**HR data Analysis using S3 datalake,AWS Glue and Athena**

In this Project, we will demonstrate how to detect real-time stock exchange data generated by stock market. To perform this task, Apache Kafka will be used as a message queue, S3 as Data Lake and Spark-SQL for performing the analytics. This combination is well known for its usability, high throughput and low-latency characteristics. we will store all the historical data into S3and perform ELT and Data warehouse query using Spark and Athena

**Project Description:**

### **Goal:**

Design a HR Intelligence system that provide enable customer to build dashboard and perform search for specific stock data.

### **Objective:**

Based on data analytics, companies will be able to view employees details based on specific condition such as salary etc.

Please complete following tasks based on this dataset

1. Using AWS console upload employees data into S3 bucket "/user/ Mukesh/ /employees\_data" directory

2. Using AWS Glue, create a data frame with 5 columns id,age,gender,role and salary for this given employees text file data

4. Perform data cleansing activities to remove record with null values and if salary field contains string value then salary must be mentioned as 10000. After transformation write the data into a new data catalog bucketed table "Consultant\_Table\_Bucket" in the database “mukesh” having 4 buckets on the field salary. This table should store the data into columnar format ORC

5. Insert all those employees whose salary is greater than 50000 from "Consultant\_Table\_Bucket" table into new data catalog table “highsalaryEmployees” in “mukesh” database. While inserting into "Consultant\_Table\_Bucket" table you need to convert "consultant" role into "DataEngineer" role.

6.Write a AWS athena query to find out Max, min salary of " DataEngineer" from the " highsalaryEmployees " data catalog table