Term project 2

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

Team #5: Azizbek Ussenov, Guillermo Leal, Tatyana Yakushina, Yutong Liang

The project provides a comprehensive analysis of jobs advertised on LinkedIn through a multi-platform data aggregation methodology to draw meaningful insights. The "Linkedin Job Postings" dataset is retrieved from Kaggle, and has undergone advanced data processing and analytical workflows. The dataset contains over 124,000 job postings from Linkedin listed in 2023 and 2024, with the majority companies in the dataset being U.S. -based. The datasets contain 11 tables relating to the job postings and the posting companies information.

The dataset has been imported into Azure SQL Database to support scalable cloud-based solutions which is also project team collaboration friendly. In addition, a minor portion of the text-based data, company id, job_id, job posting descriptions, skills descriptions, and company descriptions are being stored in MongoDB in JSON object formats for text searching in a NoSQL database. As Azure services can be costly, a cost-efficient local database has also been established with MySQL Workbench for storage and queries in the early stage of the project and as a backup database for the Azure SQL DB. To expand the scope of analysis, the project extracted *Bureau of Labor Statistics (BLS)* data using public API. The data contains the official labor statistics - the average wage by state and industry in the United States. 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 (image) 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:

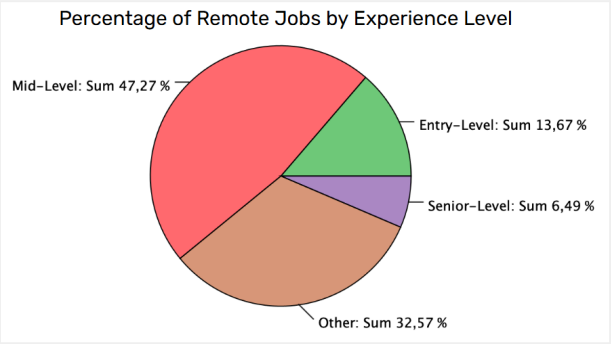
- **Extract:** retrieving data from a range of sources with the help of relevant nodes to have integrated data for our analysis. The sources list includes:
  - *Azure SQL DB (eer diagram)*: creating a Azure resource group, SQL server, and importing datasets from Kaggle into Azure SQL Database using Azure Data Studio import wizard with Azure SQL free tier subscription (details).
  - *MySQL (eer diagram)*: creating a local relational database as a copy of the Azure SQL DB for the cost-saving and backup purposes.
  - *MongoDB (data schema)*: converting the text-based columns such as "descriptions" into a JSON format and uploading to MongoDB database by splitting the JSON files into smaller chunks to adhere to MongoDB's 16MB document size limit (details).
  - *API*: requesting API from BLS website for extracting wage data per states and industries.
- **Transform:** We performed text clustering with Python to categorize all industries to 25 unique ones for both BLS and Linkedin dataset (artifacts), then we exported the cleaned and transformed data into csv files.
  - *TF-IDF Vectorization* for converting to numerical vectors, *TruncatedSVD* for dimensionality reduction, *K-Means clustering* to group into clusters, *Sentence-BERT* for name generation to embed pre-defined cluster names and industry names, *Cosine Similarity* for similarity check between embeddings.

We connect Azure SQL DB, MongoDB, API called BLS data using KNIME python node, and the preprocessed clustering csv files to KNIME. Initial data cleaning was performed such as removing N/A values, correcting numerical/boolean variable types, and renaming the necessary values.

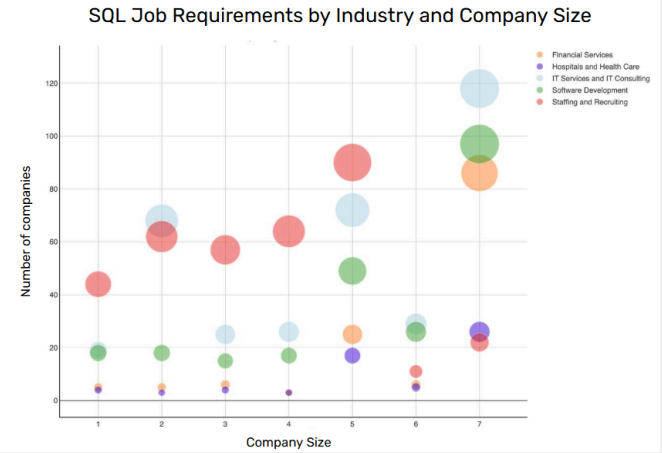
- **Load:** Then, we performed filtering, joining, grouping, string manipulation, rule engine, sorting, json to table, etc on the analytical layer. We analyzed the data given hypothesis proposed and produced tabular and visualization output such as bar chart, pie chart, and bubble chart as the analytical output.

As our final results, we verified four hypotheses throughout the analysis process and created 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. On average, LinkedIn wages are consistently higher than BLS wages across all U.S. states due to a focus on professional roles. (see chart in Hypothesis 3)
4. On average, 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. (see chart in Hypothesis 4)

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. You may find the detailed readme file which contains both technical operation manuals for reproducibility purposes and analytics. The project report is stored in the "Term2" folder. In the same folder, you may also find .knwf KNIME workflow and its svg image, the scripts for creating a local backup database, the presentation, and the Data_Preprocessing folder containing all the other relevant artifacts to the project.