```
READY
Spark Scala HDFS Hive and Impala
   hdfs dfs -put -f /home/cloudera/Case_Study/yellow_tripdata_2020-01.csv /nyctaxi/data
hdfs dfs -put -f /home/cloudera/Case_Study/yellow_tripdata_2020-02.csv /nyctaxi/data
hdfs dfs -put -f /home/cloudera/Case_Study/taxi+_zone_lookup.csv /nyctaxi/lookup
   hdfs dfs -ls /nyctaxi/data
hdfs dfs -ls /nyctaxi/lookup
  %spark
sc
                                                                                                                                                                    READY
  res1: org.apache.spark.SparkContext = org.apache.spark.SparkContext@49985d2
   %spark
                                                                                                                                                                    READY
   trips01_DF.printSchema()
trips02_DF.printSchema()
trips01_DF.show(5)
trips02_DF.show(5)
   |-- VendorID: integer (nullable = true)
   -- tpep_pickup_datetime: timestamp (nullable = true)
   |-- tpep dropoff datetime: timestamp (nullable = true)
    -- passenger_count: integer (nullable = true)
   |-- trip_distance: double (nullable = true)
|-- RatecodeID: integer (nullable = true)
   |-- store_and_fwd_flag: string (nullable = true)
|-- PULocationID: integer (nullable = true)
   |-- DOLocationID: integer (nullable = true)
   |-- payment type: integer (nullable = true)
   -- fare_amount: double (nullable = true)
   |-- extra: double (nullable = true)
   |-- mta_tax: double (nullable = true)
   |-- tip_amount: double (nullable = true)
       tolls_amount: double (nullable = true)
   |-- improvement_surcharge: double (nullable = true)
   %spark val trips_union_DF = trips01_DF.union(trips02_DF).toDF
                                                                                                                                                                    READY
  trips_union_DF: org.apache.spark.sql.DataFrame = [VendorID: int, tpep_pickup_datetime: timestamp ... 16 more fields]
                                                                                                                                                                    READY
      trips_DF = trips_union_DF.withColumn("year", year(to_date($"tpep_pickup_datetime")))
.withColumn("month", month(to_date($"tpep_pickup_datetime")))
  trips_DF: org.apache.spark.sql.DataFrame = [VendorID: int, tpep_pickup_datetime: timestamp ... 18 more fields]
  %spark
trips_DF.agg(min("passenger_count"),max("passenger_count"), min("trip_distance"),max("trip_distance"),min("total_amount"),max("total_amount")).show()
                                                                                                                                                                    READY
  +------
  |min(passenger_count)|max(passenger_count)|min(trip_distance)|max(trip_distance)|min(total_amount)|max(total_amount)|
  | 0| 9| -30.62| 210240.07| -1242.3|
   READY
  trips_clean_DF: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [VendorID: int, tpep_pickup_datetime: timestamp ... 18 more fields]
                                                                                                                                                                    READY
   trips_clean_DF.agg(min("passenger_count"), max("passenger_count"), min("trip_distance"), max("trip_distance"), min("total_amount"), max("total_amount")).show()
  |min(passenger count)|max(passenger count)|min(trip distance)|max(trip distance)|min(total amount)|max(total amount)|
                                                        1.0
                                                                     369.94
                   1
                                        9|
                                                                                        0.0
                                                                                                       6061.42
                                                                                                                                                                    READY
   READY
```

hdfs dfs -ls -R -h /nycdata

```
7.3 M 2022-06-16 16:48 /nycdata/year=2020/month=1/part-00004-d2514daa-c589-4357-a932-224eb3716069.c000.snappy.parquet
   -rw-r--r--
                   1 root supergroup
                                               10.7 K 2022-06-16 16:49 /nycdata/year=2020/month=1/part-00005-d2514daa-c589-4357-a932-224eb3716069.c000.snappy.parquet
   -rw-r--r--
                   1 root supergroup
                                                     0 2022-06-16 16:53 /nycdata/year=2020/month=2
   drwxr-xr-x
                     root supergroup
                                                4.3 K 2022-06-16 16:44 /nycdata/year=2020/month=2/part-00001-d2514daa-c589-4357-a932-224eb3716069.c000.snappy.parquet
   -rw-r--r--
                   1 root supergroup
                                                4.4 K 2022-06-16 16:46 /nycdata/year=2020/month=2/part-00002-d2514daa-c589-4357-a932-224eb3716069.c000.snappy.parquet 5.3 K 2022-06-16 16:48 /nycdata/year=2020/month=2/part-00004-d2514daa-c589-4357-a932-224eb3716069.c000.snappy.parquet
   -rw-r--r--
                   1 root supergroup
   -rw-r--r--
                   1 root supergroup
   -rw-r--r--
                  1 root supergroup
                                               20.1 M 2022-06-16 16:49 /nycdata/year=2020/month=2/part-00005-d2514daa-c589-4357-a932-224eb3716069.c000.snappy.parquet
                                               20.0 M 2022-06-16 16:50 /nycdata/year=2020/month=2/part-00006-d2514daa-c589-4357-a932-224eb3716069.c000.snappy.parquet
   -rw-r--r--
                  1 root supergroup
                  1 root supergroup
   -rw-r--r--
                                               20.1 M 2022-06-16 16:52 /nycdata/year=2020/month=2/part-00007-d2514daa-c589-4357-a932-224eb3716069.c000.snappy.parquet
   -rw-r--r--
                                               20.2 M 2022-06-16 16:53 /nycdata/year=2020/month=2/part-00008-d2514daa-c589-4357-a932-224eb3716069.c000.snappy.parquet
                  1 root supergroup
    %spark
trips_DF.count()
                                                                                                                                                                                                                       READY
   res10: Long = 12704362
    %spark
trips_clean_DF.count()
                                                                                                                                                                                                                        READY
   res11: Long = 9103468
   %sh
hdfs dfs -get /nycdata/year=2020/month=1/part-00000-d2514daa-c589-4357-a932-224eb3716069.c000.snappy.parquet /home/cloudera/Downloads/102.parquet
                                                                                                                                                                                                                       READY
    %sh
                                                                                                                                                                                                                       READY
    parquet-tools cat --json /home/cloudera/Downloads/102.parquet
   {"VendorID":1,"tpep_pickup_datetime":"AHakK7wbAADihCUA","tpep_dropoff_datetime":"ALbHOf8bAADihCUA","passenger_count":1,"trip_distance":1.2,"RatecodeID":1,"store_and_fwd_flag":"N","PULocationID":238, "DOLocationID":239, "payment_type":1,"fare_amount":6.0, "extra":3.0, "mta_tax":0.5, "tip_amount":1.47, "tolls_amount":0.0, "improvement_surcharge":0.3, "t
   otal_amount":11.27,"congestion_surcharge":2.5}
   {"VendorID":1,"tpep_pickup_datetime":"AMMajCMcADihCUA","tpep_dropoff_datetime":"APArKIscAADihCUA","passenger_count":1,"trip_distance":1.2,"RatecodeID":1,"store_and_fwd_flag":"N","PULocationID":239,"DOLocationID":238,"payment_type":1,"fare_amount":7.0,"extra":3.0,"mta_tax":0.5,"tip_amount":1.5,"tolls_amount":0.0,"improvement_surcharge":0.3,"to
  g: N, PolocationID: 1239, DolocationID: 1238, payment_type: 1, fare_amount: 7.0, extra: 3.0, mta_tax: 0.5, tip_amount: 1.5, tolis_amount: 0.0, improvement_surcharge: 0.3, to tal_amount: 11.5, "tope_pickup_datetime: "AJa/bDQdAADihCUA", "tpep_dropoff_datetime: "AB6mDCAeAADihCUA", "passenger_count: 2, "trip_distance: 2.4, "RatecodeID: 1, "store_and_fwd_flag": "N", "PULocationID: 1246, "DOLocationID: 79, "payment_type: 1, "fare_amount: 12.0, "extra: 3.0, "mta_tax: 0.5, "tip_amount: 1.75, "tolls_amount: 0.0, "improvement_surcharge: 0.3, "total_amount: 17.55, "congestion_surcharge: 2.5}
   {"VendorID":1,"tpep_pickup_datetime":"AI6zHkYdAADihCUA","tpep_dropoff_datetime":"AJDyUqceAADihCUA","passenger_count":1,"trip_distance":3.3,"RatecodeID":1,"store_and_fwd_fla
   g":"N", "PULocationID":161, "DOLocationID":144, "payment_type":1, "fare_amount":17.0, "extra":3.0, "mta_tax":0.5, "tip_amount":4.15, "tolls_amount":0.0, "improvement_surcharge":0.3, "total_amount":24.95, "congestion_surcharge":2.5}
   {"VendorID":2, "tpep_pickup_datetime": "ANRFdmMbADihCUA", "tpep_dropoff_datetime": "AL4J7bEbAADihCUA", "passenger_count":1, "trip_distance":1.07, "RatecodeID":1, "store_and_fwd_flag": "N", "PULocationID":43, "DOLocationID":239, "payment_type":1, "fare_amount":6.0, "extra":0.5, "mta_tax":0.5, "tip_amount":1.96, "tolls_amount":0.0, "improvement_surcharge":0.3, "to
   tal_amount":11.76, "congestion_surcharge":2.5}
{"VendorID":2, "tpep_pickup_datetime": "ANr2m@QcAADihCUA", "tpep_dropoff_datetime": "AFpeJk4eAADihCUA", "passenger_count":1, "trip_distance":7.76, "RatecodeID":1, "store_and_fwd_fla
   g":"N","PULocationID":143,"DOLocationID":25,"payment_type":1,"fare_amount":28.5,"extra":0.5,"mta_tax":0.5,"tip_amount":4.84,"tolls_amount":0.0,"improvement_surcharge":0.3,"t
   Paragraph received a SIGTERM
   ExitValue: 143
                                                                                                                                                                                                                       READY
    parquet-tools head -n 2 /home/cloudera/Downloads/102.parquet
   tpep pickup datetime = AHakK7wbAADihCUA
   tpep_dropoff_datetime = ALbHOf8bAADihCUA
   passenger_count = 1
   trip_distance = 1.2
   RatecodeID = 1
   store and fwd flag = N
   PULocationID = 238
   DOLocationID = 239
  payment_type = 1
   fare amount = 6.0
   extra = 3.0
   mta_tax = 0.5
   tip_amount = 1.47
   tolls_amount = 0.0
   improvement surcharge = 0.3
   total_amount = 11.27
                                                                                                                                                                                                                       READY
    parquet-tools schema /home/cloudera/Downloads/102.parquet
   message spark_schema {
     optional int32 VendorID:
      optional int96 tpep_pickup_datetime;
     optional int96 tpep_dropoff_datetime;
     optional int32 passenger_count;
     optional double trip_distance; optional int32 RatecodeID;
     optional binary store_and_fwd_flag (UTF8); optional int32 PULocationID;
      optional int32 DOLocationID;
     optional int32 payment_type;
      optional double fare_amo
     optional double extra;
      optional double mta_tax;
     optional double tip_amount;
optional double tolls_amount;
      optional double improvement_surcharge;
    %spark
                                                                                                                                                                                                                       READY
```

-rw-r--r--

lookup\_DF.printSchema()

1 root supergroup

- root supergroup

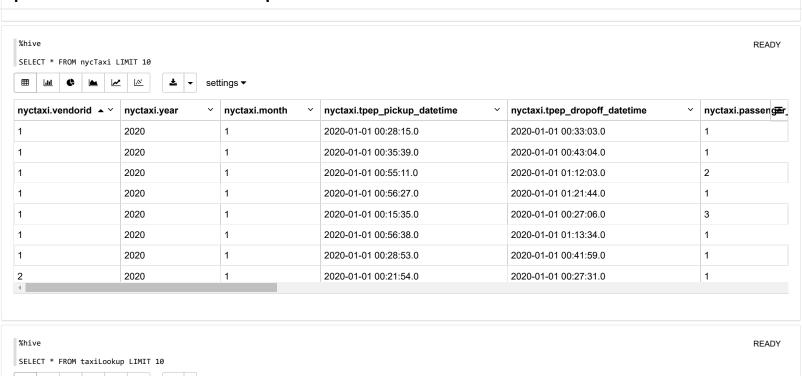
0 2022-06-16 16:53 /nycdata/ SUCCESS

0 2022-06-16 16:53 /nycdata/year=2020

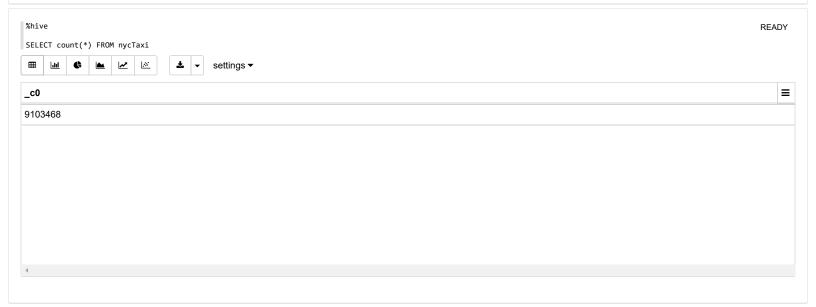
```
|-- LocationID: integer (nullable = true)
|-- Borough: string (nullable = true)
|-- Borough: string (nullable = true)
|-- Borough: string (nullable = true) |
    +-----
      LocationID| Borough| Zone|service_zone|
    |LocationID|
                                  EWR |
                                               Newark Airport
                  1|
                                                                                      EWR|
                  2|
                              Queens|
                                                      Jamaica Bay|
                                                                              Boro Zone
                                 Bronx|Allerton/Pelham G...|
                                                                              Boro Zone
                  3
                         Manhattan
                                 hattan| Alphabet City| Yellow Zone|
Island| Arden Heights| Boro Zone|
                 5|Staten Island|
   only showing top 5 rows
                    ... ....b. ....b. ... Nakara... fillestanto, del nacionel akada...
    %spark
lookup_DF.write
.mode("overwrite")
.parquet("/nyclookup")
                                                                                                                                                                                                                                                              READY
    %sh
                                                                                                                                                                                                                                                              READY
    hdfs dfs -ls -R -h /nyclookup/
    -rw-r--r-- 1 root supergroup
-rw-r--r-- 1 root supergroup
                                                               0 2022-06-17 12:17 /nyclookup/_SUCCESS
                                                         5.6 K 2022-06-17 12:17 /nyclookup/part-00000-2345d162-0606-4b62-b43a-59233588d9d6-c000.snappy.parquet
    %hive
                                                                                                                                                                                                                                                              READY
    SET hive.execution.engine = spark
   Query executed successfully. Affected rows : -1
     %hive
                                                                                                                                                                                                                                                              READY
     CREATE DATABASE rides
    %hive
                                                                                                                                                                                                                                                              READY
    USE rides
   Ouery executed successfully. Affected rows : -1
    %hive
                                                                                                                                                                                                                                                              READY
    drop table nycTaxi
   Query executed successfully. Affected rows : -1
     %hive
                                                                                                                                                                                                                                                              READY
    drop table taxiLookup
   Query executed successfully. Affected rows : -1
    %hive
                                                                                                                                                                                                                                                              READY
     \verb|create| external table nycTaxi(|\\
           ate external table nycTaxi(
VendorID int,
tpep_pickup_datetime timestamp,
tpep_dropoff_datetime timestamp,
passenger_count int,
trip_distance double,
RatecodeID int,
store_and_fwd_flag string,
PULocationID int,
DOLocationID int,
nawment type int
    DOLocationID int,
payment_type int,
fare_amount double,
extra double,
mta_tax double,
tip_amount double,
tip_amount double,
improvement_surcharge double,
total_amount double,
congestion_surcharge double)
partitioned by (year int, month int)
stored as parquet
     stored as parquet
location '/nycdata/'
TBLPROPERTIES ("parquet.compression"="SNAPPY")
   Query executed successfully. Affected rows : -1
     %hive
                                                                                                                                                                                                                                                              READY
    create external table taxiLookup(
   LocationID int,
   Borough string,
   Zone string,
   service_zone string)
stored as parquet
location '/nyclookup/'
TBLPROPERTIES ("parquet.compression"="SNAPPY")
   Query executed successfully. Affected rows : -1
     %hive
                                                                                                                                                                                                                                                              READY
    MSCK REPAIR TABLE nycTaxi
```

Query executed successfully. Affected rows : -1

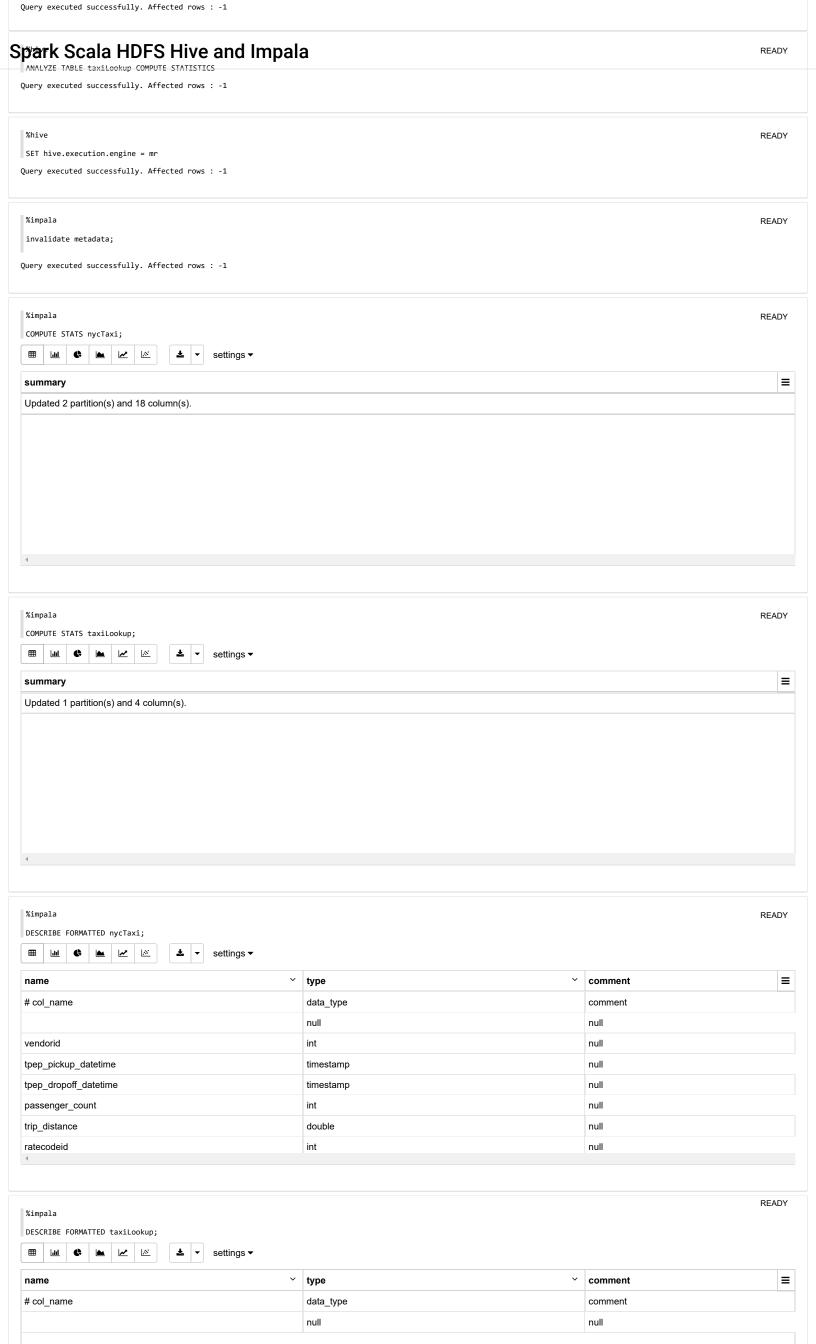
## MSCK REPAIR TABLE taxilookup SparkeScalasHDFAScHive:and Impala

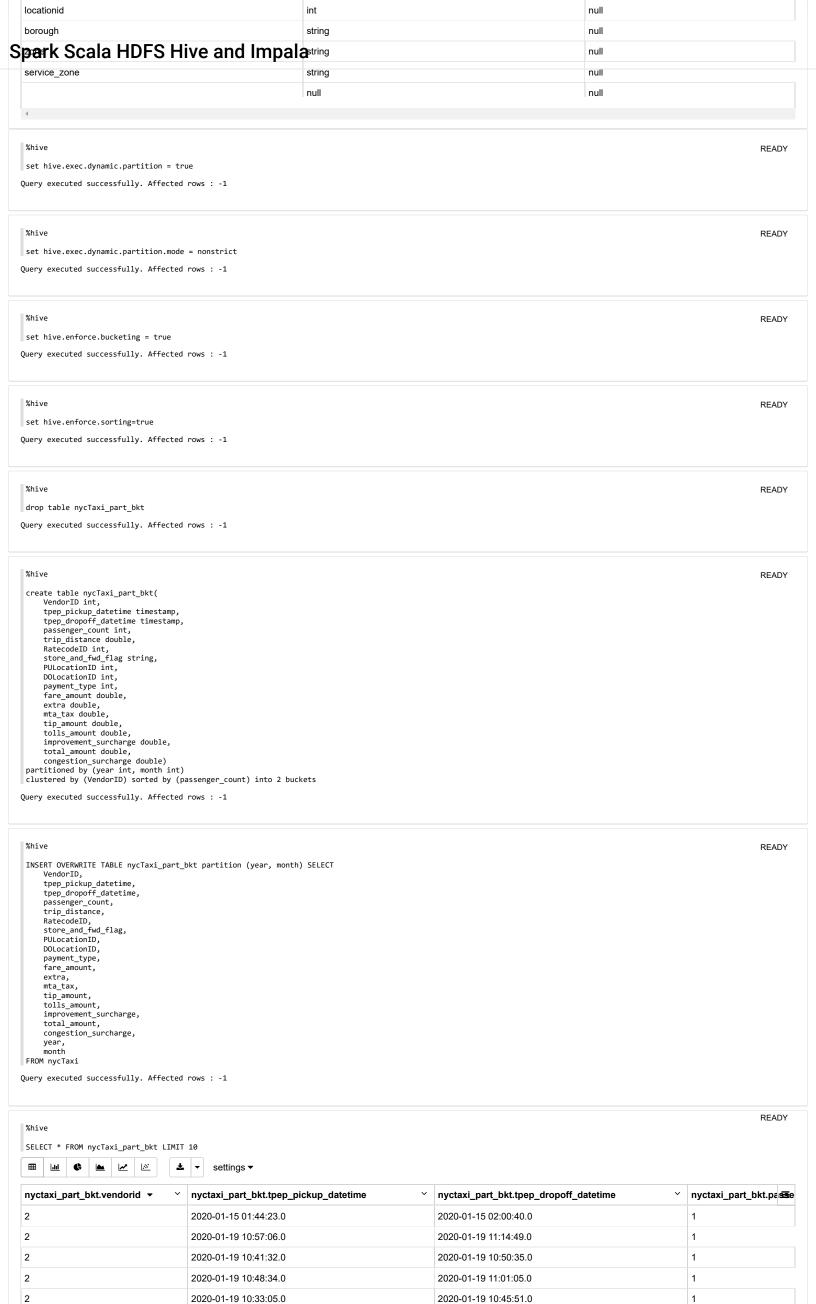






%nive	READY
SELECT count(*) FROM taxiLookup	
■ Lul C Le Le settings ▼	
_c0	=
265	

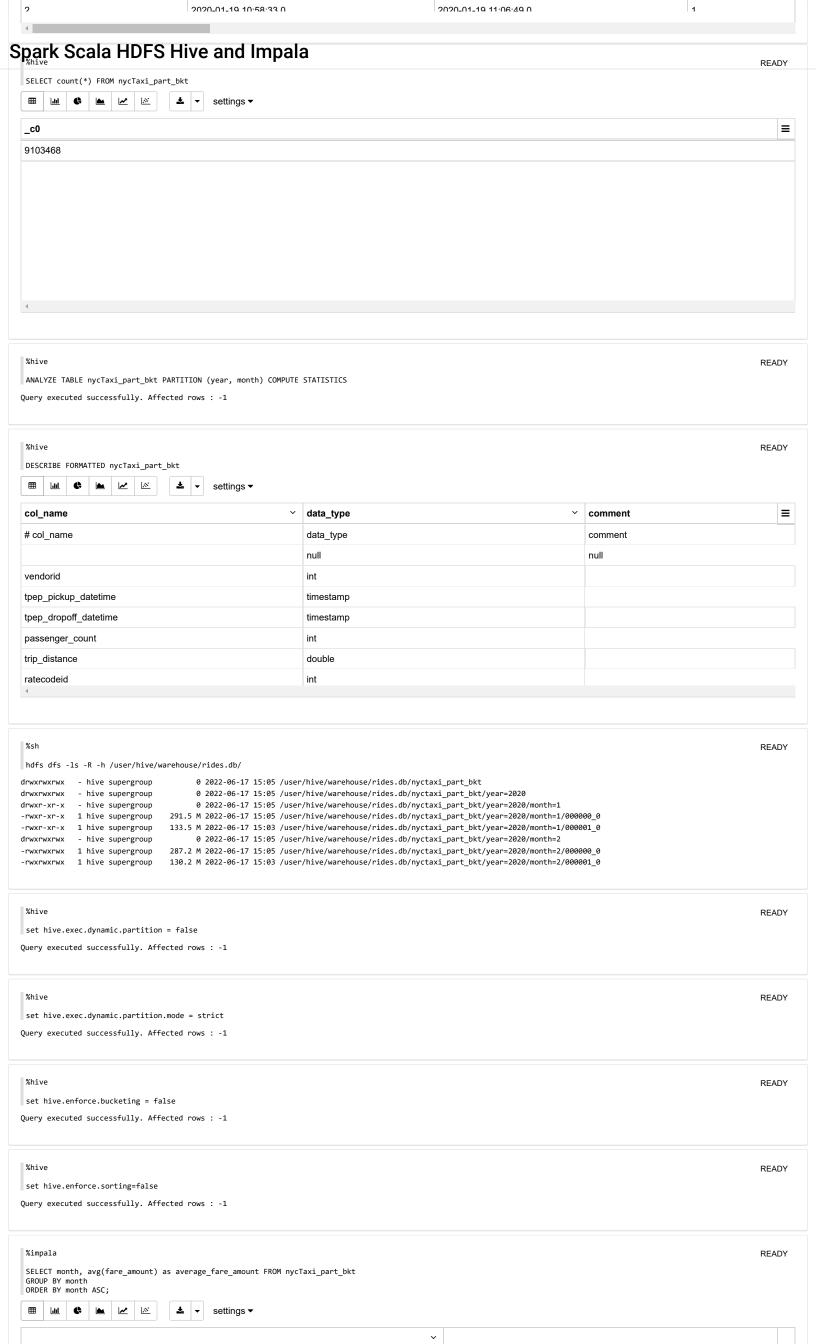


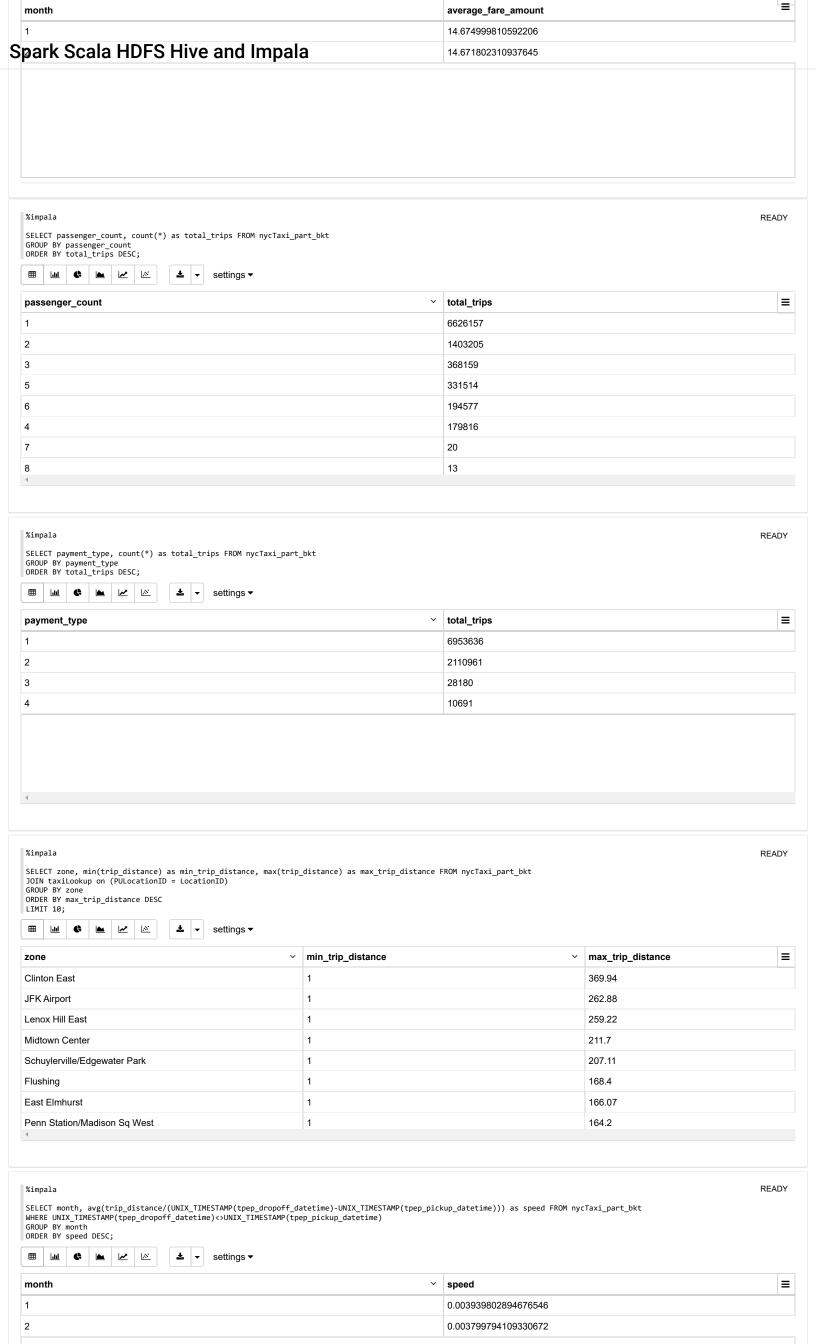


2020-01-19 10:13:21.0

2

2020-01-19 10:04:28.0





Spark Scala HDFS Hive and Impala	
4	
%impala	READY