**Session 27**

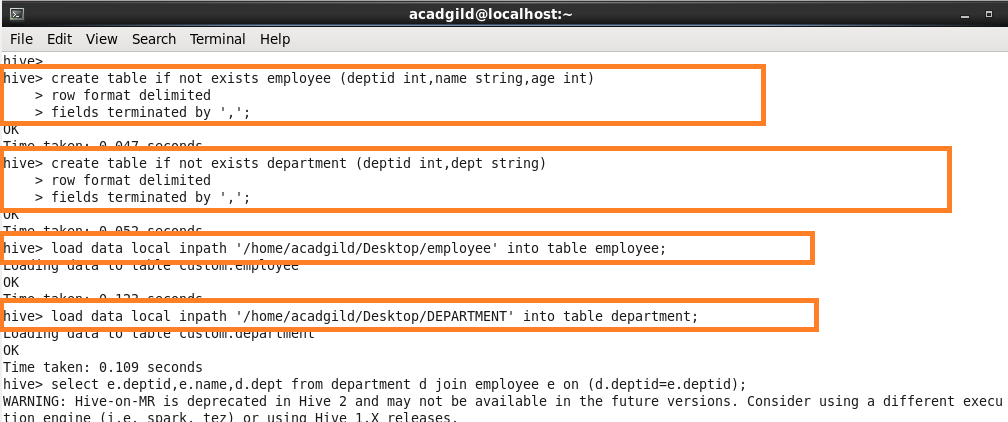
**Assignment 5**

**Problem Statement:**

Perform join optimizations in hive by following the steps in the below blog link

**Join table ordering (Largest table last)**.

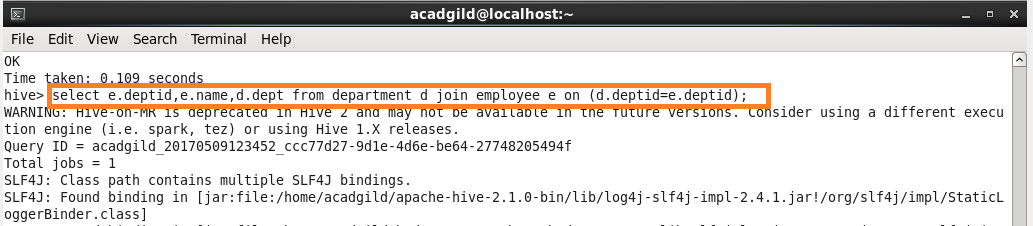
**Step 1:** Creating tables and loading data into the tables



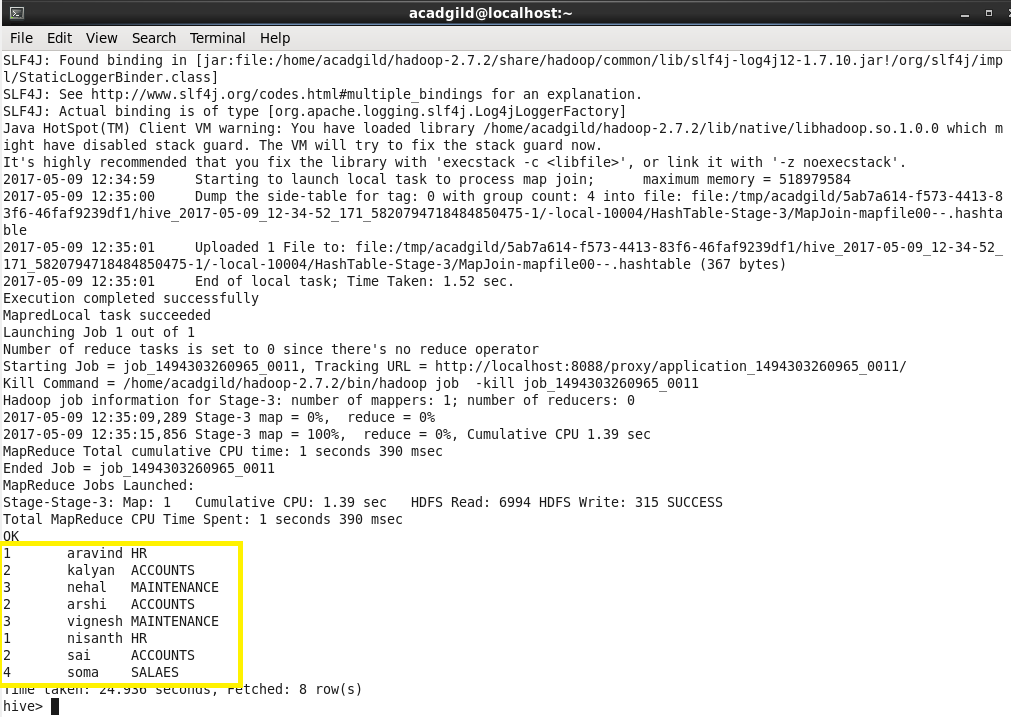
**Logic:**

* When Hive executes a join, it needs to select which table is streamed and which table is cached. Hive takes the last table in the JOIN statement for streaming, so we need to ensure that this streaming table is largest among the two.
* We have 2 tables. They are Employee and department. Employee table is the largest table and department is the smallest one. We need to set the largest table at the last for streaming purpose. Hence employee table is set at the last.
* This is one of the ways to optimize joins in hive

**Step 2:** Joining the tables department and employee on the basis of deptid.



**Output:**



**Step 3: Map Side Join.**

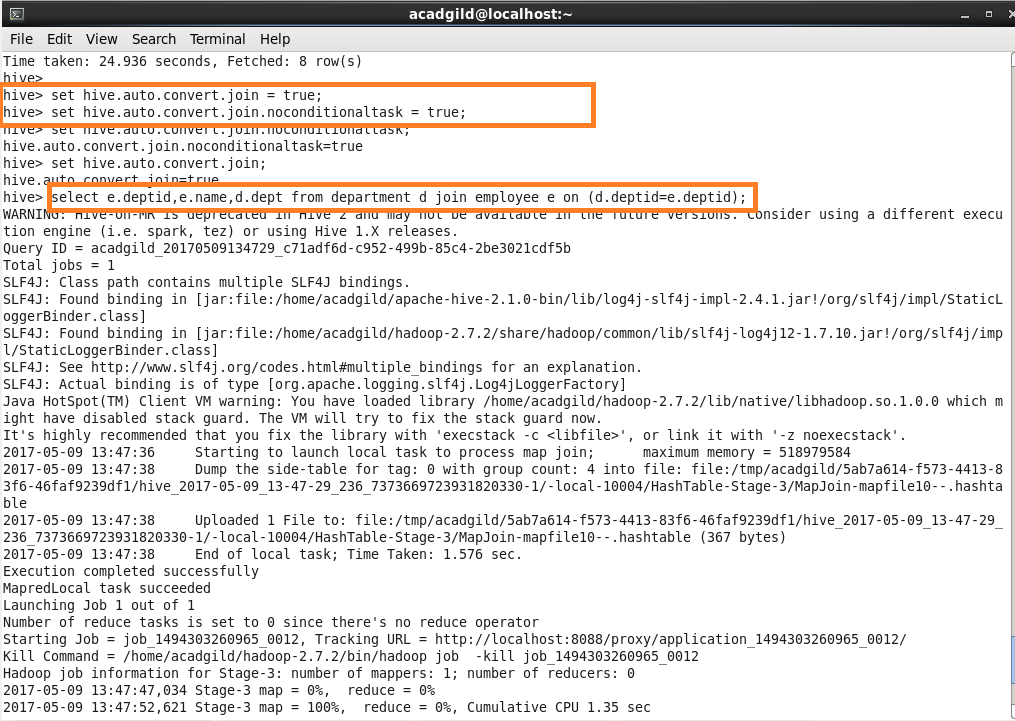
**In map side join we need to set this in shell**

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

**If we set this, automatically map only job gets started and runs.**

**If there is no reducer, obviously the time gets reduced.**

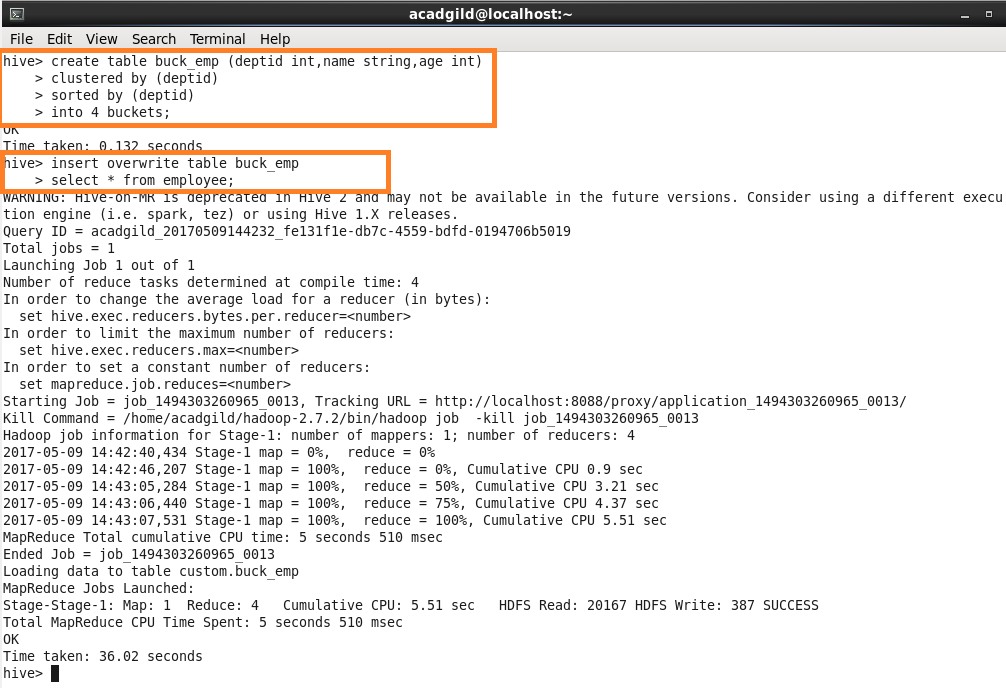
From the below pic you can see that there is no reduce operation.



**Step 4: Sort-Merge-Bucket (SMB) Map Join.**

**We cannot use normal tables for this join. We need to create a bucketed tables, sort the contents and then only we should join those tables.**

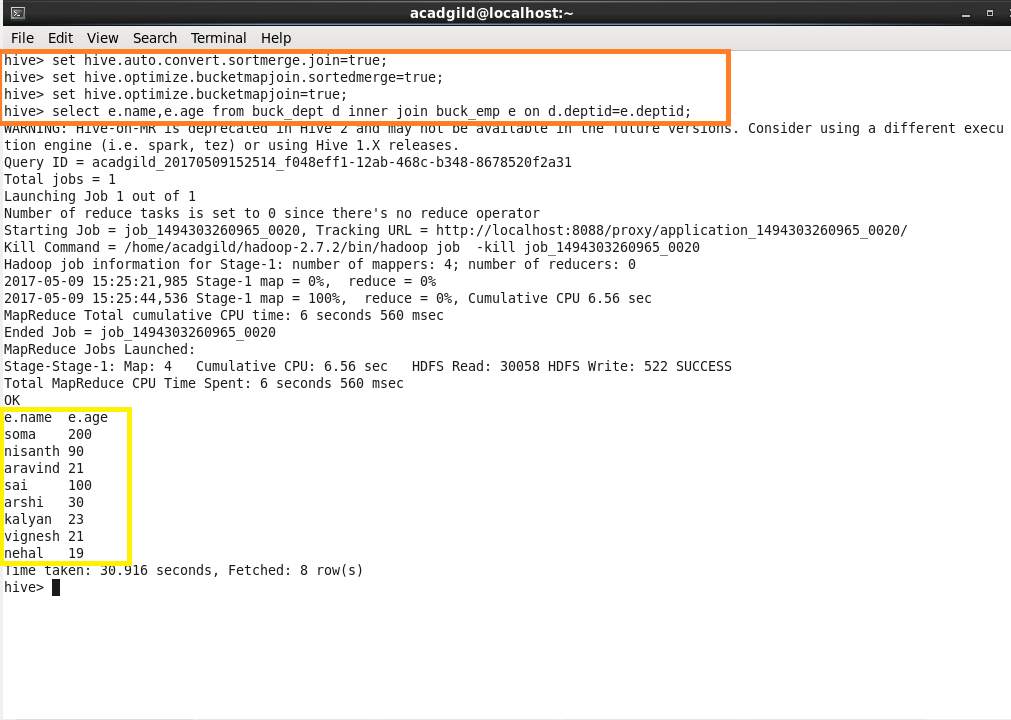
**So creating a bucketed table with 4 buckets, sorted it by deptid and loading data into it using insert command.**



**Creating another bucketed table with 4 buckets, sorted it by deptid and loading data into it using insert command.**



**Output:**



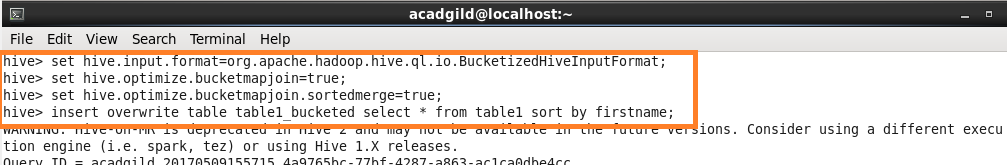
**Step 5:** SORT-MERGE BUCKET JOIN:

Sort-Merge bucket join is like reduce side join. Before joining we’ll have to sort the data and then only merge.

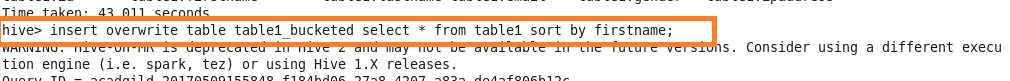
Moreover in this join we should set

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

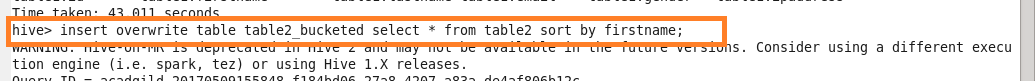
because we want reducer to run.



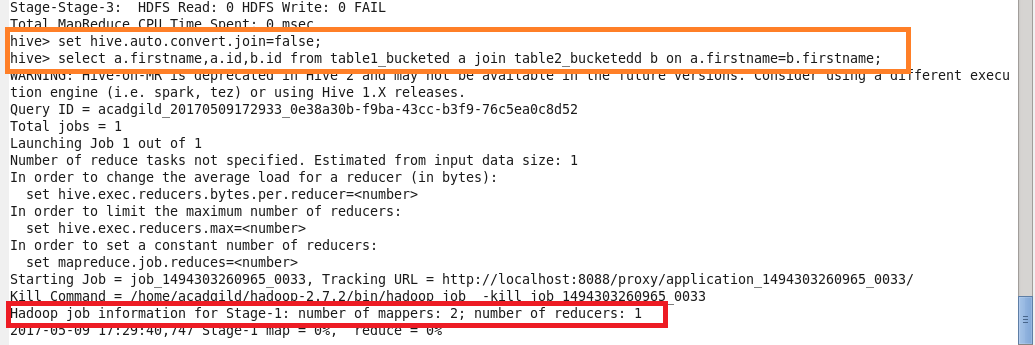
Sorting the contents of bucketed table 1



Sorting the contents of bucketed table 2



Writing query for extracting data



**Output:**

