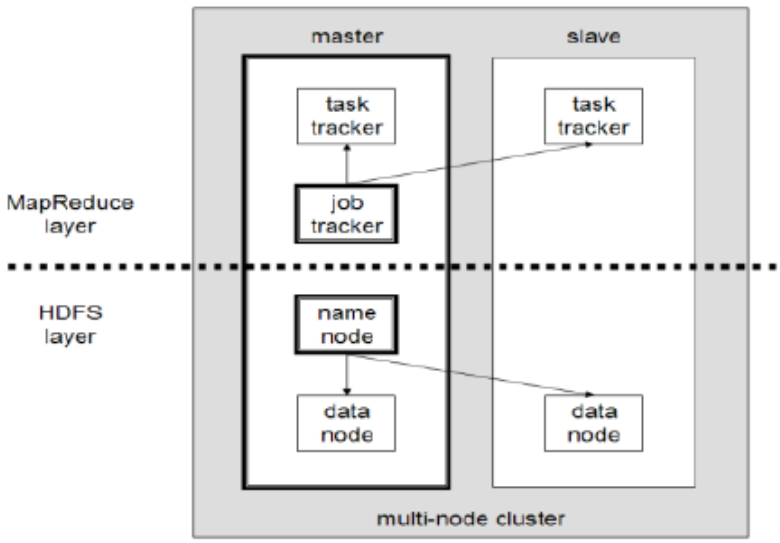
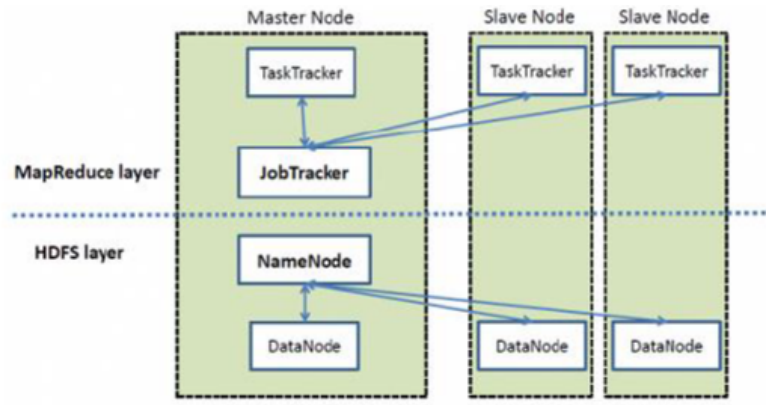
Platform

- Keras
  - Sequential model
    - Super simple
    - Only for single-input, single-output
    - sequential layer stacks
    - Good for 70+% of use cases
  - Functional API
    - Works like playing Lego bricks
    - Multi-input, multi-output, arbitrary static
    - graph topologies
    - Good for 95% of use cases
  - Model subclassing
    - Maximum flexibility
    - Larger potential error surface
- Hadoop
  - Hadoop is the open source implementation of the google file system and Map-Reduce distributed computation;
  - Hadoop masks the complexities that are involved in working with distributed systems and provides a fault tolerant distributed file system
  - (HDFS) as well as an API that makes programming in Hadoop a lot easier.
  - HDFS: Hadoop Distributed File System is based upon Google File System.
  - Hadoop is divided into HDFS and Map-Reduce. HDFS is used for storing the data and Map-Reduce is used for processing data.
  - Hadoop deals with files; It is not a database.
  - Hadoop HDFS Architecture

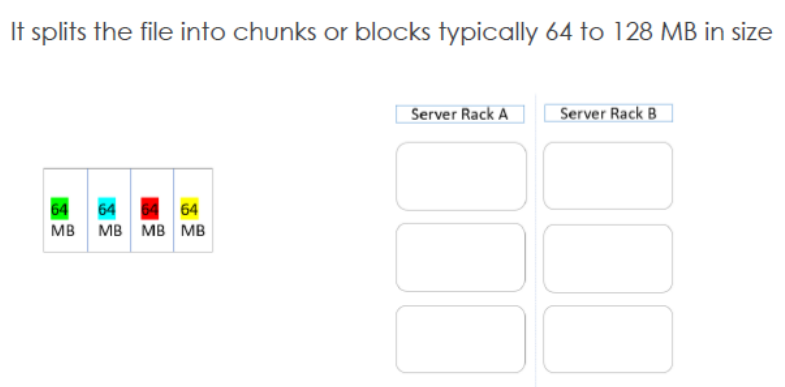


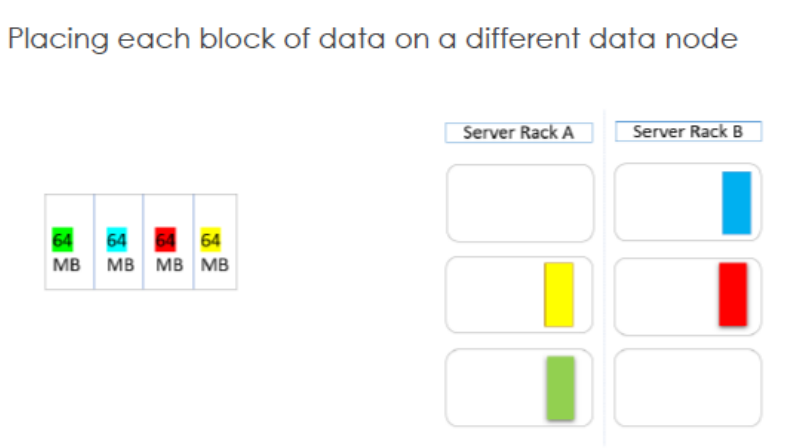
- The Hadoop distributed file system (HDFS) is a distributed, scalable, and portable file system written in Java for the Hadoop framework.
- HDFS has five services as follows:
  - 1. name node
    - HDFS consists of only one Name Node that is called the Master Node. The master node can track files, manage the file system and has the metadata of all of the stored data within it.
  - 2.Data node
    - This is also known as the slave node and it stores the actual data into HDFS.
  - 3.Job tracker
    - Job tracker talks to the Name Node to know about the location of the data that will be used in processing.
  - 4.Secondary Name Node
    - this is also known as the checkpoint Node. It is the helper Node for the Name Node.
  - 5.Task Tracker
    - It is the Slave Node for the Job Tracker and it takes the task from the Job Tracker
- Server Cluster

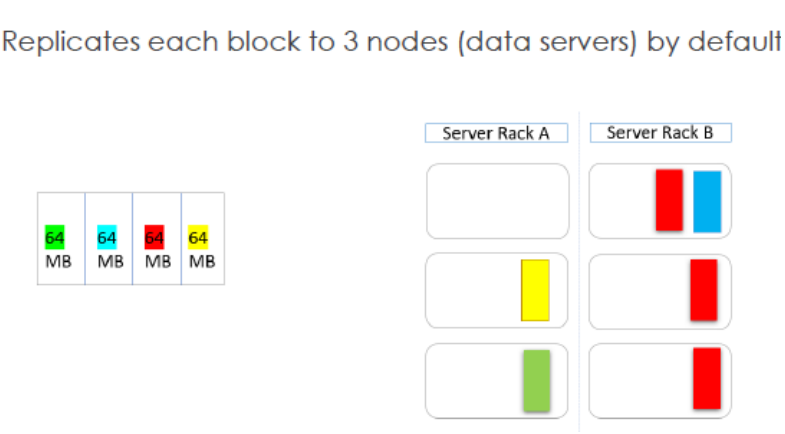


- Name Node is a master node and Data node is its corresponding Slave node and can talk with each other
- Enables multiple machines to perform computation of data;

- Example
  - Hadoop does two main things when a file is to be saved on HDFS:
    - It splits the file into chunks or blocks typically 64 to 128 mb data sizes;
    - Replicate each block of data (x3) and place them on a different data node;
  - With the replication factor of 3, the cluster has the power to choose among 3 servers, which one to use for computation.
  - In case any node goes down, there is another copy of the data that is available, achieving fault tolerance;
  - Pipeline

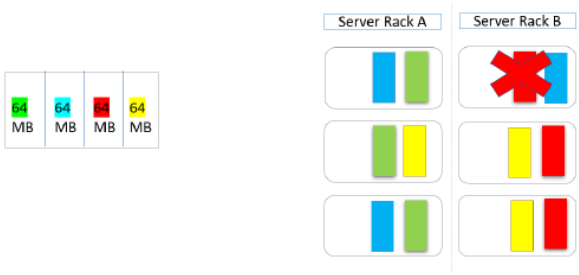






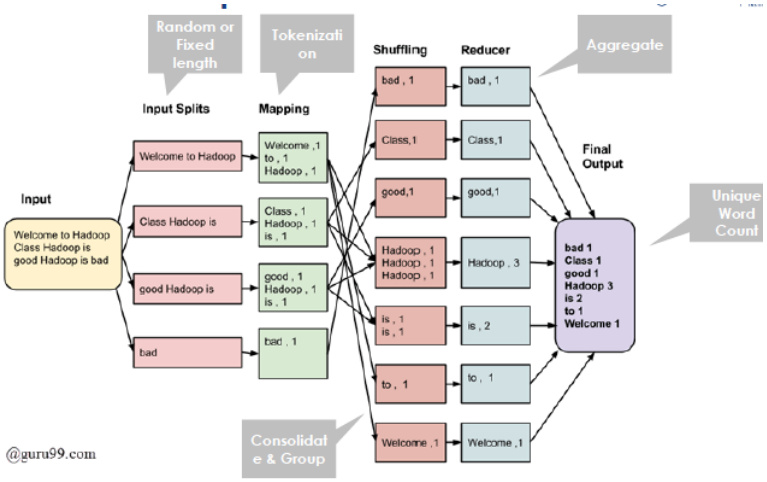
Enables multiple machines to perform computation of data

Fault recovery: in case any node goes down, there's another copy of the data that is available, achieving fault tolerance

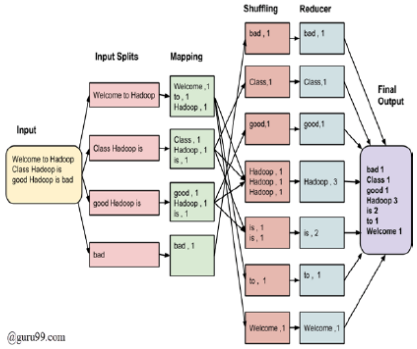


It tries to serve data nodes on the same rack, to avoid network traffic between nodes, thus called: rack aware

- Map Reduce
  - Map Reduce is a software framework used for parallelly computing or processing data that we store in HDFS;
  - Framework where (Hive Sql) queries are interpreted
  - A Map Reduce job has three phases: (divide & conquer method)
    - 1.Map
    - 2.Shuffle & Sort
    - 3.Reduce
  - example



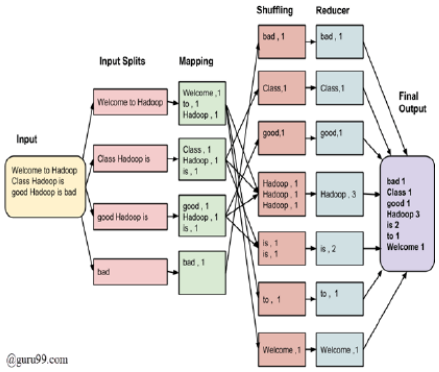
- A ***map-reduce job*** can use one or more mapper depending on the number of blocks the file spreads across.
- A ***mapper*** is assigned to each block. Here parallel distributed processing takes place given a file is split into blocks across multiple servers.
- Mappers take elements as a ***key*** and ***value*** and process them one at a time.



Maps are associated with the number of blocks of data required to read the input.

Shuffle and sort has the task of sorting the key-value pairs.

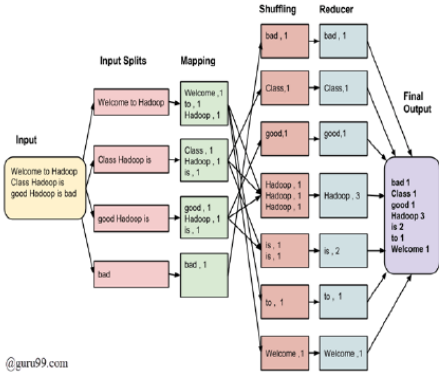
It uses the hash value of the key and splits keys into buckets according to the number of reducers.



Mapper makes sense of data (calculation) while ***reducers*** get the key value pair as an output from mappers and ***aggregate*** the results together.

No reducers are not dependent on the number of mappers. It can be configurated in the map-reduce job.

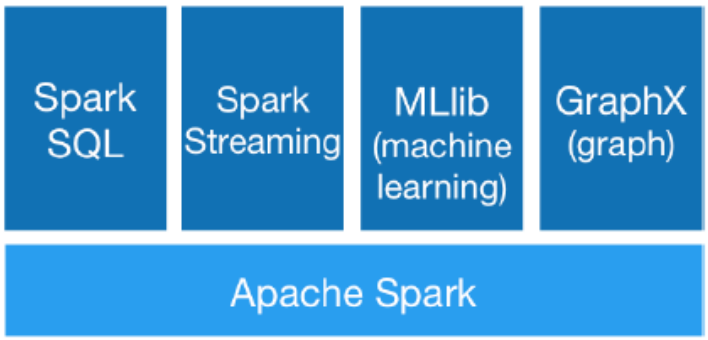
Output of reducer is sent to a file directory in HDFS.



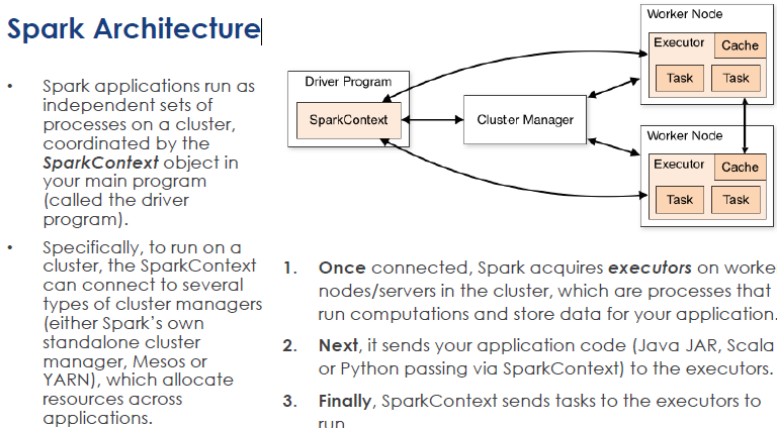
- Spark
  - Brief
    - Apache Spark is a unified analytics engine for large scale data processing.
    - It was the first unified analytics engine
    - Spark simplifies working with data by supporting different languages(e.g. SQL, Java, Python,etc)

- Spark brings data processing and analytics to one platform.
- Speed: Spark runs workloads 100x faster than Hadoop.
- Apache Spark achieves high performance for both batch and streaming data, using in memory computation compared with Hadoop's compute on disk

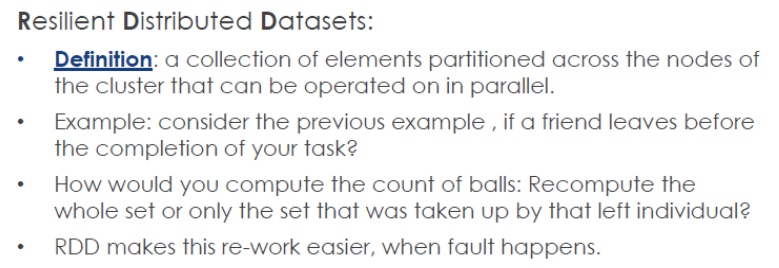
- Spark powers a stack of libraries including SQL and DataFrames , MLlib for machine learning, GraphX , Spark Streaming, interactive dashboarding, and advanced analytics.



- Architecture



- RDD



- Spark Summary
  - RDD is distributed and stored in various clusters on your system;
  - Each executor (data node server) works on their part of the data, the results are then aggregated and send back to the driver (name node server);
  - Dataframe in Spark inherits RDD properties (resilient and distributed) and metadata;
  - SparkSQL commands execute against dataframes (table alike);
  - Spark is not a dataframe, it's a in-memory compute engine that can read from databases;
  - Data is ephemeral: you never lose your data even when spark is down;

- Dataframe aggregation takes shorter time;