Tech Job Market in Canada

Karim El Hoshy

McGill University

karim.elhoshy@mail.mcgill.ca

Introduction

In the rapidly evolving landscape of technology and data-driven decision-making, understanding job market trends is crucial for both job seekers and employers. My recent project aimed to create a comprehensive understanding of the tech job market in Canada. This report outlines the process of data collection, analysis, transformation and visualization using cutting-edge tools and technologies. All data collection, processing and analysis tasks were conducted on Databricks, leveraging its robust capabilities for handling large datasets and performing complex computations.

Data Collection

The data for this project was sourced by scraping LinkedIn job postings. The scraping process involved collecting job posting data across different industries, focusing on key information such as job titles, required skills, industry, and job descriptions.

Tools Used:

- **Python:** Utilized for scripting and data extraction.
- **BeautifulSoup:** Used for web scraping.
- **Requests:** Used to make HTTP requests to retrieve web pages for scraping.
- **Pandas:** Utilized for data manipulation and analysis.
- **Amazon S3:** Used for scalable and secure storage of the scraped data.

The data collection scripts are detailed in the Jupyter notebook ‘dataCollection.ipynb’.

Data Processing and Exploration

After collecting the data, the next step was to explore the dataset to gain a better understanding of its structure and content. This involved:

- **Creating a Requirements CSV:** To have a structured view of the data.
- **Creating Tables on the SQL Warehouse on Databricks:** To perform detailed queries and obtain deeper insights.
- **Generating Initial Statistics and Visual Summaries:** To identify key patterns and trends in the data.

The data processing and exploration steps are documented in the Jupyter notebook ‘dataExploration.ipynb.’

Data Cleaning

The collected and explored data then underwent extensive cleaning to remove inconsistencies, duplicates, and irrelevant entries. The key steps included:

- **Extracting Skills from Requirements:** Implementing a function to parse job descriptions and identify relevant degrees, experience, and skills including programming languages, frameworks, tools and soft skills. Using Natural Language Processing techniques to extract and categorize skills from job descriptions.
- **Standardizing Job Titles and skill Names:** Ensuring consistency across the dataset.
- **Handling Missing Values and Erroneous Data Entries:** Maintaining the quality of the dataset.
- **Creating Job Area Column:** Grouping roles into specific job areas (e.g., data science, cybersecurity, software engineering) to enable focused analysis on different job areas.
- **Grouping Industries:** Standardizing industry names and categories for better segmentation and analysis.
- The data cleaning process is documented in the Jupyter notebook ‘dataCleaningandTransformation.ipynb’.
- **Combining and Unpivoting CSV Files**: Utilized Apache Spark on Databricks to combine and unpivot CSV files, facilitating seamless usage in Power BI.

The detailed code for these steps is in the Jupyter notebook ‘dataCleaningandTransformation.ipynb.’

Visualization

The visualization component of the project was developed using Power BI. The dashboard provides interactive and user-friendly insights into job market trends.

Key Visuals:

- **Cloud Platform Demand**: A donut chart showing the demand for different cloud services (AWS, Azure, GCP).
- **Top 10 Most Common Skills in Tech Job Postings**: A word cloud highlighting the most common skills required in tech job postings.
- **Top 10 Most Common Skills in Data Science**: A bar chart displaying the most common skills in data science job postings.
- **Percentage of Openings Per Seniority Level**: A pie chart showing the distribution of job openings by seniority level.
- **Top 10 Industries with the Most Tech Opportunities**: A bar chart indicating the industries with the highest number of tech job opportunities.
- **Number of Jobs, Skills Extracted, Industries, and Job Areas**: Key metrics showcasing the total number of jobs studied after all the filtering and cleaning (3,537), skills extracted (20,750), industries (26), and job areas (12) covered in the analysis.

The dashboard provides a comprehensive view of the job market, making it easier for job seekers and employers to identify trends and make informed decisions.

Results

The Job Market Analysis Dashboard revealed several key insights:

- **Top Job Titles**: Data Engineer, Software Engineer, and Security Engineer were among the most common job titles across multiple industries.
- **In-Demand Skills**: Skills such as Python, SQL, and Cloud Computing were highly sought after.
- **Industry Opportunities**: Technology and IT Services, Finance, and Healthcare were leading industries with a high demand for data-related roles.
- **Cloud Platform Preferences:** AWS is the most in-demand cloud platform, followed by Azure and GCP, indicating a strong preference for AWS among employers.
- **Seniority Level Distribution:** The majority of job openings are for Mid-Senior level positions, suggesting a demand for experienced professionals.

Conclusion

This project successfully demonstrated the power of data scraping, processing, and visualization to provide actionable insights into the job market. It also showcased the positive effect LLMs can have on Data Scientists' and Data Engineers' productivity when used efficiently. The Job Market Analysis Dashboard is a valuable tool for job seekers to understand the skills in demand and for employers to align their job postings with market trends.

Future Enhancements

1. **Expand Data Sources:**

To enhance the comprehensiveness of the dataset, **including additional job portals** such as Indeed, Glassdoor, and Monster can be beneficial. However, integrating APIs from these job portals proved challenging as they did not provide the specific data needed for this project. Alternative methods, such as direct web scraping from these portals, could be explored to capture a broader range of job postings, ensuring the analysis reflects a more diverse job market.

2. **Real-Time Updates:**

Implementing real-time data scraping and dashboard updates would significantly improve the dashboard's utility. By continuously updating the job postings and skill requirements, job seekers can access the most current information, enabling them to tailor their applications more effectively. This requires setting up a robust **automated data pipeline** with scripts running at scheduled intervals, ensuring the dashboard reflects the latest job market trends.

3. **Advanced Analytics:**

Integrating predictive analytics can provide valuable insights into job market trends. Predictive models can forecast future demand for specific skills and job titles based on historical data. To achieve this, it is essential to collect data over a more extended period, beyond the current dataset that mostly covers May to July 2024. Using time series analysis and machine learning models, such as ARIMA, Prophet, or LSTM, can help predict future trends and assist job seekers in identifying emerging skills.

4. **Building a Job Market Analysis Website:**

Creating a website that hosts the Job Market Analysis Dashboard can serve as a powerful tool for job seekers. The website can provide various features, such as:

- **Interactive Dashboards**: Allowing users to filter and explore data based on their interests.

- **Job Search Tools**: Enabling users to search for jobs based on their skills, preferred industries, and locations.
- **Career Advice**: Offering tips and resources on how to tackle the job market, including resume building and interview preparation.
- **Notifications and Alerts**: Users can sign up for alerts on new job postings matching their criteria, keeping them updated in real time.

5. **Improving Data Collection and Filtering:**

Enhancing the data collection process to ensure higher accuracy and relevance can involve:

- **Advanced Scraping Techniques**: Using more sophisticated scraping methods to handle dynamic and complex web pages.
- **Data Validation**: Implementing validation checks to ensure the quality and integrity of the collected data.
- **Filtering Mechanisms**: Developing advanced filtering mechanisms to remove noise and irrelevant data, focusing on high-quality job postings.

These improvements would require additional time and resources but would significantly enhance the overall effectiveness and reliability of the project.