Big Data for Labour Market Information

Session 7

Architecture: solutions for real-time LMI (based on KDD)

Alessandro Vaccarino – Fabio Mercorio

Big Data for Labour Market Information – focus on data from online job vacancies – training workshop
Milan, 21-22 November 2019
Topics

1. Goal & context
2. Challenges
   1. The functional architecture
   2. Why use micro-services
   3. The Team and the pipeline design
   4. How handle infrastructure costs
Skills OVATE
Online Vacancy Analysis Tool for Europe

Seven countries
Czechia, France, Germany, Italy, Ireland, Spain & United Kingdom

Data classification
Merging different vacancy templates, styles and vocabularies into single language, using European and global taxonomies

Occupation:
up to ISCO 4-digit 400+ occupations

Skill:
ESCO version 1 2000+ skills

Sector:
NACE rev. 2 20+ sectors

Region:
NUTS-2 276 regions

Jobs demand
Shop assistants
Admin. secretaries
Junior accountants

High skilled: 56%

Software developers
System analysts
Material & quality engineers

Skilled non-manual
Truck drivers
Vehicle mechanics
Metal machine operators

Skilled manual
Freight handlers
Manufacturing labourers

Low skilled
Office cleaners

Top skills requested
Be adaptive to change
Work well in team
Use office software
Assist customers
Use a computer
Solve problems
Communicate well

Cross-cutting skills

Importance of data analysis skill across jobs

Financial analysts
Chemists
Mathematicians
Finance managers
Database specialists
Clearing & forwarding agents
Medical lab technicians

Stay tuned for more!
28 countries covered by end 2019
Fully operational system in 2020
Collaboration with ESCO & ESTAT

Covering periods
1 July to 31 December 2018

rtmi@cedefop.europa.eu
www.cedefop.europa.eu/skills-online-vacancies (end March 2019)
Challenges

• Handle a huge amount of near real time data
• Data coming from web → Need to detect and reduce noise
• Multi language environment
• Need to relate to classification standards
• Find a way to summarize and present a wide and complex scenario
Topics

1. Goal & context
2. Challenges
   1. Stakeholders
   2. The functional architecture
   3. Why use micro-services
   4. The Team and the pipeline design
Stakeholders

- Project Leader
- Key Users
- Domain Experts
- End Users
Project leader

- Lead the project with the steering committee
- Define the scope of the project
- Define key organizations
- Maintain relations with stakeholders
- Provide advice
Key Users

- Define requirements
- Monitor quality of the project
- Provide input to the development of the project
- Manage the source landscaping
- Validate overall data flow and methodology
Domain Experts

• International Country Experts
  • Provide the knowledge and expertise
  • Execute the landscaping
  • Understand the language/terms of their context
  • Evaluate the accuracy of the results
  • Test the product
  • Provide feedback
End Users

- Decision Makers and Business Users
  - (Visual) Explore dataset, analysis and aggregate data
  - Define new analysis processes
  - Produce Data storytelling
  - Make decisions by exploring data
- Data Scientists
  - Apply new machine learning models and AI techniques
  - Extract new insights from the data
  - Apply advanced data modelling to the dataset
- Data Analysts
  - Interprets data and turns it into information
  - Identifying patterns and trends
  - Extract and analyze aggregate data
  - Publish and share their analysis
Topics

1. Goal & context
2. Challenges
   1. Stakeholders
   2. The functional architecture
   3. Why use micro-services
   4. The Team and the pipeline design
Overall Data Flow

Ingestion

Data Ingestion

Processing

Pre-Processing

Information Extraction

ETL

Front end

Presentation Area
Conceptual architecture

Data ingestion
- Direct access
- Crawler
- Scraper
- Monitor and scheduler

Data processing
- Data quality
- Data processing and classification
- ETL

Data analysis
- Visual interface
- Data lab
- Data Supply

Backup
Logical view

- Employment Agencies and Public Employment Services
- Job Portals
- Classified Ads Sites
- Newspaper, Companies, University Job Placement
- Web Scraper
- Web Crawler
- Direct Access
- Pre-Processing
- Information Extraction and Classification
- Data Management and Presentation
- Document store
- Interactive Data Analytics
- Labour Market Analysts

- Job Vacancies Classified on ISCO
- Recognised NUTs
- Other dimension (contract, sector, education, …)

Logical view
Physical view

Data Ingestion → Data Processing → Modelling, Machine Learning, AI → Data visualization

Input: Unstructured Data

System and process monitoring

Automation & management

Output: Dashboard and interactive report, Machine to machine, Web App

Data storage & archiving
Key design projects

- Micro-services
- Componentization
  - Component specialization
  - Small applications
  - Portability
  - Reuse
  - Maintenance
- Scale Out
  - Performance
Key components

• **Data ingestion**: collect raw data from OJV in both structured and unstructured (raw text) formats
• **Data processing**: classify data through machine learning techniques
• **Data analysis**: extract information from data and make it available through visualization
• **Backup**: store data in a safe environment to allow warm and cold restore
Infrastructure Challenges

- Manage multiple parallel ingestion activities
- Availability of high performance computational infrastructure at a glance
- High memory requirements
- High storage volumes to store source and staging data
- Big data environment
- Scalable architecture
Topics

1. Goal & context
2. Challenges
   1. Stakeholders
   2. The functional architecture
   3. Why use micro-services
   4. The Team and the pipeline design
Microservices

A particular way of designing software applications as suites of independently deployable services
Context

Manutability

Monitoring

Scability

Updates

Onboarding
Pre-Processing Microservices

- Language Detector
- Spam Filter
- No-Vacancy Filter
- Stemmer
- Deduplication component
- N-gram component
- Text Cleaner
- Merge Vacancy
- TF-IDF Transformer
- Document2Vec
- Tokenizer
- StopWords Removers
Classification Microservices

- Skills Classifier
- Occupation Classifier
- Industry Classifier
- Education Requirements Classifier
- WorkingHours Detector
- Contract Detector
- Locations Detector
- Salary Extractor
- Experience Extractor
- Dates Extractor
Technology requirements

1. Services on request
2. Network access
3. Resource pooling
   1. Governance
4. Quick elasticity
5. Measurement of services
   1. Data Quality
   2. Performance
6. Portability (on-premises and different cloud services)
7. Polyglot
   1. Computer programming languages
   2. Technologies
Topics

1. Goal & context
2. Challenges
   1. Stakeholders
   2. The functional architecture
   3. Why use micro-services
   4. The Team and the pipeline design
The team

1. Cloud Architects
2. Software Architects and Developers
3. Big Data Engineers
4. Data Scientists
5. Domain & Ontology Experts
Organizations

Organize around business services

Team

Service
Execution

Cloud
Infrastructure

Define

Service

Micro-service

Design

Develop

Components

Deploy
Organize around business services

**Language Detector**
- Data Scientist
- Domain Expert
- Big Data Engineer
- Cloud Architect
- Software Architect

**Occupation Classifier**
- Data Scientist
- Domain Expert
- Big Data Engineer
- Cloud Architect
- Software Architect

**Salary Extractor**
- Data Scientist
- Domain Expert
- Big Data Engineer
- Cloud Architect
- Software Architect

**Skills Classifier**
- Data Scientist
- Domain Expert
- Big Data Engineer
- Cloud Architect
- Software Architect