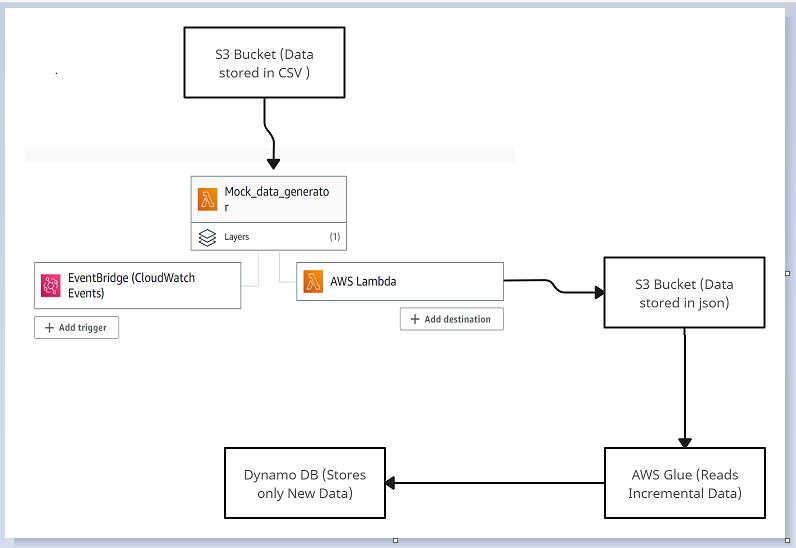
**AWS Data Pipeline Project Version 1**

**Overview of the Project:**

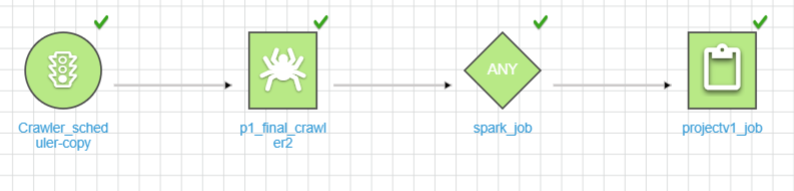
Data stored in **S3 Bucket** in CSV format is read by First **Lambda** Function using pandas library and , in this lambda function data will be shuffled and ¼ th of the data is sent to Second Lambda Function in JSON format. From second Lambda function data will be dumped to a new S3 Bucket. From this S3 Bucket, Incremental Data will be read by **AWS Glue** jobs which is created using **Spark** (Crawlers are used to extract metadata information of the stored objects in s3) and dumped into DynamoDB which receives only new data which is not present in **DynamoDB**.

A workflow is created to automate the crawler and spark job executions. And the Part1 and Part2 Diagrams are scheduled using **AWS Event Bridge. AWS IAM** roles are necessary between each aws services to make the flow of data between different aws services possible without restrictions.

**Architecture:**



Part 1: Flow of Data



Part 2: Workflow (crawler and spark tasks)

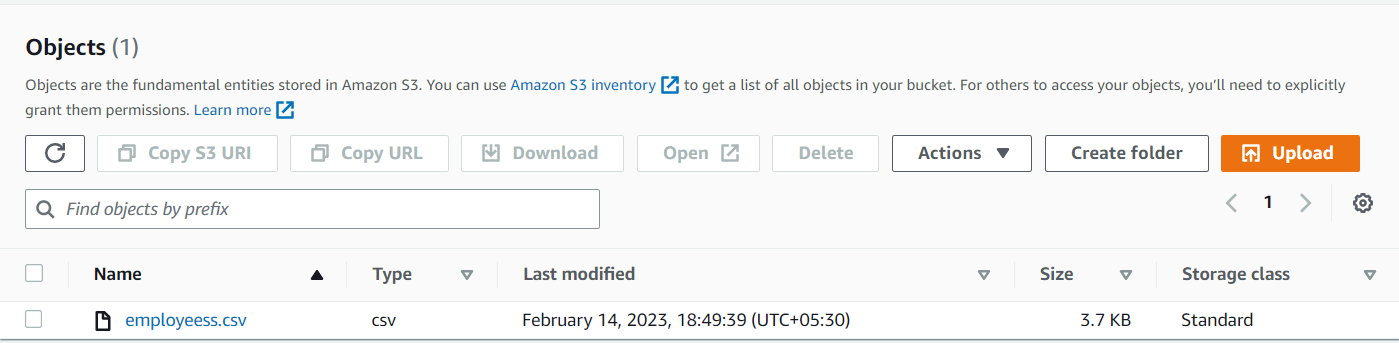
## Prerequisites:

## Requires and AWS account with appropriate AWS Identity and IAM permissions.

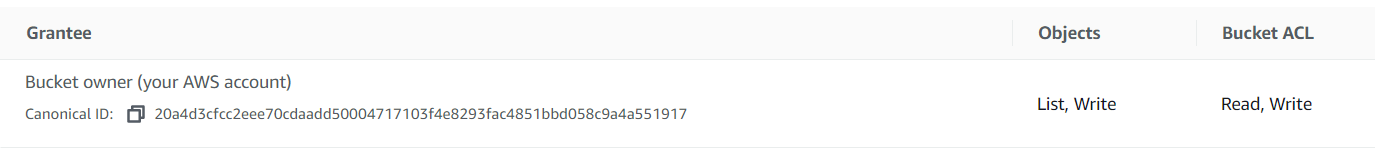
**Deploying Solution:**

**Step1 :**

1.Create S3 Bucket and dump data with CSV format.



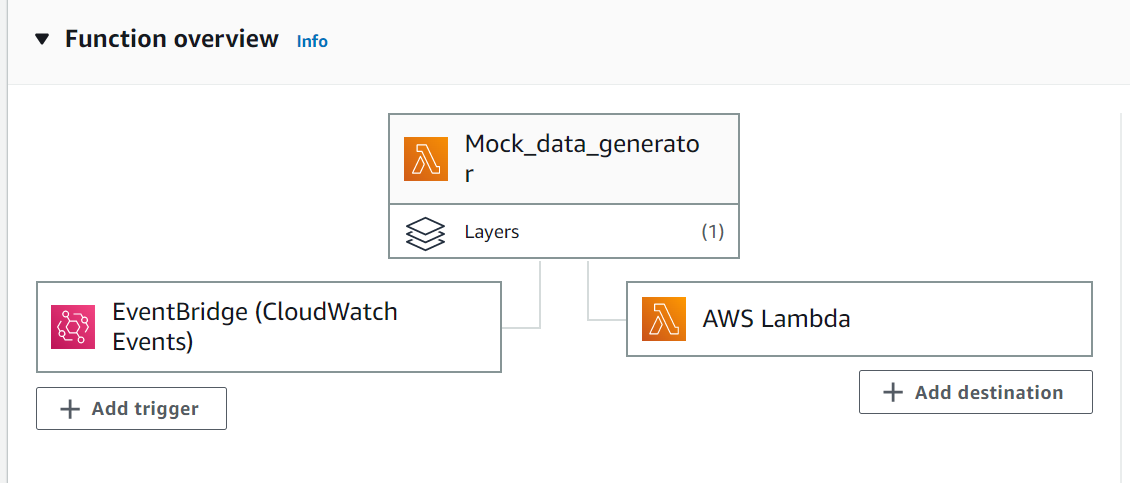
2.Grant access to Read,List and Write Data from S3.



**Step2 :**

1.Create a Lambda Function which reads the data from s3, shuffles the data and writes ¼ th of the data to another Lambda Function in Json Format.

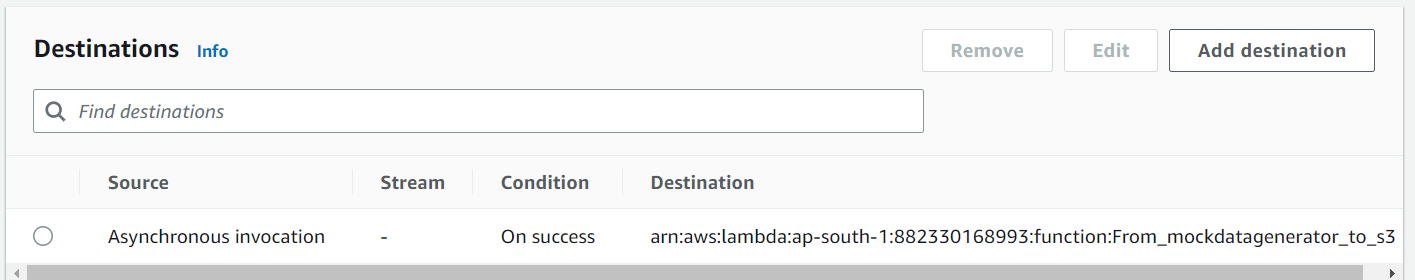
**First Lambda Function File name**: Mock\_data\_generator.py



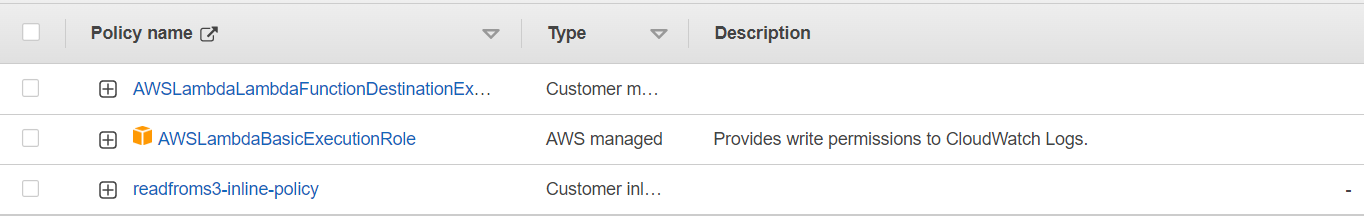
2.Add Event Bridge as a Trigger to schedule automation of Lambda Functions, And Second Lambda function is added as destination where Transformed data will be sent.From Second Lambda Function, data will be dumped to s3.

**Second Lambda Function File name**: Lambda\_to\_s3.py

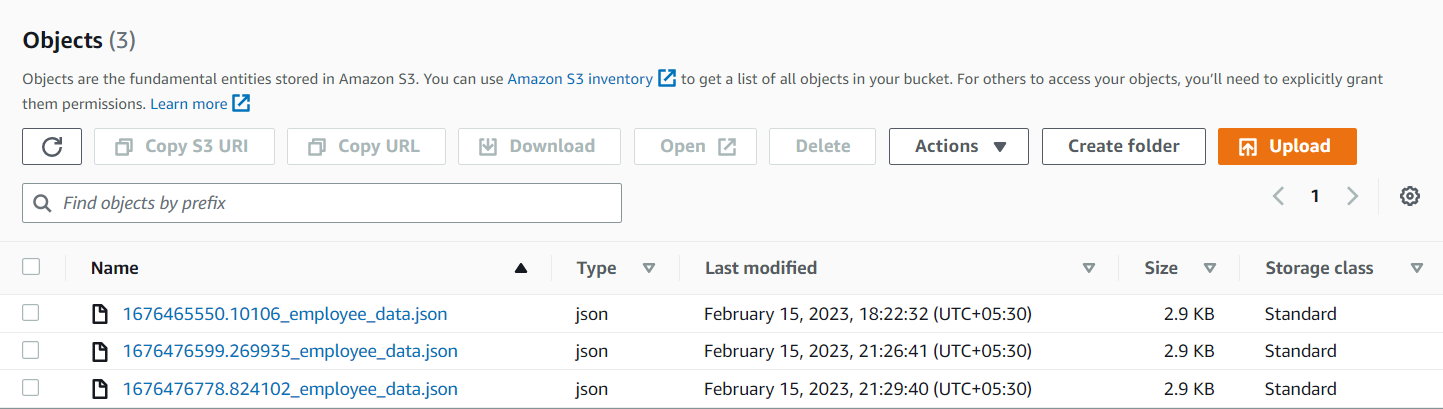
On success of First Lambda Function data will be sent to Second lambda Function.PFB screenshot for reference



3.Grant Necessary Permissions to Lambda Functions to read data from s3 and to generate **cloudwatch Logs**.

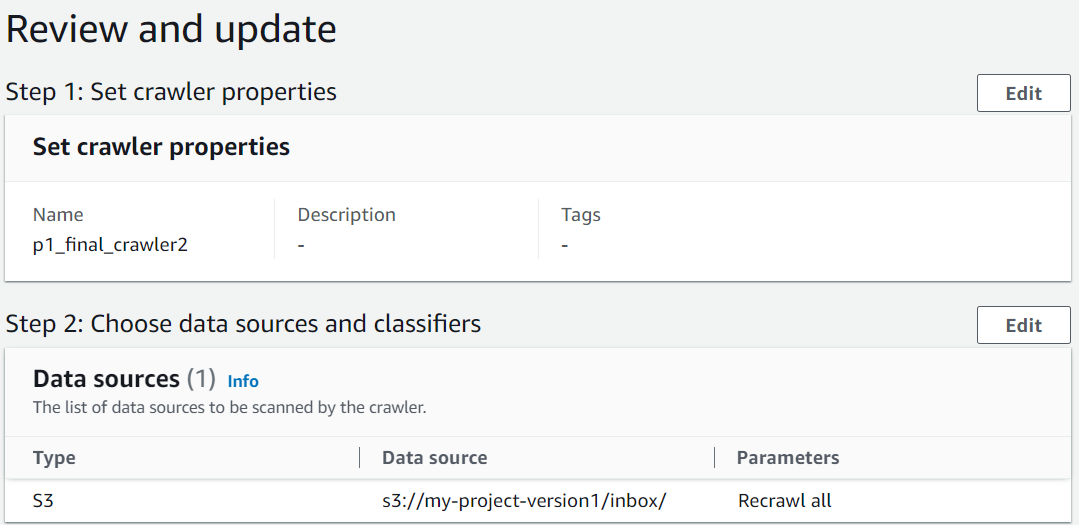


4.Based on the Frequency of EventBridge execution data will be generated in S3.

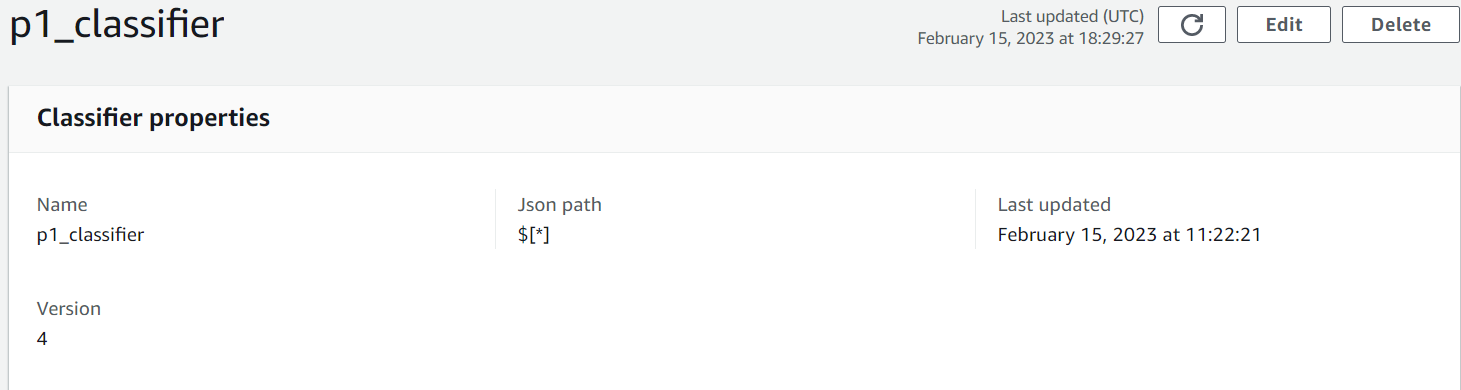


**Step3:**

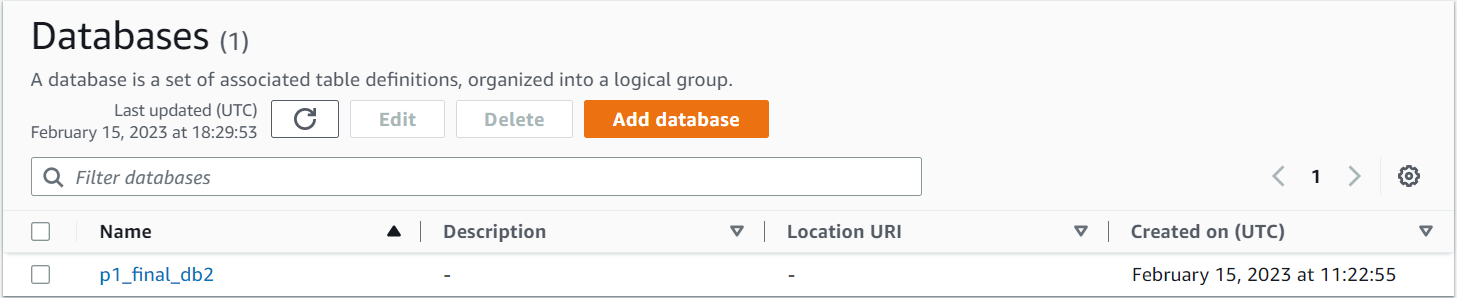
1.Create a crawler in AWS Glue to extract the metadata information of object stored in S3, And create Database and Datasource and point the crawler for the specific s3 location from where you need to read data.



2.Specify Classifier if data format is other than CSV.

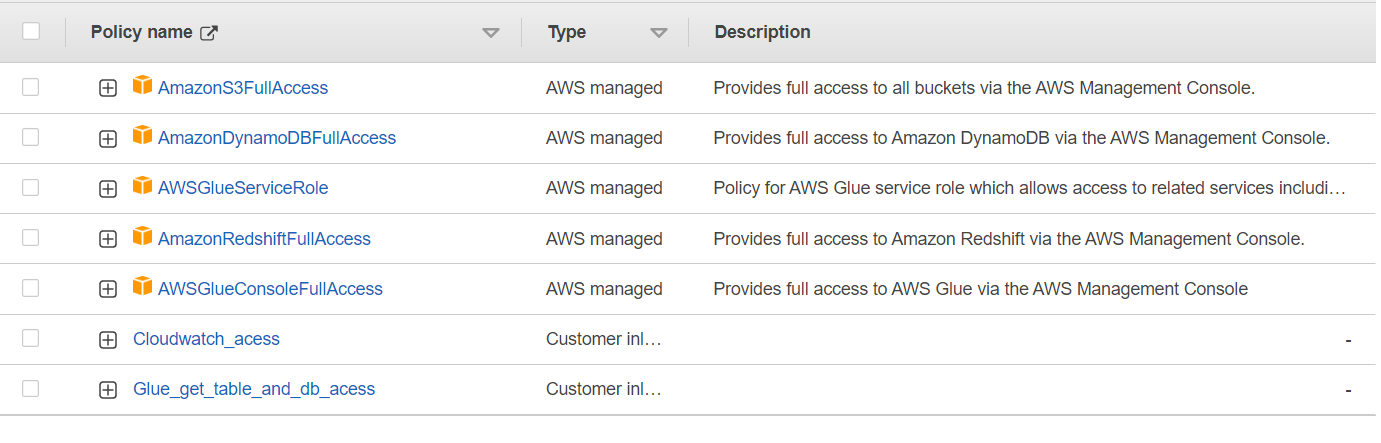


3.Create a Database where you need to store the metadata of the object you are reading.

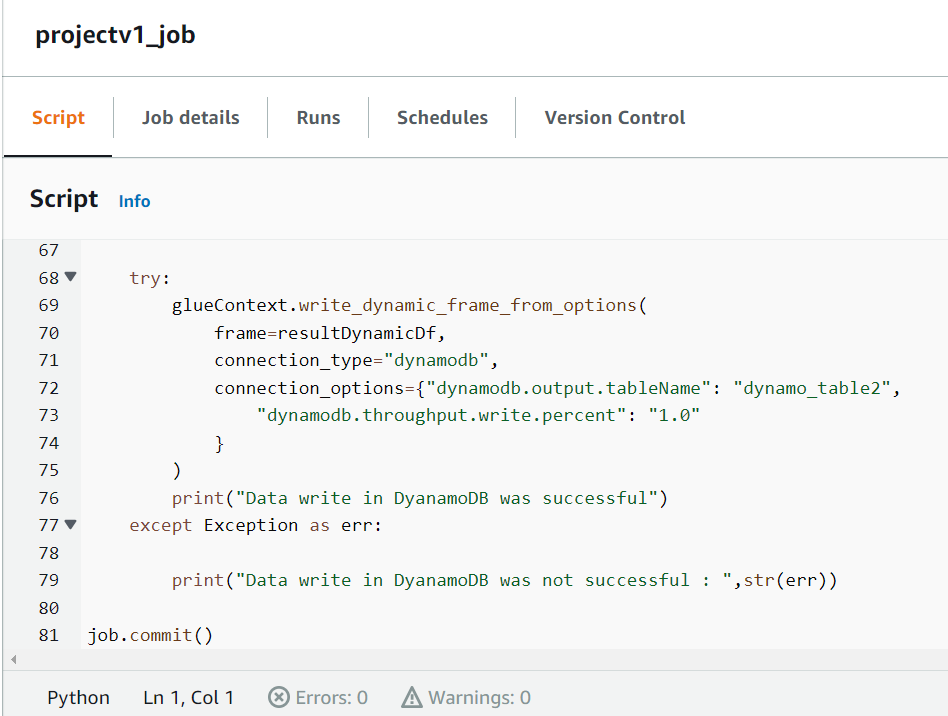


**Step4:**

1.Create a spark job in AWS Glue, which reads data from table which was created from crawler. Specify the IAM roles need to execute the job.



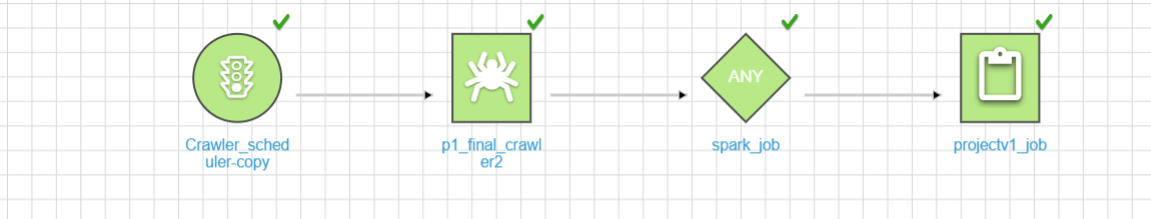
2.Enable job Book mark to read data incrementally from s3 and write query such a way that it removes duplictates from incoming data.



**Spark Code File Name**: Spark\_Code.py

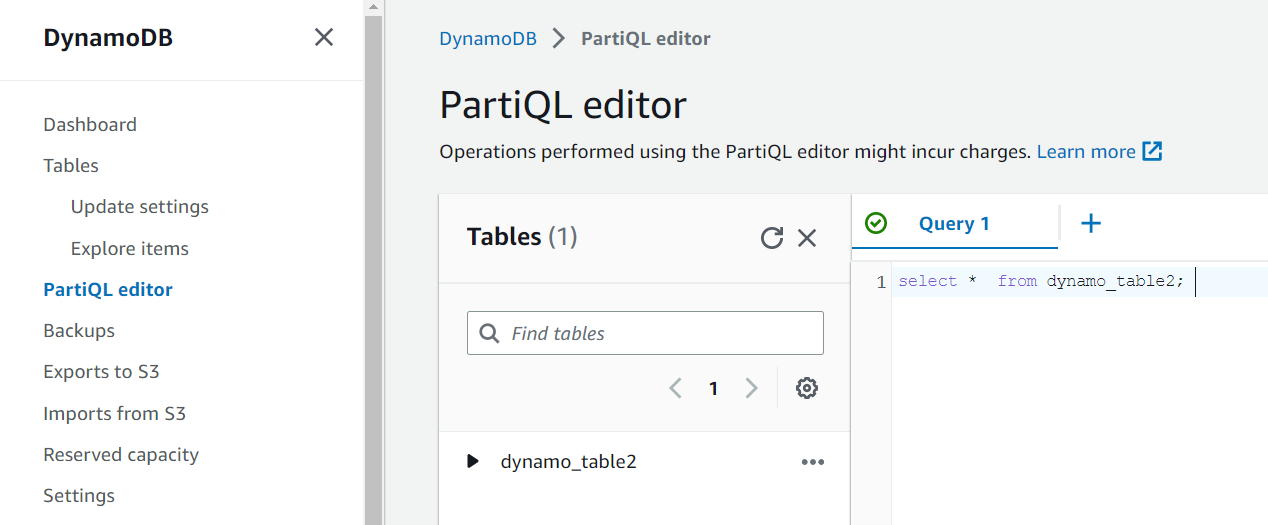
**Step5:**

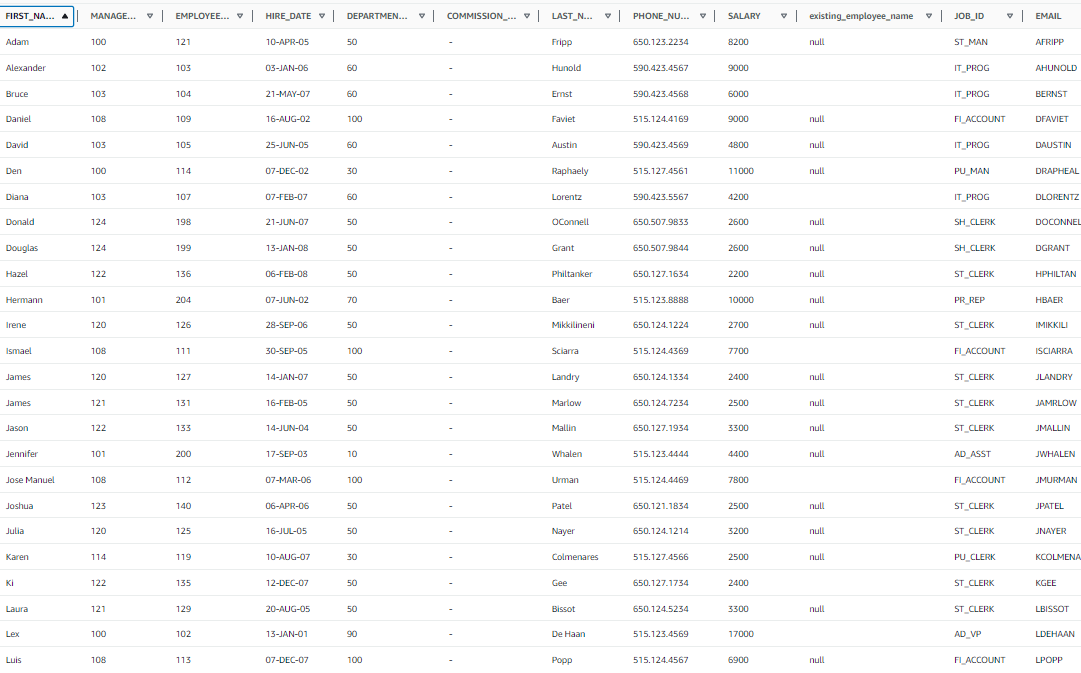
Create a Workflow which automates crawler and spark job execution and schedule the workflow using Event Bridge.



**Step6:**

Create DynamoDB where you need to store the destination data and Load the data into other aws services or query the data using query editor or other service.





Final Data stored in DynamoDB