Rannvijay Singh Rawat SAP-1000011240

Roll no.- 180109029

BIG DATA ANALYSIS

ASSIGNMENT-2

Q-1 - What are the four storage formats of Big SQL. Explain.

Ans - The four storage formats of Big SQL are:-

(i) Optimized Row Columnar (ORC) → This format provides a highly effiient data way to store data. ORC files stores collection of rows in a columnar format, which enables parallel processing of row collections across the cluster. The ORC file format uses type-specific encoders for each column and divides the file into large stripes. The recommended compression type for this file format is zlib (the default).

(ii) Record Columnar (RC) → The RC file format is an effiient high performance format that uses binary key/value pairs. It partitions rows horizontally into row splits and partitions each row split vertically. The following compression types are recommended: bzip2, deflate, gzip, snappy.

(iii) Parquet → This file format is an Open source columnar storage format for Hadoop that supports effiient compression and encoding schemes. During load and insert operations, the following values are set as the default values for the Parquet format. You can change these values by using the SET HADOOP PROPERTY command before you run the LOAD HADOOP statement:

- SET HADOOP PROPERTY 'dfs.blocksize' = 268435456;
- SET HADOOP PROPERTY 'parquet.page.size' = 65536;
- SET HADOOP PROPERTY 'parquet.compression' = 'SNAPPY'

Scanned by TapScanner

Rannvijay Singh Rawat
Roll no.:- 180109029

SAP-1000011240

(iv) Sequence → This file format is used to hold arbitrary data that might not otherwise be splittabb. A sequence file maintains additional metadata to recognize record boundaries. There are 2 types of sequence files:

- A binary sequence file stores data in a binary format by using the Hive. With binary storage, the data requires very little conversion processing while being read.
- A text sequence file stores delimited data within the sequence file format, which enables the use of compression algorithms on tenual data that would not otherwise be splittabb.

The following compression types are recommended: bzip2, deflate.

Q2 - What is the difference between static partioning and dynamic partioning.

Ans →

Static Partition	Dynamic Partition
(i) Insert input data files individually into a partition table is Static Partition.	(i) Single insert to Partition table is known as dynamic Partition.
(ii) Static Partition saves your time in loading data compared to dynamic Partition.	(ii) Dynamic Partition takes more time in loading data compared to static Partioning.
(iii) We can alter the partition in static partition.	(ii) We can alter the partition in Dynamic partition.
(iv) Usually when loading big files into Hive tables static Partitions are preferred	(iv) Usually dynamic partition load the data from non partioned table.

Rannvijay Singh Rawat SAP - 1000011240

Roll no - 180109029

Q-3 - How buckets are defined explain with real time example.

Ans - Mechanism to query and examine random sample of data. Break data into a set of buckets based on a hash function of a "bucket column". Capability to execute queries on a subset of random data.

Basically, this concept is based on hashing function on the bucketed column. Along with mod (by the total number of buckets).

(i) Where the hash_function depends on the type of bucketing column.

(ii) However the records with the same bucketed column will always be stored in the same bucket.

(iii) Moreover, to divide the table into buckets we use CLUSTERED BY clause.

(iv) Along with portioning or Hive tables bucketing can be done and even without partioning.

(v) Moreover, Bucketed tables will create almost equally distributed data file parts.

A-4 - Explain in detail about SequenceFile in Hadoop.

Ans - Sequence files are flat files consisting of binary key-value pairs. When Hive converts queries to MapReduce jobs, it decides on the appropriate key-values pairs to be used for a given record. Sequence files are in the binary format which can be split and the main use of these files is to club two or more smaller files and make them as one sequence file. Sequence files acts as a container to store the small files. The Sequence files provides a Writer, Reader and Sorter classes for writing, reading and sorting respectively.

We know that Hadoop's performance is drawn out when we work with the small no. of files with big size rather than the large no. of files with small size. Due to this, a no. of Metadata increases

Scanned by TapScanner

Rannvijay Singh Rawat SAP- 1000011240
Roll no - 18010 9029

which will become an overhead to the NameNode, To solve this problem Sequence files are introduced in Hadoop.

Q - 5 - Explain relationship between DB2 and BigSql.

Ans - The relationship between DB2 and BigSql

- BigSql and DB2 have the same "DNA"
- "Native Tables" with full transactional support on the Head Node.
- Row Oriented, tradition DB2 tables. ~~in~~
- BLU Columnar, In memory tables (on HeadNode Only)
- Materialized Query tables.
- GET SNAPSHOP / snapshot table functions
- Row and Column security.
- Federation / Fluid Query
- Views
- Workload Manager.
- System Temporary Table Spaces to support sort overflows.
- SQL PL stored procedures and UDFs.