**Session 27**

**Assignment 6**

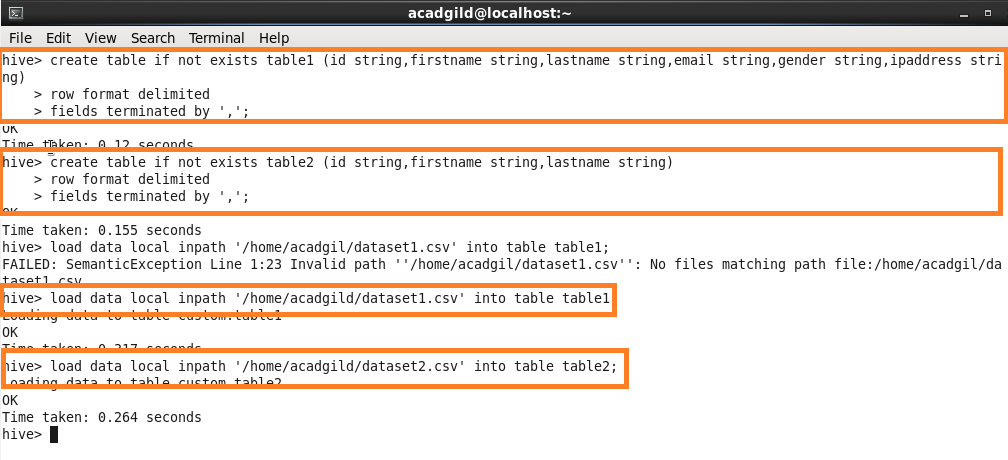
**Problem Statement:**

Using the below given two datasets you need to give a demo on the below joins in hive.

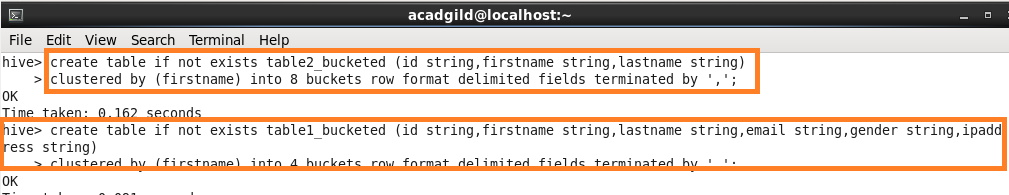
* Bucket Map join
* Sort-Merge Bucket join
* Sort-Merge Bucket Map join
* Left semi join

**Part 1:** Bucket Map Join

**Step 1:** Created 2 tables and loaded the data into the tables



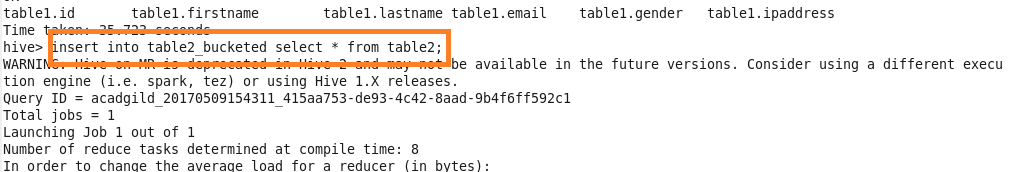
**Step 2:** To perform bucketing, we need to have bucketed tables.so I’m creating 2 bucketed tables. One with 4 buckets and one with 8 buckets.



**Step 3:** Inserting the data in table1 into the bucketed table table1\_bucketed



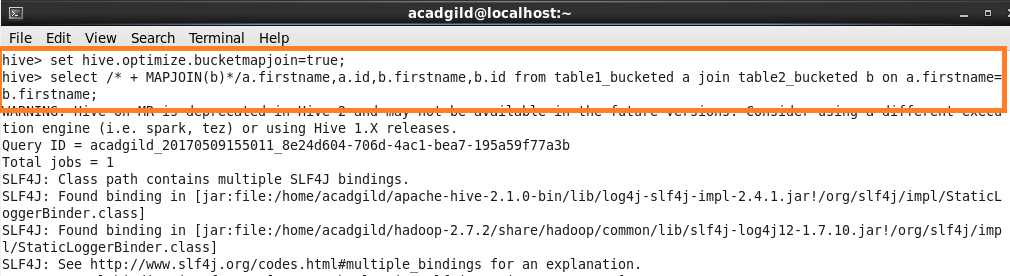
**Step 4:** Inserting the data in table2 into the bucketed table table2\_bucketed .



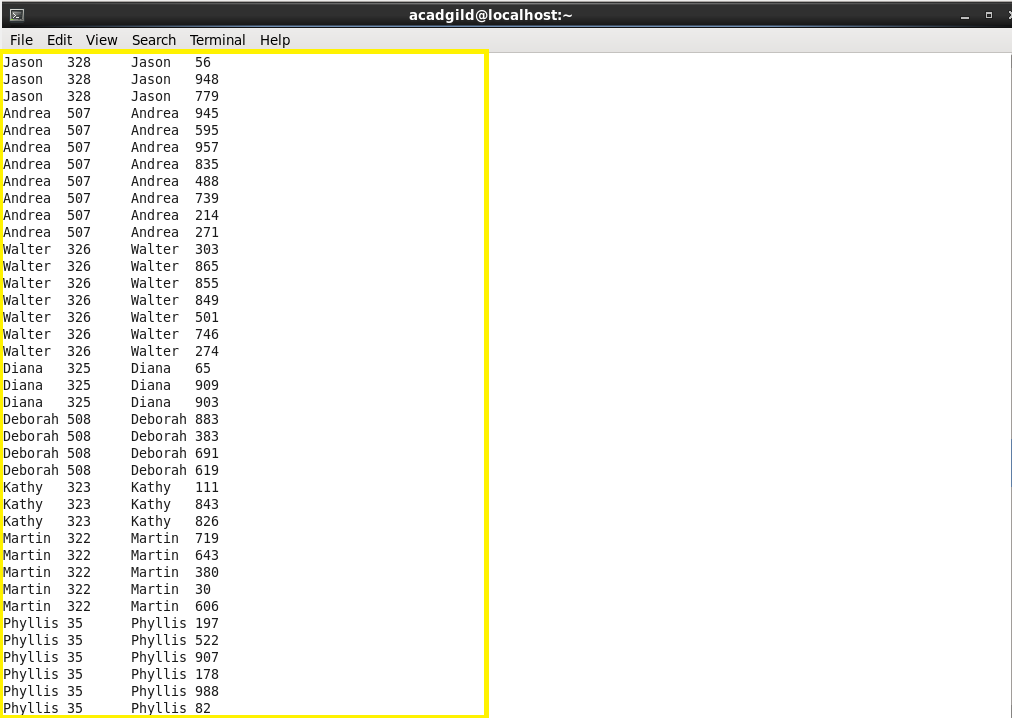
**Step 5:** For performing Bucket-Map join, we need to set this property in the Hive shell.

* set hive.optimize.bucketmapjoin = true

and writing query for extracting data.



**OUTPUT:**



**Part 2:** Sort-Merge Bucket Join

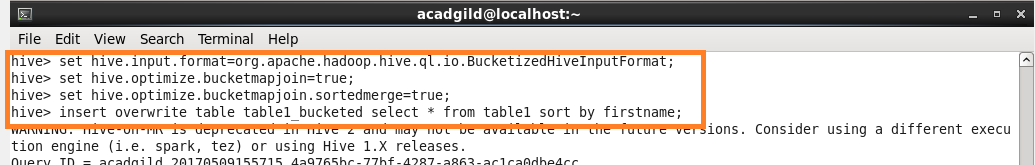
Sort-Merge bucket join is like reduce side join.

**Step 1:** Before joining we’ll have to sort the data and then only merge.

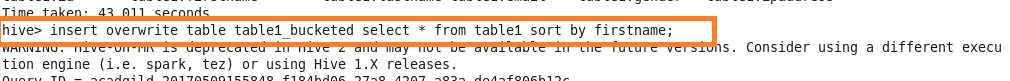
Moreover in this join we should set

* set hive.auto.convert.join= false;

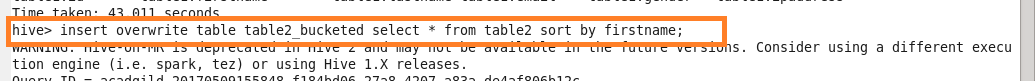
because we want reducer to run.



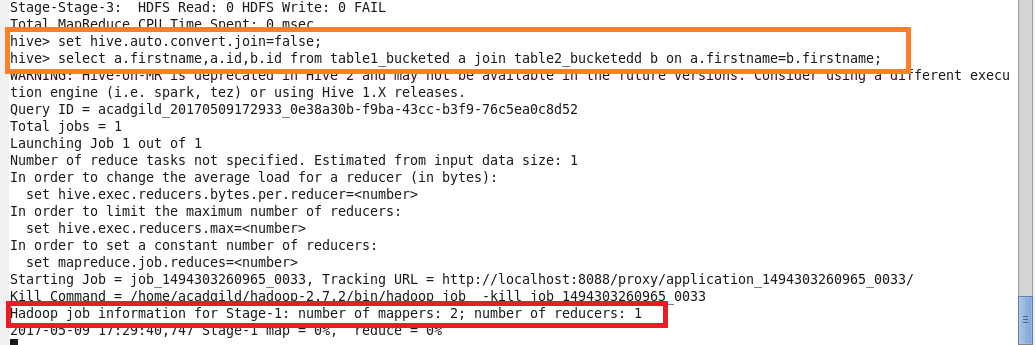
**Step 2:** Sorting the contents of bucketed table 1



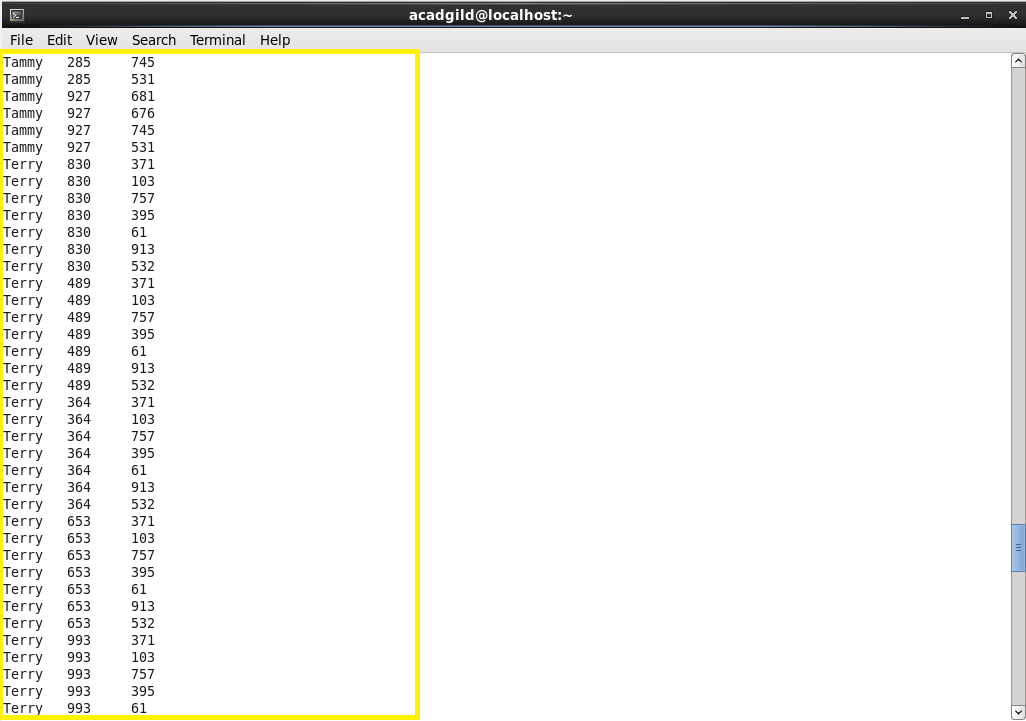
**Step 3:** Sorting the contents of bucketed table 2



**Step 4:** Writing query for extracting data



**Output:**

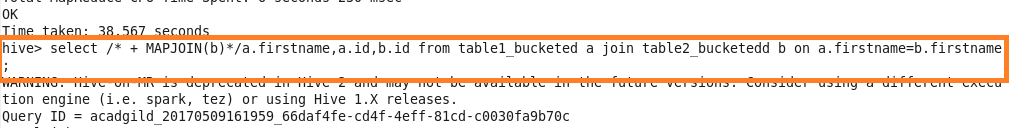


**Part 3:** Sort-Merge Bucket Map Join

**Step:** In sort-merge bucket map join we need to set

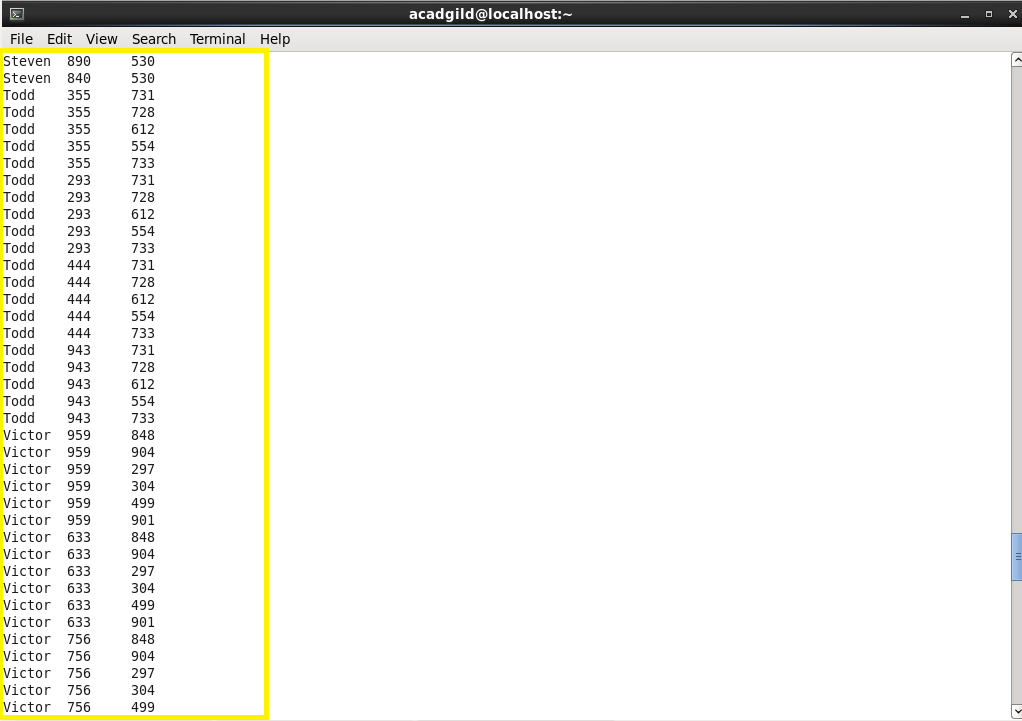
* set hive.auto.convert.join= true;

but without setting this also we can run a map only job by typing like this



This query runs a map only job. The same bucketed tables are used that were created previously for this join.

**Output:**



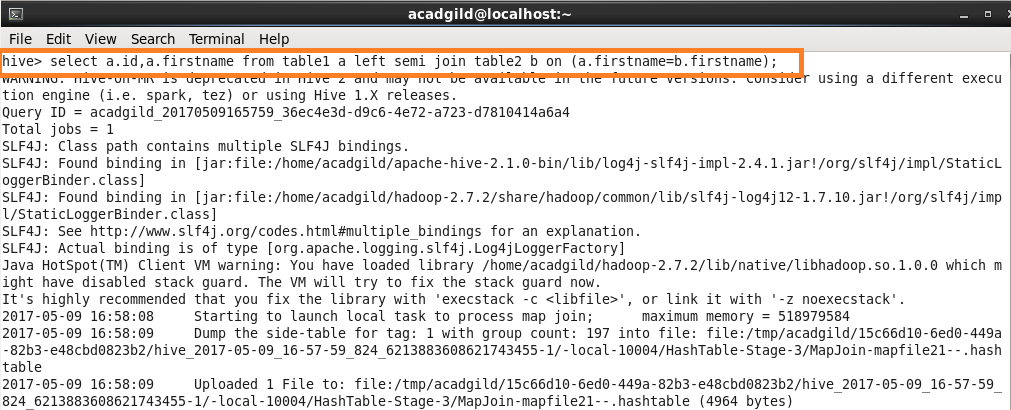
**Part 4:** Left semi join

Left semi join is more or less like sub query in SQL.

The left semi join is used in place of the IN/EXISTS sub-query in Hive. In a traditional RDBMS, the IN and EXISTS clauses are widely used whereas in Hive, the left semi join is used as a replacement of the same.

Left semi join eliminates the use of where and in keywords that we use in typical sub queries.

**Step:** The query used for this join is shown below.



**Output:**

