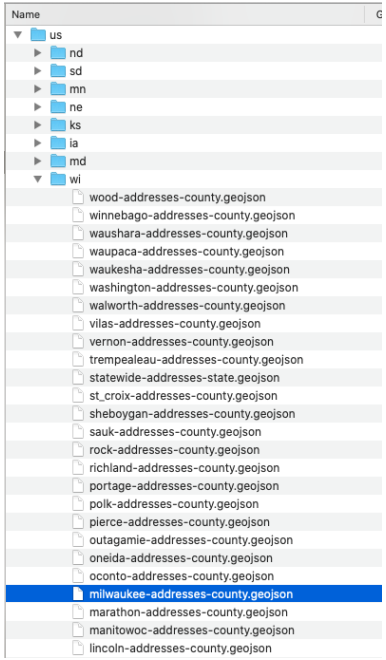
Use OpenAdresses Data To Validate Adresses

Practical Exam

Goal

OpenAddresses.io provides regulary exports of worldwide adresses (we will focus on US south/west/midwest/northeast for now):

- https://batch.openaddresses.io/data



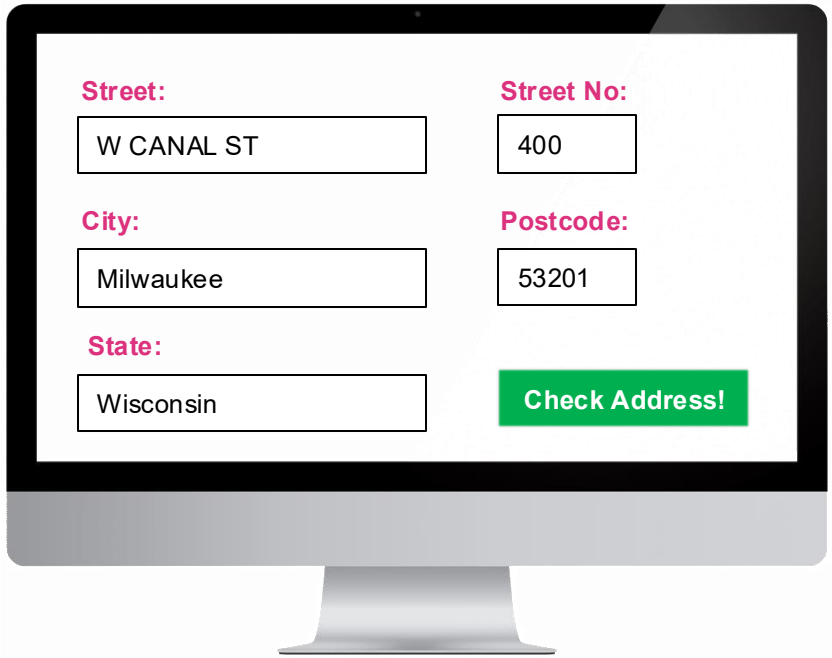

```
{"type":"Feature","properties":{"hash":"394d6a8e3e6cecbf","number":"7705","street":"W LINCOLN AVE","unit":"1","city":"West Allis","district":"","region":"","postcode":"53219","id":""},"geometry":{"type":"Point","coordinates":[-88.0088621,43.0025845]}}
{"type":"Feature","properties":{"hash":"6101cecbe71c7bbe","number":"7705","street":"W LINCOLN AVE","unit":"2","city":"West Allis","district":"","region":"","postcode":"53219","id":""},"geometry":{"type":"Point","coordinates":[-88.0088621,43.0025845]}}
{"type":"Feature","properties":{"hash":"81e3634e904916db","number":"1060","street":"N 115TH ST","unit":"106","city":"Wauwatosa","district":"","region":"","postcode":"53226","id":""},"geometry":{"type":"Point","coordinates":[-88.0551894,43.0441061]}}
{"type":"Feature","properties":{"hash":"fbf0248cdd1623ad","number":"12137","street":"W BURLEIGH ST","unit":"2","city":"Wauwatosa","district":"","region":"","postcode":"53222","id":""},"geometry":{"type":"Point","coordinates":[-88.0649444,43.0741544]}}
{"type":"Feature","properties":{"hash":"6c5867d0d98b7e9a","number":"11515","street":"W CLEVELAND AVE","unit":"B231","city":"West Allis","district":"","region":"","postcode":"53227","id":""},"geometry":{"type":"Point","coordinates":[-88.0560418,42.9946529]}}

[...]
```

Goal

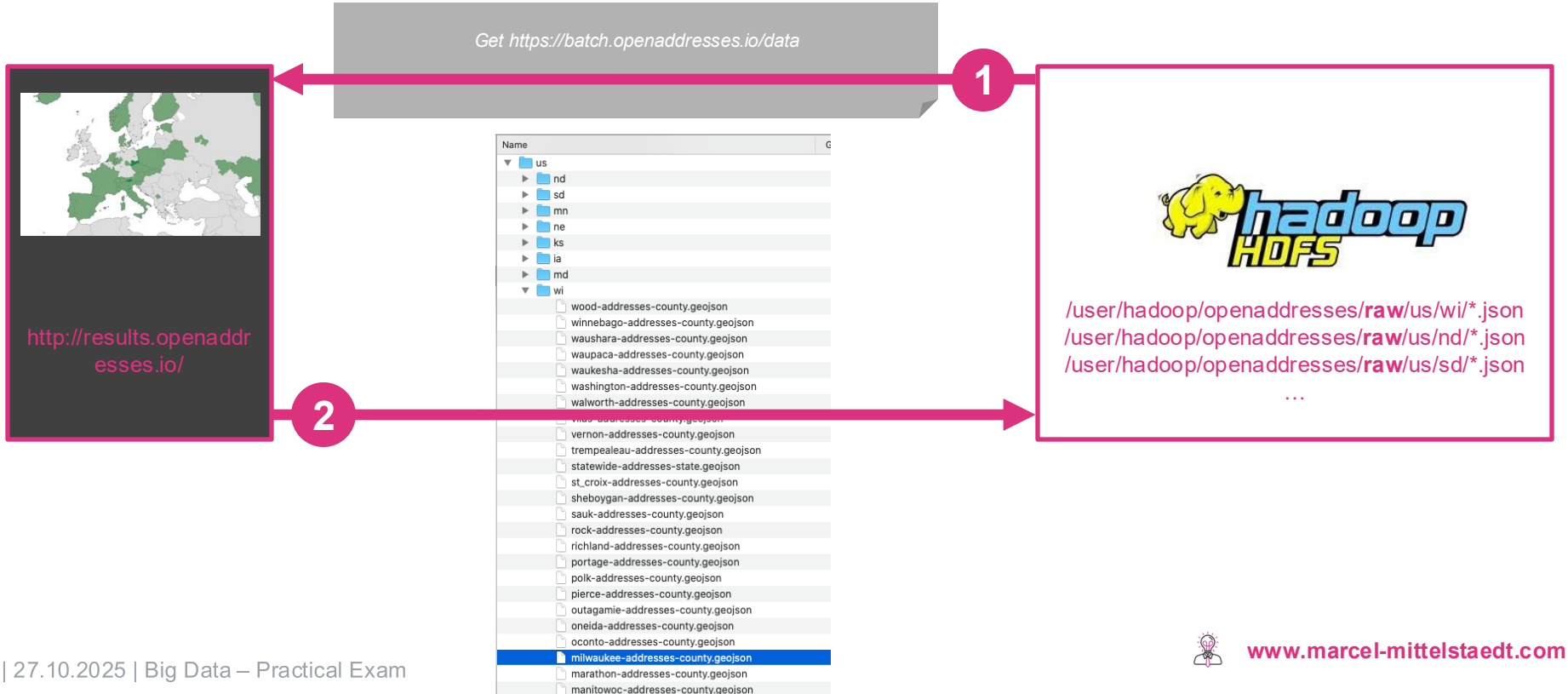
We want to make use of this data to validate adresses entered on a website, to check whether they are real or not.

Workflow:

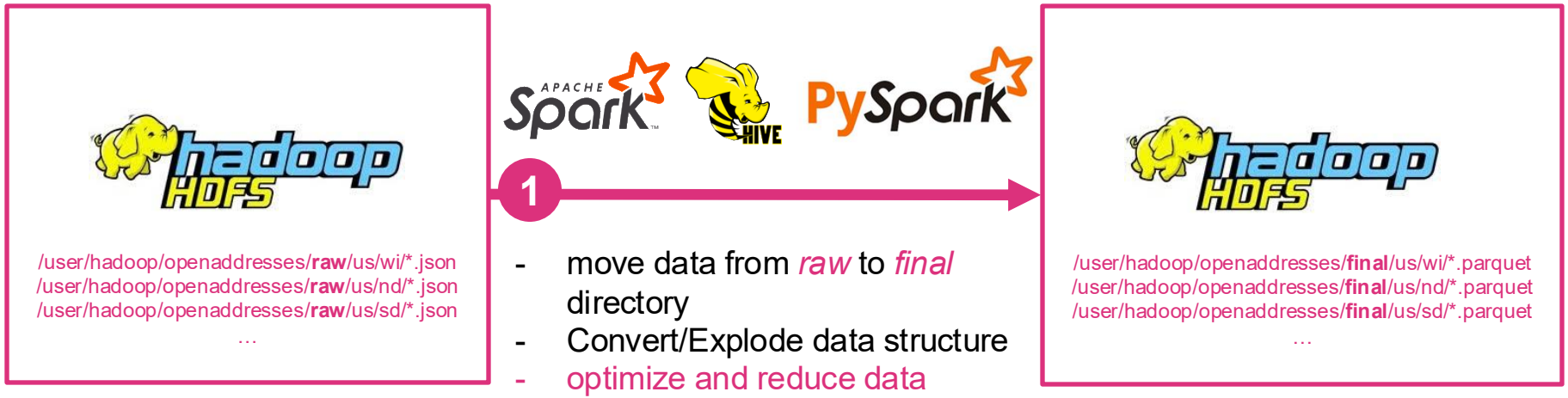
- **Gather data** from OpenAddresses.io
- **Save raw data** (*JSON files*) to HDFS (partitioned by state
- shortcut, e.g. *wi, nd, sd*…)
- **Optimize**, **reduce** and **clean raw data** and save it to **final** directory on HDFS
- **Export** address data **to end-user database** (e.g. MySQL, MongoDB…)
- Provide a simple **HTML Frontend** which is able to:
  - read from end-user database
  - process user input (Street, City, Postcode…)
  - validate user input against OpenAddress data in end-user database
  - Display result (real or non real address)
- The whole data workflow **must be implemented** within an ETL **workflow tool** (e.g. **Pentaho Data Integration** or **Airflow**) and **run automatically**



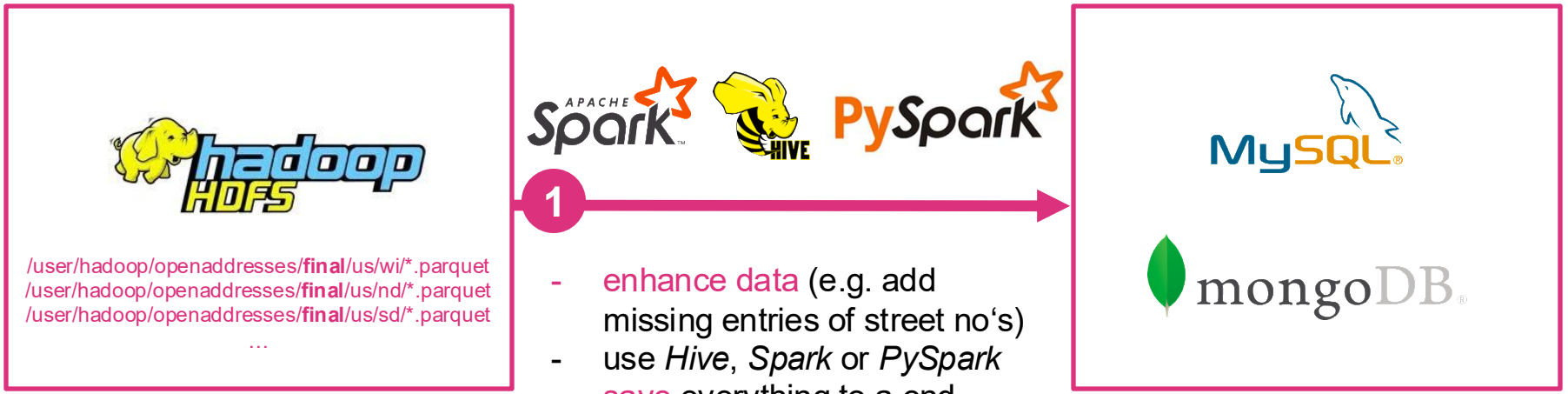
Dataflow: 1. Get Address Data



Dataflow: 2. Raw To Final Transfer

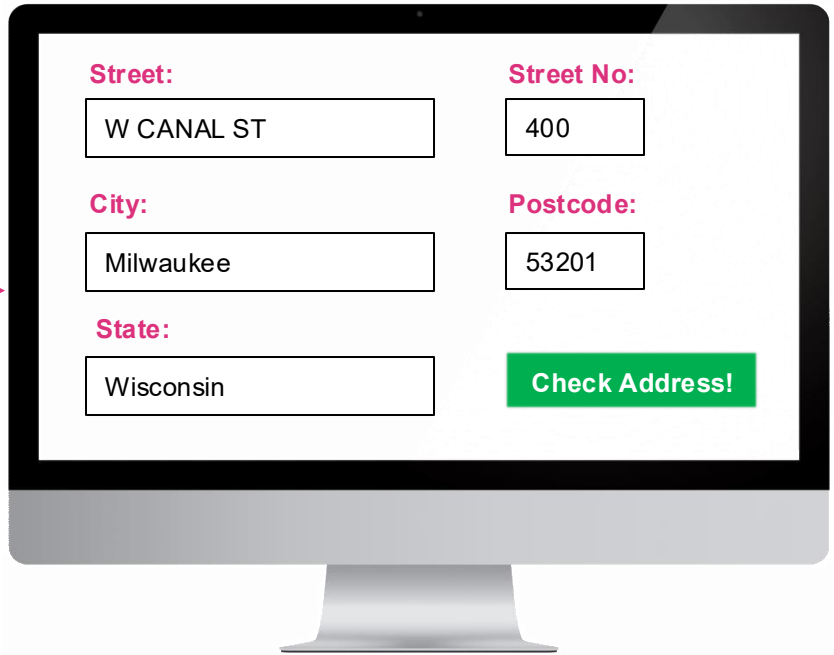


Dataflow: 3. Enhance Data And Save Results



Dataflow: 4. Provide Simple Web Interface

1



- Provide a simple **HTML Frontend** which is able to:
  - read from end-user database
  - process user input (Street, City, Postcode…)
  - validate user input against OpenAddress data in end-user database
  - Display result (real or non real address)