**Data Pipeline**

**Scenario**

There is a Famous Restaurant Owner who wants to know how much his overall rating of each item that is sold in his restaurants. He is owning many branches in Kerala, Karnataka, TamilNadu. as the data is not stored in an appropriate format he approaches the Data Engineer to process the Data from him as per his requirement

**Input :**

The data is stored in 3 CSV files Region wise  **Kerala, Karnataka, TamilNadu**.

**CSV format**

**"So\_no","Customer","ItemId","Rating"**

**Output:**

The Owner wants to store it in two Data Base Tables

**Table 1 => Userreview =>** where all the Users review will be stored irrespective of the region.

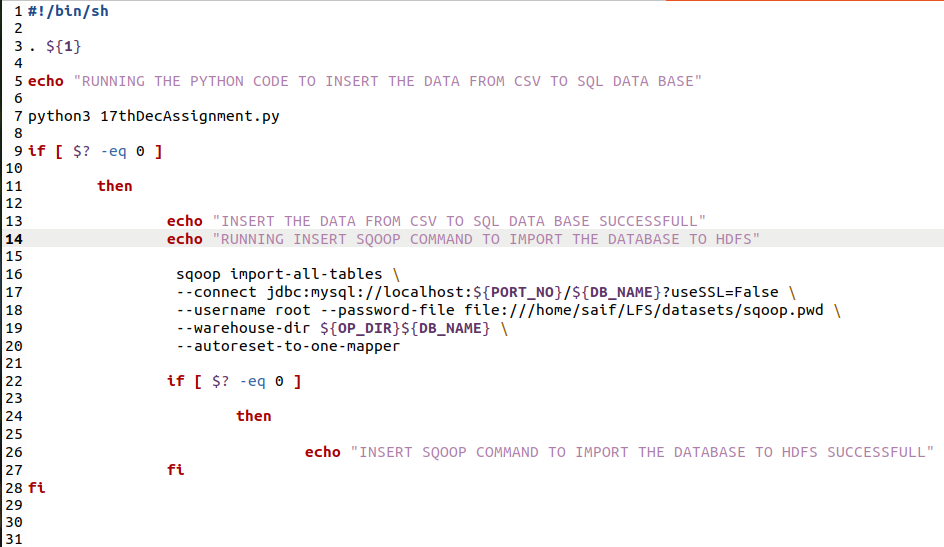
**Table 2 => Item\_review\_rating =>** Item wise the Review rating Aggregate will be stored.

The Owner also wants to **store this table in HDFS** as well.

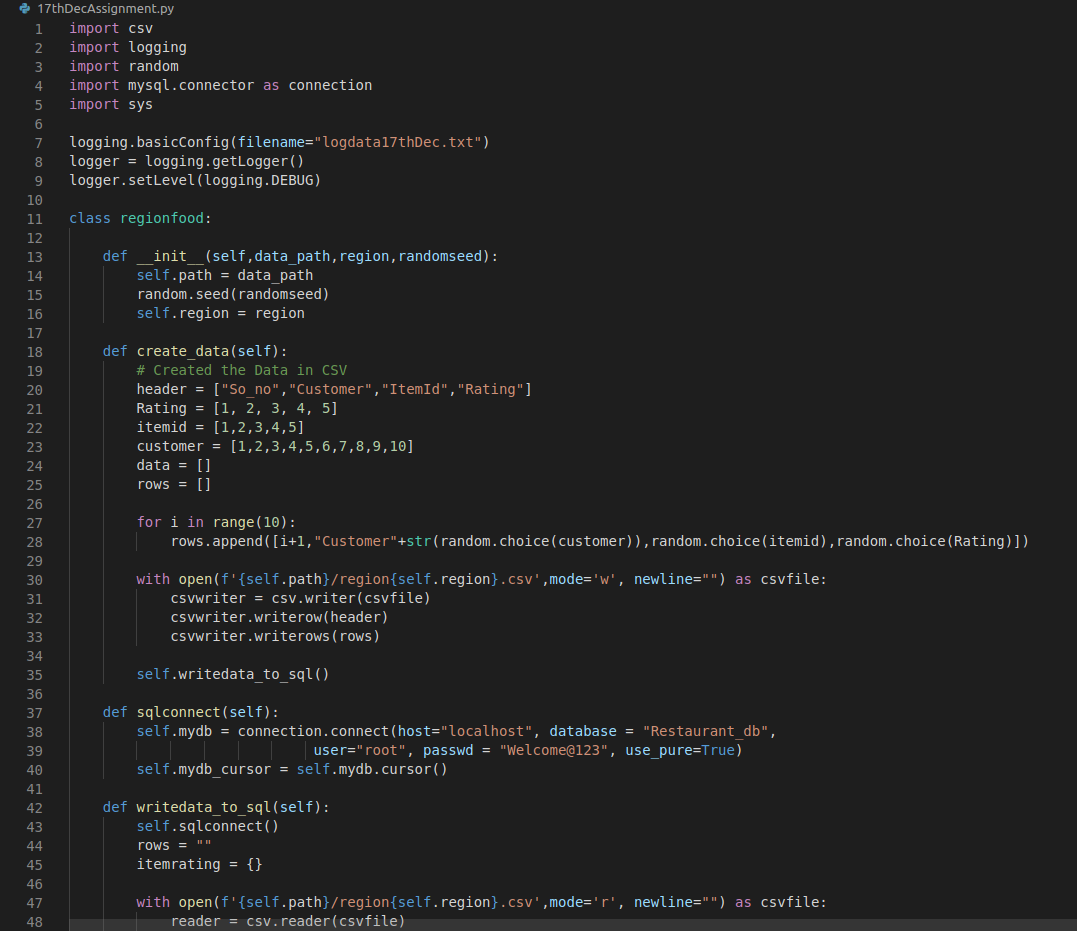
**My Approach:**

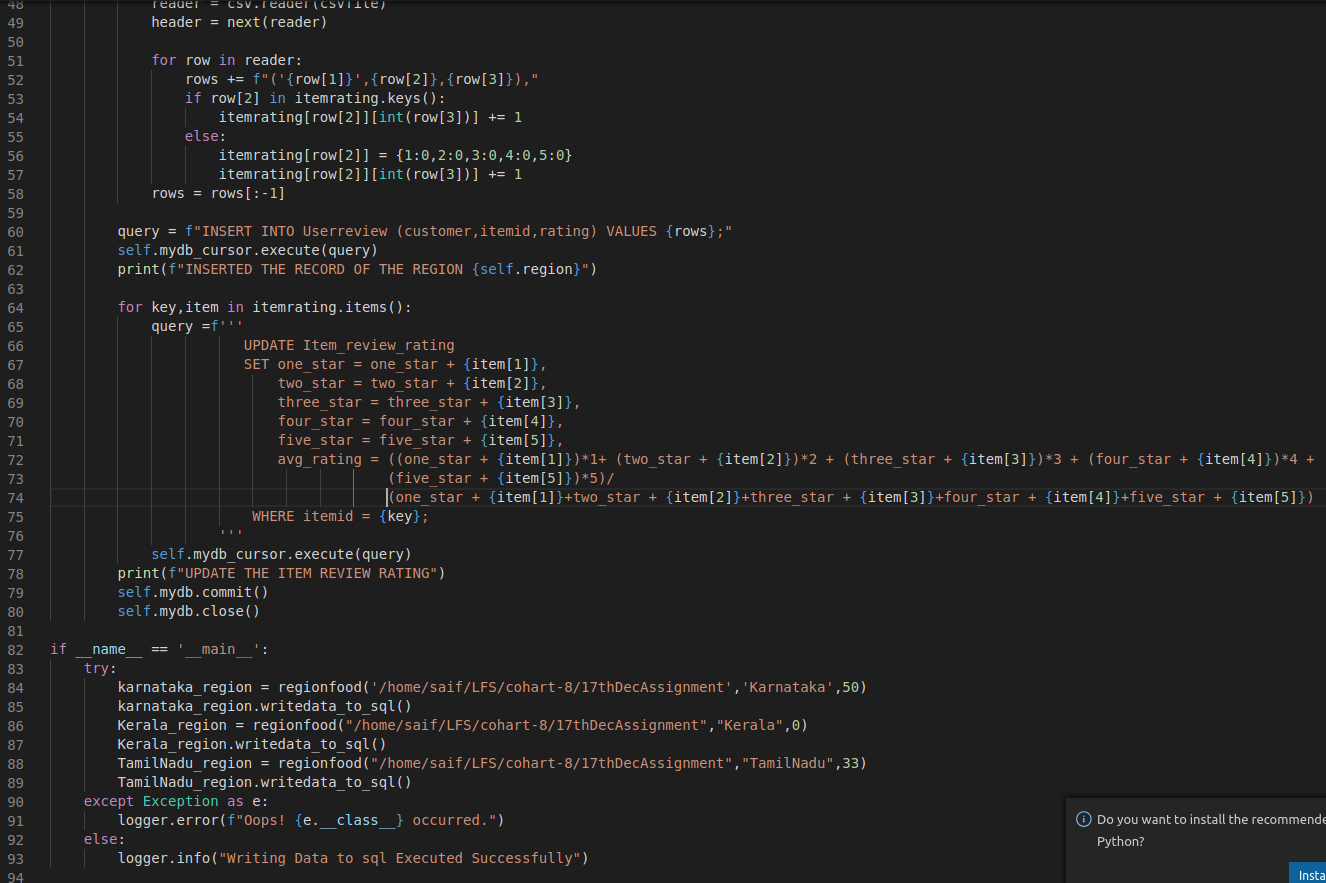
* Using python read the CSV file one by one insert the record directly in the **Userreview Table** and calculate the aggregate of the Rating Item wise update it in the **Item\_review\_rating Table**
* Using the Sqoop write the command to Insert Entire Data Base into the HDFS
* Write a common shell script that executes Python and the Sqoop in one file execution

**SHELL SCRIPT:**

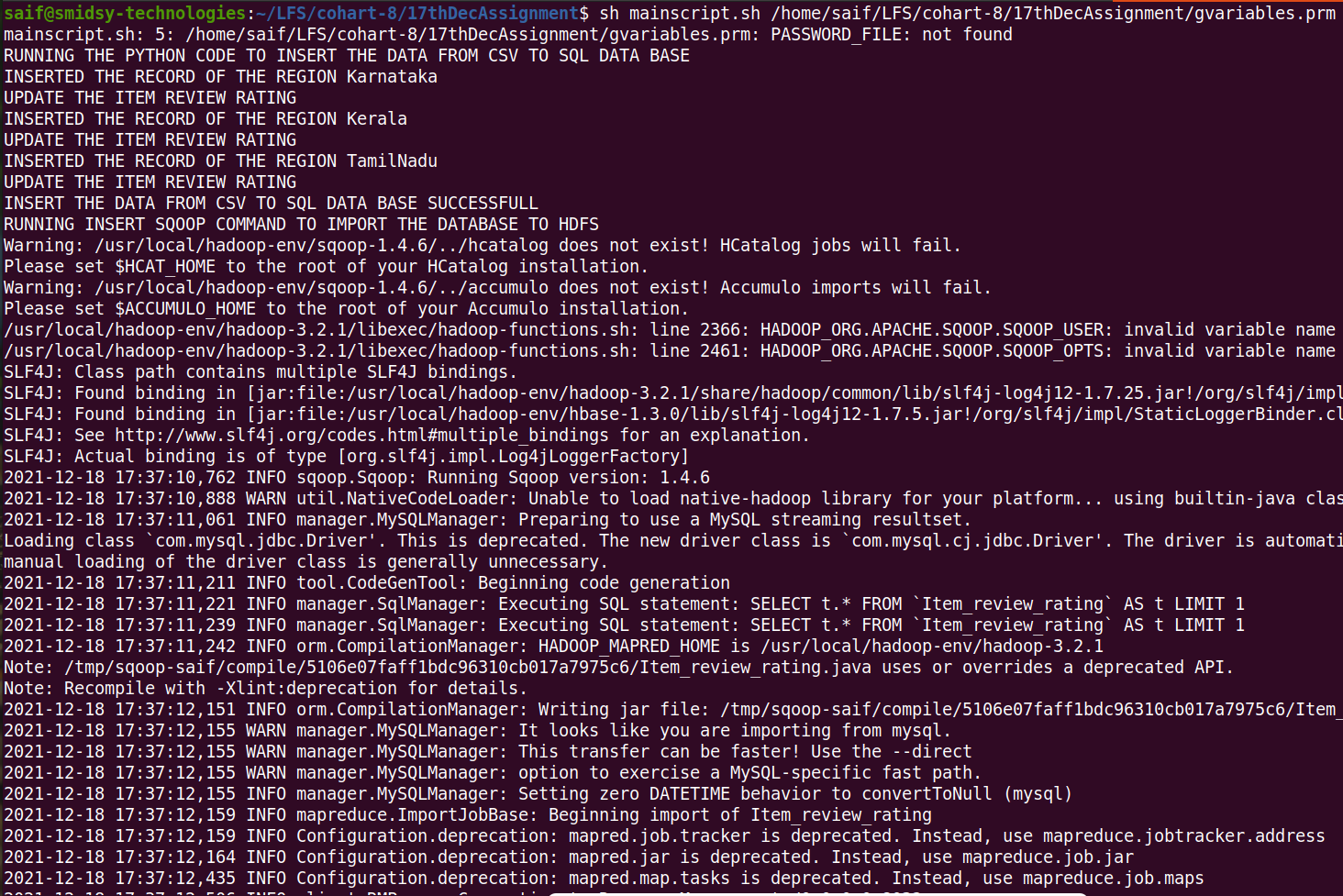


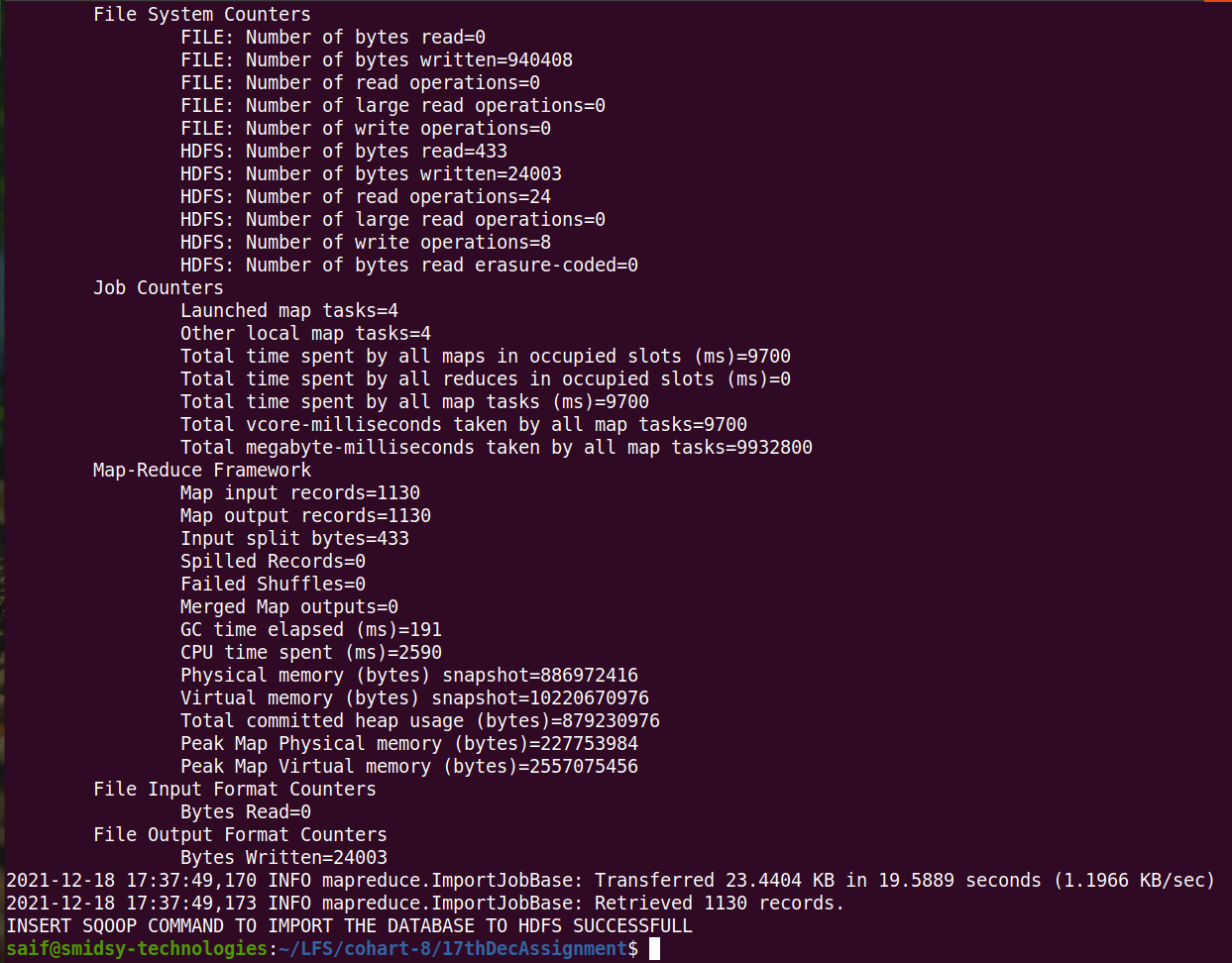
**PYTHON CODE:**



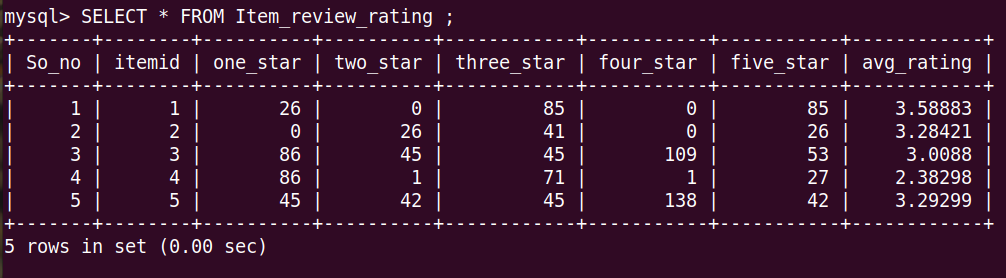


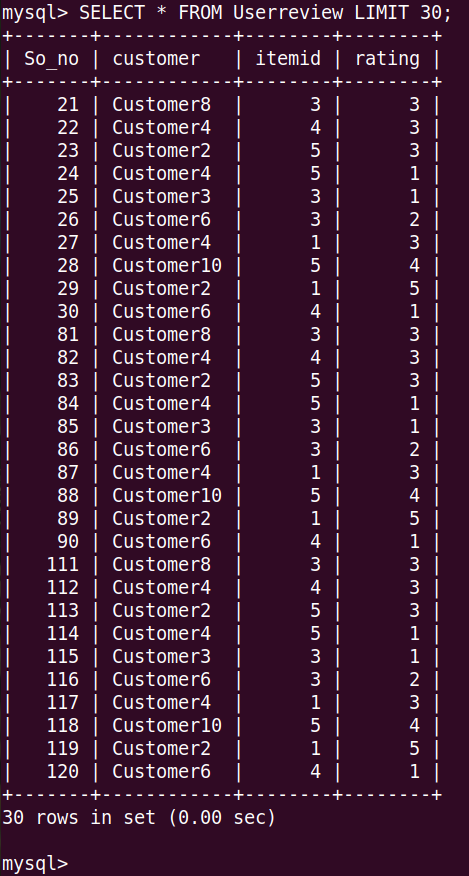
**CODE EXECUTION:**





**TABLES:**





**HDFS:**

