

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://d37ci6vzurychx.cloudfront.net/trip-data/yellow_tripdata_2022-01.parquet
 - https://d37ci6vzurychx.cloudfront.net/trip-data/yellow_tripdata_2022-02.parquet
 - https://d37ci6vzurychx.cloudfront.net/trip-data/yellow_tripdata_2022-03.parquet
 - ...

```
VendorID tpep_pickup_datetime tpep_dropoff_datetime passenger_count trip_distance RatecodeID store_and_fwd_flag ... mta_tax tip_amount tolls_amount improvement_surcharge total_amount congestion_surcharge airport_fee
0 1 2022-01-01 00:35:40 2022-01-01 00:53:29 2.0 3.80 1.0 N ... 0.5 3.65 0.0 0.3 21.95 2.5 0.0
1 1 2022-01-01 00:33:43 2022-01-01 00:42:07 1.0 2.10 1.0 N ... 0.5 4.00 0.0 0.3 13.30 0.0 0.0
2 2 2022-01-01 00:53:21 2022-01-01 01:02:19 1.0 0.97 1.0 N ... 0.5 1.76 0.0 0.3 10.56 0.0 0.0
3 2 2022-01-01 00:25:21 2022-01-01 00:35:23 1.0 1.09 1.0 N ... 0.5 0.00 0.0 0.3 11.80 2.5 0.0
4 2 2022-01-01 00:36:48 2022-01-01 01:14:20 1.0 4.30 1.0 N ... 0.5 3.00 0.0 0.3 30.30 2.5 0.0
...
2463926 2 2022-01-31 23:36:53 2022-01-31 23:42:51 NaN 1.32 NaN None ... 0.5 2.39 0.0 0.3 13.69 NaN NaN
2463927 2 2022-01-31 23:44:22 2022-01-31 23:55:01 NaN 4.19 NaN None ... 0.5 4.35 0.0 0.3 24.45 NaN NaN
2463928 2 2022-01-31 23:39:00 2022-01-31 23:50:00 NaN 2.10 NaN None ... 0.5 2.00 0.0 0.3 16.52 NaN NaN
2463929 2 2022-01-31 23:36:42 2022-01-31 23:48:45 NaN 2.92 NaN None ... 0.5 0.00 0.0 0.3 15.70 NaN NaN
2463930 2 2022-01-31 23:46:00 2022-02-01 00:13:00 NaN 8.94 NaN None ... 0.5 6.28 0.0 0.3 35.06 NaN NaN
```

[2463931 rows x 19 columns]
[...]

yellow_tripdata_2022-01.parquet



www.marcel-mittelstaedt.com

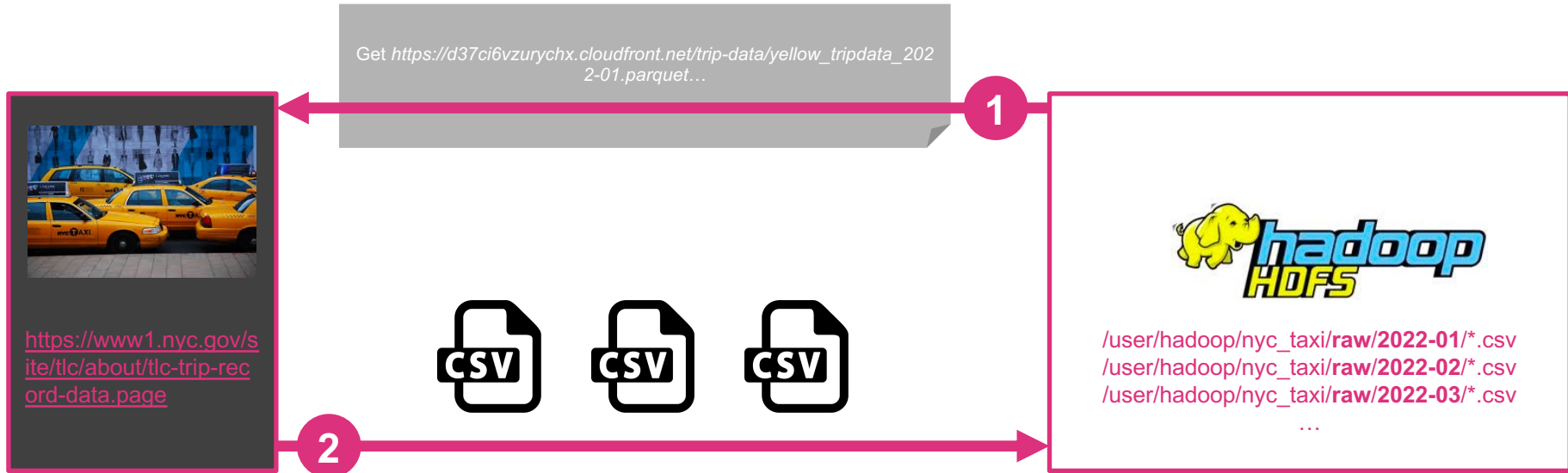
Goal

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

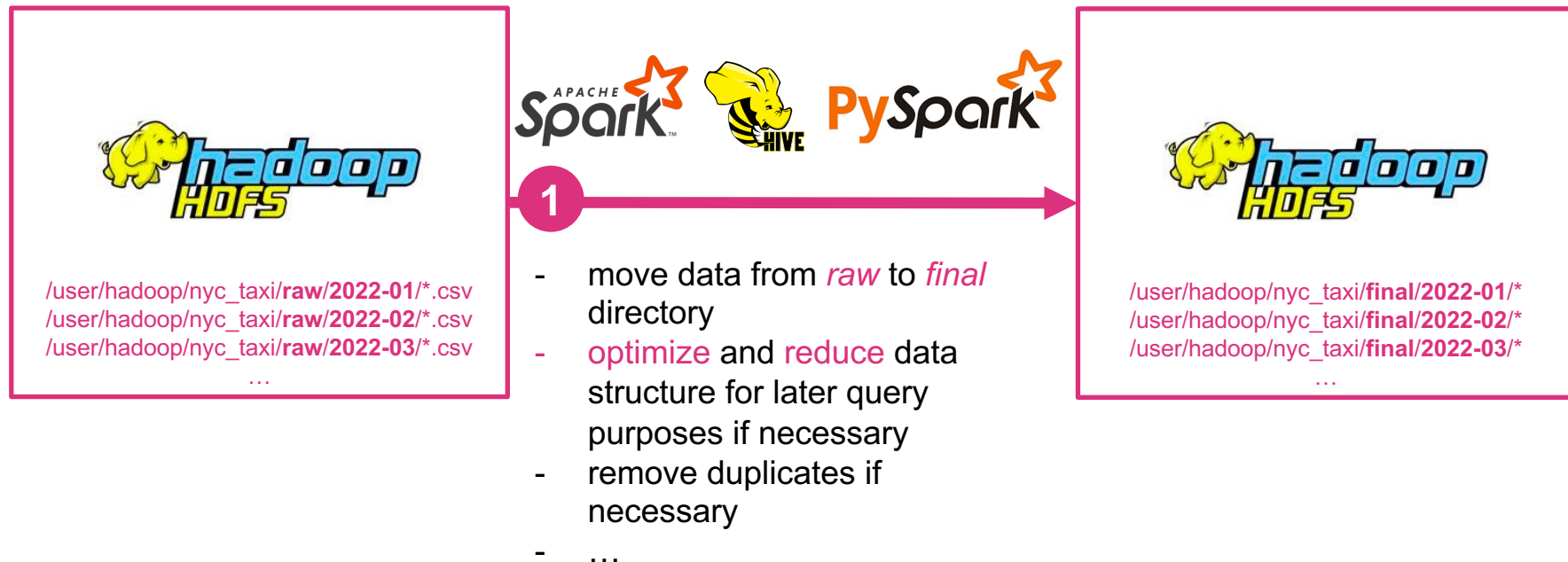
Workflow:

- **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



Dataflow: 2. Raw To Final Transfer



Dataflow: 3. Calculate And Export KPIs



/user/hadoop/nyc_taxi/final/*

...



1

- 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