Partitioning and Bucketing in Hive

@void_io

Partitioning in Hive

Partitioning in Hive offers a way of segregating hive table data into multiple directories. Partitioning gives effective results when:

- There are limited number of partitions
- Comparatively equal sized partitions

- Syntax for Partition Table:
- ```
  CREATE TABLE table_name(column1 dataType,column2 dataType)
  PARTITIONED BY(column3 dataType);
  ```
  ```

Static Partition in Hive

Static Partition in Hive

- Insert input data files individually into a partition table is Static Partition
- Usually when loading files (big files) into Hive Tables static partitions are preferred
- Static Partition saves your time in loading data compared to dynamic partition
- You “statically” add a partition in table and move the file into the partition of the table.
- We can alter the partition in static partition

```
LOAD DATA [LOCAL] INPATH ['path_name'] OVERWRITE INTO TABLE
[table_name] PARTITION (partition_column='value');
```

Here we have to give the partition column **value** explicitly whenever we want to create new partition

Dynamic Partition in Hive

Dynamic Partition in Hive

- Single insert to partition table is known as dynamic partition
- Usually dynamic partition load the data from non partitioned table
- Dynamic Partition takes more time in loading data compared to static partition
- When you have large data stored in a table then Dynamic partition is suitable.
- If you want to partition number of column but you don't know how many columns then also dynamic partition is suitable

```
INSERT OVERWRITE TABLE [table_name] PARTITION (partition_column)
SELECT * from reference_table;;
```

* We have to set 2 properties for loading data in dynamic partition

```
SET hive.exec.dynamic.partition= true;
SET hive.exec.dynamic.partition.mode= nonstrict;
```

Bucketing in Hive

Bucketing is a method to evenly distributed the data across many files. Create multiple buckets and then place each record into one of the buckets based on some logic mostly some hashing algorithm.

```
[Hash(column(s))] MOD [Number of buckets]
```

Syntax for Bucketing:

```
CREATE TABLE table_name(column1 dataType,column2 dataType)
PARTITIONED BY(column3 dataType) CLUSTERED BY (columnname) INTO 32
BUCKETS;
```

* We have to set 1 property for loading data in bucketing

```
SET hive.enforce.bucketing=true;
```

Thank you ☺

@void_io