

## Ingestion Task

- **Command for running sudo**  
sudo - i
- **Command for entering inside hbase**  
hbase shell
- **Command for creating hbase table**  
create 'taxi', 'cf'
- **Command for listing all tables**  
list

```
root@ip-172-31-90-95:~  
E::::E      M::::M      M::M      M::::M      R::R      R::::R  
E::::E      EEEEE M::::M      MMM      M::::M      R::R      R::::R  
EE:::::EEEEEEEEE:E M::::M      M::::M      R::R      R::::R  
E:::::EEEEEEEEE:E M::::M      M::::M      RR:::R      R::::R  
EEEEEEEEEEEEEEEEEE MMMMMM      MMMMMM      RRRRRR      RRRRRR  
  
[root@ip-172-31-90-95 ~]# hbase shell  
HBase Shell  
Use "help" to get list of supported commands.  
Use "exit" to quit this interactive shell.  
Version 1.4.13, rUnknown, Wed Jun  8 00:30:30 UTC 2022  
  
hbase(main):001:0> hbase 'taxi','cf'  
NoMethodError: undefined method `hbase' for #<Object:0x37a9b687>  
  
hbase(main):002:0> create 'taxi','cf'  
0 row(s) in 2.5980 seconds  
  
=> Hbase::Table - taxi  
hbase(main):003:0> list  
TABLE  
taxi  
1 row(s) in 0.0160 seconds  
  
=> ["taxi"]  
hbase(main):004:0> █
```

- **Command for installing mysql on EMR cluster**

wget <https://de-mysql-connector.s3.amazonaws.com/mysql-connector-java-8.0.25.tar.gz>

tar -xvf mysql-connector-java-8.0.25.tar.gz

cd mysql-connector-java-8.0.25

sudo cp mysql-connector-java-8.0.25.jar /usr/lib/sqoop/lib/

- **Command for importing table contents from RDS mysql to Hbase table**  
sqoop import --connect jdbc:mysql://upgradassignmentdb.cdwr877prmiq.us-east-1.rds.amazonaws.com:3306/mapreduce1 --username admin --password admin123 --table Taxi --hbase-table taxi --column-family cf --hbase-row-key Trip\_no -m 1

```
root@ip-172-31-93-25:~  
Current count: 10217000, row: 29210512  
Current count: 10218000, row: 29211512  
Current count: 10219000, row: 29212512  
Current count: 10220000, row: 29213512  
Current count: 10221000, row: 29214512  
Current count: 10222000, row: 29215512  
Current count: 10223000, row: 29216512  
Current count: 10224000, row: 29217512  
Current count: 10225000, row: 29218512  
Current count: 10226000, row: 29219512  
Current count: 10227000, row: 29220512  
Current count: 10228000, row: 29221512  
Current count: 10229000, row: 29222512  
Current count: 10230000, row: 29223512  
Current count: 10231000, row: 29224512  
Current count: 10232000, row: 29225512  
Current count: 10233000, row: 29226512  
Current count: 10234000, row: 29227512  
Current count: 10235000, row: 29228512  
Current count: 10236000, row: 29229512  
10236828 row(s) in 493.1410 seconds  
  
=> 10236828  
hbase(main):002:0>
```

- **Installing happybase**  
sudo yum install python3-devel  
pip install happybase
- **Running batch\_ingest.py**  
python batch\_ingest.py
- **Count after batch wise insertion of data to taxi table in hbase**  
count 'taxi'

```
root@ip-172-31-93-114:~  
Current count: 10028000, row: 9982776  
Current count: 10029000, row: 9983676  
Current count: 10030000, row: 9984576  
Current count: 10031000, row: 9985476  
Current count: 10032000, row: 9986376  
Current count: 10033000, row: 9987276  
Current count: 10034000, row: 9988176  
Current count: 10035000, row: 9989076  
Current count: 10036000, row: 9989977  
Current count: 10037000, row: 9990876  
Current count: 10038000, row: 9991776  
Current count: 10039000, row: 9992676  
Current count: 10040000, row: 9993576  
Current count: 10041000, row: 9994476  
Current count: 10042000, row: 9995376  
Current count: 10043000, row: 9996276  
Current count: 10044000, row: 9997176  
Current count: 10045000, row: 9998076  
Current count: 10046000, row: 9998977  
Current count: 10047000, row: 9999877  
10047135 row(s) in 237.8270 seconds  
  
=> 10047135  
hbase(main):002:0>
```

```
hadoop@ip-172-31-93-114:~  
.3.5)  
Requirement already satisfied: pytz>=2017.3 in /usr/local/lib/python3.7/site-packages (from pandas) (2022.1)  
Requirement already satisfied: numpy>=1.17.3; platform_machine != "aarch64" and platform_machine != "arm64" and python_version < "3.10" in /usr/local/lib64/python3.7/site-packages (from pandas) (1.20.0)  
Requirement already satisfied: python-dateutil>=2.7.3 in ./local/lib/python3.7/site-packages (from pandas) (2.8.2)  
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.7/site-packages (from python-dateutil>=2.7.3->pandas) (1.13.0)  
[hadoop@ip-172-31-93-114 ~]$ python batch_ingest.py  
Traceback (most recent call last):  
  File "batch_ingest.py", line 68, in <module>  
    insert_first_column(f)  
  File "batch_ingest.py", line 32, in insert_first_column  
    df.insert(0, 'New_ID', range(rowc, rowc + len(df)))  
UnboundLocalError: local variable 'rowc' referenced before assignment  
[hadoop@ip-172-31-93-114 ~]$ vi batch_ingest.py  
[hadoop@ip-172-31-93-114 ~]$ python batch_ingest.py  
Connect to HBase. table name: taxi, batch size: 1000  
Connected to file. name: ['/home/hadoop/yellow_tripdata_2017-03.csv', '/home/hadoop/yellow_tripdata_2017-04.csv']  
Done. row count: 10047136, duration: 1472.217 s  
[hadoop@ip-172-31-93-114 ~]$
```