Use NYC Taxi Trip Record Data To Calculate Performance KPIs

Practical Exam

Goal

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

- https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page
- Latest Full Dumps:
  - *https://s3.amazonaws.com/nyc-tlc/trip+data/yellow_tripdata_2019-06.csv*
  - *https://s3.amazonaws.com/nyc-tlc/trip+data/yellow_tripdata_2019-05.csv*
  - *https://s3.amazonaws.com/nyc-tlc/trip+data/yellow_tripdata_2019-04.csv*
  - *…*

```
1,2019-01-01 00:21:05,2019-01-01 00:24:23,1,.50,1,N,41,24,2,4.5,0.5,0.5,0,0,0.3,5.8
1,2019-01-01 00:44:55,2019-01-01 01:03:05,1,2.70,1,N,239,140,2,14,0.5,0.5,0,0,0.3,15.3
1,2019-01-01 00:08:26,2019-01-01 00:14:21,2,.80,1,N,262,141,1,6,0.5,0.5,1,0,0.3,8.3
1,2019-01-01 00:20:22,2019-01-01 00:52:51,1,10.20,1,N,140,257,2,33.5,0.5,0.5,0,0,0.3,34.8
1,2019-01-01 00:09:18,2019-01-01 00:27:06,2,2.50,1,N,246,239,1,12.5,0.5,0.5,2.75,0,0.3,16.55
1,2019-01-01 00:29:29,2019-01-01 00:32:48,3,.50,1,N,143,143,2,4.5,0.5,0.5,0,0,0.3,5.8
1,2019-01-01 00:38:08,2019-01-01 00:48:24,2,1.70,1,N,50,239,1,9,0.5,0.5,2.05,0,0.3,12.35
1,2019-01-01 00:49:29,2019-01-01 00:51:53,1,.70,1,N,239,238,1,4,0.5,0.5,1,0,0.3,6.3
[…]
```

yellow_tripdata_2019-01.csv

Goal

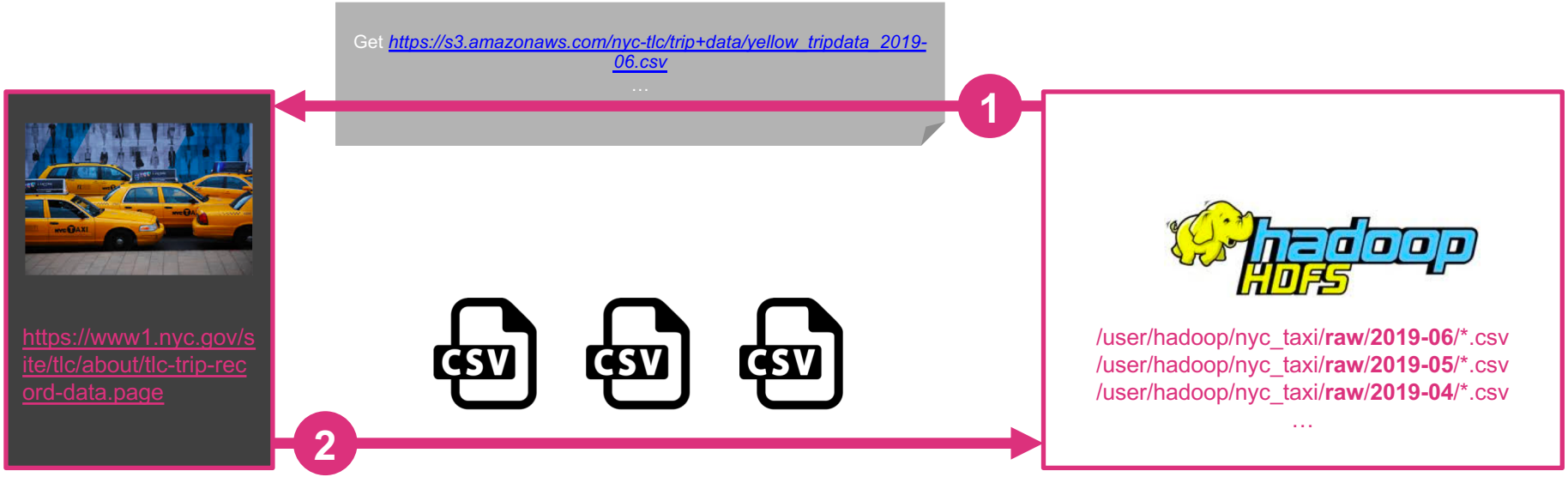
We want to make use of this data to calculate some KPIs

Workflow:

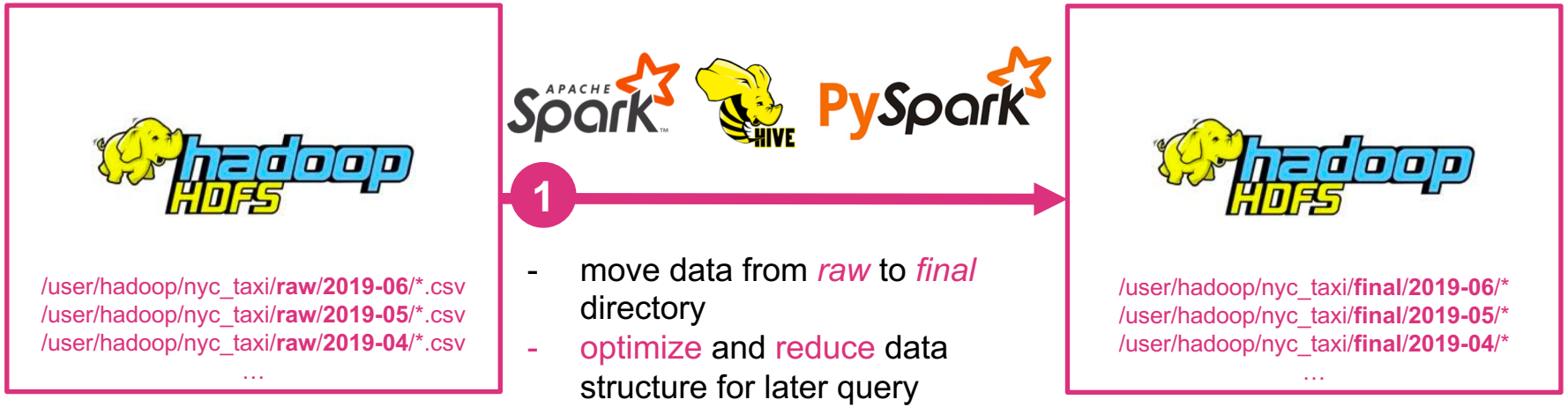
- **Gather data** from https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page
- **Save raw data** (*CSV files*) to HDFS (partitioned by *YYYY-MM*)
- **Optimize**, **reduce** and **clean raw data** and save it to **final** directory on HDFS
- **Calculate KPIs** and **Export** them to an **Excel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

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

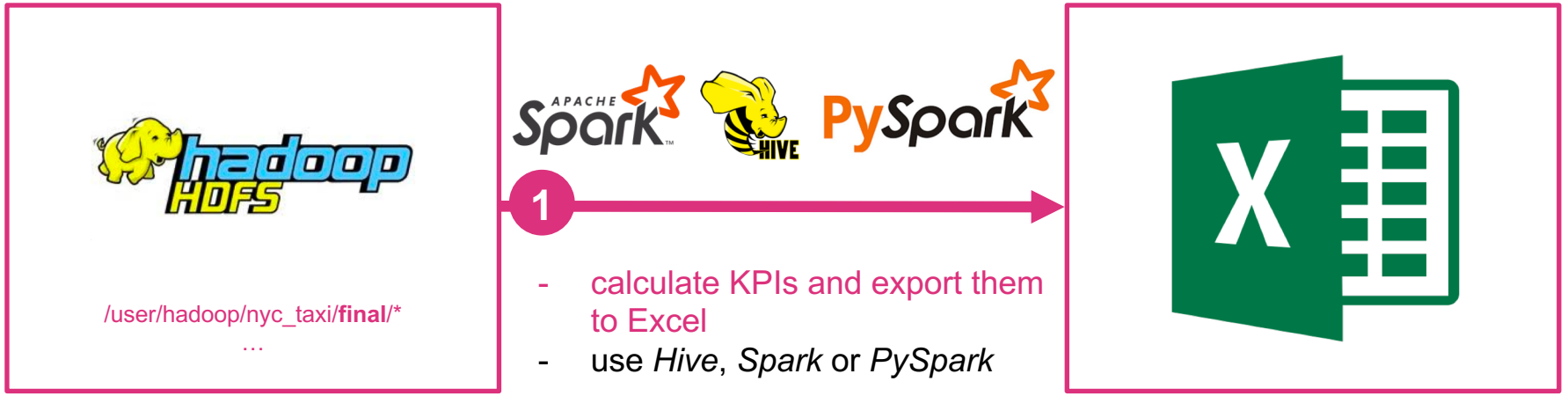
Dataflow: 1. Get TLC NYC Taxi Data



Dataflow: 2. Raw To Final Transfer



Dataflow: 3. Calculate And Export KPIs



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