

Data Engineering Cohort 1 Module 5 Assignment 5.4 Earning Prediction AWS

Member #1

Name: Muhammad Humza

Roll Number: 2211-018-KHI-DEG

Member #2

Name: Muhammad Ovais

Roll Number: 2211-023-KHI-DEG

Task

Activity tasks/4_data_pipelines/day_4_data_lake/data/output_data/employee_earnings

Using the data manipulation tool of your choice (eg. Python) simulate the earnings predictions for 2 more days. Load it to the Data Lake that you've created today

Rerun queries from Task 3 and Task 4 and see how the results change with this new data.

Create a new query in Athena that calculates the % change in earnings for every employee from a given day compared to the previous day.

Solution

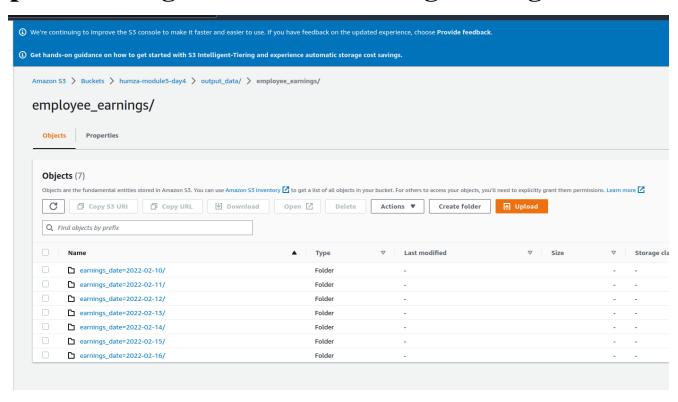
Step 1: using python file to predict earnings

The python file is attached in the .zip file you can find it in Assignment.ipynb format

Step 2: 5.2 Assignment

All the steps are similar to assignment 5.2 you can refer to it for your understanding

Step 3: Creating a bucket and adding earnings files in it



Step 3: SQL Query to find difference in earnings

QUERY

SELECT emp_id, full_name, earnings, earnings_date, (earnings - previous_earnings) / previous_earnings FROM (SELECT emp_id, CONCAT(first_name, ' ', middle_initial, ' ', last_name) AS full_name, LAG(CAST(earnings as double), 1) OVER (PARTITION BY emp_id ORDER BY earnings_date FROM "ovaissaleem_glue_database". "muhammadhumza_employee_earnings") t WHERE earnings_date BETWEEN '2022-02-14' AND '2022-02-16'

