A Project on

LinkedIn Job Postings

Supervised By

Roy Kucukates

Prepared By

Nabila Noor
Ugochinyere Okehi
Obinna Onyema
Damilola Agbolabori

Toronto Metropolitan University
2023

Table of Contents

Introduction

In the fast-evolving landscape of the job market, gaining insights into job postings is critical for job seekers, employers, educational institutions, students, and policymakers. This project leverages the LinkedIn Job Postings Dataset from Kaggle (Raj, 2023) to comprehensively analyze and present the dynamics of job postings.

The dataset consists of multiple files, including job postings, company details, industry information, employee counts, benefits, and more. By addressing key objectives, the project aims to provide actionable insights:

- In-Demand Roles and Skills: Identifying and describing the most sought-after job roles and skills in the current market.
- Dominant Industry Sectors: Determining which industry sectors are actively hiring and understanding their impact on the job market.
- Regional Job Demand: Analyzing job demand based on geographical regions, providing valuable information for location-based career decisions.
- Qualifications and Experience: Highlighting typical qualifications and experience required for various roles, aiding both job seekers and employers in making informed decisions.

By addressing these requirements, the project facilitates a nuanced understanding of the job market dynamics, empowering stakeholders to navigate and adapt to the ever-changing employment landscape. The findings aim to bridge the gap between job market participants and provide a foundation for informed decision-making and strategic planning.

Review of Related Literature

The intricate dynamics of the job market form a nexus that intertwines labor economics, human resource management, and workforce planning. This literature review seamlessly integrates pivotal research contributions, unraveling the tapestry of job market trends, demands, and skills prerequisites.

Strohmeyer et al. (2018) lay a foundation by emphasizing the critical role of job market analysis in predicting employment trends. Their framework ingeniously combines traditional economic

indicators with unconventional data from online job postings and social media, enhancing the precision of predictions.

Moving in tandem, McGuinness and Bergin's (2019) exploration of skills mismatch underscores its profound impact on job vacancies and overall labor market efficiency. The study accentuates the pivotal role of aligning job seekers' skills with employers' requirements for effective job market functioning.

Journeying geographically, Faggian et al. (2020) employ spatial analysis techniques to unveil variations in job markets. Regional patterns in job demand come to the forefront, shedding light on factors contributing to disparities in employment opportunities across diverse locations.
Temporal nuances in job requirements take center stage in the work of Khan et al. (2017). Through the lens of machine learning algorithms, the study delves into historical job data, unraveling the evolution of skill demands over time and identifying emerging trends in the job market.

Bertrand and Mullainathan's (2003) seminal work delves into the intricacies of qualifications and experience in job advertisements. Investigating biases in hiring processes and discerning the impact of educational and experiential requirements on candidate selection, their research adds a layer of understanding to the hiring landscape.

In the realm of big data analytics, Acar and Turetken (2015) navigate the intersection of large-scale datasets and job market research. Their exploration underscores how datasets, including online job postings and resumes, can offer profound insights into job market dynamics, elevating decision-making processes.

A harmonious undertone is maintained as the LinkedIn Job Postings Dataset itself becomes a subject of analysis by various researchers. Insights derived from this dataset encompass job categorization, salary trends, and company characteristics.
This literature review weaves an intricate yet coherent tapestry, each thread contributing to a nuanced understanding of the multifaceted dynamics within the realm of job markets.

In-Demand Job Roles & Skills

The dynamism of the contemporary job market necessitates a comprehensive understanding of the in-demand job roles and the specific skills that recurrently surface in job postings.

Researchers have investigated various aspects, including occupational trends, spatial dynamics in job search, and skills mismatch. Strohmeyer et al. (2018) provide insights into job market predictions, while Faggian et al. (2020) emphasize spatial considerations. Bertrand and Mullainathan (2003) address labor market discrimination, revealing insights

into the employability of different demographic groups. Temporal trends in job skill requirements are discussed by Khan et al. (2017), and McGuinness and Bergin (2019) delve into the concept of skills mismatch. Integrating these studies offers a holistic understanding of the prevailing job roles and skills, enabling stakeholders to navigate the dynamic job market effectively.

Dominant Industry Sectors

Acar and Turetken (2015) highlight the business impacts of big data, providing a foundation for analyzing large datasets, such as the LinkedIn Job Postings Dataset. Researchers often explore the spatial aspects of job search and matching, as demonstrated by Faggian et al. (2020). By understanding the industries associated with job postings, analysts gain a comprehensive view of sector-specific demands. Integrating these perspectives, this literature review contributes to the exploration of industry sectors, shedding light on recruitment activities and potential shifts in hiring patterns over time.

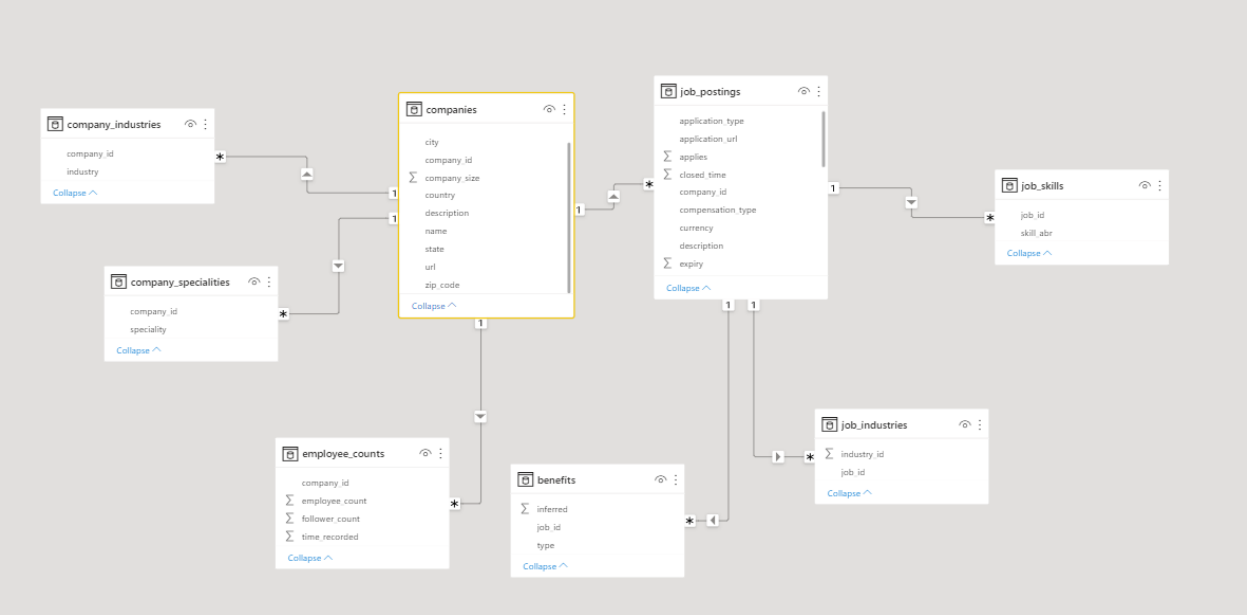
Regional Job Demand

Strohmeyer et al. (2018) provide lessons from predicting job offers using LinkedIn, indicating the platform's relevance for understanding job market trends. McGuinness and Bergin (2019) contribute by examining skills mismatch, a factor that can influence regional job demand. Analyzing the spatial distribution of job opportunities, researchers explore how specific regions experience pronounced growth, providing a nuanced understanding of regional employment trends. By drawing on these studies, this literature review aims to unravel the complexities of regional job demand, identifying key factors influencing job growth across diverse geographical locations.

Qualifications and Experience

Examining the qualifications and experience requirements within job postings is crucial for understanding the expectations of employers and the skills deemed essential for various roles. Faggian et al. (2020) contribute to this topic by emphasizing the importance of employment networks in spatial job search and job matching. While their focus is on spatial dynamics, the underlying premise acknowledges the influence of qualifications and experience in job matching.

Data description



The Linkedin Job Postings data set consists of 8 csv files which were loaded in 8 different tables in Hive. Details of the content of each table is highlighted below.

1. job_postings: This table contains 15,886 records. Some of the columns in the table contain missing values. However, the columns we need for the questions we intend to answer in this project have less than 20% missing values, hence we will be ignoring the missing values.These columns include the job_id, company_id, title, description and views
2. companies: This table contains detailed information about each company that posted a job, including the company name, website, description, size, location, and more. All the columns in the table have less than 10% missing values.
3. company_industries: This table contains the industries associated with each company. The two columns in this table have no missing values.
4. company_specialities: This table contains the specialties associated with each company. The two columns in this table have no missing values.
5. employee_counts: This table contains the employee and follower counts for each company. There are no missing values in this table.
6. benefits: This table contains the benefits associated with each job. There are no missing values in this table.
7. job_industries: This table contains the industries associated with each job. There are no missing values in this table.
8. job_skills: This table contains the skills associated with each job. There are no missing values in this table.

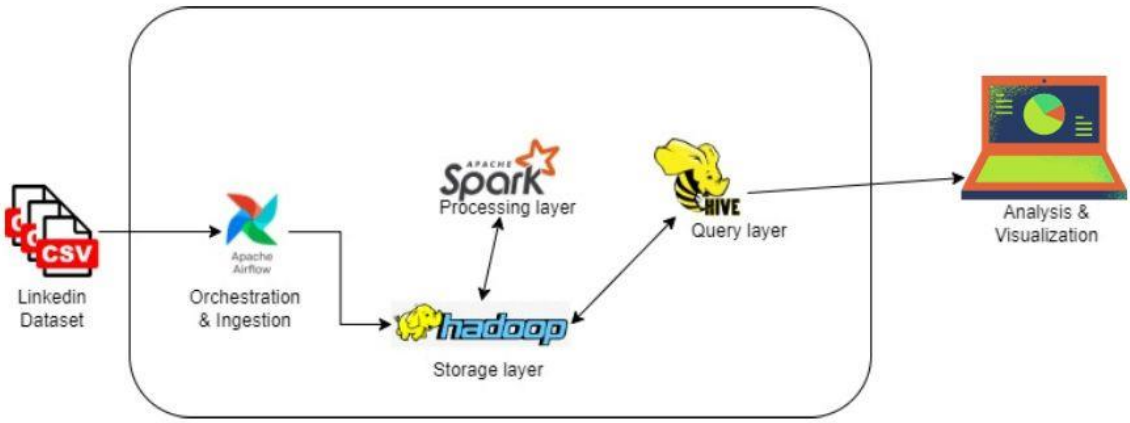
Work distribution

The project team worked together to deliver the solution with roles overlapping for peer review and collaboration.

- Nabila Noor
  Exploratory Data Analysis, Pyspark processing/cleaning, Presentation, Report
- Ugochinyere Okehi
  Exploratory Data Analysis, Presentation, Report
- Obinna Onyema
  Airflow Pipeline, Presentation, Report
- Damilola Agbolabori
  Data Modelling, Database set up, Pyspark processing/cleaning, Presentation, Report

Solution description

In implementing this solution, we used big data tools in a way that has the potential to handle high volumes of data efficiently. This is because data volumes and speed requirements in industry continue to push new boundaries. The average company manages 162TB of data (IDG, 2016) therefore it is imperative to consider volume, velocity and variety (Laborde, 2020) when designing modern data pipelines.



Solution Architecture

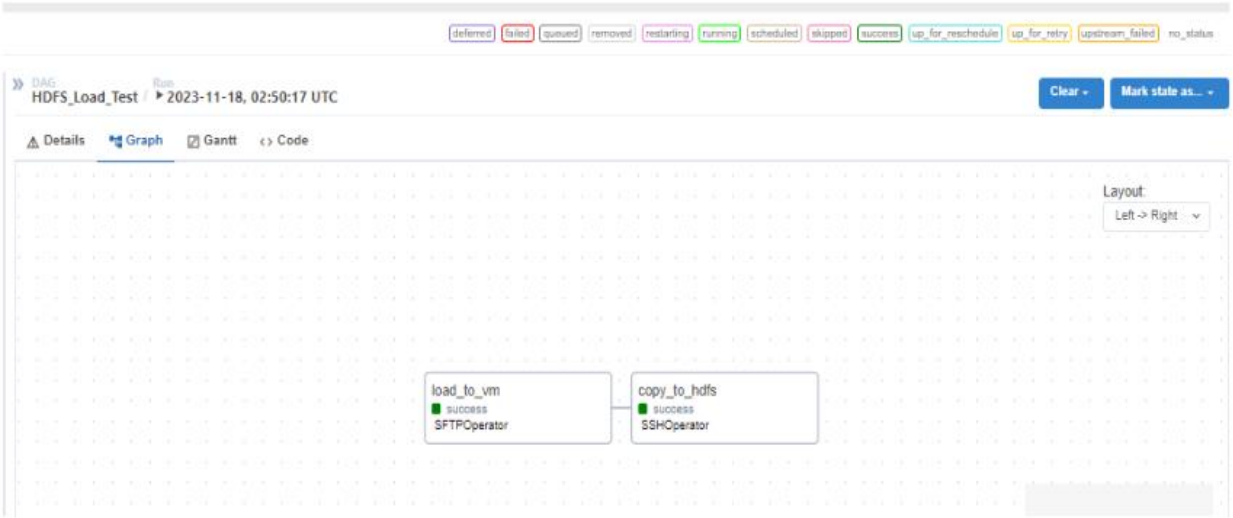
Tools

1. Airflow: for the data pipeline. Airflow handles workflow management with directed acyclic graphs (DAGs). It has the ability to handle concurrency and scale.

2. Hadoop: for data storage. The hadoop file system enables distributed computing which is valuable for big data processing.
3. Hive: for database management and ad hoc queries. Hive runs on the hadoop file system.
4. Spark: for processing & data analysis. Spark is powerful for handling large volumes of data. It supports multiple languages but we used python (PySpark) in most of our interactions with Spark in this project.
5. Jupyter: we used jupyter for visualization to maintain simplicity.
6. Github: to facilitate collaboration and code sharing.

Data Extraction, Loading & Storage

➢ Airflow was used to extract csv files from source and load into the Hadoop file system (HDFS). A connection was set up in airflow to connect to the virtual machine running hadoop via SSH. Tasks were created using SSHOperator and SFTPOperator to start hadoop and to load files into HDFS, respectively.



➢ The csv files were then loaded into dataframes using PySpark and then saved into tables in the Linkedin database that was created in Hive. See below sample code snippet for the creation of the benefits table, link to the code used for cresting all tables can be found in the appendix section.

```
hive> CREATE DATABASE IF NOT EXISTS LINKEDIN;
OK
Time taken: 8.42 seconds
```

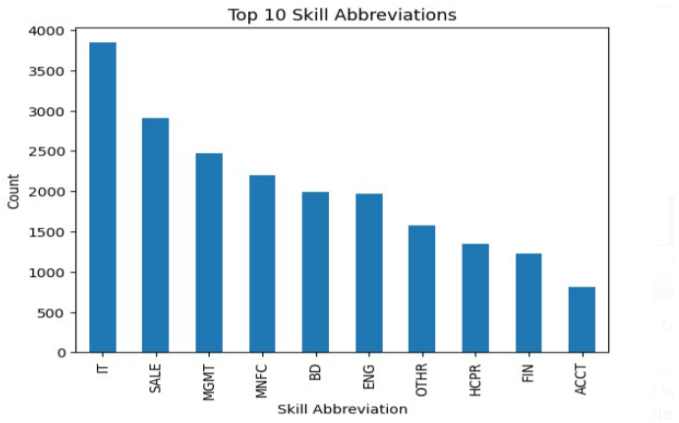
```
>>> benefits_df = (spark.read
...
...       .option("multiline", "true")
...
...       .option("quote", '"')
...
...       .option("header", "true")
...
...       .option("escape", "\\")
...
...       .option("escape", '"')
...
...       .csv('/user/root/Linkedin/benefits.csv')
...
... )
>>> benefits df.write.saveAsTable("Linkedin.benefits")
hive> select * from benefits limit 10;
OK
3690843087      0       Medical insurance
3690843087      0       Dental insurance
3690843087      0       401(k)
3690843087      0       Paid maternity leave
3690843087      0       Disability insurance
3690843087      0       Vision insurance
3691763971      1       Dental insurance
3691763971      1       Disability insurance
3691763971      1       401(k)
3691775263      0       Medical insurance
Time taken: 0.415 seconds, Fetched: 10 row(s)

hive> show tables;
OK
benefits
companies
company_industries
company_specialities
employee_counts
job_industries
job_postings
job_skills
Time taken: 0.356 seconds, Fetched: 8 row(s)
```

Insights and Discussion

- In-Demand Skills - In this section we identified the top skills that recurrently manifest in job postings.

The data indicates that the most common skill is IT, which is followed closely by Sales, Management, Manufacturing and Business Development skills.

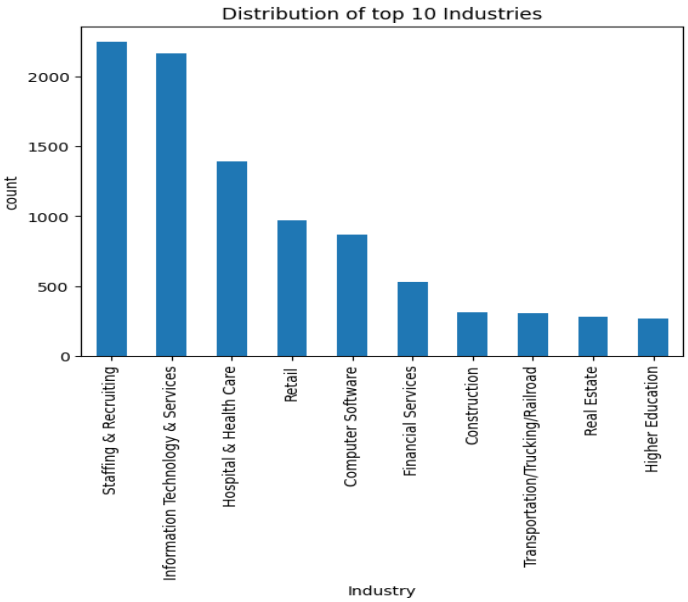
cnt	skill_abr	percentage_of_total
3841	IT	13.77
2904	SALE	10.41
2467	MGMT	8.84
2195	MNFC	7.87
1993	BD	7.14
1974	ENG	7.08
1574	OTHR	5.64
1346	HCPR	4.82
1227	FIN	4.4
813	ACCT	2.91



- Dominant Industries - In this section we identified the industries are hiring the most based on the provided data.

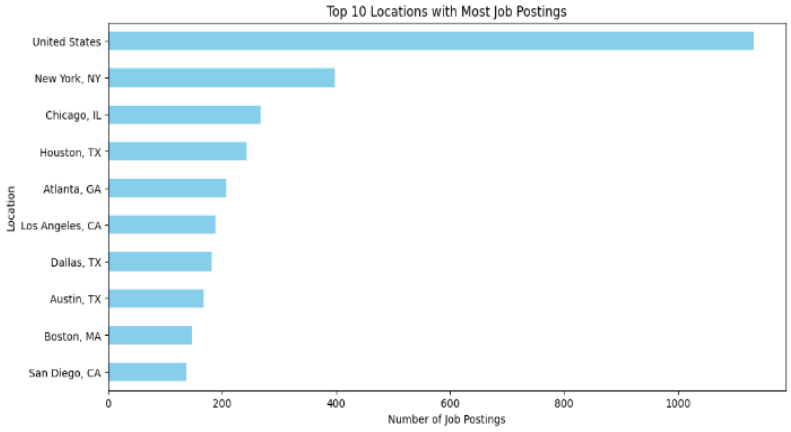
The results from the analysis of data available in the company_industries table also shows that the IT industry is among the top industries of the companies that post job openings on Linkedin.

cnt	industry	percentage_of_total
2245	Staffing & Recrui...	14.14
2164	Information Techn...	13.63
1392	Hospital & Health...	8.77
968	Retail	6.1
868	Computer Software	5.47
531	Financial Services	3.34
315	Construction	1.98
305	Transportation/Tr...	1.92
285	Real Estate	1.79
267	Higher Education	1.68

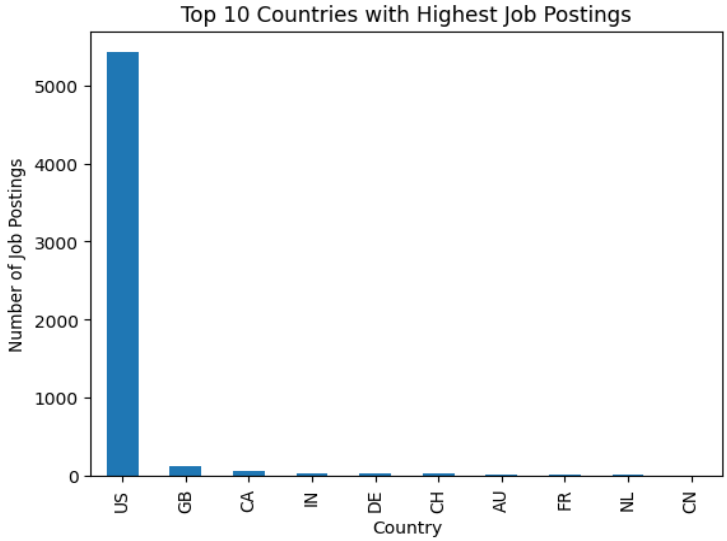


- Regional Job Demand - In this section we addressed the following questions:
  - How does job postings vary across different locations?
  - Are there specific regions experiencing higher job growth?

location	count
United States	1133
New York, NY	398
Chicago, IL	267
Houston, TX	243
Atlanta, GA	207
Los Angeles, CA	188
Dallas, TX	182
Austin, TX	168
Boston, MA	147
San Diego, CA	137

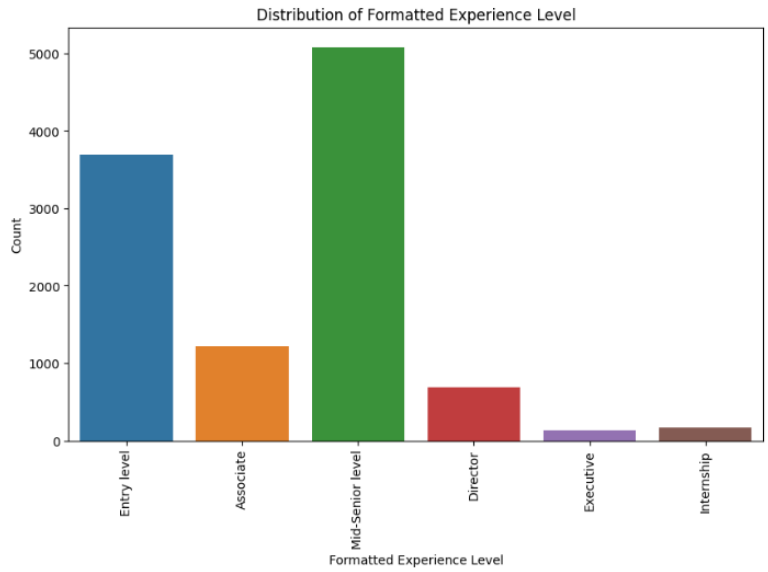


cnt	country	percentage_of_total
5419	US	91.74
129	GB	2.18
71	CA	1.2
39	IN	0.66
35	DE	0.59
29	CH	0.49
19	AU	0.32
19	FR	0.32
15	NL	0.25
11	CN	0.19



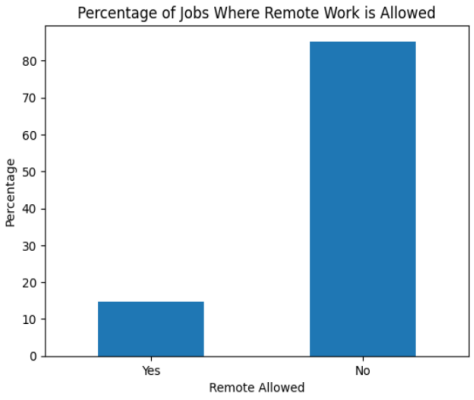
- Level of Experience - In this section we addressed the follow questions:
  - How does the level of experience vary for different positions?

formatted_experience_level	cnt	percentage_of_total
Mid-Senior level	5083	46.28
Entry level	3694	33.63
Associate	1220	11.11
Director	687	6.25
Internship	166	1.51
Executive	134	1.22

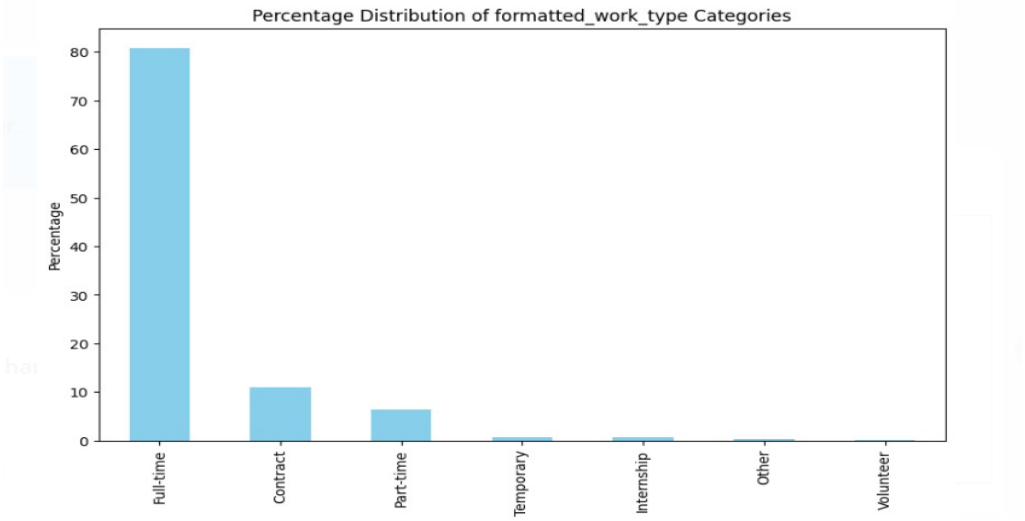


- Work Mode - In this section we addressed the follow questions:
    - What is the top work mode for job postings within the contemporary job market?
    - What are the prevalent work types within the contemporary job market?

remote_allowed	count	percentage
1.0	2340	14.729950900163665
0	13546	85.27004909983633



formatted_work_type	cnt	percentage_of_total
Full-time	12844	80.85
Contract	1739	10.95
Part-time	1010	6.36
Temporary	121	0.76
Internship	111	0.7
Other	53	0.33
Volunteer	8	0.05



- The top 10 companies with the highest number of job posting
  The cleaning of the data:

There was a total of 15886 records in the job_posting dataframes, and 24032 in the companies dataframes. The cleaning revealed 366 records without company ids in the job_posting dataframe.

```
>>> job_postings_df = (spark.read.option("multiline", "true").option("quote",'"').option("header", "true").option("escape", "\\
").option("escape", '"').csv("file:///home/data/project/job_postings.csv"))
>>> companies_df=spark.read.csv("file:///home/data/project/companies.csv", header=True)
>>> job_postings_df = job_postings_df.withColumn("company_id",job_postings_df.company_id.cast(IntegerType()))
>>> job_postings_df = job_postings_df.withColumn("views",job_postings_df.company_id.cast(IntegerType()))
>>> job_postings_df.count()
15886
>>> companies_df.count()
24032
>>> job_postings_df = job_postings_df.filter(job_postings_df.company_id.isNotNull())
>>> companies_df = companies_df.filter(companies_df.company_id.isNotNull())
>>> job_postings_df.count()
15520
>>> companies_df.count()
24032
```

An Inner join of the two dataframes

```
# An Inner Join of job_posting_df and companies_df
job_cols = job_postings_df.columns
company_cols = companies_df.columns

job_postings_df = job_postings_df.selectExpr([col + ' as job_' + col for col in job_cols])
companies_df = companies_df.selectExpr([col + ' as company_' + col for col in company_cols])
join_job_companies_df = job_postings_df.join(companies_df, job_postings_df.job_company_id == companies_df.company_company_id, "inner")
```

```
>>> highest_job_postiting.orderBy(col('count').desc()).show(10)
+-------------+------------------+-----+
|job_company_id|      company_name|count|
+-------------+------------------+-----+
|      3570660|     City Lifestyle|  161|
|         1103|           Verizon|  113|
|        11056|     Insight Global|  108|
|         1586|            Amazon|   93|
|         1441|            Google|   93|
|     10420321|    The Mom Project|   92|
|     18506580|      Vivian Health|   71|
|         1403|Booz Allen Hamilton|   70|
|         6176|          7-Eleven|   68|
|         1681|       Robert Half|   56|
+-------------+------------------+-----+
only showing top 10 rows
```

- The top 20 job titles with the highest number of applicant. The Junior Software Engineer role at Brooksource has the highest number of job applicants in this dataset for this time period.

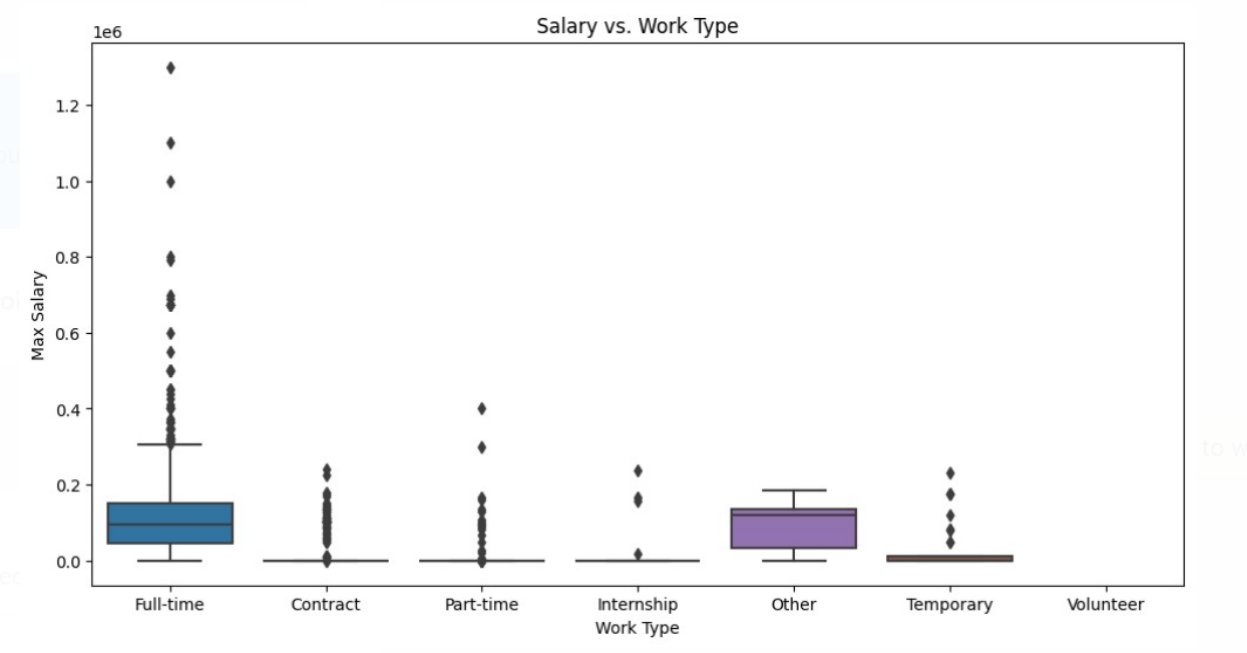
```
# top 20 companies with the highest number of applicant grouped by Title of job
join_job_companies_df = join_job_companies_df.filter(join_job_companies_df.job_applies.isNotNull())
highest_applicant_posting = join_job_companies_df.groupby("job_company_id", "company_name", "job_title").sum("job_applies")
highest_applicant_posting.orderBy(col('sum(job_applies)').desc()).show(20, truncate=False)
```

```
>>> highest_applicant_posting.orderBy(col('sum(job_applies)').desc()).show(20, truncate=False)
+-------------+------------------------------+-----------------------------------------------------------------------------------------+---------------+
|job_company_id|company_name                  |job_title                                                                                |sum(job_applies)|
+-------------+------------------------------+-----------------------------------------------------------------------------------------+---------------+
|18476        |Brooksource                   |Junior Software Engineer                                                                 |1615           |
|2503130      |Noom                          |Customer Success Manager                                                                 |1420           |
|4787         |Apex Systems                  |Customer Service Representative                                                          |983            |
|35676539     |ClearSky Health               |Talent Acquisition Manager                                                               |924            |
|5310955      |Curate Partners               |Scrum Master / Project Manager (Healthcare/Call Center) - 100% Remote                    |832            |
|10577525     |The CARIAN Group              |Power BI Data Analyst - (Remote)                                                         |742            |
|10229031     |RMS Beauty                    |Digital Marketing Manager                                                                |669            |
|2684081      |Optomi                        |Business Intelligence Analyst                                                            |646            |
|18715975     |AltaView Advisors LLC         |Director of Operations                                                                   |618            |
|6396497      |Coast Medical Service         |Healthcare Recruiter                                                                     |611            |
|35471087     |Riverside Insights            |Recruiter                                                                                |588            |
|76323674     |OERTZENGroup                  |Recruiter                                                                                |586            |
|10577525     |The CARIAN Group              |Data Scientist - (Remote)                                                                |577            |
|25742        |AHIMA                         |Talent Acquisition Specialist                                                            |561            |
|1265059      |Carol Olsby & Associates      |Executive Assistant - Global Programs - Non-profit - US Eastern Time Zone Candidates Only|554            |
|95019864     |Spyre Therapeutics            |Head of Information Technology                                                           |532            |
|2502526      |LS Direct                     |Data Analyst                                                                             |500            |
|28923        |Harbor Freight Tools          |Manager, Workforce Management (Remote)                                                   |486            |
|2473798      |Fulcrum Technology Solutions, LLC|Corporate Recruiter                                                                      |476            |
|323887       |Jackson Therapy Partners      |National Recruiter - $5,000 Sign on Bonus!                                               |473            |
+-------------+------------------------------+-----------------------------------------------------------------------------------------+---------------+
only showing top 20 rows
```

Noom company has the highest number of applicants from all their job postings from around the world.

```
>>> highest_applicant_posting.orderBy(col('sum(job_applies)').desc()).show(10)
+--------------+---------------+----------------+
|job_company_id|   company_name|sum(job_applies)|
+--------------+---------------+----------------+
|       2503130|           Noom|            3039|
|         11056| Insight Global|            2747|
|         18476|    Brooksource|            2137|
|          4787|   Apex Systems|            1892|
|          1681|    Robert Half|            1483|
|          1586|         Amazon|            1451|
|       2684081|         Optomi|            1420|
|          1441|         Google|            1358|
|      10577525|The CARIAN Group|            1319|
|         11229|           Vaco|            1141|
+--------------+---------------+----------------+
only showing top 10 rows
```

- Salary Vs Work Type



Recommendations for future work

It is recommended that future implementation includes additional data which is more up to date so as to make more time-relevant analysis.

Data sourcing can also be automated with the use of APIs for obtaining data from LinkedIn so that csv files need not be manually dumped to a folder from where they will be picked up by Airflow. Alternatively, an automated system such as web scraping may be used to obtain the data and store the output as CSV files in a location (such as SFTP) from which the Airflow pipeline can pick up these files.

Elastic search can be integrated to front-end users for searches at scale.

Conclusion

From the work done, it is evident that

- The most in demand skills are in IT, Sales, Management, Manufacturing and Business Development.
- The United States has the bulk of job listings: over 90%
- Mid-level and entry-level listings make up 80% of job listings
- 15% of jobs permit remote work

About the big data tools, we used in this project, we have made the following observations:

- Some source files had unique quotes and escape sequences. Spark data frames loaded these files with ease, compared to Hive.
- Apache Airflow is great for automating the entire workflow and has the functionality to handle multiple concurrent tasks to manage higher scale.
- Big data tools can be resource-intensive. We battled with preparing data in the small-sized Azure VMs. There were speed and memory limitations. It is recommended that in production deployments; adequate capacity is provided to enable scalability.

References

- IDG. (2016). *Data and Analytics Survey*. International Data Group. Retrieved Nov 18, 2023, from https://cdn2.hubspot.net/hubfs/1624046/IDGE_Data_Analysis_2016_final.pdf
- Laborde, R. (2020, January 23). *The Three V's of Big Data: Volume, Velocity, and Variety*. Oracle Blogs. Retrieved November 18, 2023, from https://blogs.oracle.com/life-sciences/post/the-three-vx27s-of-big-data-volume-velocity-and-variety
- Raj, R. (2023). *LinkedIn Job Postings Dataset*. Kaggle. Retrieved November 19, 2023, from https://www.kaggle.com/datasets/rajatraj0502/linkedin-job-2023
- Strohmeyer, R., et al. (2018). "Job Market Predictions: Lessons from Predicting Job Offers Using LinkedIn." *arXiv preprint arXiv:1806.04460*.
- McGuinness, S., & Bergin, A. (2019). "Skills Mismatch: Concepts, Measurement and Policy Approaches." *Oxford Research Encyclopedia of Economics and Finance*.
- Faggian, A., et al. (2020). "Spatial Job Search, Job Matching, and the Importance of Employment Networks." *Regional Studies, 54*(7), 913-925.
- Khan, A., et al. (2017). "Temporal Trends in Job Skill Requirements: A Text Mining Perspective." *Expert Systems with Applications, 78*, 112-125.
- Bertrand, M., & Mullainathan, S. (2003). "Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination." *American Economic Review, 94*(4), 991-1013.
- Acar, A. Z., & Turetken, O. (2015). "Big Data and Its Business Impacts: A Brief Research Agenda." *Journal of Organizational Computing and Electronic Commerce, 25*(3), 246-257.
- Dr Roy, K. (2023). Data Processing: Spark DS8003 – MGT OF BIG DATA AND TOOLS [PowerPoint slides]. Toronto Metropolitan University.
- Dr Roy, K. (2023). Data Processing: Spark DS8003 – MGT OF BIG DATA AND TOOLS [PowerPoint slides]. Toronto Metropolitan University.
- Dr Roy, K. (2023). DISTRIBUTED COMPUTING, HADOOP, HDFS: DS8003 – MGT OF BIG DATA AND TOOLS [PowerPoint slides]. Toronto Metropolitan University.
- Kaggle. (n.d.). "LinkedIn Job Postings 2023 Dataset." https://www.kaggle.com/datasets/rajatraj0502/linkedin-job-2023