Term project 2

Jobs, Skills, and Salaries:
A Data Engineering Dive into LinkedIn Insights

Team:
Azizbek Ussenov
Guillermo Leal
Tatyana Yakushina
Yutong Liang

Supervisor: Laszlo Sallo

The project provides a comprehensive analysis of jobs advertised on LinkedIn through a multi-platform data aggregation methodology to draw meaningful insights. The dataset is retrieved from Kaggle, and has undergone advanced data processing and analytical workflows. The dataset has been uploaded into Azure to support scalable cloud-based solutions. In addition, a minor portion of the data is being stored in MongoDB for text analysis and exploration purposes. Because Azure services are quite costly, a cost-efficient local database has been established with MySQL Workbench for storage and queries. To expand the scope of analysis, the project extracted *Bureau of Labor Statistics (BLS)* data using API. This would allow us to dive into the comparison of official wage data to those of actual company offers appearing on LinkedIn, resulting in deep analysis with reliable outside references.

The KNIME workflow provides a comprehensive analysis of job trends, wage comparisons, and skill demands across states and clustered industry names. It covers all the ETL processes to showcase the whole pipeline for analysis:

- **Extract:** retrieving data from a range of sources such as Azure, MongoDB, MySQL Workbench local database, and API with the help of relevant nodes to have integrated data for our analysis.
- **Transform:** cleaning, filtering, grouping, mathematical calculations like average wages, and also creating 25 clusters with unique names to categorize all industry names from both API and Linkedin dataset.
- **Load:** exporting cleaned and transformed data into csv and json files for further analysis, and generating different charts like bar chart, pie chart, and bubble chart as an output for illustration purposes.

As our final results we verified four hypotheses throughout the analysis process and presented the outputs using the KNIME workflow and charts:

1. Senior roles are less likely to offer remote work options compared to entry- and mid-level positions.
2. SQL is more frequently required in IT and data-centric industries such as finance, healthcare, and e-commerce compared to other sectors.
3. LinkedIn wages are consistently higher than BLS wages across all U.S. states due to a focus on professional roles.
4. LinkedIn-reported wages are consistently higher than BLS wages across key industries due to LinkedIn's focus on specialized, professional, and high-demand roles, compared to the broader coverage of the BLS data.

Overall, the project succeeded in integrating data from different sources and analyzing the Linkedin and BLS data for better workforce planning and policy analysis. Please visit Github repository for in-depth details: Term-2