Big Data Hadoop and Spark Developer

Introduction to Big Data and Hadoop




Learning Objectives

- Discuss the basics of big data with a case study
- Explain the basics of Hadoop
- Describe the components of the Hadoop Ecosystem

Introduction to Big Data and Hadoop

Topic 1—Introduction to Big Data

Data Is Exploding

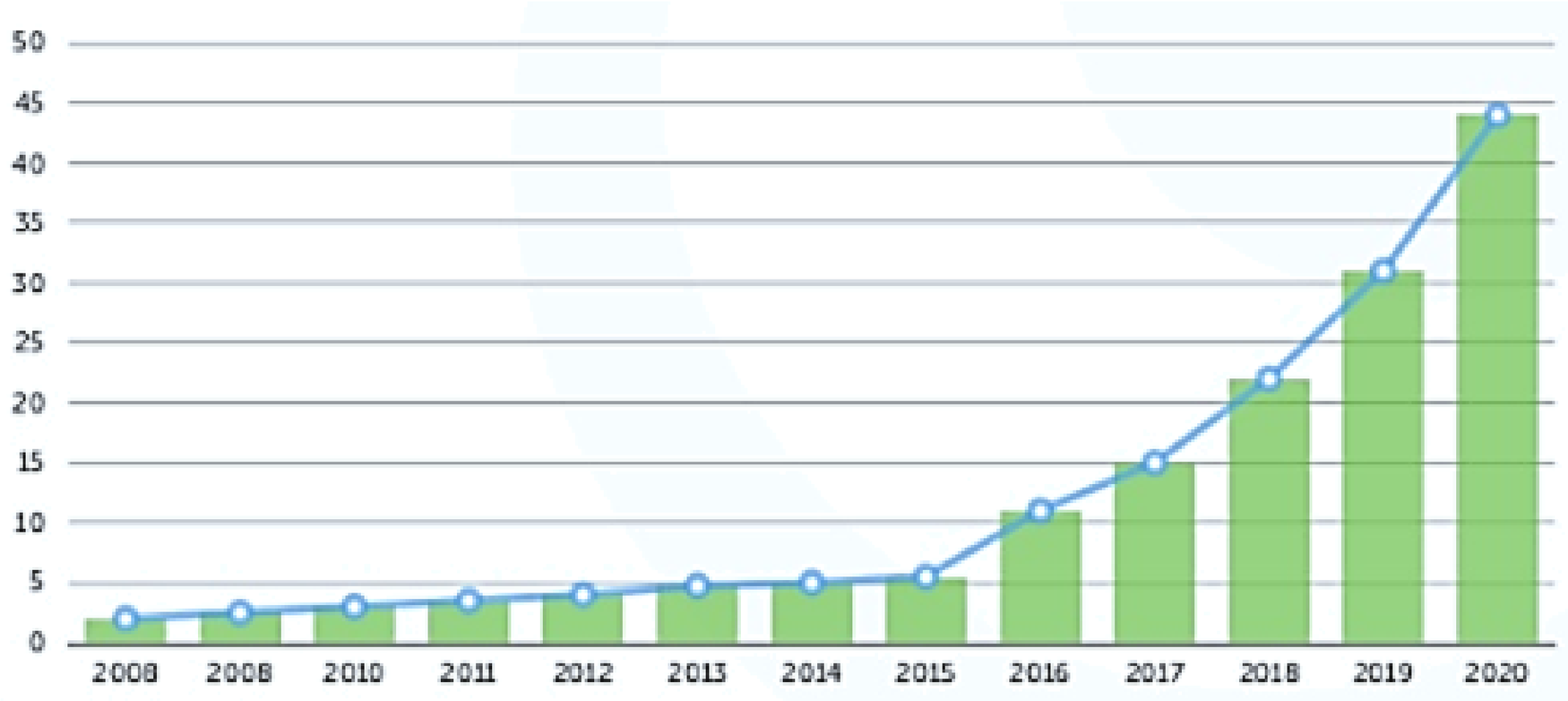
IBM reported that 2.5 billion gigabytes of data was generated every day in 2012.
It is predicted that by 2020:

- About 1.7 megabytes of new information will be generated for every human, every second
- 40,000 search queries will be performed on Google every second
- 300 hours of video will be uploaded to YouTube every minute
- 31.25 million messages will be sent and 2.77 million videos viewed by Facebook users
- 80% of photos will be taken on smartphones
- At least a third of all data will pass through Cloud

Data Is Exploding(Contd.)

By 2020, data will show an exponential rise!

Data in Zettabytes (ZB)



What Is Big Data?

Big data refers to the large volume of structured and unstructured data. The analysis of big data leads to better insights for business.

Big Data: Case Study

NETFLIX

Netflix is one of the largest providers of commercial streaming video in the US with a customer base of over 29 million.

It receives a huge volume of behavioral data.

- When do users watch a show?
- Where do they watch it?
- On which device do they watch the show?
- How often do they pause a program?
- How often do they re-watch a program?
- Do they skip the credits?
- What are the keywords searched?

Big Data: Case Study

NETFLIX

Traditionally, the analysis of such data was done using a computer algorithm that was designed to produce a correct solution for any given instance.

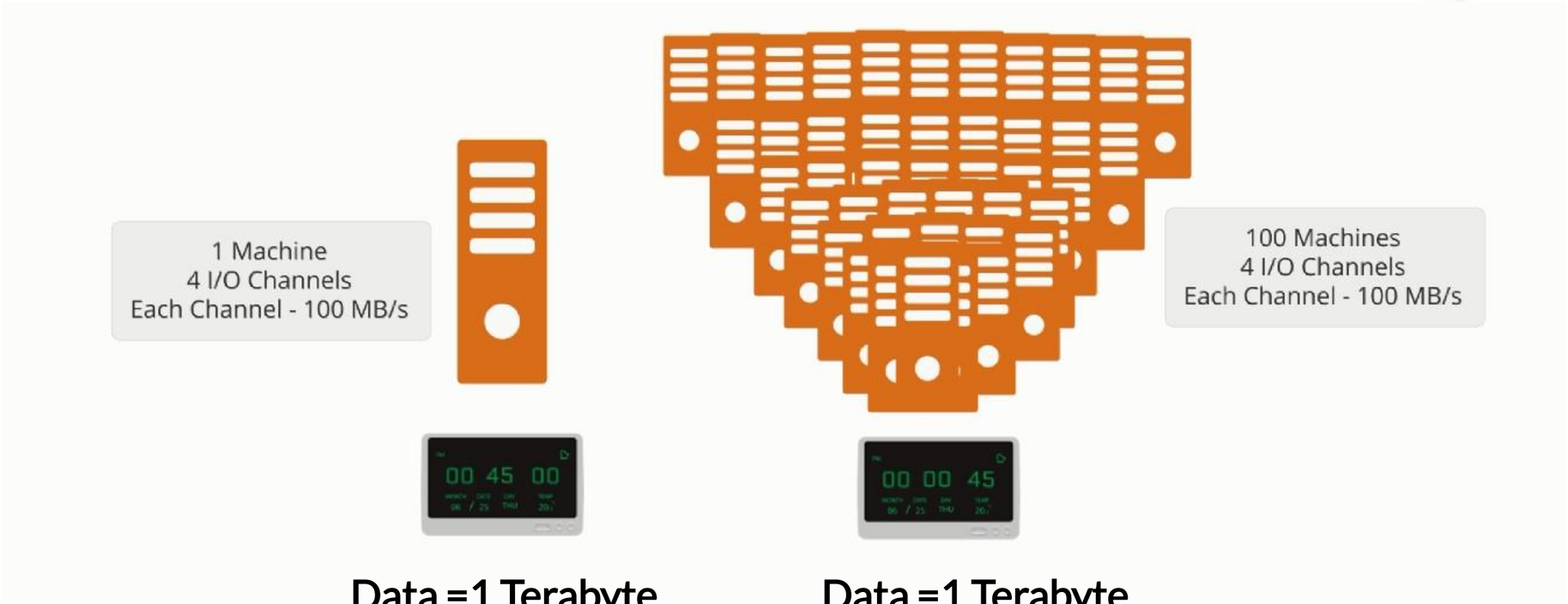
As the data started to grow, a series of computers were employed to do the analysis. They were also known as distributed systems.

Distributed Systems

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

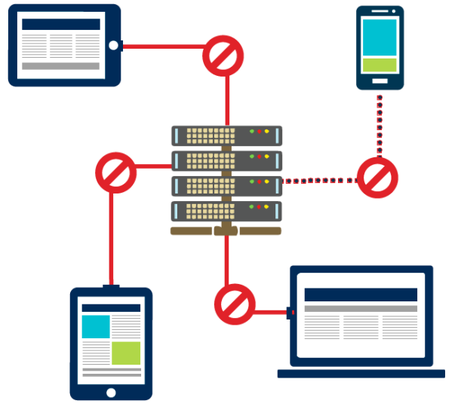
https://en.wikipedia.org/wiki/Distributed_computing

How Does a Distributed System Work?



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

Challenges of Distributed Systems



1. High chances of system failure
2. Limited bandwidth
3. High programming complexity

HADOOP is used to overcome these challenges!

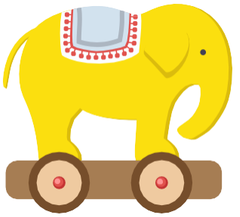
Introduction to Big Data and Hadoop

Topic 2—Introduction to Hadoop

What Is Hadoop?

> Hadoop is a framework that allows distributed processing of large datasets across clusters of 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.

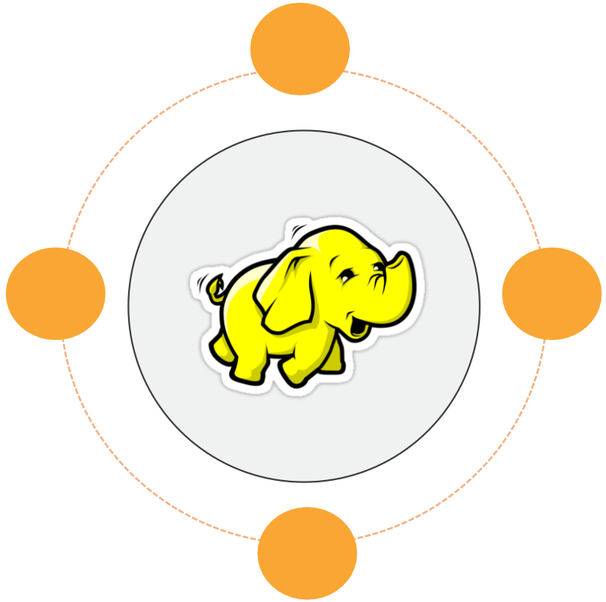


https://twitter.com/cutting

Characteristics of Hadoop

Scalable: Can follow both horizontal and vertical scaling

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



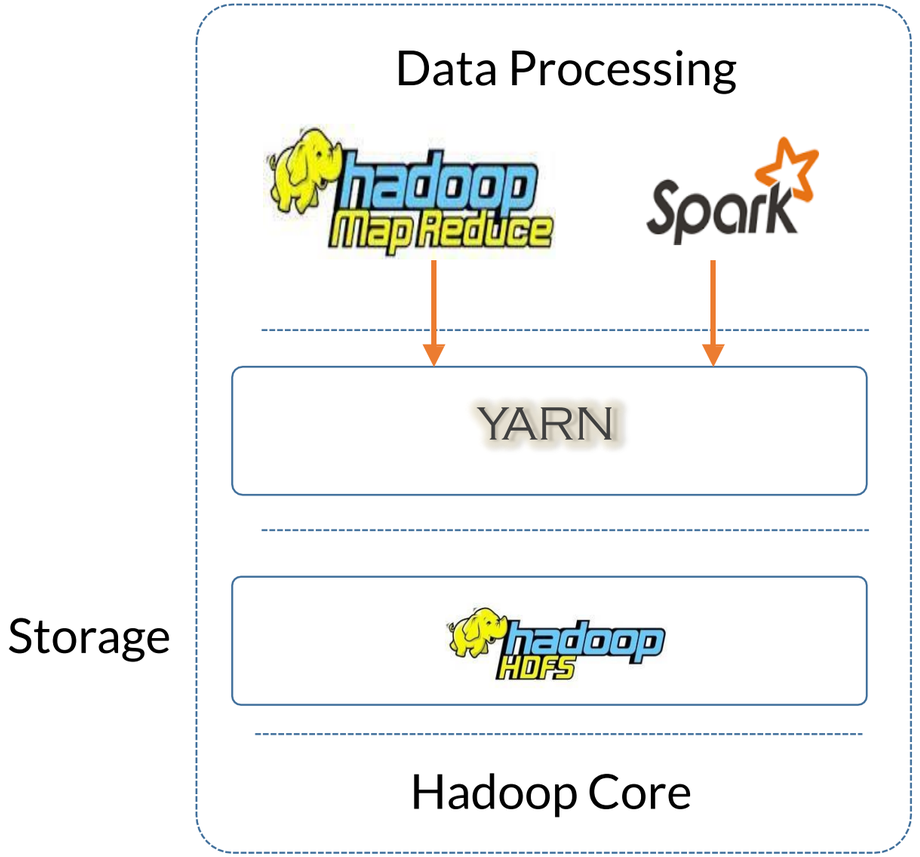
Flexible: Stores a lot of data and enables you to use it later

Economical: Ordinary computers can be used for data processing

Traditional Database Systems vs. Hadoop

Traditional Database Systems	Hadoop
Data is stored in a central location and sent to the processor at run time.	In Hadoop, the program goes to the data. It initially distributes the data to multiple systems and later runs the computation wherever the data is located.
Traditional Database Systems cannot be used to process and store a large amount of data (big data).	Hadoop works better when the data size is big. It can process and store a large amount of data easily and effectively.
Traditional RDBMS is used to manage only structured and semi-structured data. It cannot be used to manage unstructured data.	Hadoop has the ability to process and store a variety of data, whether it is structured or unstructured.

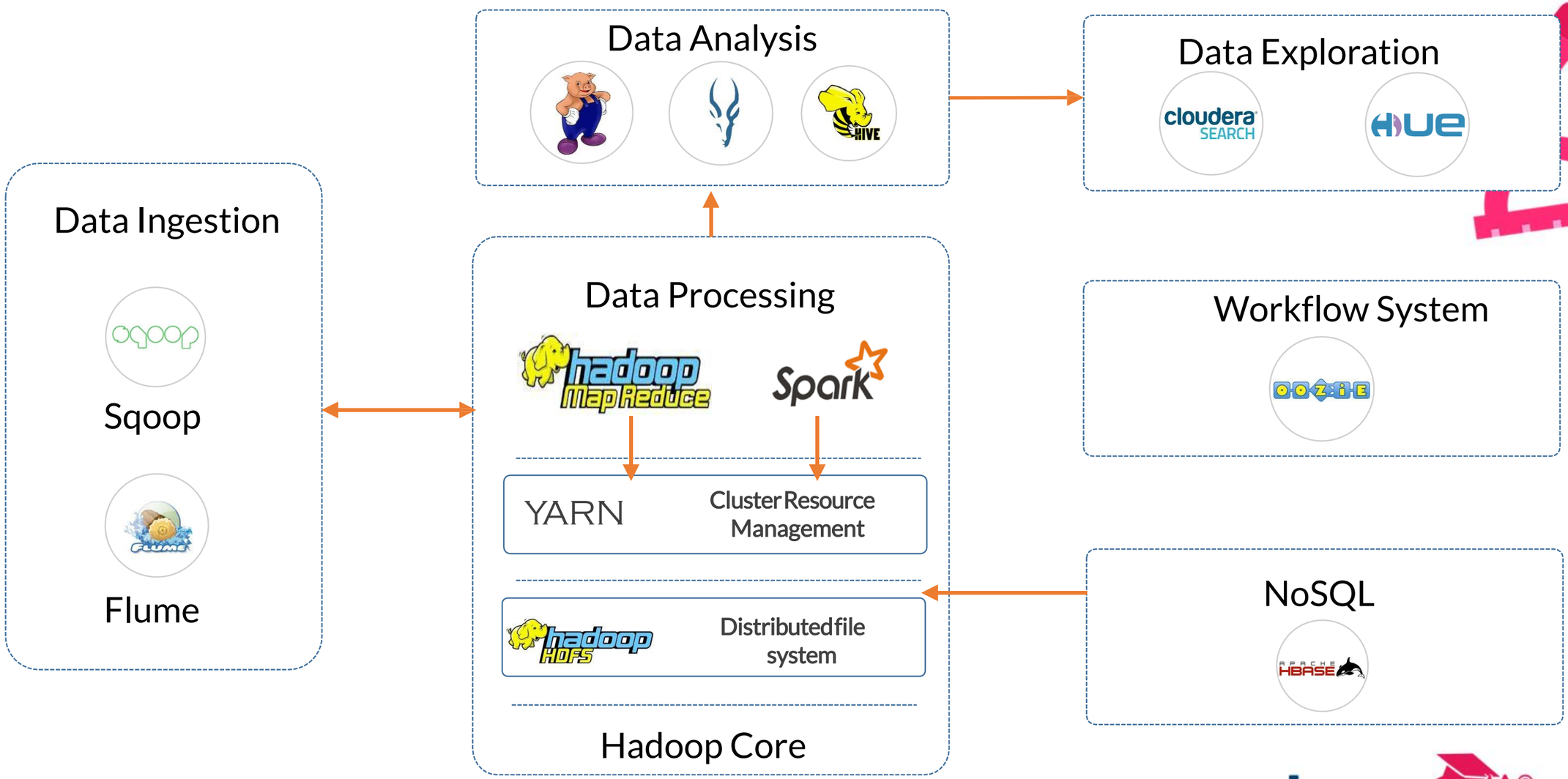
Hadoop Core Components



Introduction to Big Data and Hadoop

Topic 3—Components of Hadoop Ecosystem

Components of Hadoop Ecosystem

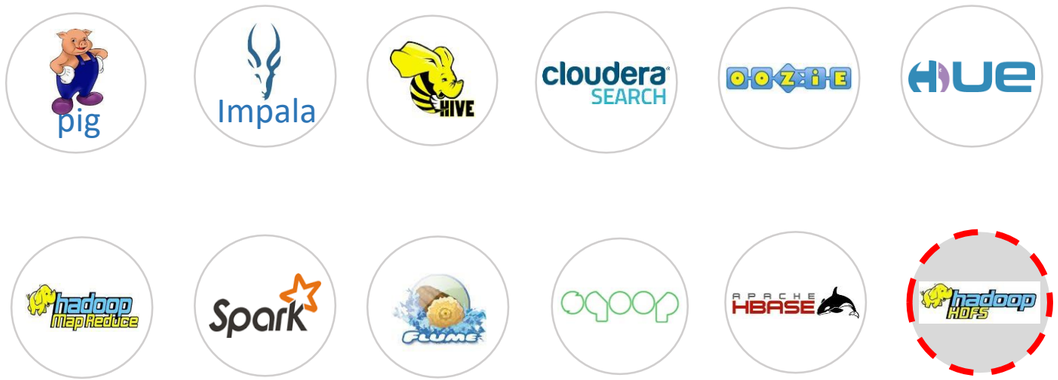


Components of Hadoop Ecosystem

HDFS (HADOOP DISTRIBUTED FILE SYSTEM)

- HDFS is a storage layer of Hadoop suitable for distributed storage and processing.

- It provides file permissions, authentication, and streaming access to file system data.

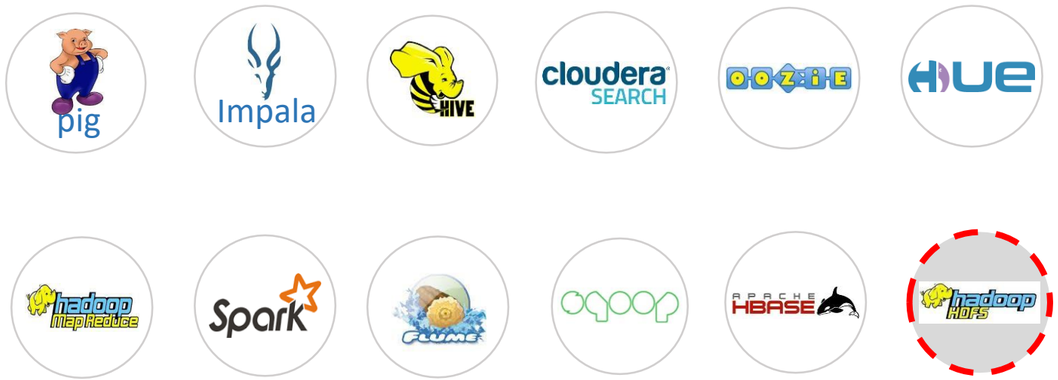


HDFS can be accessed through Hadoop command line interface.

Components of Hadoop Ecosystem

HBase

- HBase is a NoSQL database or non-relational database that stores data in HDFS.
- It provides support to high volume of data and high throughput.
- It is used when you need random, real-time read/write access to your big data.

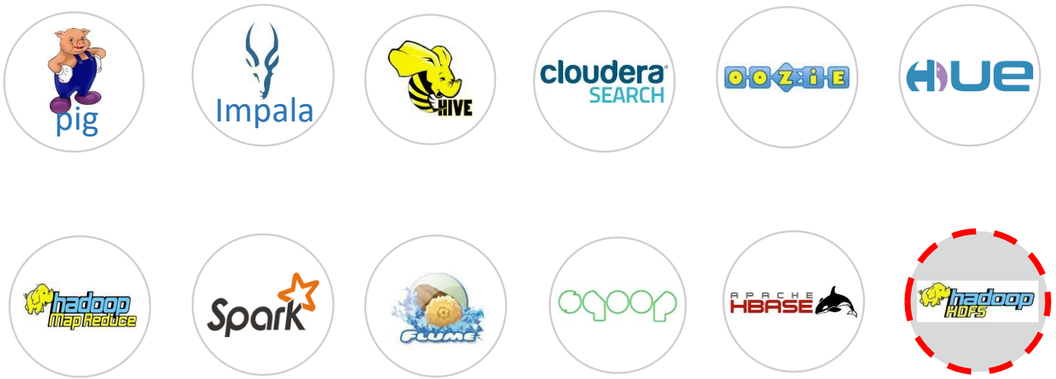


HBase tables can have thousands of columns.

Components of Hadoop Ecosystem

SQOOP

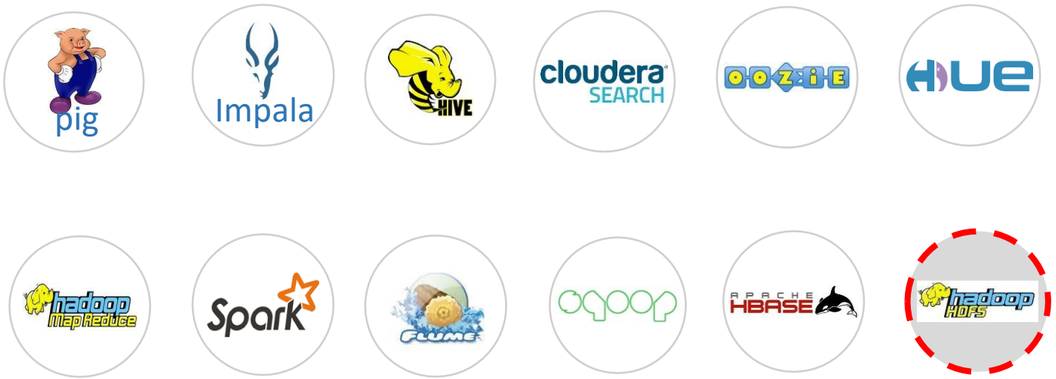
- Sqoop is a tool designed to transfer data between Hadoop and relational database servers.
- It is used to import data from relational databases such as Oracle and MySQL to HDFS
- and export data from HDFS to relational databases.



Components of Hadoop Ecosystem

FLUME

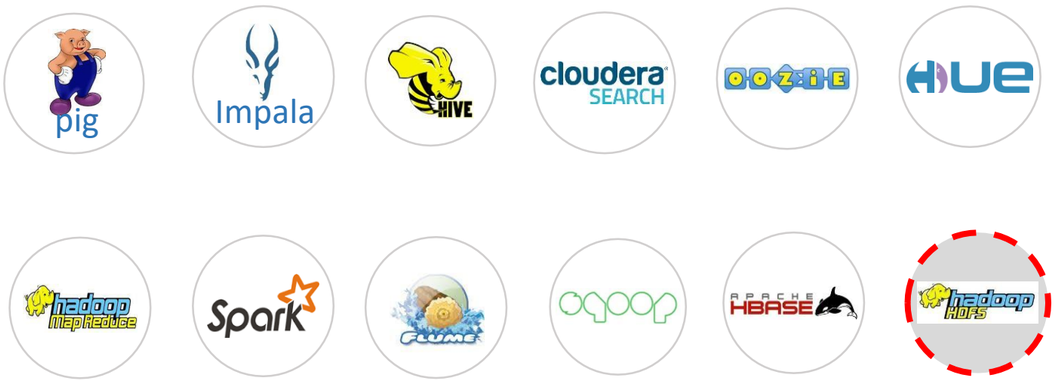
- Flume is a distributed service for ingesting streaming data suited for event data from multiple systems.
- It has a simple and flexible architecture based on streaming data flows.
- It is robust and fault tolerant and has tunable reliability mechanisms.
- It uses a simple extensible data model that allows for online analytic application.



Components of Hadoop Ecosystem

SPARK

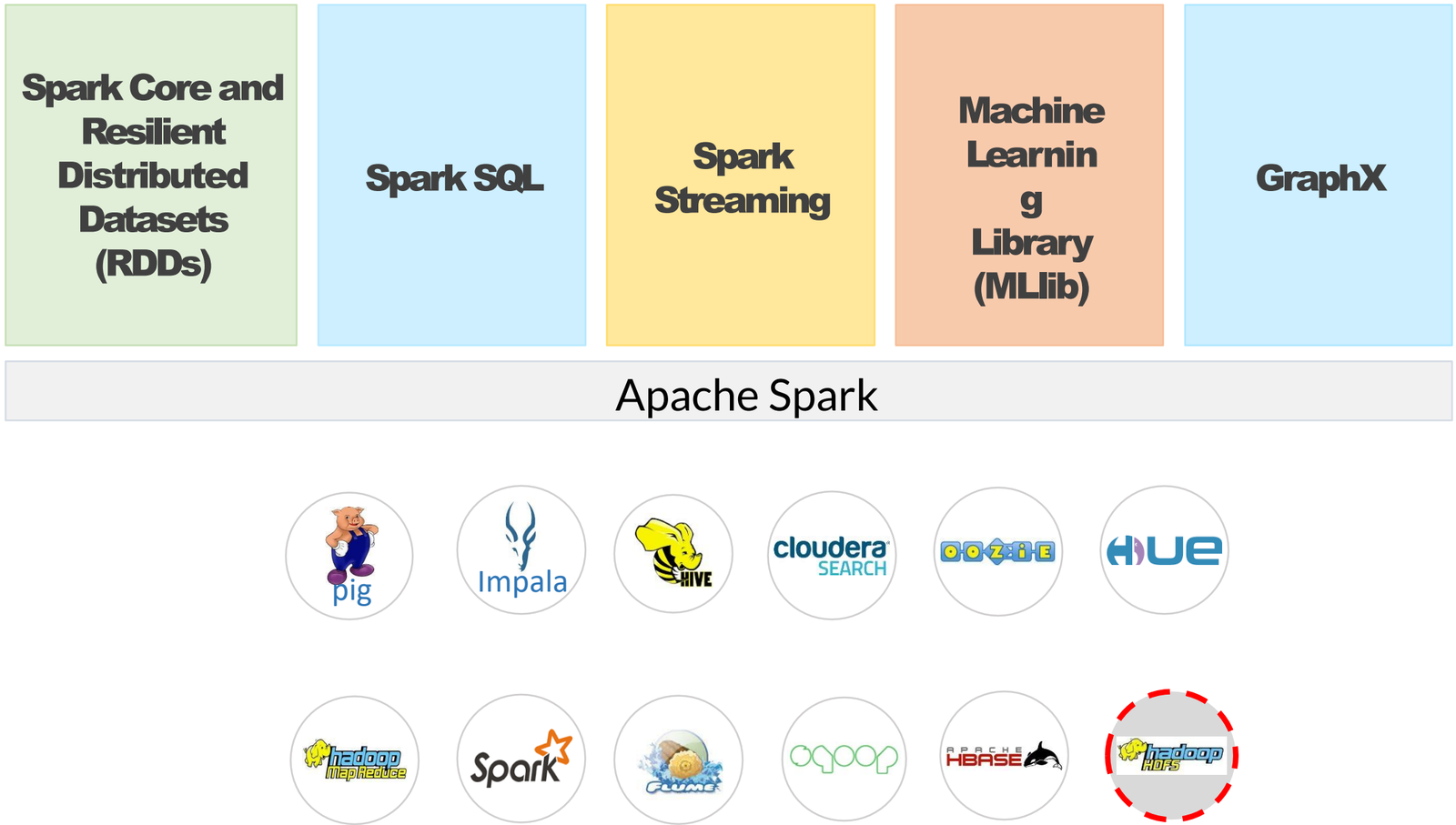
- Spark is an open-source cluster computing framework that supports Machine learning,
- Business intelligence, Streaming, and Batch processing.
- Spark solves similar problems as Hadoop MapReduce does but has a fast in-memory approach and a clean functional style API.



Spark and MapReduce will be discussed in the upcoming lessons.

Components of Hadoop Ecosystem

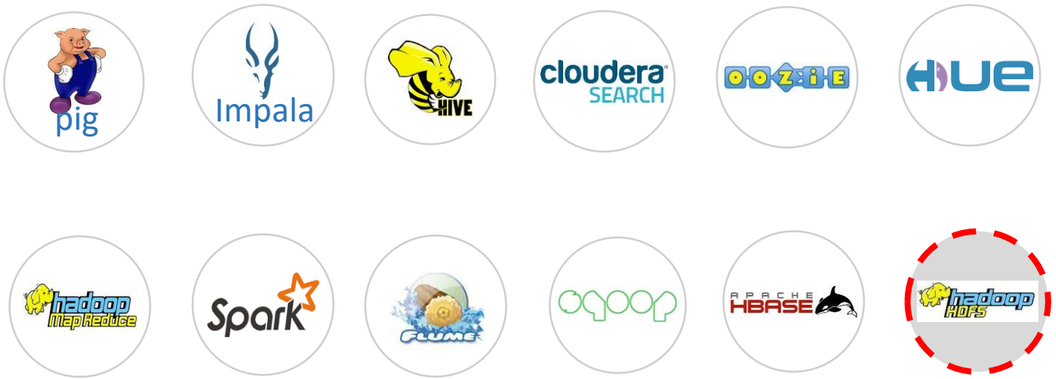
SPARK: COMPONENTS



Components of Hadoop Ecosystem

HADOOP MAPREDUCE

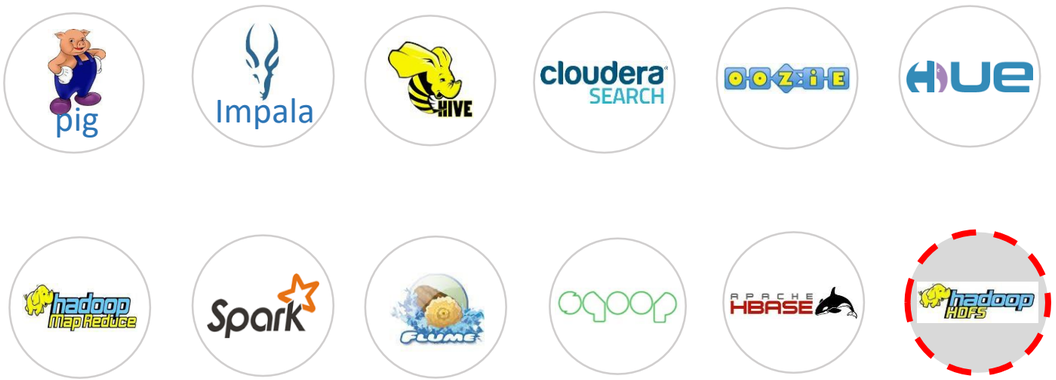
- Hadoop MapReduce is a framework that processes data. It is the original Hadoop processing engine, which is primarily Java-based.
- It is based on the map and reduce programming model.
- It has an extensive and mature fault tolerance.
- Hive and Pig are built on map-reduce model.



Components of Hadoop Ecosystem

PIG

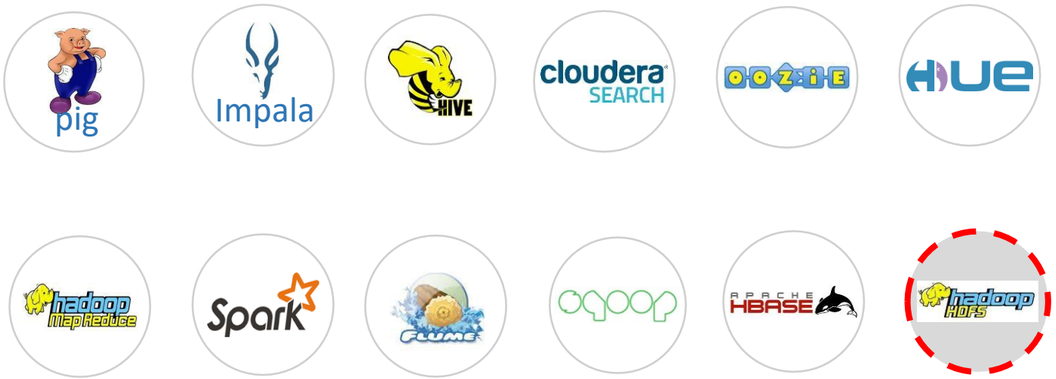
- Once the data is processed, it is analyzed using an open-source high-level dataflow
- system called Pig.
- Pig converts its scripts to Map and Reduce code to reduce the effort of writing complex map-reduce programs.
- Ad-hoc queries like Filter and Join, which are difficult to perform in MapReduce, can be
- easily done using Pig.



Components of Hadoop Ecosystem

IMPALA

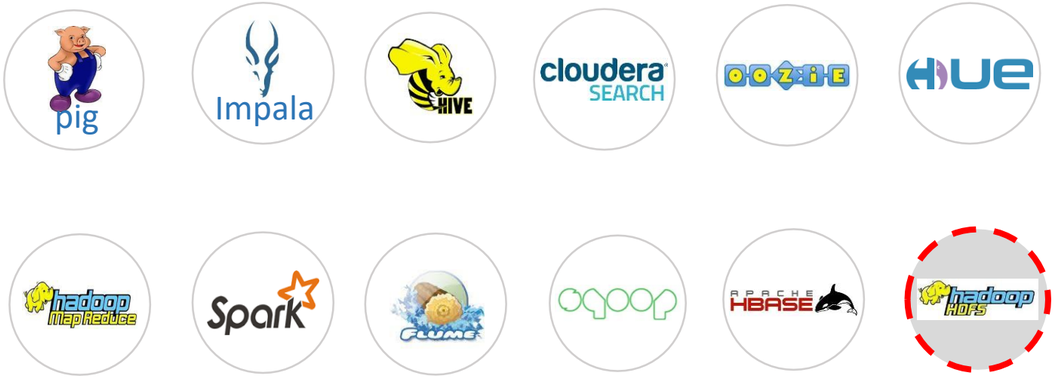
- It is an open-source high performance SQL engine that runs on the Hadoop cluster.
- It is ideal for interactive analysis and has very low latency, which can be measured in milliseconds.
- Impala supports a dialect of SQL, so data in HDFS is modeled as a database table.



Components of Hadoop Ecosystem

HIVE

- Hive is an abstraction layer on top of Hadoop that executes queries using MapReduce.
- It is preferred for data processing and ETL (Extract Transform Load) and ad hoc queries.

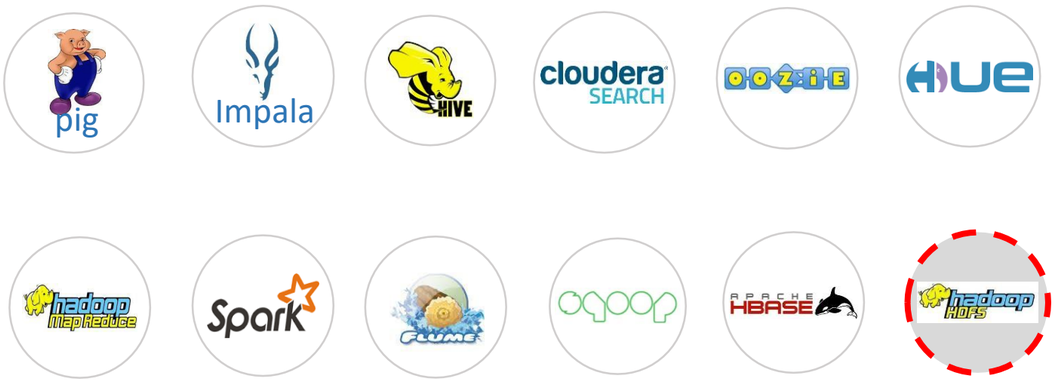


Components of Hadoop Ecosystem

CLOUDERA SEARCH

- It is Cloudera's near-real-time access product that enables non-technical users to search and explore data stored in or ingested into Hadoop and HBase.

- Cloudera Search is a fully integrated data processing platform. It uses the flexible, scalable, and robust storage system included with CDH or Cloudera's Distribution, including Hadoop.

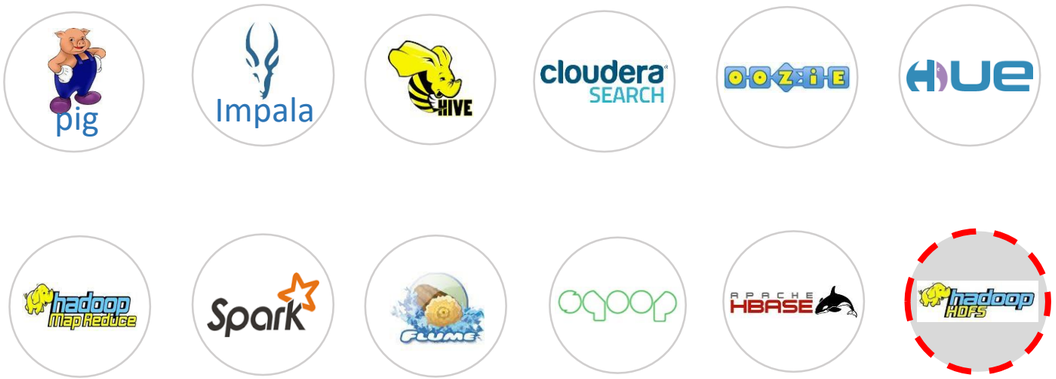


Components of Hadoop Ecosystem

OOZIE

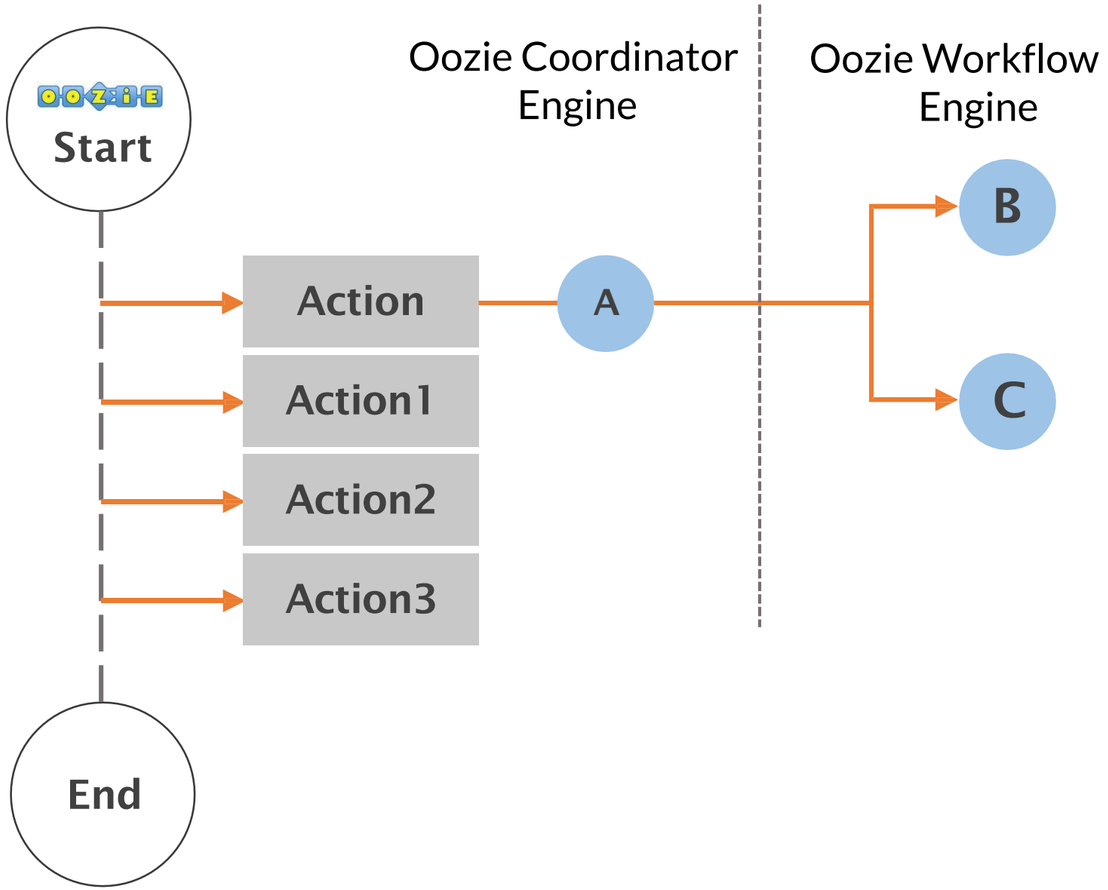
- Oozie is a workflow or coordination system used to manage the Hadoop tasks.

- Oozie coordinator can trigger jobs by time (frequency) and data availability.



Components of Hadoop Ecosystem

OOZIE APPLICATION LIFECYCLE

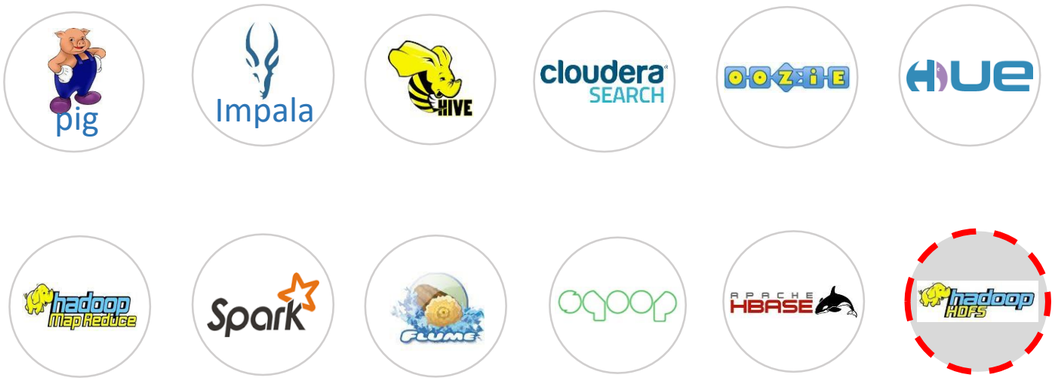


Components of Hadoop Ecosystem

HUE (HADOOP USER EXPERIENCE)

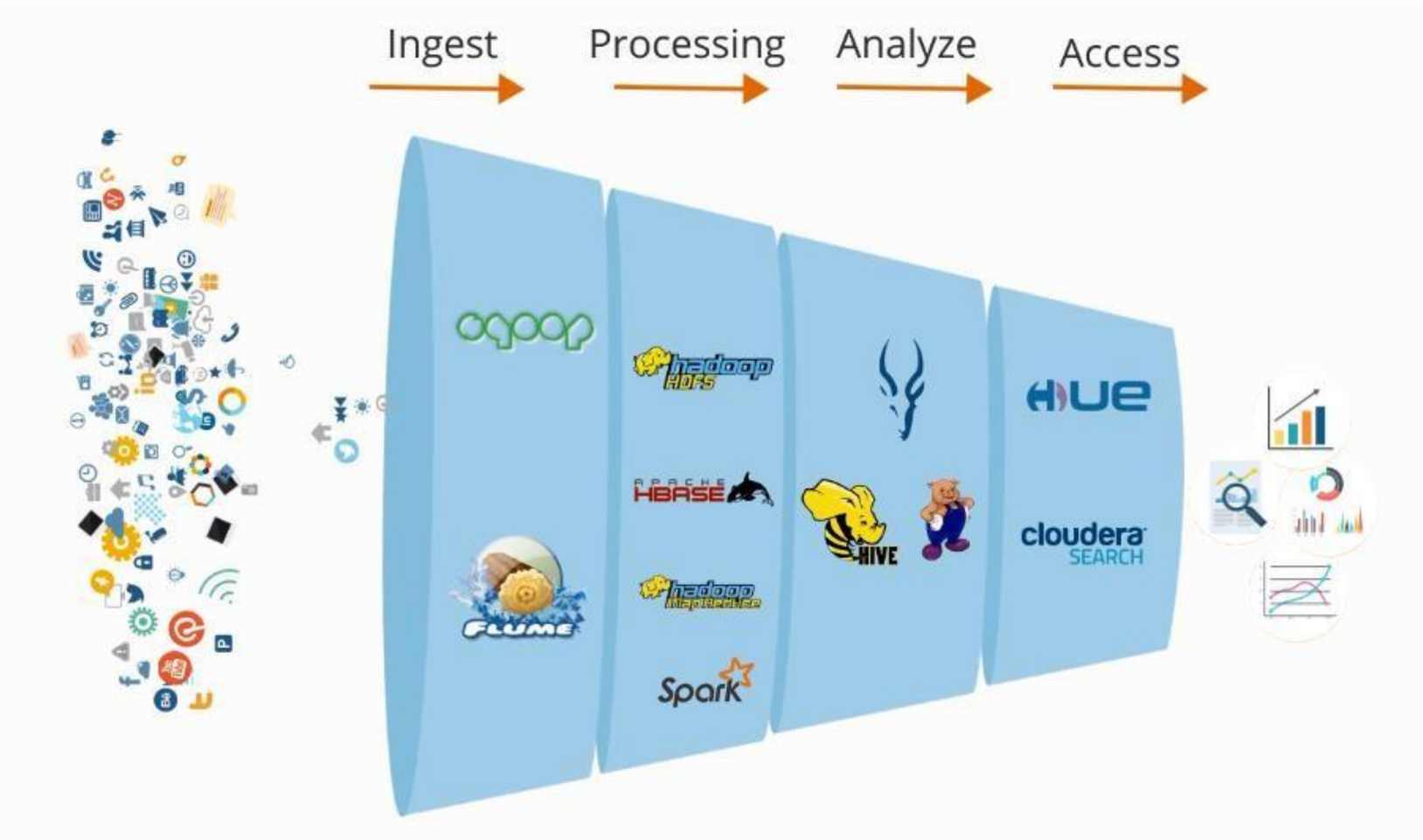
- Hue is an acronym for Hadoop User Experience. It is an open source Web interface for analyzing data with Hadoop.

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



Big Data Processing

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



Key Takeaways

Hadoop is a framework for distributed storage and processing.

Core components of Hadoop include HDFS for storage, YARN for cluster-resource management, and MapReduce or Spark for processing.

The Hadoop ecosystem includes multiple components that support each stage of big data processing:

- Flume and Scoop ingest data
- HDFS and HBase store data
- Spark and MapReduce process data
- Pig, Hive, and Impala analyze data
- Hue and Search help to explore data
- Oozie manages the workflow of Hadoop tasks

Quiz

Quiz 1

What is a Distributed system?

a. One machine processing a file

b. Multiple machines processing a file

c. A Traditional system

d. In-memory computation

Quiz 1

What is a Distributed system?

a. One machine processing a file

b. Multiple machines processing a file

c. A Traditional system

d. In-memory computation

The correct answer is b.

In distributed systems, you use multiple machines to process one file.

Quiz 2

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 for 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.

Quiz 2

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 for 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 for distributed processing of large datasets across clusters of commodity computers using a simple programming model.

Quiz 3

Which of the following is NOT a key characteristic of Hadoop?

a. Economical

b. Adaptable

c. Flexible

d. Reliable

Quiz 3

Which of the following is NOT a key characteristic of Hadoop?

a. Economical

b. Adaptable

c. Flexible

d. Reliable

The correct answer is **b.**

The four key characteristics of Hadoop are that it is economical, reliable, scalable, and flexible.

Quiz 4

Which of the following is used in the data storage processing stage?

a. Impala

b. Spark

c. Hive

d. HDFS/HBase

Quiz 4

Which of the following is used in the data storage processing stage?

a. Impala

b. Spark

c. Hive

d. HDFS/HBase

The correct answer is d.

HBase/HDFS is used in the data storage processing stage.

Quiz 5

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 MapReduce

c. Enable non-technical users to search and explore data stored in or ingested into Hadoop and Hbase

d. Stream event data from multiple systems

Quiz 5

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 MapReduce

c. Enable non-technical users to search and explore data stored in or ingested into Hadoop and Hbase

d. Stream event data from multiple systems

The correct answer is **a.**

Scoop is used to import data from relational databases to Hadoop HDFS and export from Hadoop file system to relational databases.

This concludes
"Introduction to Big Data and Hadoop."

Thank You!