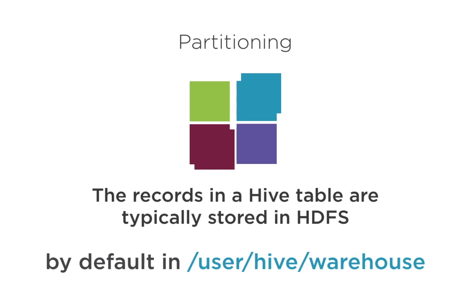
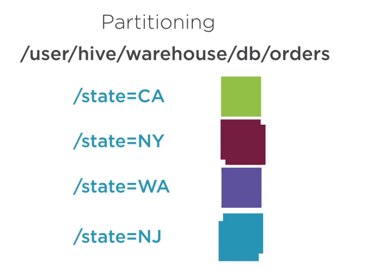
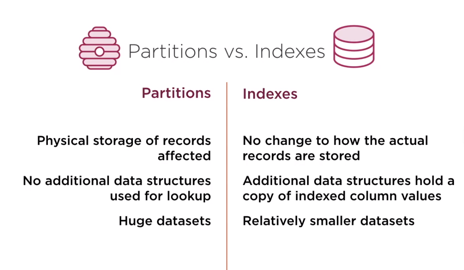
Hive stores data is below path in HDFS:

**/user/hive/warehouse**

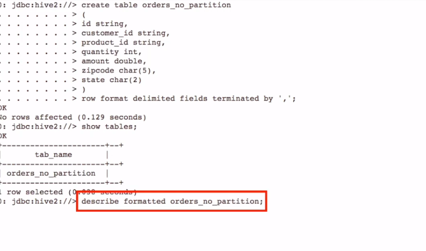


Data can be stored like below for Orders partitioned by States in USA like below:

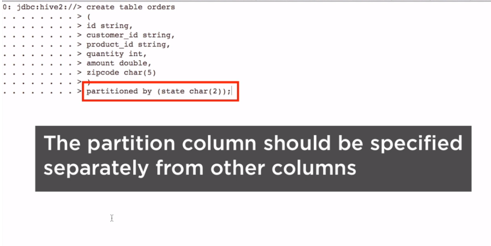




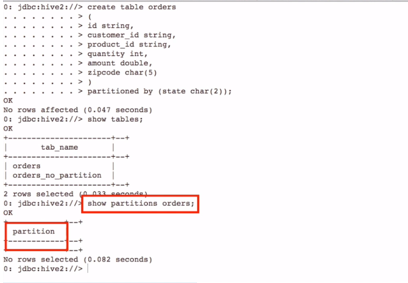
**Below we can create a table with no partitions:**



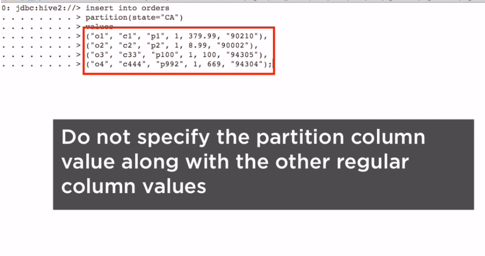
**Now with partitions:**



**Now show the partitions for the table:**

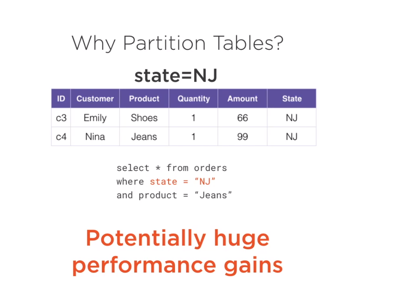


Now let’s isnert data based on partition data, now lets insert only for california.

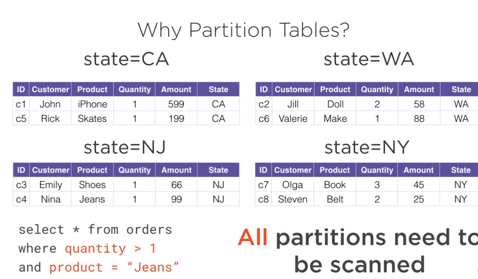


Now to list the directory path of this database , please use below command:

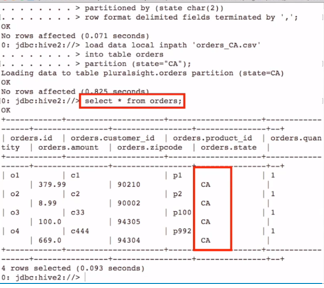
* Hadoop fs –ls /user/hive/warehouse/pluralsight.db/orders



Below all partitions will be scanned.



Load command in Hive:



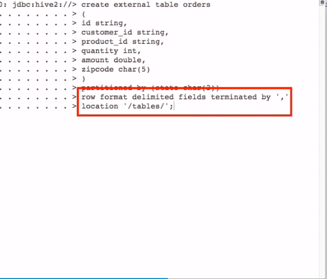
Now create an external table and then control/alter it through hive.

Table with and without buckets:

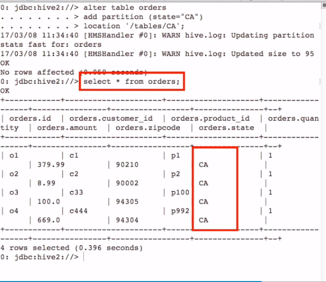


**Now let’s create an external table:**

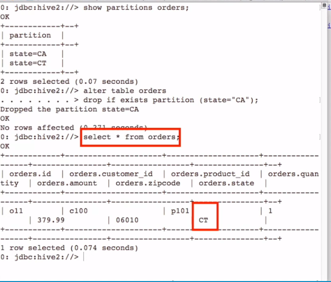
**We need to mention keyword ‘external’**



Now we need to alter the table to point it to the .csv file that we palced in the partitions.

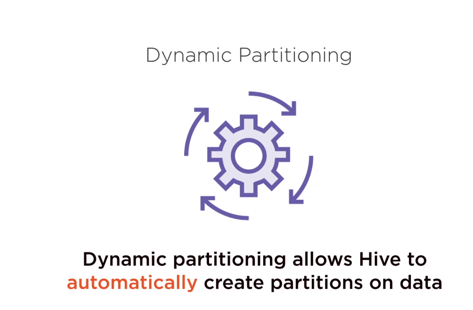


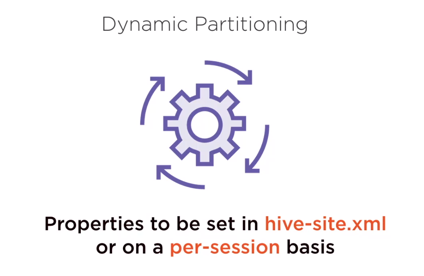
We can also drop a partition by below:



**Dynamic Partitioning:**

It is better to have your partitioning dynamic as namual process is hectic and not worth real time:



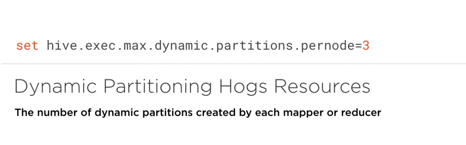


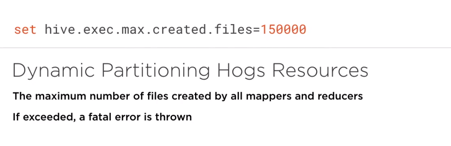
**To enable dynamic partitioning:**



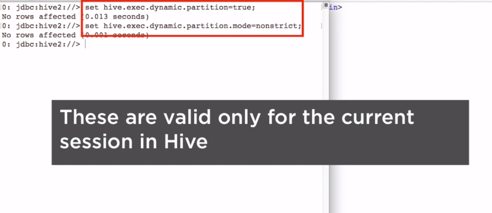




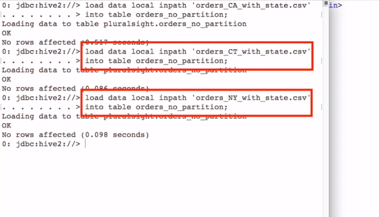


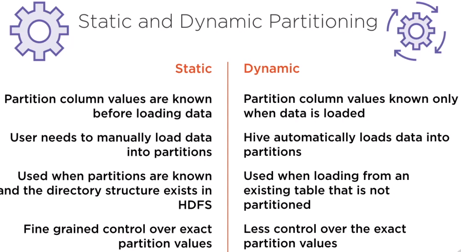


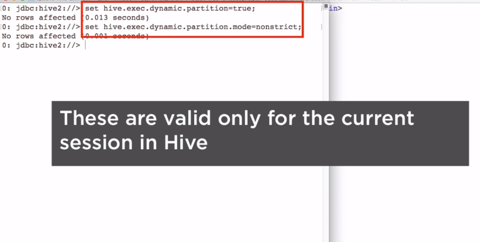
Now lets see practical and we shall set it in dynamic mode.

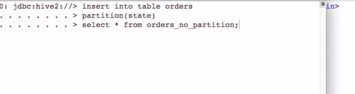


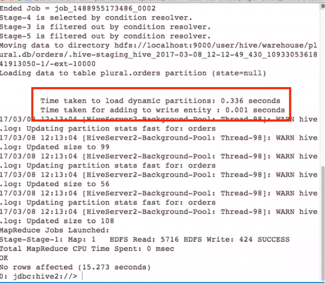
Now load the data in the new table without partition:



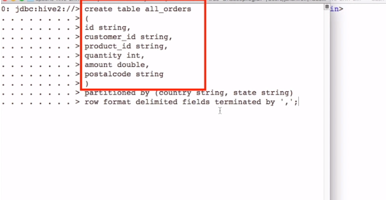




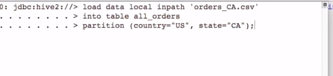




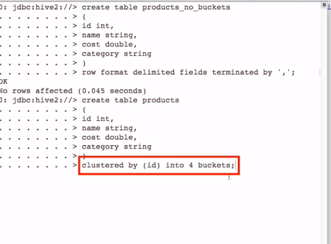
Multi partition Table



Here we can load the data and file paths are automatically created.



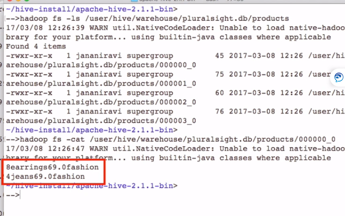
**Below is an example of a table with and without Bucket:**

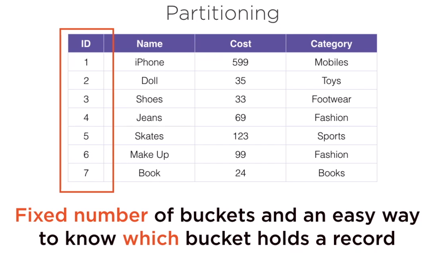


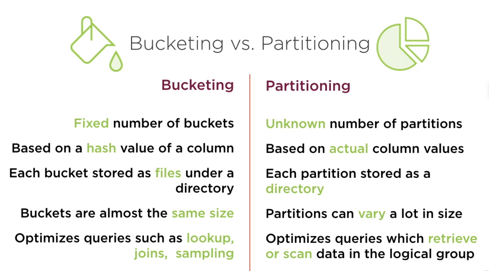
Sample file:



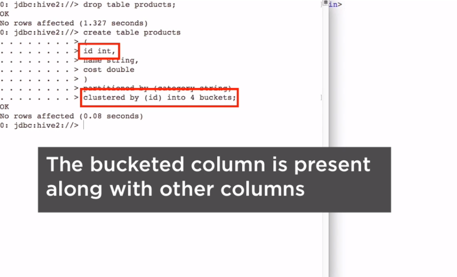
Wec annot load a table which is bucketed directly, we can only insert from a non-bucketed table to a bucketed table and notice below there are 4 bucket file systems created after inserting the data to a bucketed table;



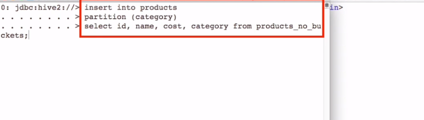




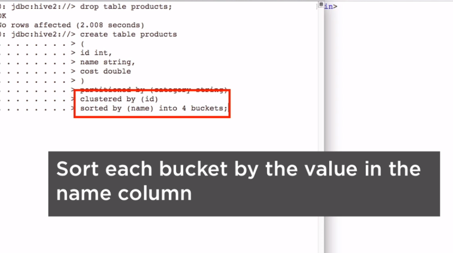
Now the table is bucketed as well as partitioned.



We need mention about the partition column in the insert column and make sure dynamic partition is enabled:



Now lets design a table with partition,bucket and sort

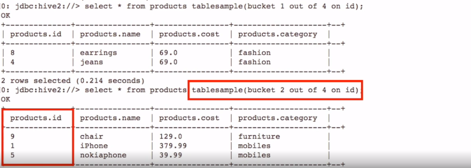


**It’s ideal to sort a column on which you have bucketed to get the best results:**

Hive samling data:



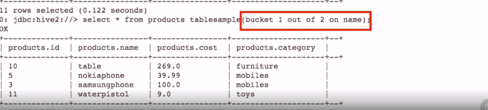
**Bucket Sampling:**



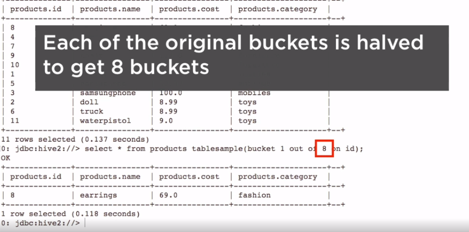
Even though a table has 4 buckets , we can forge it as 2 buckets so that the sampling can come out of the combined bucket out of the combined 2 buckets as below:



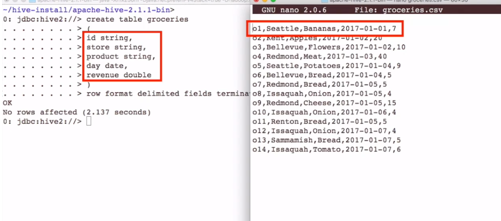
It can also bucket a data on the fly where it buckets the name field



It can also half the buket like 4 buckets can be halved to 8 buckets and we can querty like below:

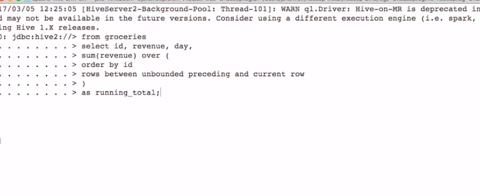


Window functions:



Now load the data.

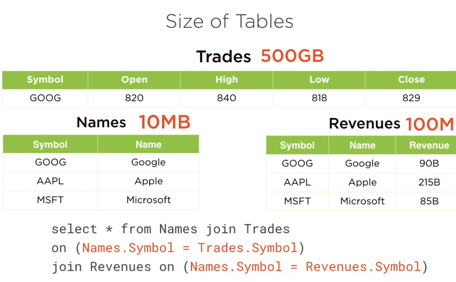




Joins in HIVE:

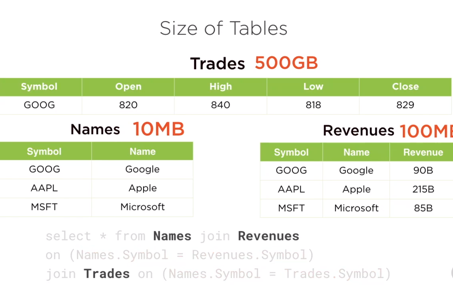




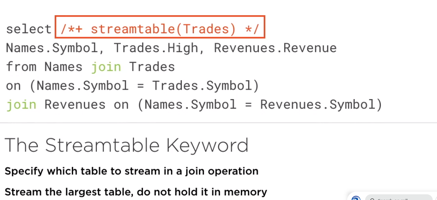


The first 2 tables re laoded in memory and the last table’s data is steamed to DISK and s always keep the smaller tables in the first join and have the large table is at the very end.

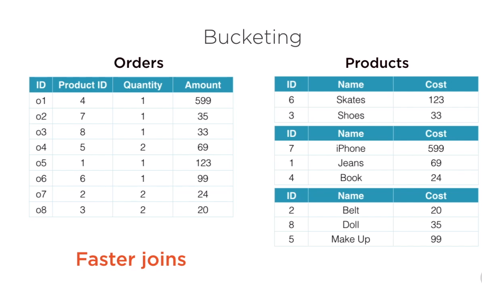
Rewrite as below where the large table is at the end.



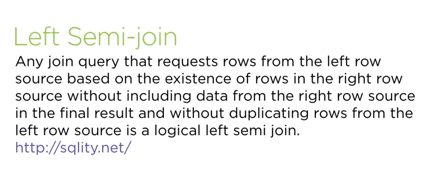
Mention the table which needs to be streamed as an option.



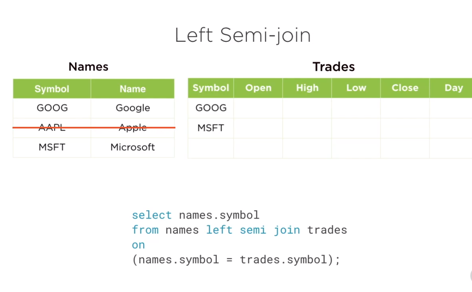
Bucketing is good for faster Joins.



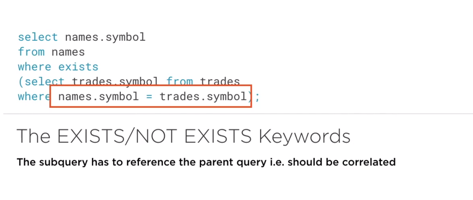
Left Semi Join



Apple will not be in the final result

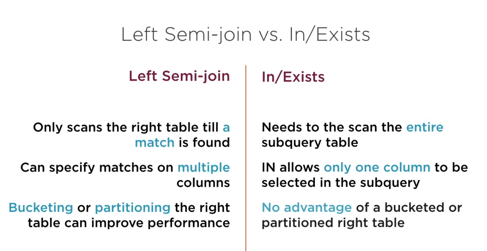


Exists query:

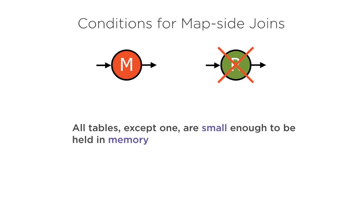


The left semi join is faster than the in or exist sub queries:

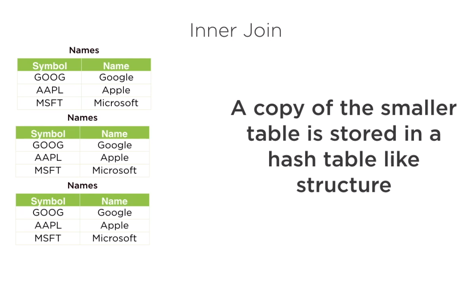


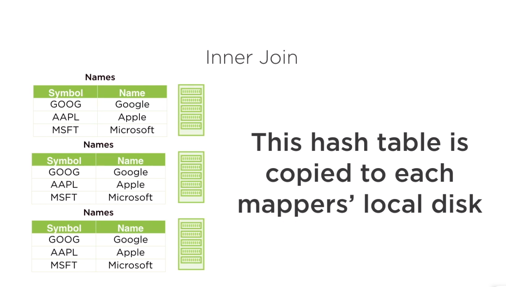


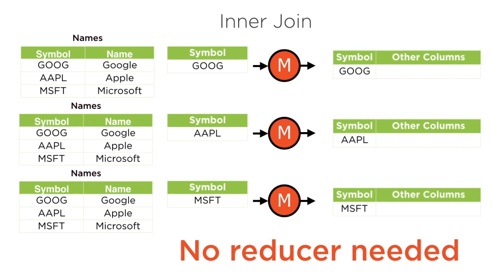
Map Only join or Map side Joins



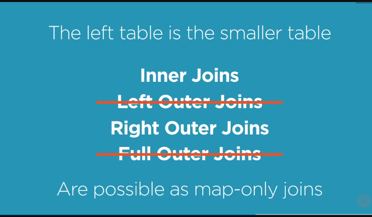
The small table is brought into memory as a hash table and the hash table is copied over to all machines where the second table is searched upon.

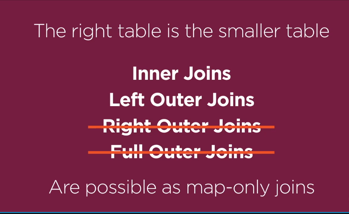






When the ledt table is smaller than the right table then only brlow joins can have map side or map oly joins and can remove the redcue phase:

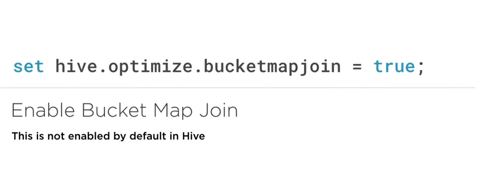




**Bucket Map Join:**

**This can be done only when the joins in the buckets are an integral multiple of the other bucket like 2 is to 4 below.**



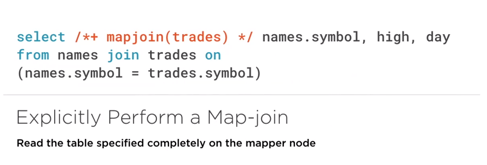




**TO perform merge-sort operation:**



**To specify explicitly to do only map-only join , need to use below:**



**To do SCD1 type update in HiveQL, Use below by creating a view:**

**A view created between the initial file and the modified file base don timestamp will give the latest updated record in the view and the same is stored in the external table.**

