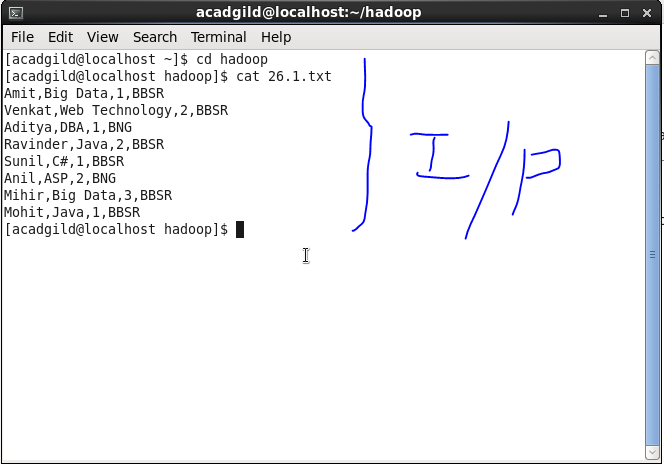
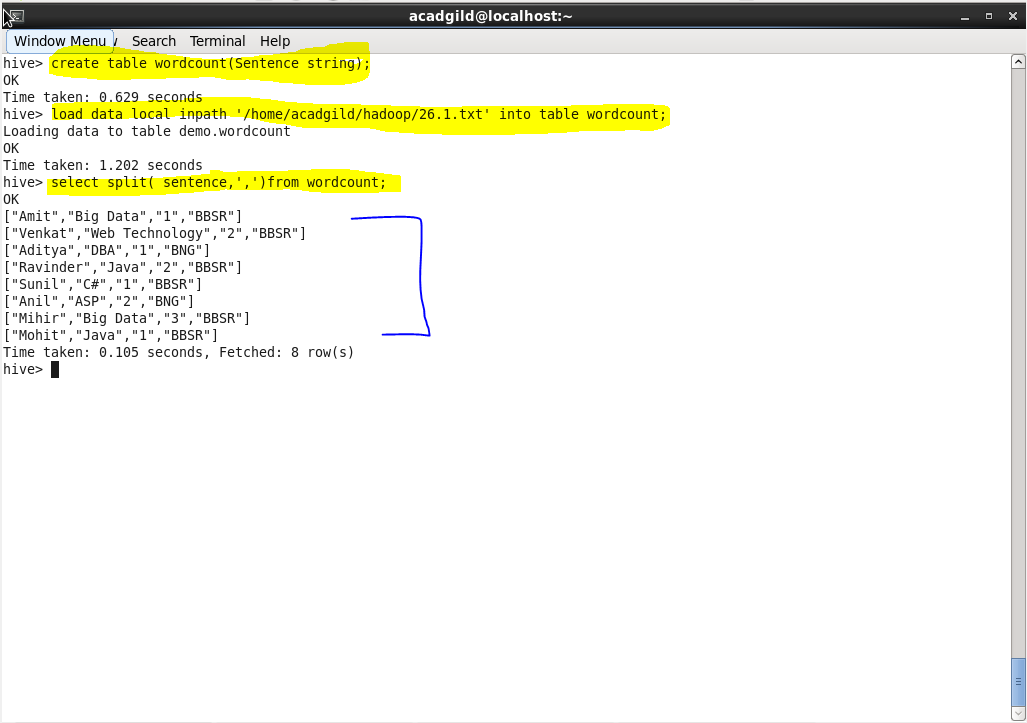
**Problem Statement:**

**-Perform word count in hive.**

**Input File-**

****

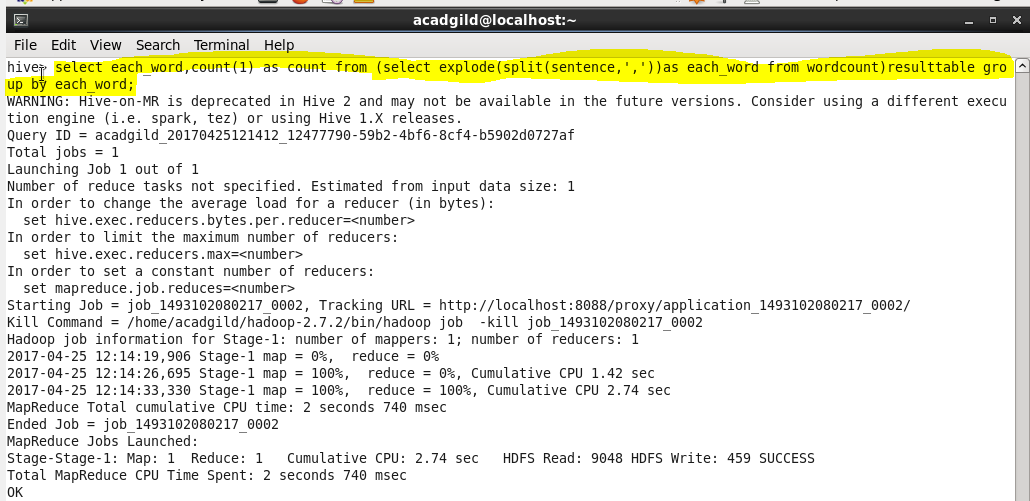
1. **Word Count-**
2. We have created a table named wordcount with one field sentence.
3. After creating table we have loaded data in the table.
4. After that split command is used to split each word from the table and separated it by comma(,).

****

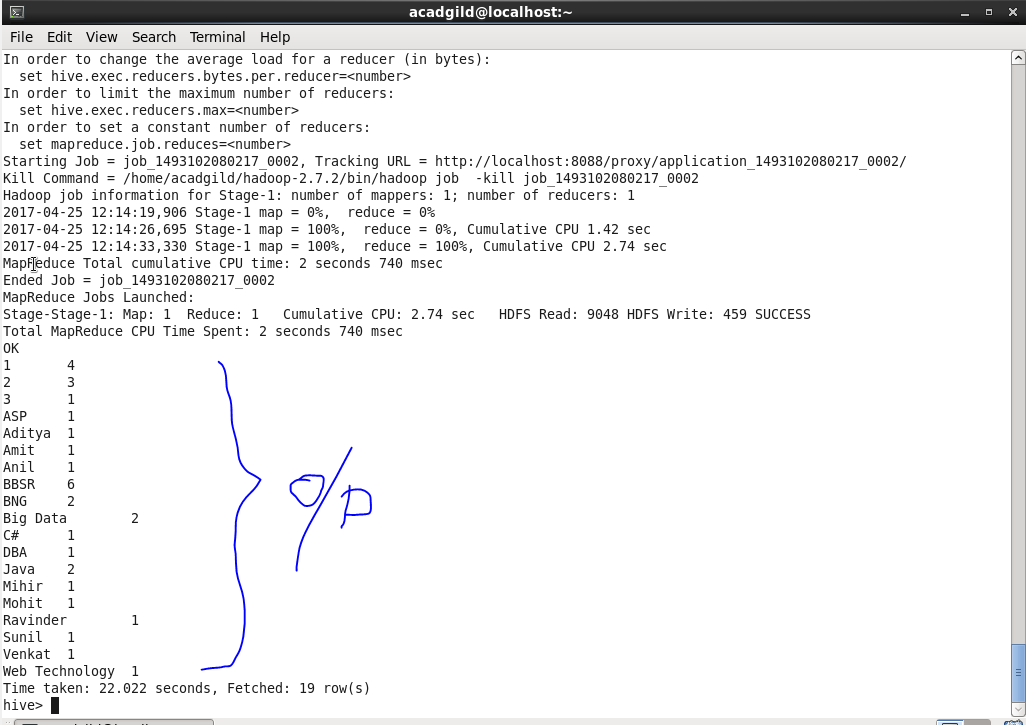
1. We used explode command, explode takes an array as input and givers output in the form of map as separate rows for each word.

****

1. We used count and group by command to count each word from each row and group them by each word to get the count of each word.

****

**Output-** By following the steps we get the output as-

****

1. **Explain the working of Partitioning in brief.**

Partitioning is the process of dividing data stored in tables into some parts based on values of particular columns.

* Hive organizes tables horizontally into partitions.
* It is a way of dividing a table into related parts based on the values of partitioned columns such as date, city, department etc.
* Using partition, it is easy to query a portion of the data.
* Partitioning can be done based on more than column which will impose multi-dimensional structure on directory storage.
* In Hive, partitioning is supported for both managed and external tables.
* The partition statement lets Hive alter the way it manages the underlying structures of the table’s data directory.
* In case of partitioned tables, subdirectories are created under the table’s data directory for each unique value of a partition column.
* When a partitioned table is queried with one or both partition columns in criteria or in the WHERE clause, what Hive effectively does is partition elimination by scanning only those data directories that are needed.
* If no partitioned columns are used, then all the directories are scanned (full table scan) and partitioning will not have any effect.

Advantages-

* It is used to distribute execution load.
* The data is stored in parts, so query response time is faster to process the small parts of data instead looking for entire data.

Limitations-

* Too many partitions in a table creates large number of files and directories in HDFS, which is an overhead to Name Node.
* Partitions may optimize some queries by where clause but may be less responsive or clauses like group by.

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

Classification of partitioning

• Static partitioning

• Dynamic Partitioning

1. **Static Partitioning**

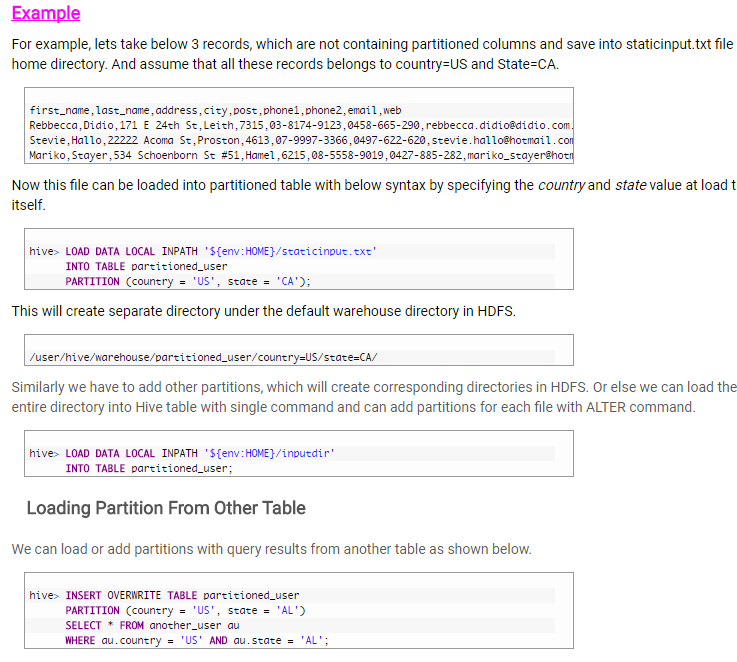
* Inserting input data individually into a partition table is called static partitioning.
* Static partitioning needs to be applied when we know data (supposed to be inserted) belongs to which partition.
* Static partitioning is used generally for big files.
* We manually add a partition in table and move the file into that partition.
* We can even alter that partition.
* For using static partition we have to set the flag in hive-site.xml-

Set hive.mapred.mode=strict

1. **Dynamic Partitioning**

* Single insert to partition table is known as dynamic partition
* In static partitioning, every partitioning needs to be backed with individual hive statement which is not feasible for large number of partitions as it will require writing of lot of hive statements.
* In that scenario dynamic partitioning is suggested as we can create as many number of partitions with single hive statement.
* It takes more time to load data as compared to static partition.
* We cant perform alter in dynamic partition.

**Example of Static Partitioning-**

****

**Example of Dynamic Partitioning-**

