CAE - II

Name: Mohnish Devaraj

Reg No: 39110636

Roll No: 195115398

Date: 06 Apr 2022

Subject Name: Big Data Analytics

Subject Code: SCSA1603

Total pages: 8

PART - B

(6) Cassendra Data Model

Cassendra is a NOSQL database, which is a key-value store. Some of the features of Cassendra data Model are as follows:

→ Data in Cassendra is stored as a set of rows that are organised into tables.

→ Tables are also called Column families

→ Each row is identified by a primary key value.

Cassendra dat Model provides a mechanism for data storage. The components of Cassendra data model are key-space, tables, and Columns.

a) Key spaces

Cassendra Data Model consists of Keyspace at the highest level. keyspaces are the containers of data, similar to the schema or database in a relational database. Some of the features of key spaces are:

→ A key space needs to be defined before creating tables, as there is no default keyspace.

→ A keyspace can contain any number of tables belongs to only to one key space. This represents a one-to-many relationship.

→ Replication is specified at the key space level. For example, replication of three implies that each data row in the key space will have three compies.

b) Tables

Within the keyspaces, the tables are defined. Tables are also refferred to as Column Families in the carlier versions of Cassendra. Tables contain a set of Columns and a primary key, and they store data in a set of rows. Some features of tables are:

→ Tables have multiple rows and columns. As mentioned earlier, a table is also called Column Family in the earlier versions of Cassandra.

→ It is still referred to as Columns family in some of the error messages and documents of Cassandra.

→ It is important to define a primary key for a table.

Columns

Columns define the structure of data in a table. Each column has an associated, type, such as integer, text, double and Boolean. These Cassandra data model components will be discussed in detail. some of its features are:

→ Columns consists of various types, such as integer, text, float, double, and Boolean.

→ Cassandra also provides collection types such as set, list and map

→ This time stamp can be retrieved using the function write time.

Features:

→ Distributed
→ Fault-tolerance
→ support replication
→ Map Reduce support
→ Scalability
→ Query Language

(9) Hadoop Ecosystem is a platform or a suite which provides various services to solve the big data problems. It includes Apache projects and various commercial tools and solution. Most of the tools or solutions are used to supplement or suppor the major elements.

a) HDFS:

HDFS is the primary or major Component of Hadoop Ecosystem and is responsible for storing large data sets of structured and unstructured data across various nodes and there by maintaining the metadata in the form of log files.

HDFS Consists of two nodes: Name node and Data node

b) YARN:

Yet Another Resoure Negotiator, as the name implies, YARN is the one who helps to manage the resources across the clusters. In short, it performs scheduling and resource allocation for the Hadoop System.

c) Map Reduce

By Making the use of distributed and paralled algorithms, MapReduce makes it possible to carry over the processing logic and helps to write application which transforms big data sets into a manapable one

d) PIG:

Pig was basically developed by Yahoo which works on a Pig Latin Language, which is Query based Language similar to SQL.

e) Hive

With the help of SQL methodology and interface, HIVE performes reading and writing of large sets. However, it's query language is called HQL.

f) Mahout:

Mahout, allows machine Learnability to a system or application. Machine Learning, as the name suggests help the system to develop itself based on some patterns.

g) Apache spark

It's a platform that handles all the process consumptive tasks like batch processing, interactive or iterative ~~real~~ real-time processing, graph conversions and visualization, etc.

h) Apache HBase

It's NoSQL database which supports all kind of data and thus capable of handling anything of Hadoop Database.

Read:

step 1: The client opens the file it wishes to read by calling open() on the file system Object.

step 2: DFS calls the name node, using remote procedure calls, to determine the locations of the blocks in the file.

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()

step 5: When the end of the block is reached, DFSInput stream will close the connection to the data node.

step 6: When the client has finished, reading the file, a function is called close()

Write:

step 1: The client creates a file by calling create

step 2: DFS makes an RPC call to the name node to create a new file in the file system namespace.

step 3: Because the client writes the data, the DFS output stream splits it into packets, which it writes to an indoor queue.

step 4: The second data node stores the packet and forwards it to the third dat node

step 5: The DFSOutput stream sustains an internal queue of packets.

step 6: This action sends up all the remaining packets to the datanode pipeline.

PART-A

① MangoDB features

→ Ad-hoc queries for optimized, realtime analytics.
→ indexing approprately for better query execution.
→ Replication for better data availabity & stabitity
→ sharing
→ load balancing

② On a dash board, double-click the file in the Edit title dealog box that appears, format the title as you. Enter a new title for the dashboard or a description.

③ There are three components of Hadoop.

* Hadoop HDFS - Hadoop Distributed File system (HDFS) is a storage unit of Hadoop.

* Hadoop map Reduce - Hadoop map Reduces is the processing unit of Hadoop

* Hadoop YARN - Hadoop YARN is a resource manager unit of Hadoop.

(4) the Name Node is the center piece of an HDFS file system. It keeps the Directory tree of all file in the file system, & tracks where across the cluster the file data is kept. It does not store the data of these files itself.

(5) Job tracker

Job tracker is a master which creates & runs the job. Job tracker which can run on the Name node allocates the job to task trackers. It is tracking resource availability & task like cycle management, tracking is process, fault tolerance, etc.

Task tracker

Task tracker run the tasks & report the status of task to job tracker. Task tracker run on Data nodes. It has function of following the orders of the job tracker & updating the ~~Job tra~~ job tracker with its progress status periodically.