Explain the working of Partitioning in brief

When any user wants data contained within a table to be split across multiple sections in hive table, use of partition is suggested.

Hive organizes tables into partitions. It is a way of dividing a table into related parts based on the values of partitioned columns such as date, city, and department. Using partition, it is easy to query a portion of the data.

The entries for the various columns of dataset are segregated and stored in their respective partition. When we write the query to fetch the values from table, only the required partitions of the table are queried, which reduces the time taken by query to yield the result.

**Example:**

If any user wants to do some analysis on purchasing patterns of customer in some specific date range and if his query has to run on entire dataset then efficiency will be low in this case.

But running the hive queries on data with in specified date range will increase the efficiency when compared to running the data on entire dataset as the query will run on certain chunk of dataset, so time taken to retrieve the result will be less.

Explain the difference between Static and Dynamic Partitioning in Hive with an example.

Static Partitioning:

* Insert input data files individually into a partition table is Static Partition
* Usually when loading files (big files) into [Hive tables](http://www.hadooptpoint.com/hive-create-table-examples/) 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
* You can get the partition column value form the filename, day of date etc. without reading the whole big file.
* If you want to use Static partition in hive you should set property set hive.mapred.mode = strict  This property set by default in hive-site.xml
* Static partition is in Strict Mode
* You should use where clause to use limit in static partition
* You can perform Static partition on Hive Manage table or external table.

Dynamic Partitioning:

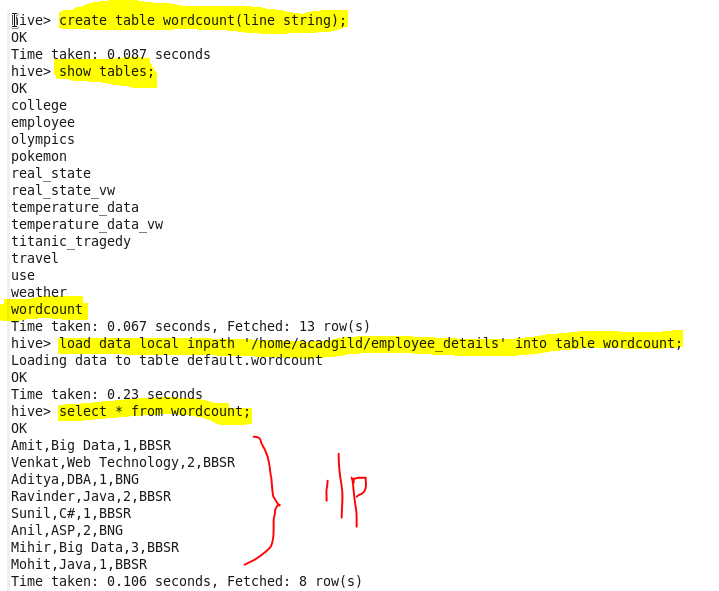
* single insert to partition table is known as dynamic partition
* Usually dynamic partition load the data from non partitioned table
* Dynamic Partition takes more time in loading data compared to static partition
* When you have large data stored in a table then Dynamic partition is suitable.
* If you want to partition number of column but you don’t know how many columns then also dynamic partition is suitable
* Dynamic partition there is no required where clause to use limit.
* we can’t perform alter on Dynamic partition
* You can perform dynamic partition on hive external table and managed table
* If you want to use Dynamic partition in hive then mode is in nonstrict mode
* Here is hive dynamic partition properties you should allow

Example:

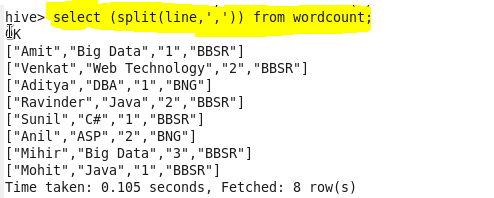
For example, let’s say you have a huge dataset accumulated over many years. You might be concerned about query performance against the Hive table created for this dataset. You could use one or more columns to partition the table created for this underlying data. One possibility would be to partition the data by year. If you already know that the data is nicely segregated by year before it’s loaded into Hive, static partitioning can be used. However, imagine a situation where the year values are not known in advance of loading the data, and you want to avoid the pain of digging into the data to extract year values. In this case, dynamic partitioning is the best approach.

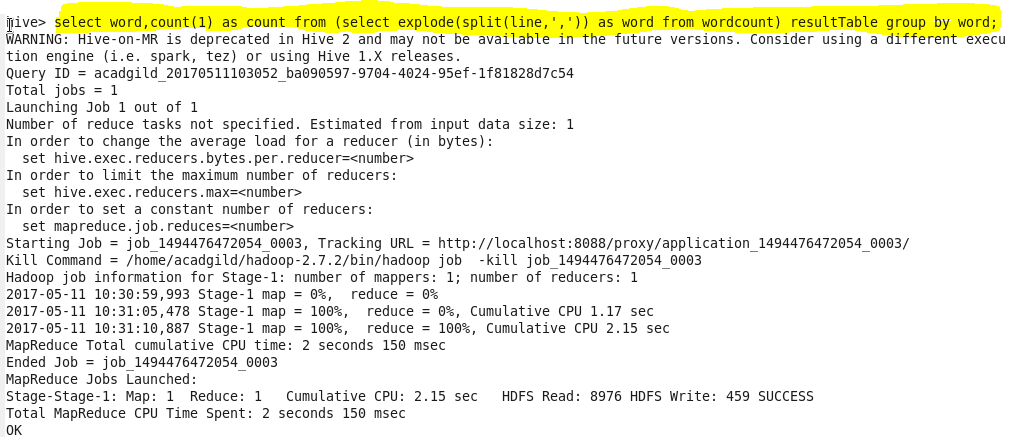
Perform word count in Hive for above given dataset.

Creating the table:



Code:





Output: