Cloud Project

IBM HR Analytics Employee Attrition & Performance

Group Number: 5
Naresh Gajula

Problem Setting:

Employee turnover is a significant challenge for organizations, leading to increased recruitment and training costs, lost productivity, and decreased team morale. Understanding the factors that lead to attrition can help companies develop effective strategies to retain valuable employees. However, relevant data is often scattered across various HR records, making it challenging to perform comprehensive analysis and derive actionable insights.

Problem Definition:

Our goal is to consolidate and analyze employee attrition data to understand the key drivers of employee turnover. This involves gathering employee demographic data, job roles, satisfaction levels, performance metrics, and more, into a single, organized source. By combining these data points, we aim to help HR professionals and management make informed decisions to improve employee retention and workplace satisfaction.

Objective:

Our main objective is to create a system that collects, organizes, and analyzes employee attrition data. This includes cleaning the data, structuring it for analysis, and generating insights on factors influencing attrition. Through this analysis, we aim to provide a clearer understanding of the dynamics behind employee turnover, allowing HR teams to proactively address attrition risks and implement targeted retention strategies.

End Goals

At the end of the project, we aim to have a reliable and user-friendly system that can be used not only for this dataset but also adapted for future HR data. By providing insights into employee satisfaction, job involvement, and other factors impacting attrition, we hope to support organizations in making data-driven decisions to improve workforce stability and overall job satisfaction.

Facts and Dimensions

The dataset is structured to provide both factual and dimensional data points that support a comprehensive view of employee behavior and satisfaction:

Fact Table:

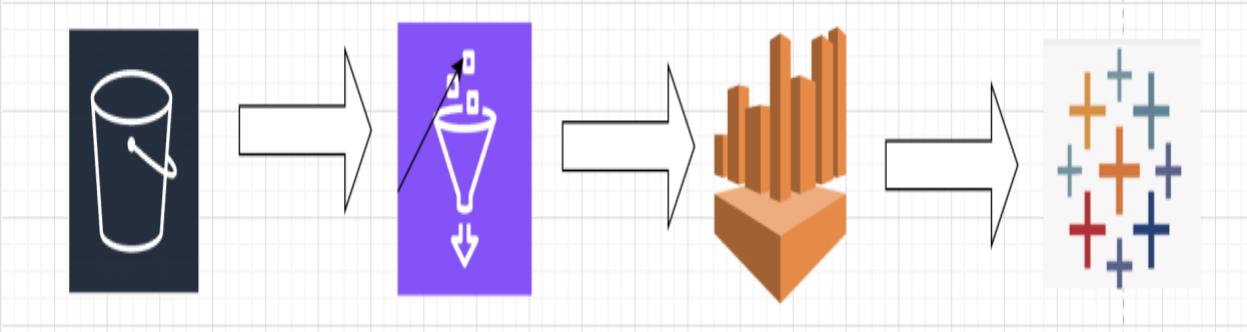
Fact Employee Transactions: Contains key metrics related to employee performance and attrition, including attributes such as Attrition, Monthly Income, Performance Rating, and Years at Company.

Dimensions:

• Employee: Includes specific details about employees, such as Employee Number, Age, Gender, and Marital Status.
• Job: Captures job-related information, such as Job Role, Department, and Job Satisfaction.
• Business: Provides business-specific data, including operational and organizational details.
• Time: Tracks temporal information, such as Date, Month, and Year.

Insights and Strategic Goals

By analyzing this dataset, we aim to:
• Identify the main drivers of employee attrition.
• Understand the impact of factors like job satisfaction, work-life balance, and compensation on turnover.
• Support HR teams in making data-driven decisions to enhance employee retention.



Data Pipeline Description:

The data pipeline represents a detailed transformation and processing workflow using AWS services, designed to ingest, join, query, and store data effectively.

Data Pipeline Steps:

Data Source (Amazon S3):
The pipeline begins with data ingestion from two distinct Amazon S3 buckets. These buckets store raw datasets, serving as the input for further processing.

2. Transform - Join:
The two data sources are combined using a Join operation to create a unified dataset. This step integrates the datasets based on common keys, ensuring a consolidated view of the data.

3. Transform - SQL Query:
After joining, an SQL Query transformation is applied to perform filtering, aggregation, and data manipulation tasks. This prepares the data for downstream analysis.

4. Further SQL Queries:
The transformed dataset undergoes additional SQL Queries, which involve converting OLTP schema to OLAP for further analysis. Each query targets a specific aspect of the data for analysis.

5. Data Target (Amazon S3):
The results of the transformations are saved back to multiple Amazon S3 buckets. Each bucket may store data outputs optimized for specific use cases or further analysis.

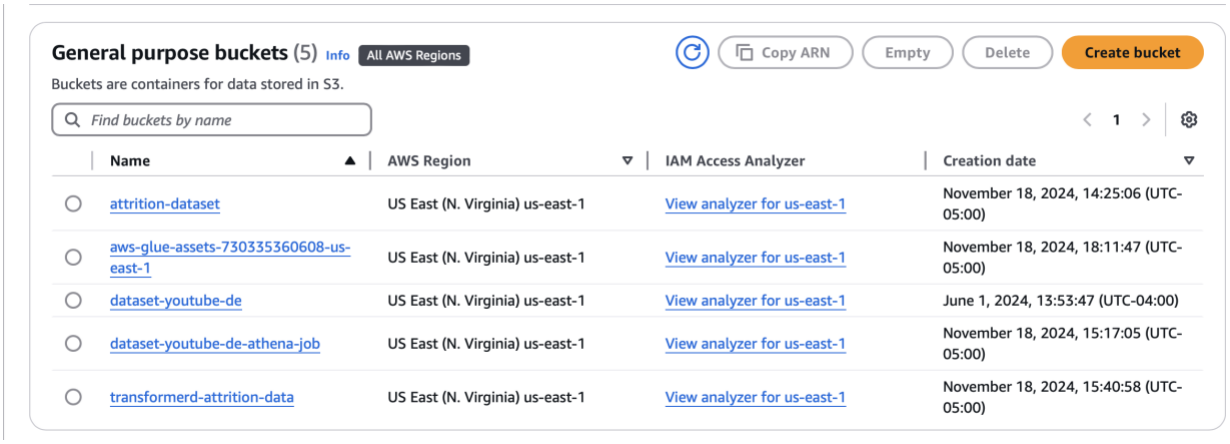
6. Athena:
Using Athena, we performed various analyses, extracted insights, and generated reports by querying the tables generated by Glue's crawlers. This streamlined our data exploration process and provided real-time access to our transformed dataset. By integrating these AWS services seamlessly, we constructed a resilient data pipeline that facilitated data ingestion, transformation, and analysis, enabling us to derive valuable insights from our raw data with ease.

Key Benefits of This Pipeline:
• **Integration**: Seamlessly integrates raw data from multiple sources into a unified dataset.
• **Scalability**: Leverages Amazon S3's scalability for both input and output storage.

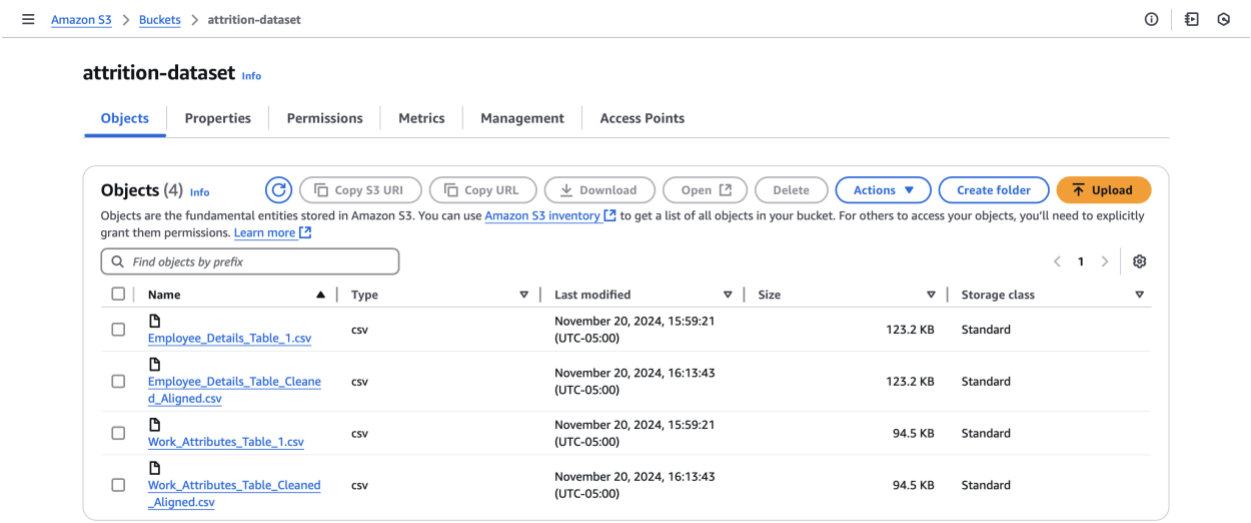
• **Flexibility**: SQL-based transformations allow dynamic customization of queries.
• **Automation**: Provides an automated workflow for data transformation and storage.

ETL Implementation:

1. Buckets Created

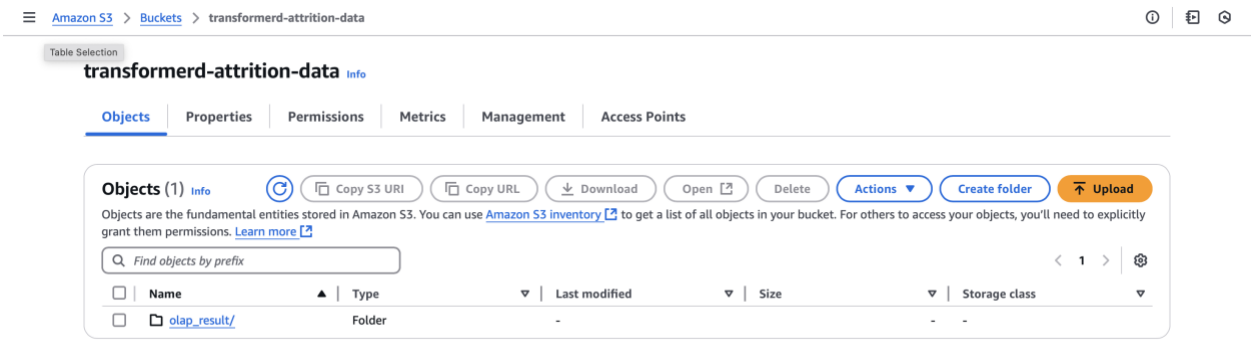


Note: We have two buckets one that stores the un-transformed data and the second one for the transformed data.



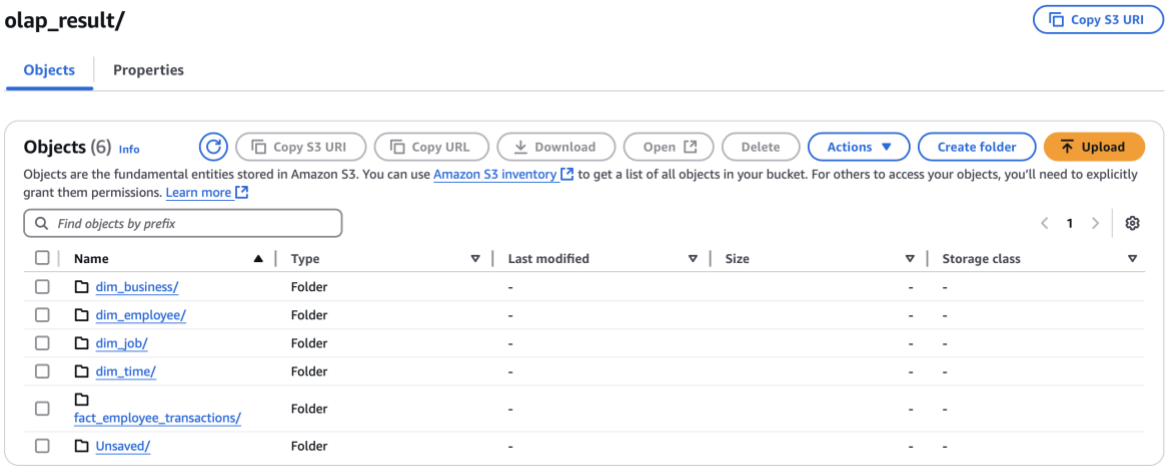
Note: We have two sources. One that holds the employee details and the other where the final transformed data is stored.

SOURCE BUCKET



DESTINATION BUCKET

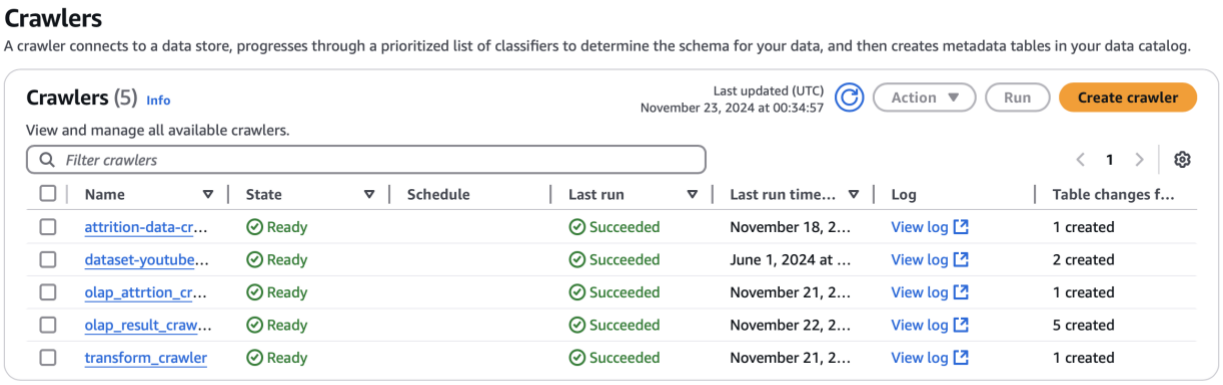
Holds the final transformed dimensional data.



TRASNFORMERD DATA

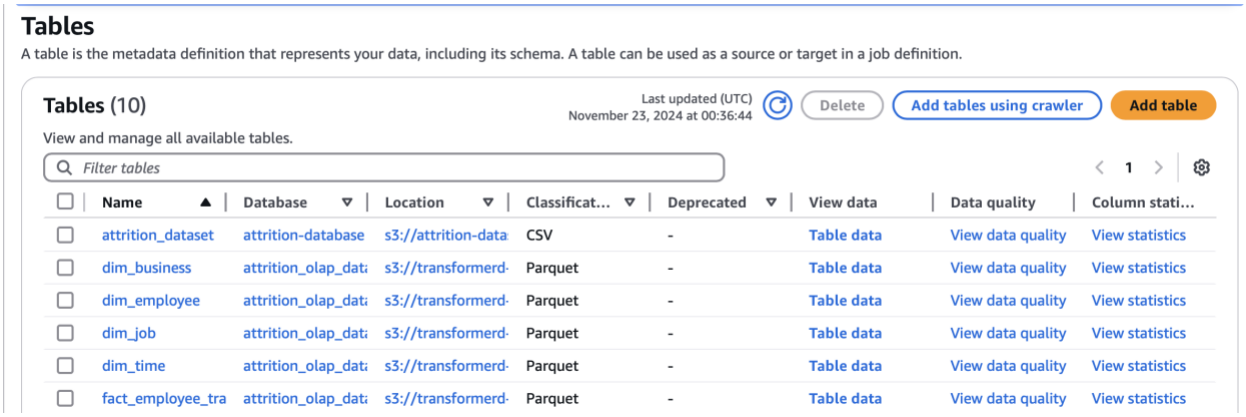
These are the 5-dimensional tables

2. CRAWLERS

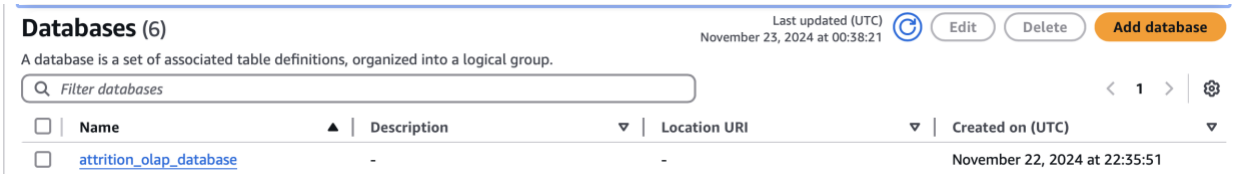


CRAWLERS USED

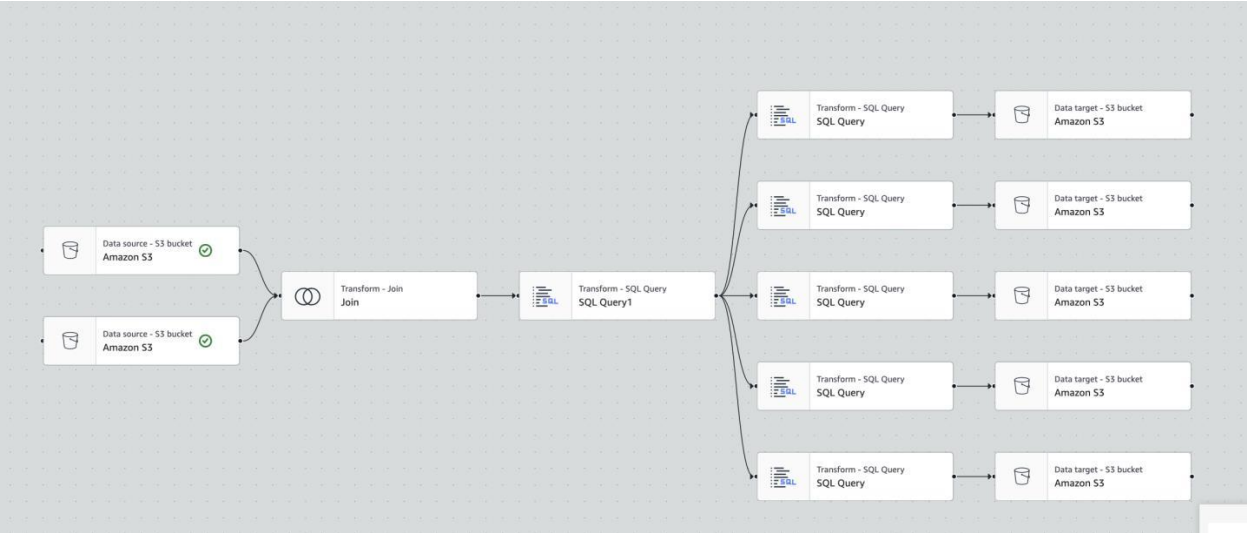
3. TABLES CREATED FROM THE CRAWLER



4. DATABASE HOLDING THE TABLES

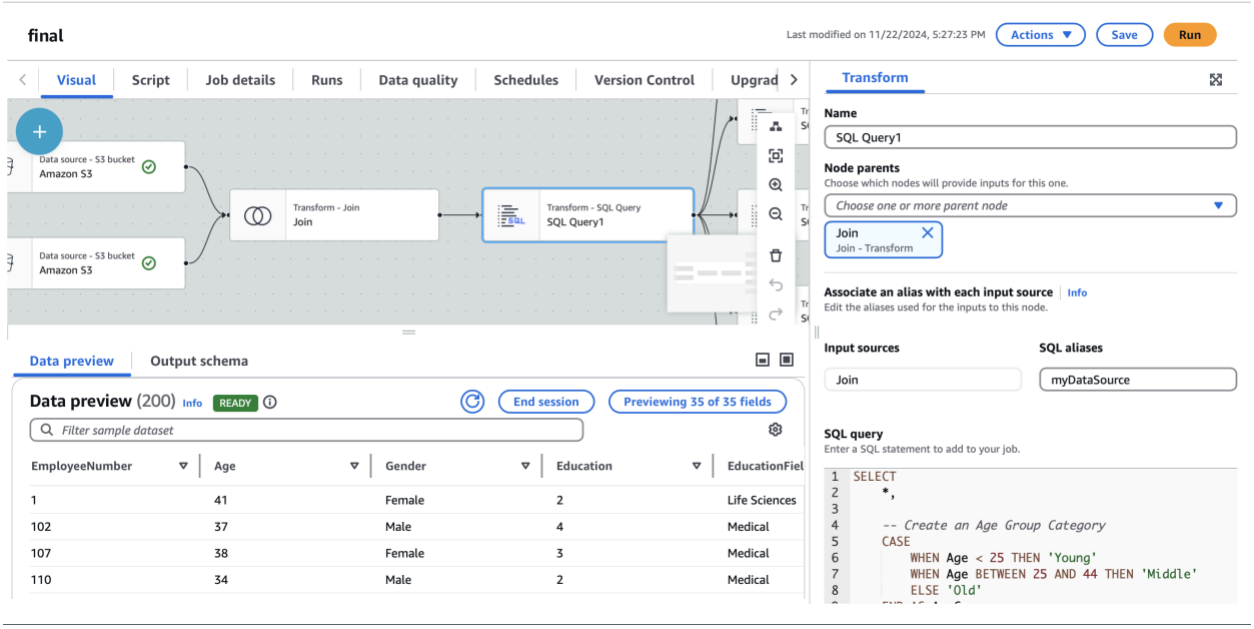


5. VISUAL OF ETL

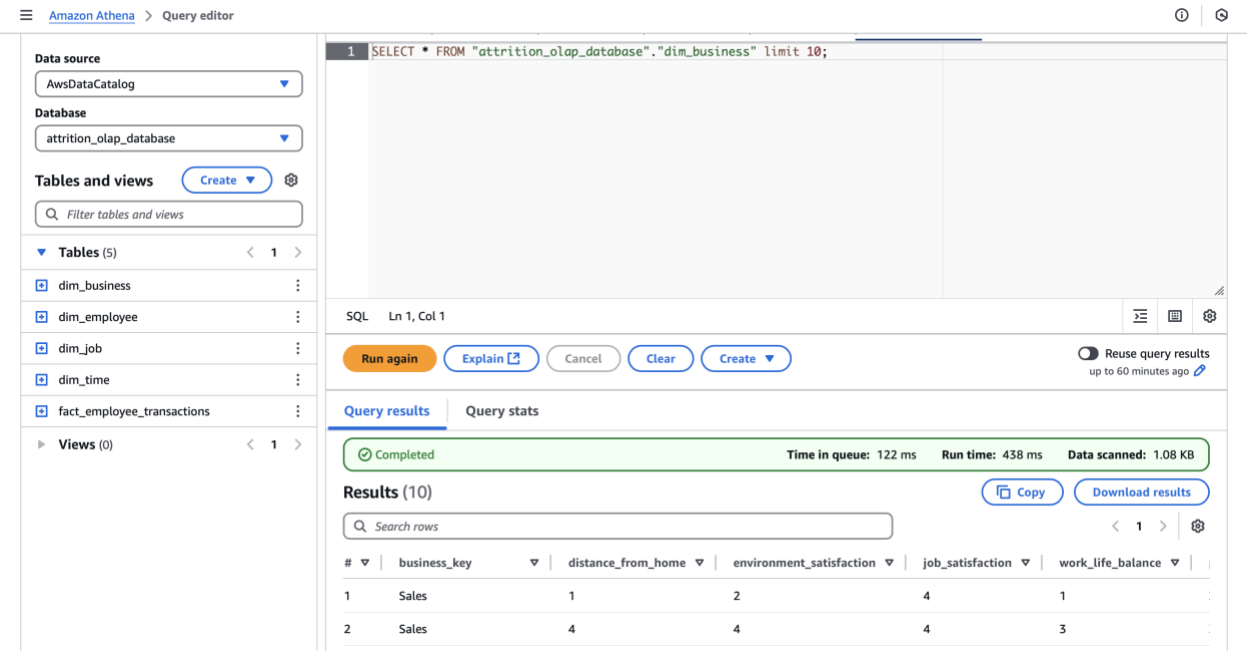


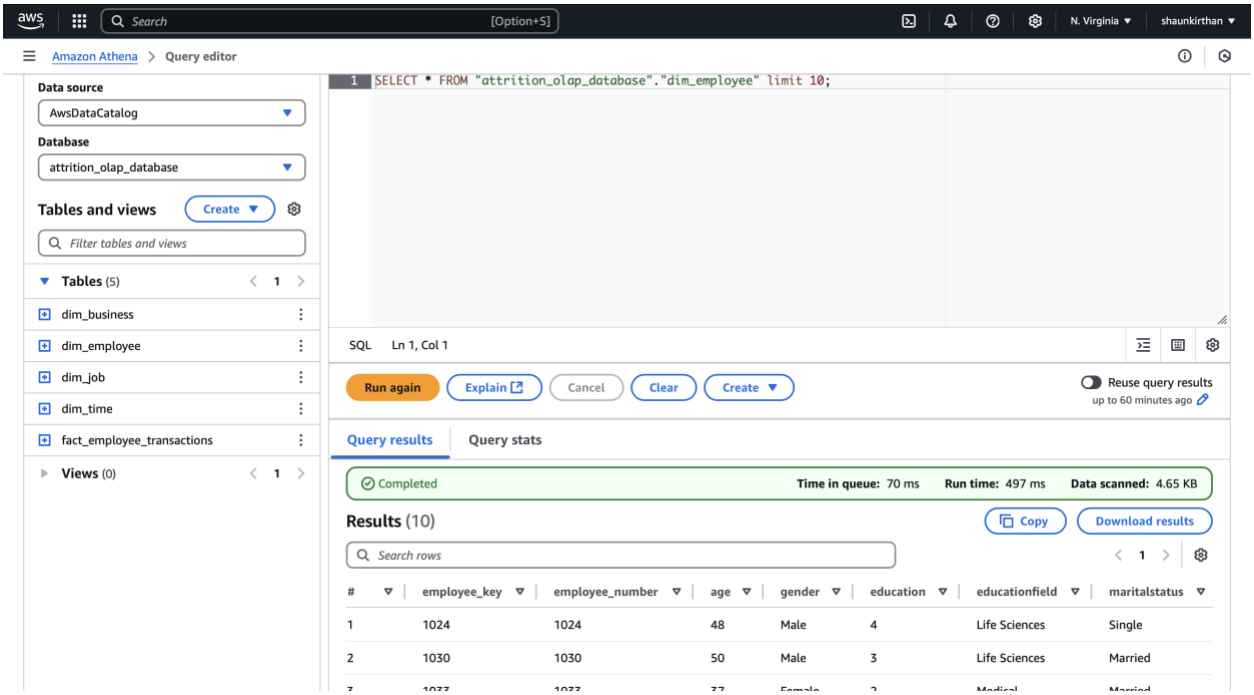
Two tables are retrieved from Amazon S3 buckets and joined based on a common key to create a unified dataset. Subsequent transformation steps involve data cleaning and enrichment, where unnecessary features are dropped, new features are derived, and data is aggregated to produce meaningful metrics. A specialized transformation step further converts the data from an OLTP (Online Transaction Processing) format to an OLAP (Online Analytical Processing) format, optimizing it for multidimensional analysis. Finally, the transformed datasets are stored back into Amazon S3 buckets, ready for querying or integration into downstream analytics tools. This streamlined workflow ensures efficient data preparation for advanced reporting and visualization.

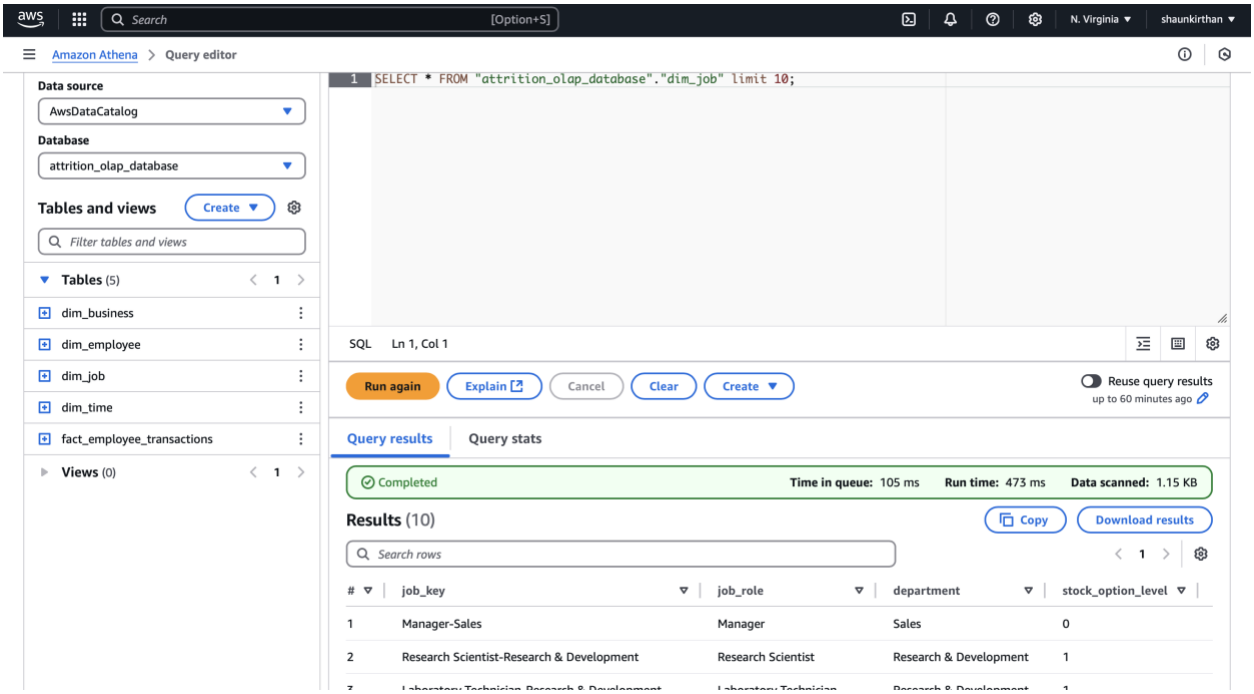
6. SQL Transformation

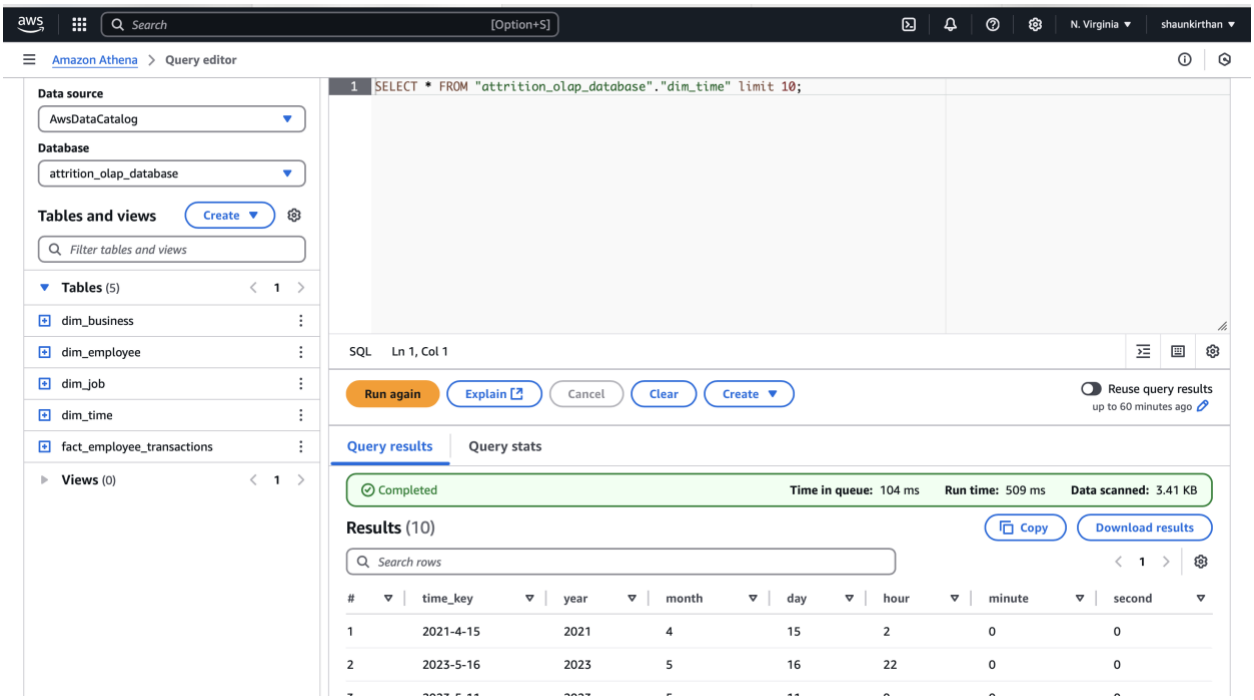

final
Last modified on 11/22/2024, 5:27:23 PM
Actions
Save
Run
Visual
Script
Job details
Runs
Data quality
Schedules
Version Control
Upgrad
Data source - S3 bucket
Amazon S3
Data source - S3 bucket
Amazon S3
Transform - Join
Join
Transform - SQL Query
SQL Query1
Transform
Name
SQL Query1
Node parents
Choose which nodes will provide inputs for this one.
Choose one or more parent node
Join
Join - Transform
Associate an alias with each input source
Info
Edit the aliases used for the inputs to this node.
Input sources
SQL aliases
Join
myDataSource
Data preview
Output schema
Data preview (200)
Info
READY
End session
Previewing 35 of 35 fields
Filter sample dataset
EmployeeNumber
Age
Gender
Education
EducationFiel
1
41
Female
2
Life Sciences
102
37
Male
4
Medical
107
38
Female
3
Medical
110
34
Male
2
Medical
SQL query
Enter a SQL statement to add to your job.
SELECT
*,
-- Create an Age Group Category
CASE
WHEN Age < 25 THEN 'Young'
WHEN Age BETWEEN 25 AND 44 THEN 'Middle'
ELSE 'Old'

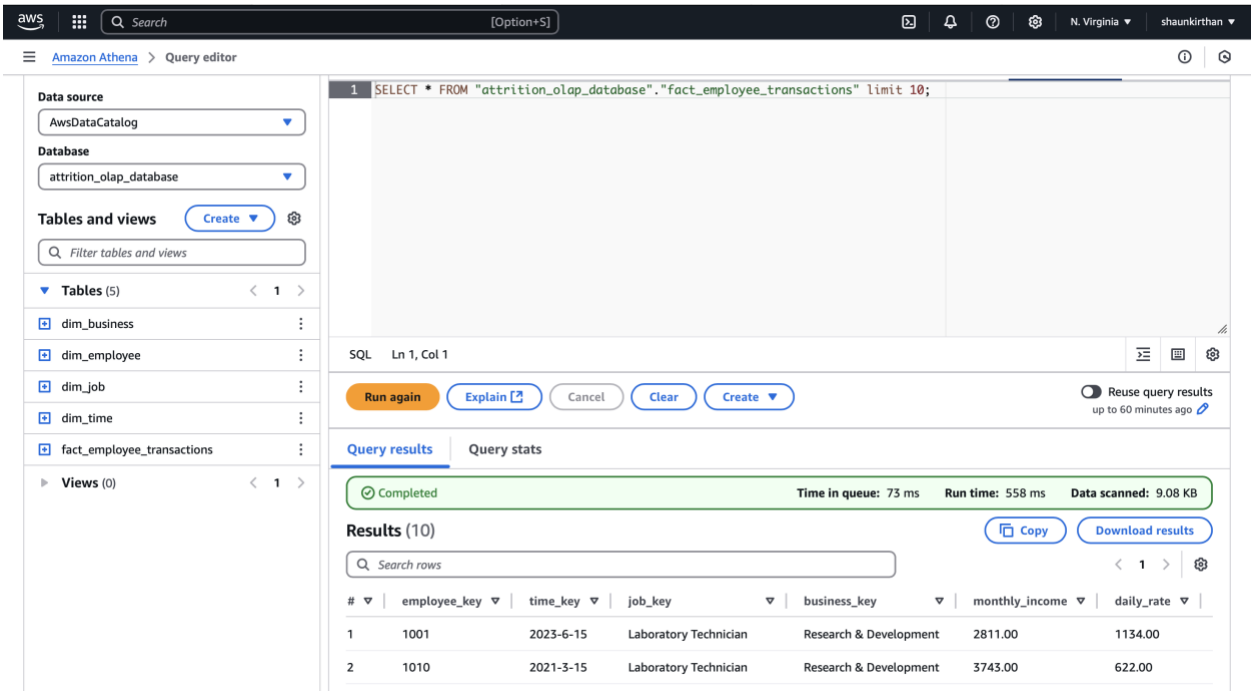
7. The Tables


Amazon Athena
Query editor
Data source
AwsDataCatalog
Database
attrition_olap_database
Tables and views
Create
Filter tables and views
Tables (5)
dim_business
dim_employee
dim_job
dim_time
fact_employee_transactions
Views (0)
SELECT * FROM "attrition_olap_database"."dim_business" limit 10;
SQL
Ln 1, Col 1
Run again
Explain
Cancel
Clear
Create
Reuse query results
up to 60 minutes ago
Query results
Query stats
Completed
Time in queue: 122 ms
Run time: 438 ms
Data scanned: 1.08 KB
Results (10)
Copy
Download results
Search rows
business_key distance_from_home environment_satisfaction job_satisfaction work_life_balance
1 Sales 1 2 4 1
2 Sales 4 4 4 3


aws
Search
[Option+S]
N. Virginia
shaunkirthan
Amazon Athena
Query editor
Data source
AwsDataCatalog
Database
attrition_olap_database
Tables and views
Create
Filter tables and views
Tables (5)
dim_business
dim_employee
dim_job
dim_time
fact_employee_transactions
Views (0)
1 SELECT * FROM "attrition_olap_database"."dim_employee" limit 10;
SQL
Ln 1, Col 1
Run again
Explain
Cancel
Clear
Create
Reuse query results
up to 60 minutes ago
Query results
Query stats
Completed
Time in queue: 70 ms
Run time: 497 ms
Data scanned: 4.65 KB
Results (10)
Copy
Download results
Search rows
employee_key employee_number age gender education educationfield maritalstatus
1 1024 1024 48 Male 4 Life Sciences Single
2 1030 1030 50 Male 3 Life Sciences Married

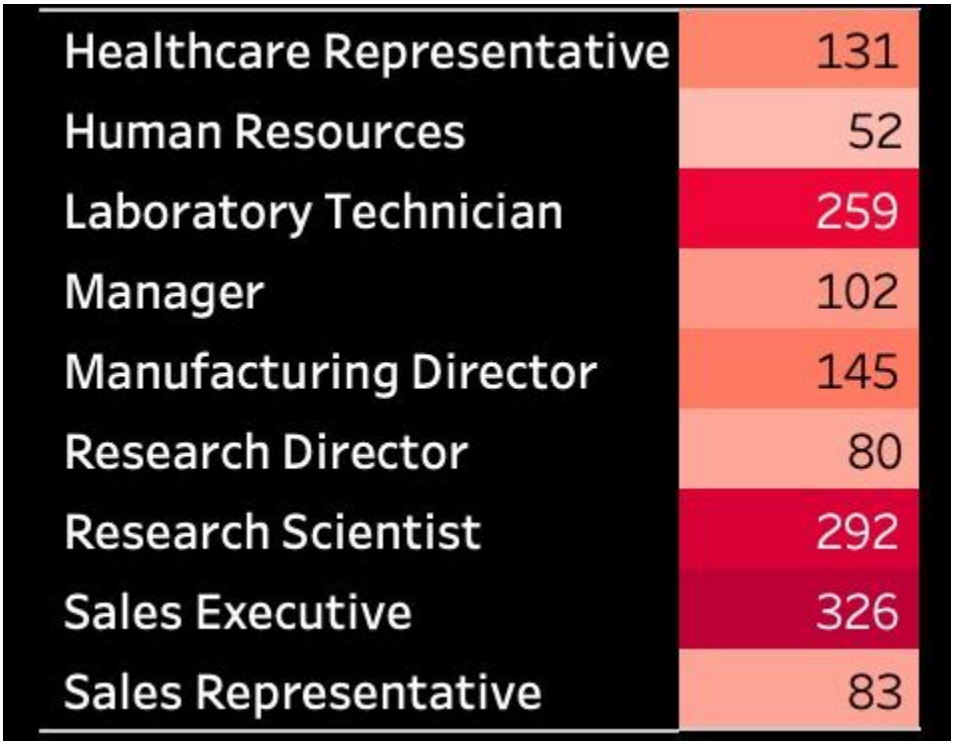

aws
Search
[Option+S]
N. Virginia
shaunkirthan
Amazon Athena
Query editor
Data source
AwsDataCatalog
Database
attrition_olap_database
Tables and views
Create
Filter tables and views
Tables (5)
dim_business
dim_employee
dim_job
dim_time
fact_employee_transactions
Views (0)
1 SELECT * FROM "attrition_olap_database"."dim_job" limit 10;
SQL
Ln 1, Col 1
Run again
Explain
Cancel
Clear
Create
Reuse query results
up to 60 minutes ago
Query results
Query stats
Completed
Time in queue: 105 ms
Run time: 473 ms
Data scanned: 1.15 KB
Results (10)
Copy
Download results
Search rows
job_key job_role department stock_option_level
1 Manager-Sales Manager Sales 0
2 Research Scientist-Research & Development Research Scientist Research & Development 1





KPIs, METRICS AND DASHBOARDS FOR ON CLOUD PROJECT:

KPIs:



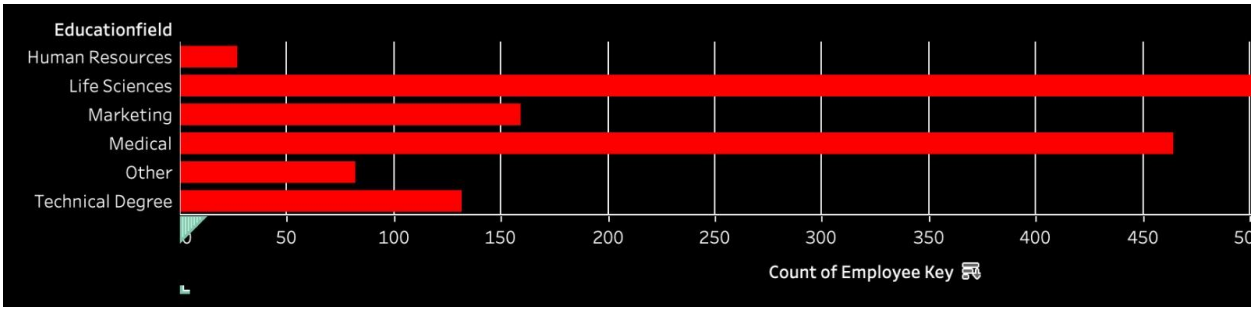
KPIs:

- The gender breakdown (588 females and 882 males) and yearly attrition trends (2020-2023 counts) from the dashboard are measurable outcomes used as KPIs.
- Job roles like Sales Executive, Research Scientist, etc., shown in the heatmap/tree map, represent role-based attrition metrics.

KPIs:

The "Attritions: 1,470" is a direct indicator of total employee attrition, which is a clear KPI.

METRICS:



Attrition by Education Field: A key performance indicator derived from the chart is the total attrition for each education field:

- Life Sciences: Highest attrition count (nearly 500).
- Marketing: Significant attrition count (second highest).
- Medical, Technical Degree, and Other: Moderate attrition counts.
- Human Resources: Lowest attrition count.

Dominant Education Field Impact: The chart identifies Life Sciences and Marketing as the most impacted fields, which could guide targeted retention strategies.

Metrics

1. Count of Employee Key: The count represents the total number of employees from each education field who left the organization.
2. Field Comparison: Provides insight into which education field contributes most to attrition relative to others.

DASHBOARDS:

Dashboard Explanation:

1. **Attrition Overview (Top-Left: Attrition Count & Gender Breakdown):**
    - Total Attrition Count: A bold number (1,470) displayed prominently indicates the severity of attrition.
    - Gender Breakdown:
        - Pie chart and numbers show that male attrition (882) is higher than female attrition (588).
        - This could highlight gender-specific challenges in the workplace, such as different expectations or work-life balance.

Insight: HR policies might need to address specific issues that affect males more significantly to balance attrition rates.

2. **Total Working Years vs. Attrition (Line Chart):**
    - The graph shows attrition peaks for employees with less than 10 years of total working experience.
    - A steady decline is observed as working years increase, with minimal attrition for those with 20+ years.

Insight: Newer employees are at higher risk of leaving. Enhanced onboarding, mentorship, or early career development programs could improve retention.

3. **Overtime/Marital Status Correlation (Bar Chart):**
    - Married employees show the highest attrition, followed by single and divorced employees.
    - This could imply that married employees may feel greater pressure balancing work-life responsibilities, especially when required to work overtime.

Insight: Flexible schedules, better overtime policies, or work-life balance initiatives could help reduce attrition among married employees.

4. **Yearly Attrition Trend (Bar Chart):**
    - A steady attrition trend is observed from 2020–2023:
        - 2020 has the highest attrition (394), slightly decreasing in 2021 and 2022, before rising again in 2023 (376).
    - These trends might indicate external factors influencing attrition, such as changes in workplace policies, economic conditions, or organizational restructuring.

Insight: Analyze specific organizational changes in these years to understand spikes and dips in attrition.

5. **Education Field Analysis (Bottom-Right Bar Chart):**
    - Life Sciences and Marketing education fields show the highest attrition, while Human Resources has the lowest.

- This suggests that certain educational backgrounds might be more susceptible to job dissatisfaction or misalignment with roles.

Insight: Investigate job-role alignment for these fields and introduce tailored retention strategies.

6. **Age Distribution (Histogram):**
    - Most attrition occurs in the 25–35 age range, typically representing early- to mid-career employees.
    - Attrition decreases after age 35 and becomes negligible after age 50.

Insight: Young employees might be leaving due to career growth opportunities or dissatisfaction. Programs targeting career advancement, skill-building, and mentorship can mitigate this.

Strengths of the Dashboard

1. Comprehensive Visualization:
    - Incorporates multiple aspects of employee demographics, work experience, and job roles.
    - Effective use of visualizations like bar charts, histograms, line graphs, and tree maps.
2. Actionable Data:
    - Directly links attrition factors to employee attributes (gender, marital status, role, age, etc.), making insights easy to derive.
3. Balanced View:
    - Combines high-level overviews (e.g., total attrition) with detailed breakdowns (e.g., role- or age-specific insights).

Conclusion:

The project successfully addresses the challenge of employee attrition by consolidating disparate datasets and applying advanced analytical techniques to derive actionable insights. Utilizing AWS services, a scalable and efficient data pipeline was developed to transform raw HR data into structured, analytical-ready formats. This solution not only identifies key drivers of employee turnover, such as job satisfaction, work-life balance, and compensation, but also provides organizations with a powerful tool for making data-driven decisions.

The approach ensures seamless data integration, optimized data processing, and meaningful visualization of insights, ultimately supporting HR teams in proactively addressing attrition risks and fostering a stable and satisfied workforce. This system sets the stage for future enhancements, including predictive analytics and broader HR applications, making it an invaluable resource for strategic decision-making in workforce management.