Data Loading Script

```
use bus_transport;
```

INTERNAL TABLES:

Tables stored as ORC or PARQUET needs to be loaded with usage of other tables stored as text file – as we are loading data from .txt files

```
CREATE TABLE route_txt (
   route_id INT,
   route_name STRING,
   metrics MAP<STRING, INT>
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
COLLECTION ITEMS TERMINATED BY '|'
MAP KEYS TERMINATED BY ':'
STORED AS TEXTFILE;

LOAD DATA INPATH 'hdfs:///user/mmajewska/Route_2020.txt' overwrite INTO TABLE route_txt;
INSERT OVERWRITE table route SELECT * FROM route_txt;
drop table route_txt;



CREATE TABLE junk_txt (
   junk_id INT,
   satisfaction_level_category STRING,
   occupation_level_category STRING
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE;
```

```
LOAD DATA INPATH 'hdfs:///user/mmajewska/Junk.txt' overwrite INTO TABLE junk_txt;

INSERT OVERWRITE table junk SELECT * FROM junk_txt;

drop table junk_txt;
```

EXTERNAL TABLE:

Placing file Service_Office_2020.txt in directory hdfs:///user/mmajewska/database

STATIC PARTITIONING:

Here the same situation as for internal tables - we are creating tables for .txt files (table format is ORC)

```
CREATE TABLE bus_txt (

  bus_id INT,

   bus_registration STRING,

   bus_office_id INT,

  additional_equipment ARRAY<STRING>,

  bus_type STRING

)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

COLLECTION ITEMS TERMINATED BY '|'

STORED AS TEXTFILE;
```

```
LOAD DATA LOCAL INPATH 'Bus_low_floor.txt' INTO TABLE bus_txt;

INSERT OVERWRITE TABLE bus PARTITION (bus_type='low floor')

SELECT bus_id, bus_registration, bus_office_id, additional_equipment FROM bus_txt;

TRUNCATE TABLE bus_txt; -- to clean temporary table
```

```
LOAD DATA LOCAL INPATH 'Bus_standard.txt' INTO TABLE bus_txt;

INSERT INTO TABLE bus PARTITION (bus_type='standard')

SELECT bus_id, bus_registration, bus_office_id, additional_equipment FROM bus_txt;

TRUNCATE TABLE bus_txt;


LOAD DATA LOCAL INPATH 'Bus_minibus.txt' INTO TABLE bus_txt;

INSERT INTO TABLE bus PARTITION (bus_type='minibus')

SELECT bus_id, bus_registration, bus_office_id, additional_equipment FROM bus_txt;

drop table bus_txt;
```

DYNAMIC PARTITIONING:

```
set hive.exec.dynamic.partition=true;

set hive.exec.dynamic.partition.mode=nonstrict;
```

Creation of temporary tables – to load all data from file (including partitioning field), need to change storing format to PARQUET or ORC

```
CREATE TABLE date_tmp (

   date_id INT,

   date_format DATE ,

   year INT,

   month STRING,

   month_no INT,

  day_type STRING

)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

STORED AS TEXTFILE;


LOAD DATA INPATH 'hdfs:///user/mmajewska/Date.txt' overwrite INTO TABLE date_tmp;

INSERT OVERWRITE TABLE date_dim partition(month_no, day_type)
```

```
SELECT date_id, date_format, year, month, month_no, day_type FROM date_tmp;
drop table date_tmp;

CREATE TABLE time_tmp (
    time_id INT,
    hour INT,
    minutes INT,
   time_of_day STRING
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE;

LOAD DATA INPATH 'hdfs:///user/mmajewska/Time.txt' overwrite INTO TABLE time_tmp;
INSERT OVERWRITE TABLE time_dim partition(time_of_day)
SELECT time_id, hour, minutes, time_of_day FROM time_tmp;
drop table time_tmp;

CREATE TABLE travel_tmp (
    bus_id INT,
    route_id INT,
   departure_time INT,
    arrival_time INT,
   tickets_validated INT,
   bus_capacity INT,
   avg_satisfaction_level_received INT,
   satisfaction_surveys_number INT,
   junk_id INT,
  travel_date INT
)
ROW FORMAT DELIMITED
```

```
FIELDS TERMINATED BY ','

STORED AS TEXTFILE;


LOAD DATA INPATH 'hdfs:///user/mmajewska/Travel_2020.txt' overwrite INTO TABLE travel_tmp;

INSERT OVERWRITE TABLE travel partition(travel_date)

SELECT bus_id, route_id, departure_time, arrival_time , tickets_validated,
bus_capacity, avg_satisfaction_level_received, satisfaction_surveys_number, junk_id, travel_date
FROM travel_tmp;

drop table travel_tmp;
```