Om Shree
20060077

IOT Assignment - Chapter 10

① What do you mean by data analytics? Why is cleaning of data required?

Ans) Data analytics is the process of examining data sets in order to find trends and draw conclusions about the information they contain.

Data cleaning is required because :—

- Removal of errors when multiple sources of data are at play.
- Ability to map different functions and what your data is intended to do.
- Increases overall productivity
- Allows for highest quality information in your decision making.

② What is big data? What are characteristics, classifications & challenges?

Ans) Big data is defined as data arriving in increasing volumes and with more velocity. This data contains high variety.

- Characteristics of big data :—

(1) Volume

(2) Value

(3) Variety

(4) Velocity

(5) Veracity

- classifications of big data :—

(1) Un-structured

(2) Semi-structured

(3) Structured

- Challenges of big data :—

(1) Data quality

(2) Data storage

(3) Lack of data science professionals

(4) Validating data

(5) Accumulating data from various sources.

(3) What was traditional approach for storing & processing data ?

Ans) Traditional approach made use of ETL software.

① Data generated out of organizations is given as an input to ETL system.

② ETL system extracts the data & converts it into proper format. This data is then loaded into database.

③ End users can generate reports & perform analytics by acquiring this data.

④ Exponential growth of data can't be managed using ETL.

④ What do you mean by Hadoop, Hadoop cluster and features of Hadoop.

Ans) Hadoop is an open source framework that is used to efficiently store and process large datasets ranging in size from gigabytes to petabytes.

- Hadoop cluster is a collection of computers, known as nodes, that are networked together to perform these kinds of parallel computations of big datasets.

- Features of Hadoop :—

① Cost effective system
② Large cluster of nodes
③ Parallel processing
④ Distributed data
⑤ Automatic fail over management
⑥ Data locality optimization
⑦ Scalability
⑧ Heterogenous cluster

(5) How does Hadoop work? Explain Hadoop ecosystem.

Ans) Hadoop ecosystem is a platform which provides various services to solve big data problems. There are 4 major elements of Hadoop :—

(1) HDFS

(2) mapReduce

(3) YARN

(4) Hadoop common

In addition to these, we have :—

(5) HBase

(6) PIG/HIVE

(7) Spark

- The data is submitted to Hadoop. HDFS stores the data and mapReduce processes the data. Yarn is responsible for division of tasks.

(6) What is HDFS? mention its pros & cons.

Ans) HDFS stands for Hadoop Distributed File System. HDFS follows a master-slave topology.

- Pros of HDFS :—

(1) Cost

(2) Scalability

(3) Flexibility

(4) Speed

(5) Fault tolerance

⑥ High throughput

⑦ Minimum network traffic

• Cons of HDFS :—

① Problem with small files

② Vulnerability

③ Low performance in small data surrounding.

④ Lack of security

⑤ High up processing

⑥ Supports only batch processing

⑦ What do you mean by HDFS daemons?

Ans) ~~HDFS dae~~ ~~HDFS do~~ HDFS daemons are processes. All HDFS daemons follow the master-slave architecture. HDFS consists of 3 daemons, which are :—

① Namenode

② Datanode

③ Secondary namenode

• Namenode is the master node whilst there are multiple slave datanodes. Namenode is used for storing metadata of HDFS. The datanodes are responsible for retrieving & storing information as instructed by namenode.

⑧ Explain secondary name node.

Ans) Namenode holds metadata for HDFS like block information, size, etc. This information is stored in main memory as well as disk for persistance damage. Information is stored in 2 different files :-

- Edit Logs : keeps track of each & every changes to HDFS
- Fsimage : stores the snapshot of file system

→ Namenode is also a single point of failure. To solve this problem, the secondary name node gets edit logs from namenode periodically & copies it to fsimage. The purpose of secondary name node is to have checkpoints in HDFS. Hence, it is called checkpoint node.

⑨ How does reading & writing happen in Hadoop system?

Ans) Anatomy of file read in HDFS :-

Step 1 : Client opens the file to read by calling open() on the file system.

Step 2 : HDFS calls the namenode using remote procedure calls to determine the location of first few blocks. DFS returns an FSDataInput stream.

Step 3: The client then calls read() on the stream.

Step 4: Data is streamed from the data node back to the client, which calls read() repeatedly on the stream.

Step 5: When the end of block is reached DFSInputStream will close the connection.

Step 6: Client calls the close function.

→ Anatomy of file write in HDFS:—

Step 1: Client calls the create() function.

Step 2: DFS makes an RPC call to the name node to create a new file in the file system's namespace, with no blocks associated with it. DFS returns an DFSOutputStream for client to start writing data to.

Step 3: Data is split into packets which are written to info queue. Data streamer streams the data to primary data node.

Step 4: DFSOutputStream sustains an internal queue of packets which are waiting to be acknowledged.

(10) Hadoop rack arrangement & Hadoop FS metadata.

Ans) The rack is a collection of around 40-50 datanodes connected using the same network switch. If the network goes down, the whole rack will be unavailable. A large Hadoop cluster is deployed in multiple racks.

→ Hadoop FS metadata consists of

editlog

fsimage

- fsimage contains serialized form of all directory & file in the file system. fsimage is stored as a file in the NameNode's local file system.
- EditLog : This is a transaction log, which logs every change in the file system.