Overview

Process to build a data pipeline to automate data ingestion processing and transformation

* Extract streaming data from APIs, process and transform using Apache Spark (Pyspark)
* Write the streaming output processed data into AWS S3 bucker.
* The process of extraction to saving data is orchestrated using Amazon Managed Workflows for Apache Airflow (MWAA)
* Parse the data from S3 bucket into AWS Athena using IAM roles
* Query data using AWS Athena.
* Create visualization and analysis of data using ML algorithms for real time predictions

Languages: SQL, Python3, Pyspark

Services: Amazon Managed Workflows for Apache Airflow (MWAA), AWS S3, AWS Athena, Tableau or AWS Quicksight

Business Impact of building data pipelines

* Saves time and effort for BI analysts
* Manual process of cleaning, validating of raw data takes on an average of 3-4 hours of time for each and every raw file received at a certain interval of time.
* Automating this simple process saves 3-4 hours for each run and saves effort and time for Data engineers as well as Analysts
* Only pipeline monitoring and maintenance needs to be taken care of once the pipeline is built.
* Grabbing the data from source to destination-sources being APIs. Destination systems AWS S3 bucket for real time streaming data

Data Streaming Architecture

