CS226 (Mis)Fortune-500 Final Project Report
Exploring the Global Job Market: A Big Data Analysis of International Job Postings

SUMEDHA GIRISH ATREYSA, Student ID: 862395753, Email: satre002@ucr.edu
TEJAS MILIND DESHPANDE, Student ID: 862393211, Email: tdesh006@ucr.edu
MANISH DEEPAK CHUGANI, Student ID: 862388066, Email: mchug002@ucr.edu
NUNNA LAKSHMI SARANYA, Student ID: 862394536, Email: nsara014@ucr.edu
XINLE CHEN, Student ID: 862334534, Email: xchen440@ucr.edu

The advancement of Artificial Intelligence Technologies has taken the world of Information Technology Software by storm. Job security and procurement have developed an air of uncertainty and a fear of the unknown. Can AI replace human developers? Is the advent of Large Language models combined with the current recession causing a reduction in the number of jobs for Software Developers, Data Scientists, Data Analysts, and other such software-oriented job titles? The proposed approach aims to provide users with a query engine that allows them to answer their burning questions without having to do an in-depth analysis involving data procurement, pre-processing and integration. Using various data sources, we accumulated the data into a single data store, allowing the user to query the job postings data store within minutes using the power of Big Data Technology. This project has used cloud-based processing technologies and in its current state, does not provide a fully functional front-end. However, the developers of the proposed approach have a functional back-end that uses the power of cloud-based processing (Google Colab) and provides them the platform to provide analytics as shown in the results section.

Additional Key Words and Phrases: Big Data Management, Data Warehousing, Data Integration, Data Pre-processing, PySpark

ACM Reference Format:
Sumedha Girish Atreysa, Tejas Milind Deshpande, Manish Deepak Chugani, Nunna Lakshmi Saranya, and Xinle Chen. 2023. CS226 (Mis)Fortune-500 Final Project Report Exploring the Global Job Market: A Big Data Analysis of International Job Postings. In . ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

In this current interconnected world, the global labor market presents a unique opportunity for both professionals and businesses to tap into a diverse and skilled talent pool. International job listings provide valuable insights into the evolving nature of employment trends, reflecting the geographical shifts in demand, skill requirements, and sector-specific needs. Understanding these dynamics is crucial for both job seekers and companies to navigate the complexities of this ever-changing landscape.

Data gleaned from global job listings offers a lot of information that shows how employment opportunities across different locations are distributed, which types of positions are currently in high demand, and which qualifications are

Manuscript submitted to ACM

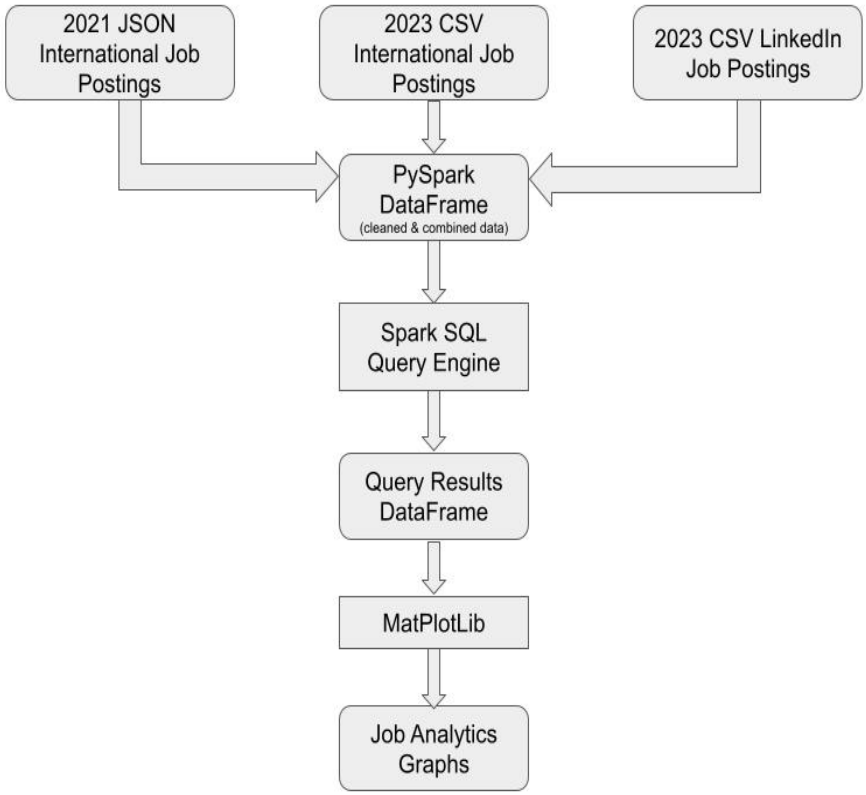

Fig. 1. Proposed Approach: MisFortune-500

essential in distinct global regions. The labor market, by its very nature, is a dynamic and adaptable system. Grasping these trends is of immense importance to policymakers, educators, career counselors, and anyone seeking employment.

The proposed approach aims to delve into the global labor market, uncovering emerging trends and providing actionable insights to guide companies and job seekers in their decision-making processes. To enhance the effectiveness of career matching and recruiting on an international scale, the proposed idea is to explore and extract valuable information from data on international job postings.

The objectives of the proposed idea is:

(1) The proposed approach performs a thorough and perceptive examination of the worldwide labor market with a particular emphasis on employment titles like "Software Development Engineer," "Data Analyst," "Data Scientist," and "Data Engineer," amongst others. The proposed approach provides the existing back-end. By leveraging BigData Technologies incorporated into the back-end, a user can query a data store containing job postings from 2021 and 2023 across international job boards.

(2) The proposed approach also performs a comparative analysis of the data from 2021 and 2023 job postings. The results of this investigation show how the global labor market has changed since September 2021. It pinpoints

the changes in upcoming skills for the chosen job roles, variations in demand, and the growth or decline of requirements for any specific employment prospects.

The pipeline for the proposed approach is shown in Figure 1. The proposed approach builds a query processing framework which allows fast and efficient query execution on voluminous datasets. The framework also heavily outperforms traditional RDBMS architectures like MySQL as will be shown in Section 5.

The remainder of this report is structured as follows. Section 2 delves into the existing body of research relevant to our project. Section 3 describes the data used in this study and the preprocessing steps taken to prepare it for analysis. Section 4 describes the architecture of the system that will generate the job market analytics from the job postings data. Section 5 provides two evaluation metrics and the results based on them. Section 6 describes the future work and limitations of scalability of voluminous datasets. It also brings to light a lack of processing power available to the developers of the proposed approach. Section 7 summarizes the results and provides information of how challenges that were faced by the developers were overcome and can be mitigated. Section 8 outlines the contributions of each author to the MisFortune-500.

2 RELATED WORK

The literature survey is provided in Appendix A.

2.1 Big Data Application in Job Trend Analysis: [6]

The authors make use of Big data techniques such as Hadoop and Tableau are applied to identify job trend analysis in New York. The data set studied comprises of several gigabytes. Handling such huge data can be difficult with conventional resources. Hadoop proves its value with increased scale of data. Hence, Hive and Hadoop are applied to the study. The technologies used by the author are Hadoop, Hive, CSV Cleaning, and Tableau. The authors have evaluated their work by employing Data Visualization of Job Market trends using Big Data Technologies.

2.2 Big Data and Labor Markets: A Review of Research Topics: [13]

In this paper the assessment of the current research in big data for the labor market has been compiled to detect the research gaps and generate future research directions. A systematic review of the literature was adopted to address the study questions posed for identifying, assessing, and interpreting all available research material to address specific research issues. The authors have paid special importance to the implications of technological advancement for industry and labor market, education for the new labor market & skills for new age labor markets, tools and methods used for labor market analysis.

2.3 Better understanding of the labour market using Big Data:[19]

The article underscores the need to utilize Big Data for a deeper understanding of the labor market in the digital age. Online job postings offer valuable information on required skills, calling for the development of tools to analyze this data and facilitate better job matching. The proposed approach uses Big Data for labor market analysis, focusing on skills in job postings, skill mismatches, and improving decision-making. Data is collected from web sources to identify skill gaps and enhance the labor market experience for job seekers and employers. The authors made use of Scrapy, Airflow Task Schedule, Artificial Intelligence and Universal Sentence Encoder to achieve their tasks. They successfully evaluated their model by finding the correlation between the demand for skills and employability.

2.4 Big Data For Labor Market Intelligence:[10]

The problem defined in this work is to create a data-driven system of job market intelligence to forecast employment trends, pinpoint skills that are in demand, and evaluate the effects of digitalization. Discuss important issues pertaining to the expansion of professions and the role that soft skills play in current employment. The proposed approach selects, preprocesses and transforms data sources using statistical, technical and domain expertise criteria. It then uses data mining techniques such as Knowledge Discovery in Databases(KDD) to infer analyses of the data sources. It uses a data lake framework to query data as a data scalability solution. The authors have made use of the following technologies to reach their goals:Java, Selenium, Scrapy, MongoDB, HBase, Cassandra, HDFS, Map Reduce, and Yarn. The authors evaluate the Labor Market Intelligence analysis using statistical, technical and domain expertise paradigms on structured and unstructured raw data.

2.5 Classifying Online Job Advertisements through Machine Learning:[3]

The problem defined in this work utilizes titles and job descriptions of job postings and uses Machine Learning techniques in order to build a knowledge base for the Web Labor Market. The proposed approach uses a machine learning model for classifying multilingual Web job vacancies exploiting a single-label classifier using both titles and descriptions. It uses box plots to evaluate the f1-score measure of each classification algorithm over the first-digit of the ISCO taxonomy, which identifies 9 distinct occupation groups. Finally, it builds a Knowledge Graph with occupations and skills as nodes and the linkages between them as the edges.

2.6 Classifying the Proposed Approach

The top 5 most relevant related works to the Proposed Approach were outlined above. We now dive into classifying the Proposed Approach into the categories mentioned in Appendix A. Our Proposed Approach from here on out will be labeled as MisFortune-500, depicting the name of our group. The MisFortune-500's problem definition is the development of a scalable architecture that not only compares and demonstrates that BigData Technologies like Spark overpower the capabilities of tradition RDBMS systems but also provides Data Analytics comparisons between the Job Market in 2021 and 2023. MisFortune-500's Proposed Approach Methodology utilizes datasets hosted in various data warehouses by integrating, pre-processing and loading over 50 GB of data from various data sources into a single PySpark Dataframe and then performs SparkSQL queries to obtain the Data Analytics for the same. It also tries to load this data into an RDBMS architecture in order to draw a comparison as to why BigData Technology is needed in today's ever-growing Data horizon. The technologies used in the MisFortune-500 are PySparkSQL, Python, MatplotLib for Visualization and Python Dataframes for consolidation of results. The Factors Considered for the MisFortune-500 are execution time of reading and executing queries on BigData which has Volume, Veracity and Value.

3 DATA DESCRIPTION AND PREPROCESSING

The MisFortune-500 utilizes a data store comprised of three datasets which have been integrated using the integration techniques of the MisFortune-500. The datasets, pre-processing and cleaning procedures are described in this section: International_Job_Postings_2021 [18] (techmap-jobs-dump-2021-09.json(50 GB)), Linkedin_job_postings_2023 [14] (job_postings.csv(130 MB)) and International_Job_Postings_2023 [15] (allJobs.csv(135 MB)). The first file contains 2021 job postings in JSON format, while the second file holds 2023 job postings in CSV format. The third file is also a CSV format with 2023 job postings from a different source(International Job Boards).

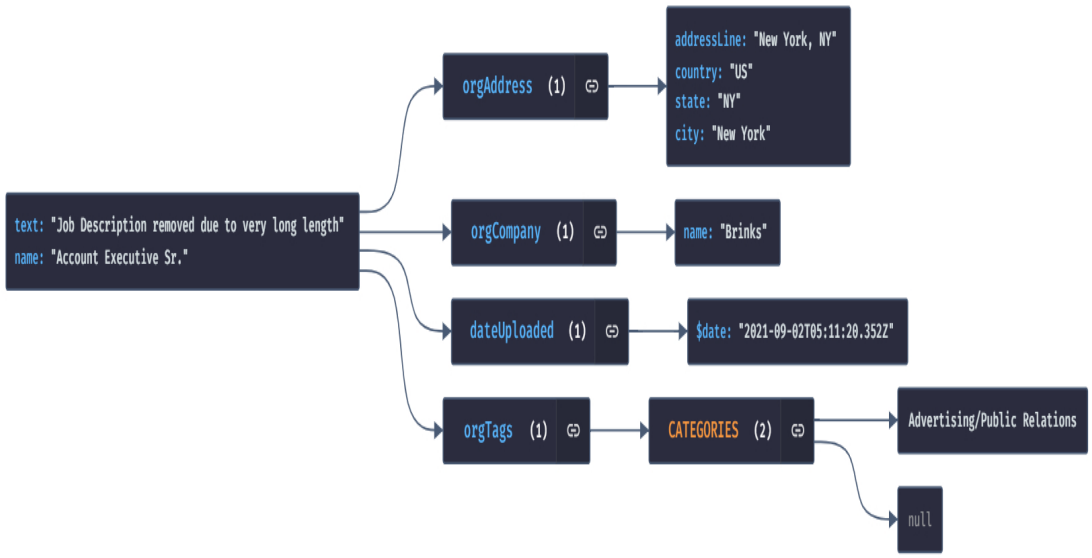

Fig. 2. Structure of Relevant Data for 2021 after initial Pre-processing

JobDate	JobLocation	HiringCompany	JobCategory	JobTitle
2021-09-02T05:11:20.352Z	New York, NY	Brinks	Advertising/Public Relations	Account Executive Sr
2021-09-02T05:11:21.117Z	San Francisco, CA	Konica Minolta	Advertising/Public Relations	Global Client Executive
2021-09-02T05:11:21.267Z	Stevens Point, WI	Sentry	Advertising/Public Relations	Retirement Plan Sales Support Analyst
2021-09-02T05:11:21.688Z	Malvern, PA	Vanguard	Advertising/Public Relations	Digital Channel Product Owner
2021-09-02T05:11:21.968Z	Chicago, IL	Skills For Chicagoland's Future	Advertising/Public Relations	Account Coordinator - HUB International

Fig. 3. Resulting Spark Dataframe sample after Pre-processing & Cleaning

Figure 2 shows the nested structure of International_Job_Postings_2021 [18] after initial pre-processing and Figure 3 shows the structure of the final DataFrame being queried for Visualization purposes.

This work also referred to Data_Analyst_Job_Postings[16] but the data was skewed for a single role and did not seem relevant due to the amount of cleaning and pre-processing required for incorporation into the data store. The fields for the dataset were not coherent with the fields of the other datasets being considered. Another dataset that was referred to was the LinkedIn_Job_Data for 2023[17] but was not incorporated due to the same reasons as described above.

3.1 Data Storage

The data is first stored as a PySpark Dataframe. This structure allows for efficient analysis and manipulation of the data. Additionally, we persist the structured data on-disk for reuse, This eliminates the need for repeated processing during subsequent analyses. Despite persisting the data on disk, python's Pandas Dataframe was unable to parse the data into a single dataframe without crashing thereby demonstrating the need for Big Data Frameworks.

```
query_string = """SELECT COUNT(*) AS JobCount,
    (CASE
        WHEN job_title LIKE "%Software%" THEN 'Software' WHEN job_title LIKE "%Data%" THEN 'Data' WHEN job_title LIKE "%Machine Learning%" THEN 'Machine Learning'
        WHEN job_title LIKE "%Artificial Intelligence%" THEN 'Artificial Intelligence' WHEN job_title LIKE "% AI %" THEN 'Artificial Intelligence'
        ELSE 'Other'
    END) AS JobCategory,
    (CASE
        WHEN date_created LIKE "%2021%" THEN '2021' WHEN date_created LIKE "%2023%" THEN '2023'
        ELSE '2022'
    END) AS Year
FROM
    JobPostings
WHERE
    (job_title LIKE "%Software%" OR job_title LIKE "%Data%" OR job_title LIKE "%Machine Learning%" OR
    job_title LIKE "%Artificial Intelligence%" OR job_title LIKE "% AI %")
GROUP BY
  JobCategory, Year
ORDER BY
    JobCount DESC;"""

cursor.execute(query_string)

end = time.time()
print(f"Time to insert records into MySQL and execute query : {(end - start):.2f} seconds")

for rec in cursor.fetchall():
  print(rec)

Mounted at /content/drive/
Time to insert records into MySQL and execute query : 3677.02 seconds
```

Fig. 4. Query Time for MySQL on the combined Dataframe: 3677.02 seconds ≈ 1 hour

3.2 Data Pre-Processing

MisFortune-500 performs data pre-processing and cleaning before integration into a consistent format from different data sources. Each data source was Few Redundant and trivial fields (namely , "job_id", "company_id", "max_salary", "med_salary", "min_salary", "pay_period", "formatted_work_type", "applies", "remote_allowed", "views", "job_posting_url", "application_url", "application_type","expiry", "closed_time", "formatted_exerience_level", "skills_desc", "posting_domain", "sponsored", "work_type", "currency", "compensation_type", "scraped", "listed_time") were removed from the job postings to reduce the size of the PySpark Dataframe. This optimization allows for faster and more efficient querying. Furthermore, duplicate job postings were identified and eliminated, ensuring the integrity and accuracy of the data. These pre-processing steps structure the data for effective and efficient analysis.

After obtaining the final integrated, cleaned and pre-processed dataframe using the methods described above and in the following section, MisFortune-500 also compares the performance of an RDBMS on this dataframe with the performance of PySpark as a Big Data Technology. The results demonstrate the need for said technologies as shown in Figure 4 since MySQL takes over an hour to read the data and process the query. As is shown later in the Evaluation section, Spark takes nearly 11 minutes for the same task.

4 IMPLEMENTATION/METHDOLOGY

This section focuses on the architecture of the system that will generate the job market analytics from the job postings data. This system will be implemented in the following parts / phases:

(1) Combine and transform this data into a consistent format.
(2) Query the data (for retrieving job records for specific category).
(3) Visualize the data

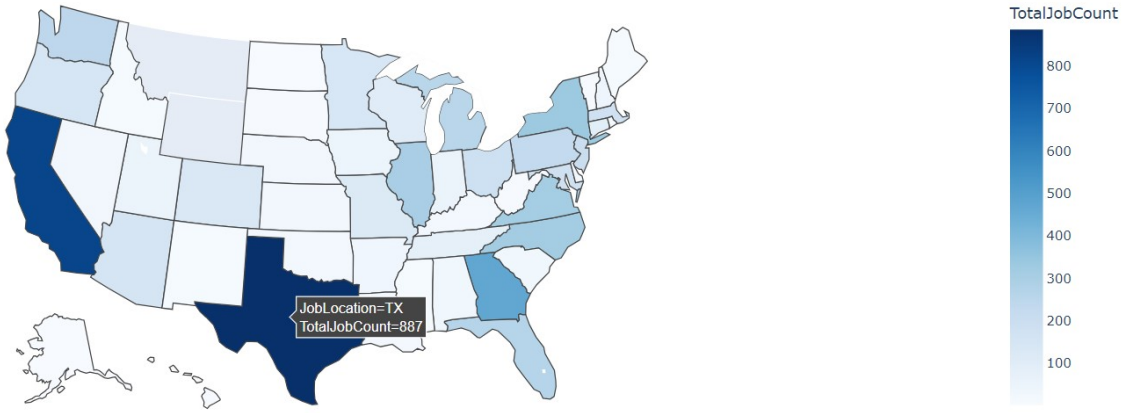

Fig. 5. State-wise Job Count Choropleth in 2021 for USA

4.1 Combine and Transform data

The process of combining and transforming data serves as a critical precursor to our analytical endeavors, with the overarching objective of establishing a robust and well-structured data foundation for subsequent analysis and exploration. Specifically, our methodology involves the transformation of individual datasets from the years 2021 and 2023 into clean, coherent dataframes. This transformation ensures that the data is standardized, cleansed of inconsistencies, and organized into a format conducive to meaningful analysis. Moreover, the integration of data from both 2021 and 2023 is executed with precision to create a unified dataset. This integration is performed in a consistent and coherent manner, emphasizing the retention of common columns or fields shared between the two datasets.

By harmonizing the datasets in this manner, we ensure that the resulting combined dataset maintains data integrity and relevance, thereby laying the groundwork for a seamless exploration of trends, patterns, and insights spanning the specified time frame. The meticulous approach to data combination and transformation is pivotal in guaranteeing a comprehensive and well-structured foundation that maximizes the analytical potential of the integrated dataset.

4.1.1 Data Transformation.

- **Data Integration:** MisFortune-500 integrated data from diverse source files into separate PySpark Dataframes. This enabled parallel processing and manipulation of the data from various sources.
- **Field Matching:** Subsequently, each Dataframe underwent transformations to achieve a consistent and structured format. This involved identifying and matching relevant field attributes across the various data sources. Matching these attributes ensures compatibility and facilitates subsequent data merging.

4.1.2 Combine Data.

- **Merging Dataframes**: Once all Dataframes have a consistent structure, they are merged into a one unified Dataframe. This gave us a comprehensive view of the combined data from all sources from 2021 and 2023.
- **Structured Data Access**: The final structure of the merged Dataframe served as the foundation for subsequent data retrieval. MisFortune-500 leverages this structured format to efficiently query and extract specific job postings based on criteria provided in the SparkSQL queries.

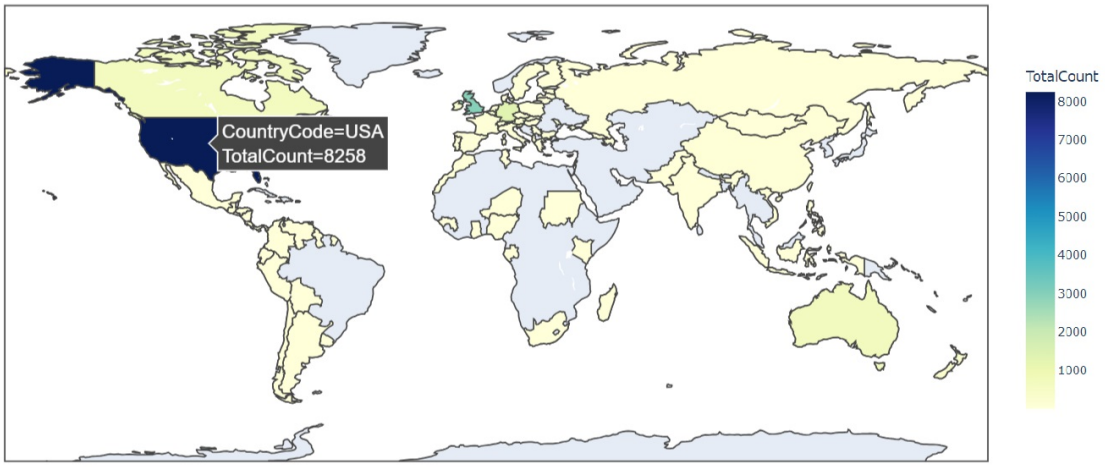

Fig. 6. Country-wise Job Count Choropleth for 2021

4.2 Data Querying

The proposed approach employs PySpark SQL as a pivotal tool for data querying, enabling the extraction of comprehensive analytics and visualizations. This strategic use of PySpark SQL is integral to our overarching objective of revealing profound insights into the intricate dynamics of the job market. By leveraging the advanced capabilities of PySpark SQL, our methodology allows for a nuanced exploration of vast datasets, empowering us to discern patterns, correlations, and anomalies within the employment landscape.

Furthermore, the proposed approach is specifically tailored to identify emerging trends within the job market. Through meticulous data analysis and visualization techniques facilitated by PySpark SQL, Pandas and Matplotlib we aim to offer a forward-looking perspective on the evolving employment landscape. This approach not only enhances our understanding of current market dynamics but also equips us with the foresight needed to adapt and respond proactively to emerging trends, thereby facilitating more informed decision-making in the realm of workforce management and strategic planning.

4.2.1 PySpark SQL for Data Queries. :

The employed methodology centers around harnessing the capabilities of PySpark SQL to query and extract pertinent information from the dataset. By leveraging PySpark SQL, our approach facilitates efficient querying based on specific criteria, enabling a targeted exploration of the data. This tool proves instrumental in not only streamlining the extraction process but also in generating a diverse range of insights derived from the intricacies of the dataset. The efficiency of PySpark SQL in handling large volumes of data ensures that our querying process is both scalable and effective, laying the foundation for a more comprehensive understanding of the underlying patterns and trends within the dataset.

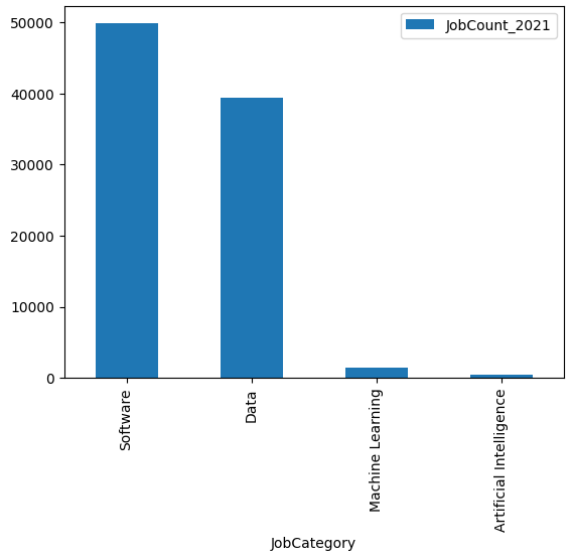

Fig. 7. Job Count by specialization in 2021

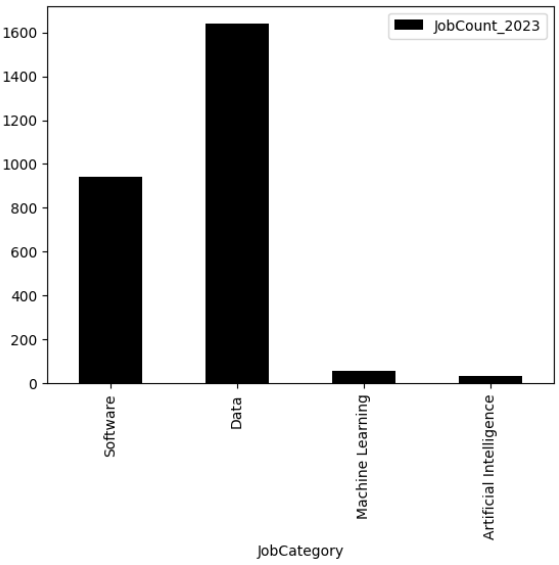

Fig. 8. Job Count by specialization in 2023

As shown by the Choropleths in Figure 5 and Figure 6, PySpark is able to execute these queries on the 2021 50 GB dataset within minutes.

4.2.2 *Analytics and Visualizations.* :

The extracted data was utilized to generate valuable analytics and visualizations, including:

- Job Postings by Position: The number of job postings for various positions that were analyzed, providing us insights into current market trends and workforce demands. Examples include Software Engineers, Data Analysts, Machine Learning Engineers, etc.
- Job Postings by Job Location: The analysis counts the number of jobs by location for two levels of hierarchy. One analysis produces number of jobs per US State and visualizes it on the map of the United States. The other analysis produces number of jobs per country of the world and visualizes it on the map of the World.
- Date-Based Trends: Job posting trends over time were analyzed to identify potential shifts in demand or fluctuations in specific job categories. This longitudinal analysis can provide valuable insights into the evolving workforce landscape.

  We also perform analyses on various job titles as shown in Figure 7, 8 and 9 to depict the changes in demand of skillsets in the Job Market in the field of Computer Science as well as the sharp decline in the number of job postings for the same categories of Job designations.

5 EVALUATION

Table 1 shows the results based on the evaluation metrics defined below:

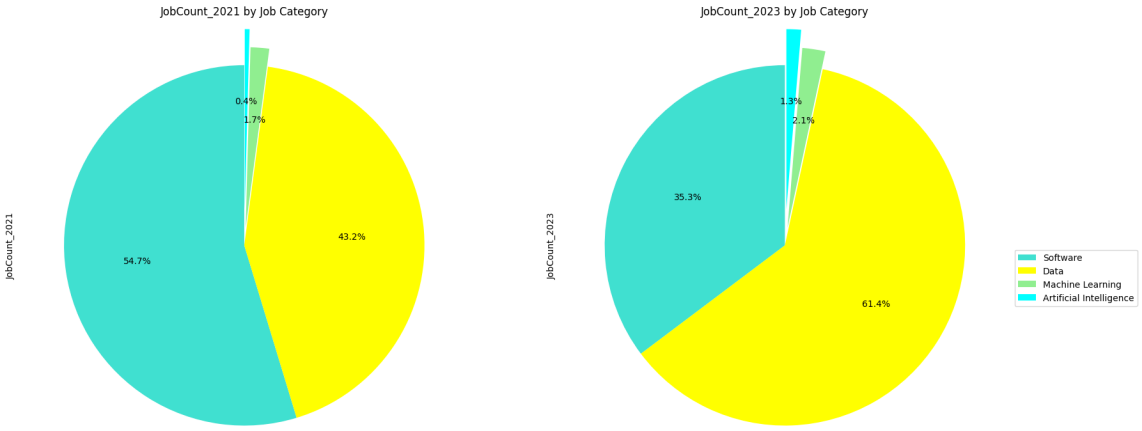

Fig. 9. Pie Charts for comparison of distribution of jobs across specializations

Dataset	Dataset Size	Query Time for SparkSQL	Query Time for RDBMS
2021 International Job Postings[18]	46.9 GB	10.54 minutes (12 GB RAM TPU)	59.35 minutes (12 GB RAM TPU)
2023 LinkedIn Job Postings[14]	69.8 MB	6.84 seconds (12 GB RAM TPU)	1.30 minutes (12 GB RAM TPU)
2023 International Job Postings[15]	133 MB	9.62 seconds (12 GB RAM TPU)	1.72 minutes (12 GB RAM TPU)
Curated Dataset (Combined & Cleaned)	42 GB	13.12 minutes (12 GB RAM TPU)	61.28 minutes (12 GB RAM TPU)

Table 1. Comparison of PySparkSQL versus MySQL

(1) As indicated in previous versions of the proposed approach, the first evaluation metric was a comparison of the capabilities of PySpark (A Big Data Technology) versus a regular Pandas Dataframe. However, due to the Pandas DataFrame not being able to parse the data due to its voluminous nature, the time taken for Spark (around 11 minutes) clearly demonstrates that Big Data Technologies make impossible tasks feasible.

(2) A comparison of a traditional RDBMS infrastructure (MySQL) versus PySpark is shown in Table 1. Spark reads, executes and outputs the results in around 13 minutes and the RDBMS takes over one hour just to read the data. However, the querying time for RDBMS is low since it performs in-memory operations. The comparison shows how SparkSQL has been particularly designed to leverage the advantages of the MySQL and Spark frameworks.

(3) Visualization plots using matplotlib are shown throughout the report and demonstrate the strong analytical power of Voluminous data and how it can depict trends of the Job Market in any particular year, trends in the United States, trends across the world and across different Job Titles indicating the trends of various technologies being deployed in the industry.

(4) The claim made that the number of job postings would have a direct correlation with the lack of jobs and the recession is successfully proved by the experiments performed using the MisFortune-500. Various figures and results demonstrate this to a satisfactory level.

6 FUTURE WORK

In the future, this project will evolve to include a wider range of employment positions and geographical areas, broadening its analytical reach. This more comprehensive viewpoint will offer a deeper comprehension of the complexities present in the employment sector, highlighting the subtleties connected to various professions and local dynamics. The initiative intends to provide perspectives that are more comprehensive and representative of the many elements impacting the employment market by expanding the inclusiveness of the dataset.

6.1 Limitations

During 2021, since the job market was booming and there were a plethora of job postings, especially in the Computer Science domain, it was possible to procure a voluminous dataset of job openings measuring 50 GB in size. But since 2022, there has been a steady decline in the number of job postings due to inflation, the recession and geopolitical factors. Thus, the data availability of job postings for the years 2022 and 2023 is extremely limited.

Since for the year 2022, we were unable to find good dataset(s) with enough job postings, it is not possible to convey the difference in job trends in 2022 compared to 2021 and 2023. Currently, it is also very difficult to scrape the data for 2022, since verifying the veracity of the data is difficult and the scraped job postings might be stale or the positions might be cancelled by the organizations due to economic downturn.

Given enough verified data, the project can definitely scale up to larger datasets and show the global job trends more accurately and precisely in any domain such as tech, manufacturing,medicine,etc.

6.2 Web Application

Due to resource limitations, MisFortune-500 is unable to host a fully functional web application as proposed in the report outline. However, the implemented design and visualization plots effectively demonstrated the capabilities of Big Data Technologies such as PySpark. Through the insightful visualizations, valuable trends of the job market landscape were obtained and this provides clear indication of a lack of job postings in the Job Market in 2023, thereby indicating a recession and fulfilling the research objectives.

A query engine front-end was attempted in development to provide the user with the ability to query the data store of MisFortune-500 as shown in Figure 10. However, due to the lack of local infrastructure resources and the infeasible nature of the task of integration of Front-End Technologies with cloud infrastructure providers like Google Colab, the effort did not yield the desired fruit. However, the attempt that was made has been demonstrated through images and is also present in the code files.

7 CONCLUSION

The MisFortune-500 outlined in this work is a successful attempt at designing a fast query engine using Big Data Technologies (PySpark, SparkSQL) that allows job seekers and employers alike to get valuable insights from the data available for the current job market. A comparison can also be drawn to the job market in previous years with insightful visualization techniques depicted by the MisFortune-500 architecture. The MisFortune-500 can be commercialized and provided to the industry if given the required computing power resources such that a front-end and back-end can be developed locally.

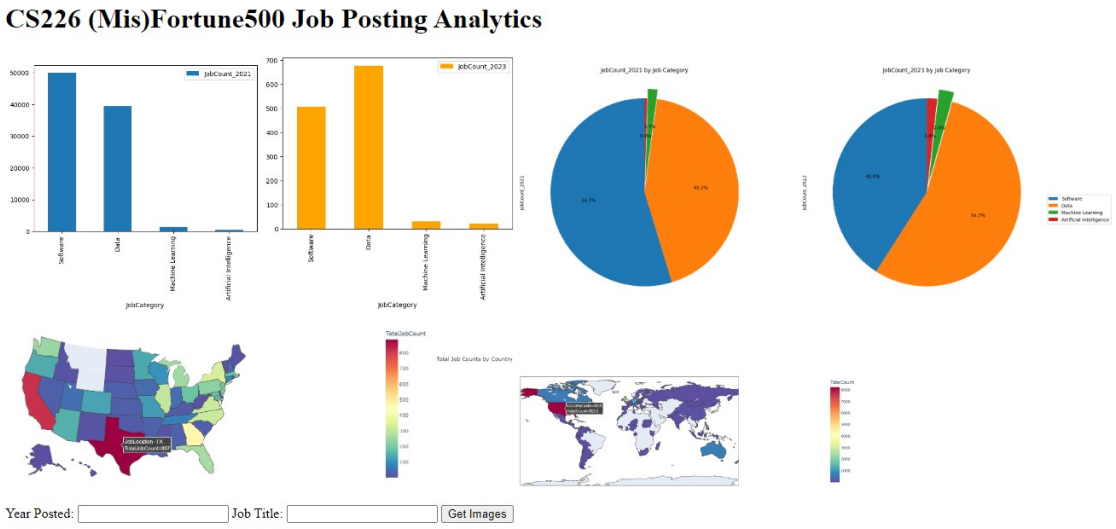

Fig. 10. In-Progress FrontEnd for MisFortune500

Thus, the MisFortune-500 architecture deploys the power of Big Data Technologies to provide insights on the current Job Market and helps clear up the clouds of uncertainty that loom over any job seeker's head when searching and applying for a job.

"If a job opportunity doesn't knock, build a Data Analysis door." ~ (Mis)Fortune-500.

8 AUTHOR CONTRIBUTIONS

All members contributed equally to the preparation of the report, presentation and literature survey selection of works to be included.

(1) **Xinle Chen** was responsible for the comparison of the MisFortune-500 with the Pandas dataframe. Xinle also was responsible for the background research of the problem and implementing the Front-End.

(2) **Sumedha Girish Atreysa** was responsible for the Data Description, Preprocessing and Storage as well as contributions to integration of the back-end with the front-end.

(3) **Nunna Lakshmi Saranya** was responsible for Combining and Transforming the Data for Query processing. She was also a contributor for the integration of the back-end with the front-end and creating the literature survey table.

(4) **Tejas Milind Deshpande** was responsible for the comparison with the traditional RDBMS architecture. The incorporation of MySQL into Jupyter and insertion of the data into the framework was carried out by Tejas. He also performed some query processing and obtained valuable insights for timing comparisons alongside strong work in deciding the schema for the final dataframe with help from Sumedha and Manish. Tejas was also responsible for the histogram visualizations.

(5) **Manish Deepak Chugani** was responsible for the ideation of the MisFortune-500 and led the architecture design of the query engine. He also provided a motivation to solve the problem since the group themselves

were struggling with employment opportunities and wanted to quell their fears. Manish also performed query processing and Data Visualization of the Choropleths and pie chart plots.

REFERENCES

[1] Armin Alibasic, Himanshu Upadhyay, Mecit Can Emre Simsekler, Thomas Kurfess, Wei Lee Woon, and Mohammed Atif Omar. 2022. Evaluation of the trends in jobs and skill-sets using data analytics: a case study. *Journal of Big Data* 9, 1 (2022), 32.

[2] Anja Bauer, Tobias Hartl, Christian Hutter, and Enzo Weber. 2021. Search processes on the labor market during the Covid-19 pandemic. In *CESifo forum*, Vol. 22. München: ifo Institut-Leibniz-Institut für Wirtschaftsforschung an der …, 15–19.

[3] Roberto Boselli, Mirko Cesarini, Fabio Mercorio, and Mario Mezzanzanica. 2018. Classifying online job advertisements through machine learning. *Future Generation Computer Systems* 86 (2018), 319–328.

[4] Andrea De Mauro, Marco Greco, Michele Grimaldi, and Paavo Ritala. 2018. Human resources for Big Data professions: A systematic classification of job roles and required skill sets. *Information Processing & Management* 54, 5 (2018), 807–817.

[5] Erica L Groshen and Harry J Holzer. 2021. Labor market trends and outcomes: What has changed since the Great Recession? *The ANNALS of the American Academy of Political and Social Science* 695, 1 (2021), 49–69.

[6] Priyanka Kale and Shilpa Balan. 2016. Big data application in job trend analysis. In *2016 IEEE International Conference on Big Data (Big Data)*. IEEE, 4001–4003.

[7] Ioannis Karakatsanis, Wala AlKhader, Frank MacCrory, Armin Alibasic, Mohammad Atif Omar, Zeyar Aung, and Wei Lee Woon. 2017. Data mining approach to monitoring the requirements of the job market: A case study. *Information Systems* 65 (2017), 1–6. https://doi.org/10.1016/j.is.2016.10.009

[8] Kai S Koong, Lai C Liu, and Xia Liu. 2002. A study of the demand for information technology professionals in selected internet job portals. *Journal of Information Systems Education* 13, 1 (2002), 21–28.

[9] Raymond Blanch Mbah, Manjeet Rege, and Bhabani Misra. 2017. Discovering job market trends with text analytics. In *2017 International Conference on Information Technology (ICIT)*. IEEE, 137–142.

[10] Mario Mezzanzanica and Fabio Mercorio. 2019. Big data for labour market intelligence: An introductory guide. *European Training Foundation* (2019).

[11] Pekka Pääkkönen and Daniel Pakkala. 2015. Reference architecture and classification of technologies, products and services for big data systems. *Big data research* 2, 4 (2015), 166–186.

[12] David Smith and Azad Ali. 2014. Analyzing computer programming job trend using web data mining. *Issues in Informing Science and Information Technology* 11, 1 (2014), 203–214.

[13] Lejla Turulja, Dalia Suša Vugec, and Mirjana Pejić Bach. 2023. Big Data and Labour Markets: A Review of Research Topics. *Procedia Computer Science* 217 (2023), 526–535.

[14] https://www.kaggle.com/datasets/arshkon/linkedin-job-postings/data 2023. LinkedIn Job Postings - 2023.

[15] https://www.kaggle.com/datasets/dilshaansandhu/international-jobs-dataset 2023. International Job Postings 2023.

[16] https://www.kaggle.com/datasets/lukebarousse/data-analyst-job-postings-google-search Nov 2022 - Present. Data Analyst Job Postings.

[17] https://www.kaggle.com/datasets/shashankshukla123123/linkedin-job-data 2023. LinkedIn Job Data.

[18] https://www.kaggle.com/datasets/techmap/international-job-postings-september-2021 2021. International Job Postings 2021.

[19] Alena Vankevich and Iryna Kalinouskaya. 2021. Better understanding of the labour market using Big Data. *Ekonomia i prawo. Economics and law* 20, 3 (2021), 677–692.

[20] Wei Lee Woon, Zeyar Aung, Wala AlKhader, Davor Svetinovic, and Mohammad Atif Omar. 2015. Changes in occupational skills-a case study using non-negative matrix factorization. In *Neural Information Processing: 22nd International Conference, ICONIP 2015, Istanbul, Turkey, November 9-12, 2015, Proceedings Part III 22*. Springer, 627–634.

Appendix A

Paper Title	**Author(s)**	**Problem Definition**	**Proposed Approach Methodology**	**Technologies Used**	**Factors Considered Important/Evaluation Metrics**
Evaluation of the trends in jobs and skill-sets using data analytics: a case study[1]	Armin Alibasic et. al.	Development of an innovative data-driven method using a case analysis in the energy sector to pinpoint trending employment.	Data Scraping and Data Preprocessing using Stop Word Removal and TF-IDF scores. Latent Semantic Analysis, LDA and NMF.	Flask, Bootstrap, PythonAnywhere, Data Mining Techniques	Demand of required skill sets in the oil and gas industry. Leveraging Data Mining Techniques to obtain trends in data. Studying the actual market behavior and future needs based on real data acquired.
Search Processes on the Labor Market during the COVID-19 Pandemic[2]	Anja Bauer et. al.	Examines the effects of the pandemic recession on young workers, whom we define as being 21–30 years old, relative to older age groups.	Regression model for comparison of how different worker characteristics affect the probability of (full-time) employment over time.	Statistical Data Mining Techniques such as Regression Analysis	How features affect employment rate and levels in the job market. Features include age, gender, race, etc.
Human resources for Big Data professions: A systematic classification of job roles and required skill sets[4]	Andrea De Mauroa et. al.	A novel, semi-automated, fully replicable, analytical methodology based on a combination of machine learning algorithms and expert judgement to leverage a significant amount of online job posts, obtained through web scraping, to generate an intelligible classification of job roles and skill sets.	Natural Language Processing of Job titles and descriptions using semantic analysis techniques, particularly, LDA Clustering. Web scraping for Data Collection.	Latent Dirichlet Allocation. The exact technologies are not mentioned.	Classification of demand for job roles such as Data Scientist, Business Analyst, Big Data Developer & Big Data Engineer
Labor Market Trends and Outcomes: What Has Changed since the Great Recession?[5]	Erica L. Groshen, Harry J. Holzer	Describes trends in labor force participation for the "working class", compare cyclical peaks from 1979 - 2019 of job salaries & wages while focusing on trends before, during and after The Great Recession	Theoretical analysis of job trends due to a recession and the variance of wages of workers due to the aforementioned reasons.	N/A	Median Wages of workers through and after the recession periods distributed by race, gender, education among other factors.
Big Data Application in Job Trend Analysis[6]	Priyanka Kale, Shilpa Balan	Big data techniques such as Hadoop and Tableau are applied to identify job trend analysis in New York.	The data set studied comprises of several gigabytes. Handling such huge data can be difficult with conventional resources. Hadoop proves its value with increased scale of data. Hence, Hive and Hadoop are applied to the study.	Hadoop, Hive, CSV Cleaning, Tableau	Data Visualization of Job Market trends using Big Data Technologies
Data mining approach to monitoring the requirements of the job market: A case study [7]	Ioannis Karakatsanis et. al.	To analyze the job market, researchers are increasingly turning to data science and related techniques which are able to extract underlying patterns from large collections of data	a Latent Semantic Indexing (LSI) model was developed that is capable of matching job advertisement extracted from the Web with occupation description data in the O* NET database	The BeautifulSoup library NLTK (Natural Language Toolkit) Gensim	The top 10 prevailing O* NET occupations identified. The generation of intuitive visualizations using the generated weightings

Table 2 continued from previous page

Paper Title	Author(s)	Problem Definition	Proposed Approach Methodology	Technologies Used	Factors Considered Important/Evaluation Metrics
Discovering Job Market Trends with Text Analytics[9]	Raymond Blanch Mbah et. al.	The problem being solved is collection, analysis and visualization of local job data using text mining techniques.	Collect and store company data from Glassdoor, use company names to query, collect and store company jobs from Indeed and build an inverted index to visualize for analysis.	ElasticSearch for building the inverted index, R and Shiny Server for visualization	Inverted Index Query Accuracy without a consideration for inference time. Clustering jobs by various categories.
Reference Architecture and Classification of Technologies, Products and Services for Big Data Systems[11]	Pekka Pääkkönen, Daniel Pakkala	A technology independent reference architecture for big data systems and classification of related implementation technologies and products/services,	Functionality, dataflows and data stores of implementation architectures in seven big data use cases were analysed. Subsequently, a reference architecture was constructed based on the analysis.	Hive CLI, MySQL, Microstrategy UI, Hive-Hadoop Clusters	Inductive reasoning for construction of the Big Data Architecture, Facilitation of architecture design for Big Data Technologies
Analyzing Computer Programming Job Trend Using Web Data Mining[12]	David Smith, Azad Ali	Data was collected using data mining techniques over a number of years from an online job agency. The data was then analyzed to reach a conclusion about the trends in the job market.	Uses a routine and a data mining technique where a batch file is regularly used to go through various steps to extract data from the job search web site.	Web Data Mining, Keyword Indexing	Trends of Computer Programming Job Roles. Use of data mining techniques
Big Data and Labor Markets: A Review of Research Topics[13]	Lejla Turuljaa et. al.	The assessment of the current research in big data for the labor market has been compiled to detect the research gaps and generate future research directions.	A systematic review of the literature was adopted to address the study questions posed for identifying, assessing, and interpreting all available research material to address specific research issues.	VOSviewer and R bibliometrix,	The implications of technological advancement for industry and labor market, education for the new labor market & skills for new age labor markets, tools and methods used for labor market analysis.
Changes in Occupational Skills - A Case Study Using Non-Negative Matrix Factorization[20]	Wei Lee Woon et. al.	To detect and study the underlying skill "dimensions", i.e. groups of skill or ability elements which co-occur repeatedly across multiple occupations.	Performing NMF on the O*NET database, which is a publicly available database of occupations developed for the US Department of Labor	Python and R Programming environments. Using sci-kit learn, NMF is implemented.	The factors that affect the skill content of jobs are being extracted using NMF and Factor Analysis for comparison using a skill requirement database provided by the US Government

Table 2 continued from previous page

Paper Title	Author(s)	Problem Definition	Proposed Approach Methodology	Technologies Used	Factors Considered Important/Evaluation Metrics
A Study of the Demand for Information Technology Professionals in Selected Internet Job Portals[8]	Kai S. Koong et. al.	This study identifies and classifies information technology related job listings that are disclosed in the databases of two leading e-recruiting services. Two secondary variables, written and oral communications and experience, were also collected and examined in this study. al.	In this study, two main hypotheses were investigated. The distribution of the various IT positions that were posted was the subject of the first hypothesis. The secondary variables that were gathered were the focus of the second hypothesis. Because the statistical tests incorporated observed and anticipated proportions and ranges, chi-square tests were employed to examine the two hypotheses (Koosis, 1997).	N/A (Chi-Squared Test & Statistical Data Mining were used but no specific technologies were mentioned)	Distribution of job opportunities based on skill sets
Classifying Online Job Advertisements through Machine Learning[3]	Roberto Boselli, Mirko Cesarini, Fabio Mercorio	The problem defined in this work utilizes titles and job descriptions of job postings and uses Machine Learning techniques in order to build a knowledge base for the Web Labor Market.	The proposed approach uses a machine learning model for classifying multilingual Web job vacancies exploiting a single-label classifier using both titles and descriptions. It uses box plots to evaluate the f1-score measure of each classification algorithm over the first-digit of the ISCO taxonomy, which identifies 9 distinct occupation groups. Finally, it builds a Knowledge Graph with occupations and skills as nodes and the linkages between them as the edges.	N/A (Neural Graph Learning, Statistical Data Mining, Natural Language Processing)	Skill sets mentioned in job postings
Better understanding of the labour market using Big Data[19]	Alena Vankevich, Iryna Kalinouskaya	The article underscores the need to utilize Big Data for a deeper understanding of the labor market in the digital age. Online job postings offer valuable information on required skills, calling for the development of tools to analyze this data and facilitate better job matching.	The proposed approach uses Big Data for labor market analysis, focusing on skills in job postings, skill mismatches, and improving decision-making. Data is collected from web sources to identify skill gaps and enhance the labor market experience for job seekers and employers.	Scrapy Airflow task schedule Artificial intelligence Universal Sentence Encoder	Skills and Competencies acquired knowledge correlation between the demand for skills and employability

Table 2 continued from previous page

Paper Title	Author(s)	Problem Definition	Proposed Approach Methodology	Technologies Used	Factors Considered Important/Evaluation Metrics
Big Data For Labour Market Intelligence[10]	Mario Mezzanzanica and Fabio Mercorio	Create a data-driven system of job market intelligence to forecast employment trends, pinpoint skills that are in demand, and evaluate the effects of digitalization. Discuss important issues pertaining to the expansion of professions and the role that soft skills play in current employment.	The proposed approach selects, preprocesses and transforms data sources using statistical, technical and domain expertise criteria. It then uses data mining techniques such as Knowledge Discovery in Databases(KDD) to infer analyses of the data sources. It uses a data lake framework to query data as a data scalability solution.	Java, Selenium, Scrapy, MongoDB, HBase, Cassandra, HDFS, Map Reduce, Yarn	Labor Market Intelligence analysis using statistical, technical and domain expertise paradigms on structured and unstructured raw data.