JOB MARKET ANALYSIS

Harshal Sanjiv Patil
UBID: hpatil2

Prathamesh Kishor Gadgil
UBID: pgadgil

Amey Jotiba Mangute
UBID : ameyjoti

I. Introduction:

Huge amounts of job-related data are generated regularly as a result of the automation of recruitment platforms, providing significant insights into skill requirements and trends in the job market. One of the major employment networking platforms, LinkedIn, provides a vast amount of structured and unstructured data that is easily used for machine learning and job market analysis.

The **1.3M LinkedIn Jobs & Skills dataset**, which includes job names, necessary skills, job summaries, and data including industry and location, was utilized in this research. For operations like skills tracking, job market analysis, and the generation of employment recommendation systems, the dataset is particularly relevant.

Our goal is to:

1. Perform exploratory data analysis (EDA) to understand the dataset and derive preliminary insights.
2. Use Apache Hadoop and Apache Spark to create a scalable data processing environment.
3. Insert the data into HDFS to get it prepared for further analysis and machine learning.

II. Problem Statement:

We aim to address the following machine learning problems using the dataset:

1. Job Category Classification: Apply classification techniques to identify a job's industry or category using its description and required skills.
2. Skill Demand Prediction: Apply regression models for predicting the demand for specific skills in a particular region or industry.
3. Job Clustering: To identify unseen job market trends, group similar job postings using clustering based on job descriptions and relevant skills.

III. Data Analysis Objectives:

1. Identify the Job Market Conditions: Analyzing the general structure and features of job postings in the LinkedIn dataset is the main goal. This involves studying how job titles are distributed, determining which positions are popular, and analyzing the frequency that these appear over the dataset.
2. Locate the Top Hiring Companies: Identifying which companies are providing the most job opportunities is the goal of the research with the goal to provide statistics on company demand. This makes it easier to figure out companies with high hiring volume and industry giants.
3. Analyze the Demand for Skills in Various Job Roles: The goal is to determine which technical and soft skills are frequently related to employment roles through analyzing job-related skills information. It helps us find the skills that companies are looking for in today's job market.
4. Job Opportunities by Location: Analyzing the regional distribution of job openings is the goal. Identifying where jobs are most likely to be located can be useful for job seekers.
5. Analysis and Cleaning of Data: Analyzing and cleaning the dataset by removing duplicates, unnecessary columns, and missing values is a basic goal.

IV. Results & Visualization:

The EDA offered an in-depth evaluation of the job market as it appeared in the LinkedIn dataset. Results clearly demonstrated the significant need for both technical and analytical positions, the priority of recruiting among top firms, and the effective cleansing of raw data for accurate analysis. The visual representations will be essential for predictive modeling and other data analysis tasks.

Figure 1: Top 15 Companies with Most Job Openings.

Job applicants can target companies with large recruitment networks using this visualization to quickly recognize corporate recruiting abilities.

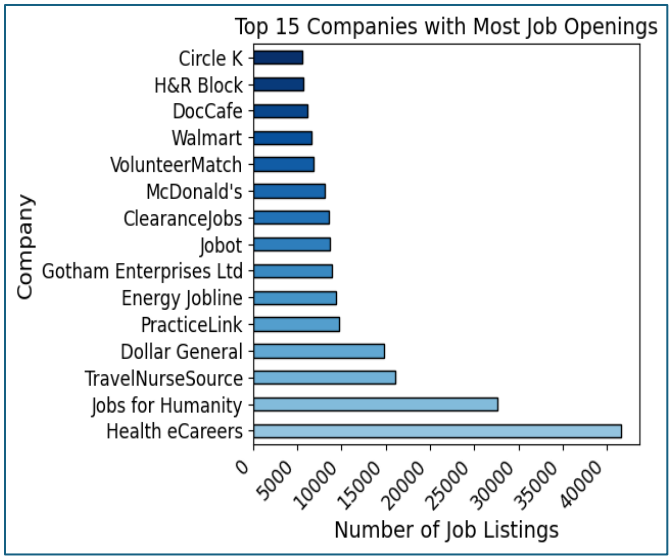


Fig 1

Figure 2: Top 20 Job Titles by Number of Job Opening:

Job popularity was represented by a vertical bar chart.

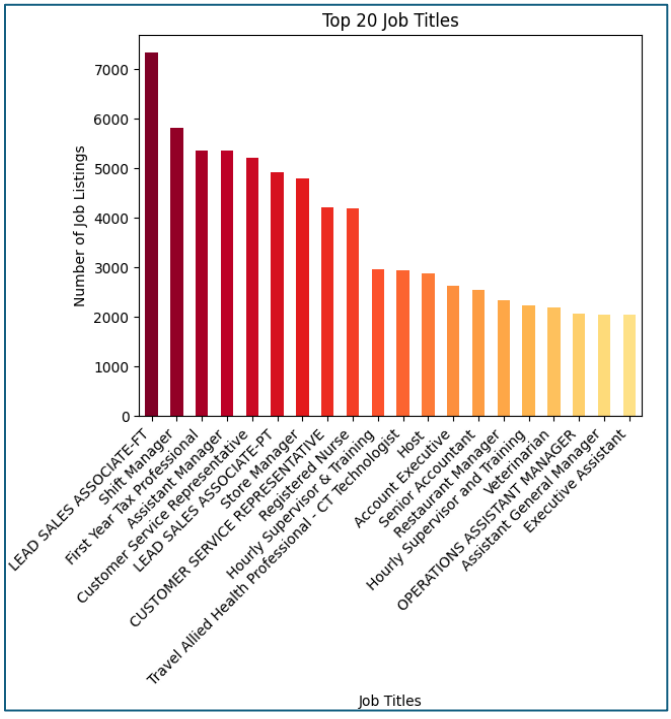


Fig 2

Figure 3: Distribution of Job Types

This bar graph shows how job types are distributed among the job listings.

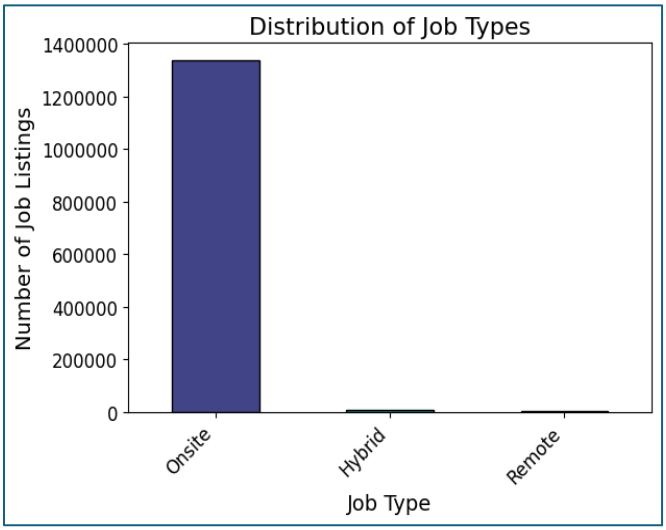


Fig 3

Figure 4: Top 10 Job Locations by Proportion

To show the Top 10 Job Locations by Proportion in the dataset a pie chart representation has been generated. New York shows the highest percentage of job listings with 13.73%.

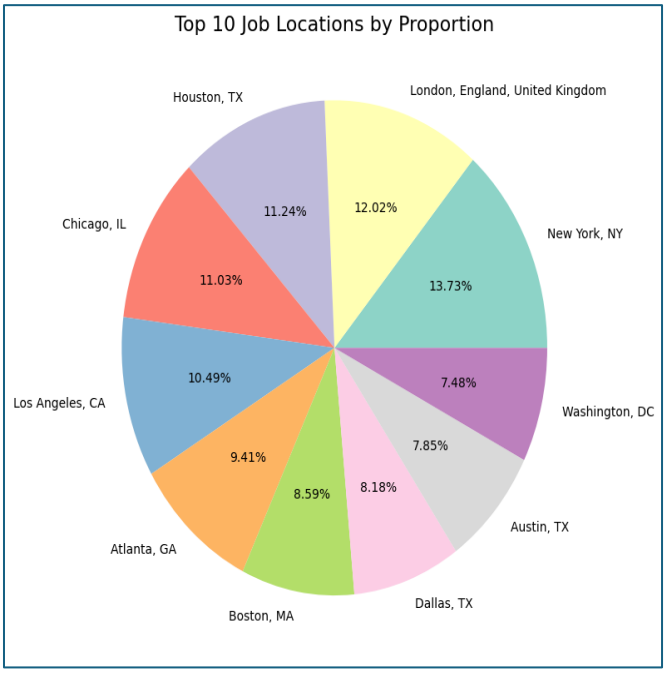


Fig 4

Figure 5: Analyze Skills Demand Across Job Roles

To show the most frequently demanded skills across all job positions a word cloud representation was generated. The most popular skills were problem-solving, communication, project management, Microsoft Office, and customer service.

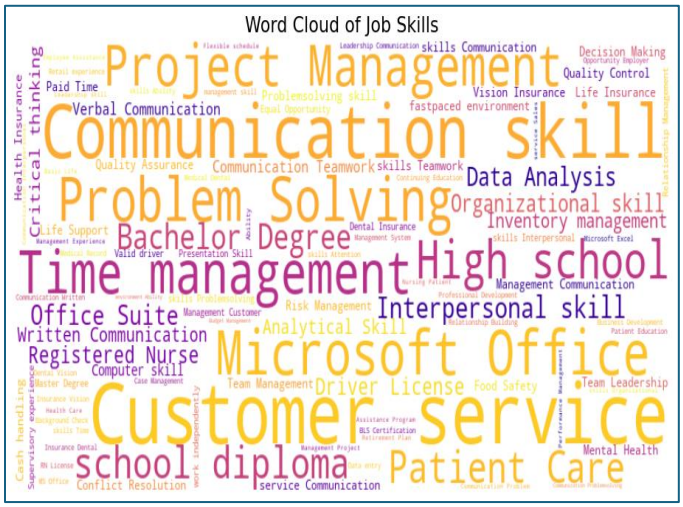


Fig 5

Figure 6: Cleaned Dataset Snapshot

The analysis was properly cleaned in order to assure its reliability and integrity. Missing values, duplicate entries, and irrelevant columns were evaluated in the dataset.

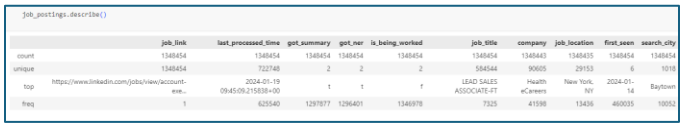
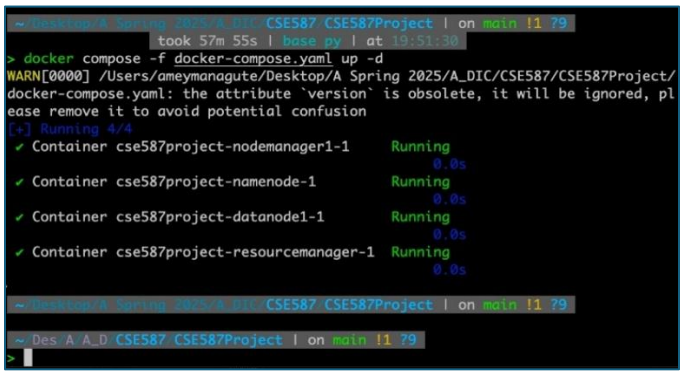


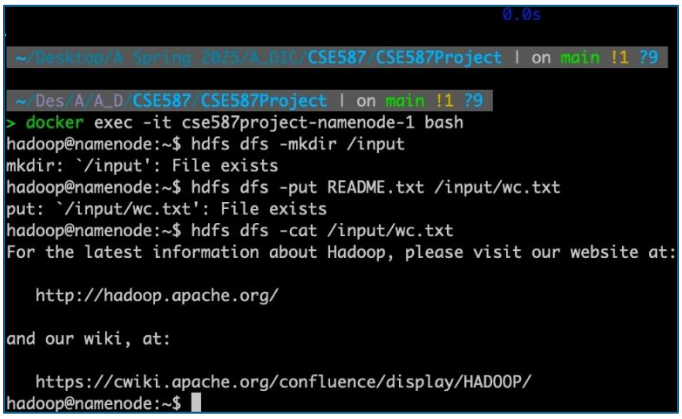
Fig 6

V. Hadoop Cluster Setup:

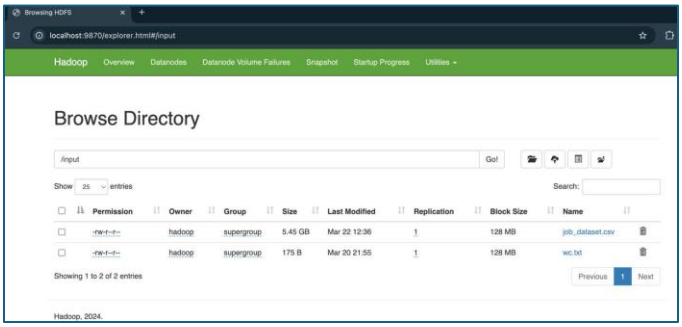
Docker compose and cluster status checkup:



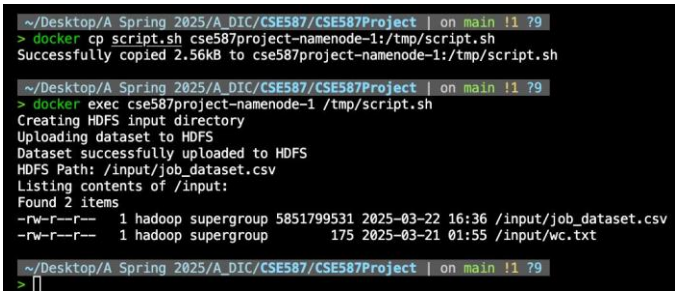
Directory creating and file upload:



Hadoop Cluster Directory:



Uploading dataset to HDFS:



VI. Views on the Ingestion and Storage of Data:

1. Utilizing hdfs dfs -put:

   For small to medium-sized datasets, such as the LinkedIn Jobs & Skills dataset, the **hdfs dfs -put** command provides an easy and effective way for uploading files to an HDFS cluster.
   It is simple to use and doesn't need any extra framework or tools. However, because of its single-threaded architecture and lack of

support for parallel processing, hdfs dfs -put may not be the best option for huge databases

2. Speed of Writing Using hdfs dfs -put:

   For this project, the hdfs dfs -put performance was sufficient; but, because due to its single-threaded architecture, it may be slower for larger datasets.
   To improve speed:
   1. Use **parallel uploads.**
   2. Optimize HDFS settings (e.g., by increasing block size).
   3. Use data ingestion tools.

3. Which Format Is Best for HDFS Data Storage:

   Because CSV format is easy and simple to use for initial exploration data analysis (EDA). But for operations including machine learning or comprehensive analysis, CSV is not the most suitable format. Improving query performance would be possible by converting the dataset to Parquet or ORC format.

VII. References:

1. https://github.com/UBCSE587/2025Spring-projectphase1
2. https://www.kaggle.com/datasets/asaniczka/1-3m-linkedin-jobs-and-skills-2024
3. https://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-common/FileSystemShell.html