Big Data - Case Study

Subject - Big Data Analytics and Architecture

PROJECT

AI Job Market Analysis

AI Job Market Analysis Using Apache Hive

Project Overview

This project focuses on performing data analysis and insight extraction from an AI Job Market dataset using Apache Hive.
The primary goal is to utilize Hive's SQL-like querying capabilities to analyze key job market trends such as industry demand, skill requirements, company hiring patterns, and salary distribution.
The project demonstrates how to manage structured job-related data on a Big Data platform (Cloudera/Hadoop) and use HiveQL for large-scale analytical querying and decision-making.

Dataset Description

The dataset, job1.csv, contains detailed information about various AI-related job postings, including:

- Job ID
- Company Name
- Industry
- Job Title
- Skills Required
- Experience Level
- Employment Type
- Location
- Salary Range (USD)
- Posted Date
- Company Size
- Tools Preferred

Objectives

The key objectives of this project are:

- To import and store CSV job market data into Hive tables efficiently.
- To perform analytical queries on job and company trends.
- To extract meaningful business insights such as:
    - Most active companies hiring for AI roles.
    - Average salary range by experience level.
    - Most in-demand skills and tools.
    - Industry-wise job distribution.
    - Experience-level demand in the job market.

Technologies Used

- Apache Hive
- Hadoop (Cloudera Environment)
- HiveQL (SQL-like Queries)
- CSV File Data Ingestion
- HDFS Storage

Steps Performed

1. Created a database and Hive table schema for the job dataset.
2. Loaded CSV data from local storage or HDFS into the Hive table.

3. Executed multiple Hive queries to summarize and visualize insights:
    - SELECT COUNT(*) → total number of job listings.
    - GROUP BY → industry and company analysis.
    - AVG() and MAX() → salary range insights by experience level.
    - ORDER BY and LIMIT → top hiring companies and skill trends.
4. Generated analytical reports summarizing job market trends and data-driven insights.

Key Insights

- Identified top industries contributing the most AI job postings.
- Found top companies hiring for AI roles globally.
- Discovered average salary variations across different experience levels.
- Highlighted most in-demand tools and skills for AI professionals.
- Observed growth in AI job postings over recent years.

Conclusion

This project showcases how Apache Hive can be leveraged for large-scale data analysis in the AI and technology job market.
By integrating structured queries with Big Data tools, analysts can derive valuable insights that support recruitment strategies, workforce planning, and industry research.

Use Database:

```
hive> CREATE TABLE job1 (
    >     job_id INT,
    >     company_name STRING,
    >     industry STRING,
    >     job_title STRING,
    >     skills_required STRING,
    >     experience_level STRING,
    >     employment_type STRING,
    >     location STRING,
    >     salary_range_usd int,
    >     posted_date int,
    >     company_size STRING,
    >     tools_preferred STRING
    > )
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ','
    > STORED AS TEXTFILE;
OK
Time taken: 0.776 seconds
```

Load Data:

```
hive> load data local inpath '/home/cloudera/Desktop/aijobmarket.csv' into table job1;
Loading data to table harsh.job1
Table harsh.job1 stats: [numFiles=1, totalSize=344125]
OK
Time taken: 4.526 seconds
```

Q.1 Total Number of Job Listings

SELECT COUNT(*) AS total_jobs FROM job1;

📊 *Insight:* Shows total number of job openings in the dataset.

```
hive> SELECT COUNT(*) AS total_jobs FROM job1;
Query ID = cloudera_20251028070303_8488ea39-15e0-499d-9a0e-8131b53e3c4b
```

Out Put:

```
Total MapReduce CPU Time Spent: 7 seconds 870 msec
OK
2001
Time taken: 64.197 seconds, Fetched: 1 row(s)
hive>
```

Q.2 Most Common Job Titles

SELECT job_title, COUNT(*) AS count

FROM job1

GROUP BY job_title

ORDER BY count DESC

LIMIT 10;

📊 *Insight:* Identifies the top 10 job roles in demand.

```
hive> SELECT job_title, COUNT(*) AS count
    > FROM job1
    > GROUP BY job_title
    > ORDER BY count DESC
    > LIMIT 10;
Query ID = cloudera_20251028072020_c8ff11e9-12d3-4b34-894f-5ac1d8b31ac9
```

Out put:

```
Total MapReduce CPU Time Spent: 5 seconds 790 msec
OK
NLP Engineer    178
Data Analyst    178
Quant Researcher        175
AI Product Manager      174
AI Researcher   161
ML Engineer     155
Data Scientist  153
Computer Vision Engineer        147
Finance 105
E-commerce      105
Time taken: 55.37 seconds, Fetched: 10 row(s)
hive>
```

Q.3 Jobs by Experience Level

SELECT experience_level, COUNT(*) AS total

FROM job1

GROUP BY experience_level

ORDER BY total DESC;

📊 *Insight:* Helps see which level (Entry, Mid, Senior) has more job opportunities.

```
hive> SELECT job_title, COUNT(*) AS count
    > FROM job1
    > GROUP BY job_title
    > ORDER BY count DESC
    > LIMIT 10;
Query ID = cloudera_20251028072020_c8ff11e9-12d3-4b34-894f-5ac1d8b31ac9
```

Out put:

```
OK
 MLflow 78
 Pandas 77
 GCP    74
 Excel  72
 Power BI       65
 TensorFlow     65
 Hugging Face   64
 SQL    64
 C++    61
 NumPy  60
 PyTorch        60
 Python 59
 Keras  57
 Flask  56
 Azure  55
 CUDA   54
 Reinforcement Learning 53
 LangChain      53
 AWS    53
 Scikit-learn   50
 R      48
 FastAPI        43
"Reinforcement Learning 38
"FastAPI        36
"Azure  36
"Power BI       36
"Keras  35
"Excel  35
"NumPy  33
"Python 33
"LangChain      33
"MLflow 32
"R      31
"CUDA   31
"GCP    30
"PyTorch        29
"Flask  29
"SQL    29
"AWS    28
"Hugging Face   27
"Scikit-learn   27
"TensorFlow     25
"C++    24
"Pandas 22
experience_level        1
Time taken: 59.167 seconds, Fetched: 45 row(s)
hive>
```

Q.4 Top 10 Companies Offering the Most Jobs

SELECT company_name, COUNT(*) AS job_count

FROM job1

GROUP BY company_name

ORDER BY job_count DESC

LIMIT 10;

📊 *Insight:* Shows the companies hiring most actively.

```
hive> SELECT company_name, COUNT(*) AS job_count
    > FROM job1
    > GROUP BY company_name
    > ORDER BY job_count DESC
    > LIMIT 10;
Query ID = cloudera_20251028073434_34334f10-88ec-4753-8217-29a3f420ad2e
```

Out put:

```
Total MapReduce CPU Time Spent: 6 seconds 300 msec
OK
"Johnson        16
"Smith  15
"Miller 13
"Williams       13
"Garcia 9
"Brown  8
"Thompson       7
"Gonzalez       6
"Anderson       6
"Walker 6
Time taken: 58.015 seconds, Fetched: 10 row(s)
hive>
```

Q.5 Job Distribution by Employment Type

SELECT employment_type, COUNT(*) AS total

FROM job1

GROUP BY employment_type;

📊 *Insight:* Breaks down jobs by type (Full-time, Contract, Internship, etc.).

```
hive> SELECT employment_type, COUNT(*) AS total
    > FROM job1
    > GROUP BY employment_type;
Query ID = cloudera_20251028073838_92a81f63-5f86-4c82-8f90-38bee7ff0182
```

Out put:

```
Total MapReduce CPU Time Spent: 3 seconds 50 msec
OK
 AWS    71
 AWS"   12
 Azure  83
 Azure" 8
 C++    68
 C++"   18
 CUDA   83
 CUDA"  17
 Excel  87
 Excel" 18
 FastAPI        70
 FastAPI"       12
 Flask  64
 Flask" 20
 GCP    71
 GCP"   9
 Hugging Face   75
 Hugging Face"  8
 Keras  77
 Keras" 16
 LangChain      76
 LangChain"     12
 MLflow 61
 MLflow"        18
 NumPy  84
 NumPy" 17
 Pandas 88
 Pandas"        14
 Power BI       73
 Power BI"      13
 PyTorch        82
 PyTorch"       17
 Python 70
 Python"        16
 R      77
 R"     20
 Reinforcement Learning 76
 Reinforcement Learning"        16
 SQL    72
 SQL"   16
 Scikit-learn   78
 Scikit-learn"  19
 TensorFlow     80
 TensorFlow"    18
employment_type 1
Time taken: 30.706 seconds, Fetched: 45 row(s)
hive>
```

Q.6 Average Salary Range by Experience Level

```
SELECT experience_level,
    AVG(CAST(SPLIT(salary_range_usd, '-')[0] AS INT)) AS avg_min_salary,
    AVG(CAST(SPLIT(salary_range_usd, '-')[1] AS INT)) AS avg_max_salary
FROM job1
GROUP BY experience_level;
```

📊 *Insight:* Estimates salary differences between junior and senior positions.

```
hive> SELECT experience_level,
    >        AVG(CAST(SPLIT(salary_range_usd, '-')[0] AS INT)) AS avg_min_salary,
    >        AVG(CAST(SPLIT(salary_range_usd, '-')[1] AS INT)) AS avg_max_salary
    > FROM job1
    > GROUP BY experience_level;
Query ID = cloudera_20251028074545_b4960422-131b-4e1c-a211-a667c8837e9f
```

Out Put:

```
Total MapReduce CPU Time Spent: 5 seconds 450 msec
OK
 AWS    NULL    NULL
 Azure  NULL    NULL
 C++    NULL    NULL
 CUDA   NULL    NULL
 Excel  NULL    NULL
 FastAPI        NULL    NULL
 Flask  NULL    NULL
 GCP    NULL    NULL
 Hugging Face   NULL    NULL
 Keras  NULL    NULL
 LangChain      NULL    NULL
 MLflow NULL    NULL
 NumPy  NULL    NULL
 Pandas NULL    NULL
 Power BI       NULL    NULL
 PyTorch        NULL    NULL
 Python NULL    NULL
 R      NULL    NULL
 Reinforcement Learning NULL    NULL
 SQL    NULL    NULL
 Scikit-learn   NULL    NULL
 TensorFlow     NULL    NULL
"AWS    NULL    NULL
"Azure  NULL    NULL
"C++    NULL    NULL
"CUDA   NULL    NULL
"Excel  NULL    NULL
"FastAPI        NULL    NULL
"Flask  NULL    NULL
"GCP    NULL    NULL
"Hugging Face   NULL    NULL
"Keras  NULL    NULL
"LangChain      NULL    NULL
"MLflow NULL    NULL
"NumPy  NULL    NULL
"Pandas NULL    NULL
"Power BI       NULL    NULL
"PyTorch        NULL    NULL
"Python NULL    NULL
"R      NULL    NULL
"Reinforcement Learning NULL    NULL
"SQL    NULL    NULL
"Scikit-learn   NULL    NULL
"TensorFlow     NULL    NULL
experience_level        NULL    NULL
Time taken: 322.627 seconds, Fetched: 45 row(s)
hive>
```

Q.7 Most Popular Tools Preferred by Companies

SELECT tools_preferred, COUNT(*) AS count

FROM job1

GROUP BY tools_preferred

ORDER BY count DESC

LIMIT 10;

📊 *Insight:* Finds the most in-demand AI tools (e.g., TensorFlow, PyTorch, etc.).

```
    > SELECT tools_preferred, COUNT(*) AS count
    > FROM job1
    > GROUP BY tools_preferred
    > ORDER BY count DESC
    > LIMIT 10;
Query ID = cloudera_20251028083838_1bdbb4c9-cb96-446d-abfa-0056eafdd365
```

Out Put:

```
Total MapReduce CPU Time Spent: 5 seconds 740 msec
OK
Internship      130
Full-time       127
Contract        118
Remote  116
Entry   62
Senior  58
Mid     48
 IQ"    9
 PL"    8
 GQ"    7
Time taken: 82.832 seconds, Fetched: 10 row(s)
hive>
```

Q.8 Most Required Skills

SELECT skills_required, COUNT(*) AS count

FROM job1

GROUP BY skills_required

ORDER BY count DESC

LIMIT 10;

📊 *Insight:* Shows which skills appear most frequently in job descriptions.

```
hive> SELECT skills_required, COUNT(*) AS count
    > FROM job1
    > GROUP BY skills_required
    > ORDER BY count DESC
    > LIMIT 10;
Query ID = cloudera_20251028084242_58a2cf75-88f3-4065-8742-f8816323c842
```

Out Put:

```
Total MapReduce CPU Time Spent: 4 seconds 820 msec
OK
ML Engineer     95
Data Analyst    93
NLP Engineer    87
Data Scientist  85
AI Product Manager      84
Computer Vision Engineer        83
Quant Researcher        76
AI Researcher   76
"FastAPI        75
"NumPy  71
Time taken: 52.568 seconds, Fetched: 10 row(s)
hive>
```

Q.9 Number of Job Postings by Year

SELECT SUBSTR(posted_date, -4) AS year, COUNT(*) AS job_count

FROM job1

GROUP BY SUBSTR(posted_date, -4)

ORDER BY year;

📊 *Insight:* Reveals trends in job postings over the years.

```
hive> SELECT SUBSTR(posted_date, -4) AS year, COUNT(*) AS job_count
    > FROM job1
    > GROUP BY SUBSTR(posted_date, -4)
    > ORDER BY year;
Query ID = cloudera_20251028084545_5168288d-de0e-4c6e-be2b-960b2c9d3428
```

Out Put:

```
Total MapReduce CPU Time Spent: 4 seconds 510 msec
OK
NULL    2001
Time taken: 49.33 seconds, Fetched: 1 row(s)
hive>
```

Q.10 Which industry has the highest number of AI-related job postings?

Hive command:

SELECT industry, COUNT(*) AS job_count

FROM job1

GROUP BY industry

ORDER BY job_count DESC

LIMIT 1;

📊 **Insight:**
This tells you which industry (e.g., Tech, Finance, Healthcare) is leading in AI job opportunities.

```
hive> SELECT industry, COUNT(*) AS job_count
    > FROM job1
    > GROUP BY industry
    > ORDER BY job_count DESC
    > LIMIT 1;
Query ID = cloudera_20251028085151_ca4b3de0-bfa3-42d7-9e05-dd38159d1dae
```

Out Put:

```
Total MapReduce CPU Time Spent: 4 seconds 360 msec
OK
Automotive      202
Time taken: 47.91 seconds, Fetched: 1 row(s)
hive>
```