Use NYC Taxi Trip Record Data To Calculate Performance KPIs

Practical Exam

Goal

NYC.gov provides monthly exports of NYC yellow taxi trip records:

- https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page
- Latest Full Dumps:
    - *https://nyc-tlc.s3.amazonaws.com/trip+data/yellow_tripdata_2020-12.csv*
    - *https://nyc-tlc.s3.amazonaws.com/trip+data/yellow_tripdata_2020-11.csv*
    - *https://nyc-tlc.s3.amazonaws.com/trip+data/yellow_tripdata_2020-10.csv*
    - …

```
VendorID,tpep_pickup_datetime,tpep_dropoff_datetime,passenger_count,trip_distance,RatecodeID,store_and_fwd_flag,PULocationID,DOLocationID,payment_type,fare_amount,extra,mta_tax,tip_amount,tolls_amount,improvement_surcharge,total_amount,congestion_surcharge
1,2020-12-01 00:07:13,2020-12-01 00:18:12,1,7.60,1,N,138,263,1,21.5,3,0.5,2.5,6.12,0.3,33.92,2.5
1,2020-12-01 00:41:19,2020-12-01 00:49:45,1,1.60,1,N,140,263,1,8,3,0.5,2.95,0,0.3,14.75,2.5
2,2020-12-01 00:33:40,2020-12-01 01:00:35,1,16.74,2,N,132,164,1,52,0,0.5,2.5,6.12,0.3,63.92,2.5
2,2020-12-01 00:02:15,2020-12-01 00:13:09,1,4.16,1,N,238,48,1,14,0.5,0.5,1,0,0.3,18.8,2.5
2,2020-12-01 00:37:42,2020-12-01 00:45:11,1,2.22,1,N,238,41,2,8.5,0.5,0.5,0,0,0.3,9.8,0
1,2020-12-01 00:27:47,2020-12-01 00:45:40,0,8.40,1,N,138,137,1,25,3,0.5,6,6.12,0.3,40.92,2.5
2,2020-12-01 00:40:47,2020-12-01 00:57:03,1,6.44,1,N,132,191,1,19.5,0.5,0.5,4.16,0,0.3,24.96,0
2,2020-12-01 00:01:42,2020-12-01 00:06:06,1,.99,1,N,234,137,1,5.5,0.5,0.5,1.86,0,0.3,11.16,2.5
2,2020-12-01 00:58:24,2020-12-01 01:36:14,2,11.81,1,N,261,7,1,36.5,0.5,0.5,1,0,0.3,41.3,2.5
1,2020-12-01 00:08:15,2020-12-01 00:16:04,2,2.70,1,N,237,107,1,9.5,3,0.5,2.65,0,0.3,15.95,2.5
2,2020-12-01 00:04:21,2020-12-01 00:29:00,1,6.28,1,N,41,68,2,23,0.5,0.5,0,0,0.3,26.8,2.5
[…]
```

yellow_tripdata_2020-12.csv

Goal

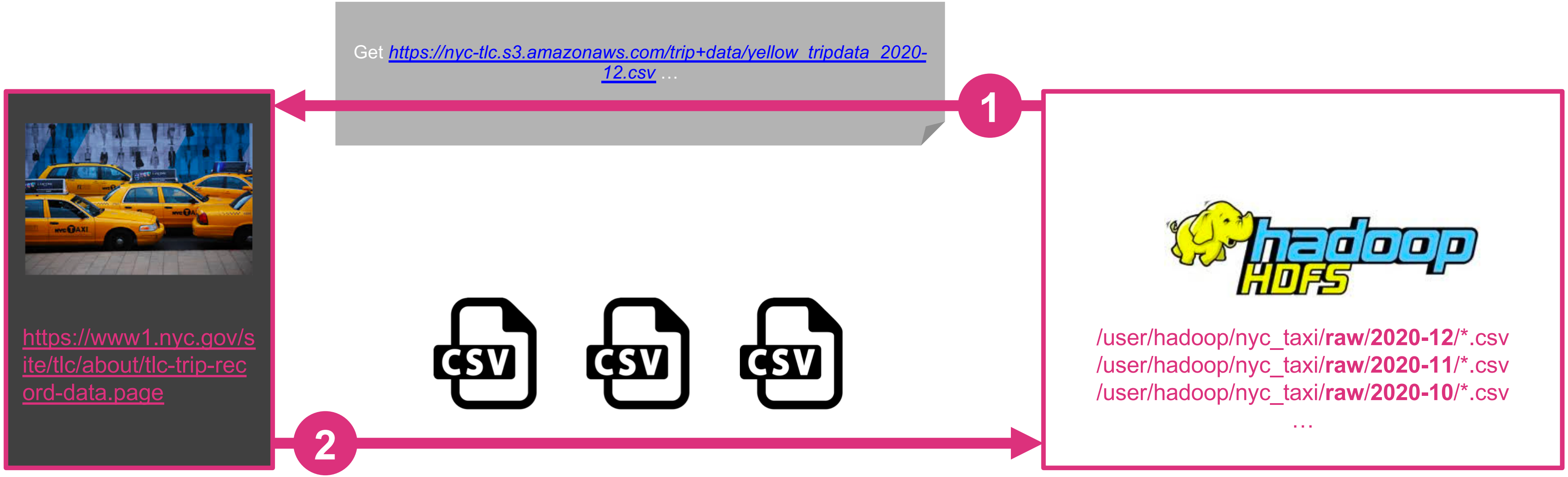
We want to make use of this data to calculate some KPIs

Workflow:

- **Gather data** from https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page
- **Save raw data** (*CSV files*) to HDFS (partitioned by *YYYY-MM*)
- **Optimize**, **reduce** and **clean raw data** and save it to **final** directory on HDFS
- **Calculate KPIs** and **Export** them to an **Excel File**

- The whole data workflow **must be implemented** within an ETL **workflow tool** (e.g. Pentaho Data Integration or Airflow) and **run automatically**

Dataflow: **1. Get TLC NYC Taxi Data**

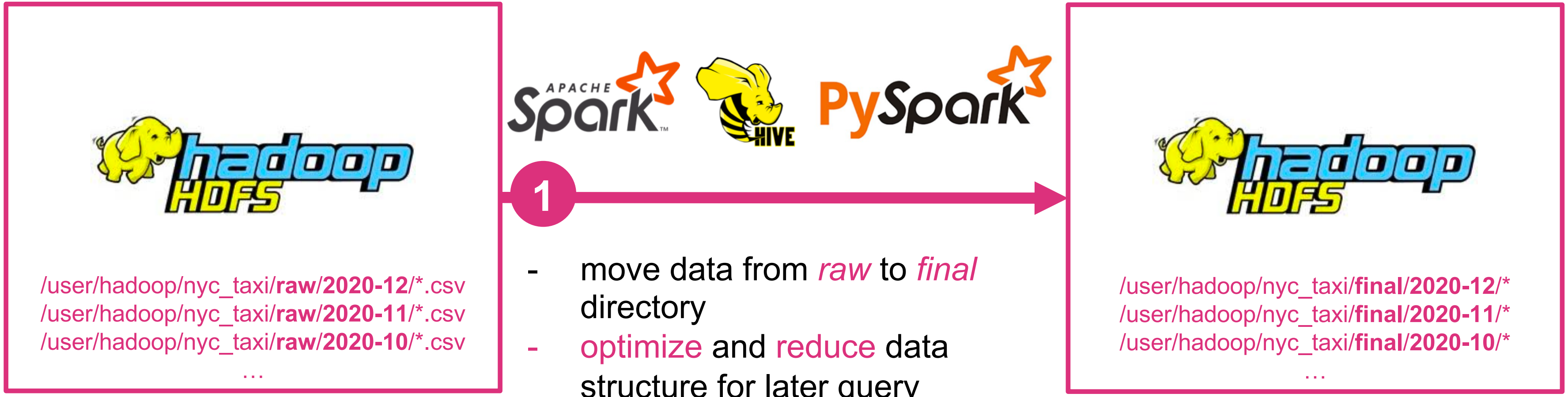


Get https://nyc-tlc.s3.amazonaws.com/trip+data/yellow_tripdata_2020-12.csv …

https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page

/user/hadoop/nyc_taxi/**raw**/**2020-12**/*.csv
/user/hadoop/nyc_taxi/**raw**/**2020-11**/*.csv
/user/hadoop/nyc_taxi/**raw**/**2020-10**/*.csv
…

Dataflow: **2. Raw To Final Transfer**



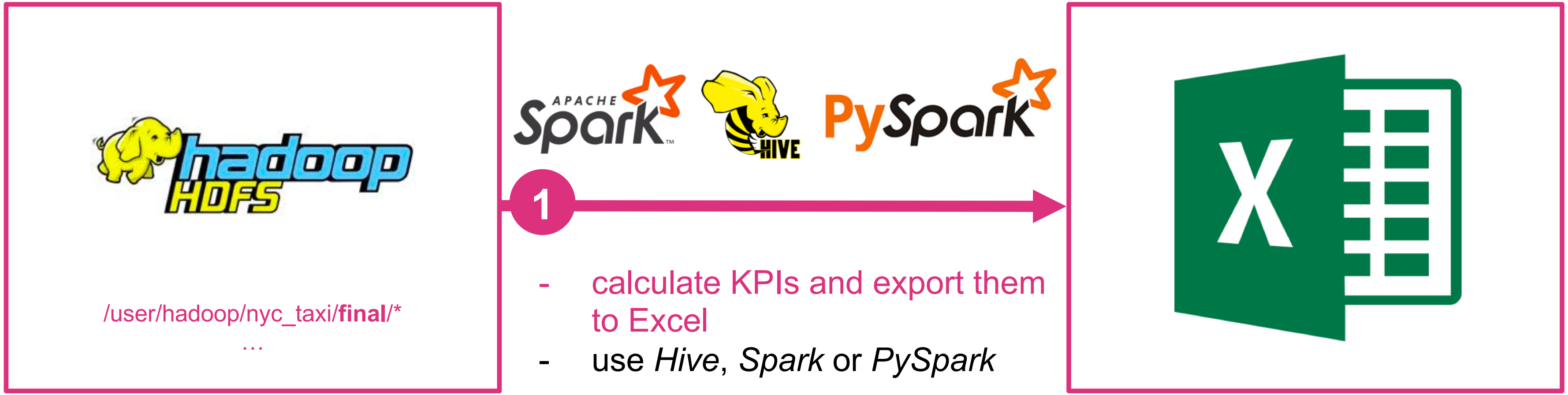
/user/hadoop/nyc_taxi/raw/2020-12/*.csv
/user/hadoop/nyc_taxi/raw/2020-11/*.csv
/user/hadoop/nyc_taxi/raw/2020-10/*.csv
…

1

/user/hadoop/nyc_taxi/final/2020-12/*
/user/hadoop/nyc_taxi/final/2020-11/*
/user/hadoop/nyc_taxi/final/2020-10/*
…

- move data from *raw* to *final* directory
- optimize and reduce data structure for later query purposes if necessary
- remove duplicates if necessary
- …

Dataflow: **3. Calculate And Export KPIs**



/user/hadoop/nyc_taxi/**final**/*
…

- calculate KPIs and export them to Excel
- use *Hive*, *Spark* or *PySpark*

Dataflow: **4. KPIs To Calculate**

Calculate per Month:

- Average Trip Duration (in minutes)
- Average Trip Distance (in miles)
- Average total amount (in USD)
- Average tip amount (in USD)
- Average passenger count (as Number)
- Usage Share by payment type (credit card, cash… in percent)
- Usage share per timeslot (in percent):
    - 00:00-06:00
    - 06:00-12:00
    - 12:00-18:00
    - 18:00-24:00