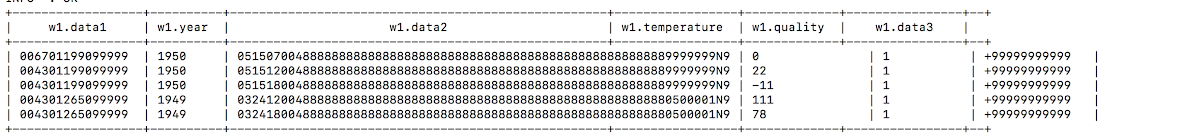
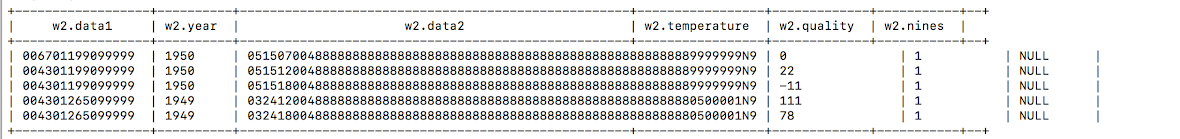
**HW 6 – ic1018**

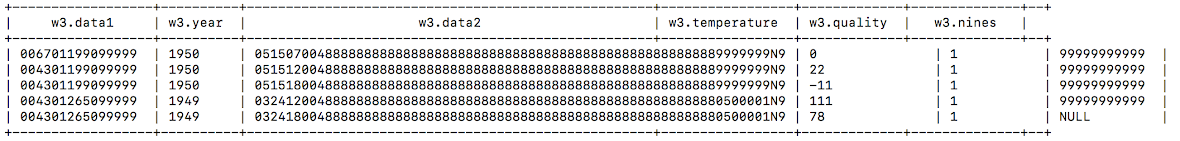
**1. In Step 4 of the Hive homework (red text), is there an unexpected result? If so, what is it and what is the reason for the result?**

 Output from w1 (date3):

Output from w2 (after using nines int):



Output from w3 (after using nines bigint):



Yes, there is are unexpected results. For w2, it didn’t show the +9999999.. all those numbers unlike for w1, and they are all NULL. Similarly, for w3, it didn’t show for the last dataset and is NULL. This is probably because it cannot read integers (int) specifically ‘int’. However it can read ‘big int’ except the last value. For the first case, the last column is read as string.

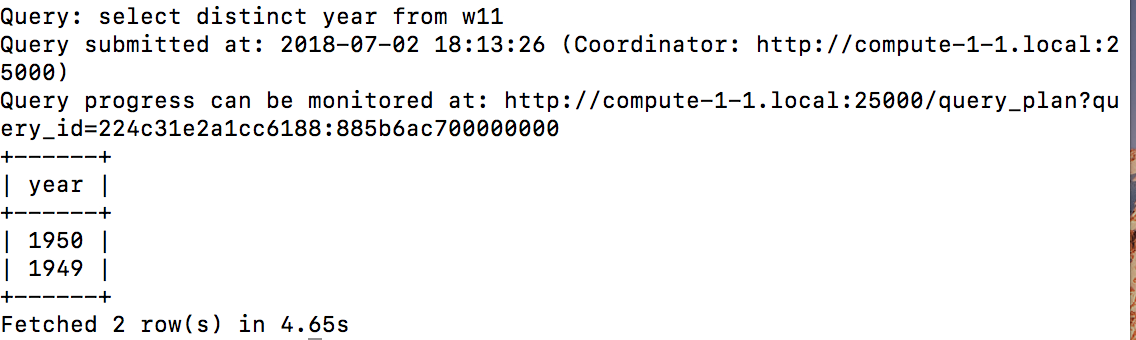
**2. In Step 3 of the Impala homework (red text):**

     a) Did you notice a difference in the time required to execute this command from Impala:

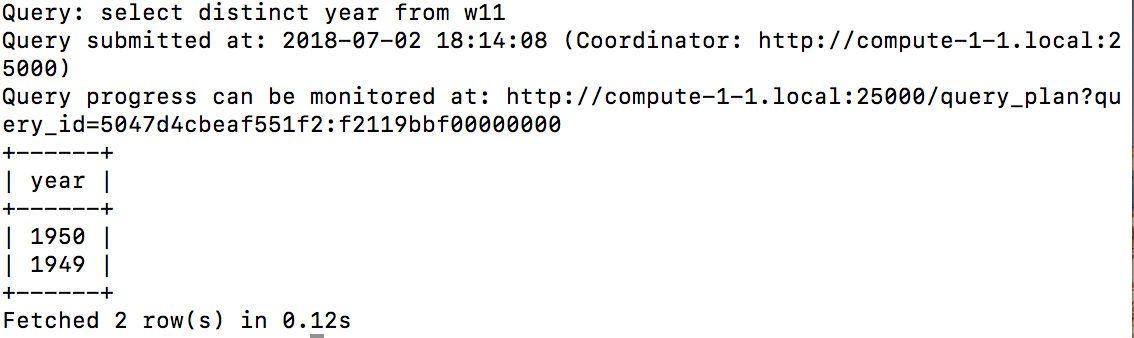
**select distinct year from w11;**

          Why do you think the timing of the second run of this command was faster?

 Query ran for the first time:



Query ran for the second time:



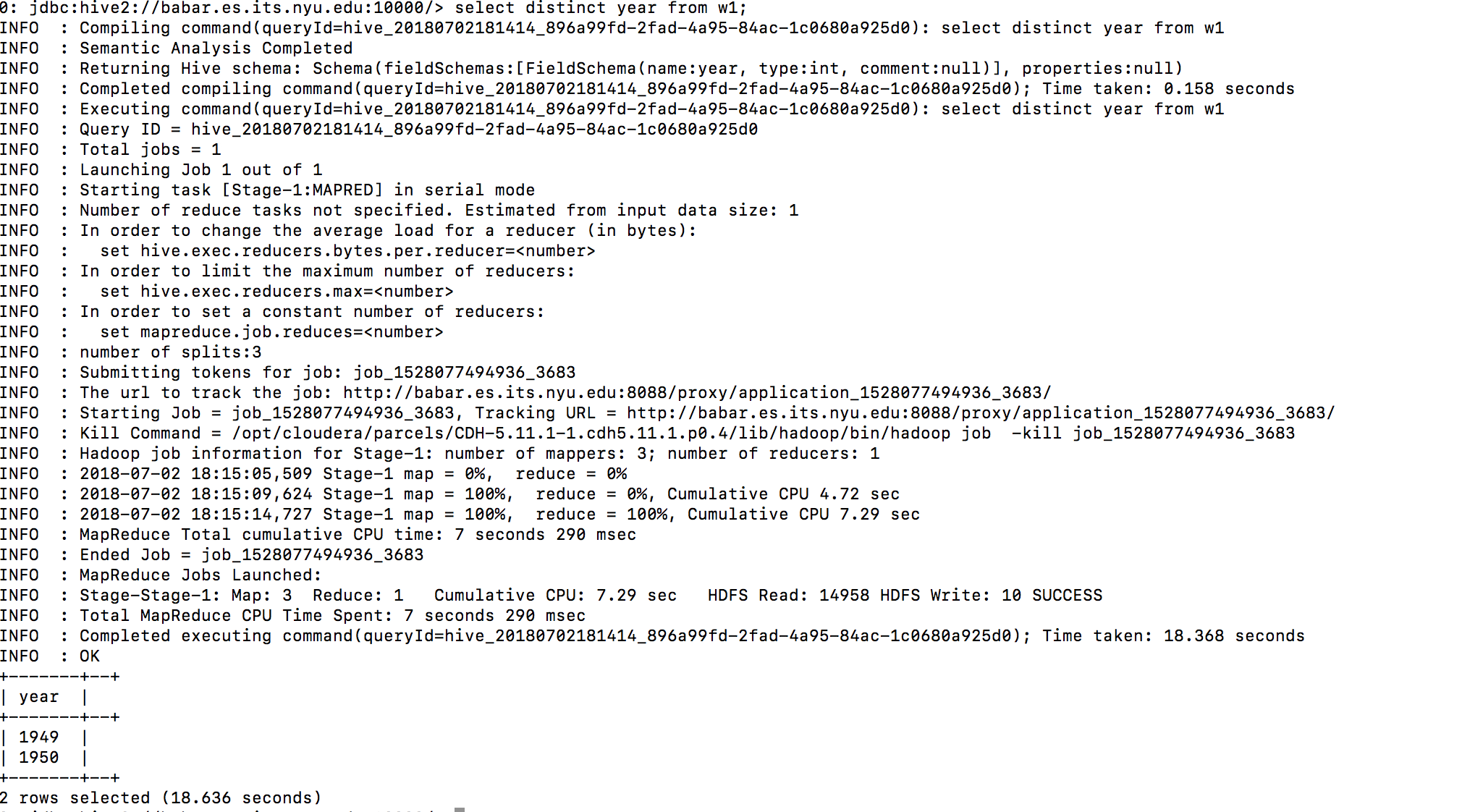
Yes, second time the query ran faster by 4.53 seconds. This is because Impala doesn’t use Hadoop at all, and is more like SQL on HDFS. It has daemons running on the cluster nodes which cache some of the data that is in HDFS, so that’s why second time the same query runs faster, as the results have been cached when the query is ran the first time.

      b) Did you notice a difference in the time required to execute this command from Hive:

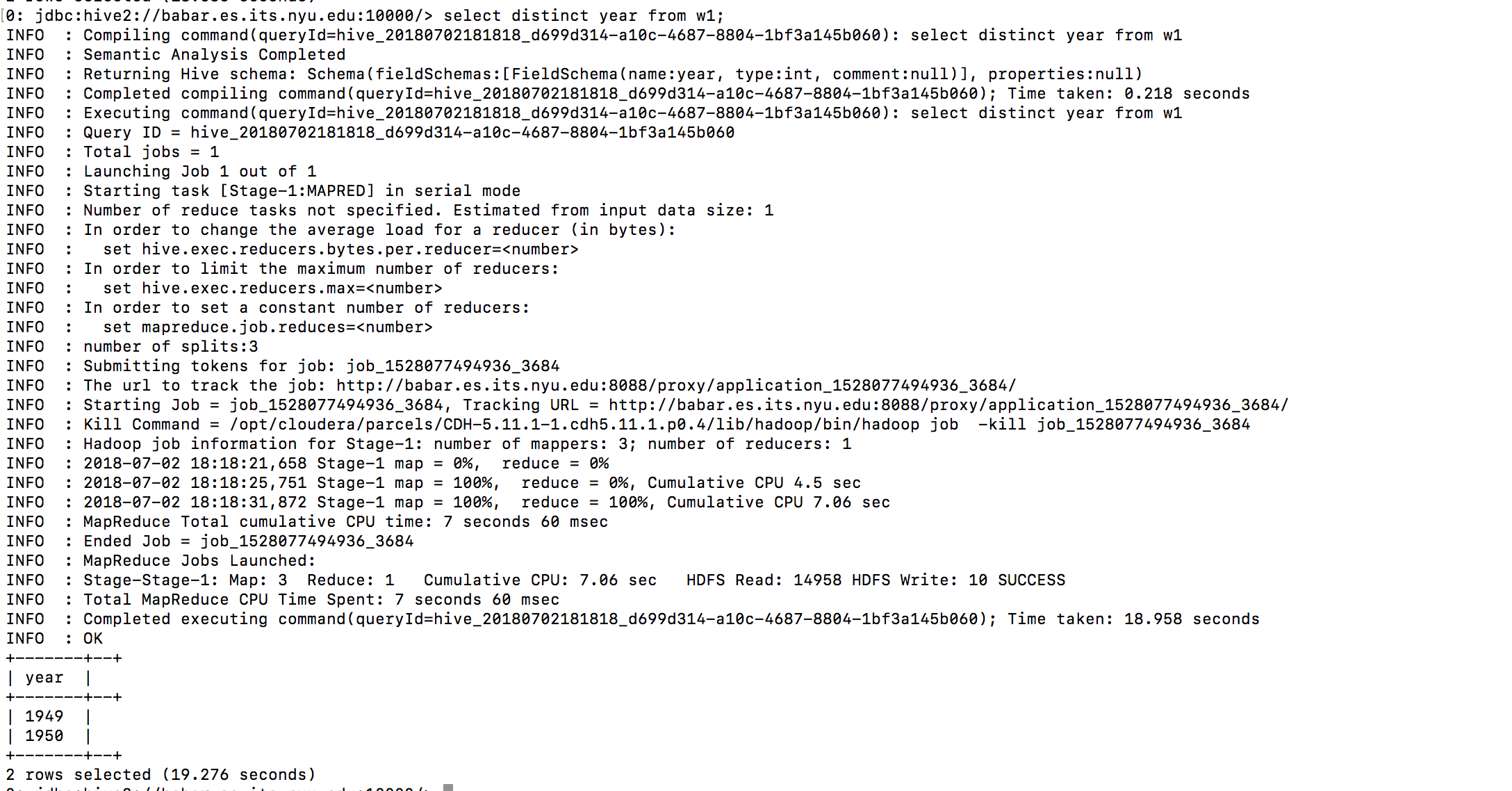
**select distinct year from w1;**

          Why do you think the timing of the second run of this command was not faster?

Query ran in hive for the first time:



Query ran in Hive for the second time:



Here, the query didn’t run faster the second time. In fact, the second time took few seconds longer than the first time. Hive is more like running SQL on Hadoop, and thus involves in running a Map/Reduce job, both the times query is ran. There is a certain overhead involved in running a Map/Reduce job in Hive. In Impala, one gains the runtime as Map/Reduce is altogether short-circuited. .