Jamia Millia Islamia

Dept. Of Computer Science

Big Data Assignment-#5

Submitted By: Wasit Shafi

Roll no: 18MCA054

Submitted To: Dr.Mansaf Alam

Q1.Discuss various Hadoop cluster mode. Give the name of Hadoop 2.x configuration files along with description.

Sol1.

Hadoop is primarily structured and designed to be deployed on a massive cluster of networked systems or nodes, featuring master nodes.

Hadoop has three deployment modes and deploy these components as follows: Local standalone mode, pseudo-distributed (single node) mode and fully distributed mode (clustered nodes).

1. **Local standalone mode:** This is the default mode if, we don't configure anything else. In this mode, all the components of Hadoop, such NameNode, DataNode, JobTracker, and TaskTracker, run in a single Java process.
2. **Pseudo-distributed mode (single node):** A single-node Hadoop deployment is considered as running Hadoop system in pseudo-distributed mode, in which all the Hadoop services, including both the master and the slave services, were executed on a single compute node.
3. **Fully distributed mode (a cluster of nodes):** A Hadoop deployment in which the Hadoop master and slave services run on a separate cluster of nodes is running in what's stated as fully distributed mode. In this architecture, Hadoop is designed to be distributed across multiple machines, some of which might act as general-purpose workers and others might be dedicated hosts for components, such as NameNode and JobTracker.

The configuration files for Hadoop is under $HADOOP_COMMON_HOME/etc/hadoop for our installation

1) HADOOP-ENV.sh-It specifies the environment variables that affect the JDK used by Hadoop Daemon (bin/hadoop). We know that Hadoop framework is wriiten in Java and uses JRE so one of the environment variable in Hadoop Daemons is $Java_Home in Hadoop-env.sh.

2) CORE-SITE.XML-It is one of the important configuration files which is required for runtime environment settings of a Hadoop cluster.It informs Hadoop daemons where the NAMENODE runs in the cluster. It also informs the Name Node as to which IP and ports it should bind.

3) HDFS-SITE.XML-It is one of the important configuration files which is required for runtime environment settings of a Hadoop. of replications can also

Q2.Discuss Big Data Customer Scenarios.

Sol2. Big Data in Education:

Education industry is flooding with huge amounts of data related to students, faculty, courses, results.

Big Data in Media and Entertainment Industry :

Some of the benefits extracted from big data in the media and entertainment industry are given below:

• Predicting the interests of audiences

• Optimized or on-demand scheduling of media streams in digital media distribution platforms

• Getting insights from customer reviews

• Effective targeting of the advertisements

Big Data in Healthcare Industry :

Healthcare is yet another industry which is bound to generate a huge amount of data. Following are some of the ways in which big data has contributed to healthcare: • Big data reduces costs of treatment since there is less chances of having to perform unnecessary diagnosis. • It helps in predicting outbreaks of epidemics and also in deciding what preventive measures could be taken to minimize the effects of the same.

Big Data in Government Sector :

Governments, be it of any country, come face to face with a very huge amount of data on almost daily basis. The reason for this is, they have to keep track of various records and databases regarding their citizens, their growth, energy resources, geographical surveys, and many more.

Welfare Schemes

- In making faster and informed decisions regarding various political programs

• To identify areas that are in immediate need of attention

Cyber Security

• Big Data is hugely used for deceit recognition.

• It is also used in catching tax evaders.

Q3.Discuss limitations of existing data analytics architecture.

Sol3.

Issue with Small Files :

Hadoop does not suit for small data. (HDFS) Hadoop distributed file system lacks the ability to efficiently support the random reading of small files because of its high capacity design.

Slow Processing Speed :

In Hadoop, with a parallel and distributed algorithm, the MapReduce process large data sets. There are tasks that we need to perform: Map and Reduce and, MapReduce requires a lot of time to perform these tasks thereby increasing latency.

Support for Batch Processing only :

Hadoop supports batch processing only, it does not process streamed data, and hence overall performance is slower.

No Real-time Data Processing :

Apache Hadoop is for batch processing, which means it takes a huge amount of data in input, process it and produces the result. Although batch processing is very efficient for processing a high volume of data, depending on the size of the data that processes and the computational power of the system, an output can delay significantly. Hadoop is not suitable for Real-time data processing.

Q4.What are Hadoop key characteristics? Discuss hadoop 2.x core components

Sol4.

Hadoop provides a reliable shared storage (HDFS) and analysis system (MapReduce).

• Hadoop is highly scalable and unlike the relational databases, Hadoop scales linearly. Due to linear scale, a Hadoop Cluster can contain tens, hundreds, or even thousands of servers.

• Hadoop is very cost effective as it can work with commodity hardware and does not require expensive high-end hardware.

• Hadoop is highly flexible and can process both structured as well as unstructured data.

• Hadoop has built-in fault tolerance. Data is replicated across multiple nodes (replication factor is configurable) and if a node goes down, the required data can be read from

another node which has the copy of that data. And it also ensures that the replication factor is maintained, even if a node goes down, by replicating the data to other available nodes.

• Hadoop works on the principle of write once and read multiple times.

Hadoop 2.x has two core Components:

1) HDFS(Hadoop Distributed File System) is a storage system.

2) YARN(Yet Another Resource Negotiator) is a processing system.