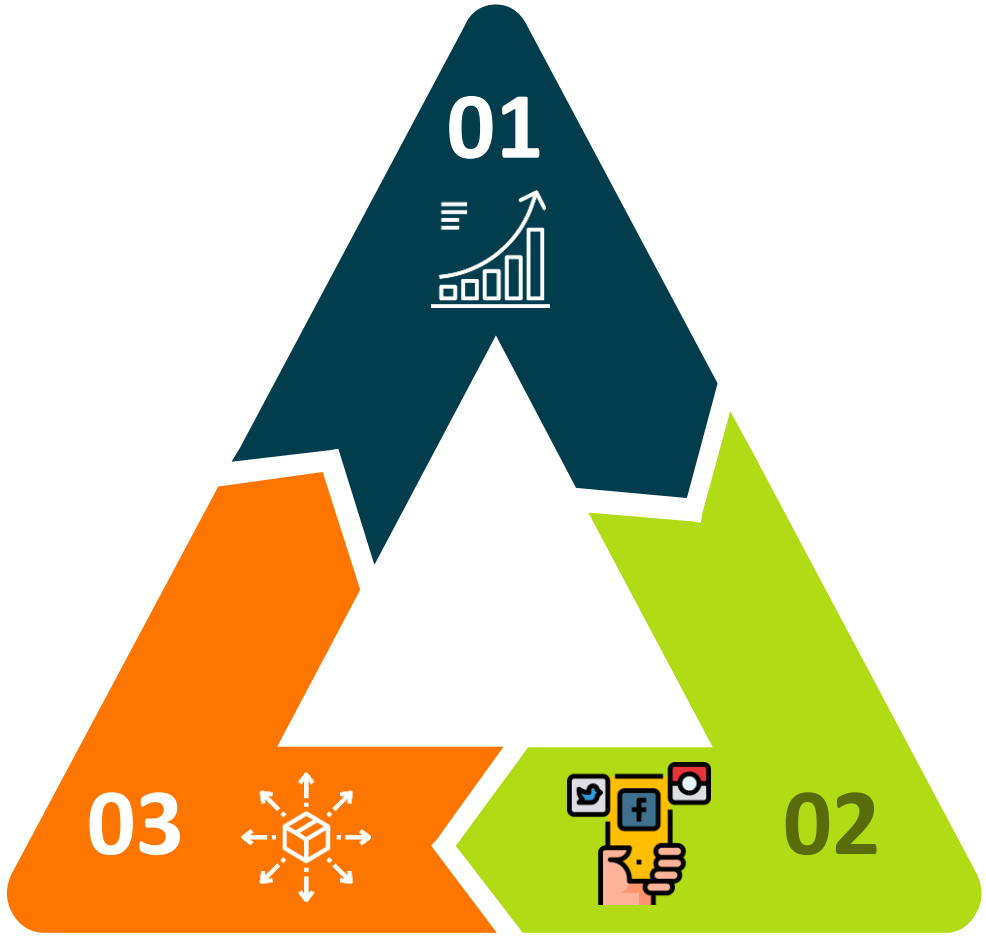
Improved customer satisfaction

Challenges of Traditional System

Challenges of Traditional Systems (RDBMS and DWH)

GROWTH RATE

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



DATA SIZE

Data ranges from terabytes ($10^{\wedge 12}$ bytes) to exabytes ($10^{\wedge 18}$ bytes).

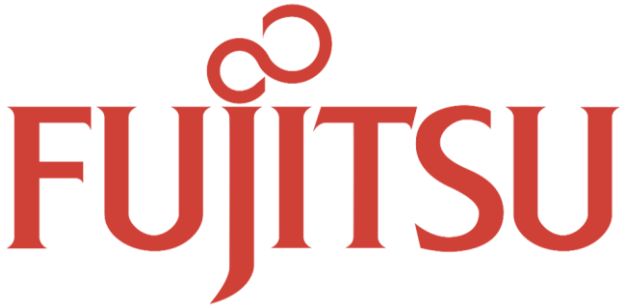
UNSTRUCTURED DATA

Relational databases can't categorize unstructured data.

Advantages of Big Data

1. Processes all types of data at scale
2. Processes huge data quickly in real-time
3. Can run anywhere and additional hardware can be added
4. Better decision-making, thanks to Hadoop

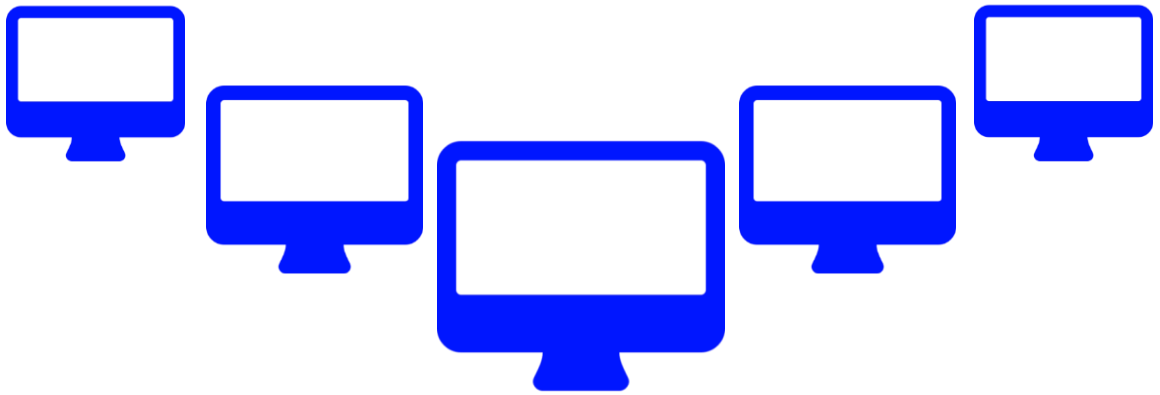
Companies Using Big Data



Big Data: Case Study

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?

Big Data: Case Study



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 were also known as distributed systems.

Features of Big Data Analytics

Scalability in Big Data

- A scalable data platform accommodates rapid changes in the growth of data, either in traffic or volume.
- It utilizes and adds hardware or software to increase the output and storage of data.
- When a company has a scalable data platform, it is prepared for the potential of growth in its data needs.

Fault Tolerance in Big Data

- **Fault tolerance** in Big data or Hadoop HDFS refers to the working strength of a system in unfavorable conditions and how that system can handle such a situation.

- HDFS also maintains the replication factor by creating a replica of **data** on other available machines in the cluster if one machine fails unexpectedly.

Data Inconsistency in Big Data

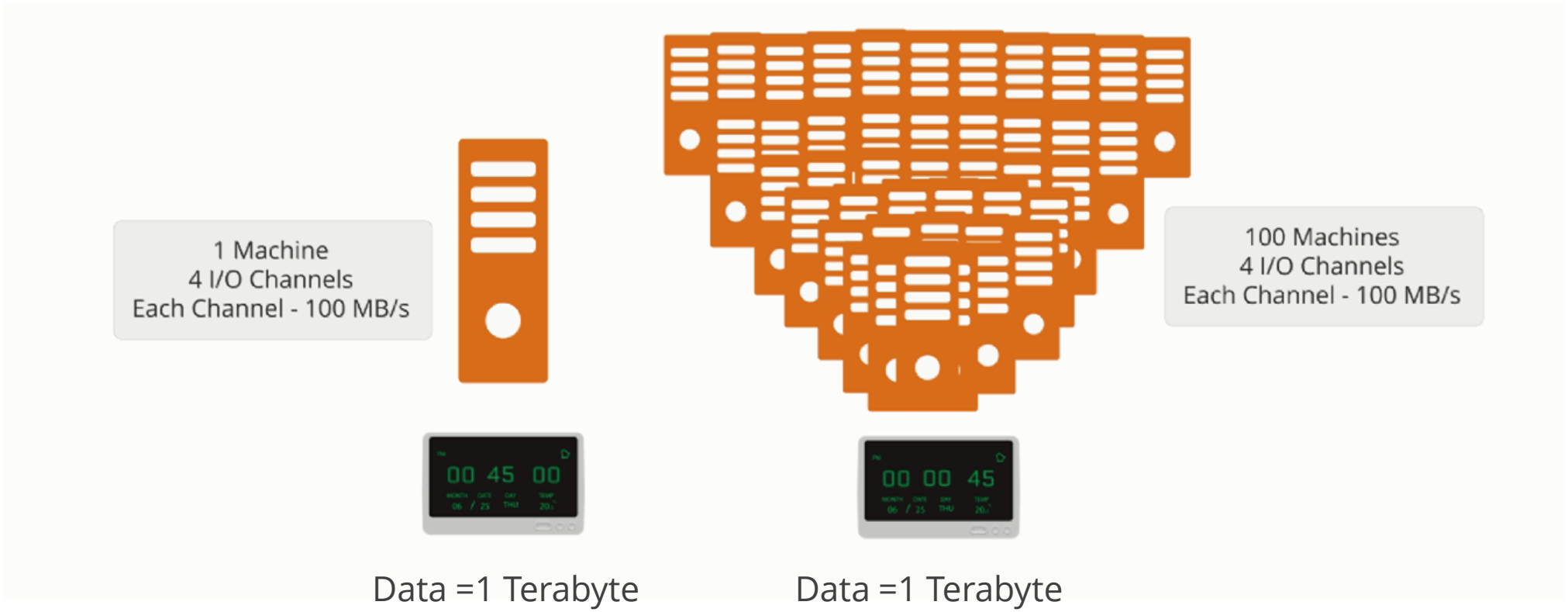
- Once data is captured in **big data**, **inconsistent** or conflicting phenomena can occur at various granularities.
- It occurs from knowledge content, **data**, information, knowledge, meta-knowledge, to expertise, and can adversely affect the quality of the outcomes in **Big data** analysis process.

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.

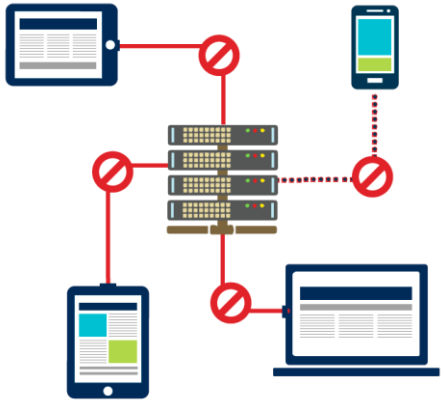
How Does a Distributed System Work?



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

Challenges of Distributed Systems

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

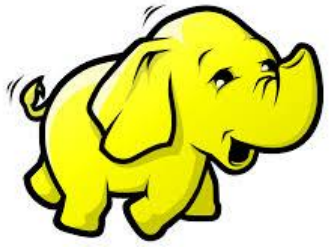


1 System failure

2 Limited bandwidth

3 High programming complexity

Any solution?



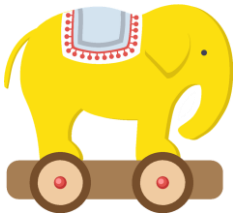
Hadoop

Introduction to Hadoop

What Is Hadoop?

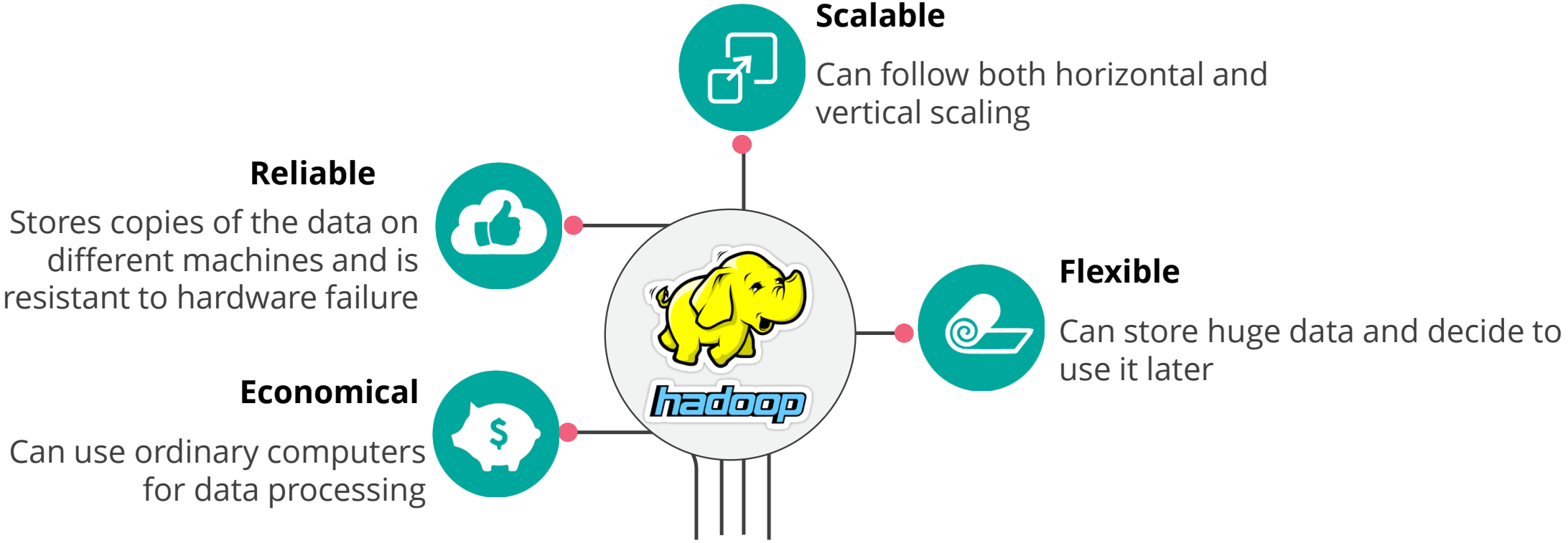
Hadoop is a framework that allows distributed processing of large datasets across clusters of commodity computers using simple programming models.

Doug Cutting discovered Hadoop and named it after his son's yellow toy elephant. It is inspired by the technical document published by Google.



Characteristics of Hadoop

The four key characteristics of Hadoop are:



Traditional Database Systems vs. Hadoop

Traditional System

Data sent to the program

VS.

Hadoop

Program sent to the data

Analogy of Traditional System and Hadoop

Human brings food toward the mouth

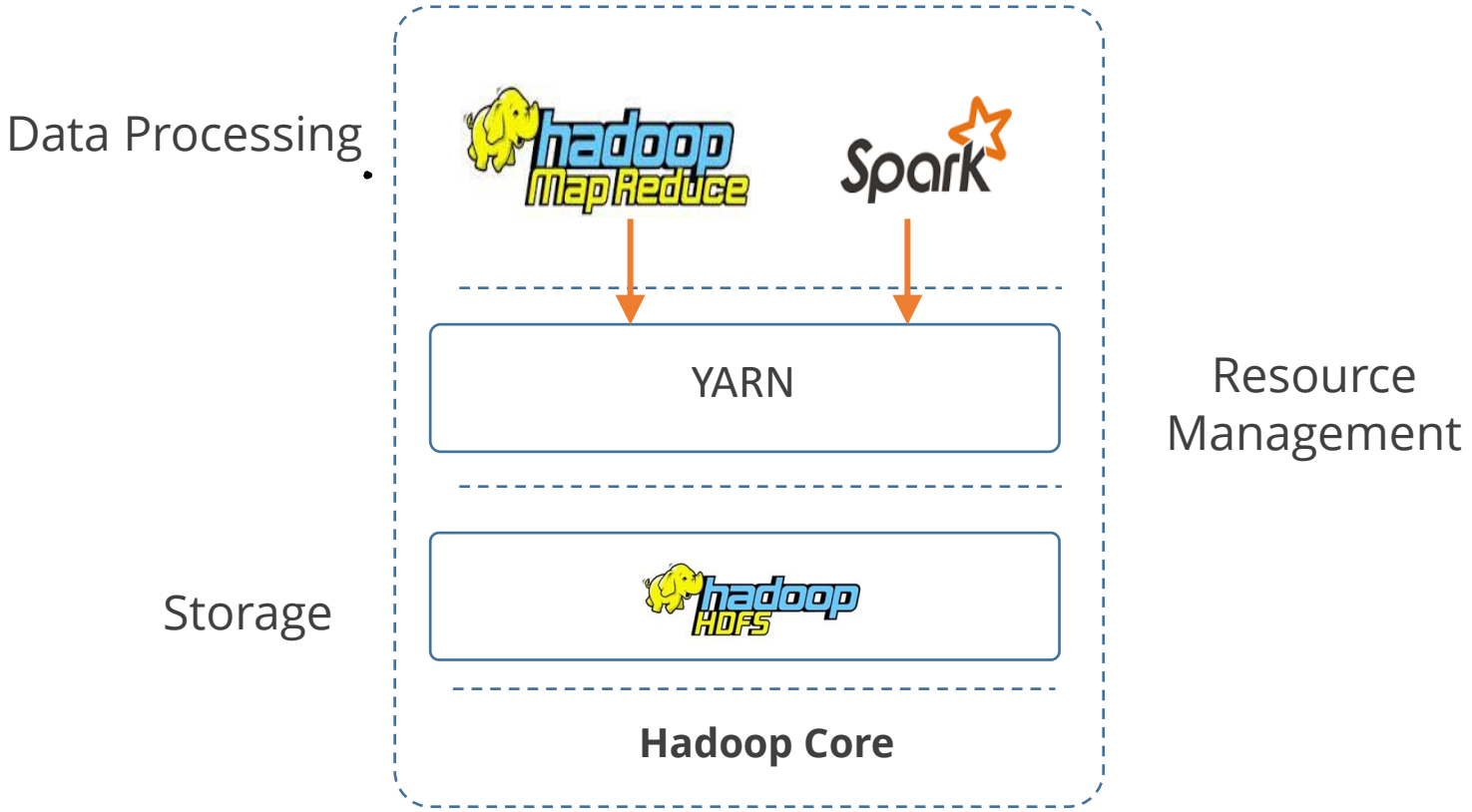
VS.

Tiger brings its mouth toward the food

Traditional Database Systems vs. Hadoop

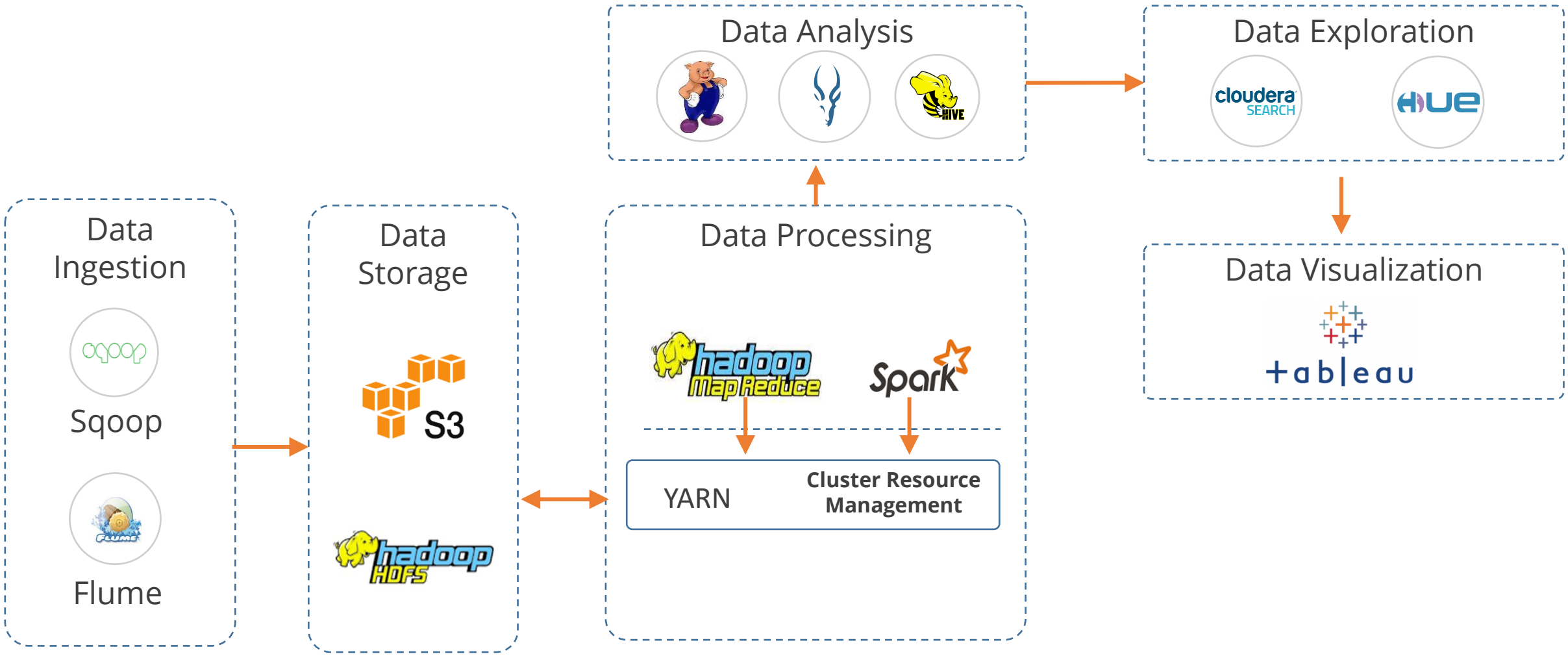
RDBMS		HADOOP
Structured	Data Types	Multi and Unstructured
Limited, No Data Processing	Processing	Processing coupled with Data
Standards and Structured	Governance	Loosely Structured
Required On Write	Schema	Required On Read
Reads are Fast	Speed	Writes are Fast
Software License	Cost	Support Only
Known Entity	Resources	Growing, Complexities, Wide
OLTP Complex ACID Transactions Operational Data Store	Best Fit Use	Data Discovery Processing Unstructured Data Massive Storage/Processing

Hadoop Core Components



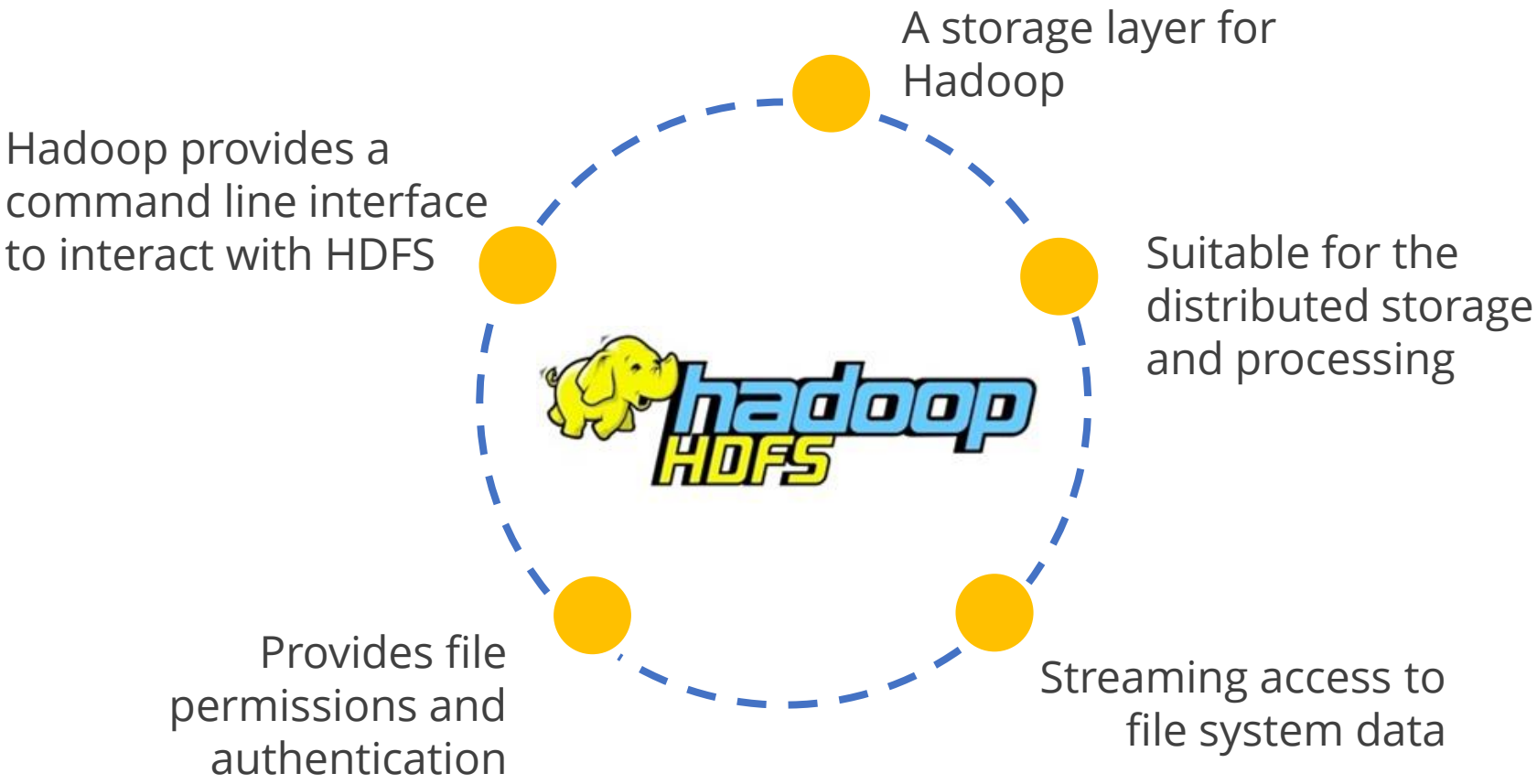
Components of Hadoop Ecosystem

Components of Hadoop Ecosystem



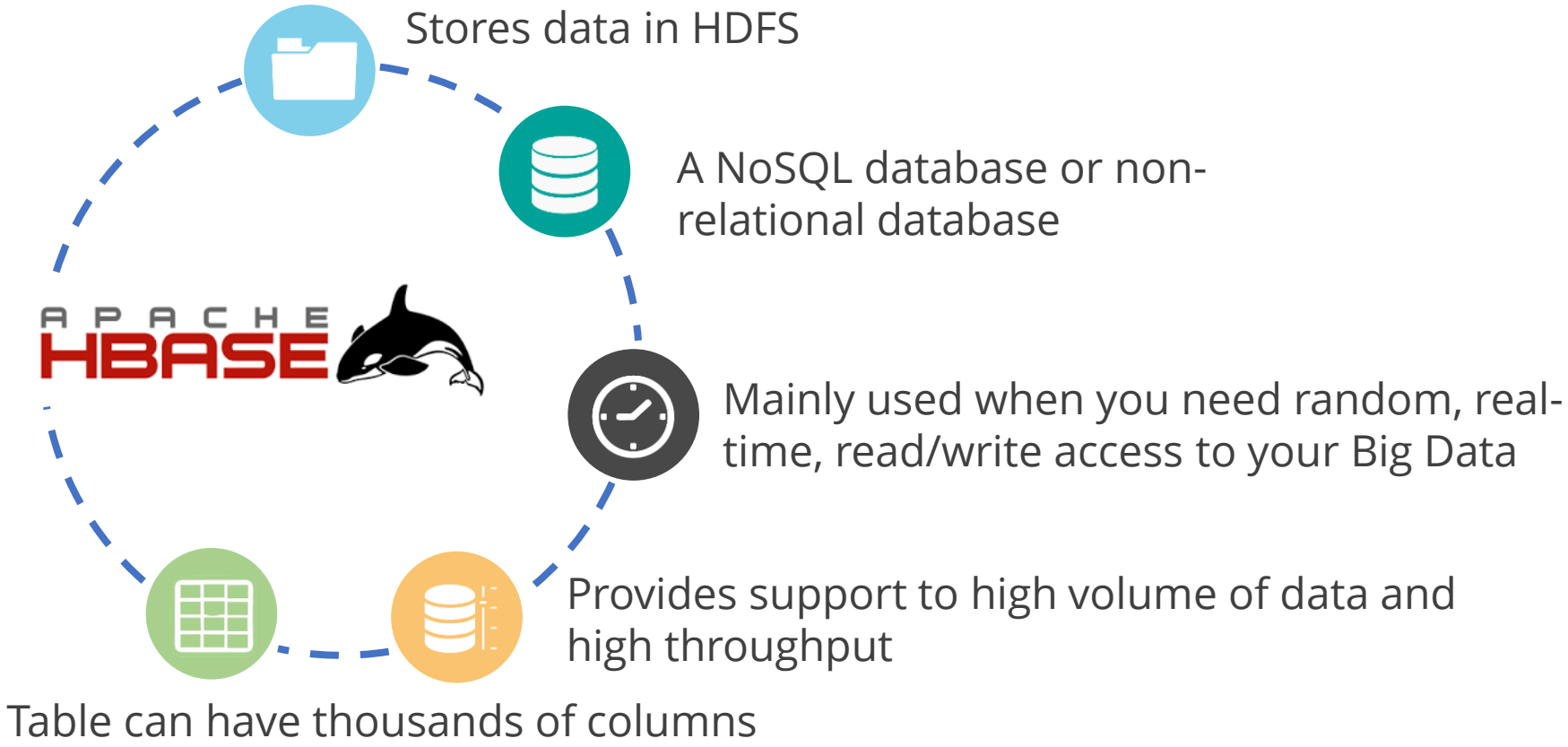
Components of Hadoop Ecosystem

HDFS (HADOOP DISTRIBUTED FILE SYSTEM)



Components of Hadoop Ecosystem

HBase



Components of Hadoop Ecosystem

SQOOP

- Sqoop is a tool designed to transfer data between Hadoop and relational database servers.

- It is used to import data from relational databases such as Oracle and MySQL to HDFS and export data from HDFS to relational databases.

Components of Hadoop Ecosystem

FLUME

If you want to ingest event data such as, streaming data, sensor data, or log files, then you can use Flume.

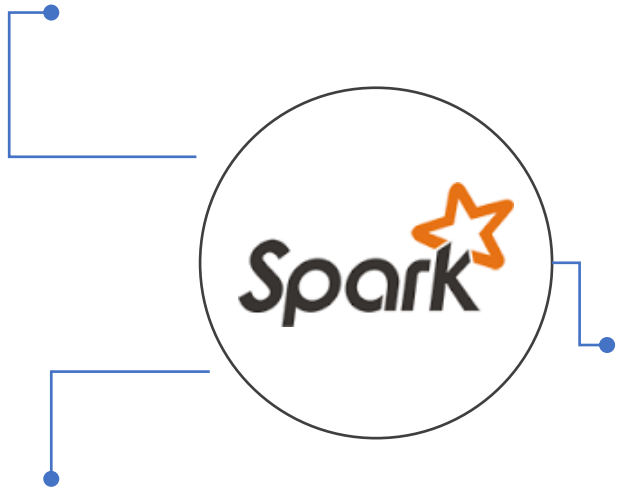
A distributed service for ingesting streaming data

Ideally suited for event data from multiple systems

Components of Hadoop Ecosystem

SPARK

An open source cluster computing framework

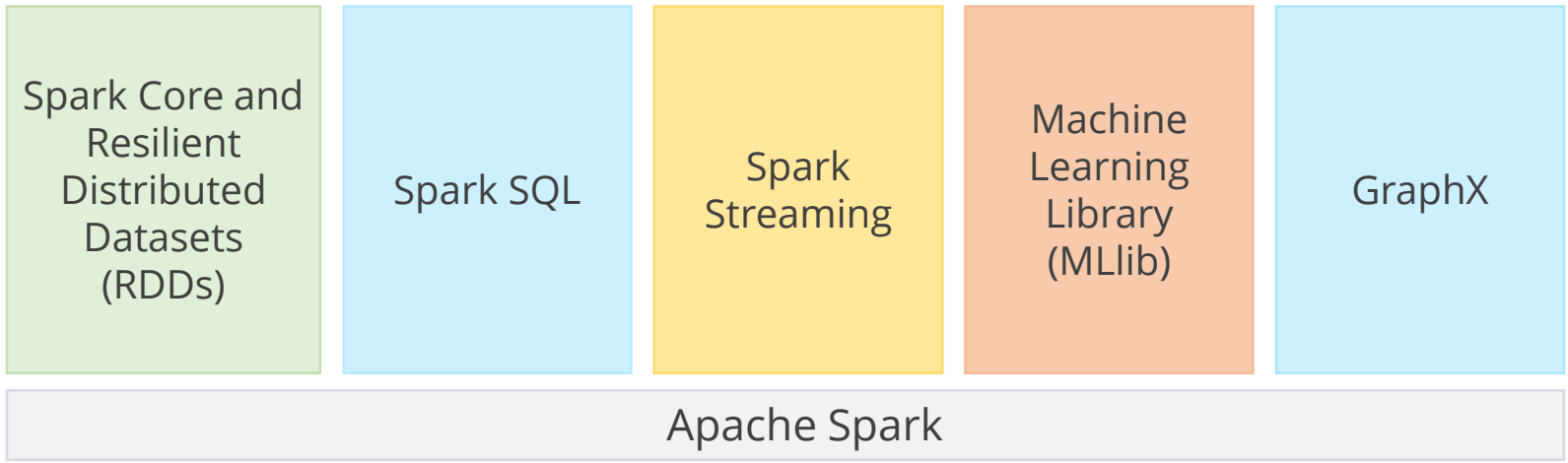


Provides 100 times faster performance than Map-Reduce

Supports machine learning, business intelligence, streaming, and batch processing

Components of Hadoop Ecosystem

SPARK: COMPONENTS

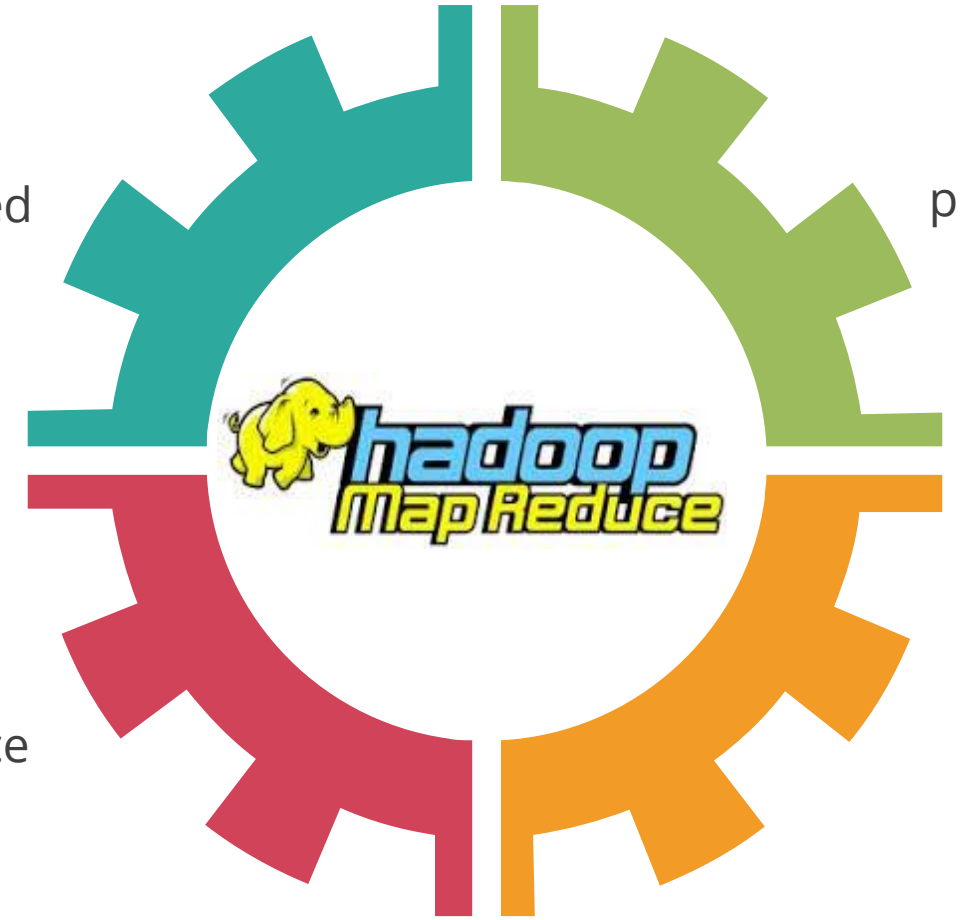


Components of Hadoop Ecosystem

HADOOP MAP-REDUCE

Commonly used

The original Hadoop processing engine which is primarily Java based



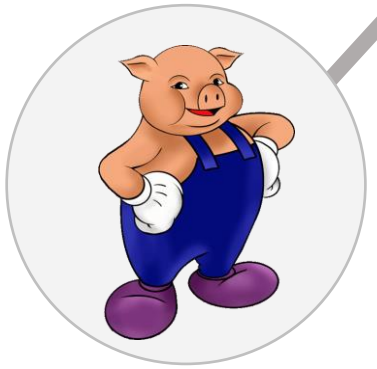
An extensive and mature fault tolerance framework

Based on the map and reduce programming model

Components of Hadoop Ecosystem

PIG

An open source dataflow system



Converts pig script to Map-Reduce code

An alternative to writing Map-Reduce code

Best for ad-hoc queries such as join and filter

Components of Hadoop Ecosystem

IMPALA

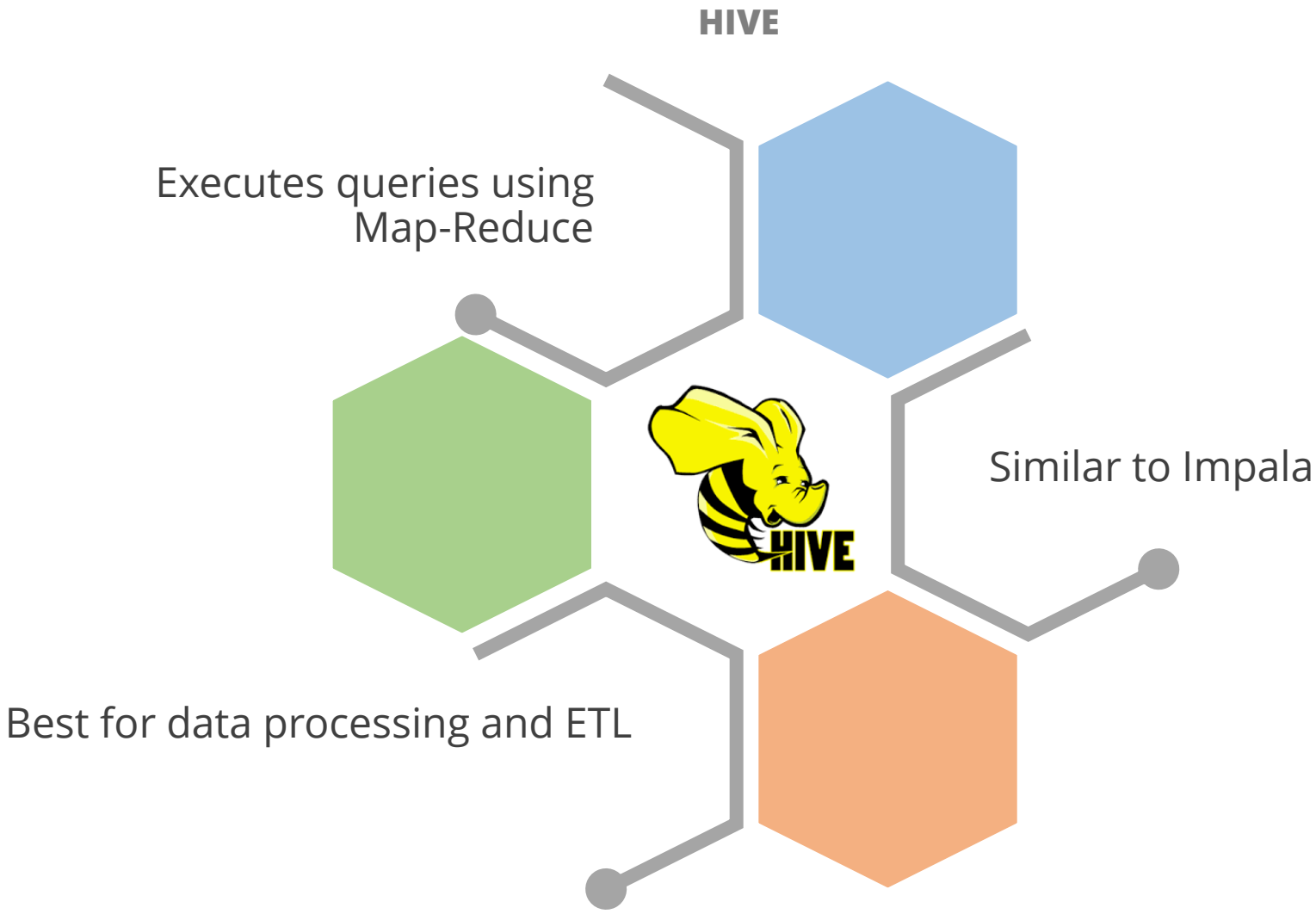
High performance SQL engine which runs on Hadoop cluster

Ideal for interactive analysis

Very low latency – measured in milliseconds

Supports a dialect of SQL (Impala SQL)

Components of Hadoop Ecosystem



Components of Hadoop Ecosystem

CLOUDERA SEARCH

One of Cloudera's near-real-time access products

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

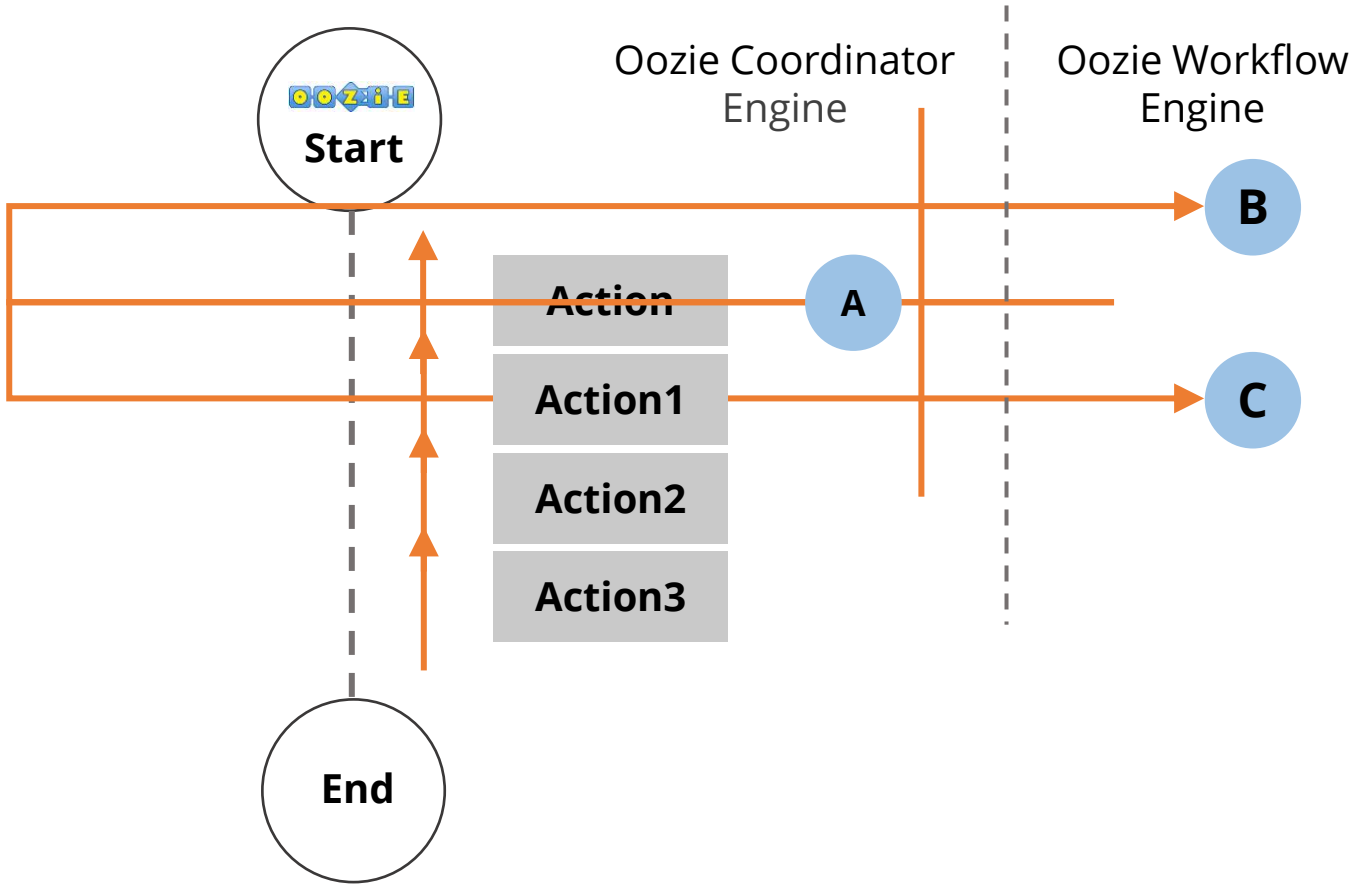
Enables nontechnical users to search and explore data stored in or ingested into Hadoop and HBase

A fully integrated data processing platform

Components of Hadoop Ecosystem

OOZIE

Oozie is a workflow or coordination system used to manage the Hadoop jobs



Components of Hadoop Ecosystem

HUE (HADOOP USER EXPERIENCE)

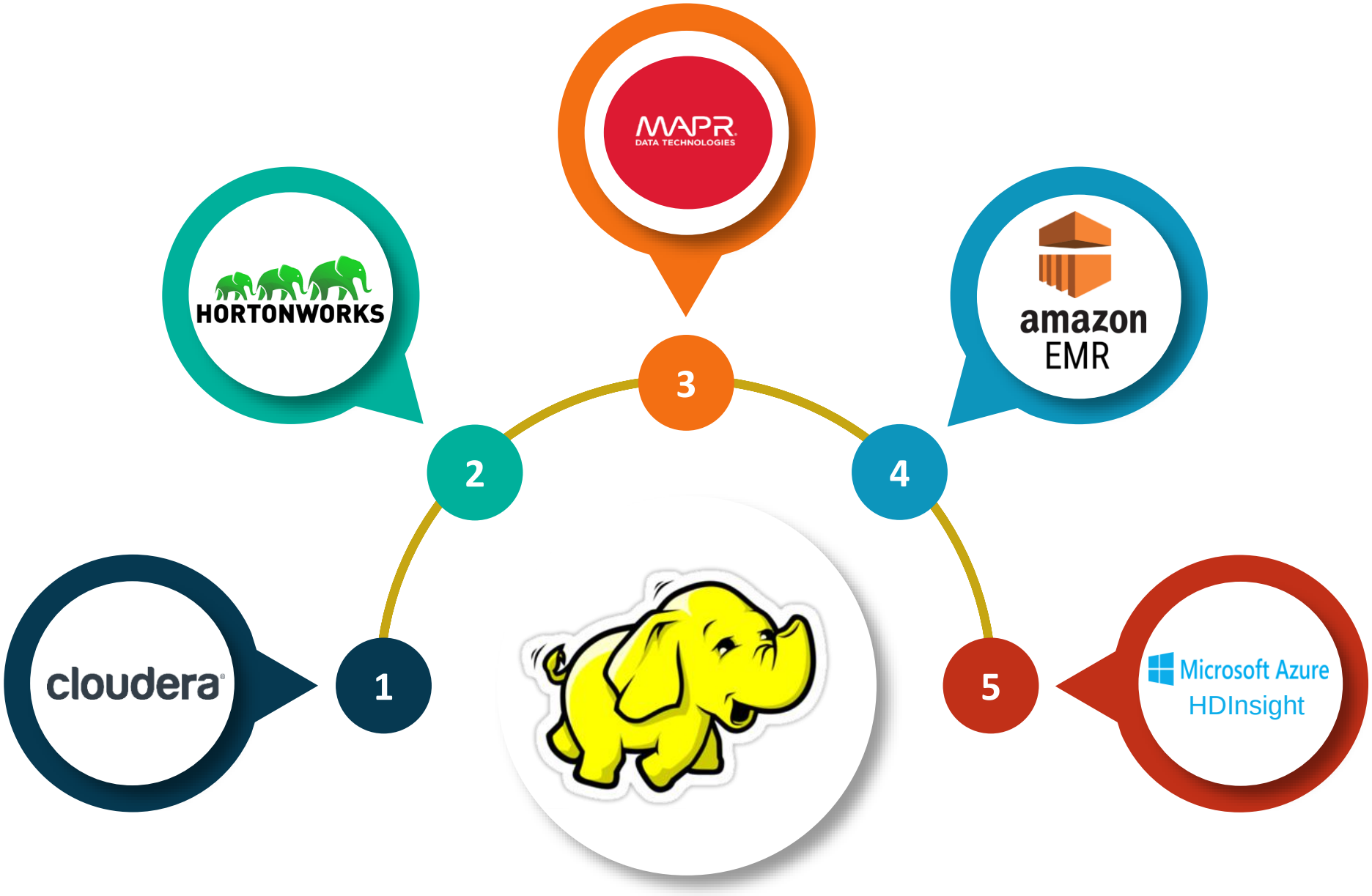
Hue is an acronym for Hadoop User Experience

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

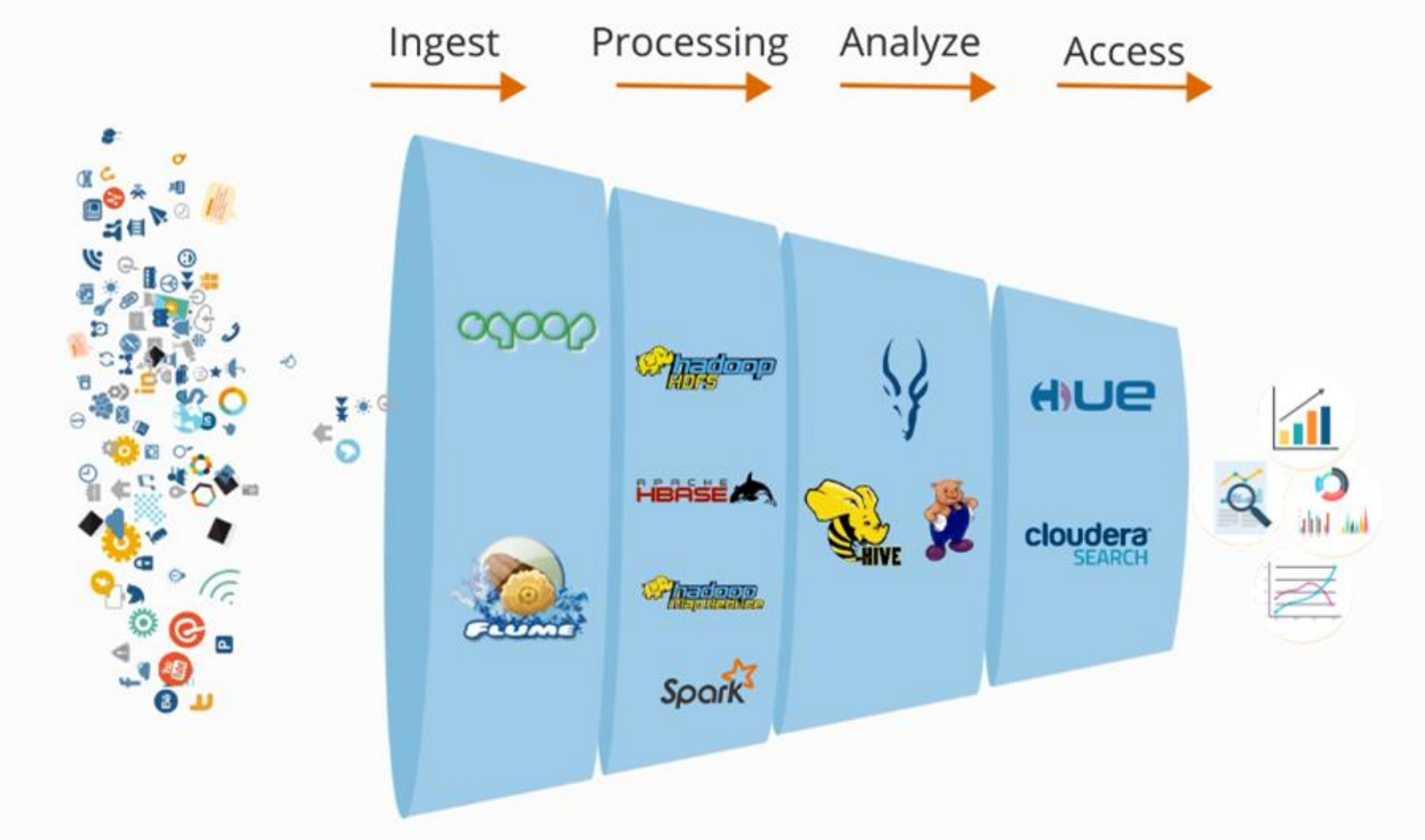
Commercial Hadoop Distributions

Various Commercial Hadoop Distributions



Big Data Processing

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



Big Data Terminology

BIG DATA

- Big data is an umbrella term for datasets that cannot reasonably be handled by traditional computers or tools due to their volume, velocity, and variety. This term is also typically applied to technologies and strategies to work with this type of data.

BATCH PROCESSING

- Batch processing is a computing strategy that involves processing data in large sets. This is typically ideal for non-time sensitive work that operates on very large sets of data. The process is started and at a later time, the results are returned by the system.

CLUSTER COMPUTING

- Clustered computing is practice of pooling resources of multiple machines and managing their collective capabilities to complete tasks.
- Computer clusters require a cluster management layer which handles communication between the individual nodes.

Big Data Terminology

STREAM PROCESSING

- Stream processing is the practice of computing over individual data items as they move through a system. This allows for real-time analysis of the data being fed to the system and is useful for time-sensitive operations using high velocity metrics.

ETL

- ETL stands for extract, transform, and load. It refers to the process of taking raw data and preparing it for the system's use. This is traditionally a process associated with data warehouses, but characteristics of this process are also found in the ingestion pipelines of big data systems.

HADOOP

- Hadoop: Hadoop is an Apache project that was the early open-source success in big data. It consists of a distributed filesystem called HDFS, with a cluster management and resource scheduler on top called YARN (Yet Another Resource Negotiator).

Key Takeaways

You are now able to:

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

Knowledge Check

1. **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

2. A bank wants to process 1000 transactions per second. Which one of the following Vs reflects this real-world use case?

a. Volume

b. Variety

c. Velocity

d. Veracity

3. Why has popularity of big data increased tremendously in the recent years?

a. Due to increased volume of data

b. Big data is an open source

c. Abundance of unstructured data

d. None of the above

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.

5. Which of the following is a column-oriented NoSQL database that runs on top of HDFS?

a. MongoDB

b. Flume

c. Ambari

d. HBase

6. Scoop is used to _____.

a. Import data from relational databases to Hadoop HDFS and export from Hadoop file system to relational databases

b. Execute queries using Map-Reduce

c. Enable nontechnical users to search and explore data stored in or ingested into Hadoop and HBase

d. Stream event data from multiple systems

Happy Learning

Thank You