Exploring the Global Job Market : A Big Data Analysis of International Job Postings

Fall 2023 - CS226
Big-Data Management
Group 18: (Mis)Fortune-500
Final Project Presentation

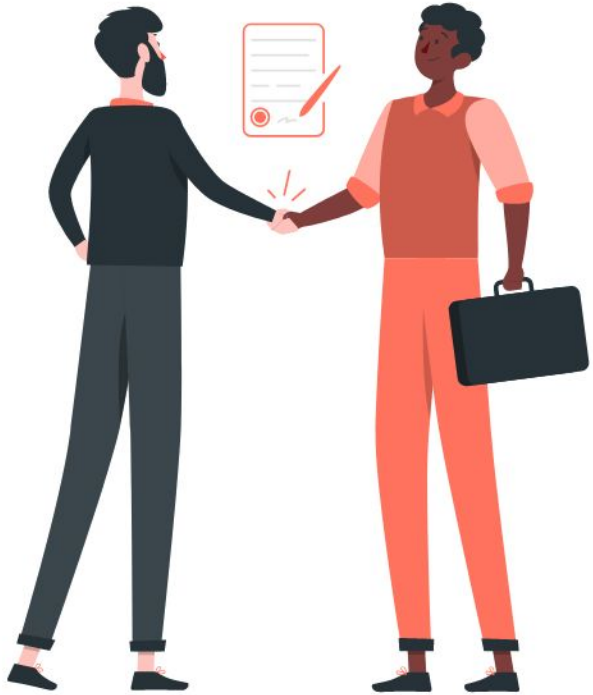
XINLE CHEN

SUMEDHA GIRISH ATREYSA

NUNNA LAKSHMI SARANYA

TEJAS MILIND DESHPANDE

MANISH DEEPAK CHUGANI



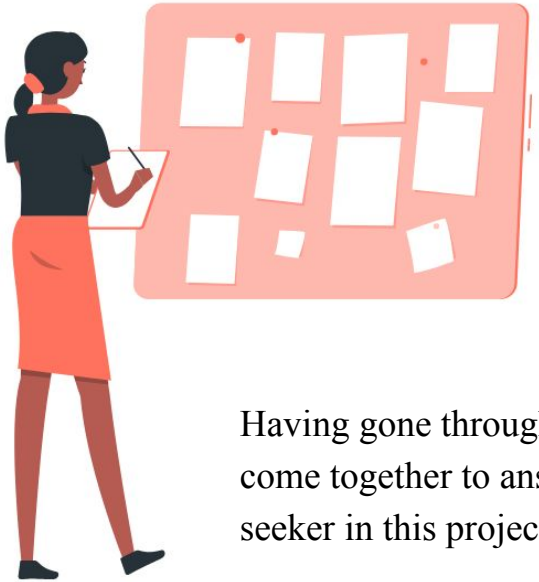
BACKGROUND “VERIFICATION”

- The advancement of LLMs and AI Technologies has taken the world of Computer Science by storm.
- Job security and procurement have developed an air of uncertainty and a fear of the unknown.
- Can advent of AI replace engineers ?
- Will the continuing recession cause a further decline in the number of jobs in Software & Data domains ?
- Data can help us get insights into the trends of the job market.

Image Link (Citation)

BACKGROUND “VERIFICATION”

- The global labor market is increasingly interconnected, offering professionals and businesses access to a diverse talent pool.
- International job listings on various job boards reflect changing demand, skills, and sector requirements, highlighting the complexities of this evolving labor market.
- It's crucial for job seekers and companies to comprehend these dynamics.
- The labor market is dynamic due to its flexibility and adaptability, making it essential for policymakers, educators, students, and anyone navigating it to understand these trends for significant benefits.



MOTIVATION FOR THE PROJECT

Having gone through the job seeker's yearly cycle, the developers of MisFortune500 decided to come together to answer their burning questions. The most pertinent question that every job seeker in this project group had was as follows:

"Are we just not good enough or is the job market really facing a recession?"

The main idea that answers this question is the comparison between the number of jobs being posted during this period versus when the job market was booming.

Thus, MisFortune500 came to life using the power of Big Data technologies such as PySpark.

“If a job opportunity doesn't knock, build a data analysis door!”

—(Mis)Fortune-500

THE PROBLEM TO BE SOLVED

- It is difficult for a job seeker to obtain a direct insight into the job market and where they stand. Understanding that the job market is trending downwards allows a job seeker some comfort when they are receiving numerous rejects.
- This project allows a job seeker quick access to the number of positions available at the current time in their area of specialization (for e.g: Software Development) as well as a comparison of the number of open positions that have been posted in an upward trending job market.
- Finally, this project allows a job seeker to obtain insights about which skill sets are in demand in the current market based on the number of open positions.
- All of this, with the power of big data technology, is available to the job seeker within minutes at the tip of their fingers.

OUR DATASETS

We will utilize three primary datasets **obtained from Kaggle** for this project:

International Job Postings - September 2021

A dataset which offers a snapshot of international job postings in September 2021.

LinkedIn job postings

a comprehensive source of contemporary job postings

GLOBAL TECH

International Jobs

The three datasets provide a rich collection of international job listings, including job descriptions, locations, required skills, qualifications, and more. The other datasets being looked at for future use are listed below

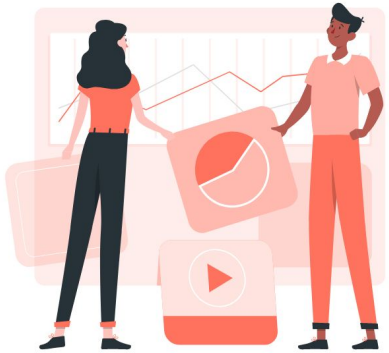
Project Overview

- Gathered large number of job advertisement from multiple sources for the years 2021 and 2023.
- Utilized BigData technologies to format and integrate these job postings to build a single scalable data store.
- Fetched counts of number of jobs posted using an efficient query processing engine powered by PySpark.
- Derived analytics and visualizations from this consolidated data to aid users in understanding job market trends.
- Compared performance and scalability of these data processing technologies with traditional architectures.

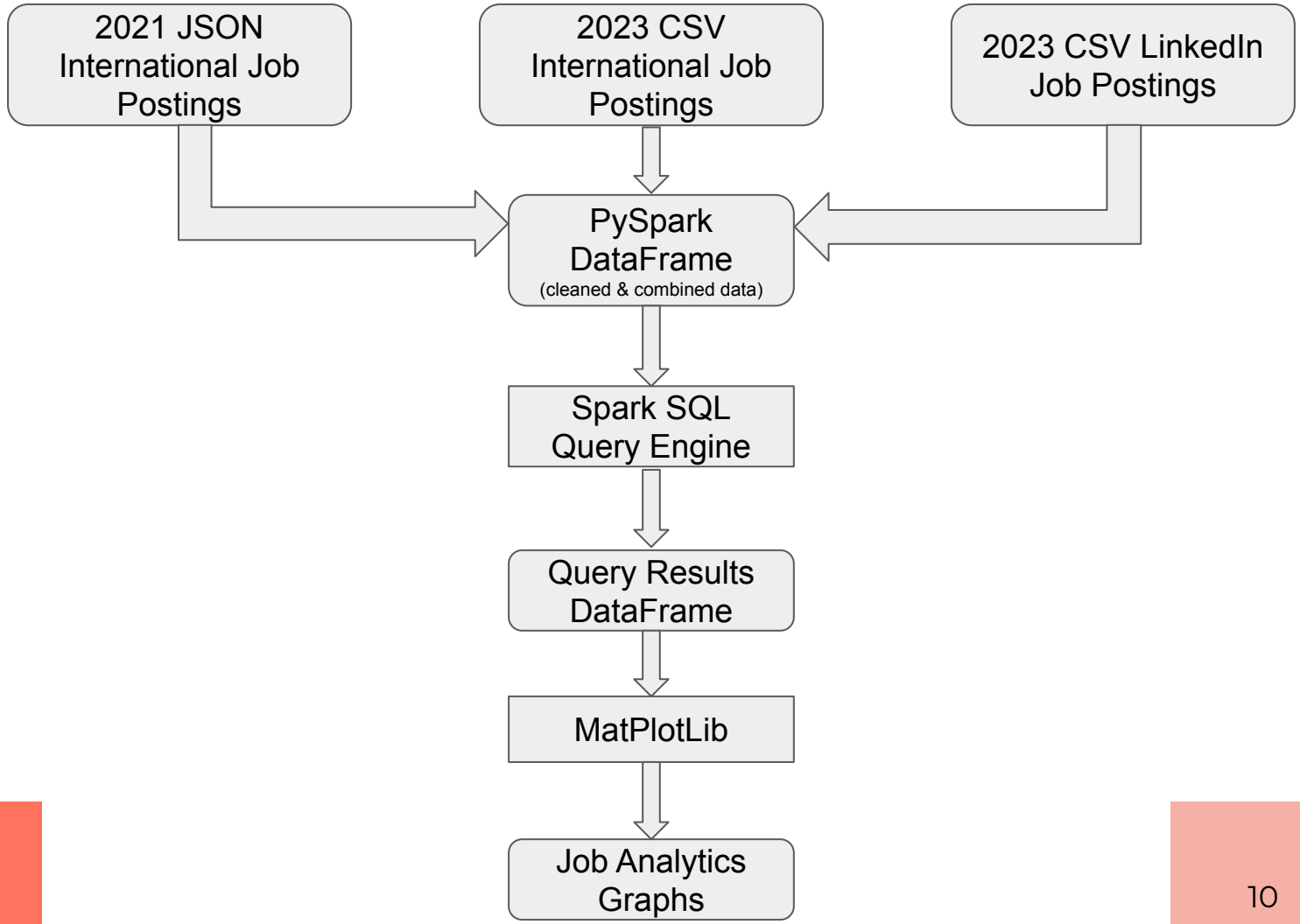
Image by pikisuperstar on Freepik

BIG DATA
RELEVANCE

- **Preprocessing and Data Cleaning:** Noise and irregularities are common with big data. Strict data purification and preprocessing methods are essential to guaranteeing data quality. The proposed approach removes null values and drops repeated, nested fields using PySpark to preprocess the data.
- **Data integration:** To make sure the data is consistent and interoperable, the proposed approach combines and harmonizes data from multiple sources. MisFortune-500 combines datasets from 2021 and 2023 after the preprocessing step into a single combined PySpark Dataframe as it's Data Integration component.
- **Data engineering:** MisFortune-500 then uses PySpark's SparkSQL framework to obtain data analytics and visualizations (matplotlib) from the data store created by the previous two steps.
- These three categories highlight the relevance of the proposed approach to Big Data Management.



Project Pipeline



MAIN WORK

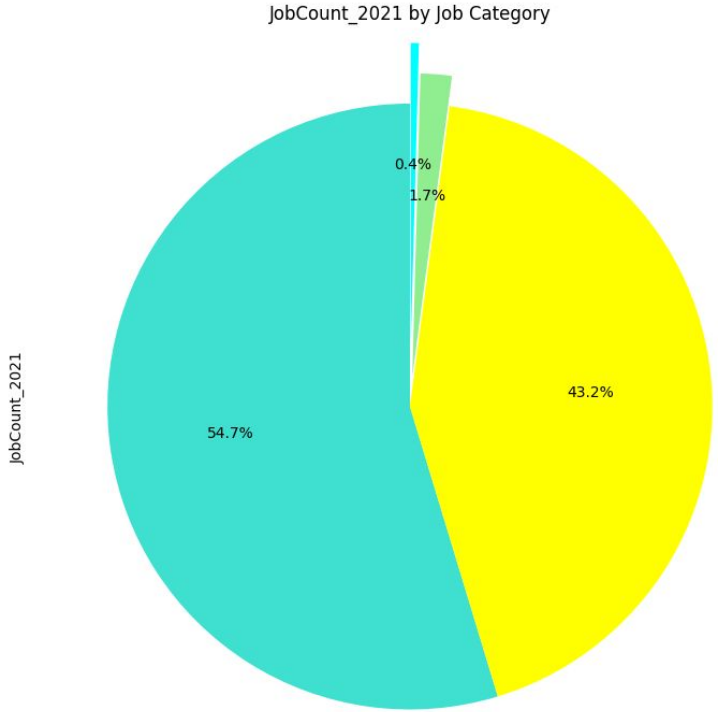
- The proposed approach identifies relevant fields which are common across multiple datasets and defines a schema containing them. This schema is used to ingest JSON data records into a PySpark DataFrame. Using a predefined schema automatically enabled discarding irrelevant fields from the JSON dataset.
- However, due to malformed records in CSV datasets, using a predefined schema led to incorrect records in the corresponding constructed DataFrame. Hence, for CSV datasets, columns were dropped after inserting complete records into the DataFrames.
- While building the DataFrames, transformations like column / field renaming, flattening nested objects, column reordering and so on were applied to the data records for achieving a consistent data structure and format across different datasets.
- These PySpark DataFrames are combined together into a single dataframe by appending records from each of them using transformations like union to achieve a single data store containing consistent and coherent data.
- Job postings falling under specific criteria are fetched from this singular dataframe using the efficient SparkSQL query processing engine which runs these queries optimally.
- These records are aggregated and passed to the MatPlotLib visualization library for constructing analytics.
- Theses analytics based on consolidated data can aid the user in understanding the job market trends as well as compare the current market situation with the past.

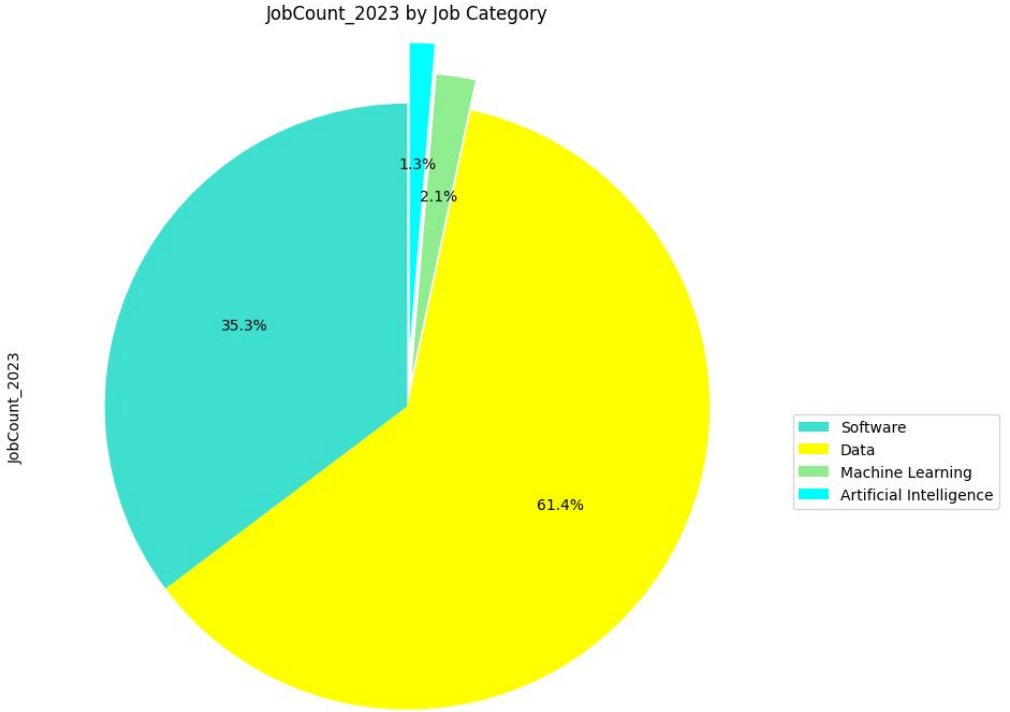
EVALUATIONS

- A pivotal metric for our project is the difference in data processing latency. We compared the latency observed using traditional data processing techniques such as Relational Database Management Systems versus that witnessed when deploying PySparkSQL.
- Another important metric is the comparison of scalability and performance of different data processing technologies such as RDBMS, PySpark and Pandas.
- This will demonstrate how the design principles and algorithms of Big Data Management Systems tackle the multiple Vs of Big-Data.

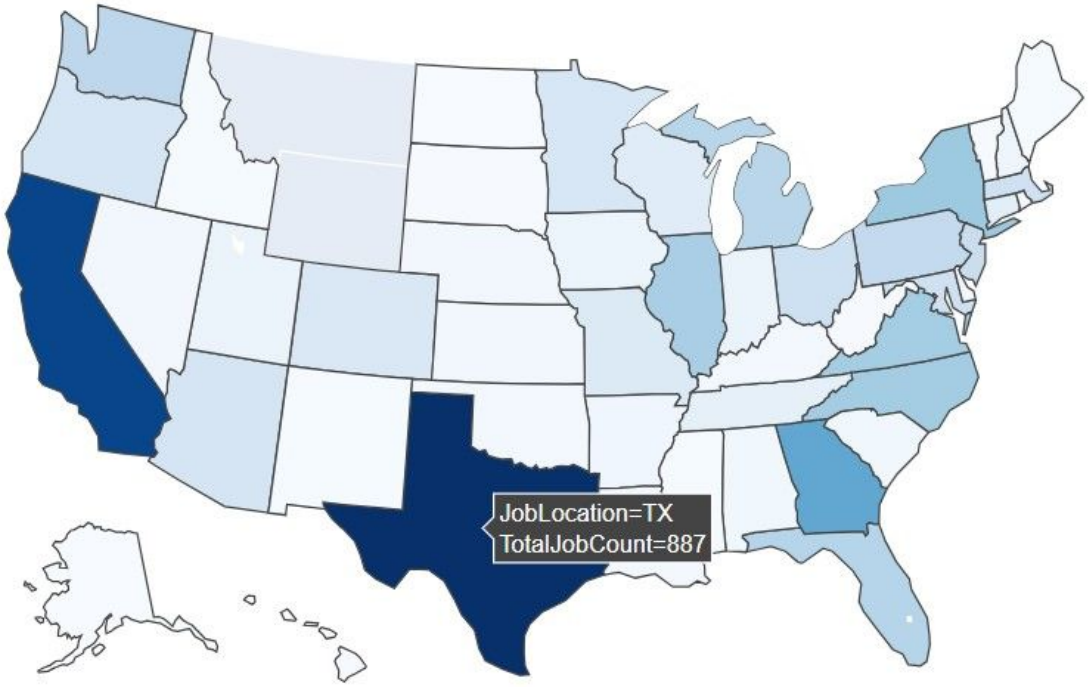
Dataset	Dataset Size	Query Time for SparkSQL	Query Time for RDBMS
2021 International Job Postings\cite{Data_2021}	46.9 GB	10.54 minutes (12 GB RAM TPU)	59.35 minutes (12 GB RAM TPU)
2023 LinkedIn Job Postings\cite{LinkedInData}	69.8 MB	6.84 seconds (12 GB RAM TPU)	1.30 minutes (12 GB RAM TPU)
2023 International Job Postings\cite{Data_2023}	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)

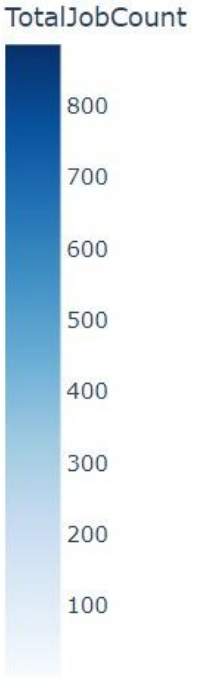
Table 1. Comparison of PySparkSQL versus MySQL


JobCount_2021 by Job Category
JobCount_2021
0.4%
1.7%
43.2%
54.7%


JobCount_2023 by Job Category
JobCount_2023
1.3%
2.1%
35.3%
61.4%
Software
Data
Machine Learning
Artificial Intelligence

Distribution of Jobs across the states of the USA in 2021





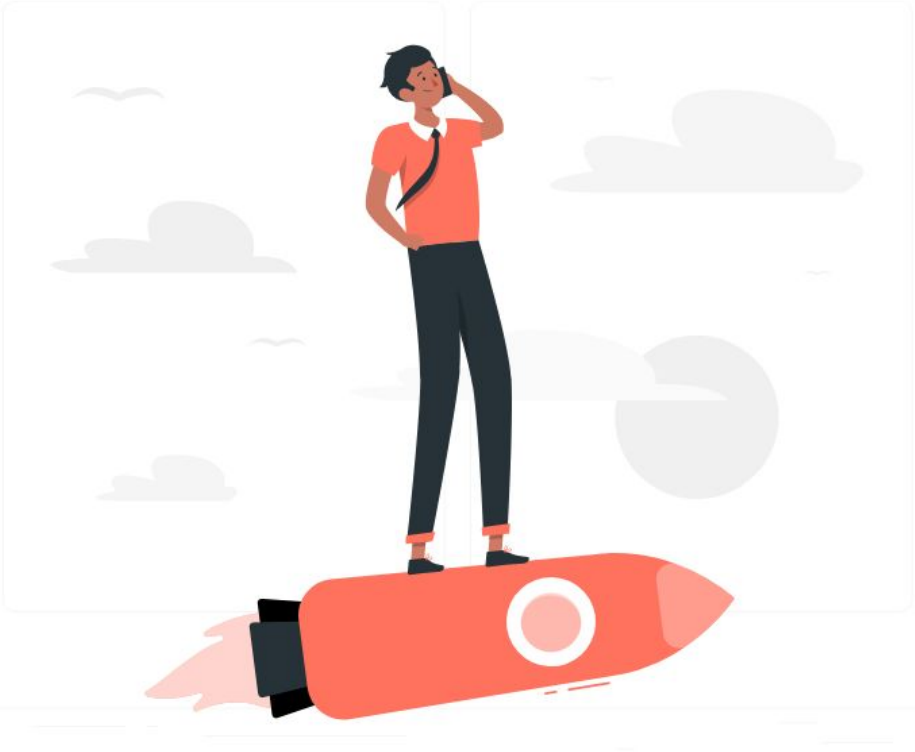
RELATED WORK

- Most of the related work in this area involves drawing valuable insights from Job Market trends by predicting the future based on the past.
- This makes it essential for the related works to use statistical techniques such as Machine Learning and even heavier processing techniques such as Neural Networks.
- However, the present status of the job market in itself tells a story when combined with the past trends.
- Our work outlines and evaluates a simple Big Data Framework without involving predictions about the future of the job market and its trends.
- The reconciliation for a job seeker that the current market is hit by recession is enough for them to continue to work towards their growth without fearing the status of an impostor (otherwise known as impostor phenomenon).
- Our work uniquely provides this insight on the Job Market.

Conclusion

- Successfully designed a fast query engine using Big Data Technologies (PySpark, SparkSQL) that allows job seekers and employers alike to get valuable insights from the available data for the current job market.
- The MisFortune 500 architecture draws a comparison between the job market in previous years with insightful visualization techniques .
- The project can be commercialized and provided to the industry if given more data and required computing power resources.

THANK YOU!!!!!



CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, infographics & images by **Freepik** and illustrations by **Stories**

.