Name:- Muhammad Arham Azam
Roll# L1S22BSCS0413.

Assignment 2

(1) What is Hadoop? Components of Hadoop and How does it work?

Hadoop is a framework essential for managing vast amounts of data generated in the digital world. The three main components of Hadoop are:-

1) HDFS (Hadoop Distributed File system)
2) MapReduce
3) YARN (Yet Another Resource Negotiater)

Hadoop is significant because it efficiently deals with the challanges posed by big data, such as volume, velocity, variety, value and veracity. The advantages of Hadoop includes it's scalability, fault tolerence and ability to handle different data formats.

2) Hadoop Ecosystem:-

Hadoop has various tools dedicated to different aspects of data management, such as storing, processing, and analyzing. The key components of Hadoop ecosystem include:-

1) Hadoop Distributed File system (HDFS):-

A filesystem designed for storing massive datasets in commodity hardware, with a master daemon (Name Node) and multiple slave daemons (Data Nodes).

2) YARN (Yet Another Resource Negotiator):-

Manages the cluster of nodes and act as Hadoop's resource management unit, allocating resources to different applications.

3) MapReduce:-
The processing unit of Hadoop, processing large volumes of data in a parallel and distributed manner through the Phases Map, shuffle and sort, and reduce.

4) Sqoop:-
Transfers data between Hadoop and external data stores like relational databases, importing data into HDFS, Hive and HBase.

5) Flume:-
A distributed service for collecting, aggregating, and moving large amounts of log data into HDFS.

6) Pig:-
Allows non-programmers to analyze and process large datasetsets using a high level data processing language called pig latin.

7) Hive:-
Utilizes SQL to read, write, and manage large datasets in distributed storage, incorporating the concepts of tables and columns.

8) Spark:-
An open-source distributed computing engine that processes and analyzes real-time data, providing in-memory computation and running faster than MapReduce.

9) Mahout:-
Create scalable and distributed machine learning algorithms for clustering, linear regression, and classification.

10) Ambari:-
An open-source tool for managing, monitoring, and provisioning Hadoop clusters.

11) **Kafka :-**

A distributed streaming platform for storing and processing streams of records.

12) **Storm:-**

Processes real-time streaming data at high speed, integrated with Hadoop for higher throughput.

13) **Ranger:-**

A framework for enabling, monitoring, and managing data security across the Hadoop platform.

14) **Knox:-**

An application gateway used in conjuction with hadoop deployments for interacting with REST APIs and UIs.

15) **Oozie:-**

A workflow scheduler system used to manage Hadoop jobs, consisting of a workflow engine and a coordinator engine.

7) HDFS Tutorial:-

The increasing volume of data has led to the need for distributed file systems, and Hadoop's HDFS addresses this by providing reliable storage for extremely large files. HDFS is used in Hadoop cluster, consisting of master (name node) and slave (data node) nodes. It resolves challenges in traditional file systems related to cost, speed, and reliability. It operates on commodity hardware, delivering high throughput and fault tolerance. It stores data in blocks of 128 MB by default, replicating across nodes for reliability and ease of access. Name nodes store the name and metadata and data node stores and replicate data blocks.