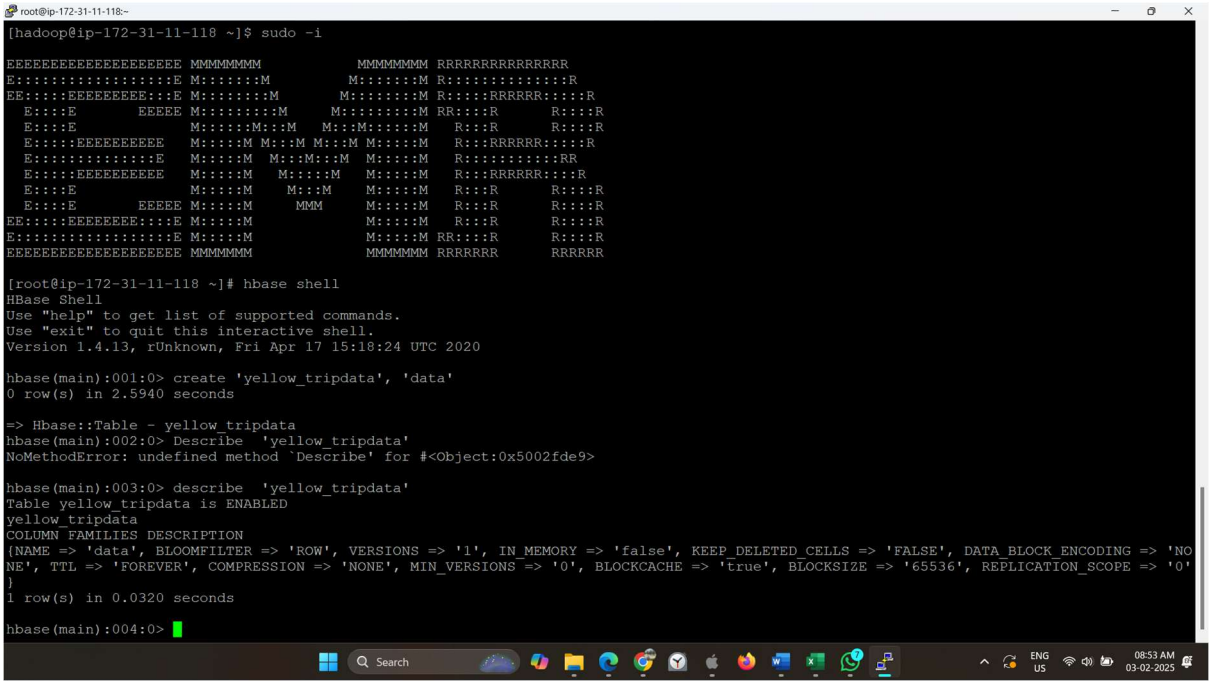
Document 02 – Ingestion Task.pdf by Huang-Yin Tso, Arun Santhosh, Vipul Kumar and Gaurav Sharma

HBase from EMR

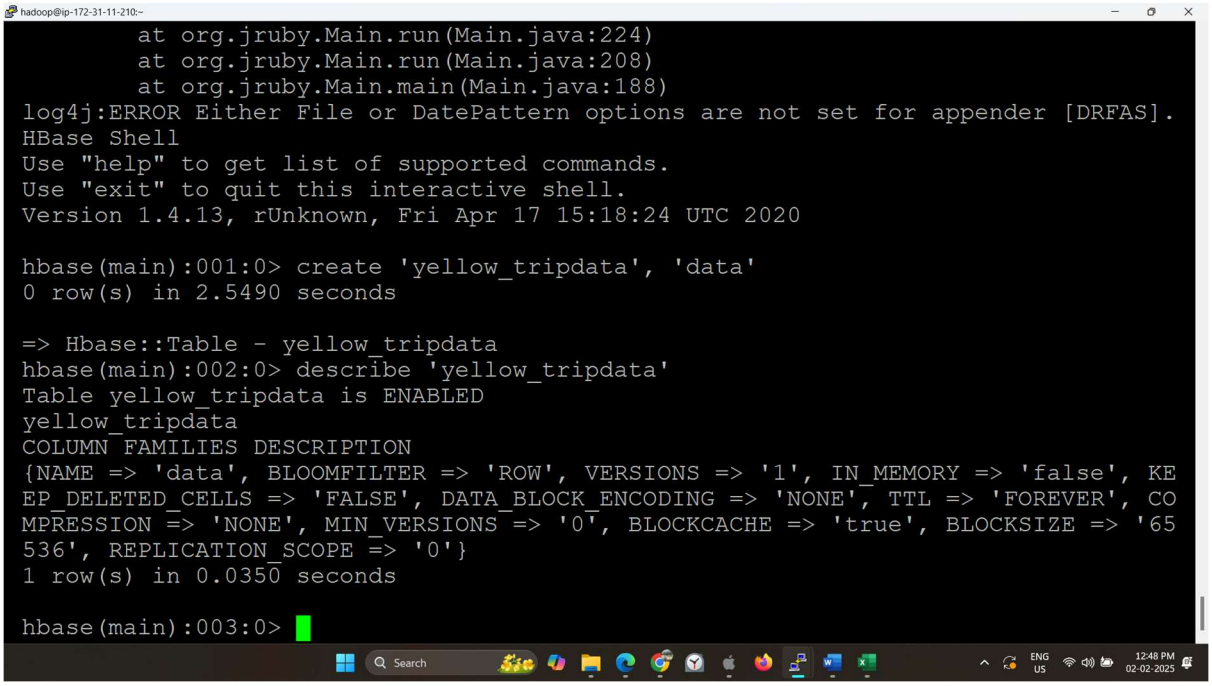
hbase shell

Creating table in HBase

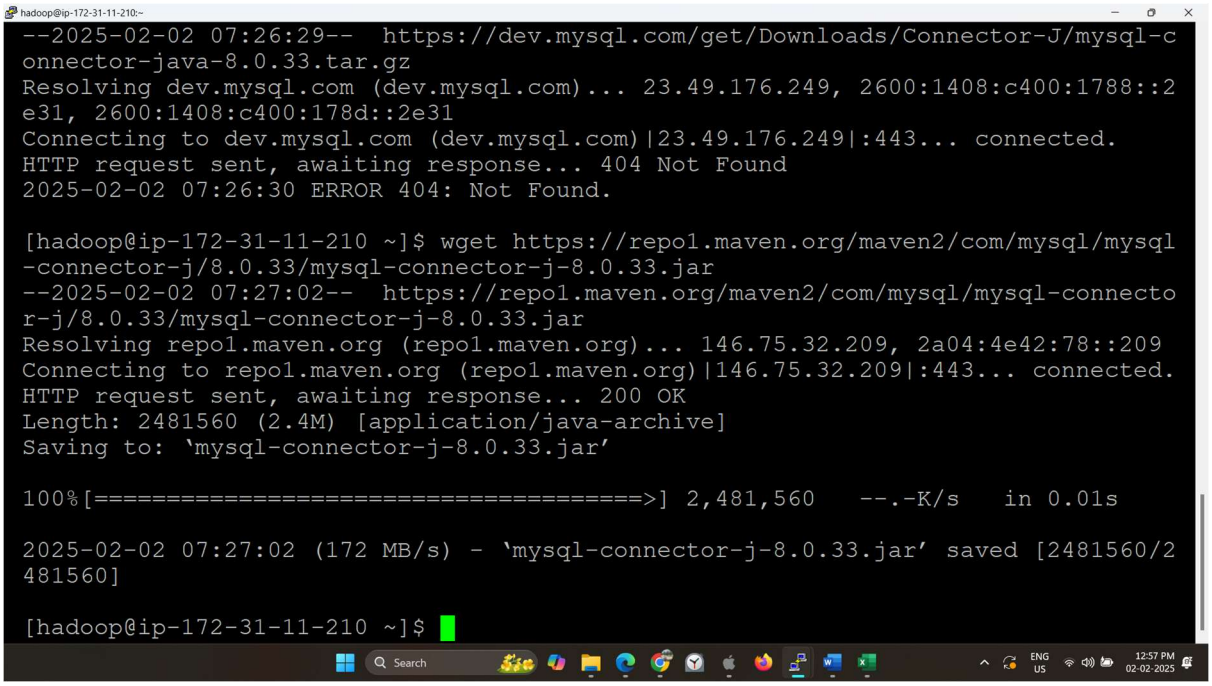
create 'yellow_tripdata', 'data'



Describe 'yellow_tripdata'

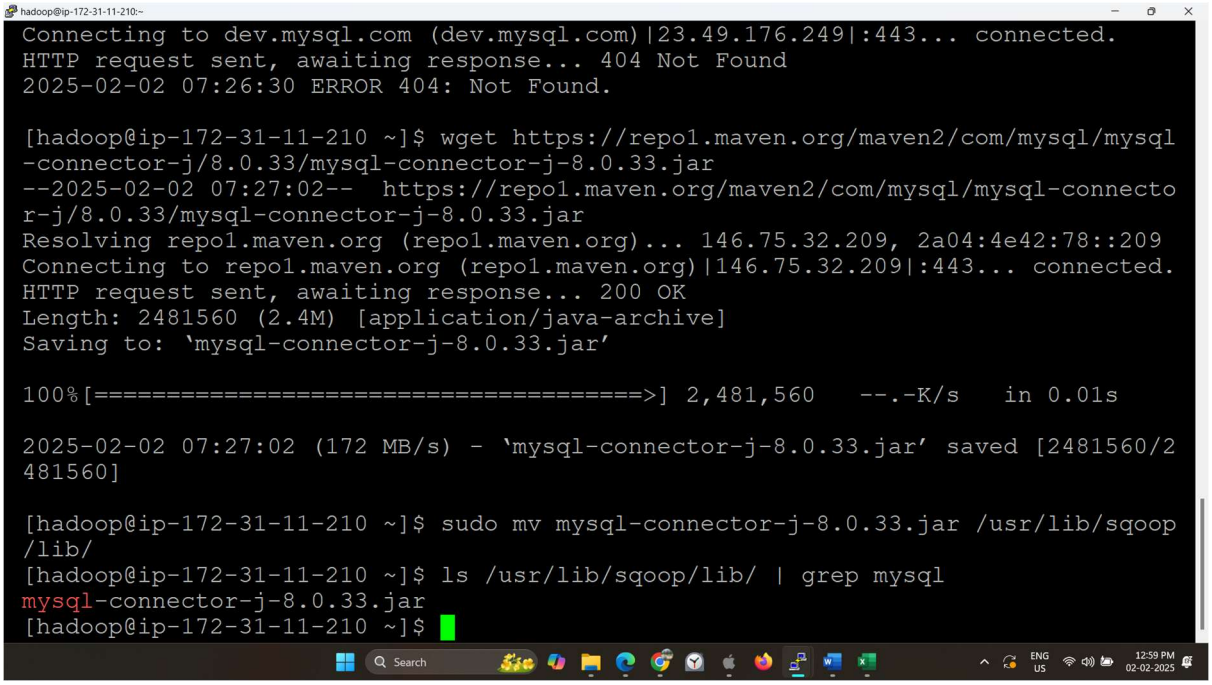


wget https://repo1.maven.org/maven2/com/mysql/mysql-connector-j/8.0.33/mysql-connector-j-8.0.33.jar



After downloading, move the JAR file to Sqoop's library directory to ensure Sqoop can access the MySQL driver:

sudo mv mysql-connector-j-8.0.33.jar /usr/lib/sqoop/lib/



Sqoop import

Since table lacks a primary key, using a **numerical column** as a split key: PULocationID (Pickup Location ID).

sqoop import \

--connect jdbc:mysql://mysqldb.cyhl71hae9td.us-east-1.rds.amazonaws.com/mysqldb \

--username admin \

--password admin123 \

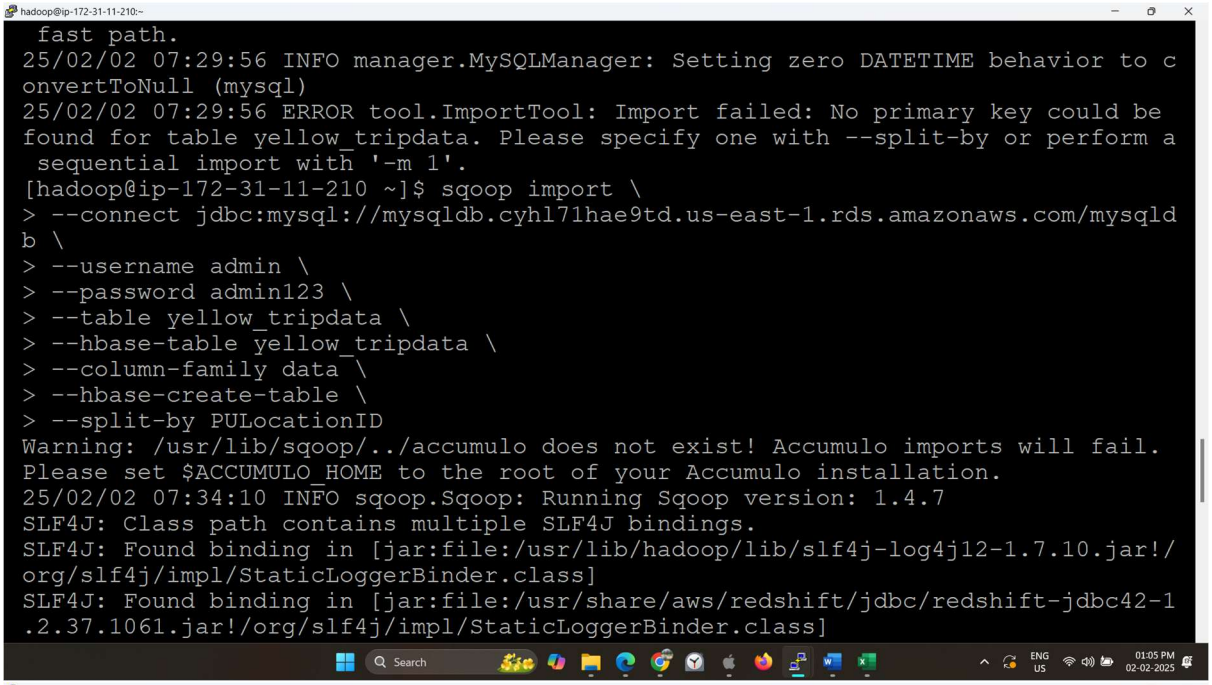
--table yellow_tripdata \

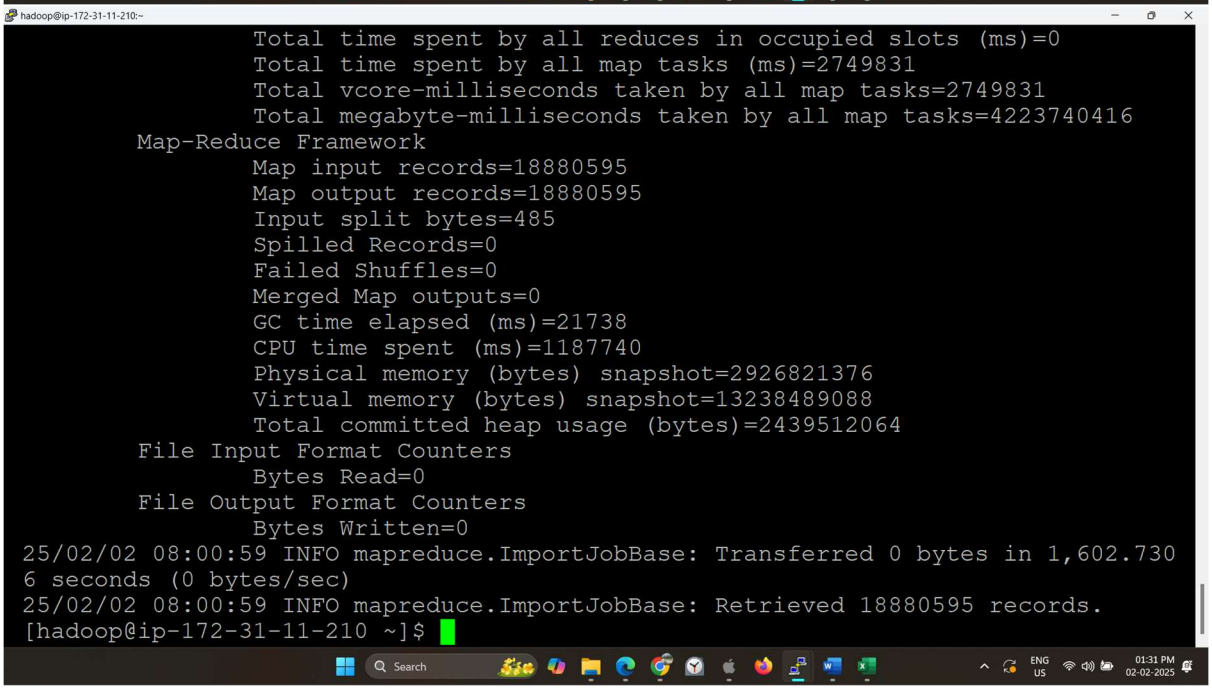
--hbase-table yellow_tripdata \

--column-family data \

--hbase-create-table \

--split-by PULocationID





scan ' yellow_tripdata', {LIMIT => 2}