Akash Mahajan

AWS DATA ENGINEER

mahajanakash935@gmail.com
9172459659

PROFILE

Data Engineer with 4 years of experience in ETL development and building Data Warehouse solution by availing technologies like Python, SQL, PySpark and AWS Services like EMR, Glue, Redshift, Athena, S3

TECHNICAL SKILLS

- **AWS Services:** EMR, Glue, Redshift, Athena, S3, EC2
- **Programming language:** Python, SQL
- **Workflow engine:** Airflow
- **Big Data Technologies:** PySpark
- **Versioning tools:** GitHub
- **Ticketing Tool:** JIRA
- **Database:** Oracle, MySQL
- **Data warehouses:** Redshift
- **Tools:** Oracle SQL developer, Jupyter, Pycharm

EDUCATION

Bachelor of Engineering

(Pune University)

PROFESSIONAL SUMMARY

- AWS Data Engineer with 4 years of experience in Python, Pyspark, SQL, and AWS services
- Experience in Airflow orchestration of ETL Data Pipeline
- Good experience in ETL development on AWS Glue and AWS EMR using PySpark
- Worked on ETL pipeline utilizing S3, Glue for Data Transformation, and Redshift for Data Warehousing
- Testing and Data Validation of transformed data using Athena
- Experienced in building Data Pipelines, Data Warehouses with good knowledge of Normalization, Slowly Changing Dimension (SCD), Fact and Dimension tables
- Good knowledge in AWS services like EMR, Glue, Redshift, S3, Athena, EC2, IAM
- Experience on Oracle Database and proficiency in writing complex SQL queries
- Good knowledge of SQL concepts like view, subquery, join, String, Window and Date functions
- Good Knowledge in Spark concepts like Advance Joins, Spark architecture, Spark submit, performance optimization, RDD, Dataframe
- Worked with different file formats such as Text files, CSV, JSON and parquet
- Good knowledge of python concepts like exception handling, flow control, OOPS concepts, Functions, Modules
- Experience of Python for Airflow DAG development and development of Utilities for ETL process
- Collaborated with cross-functional teams to understand business requirements and translate them into data engineering solutions
- Strong team player with good communication, analytical, presentation and interpersonal skills

Work Experience

Working with Adclever Solution Pvt Ltd from July 2019

Client: Tesco plc
Description: Tesco plc is a British multinational groceries and general merchandise retailer headquartered in Welwyn Garden City, England
Responsibilities:
- Developing Airflow DAG to orchestrate end to end ETL workflow from Sensing ingested files to dumping transformed data into Data Warehouse
- Developing ETL using Pyspark to transform data in S3 with Glue jobs
- Loading transformed data to Redshift data warehouse for appending incremental load in SCD2 fashion
- Converting OLTP schema to OLAP schemas like star and snowflake in form of Fact and Dimension tables
- Leveraged AWS EMR for big data processing and transformation of large datasets
- Testing and data validation of transformed data using Athena SQL Query engine
- Developing python scripts for implementing utilities needed for ETL jobs
- Analyzing Business Requirement Document and extracting insights to develop low level design
- Implement version control using GIT for efficient code management and collaboration
- Participate in daily scrum calls, sprint planning, sprint retrospection, Grooming calls

Client: Acuity Insurance
Description: Acuity Insurance is one of the largest insurers in the United States. They do not sell insurance directly, instead relying on independent agents.
Responsibilities:
- Leveraging AWS EMR for big data processing and analytics, enabling the analysis of large datasets
- Understanding the functional part of application and implement the logic in EMR steps
- Leveraging AWS S3 as Data Lake for storing intermediate results of EMR ETL jobs
- Implementing test scripts to support test driven development using Athena
- Creating Airflow DAG to automate and schedule the ETL workflow
- Creating Staging and historical tables in Redshift and loaded data from S3 to Redshift tables
- Execution and monitoring of ETL jobs along with troubleshooting in case of workflow disruption
- Involved in gathering the requirements, designing, development and testing
- Version control of transformation scripts and Airflow Dag using GIT
- Participate in daily scrum calls, sprint retrospection, sprint planning, Grooming calls

Client: Zalando Austria
Description: Zalando Austria is a subsidiary of Zalando SE which is a publicly traded German online retailer of shoes, fashion and beauty products active across Europe.
Responsibilities:
- Developing SQL Queries for creating tables, views, aggregate tables in Hive
- Running and monitoring daily ingestion job from oracle database to on premises data warehouse hive and resolving issues
- Data validation and testing of migration data by running SQL queries
- Pyspark code optimization and modularizing codes into utilities
- Developing SQL and Pyspark queries to fetch data from Data Warehouse
- Responsible for pull, push and committing the code using GIT