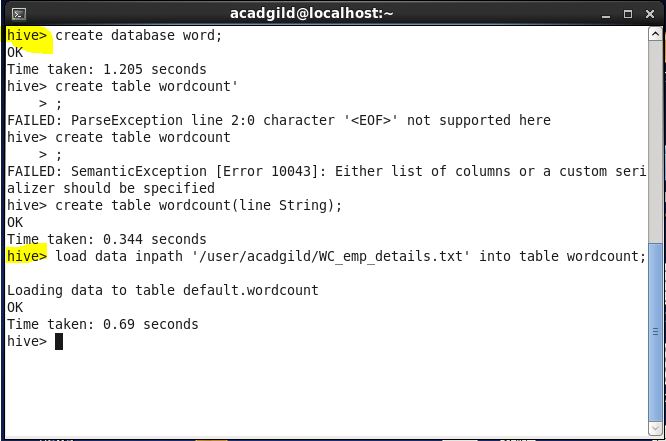
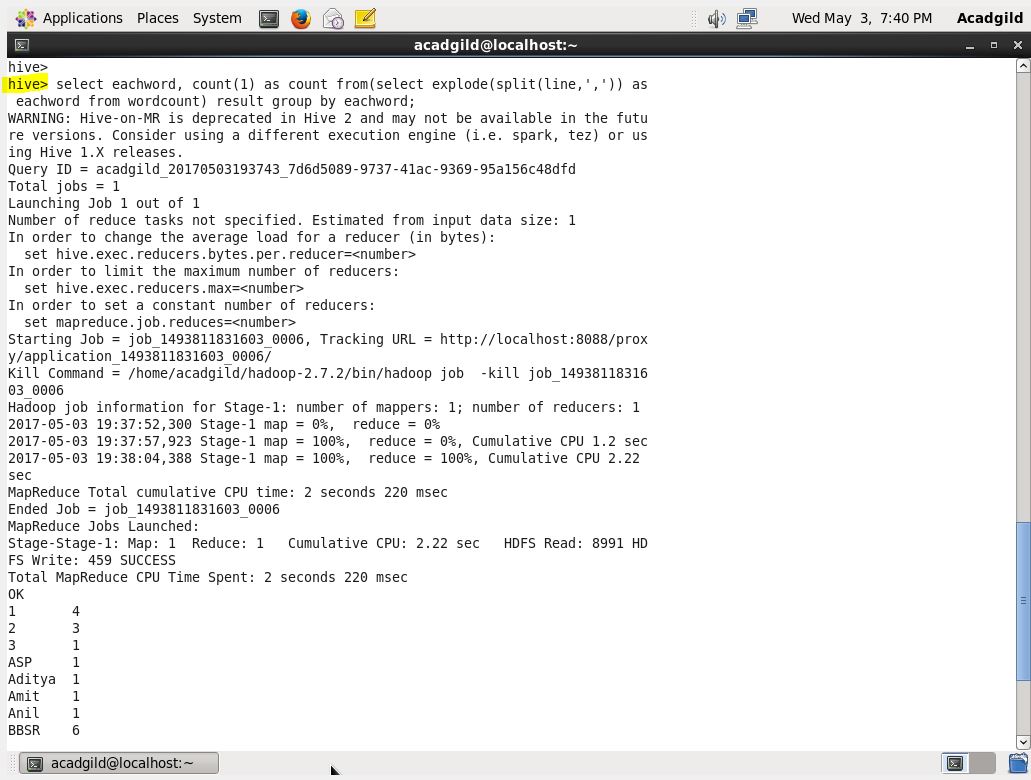
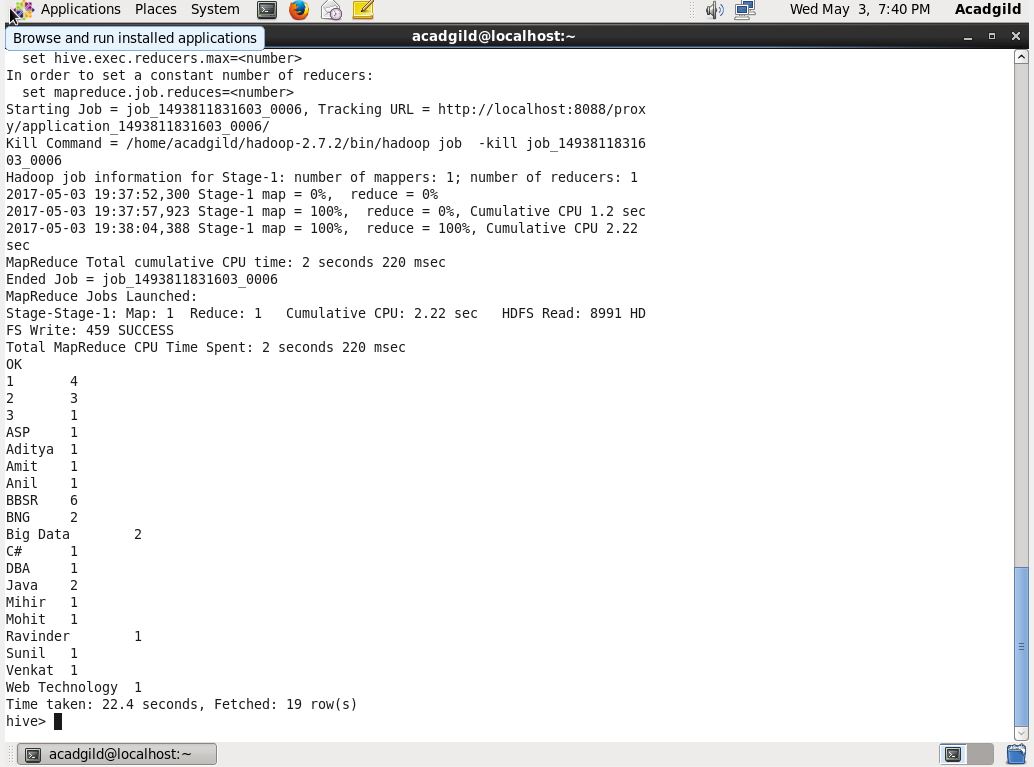
● Perform word count in Hive for above given dataset.here in this we have created a database and a table to execute the program for the word count. We have used a command explode and split and the word count is performed by splitting each line into single word and then splitting and storing as an array of words and counting is done as shown below:







1. Explain the working of Partitioning in brief.

Hive has been one of the preferred tool for performing queries on large datasets, especially when full table scan is done on the datasets. In the case of tables which are not partitioned, all the files in a table’s data directory is read and then filters are applied on it as a subsequent phase. This becomes a slow and expensive affair especially in cases of large tables. Without partitioning Hive reads all the data in the directory and applies the query filters on it. This is slow and expensive since all data has to be read. Very often users need to filter the data on specific column values. To apply the partitioning in hive , users need to understand the domain of the data on which they are doing analysis. With this knowledge, identification of the frequently queried or accessesd columns becomes easy and then partitioning feature of Hive can be applied on the selected columns.

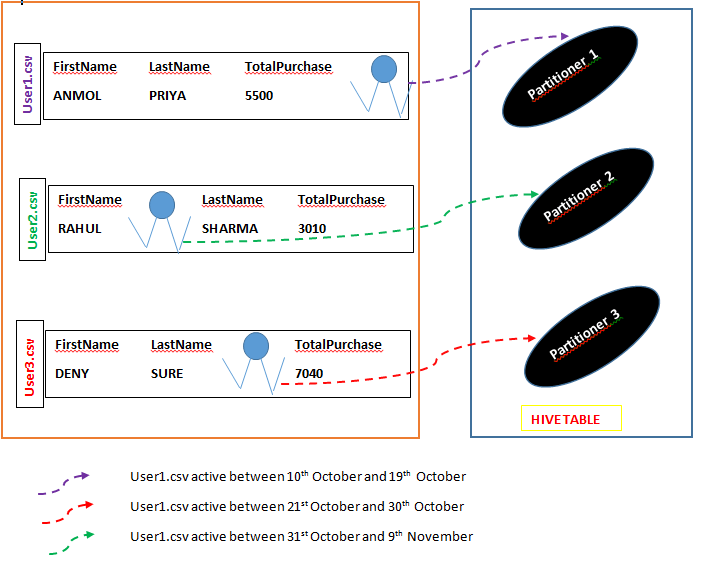
**A scenario for partitioning:**

Let’s take a scenario:

* Data is present in hdfs coming in from various ecommerce companies.
* We need to run the HiveQl queries on user buying pattern.
* We need to analyse data coming in from last 10 days.

In the above scenario instead of running the queries which involves scanning of entire table, an approach should be followed where query runs on only last 10 days of data.

**Partitioner Example**

[](https://i0.wp.com/acadgild.com/blog/wp-content/uploads/2015/11/pic1.png)

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.

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

1.Creating input files for partitioning:

Let’s take two input file:

User1:

User2:

2.Copying the input files:

The above two input files need to be copied into table further subdivided into partitions divided on the basis of country and state.

User1 data needs to be copied into the section of table having partitions as country = us and state = fl.

User2 data needs to be copied into the section of table having partitions as country = ca and state = au.

3.Retrieving the user information:

If anyone wants to retrieve the user information belonging to partition with country us and state fl, query needs to be written as shown below.

Hive will retrieve the data from the section with data having partition with country ‘us’ and state as ‘fl’ and earlier way of searching the entire table for one query will be avoided.

Similarly if the user information from country ‘ca’ and state ‘au’ has to be retrieved than we need to fire the below query.

4.Browsing the HDFS:

This section describes viewing of partitions in hive warehouse directory.

The table par\_user has been subdivided into two partitions with two different directory.

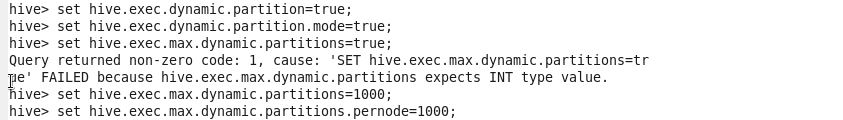
Steps for Dynamic partitioning:

1. Creating tables:

We need to create the partitioned table par\_user as shown below.

This table will be populated with the contents from table user1 and partitions will be created based on country and state.

1. Create input file for dynamic partitioning:
2. Loading input file into user1 table:
3. Setting of parameters for dynamic partitioning: To use the dynamic partitioning in hive we need to set the below parameters in hive shell or in hive-site.xml file. In this case we have set the below properties in hive shell.



1. **Retrieving data from partitioned table:**
2. We need to copy the file from user1 to partitioned table par\_user and then retriving the data from it all together using insert and select statement in one hive statement.

