**Bucketting**

In order to Increase the performance of queries Partitions are introduced in hive.So if there is a huge dataset regarding “world population” and suppose if we want to filter data by each country using where hive has to scan the entire dataset.**To ensure faster querying Partioning is made on country name** and

**Now for each country a directory will be created on hive/warehouse and the querying can be faster**

**Limitation with hive partitions:**

**No1:**

**If the dataset is so large and now for each country a directory will be created which will cause a increased overload on namenode**

**No2:**

**Now think of the above example where partitions are nade on country.Since population varies the Data say 100 GB for say 100 countries will not be equal.So again the processing on these partition will increase time if we use a group by like operation So to encounter these issues hive provides BUCKETTING**

**First problem is encountered as bucketing creates this much number of buckets so whatever may be the size the entire data will be divided among these buckets**

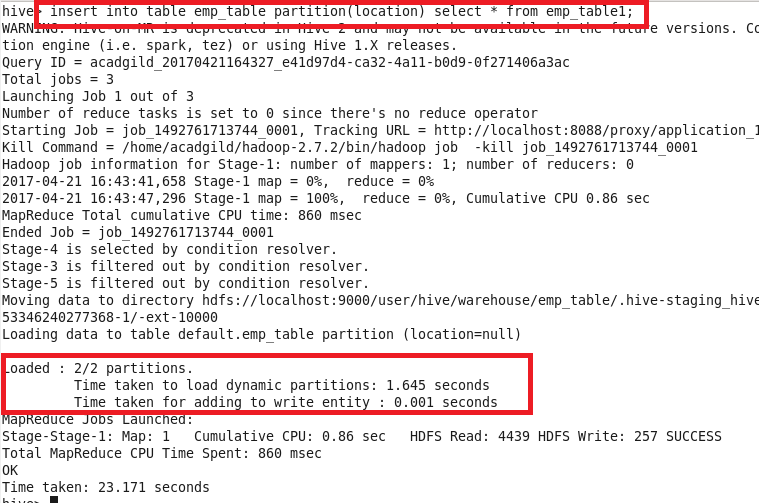
**2 Problem is encountered by since hashcode is code in the range of 1 to 10 say 10000 records all these data will be divided within these equally as 1000**

**Similarly If we want to further classify partitioned data,bucketing can be made over partitioned data and the bucketed record will be stored as files within the directory**

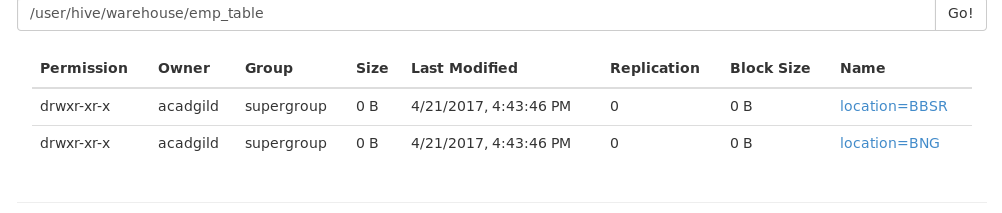
**So from above scenario we can define bucketing as**

**Hive partition divides table into number of partitions and these partitions can be further subdivided into more manageable parts known as Buckets or Clusters. The Bucketing concept is based on Hash function, which depends on the type of the bucketing column. Records which are bucketed by the same column will always be saved in the same bucket.**

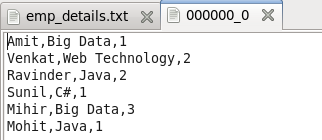
**In this example as you can see after inserting into table emp\_table 2 separate directories will be created**

****

**When we go inside the directory the files will be of varying size as shown**

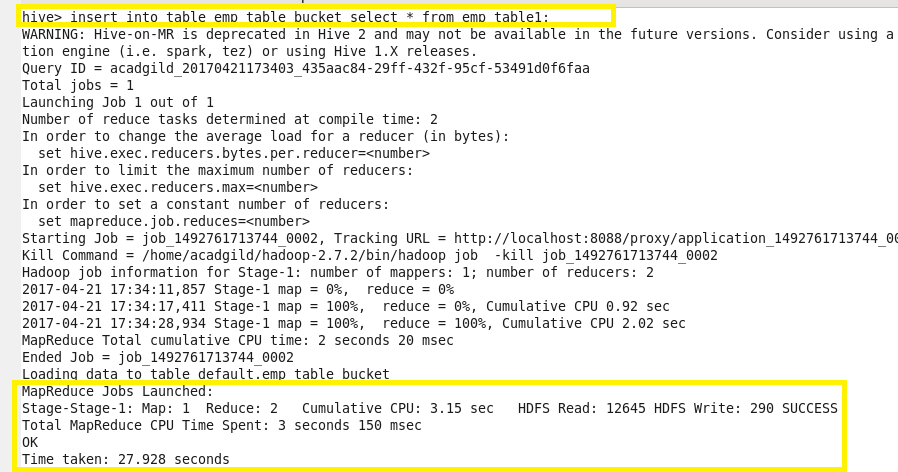
****

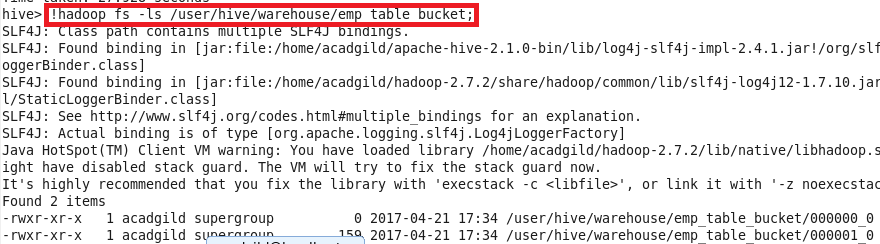
****

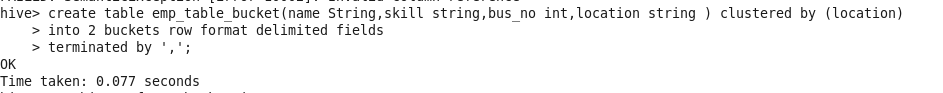
****

****

**But if we use bucketing and specify the bucketing as 2**

****

** we can see that files will be created and no additional directories will not be created**

****

**Bucketting Vs Partitioning**

|  |  |
| --- | --- |
| **BUCKETTING** | **PARTITIONING** |
| bucketing helps in organizing data in each partition into multiple files, so that the same set of data is always written in same bucket. | Partitioning helps in elimination of data, if used in WHERE clause |
| **Divides the records into file based on hashcode** | **Divides the records among directories based on key specified** |
| **It results in creation of files** | **It results in creation of Directories** |
| **It does not result in memory overhead of namenode as however large the record may be only specified files will be created**  **Say if we define 4 buckets only 4 files will be created** | **It will result in memory overhead as it depends on the type of Data** |
| **Equal distribution of data** | **Unequal distribution of Data** |
| **Since the data files are equal sized parts, map-side joins will be faster on the bucketed tables and sampling will also be faster.** |  |

**Sampling:**

**Sampling from the meaning means taking sample out of whole records since running querries on the entire record takes a larger time**

**There are 2 types of sampling in hive**

**1.Bucket Sampling**

**2.Block sampling**

**Bucket Sampling**

**In this sample data is taken from the bucketed column**

**For example if we have 64 buckets (3 out of 64 ) means 3 bucket is taken for sampling from 64 buckets created**

**Similary if we have 32 buckets and if we give the same (3 out of 64) would pick out half of the 3rd cluster as each bucket would be composed of (32/64)=1/2 of a cluster.**

**If we have 64 buckets and give (3 out of 32) 3 bucket (from first 32) and 35 bucket ( 3 bucket in second 32) will be taken for sampling**

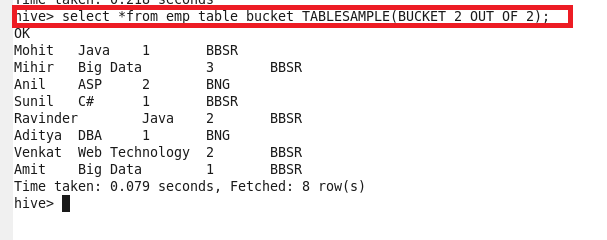
**2.Block sampling**

**Here sampling will be made on hdfs blocks say sampling (1 percent) will take 1 percent of the block size for sampling(i.e (1/100)\*128 )=1.28 Mb will be taken for sampling**

**Similarly we can directly specify as (90 MB)**

**Sampling example:**

**Now 2nd bucket is displayed**

****