BIG DATA ANALYTICS ASSIGNMENT II

Name: Aditya Mahajan
Roll No: A - 10

Reg No - 2017BCS001

17. What kind of Dataware house application is suitable?
Answer: Hive is not a full database. The design constraints and limitations of Hadoop and HDFS impose limits on what Hive can do.
Hive is most suited for data warehouse applications, where
1) Relatively static data is analyzed,
2) Fast response times are not required, and
3) When the data is not changing rapidly.

Hive doesn't provide crucial features required for OLTP, Online Transaction Processing. It's closer to being an OLAP tool, Online Analytic Processing. So, Hive is best suited for data warehouse applications, where a large data set is maintained and mined for insights, reports, etc.

18. what are Binary storage formats hive supports?
Answer: Hive natively supports the text file format, however, hive also has support for other Binary formats. Hive supports Sequence, Avro, RCFiles.
Sequencefiles: -General binary format. splittable, compressible, and row-oriented. a typical example can be. if we have lots of small files, we may use a sequence fle as a container, where the filename can be a key and content could store as a value. it supports compression which enables a huge gain in performance.
Avro datafiles:- Same as Sequence file splittable, compressible and row-oriented except support of schema evolution and multilingual binding support.

files: -Record columnar file, it's a column-oriented storage file. it breaks the table into a row split. in each split stores that value of the first row in the first column and followed sub subsequently.

19. What are the main configuration parameters in a "MapReduce" program?

Answer: The main configuration parameters which users need to specfy in the "MapReduce" framework are:

- Job's input locations in the distributed file system
- Job's output location in the distributed file system
- The input format of data
- The output format of data
- Class containing the map function
- Class containing the reduce function
- JAR file containing the mapper, reducer, and driver classes

20. What are the key steps in Big Data Solutions?

Answer: Key steps in Big Data Solutions

Ingesting Data, Storing Data (Data Modelling), and Processing data (Data wrangling, Data transformations, and querying data).

- Ingesting Data
- RDBMS Relational Database Management Systems like Oracle, MySQL, etc.
- ERPS Enterprise Resource planning (ERP) systems like SAP.
- CRM Customer Relationships Management systems like Siebel, Salesforce, etc.
- Social Media feeds and log files.
- Flat fles, docs, and images.

Storing Data

- Data Storage Formats
- Data Modelling
- Metadata management
- Multitenancy

21. What is Big Data Analysis?

Answer: It is defined as the process of mining large structured/unstructured data sets. It helps to find out underlying patterns, unfamiliar and other useful information within a data leading to business benefits.

22. Where the Mappers Intermediate data will be stored?

Answer: The mapper output is stored in the local file system of each individual mapper node. Temporary directory location can be set up in the configuration
By the Hadoop administrator.
The intermediate data is cleaned up after the Hadoop Job completes.

23. What do you mean by logistic regression?

Answer: Also known as the logit model, Logistic Regression is a technique to predict the binary result from a linear amalgamation of predictor variables.

24. How Big Data can help increase the revenue of the businesses?

Answer: Big data is about using data to expect future events in a way that progresses the bottom line. There are oodles of ways to increase profit. From email to a site, to phone calls and interaction with people, this brings information about the client's performance. Undoubtedly, a deeper understanding of consumers can improve business and customer loyalty. Big data offers an array of advantages to the table, all you have to do is use it more efficiently in order to an increasingly competitive environment.

25. What are the responsibilities of a data analyst?

Answer: Helping marketing executives know which products are the most profitable by season, customer type, region, and other feature

Tracking external trends relatives to geographies, demographics and specific products Ensure customers and employees relate well Explaining the optimal staffing plans to cater to the needs of executives looking for decision support.

26. What do you know about collaborative filtering?

Answer: A set of technologies that forecast which items a particular consumer will like depending on the preferences of scores of individuals. It is nothing but the tech word for questioning individuals for suggestions.

27. What is a Block in Hadoop Distributed File System (HDFS)?

Answer: When the file is stored in HDFS, all file system breaks down into a set of blocks and HDFS unaware of what is stored in the file. Block size in Hadoop must be 128MB. This value can be tailored for individual files.

28. Define Active and Passive Namenodes?

Answer: Active NameNode runs and works in the cluster whereas Passive NameNode has comparable data like active NameNode.

29. Which are the essential Hadoop tools for the effective working of Big Data? Answer: Ambari, "Hive", "HBase, HDFS (Hadoop Distributed File System), Sqoop, Pig, ZooKeeper, NoSQL, Lucene/SolrSee, Mahout, Avro, Oozie, Flume, GIS Tools, Clouds, and SQL on Hadoop are some of the many Hadoop tools that enhance the performance of Big Data.

30. It's true that HDFS is to be used for applications that have large data sets. Why is it not the correct tool to use when there are many small files?

Answer: In most cases, HDFS is not considered as an essential tool for handling bits and pieces of data spread across different small-sized files. The reason behind this is "Namenode" happens to be a very costly and high-performing system. The space allocated to "Namenode" should be used for essential metadata that's generated for a single file only, instead of numerous small files.

While handling large quantities of data attributed to a single file, "Namenode" occupies lesser space and therefore gives off optimized performance. With this in view, HDFS should be used for supporting large data files rather than multiple files with small data.

31. What are the main distinctions between NAS and HDFS?

Answer: HDFS needs a cluster of machines for its operations, while NAS runs on just a single machine. Because of this, data redundancy becomes a common feature in HDFS. As the replication protocol is different in the case of NAS, the probability of the occurrence of redundant data is much less.

On the other hand, the local drives of the machines in the cluster are used for saving data blocks in HDFS.

Unlike HDFS, Hadoop MapReduce has no role in the processing of NAS data. This is because computation is not moved to data in NAS jobs, and the resultant data files are stored without the same.

32. Explain "Big Data" and what are five V's of Big Data?

Answer: "Big data" is the term for a collection of large and complex data sets, that makes it difficult to process using relational database management tools or traditional data processing applications. It is difficult to capture, curate, store, search, share, transfer, analyze, and visualize Big data. Big Data has emerged as an opportunity for companies. Now they can successfully derive value from their data and will have a distinct advantage over their competitors with enhanced business decisions making capabilities.