BIG DATA
by CloudAndDataUniverse

1. BIG DATA

INTRODUCTION

Follow us on :

YouTube

LinkedIn

CDU
Cloud And Data Universe

Big data

- "Data is the new oil" ?
- Organizations and individuals who realized the value of data, have been investing in learning and adopting best practices in data industry.
- This approach has enabled them to get deeper insights and take great decisions in business which has benefitted them immensely.

Big Data Definition

Volume

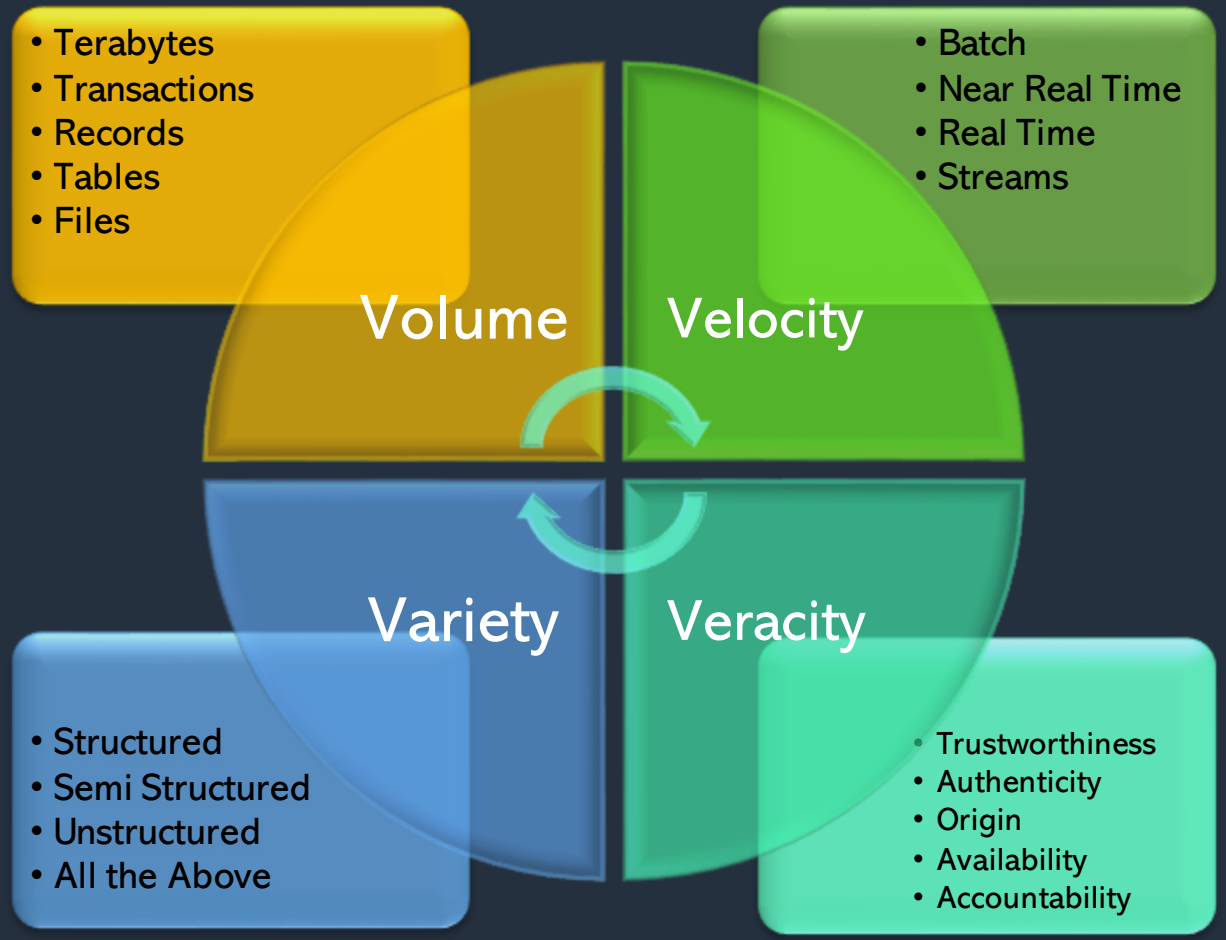
Velocity

Variety

Veracity

4 v's of Big Data

BIG DATA
by CloudAndDataUniverse



4 v's of Big Data

- Volume:- Mostly dealing with TB's of data.
- Velocity:- Speed at which data is being received/processed.
- Variety:- Structured/unstructured, images, videos.
- Veracity:- Accuracy/quality of data.

BIG DATA
by CloudAndDataUniverse

2. BIG DATA

EVOLUTION OF BIG DATA

Follow us on :

YouTube

LinkedIn

CDU
Cloud And Data Universe

Evolution

- Big Data in demand since last few years, but evolution spans across not years, rather decades.
- Good to know about the history.

Challenges

BIG DATA
by CloudAndDataUniverse

2 main challenges

- Storage
- Processing

Challenges

How to store massive amount of data?

How to process massive amount of data?

Big Data

Big Data is actually a problem!

Storage Solution

- Google comes to rescue!
- Google released a paper in 2003 on Google File system
- This paper explained how to store massive amounts of data, hence solving the storage challenge.

Processing Solution

- Again Google comes to rescue!
- Google released a paper in 2004 on MapReduce
- This paper explained how to process massive amounts of data, thereby solving the processing challenge.

Implementation

- Few Individuals at Yahoo implemented these papers and developed a framework which was named Hadoop!
- Later on Hadoop was handed over to Apache software foundation, an open-source and nonprofit corporation.
- The first version of Hadoop was released in 2006.

Implementation

- Storage solution in Hadoop was named HDFS(Hadoop distributed file system).
- Processing solution name was retained as MapReduce.

Apache Spark

- Apache Spark is an alternative to MapReduce
- Apache Spark research began in 2009.
- A paper was released in 2010.
- First version of Apache Spark was released in 2014.
- Apache Spark provides processing at lightning speed as it is memory based.

Commercial Solution

- In 2008 company named Cloudera was formed.
- It was first company to offer commercial Hadoop distributions.
- In 2011 company named HortonWorks was formed which also offered Hadoop solutions.
- Both companies later announced merger in 2018 which was completed in early 2019.

Databricks

- Creators of Apache spark founded a company and product named Databricks in 2013.
- Databricks offers a platform for Big data, Machine Learning and Lakehouse solutions.
- It is the go to choice and much demanded platform in Data Engineering!

Big Data Solutions on Cloud

- Top Cloud providers: Microsoft, Amazon and Google joined the race to provide Big Data solutions.
- All of them implemented solutions and made it much easier for users, developers and companies to work on Big data.
- Databricks is now being offered as a service on all these 3 cloud platforms: Azure, AWS & GCP.

Big Data

BIG DATA
by CloudAndDataUniverse

3. BIG DATA

DISTRIBUTED COMPUTING

Follow us on :

YouTube

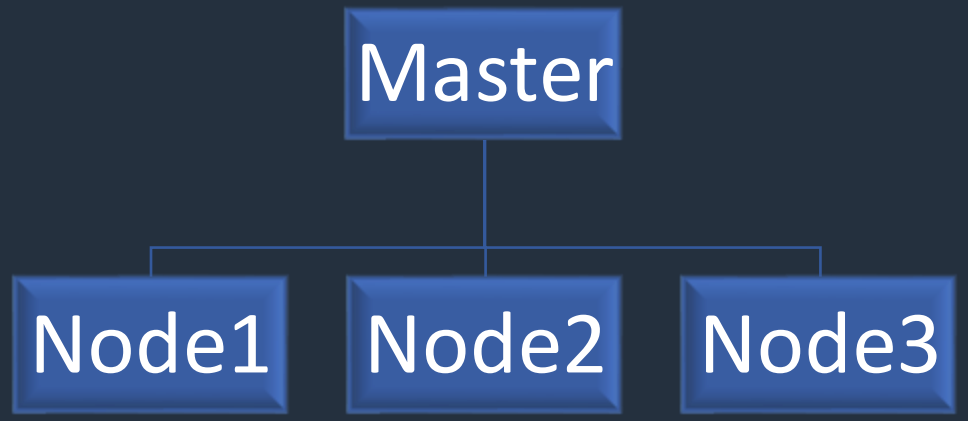
LinkedIn

CDU
Cloud And Data Universe

Distributed Computing

- If single entity can't process the work, divide it amongst multiple entities!
- Distributed computing has different types of architectures.
- We will look into Master – Slave architecture

Distributed Computing

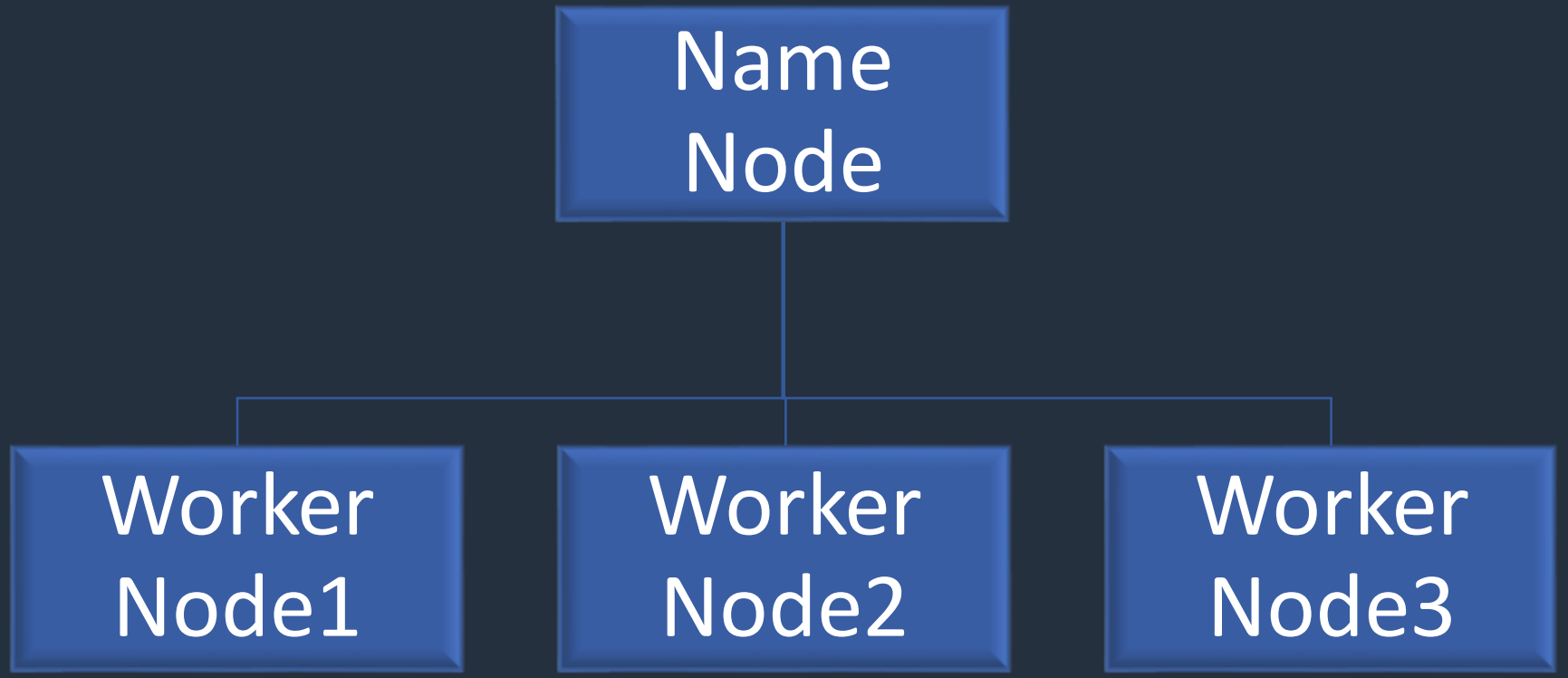


- Master – Slave Architecture contains set/group of individual machines.
- Master divides the workload and distributes amongst the slave nodes.
- Master node is also called as Name Node.
- Slave nodes are also called as worker nodes.
- Together the entire set up is called as Cluster which is a group of individual machines.

Cluster

BIG DATA
by CloudAndDataUniverse

3 Node Cluster



Cluster Hardware

- Machines are made up of Commodity hardware.
- Commodity hardware is affordable/inexpensive.
- There are rare chances of failure in Name node.

Big Data

BIG DATA
by CloudAndDataUniverse

4. BIG DATA

FEATURES

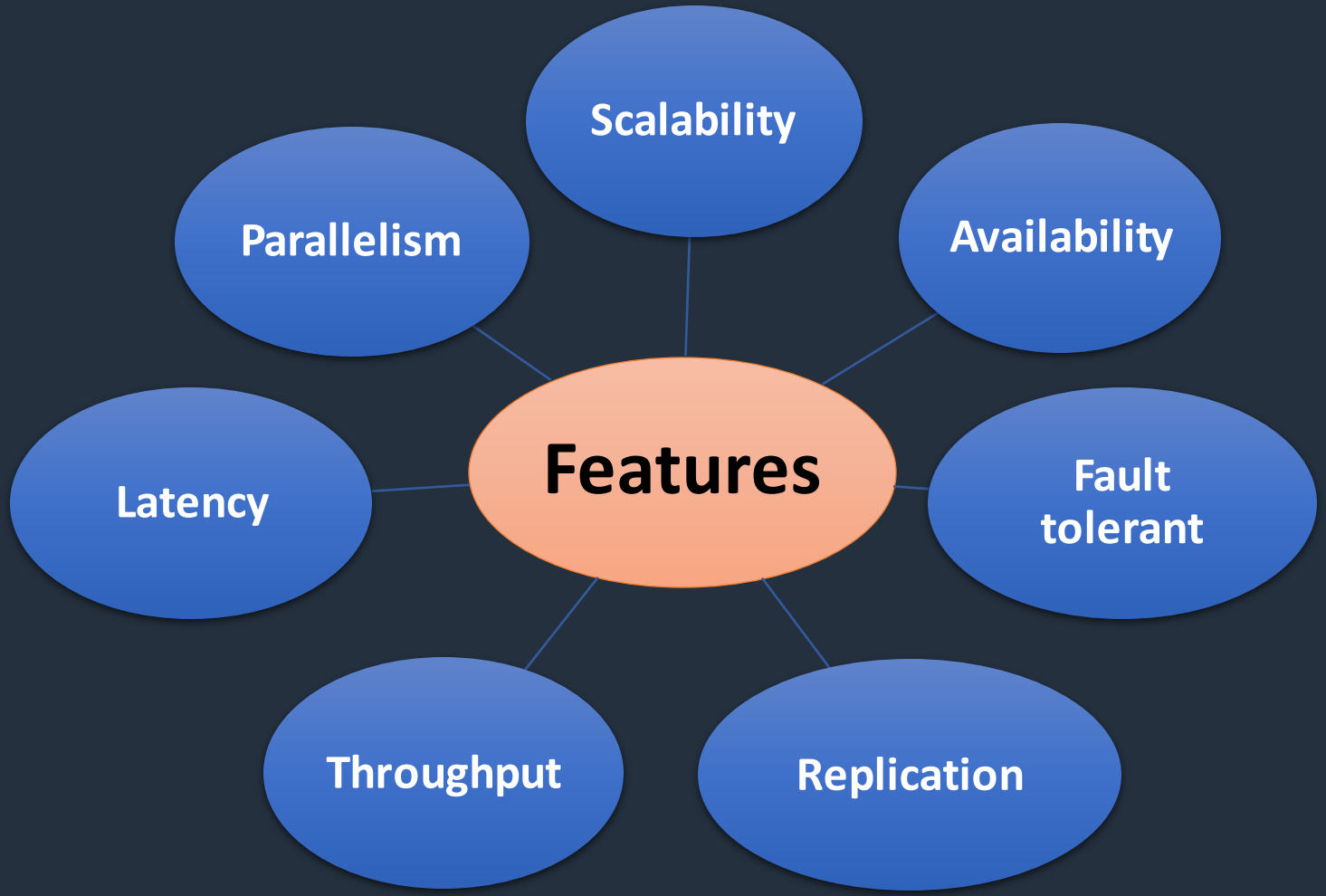
Follow us on :

YouTube

LinkedIn

CDU
Cloud And Data Universe

Features


Scalability
Parallelism
Availability
Features
Latency
Fault tolerant
Throughput
Replication

Scaling/Scalability

BIG DATA
by CloudAndDataUniverse

- It is ability of a system/cluster to increase or decrease the resources.
- 2 types: Vertical Scaling & Horizontal Scaling
- Vertical scaling: Increasing resources of existing system.
- Horizontal scaling: Adding more machines in cluster.
- Vertical scaling cannot be done beyond an extent due to limitations.
- Horizontal scaling can be done easily without disturbing the current setup.

Availability

- Distributed computing offers high availability!
- It can be defined as the time the system will be available.
- Linked to SLA(Service Level Agreement)

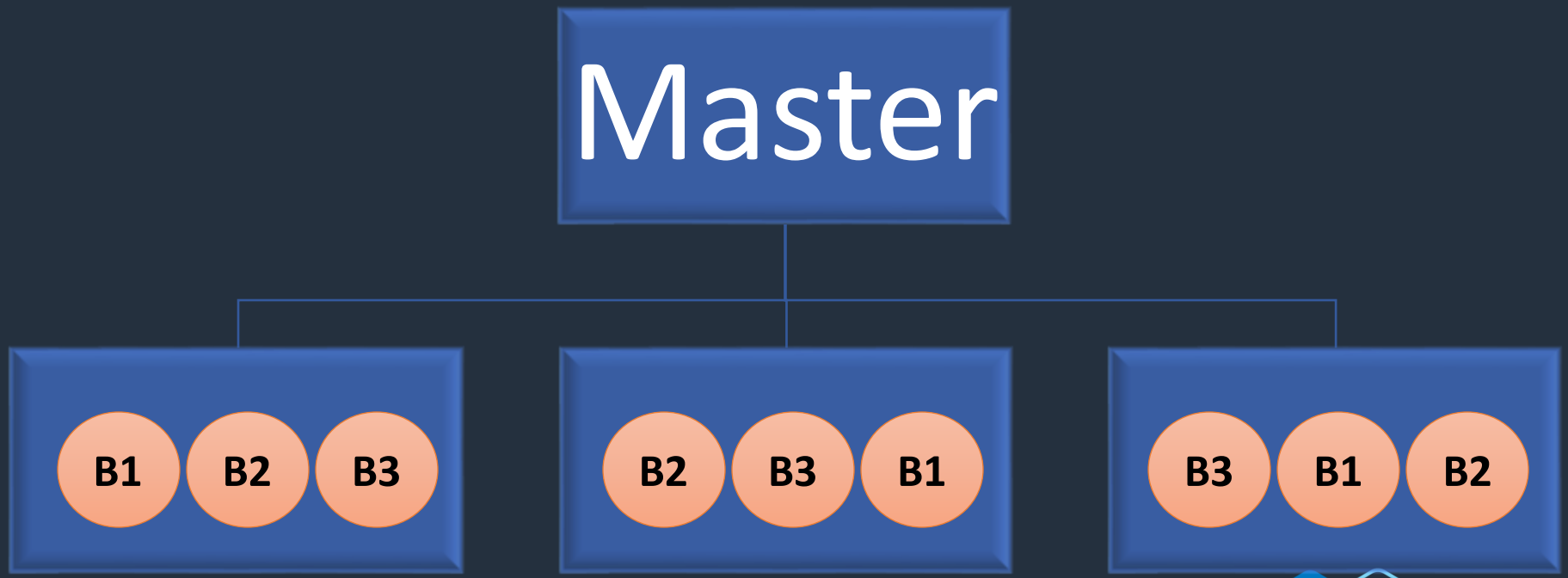
Fault Tolerant

- Fault tolerant is the ability of a system to keep running when one or more worker node is down.
- Fault tolerant system is always available!
- It comes at a cost.

Replication

BIG DATA
by CloudAndDataUniverse

- Multiple copies of data are kept on different nodes in cluster.
- Usually, replication factor is 3.



Throughput

- It is the amount of data/items which can be sent across the network or processed by system in a certain time frame

- Higher the throughput, better is the performance!

Latency

- It is the time taken to process a task.
- Low latency is always expected!

Parallelism

- It is the number of tasks/processes that can run at the same time.
- More cores more parallelism!

Big Data

BIG DATA
by CloudAndDataUniverse

5. BIG DATA

HADOOP ECOSYSTEM

Follow us on :

YouTube

LinkedIn

CDU
Cloud And Data Universe

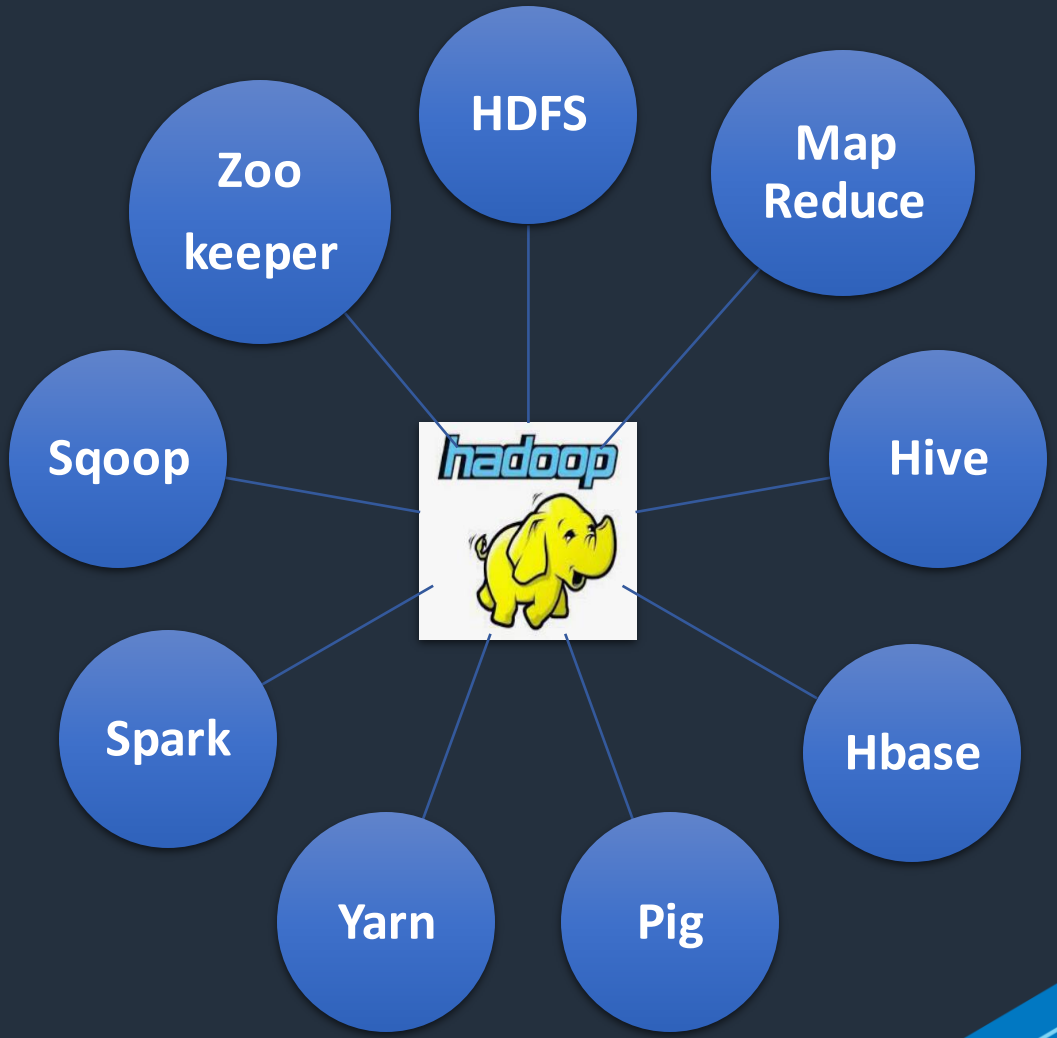
Hadoop

BIG DATA
by CloudAndDataUniverse

- Hadoop is a framework designed for Big data solutions.
- It consists of several tools/components.

Hadoop Ecosystem

BIG DATA
by CloudAndDataUniverse



HDFS

- HDFS stands for Hadoop distributed file system.
- It is the primary storage system of Hadoop.
- It runs on commodity hardware.
- It is highly fault tolerant.
- It is built on master-slave architecture.
- Files are stored across multiple nodes, also called as Data nodes.

HDFS

- HDFS is designed to store large files across multiple nodes in a cluster.
- Files are split into small chunks called as blocks.
- The default size of these blocks was 64MB in Hadoop V1 and 128 MB in Hadoop V2.
- These blocks are replicated across nodes to achieve fault tolerance.
- Both, block size and replication factor are configurable.

Map Reduce

BIG DATA
by CloudAndDataUniverse

- Map reduce is the processing solution!
- It has two stages: Map and Reduce.
- In Map, input is passed as key-value pair.
- Once map stage is completed, its output is passed to reduce stage.
- In Reduce stage you filter, sort, aggregate data.
- Map reduce programs can be developed in Java.

Sqoop

- Sqoop is a tool designed to transfer data.
- This process is known as ETL.
- Using Sqoop we can transfer data from HDFS to relational databases like MySQL, Hive, hbase.
- It uses a command line interface to process data transfer

Hive

- Hive is a data ware housing tool.
- It is a distributed data warehouse.
- We can run queries like SQL, which is known as HQL(Hive query language).
- Most demanded tool in Hadoop.

Hbase

- It is a NOSQL distributed database.
- Follows a column-oriented system.

Big Data

BIG DATA
by CloudAndDataUniverse

6. BIG DATA

TYPES OF PROCESSING

Follow us on :

YouTube

LinkedIn

CDU
Cloud And Data Universe

Types

- Broadly 2 types

Batch

Streaming

Batch processing

- Batch processing involves processing high volume of data in one go!
- Data size is known before processing begins!
- Data is processed in batches.
- Need to wait for the output until entire data is processed.
- Takes longer time to process!
- Needs huge amount of resources.

Streaming

- Streaming involves processing stream of data which is flowing continuously!
- Data size is unknown!
- Streaming processing provides real or near to real-time output.
- Stock market, social media platforms, e-commerce.
- Need less amount of resources.

Streaming tools

- Apache Kafka
- Spark streaming
- Amazon kinesis
- Google cloud dataflow
- Azure stream analytics