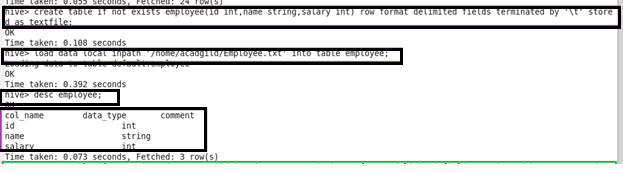
**ASSIGNMENT – 27.5**

1. Perform join optimizations in hive by following the steps in the below blog link <https://acadgild.com/blog/join-optimization-in-apache-hive/>

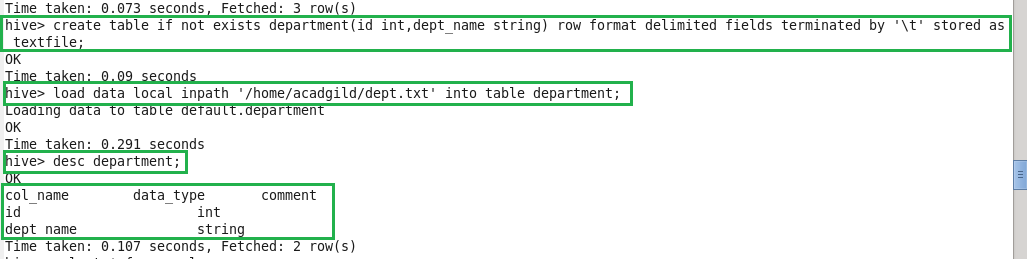
First, creating tables then loading the data into the created tables.

Then, Describing created tables :

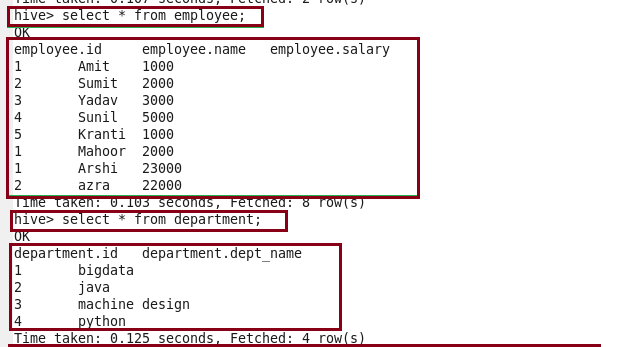
**Created table – employee :**



**Created table – department :**



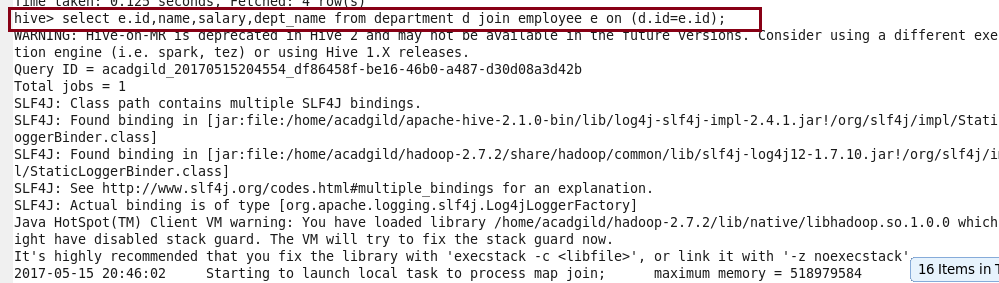
**Displaying the data of the created tables using ‘select’ keyword**

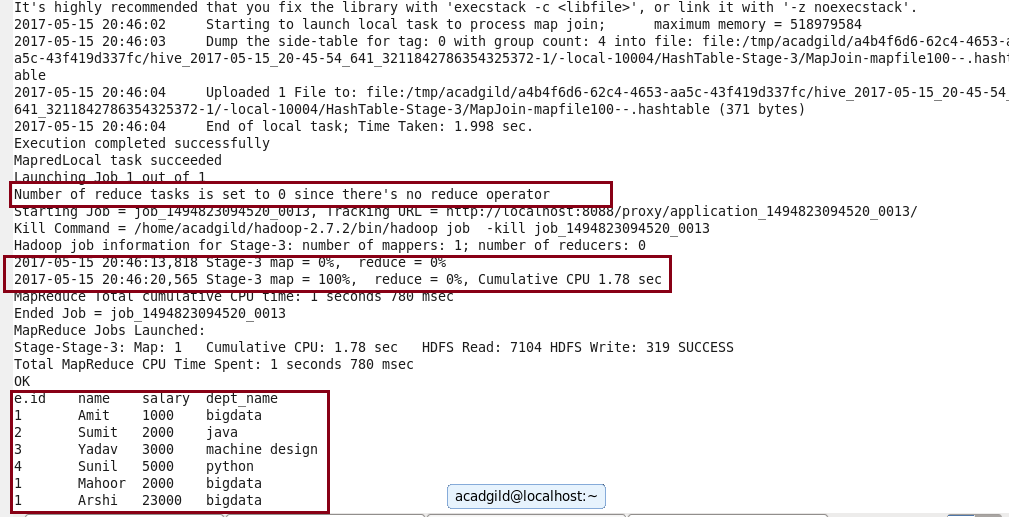


**Join table ordering (largest table last):**

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. Here, largest table is employee.

Or we can also explicitly tell Hive which table it should stream.



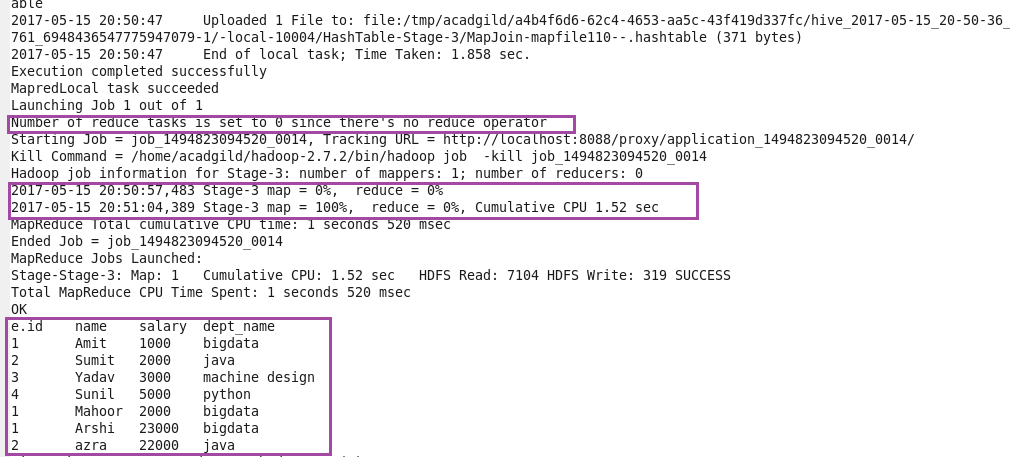
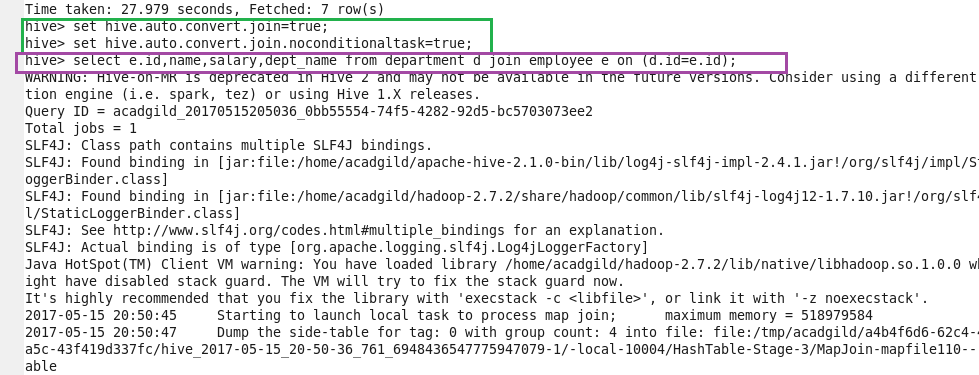


**Mapside join(join optimization) :**

To perform mapside join we need to set configurations:

Set hive.auto.convert.join=true;

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



See the highlighted part in purple color. You will find that there is no reducer phase performed in this join operation. Hence, the map-side join is faster than regular join operation.

**Another join method of join optimization in hive**

**Sort-Merge-Bucket (SMB) Map Join:**

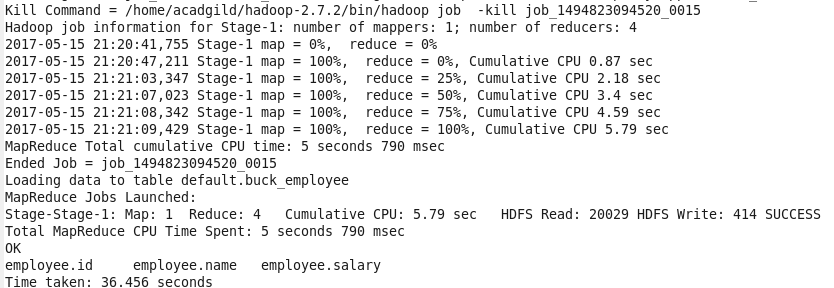
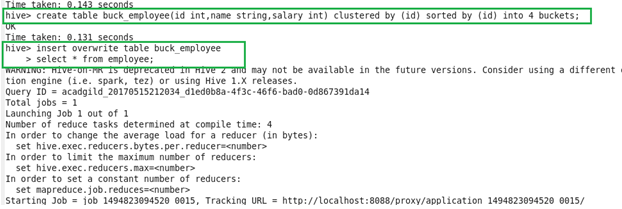
It is another Hive join optimization technique where all the tables need to be bucketed and sorted. In this case joins are very efficient because they require a simple merge of the presorted tables.

Creating bucketed tables from existing tables i.e, employee and department. Before that we need to set the configurations.

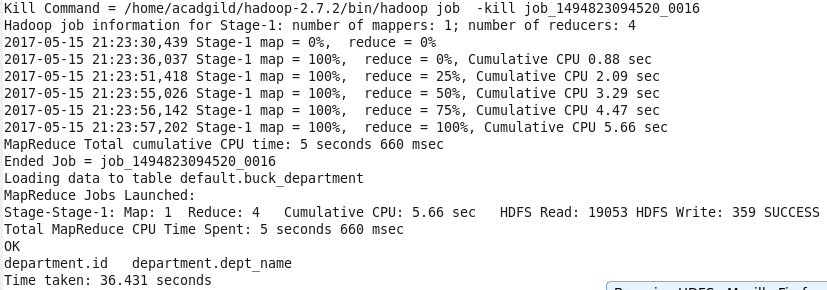
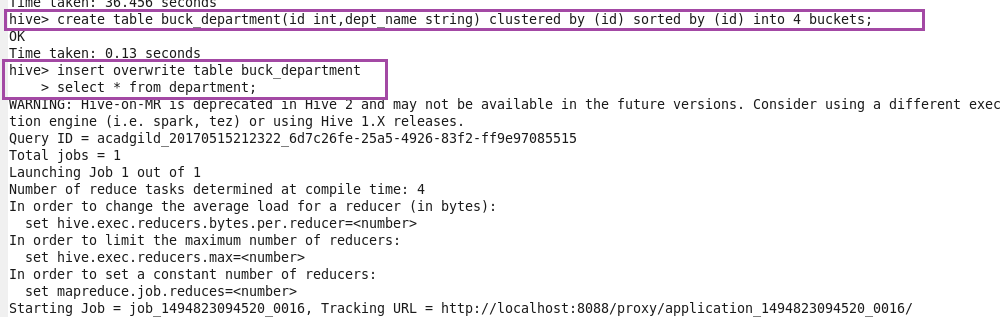
Set hive.enforce.bucketing=true;

Set hive.enforce.sorting=true;

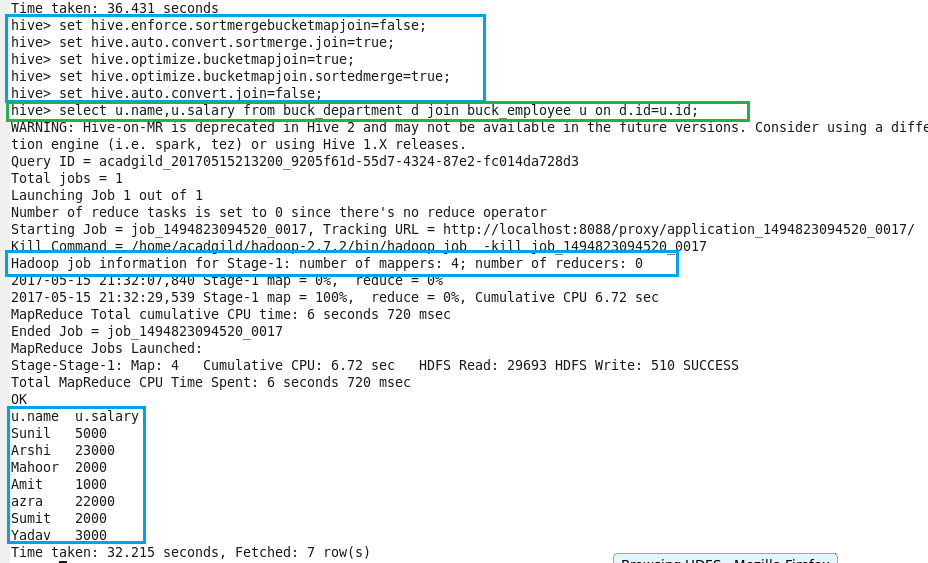
Creating bucketed table as buck\_employee and inserting to it.



Creating bucketed table as buck\_department and inserting to it.



Now the stage is set to perform SMB Map Join to optimize Hive joining. Again, make some changes in properties to perform SMB Map join.



See highlighted part in blue color. You will find that 4 mapper tasks are running (as we had 4 buckets). This helps in performing faster join operation when compared to regular.