

Labour market mismatch dashboard

Statistician, employer and employee views with matching tool

Poland

Aims

To prepare big data prototype to support decisions on reducing labour market mismatch at regional level. The dashboard has three views: for statisticians, employers and employees.

Data

- EURES: job vacancies and CVs
- EU-SILC
- PIAAC
- EU-LFS
- Cedefop data
- Commuting in Poland for Urban Audit
- Earnings database: data on country earnings by ISCED levels and regions (OECD)
- Web scraped job offers for Poland
- Administrative data source: Graduates on higher education (database)

Methods, technologies and tools

The core software is based on open source Linux Ubuntu Server platform:

- Python 3.6 language – primary language used;
- Anaconda 3 – accessing interface via Jupyter Notebook;
- Apache Spark (to execute Python scripts);
- Apache Hadoop (primary storage for csv files).

The analysis part of the prototype was also supported by:

- MySQL database (for CEDEFOP database); MS SQL Server; Jupyter Notebook; MS Excel; JavaScript, HTML5.

Additional libraries used for Python language:

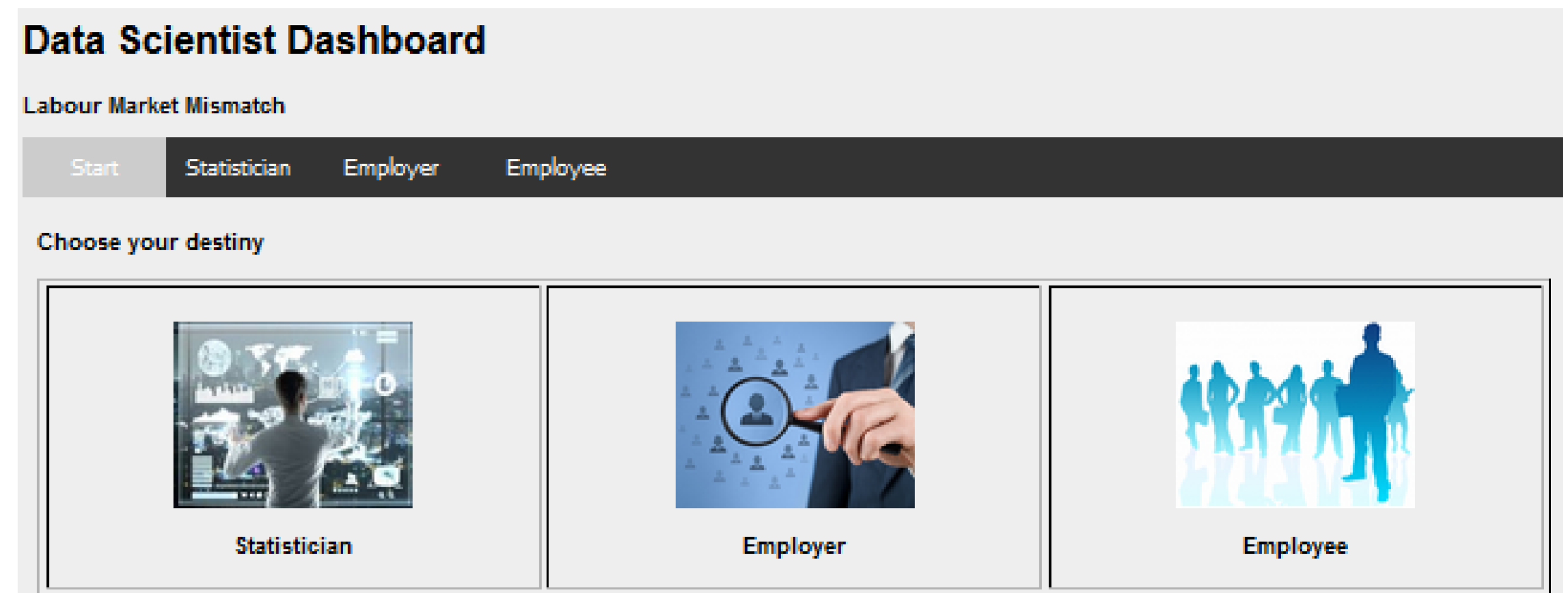
- scikit-learn; pandas; numpy; rdflib.

We used the replicated environments configured on:

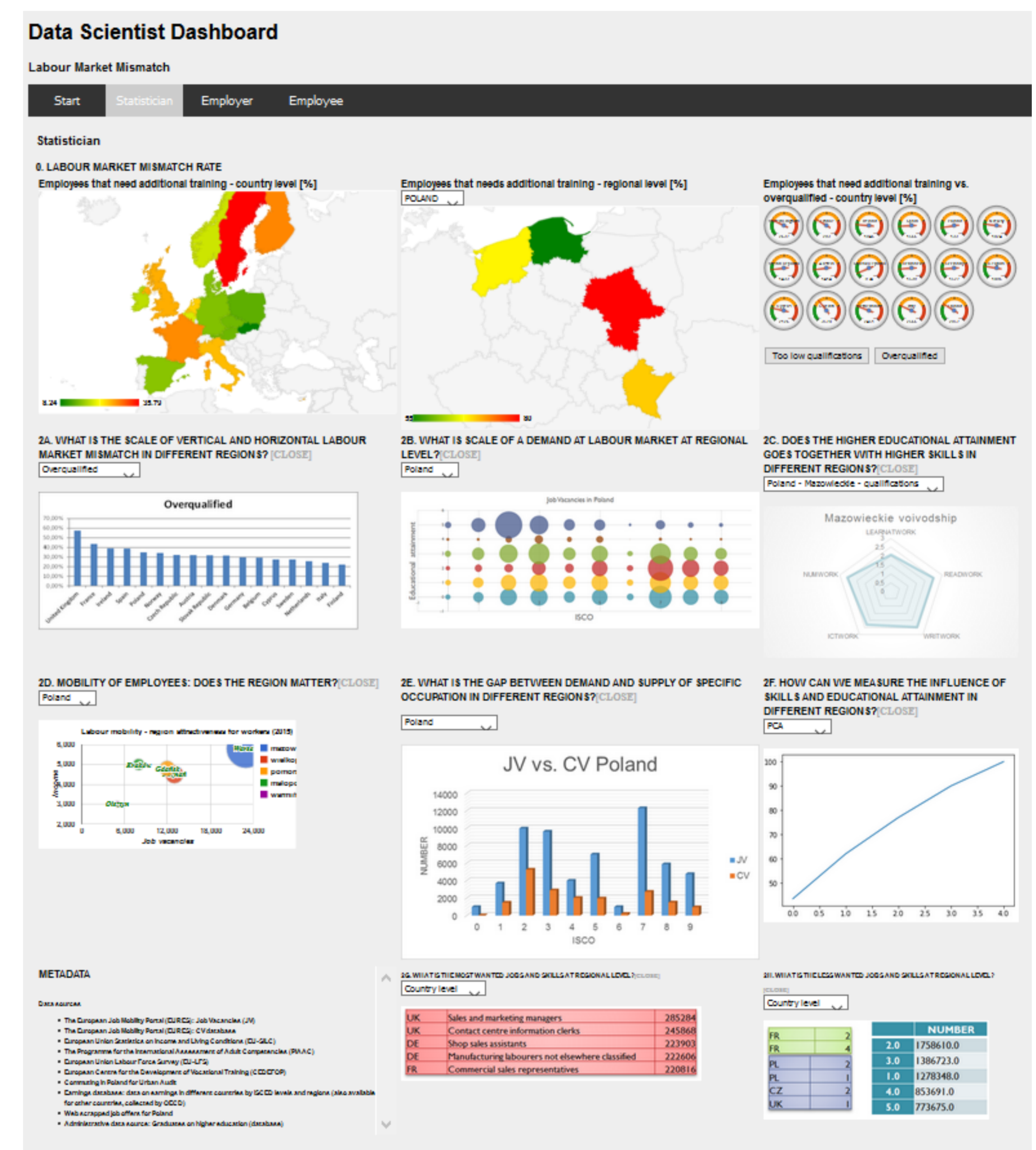
- Amazon Cloud – selected software migrated and used during hackathon;
- our own Linux Server – primary server for big data analytics for official statistics;
- MS Windows 2012 Server – secondary server for big data analytics – used to connect to MS SQL Server database.

Results

- Three views of data analysis and results



- Dashboard – statistician view



Conclusions

Key achievements:

- integration of various data sources;
- using three different perspectives of analysis: statistician, employer and employee;
- dashboard showing different aspects and dependencies between phenomena;
- web interface with numerous forms of interactive data presentation.

Main challenges:

- methodological issues of integrating specific data sources (NULL values, lack of common keys and representativeness issues do not allow integration of all data sets);
- tailoring data;
- missing or incorrect data;
- several variables in public use files (PUF) for European surveys are not present for selected countries; this makes it hard to present all the data by country.

Contact:

J.Maslankowski@stat.gov.pl