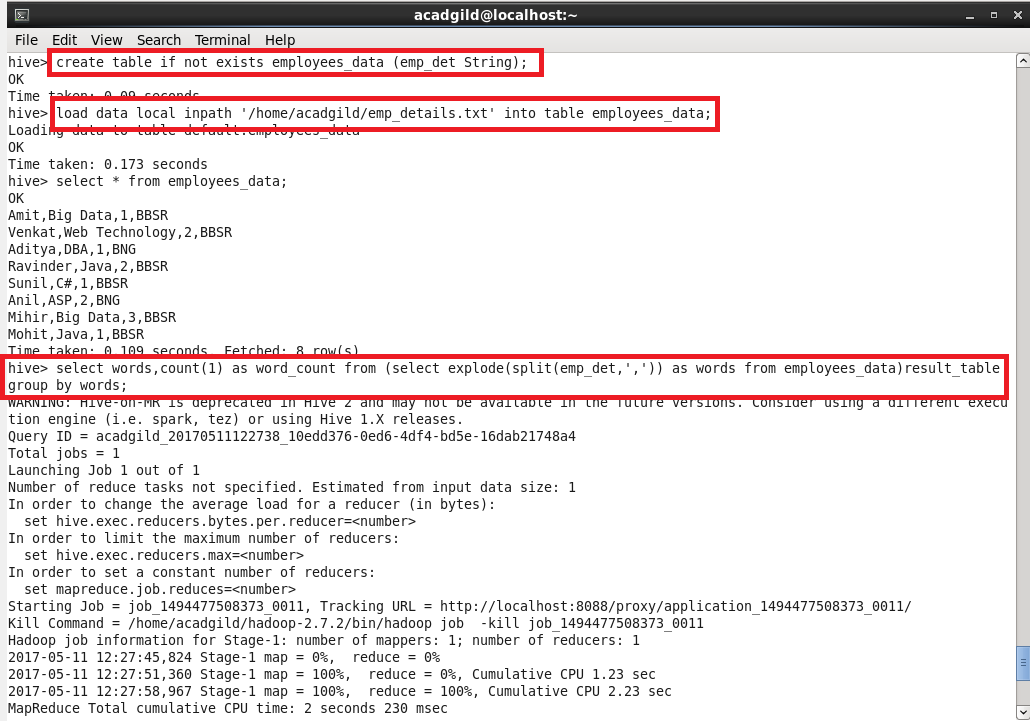
**Session 26**

**Assignment 1**

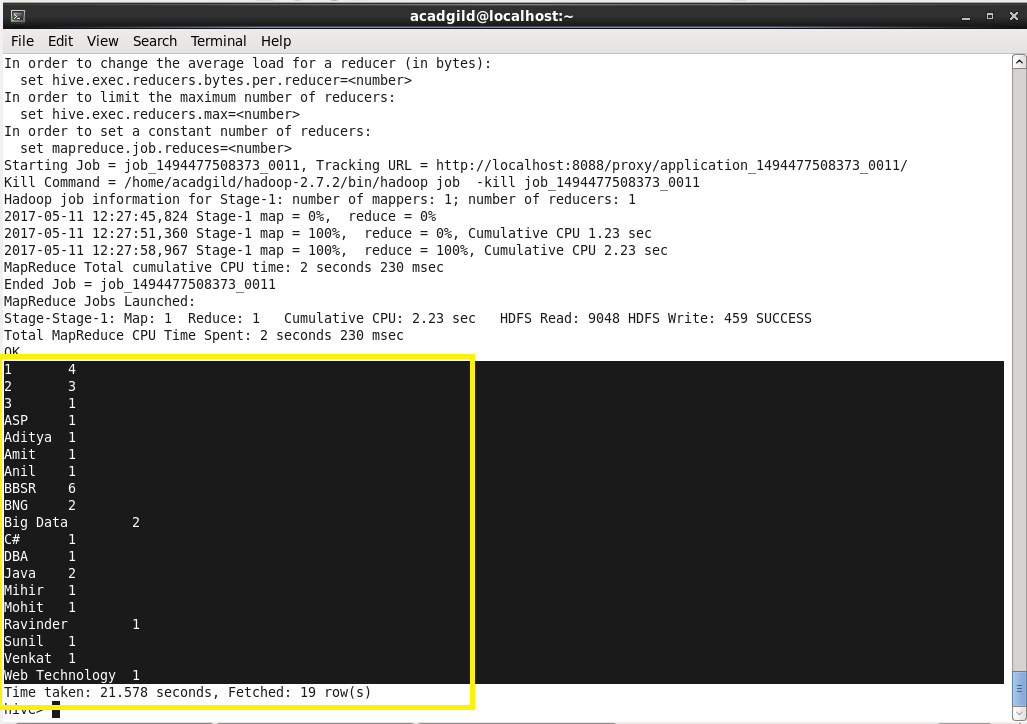
**Part 1:**

**Problem Statement:**

Perform word count in Hive for above given dataset.



**Output:**



**Part 2:**

**Problem Statement:**

Explain the working of Partitioning in brief.

Partitioning of table means dividing data in the table into some parts based on the values of particular columns like city or country, segregate the input records into different files/directories based on city or country.

Partitioning can be done based on more than column which will impose multi\_dimensional structure on directory storage. For example, in addition to partitioning log records by city column, we can also subdivide each record into country wise, so separate files will be including country column into partitioning.

There are two types in Partitioning. They are,

1) Static partitioning

2) Dynamic partitioning

**1. Static Partitioning:**

In this type the input data should contain the columns listed only in table definition but not the columns defined in partitioned by clause.

If our input column layout is given according to the expected layout and we already have separate input files for each partitioned key value pairs, like one separate file for each combination of city and country values and these files can be easily loaded into partitioned table

**2. Dynamic Partitioning:**

Instead of loading each partition with single SQL statements, which will result in writing lot of SQL statements for huge number of partition. Where HIVE supports dynamic partitioning and with that we can add any number of partitions with single SQL execution.

Hive will automatically splits our data into separate partition files based on the values of partition keys present in the input files.

**Part 3:**

**Problem Statement:**

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

1. S**tatic 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 time in loading data compared to dynamic partition when “statically” add a partition in table and move the file into the partition of the table.
* The partition in static partition can be altered as the user wants.
* The partition column value form the filename, day of date etc. can be obtained from the data without reading the whole big file. If you want to use Static partition in hive you should set property.
* Set hive.mapred.mode = strict as that property will set by default in hive-site.xml. Static partition is in Strict Mode. The user should use where clause to use limit in static partition and also can perform Static partition on Hive Manage table or external table.

1. **Dynamic Partition in Hive**

* Single insert to partition table is known as Dynamic Partition.
* Usually dynamic partition will load the data from non-partitioned tables.
* Dynamic Partition takes more time in loading data compared to static partition
* When you have large non-partitioned 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 where alter can’t be performed on Dynamic partition
* The dynamic partition can be performed on hive external table and managed table.
* If Dynamic partition wants to be used in hive then mode should be in no strict mode.