File Name	Description
emr_solution_1.txt	PySpark script for the 1st iteration. Merges new employee data with existing data using a hardcoded schema.
emr_solution_1_lambda_function.txt	AWS Lambda function triggering the 1st PySpark job on AWS EMR, aiming at processing employee data.
emr_solution_2.txt	PySpark script for the 2nd iteration. Enhances data merging with dynamic file retrieval and schema validation.
emr_solution_2and3_lambda_function.txt	Lambda function for triggering the 2nd and 3rd iterations of PySpark jobs, introducing dynamic scalability based on dataset size.
emr_solution_3.txt	PySpark script for the 3rd iteration. Introduces schema inference for flexible data processing, adjusting to evolving dataset structures.
dataset_1_existing_employee_data.csv	Existing employee dataset used in the 1st iteration of the EMR solution.
dataset_1_new_employee.csv	New employee data for merging with existing data in the 1st iteration.
dataset_1_merged_employee_data.csv	Result of merging new and existing employee datasets in the 1st iteration.
dataset_2_existing_salaries.csv	Existing salary dataset used in the 2nd iteration for schema validation and dynamic merging.
dataset_2_merged_salaries.csv	Output of the 2nd iteration showing merged salary data, showcasing dynamic handling.
dataset_2_new_salaries.csv	New salary data introduced in the 3rd iteration for testing schema inference capabilities.