**BIG DATA**

Initially there was a very less amount of data .But when the internet came there was a huge amount of data.Semi structured and unstructured data got collected in the form of email,images,audio and videos and all this data came to be known as big data.

Large amount of data.Its a popular term used to express exponential growth of data.Big data is difficult to store,collect,maintain,analyze and visualize.



There was mass amount of data and it could be handled by Hadoop using the clustering technique.

**CHARACTERISTICS OF BIG DATA**

Volume,velocity,variety

Structured data->sql

Semi-structured->xml,json

Unstructured->text,audio,video

**BIG DATA SOURCES**

1.bank

2.social media

3.instruments

4.websites

5.stock market

**USE CASE OF BIG DATA**

1.recommendation

2.fraud detection

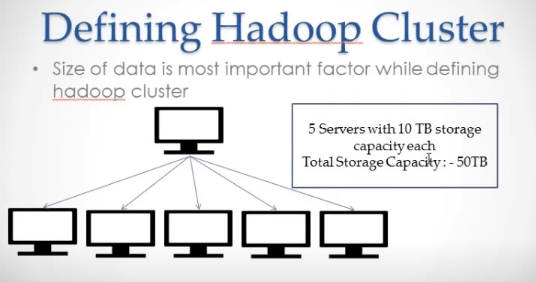
3.market basket analysis

4.sentiment analysis

**HADOOP**

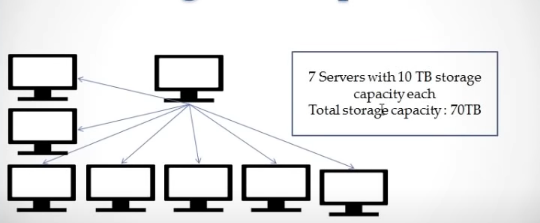
Open source framework that allows distributed processing of large datatsets on the cluster of commodity hardware.Hadoop is a data management tool and uses scale out storage.

**HADOOP CLUSTRE**

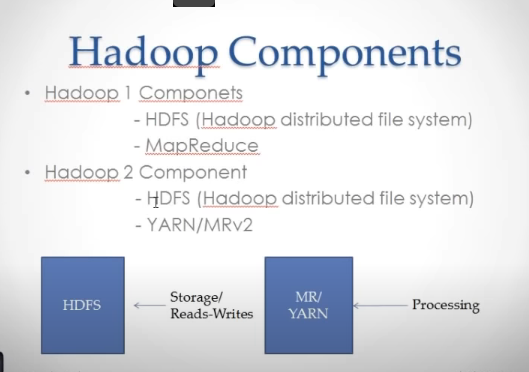
****

**A**  Hadoop cluster is a set of computers which can process a significant amount of data .suppose for the next six months our company will 50tb data to process so there will be 5 servers installed in the cluster with 10tb storage capacity each.

After 6 months 20tb more data came so thye called the Hadoop administrator to scale out their system.



**COMPONENTS OF HADOOP**

****

Hdfs is used to store the data

Mapreduce will process the data

Diagram

Description automatically generated

Similarly Hadoop 2 daemons

Diagram

Description automatically generated

Timeline

Description automatically generated

The daemons that run on master are called master daemon and the daemons that work on the slave system are called a s slave daemons.

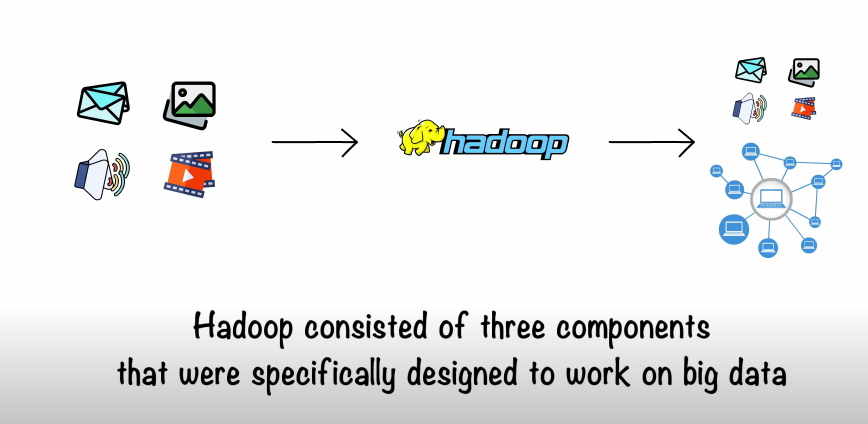
**HADOOP CLUSTER**

**Diagram

Description automatically generated**

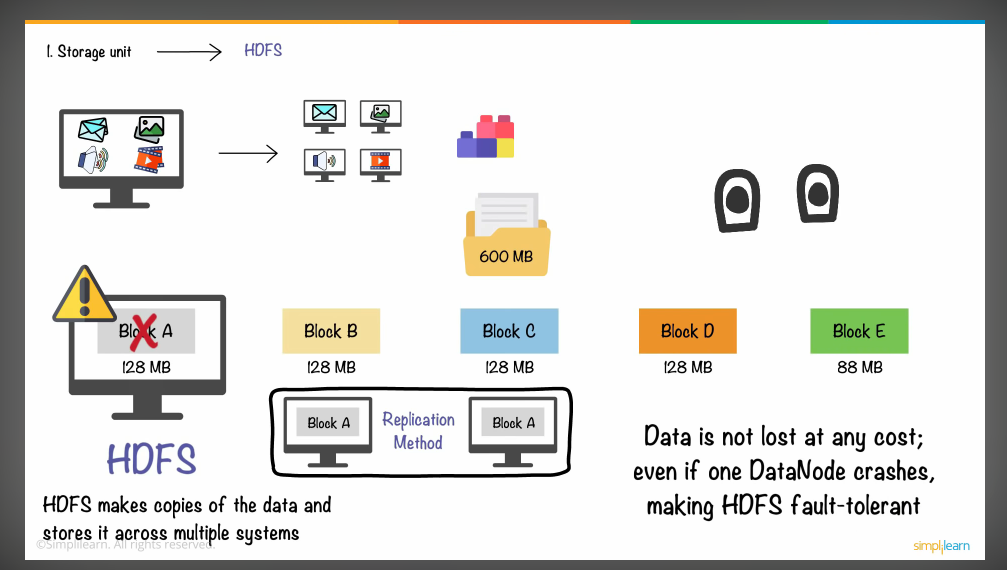
**Graphical user interface, text, application

Description automatically generated**



**HDFS**

Hdfs was used in order to store the data in an arrange dmanner.

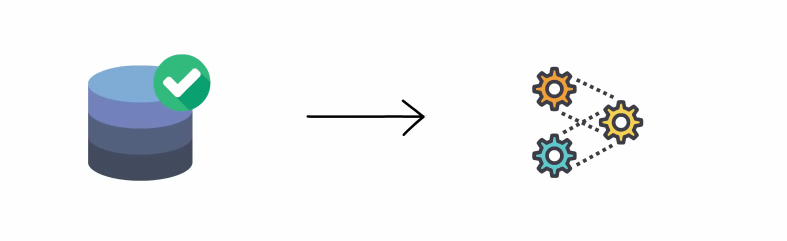


A hdfs files system stores the data in blocks of 128 mb each.so if we have a 600 mb data then it will be distributed on the diffwernet machines.

Now suppose we think that if data on block a is depleted then we will loose the data but it is not so.

A single block will be replicated to a afactor of 3 and stored at different locations.HDFS is fault tolerant.

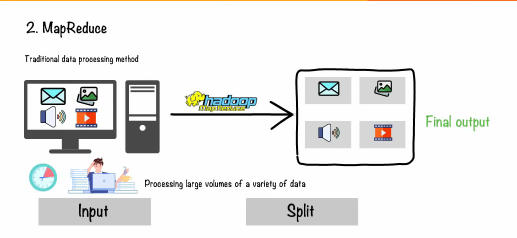
2.MAP REDUCE



After storing the data we need to get it processed.

In earlier times the data was processed on a single machine an dwas such a huge headache.This consumed time an dwas inefficient.

In order to overcome this the mapreduce separates data into parst and processes each data on separate machine.The individual results are aggregated to fomr the final output.



Diagram

Description automatically generated

Now our mapreduce ob is ready so now its time to run it on a Hadoop cluster.