# QUALITY DATA INGESTION WITH AWS GLUE

- **Bhushan Sagar**

.......................................................................................................

## Business Overview:

This project aims to ensure the quality of IMDb movie rating data before further analysis or usage. By applying data quality rules, it aims to identify and segregate data that does not meet predefined quality standards. The segregated data is stored separately for further inspection or corrective actions. The transformed and quality-assured data is loaded into an Amazon Redshift database, presumably for analytics, reporting, or other downstream processes.

## Dataset Description:

The IMDb Movies Rating Dataset contains information about various movies, including title, release year, certificate rating, runtime, genre, IMDb rating, overview, meta score, director, cast, number of votes, and gross earnings. It is used for analysing movie trends, audience preferences, and predicting movie success.

➔ Languages- ● SQL, Python3

➔ Services - ● AWS S3, AWS Glue, AWS Redshift, AWS SNS, AWS Athena, Power BI

## Redshift:

Amazon Redshift is a fully managed, petabyte-scale data warehouse service in the cloud. It allows businesses to analyse large amounts of data using SQL queries and provides high performance, scalability, and ease of management. Redshift is commonly used for data warehousing, analytics, and business intelligence applications.

## SNS:

Amazon Simple Notification Service (SNS) is a fully managed messaging service for distributed systems. It enables applications to send notifications to users or other applications via SMS, email, or push notifications. SNS simplifies the process of building and managing scalable, reliable, and highly available notification workflows. It's often used for event-driven architectures, alerts, and notifications in cloud-based applications.

## Steps:

1. Data Extraction: Extract IMDb movie rating data from an S3 bucket using AWS Glue's DynamicFrame.

2. Data Quality Evaluation: It applies a set of data quality rules to the extracted data using AWS Glue's EvaluateDataQuality functionality. These rules include checks for completeness, uniqueness, column length, specific column values, standard deviation, and row count.

3. Conditional Routing: After the data quality evaluation, the script routes the data into two groups based on the evaluation results. One group contains data that failed the quality checks, while the other contains data that passed.

4. Schema Transformation: For the data that passed the quality checks, the script applies schema changes, converting certain data types and possibly renaming columns.

5. Data Destination:

- The data that failed quality checks is written to an S3 bucket in JSON format.
- The overall evaluation outcomes, including both row-level outcomes and rule outcomes, are written to another S3 bucket in JSON format.
- The transformed and quality-assured data is loaded into an Amazon Redshift database table.

6. Notification: The script may incorporate Amazon SNS to send notifications or alerts regarding the processing status or any anomalies encountered during the process.

## Learnings/Takeaways:

1. Data Quality: Understanding the importance of maintaining data quality through checks and filters.

2. ETL: Getting insights into Extract, Transform, Load processes for data preparation.

3. AWS Glue: Learning about AWS Glue for ETL tasks, simplifying data processing.

4. Concurrency: Utilizing concurrency for faster data processing.

5. Redshift: Exploring Amazon Redshift for scalable data warehousing.

6. SNS: Leveraging SNS for efficient notification workflows in applications.

…………………………………………………………………………………………………