



Use *kaggle.com* Hubway Data To Calculate Bike Sharing Usage KPIs

Practical Exam



Goal

kaggle.com provides monthly exports of Hubway bike sharing trip records:

- <https://www.bluebikes.com/>
- Latest Full Dumps: <https://www.kaggle.com/acmeyer/hubway-data>

```
"tripduration", "starttime", "stoptime", "start station id", "start station name", "start station latitude", "start station longitude", "end station id", "end station name", "end station latitude", "end station longitude", "bikeid", "usertype", "birth year", "gender"
"133", "2015-12-01 00:01:52", "2015-12-01 00:04:06", "9", "Agganis Arena - 925 Comm Ave.", "42.351246", "-71.115639", "41", "Packard's Comer - Comm. Ave. at Brighton Ave.", "42.352261", "-71.123831", "199", "Customer", "1995", "1"
"1522", "2015-12-01 00:05:30", "2015-12-01 00:30:53", "41", "Packard's Comer - Comm. Ave. at Brighton Ave.", "42.352261", "-71.123831", "54", "Tremont St / West St", "42.354979", "-71.063348", "876", "Customer", "1983", "1"
"153", "2015-12-01 00:07:46", "2015-12-01 00:10:20", "75", "Lafayette Square at Mass Ave / Main St / Columbia St", "42.36346469304347", "-71.10057324171066", "67", "MIT at Mass Ave / Amherst St", "42.3581", "-71.093198", "757", "Subscriber", "1995", "1"
"435", "2015-12-01 00:07:48", "2015-12-01 00:15:04", "68", "Central Square at Mass Ave / Essex St", "42.36507", "-71.1031", "29", "Innovation Lab - 125 Western Ave. at Batten Way", "42.363732", "-71.124565", "853", "Subscriber", "1988", "1"
"1208", "2015-12-01 00:12:15", "2015-12-01 00:32:23", "36", "Boston Public Library - 700 Boylston St.", "42.349673", "-71.077303", "110", "Harvard University Gund Hall at Quincy St / Kirkland S", "42.376369", "-71.114025", "437", "Customer", "1982", "1"
"1117", "2015-12-01 00:16:31", "2015-12-01 00:35:09", "31", "Seaport Hotel", "42.348833", "-71.041747", "67", "MIT at Mass Ave / Amherst St", "42.3581", "-71.093198", "1161", "Subscriber", "1988", "1"
"1287", "2015-12-01 00:16:50", "2015-12-01 00:38:18", "10", "B.U. Central - 725 Comm. Ave.", "42.350406", "-71.108279", "23", "Mayor Martin J Walsh - 28 State St", "42.35892", "-71.057629", "565", "Subscriber", "1966", "1"
```

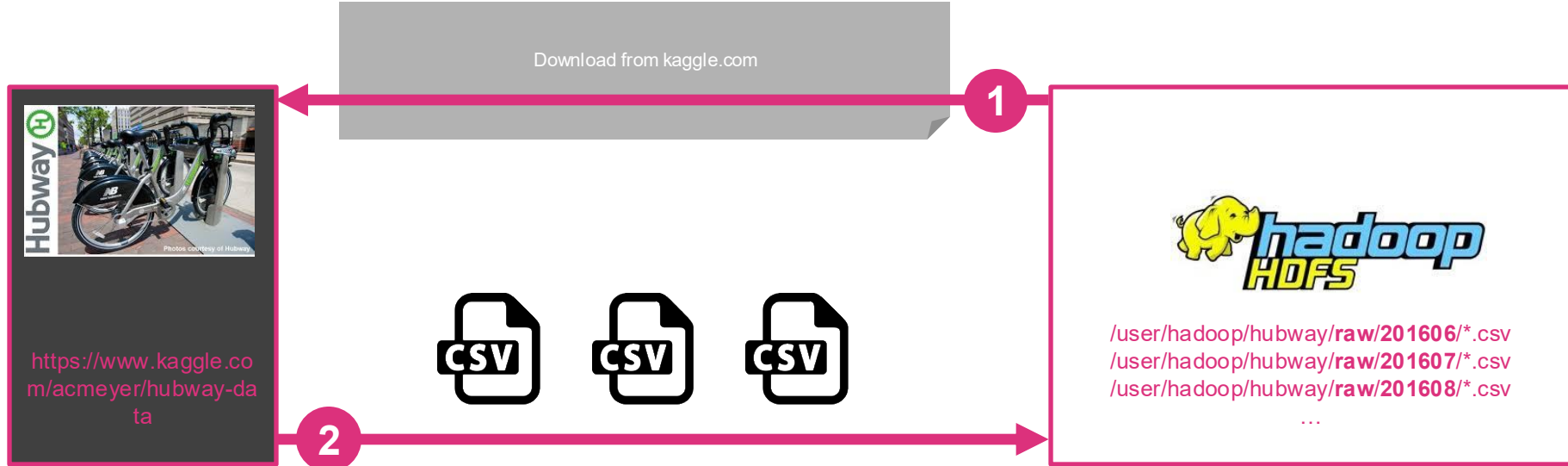
Goal

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

Workflow:

- **Gather data** from <https://www.kaggle.com/acmeyer/hubway-data>
- **Save raw data** (CSV files) to HDFS (partitioned by YYYYMM)
- **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 Hubway Bike Sharing Data



Dataflow: 2. Raw To Final Transfer



/user/hadoop/hubway/**raw**/201606/*.csv
/user/hadoop/hubway/**raw**/201607/*.csv
/user/hadoop/hubway/**raw**/201608/*.csv
...



1

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



/user/hadoop/hubway/**final**/201606/*
/user/hadoop/hubway/**final**/201607/*
/user/hadoop/hubway/**final**/201608/*
...

Dataflow: 3. Calculate And Export KPIs



/user/hadoop/hubway/final/201606/*
/user/hadoop/hubway/final/201607/*
/user/hadoop/hubway/final/201608/*
...



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 km)
- Usage Share by gender (in percent)
- Usage Share by age (in percent)
- Top 10 most used bikes
- Top 10 most start stations
- Top 10 most end stations
- Usage share per timeslot (in percent):
 - 00:00-06:00
 - 06:00-12:00
 - 12:00-18:00
 - 18:00-24:00