

The Job Intelligence Engine

Team Unattached and Unhinged

1. Varad Paradkar
2. Mitali Yadav
3. Sudeepta Bal
4. Ruthwik Dovala



The Problem

- > **The Data Deluge:** The job market is oversaturated with millions of postings, making it nearly impossible for students to identify the relevant signal (high-potential opportunities) from the noise.
- > **The Skills Gap:** Students lack actionable guidance on which skills to prioritize (e.g., Python vs. R, SQL vs. NoSQL) to gain a competitive edge in specific industries.
- > **Static Search:** Traditional job boards offer only basic keyword searches, failing to provide dynamic, data-driven insights into:
 - Industry Entry Barriers (How many skills are truly required?)
 - Optimal Learning Roadmaps (Which skills must be learned together?)
 - Real-Time Market Shifts (Which companies are hiring right now?)

The Solution

Stream

* **Ingest High-Value Corpus:** We batch process the 40K+ LinkedIn Job Dataset (Kaggle) combined with related company and skill data, providing a rich, multi-dimensional view of the market.

* **Scalable Architecture:** Apache Spark is the backbone, designed to efficiently ingest and join large datasets from multiple tables (postings, companies, skills) using complex multi-table joins.

Analyze

* **Complex Queries for Deep Insight:** We employ advanced Spark SQL Window Functions and aggregations to calculate metrics like Top Skills by Industry and Cross-Industry Skill Overlap.

* **Market Monitoring:** Spark Structured Streaming provides useful analytics on simulated data streams to instantly track key market shifts like Top Hiring Locations and active companies.

Recommend

* **Interesting Visuals:** Our product analyzes job postings to uncover key skill and industry trends and generates clear visual insights and summary datasets for students and job seekers.

* **Data-Driven Career Guidance:** Our product analyzes job descriptions and each user's past liked jobs to automatically generate personalized, content-based job recommendations.

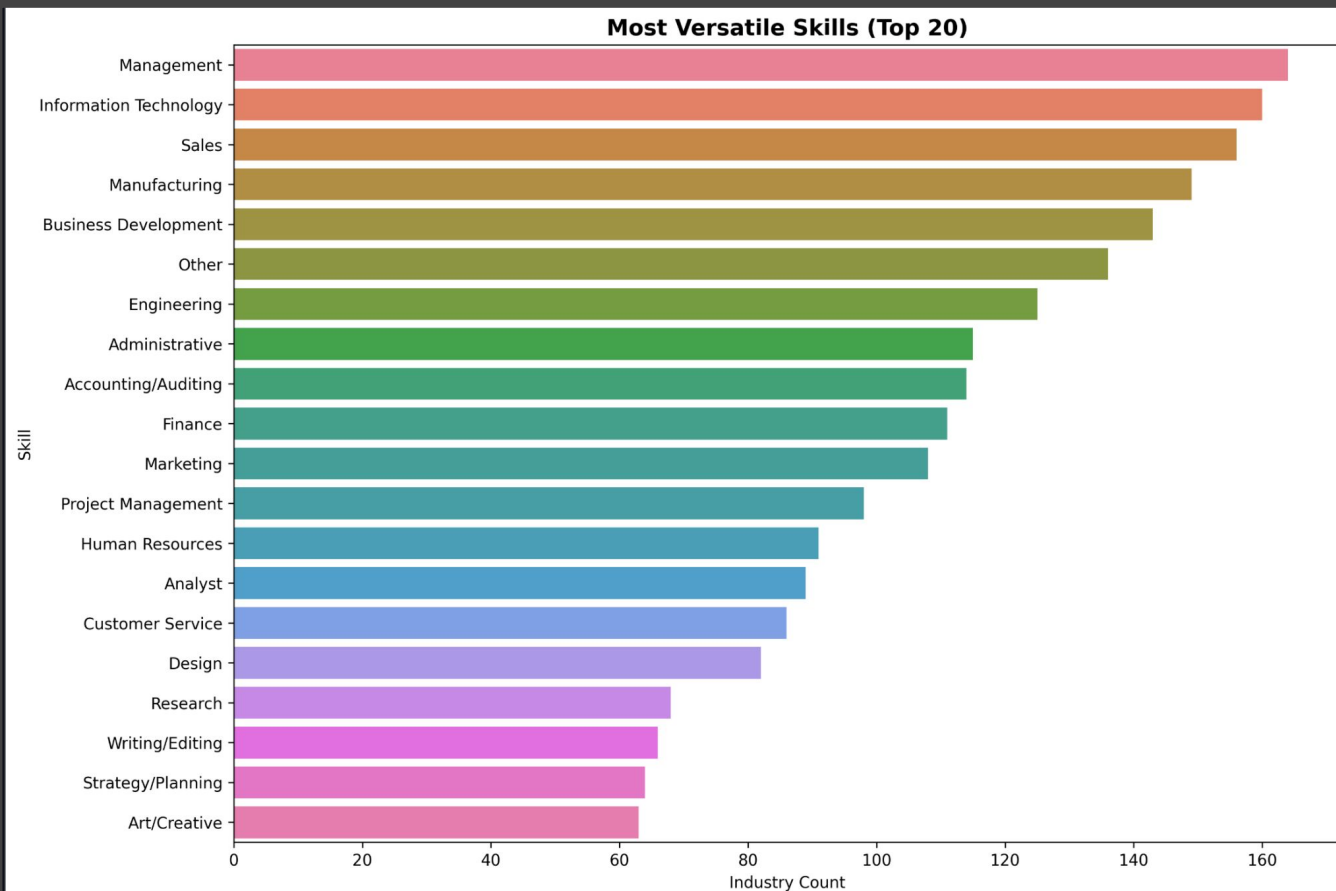
Value Proposition

1. Beyond Keywords: We don't just match keywords; we analyze the context and complexity of job roles across over more than 50 industries using multi-table joins and aggregations. This provides insights into Skill Diversity and Entry Barriers, allowing students to understand the true difficulty and required breadth of a role.

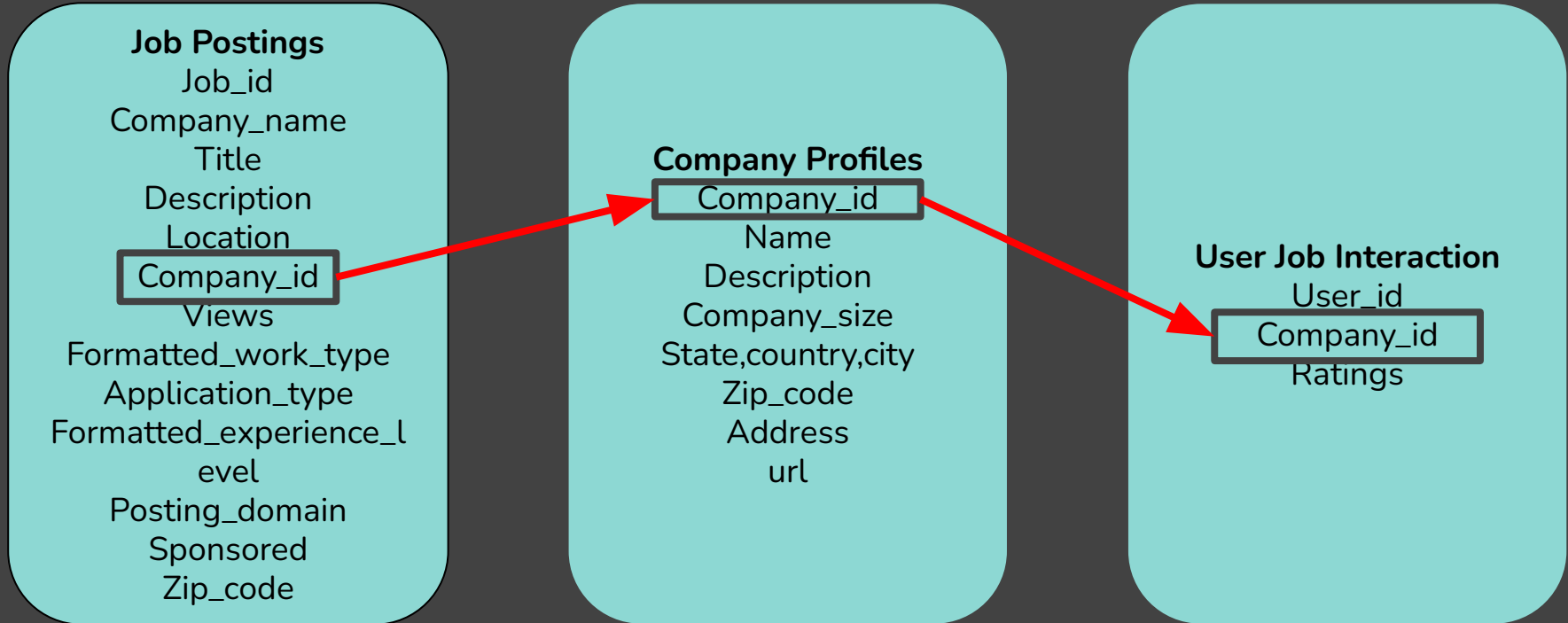
2. Personalized Ranking: Our product loads all job postings and merges key information, such as the title, description, and experience level, into a unified profile for each job. By analyzing the text of each job, the system identifies which postings are similar, essentially turning job descriptions into “meaningful fingerprints.” Using each user’s past ratings, it determines which jobs they liked most and finds new postings that closely match those preferences.

3. Actionable Intelligence: We deliver around 3 - 4 comprehensive visualizations showing clear market trends, including top-demand skills, salary potential (inferred from role complexity), and industry opportunities. Spark Structured Streaming provides a reality check on active hiring locations and companies, giving users an immediate competitive edge.

Top 20 skills extracted



The Data



The Intelligence Engine: How It Works?

Deep Insight Generation

Spark SQL to run complex, multi-stage queries
rank function to determine Top Skills by Industry by ranking skill frequency within each industrial sector
Use multi-table joins for linking job postings, company attributes, and skill dictionaries to calculate Industry Entry Barriers (average number of required skills per industry).

Near Real-Time Market Monitoring

Demonstrate the system's ability to handle live data using Spark Structured Streaming for continuous analysis
Use structured streaming to process data in micro-batches, ensuring low-latency updates

Student-Focused Output

The results of the analysis are materialized into actionable intelligence
Create a Quadrant Analysis that helps users assess the risk of skill investment (e.g., highly demanded but generalist vs. specialized).

Outputs and Results

I. Predictive Modeling Performance

Job Classification Accuracy: 97.14% Model Validation

Random Forest outperformed XGBoost by >0.2%, demonstrating successful model tuning.

II. Personalized Interview Scoring

Data Sources: Utilizes User Job Interaction history + Job Postings features.

Likelihood Score: Quantifies the possibility of securing an interview based on historical performance.

Actionable Threshold: Jobs scoring 3.0 or above are flagged as High-Potential Matches.

Github link to our project and related source codes

<https://github.com/vradcar/ITCS6190-Course-Project/tree/main>

Thank you!!!