Amol Gawale

Data Engineer

amolgawale.cloud@gmail.com 7249661101

PROFILE

"Cloud Data Engineer with 4+ Years of Industry Expertise | Empowering Data-Driven Solutions for Enhanced Business Performance"| "Transforming Raw Data into Actionable Business Intelligence"

TECHNICAL SUMMARY

- Over 4 years of combined expertise using **SQL**, **Python**, **Hadoop**, **Hive**, **Apache Spark**, and **Amazon Cloud Platform**.
- Hands-on experience working with various AWS services including **S3**, **EC2**, **Glue**, **EMR**, **Redshift**, **CloudWatch**, and **Athena**.
- Good knowledge of additional AWS services such as **Lambda**, **DynamoDB**, **RDS**, **CloudWatch**, **IAM**, and **SNS**.
- Proficient in creating and managing Glue jobs with PySpark scripts for generating ETL Jobs based on client expectations.
- Proficient in working with **Spark RDD**, **Spark DataFrames**, and **Spark SQL**.
- Led the design and development of a scalable Data Engineering solution on AWS, leveraging S3 for data storage, EMR for development and testing, Glue for **ETL** processes, and Redshift for data warehousing and analysis.
- Utilized Athena for analytical queries and data transformations on S3-resident data.
- Developed Spark job **orchestration** using **Apache Airflow**.
- Worked with various file formats, including **Text** ,**CSV**, **JSON**, **Avro**, **ORC** and **Parquet**.
- Strong understanding of Hadoop architecture, **HDFS**, and **MapReduce** concepts.
- Capable of processing large sets and high volumes of **structured** and **semi-structured data**.
- Extensive experience in **Data Cleaning** and applying transformations using PySpark based on client requirements.
- Good knowledge of database concepts such as **Fact Tables**, **Dimension Tables**, and experience with **OLAP and OLTP** systems
- Proficient in Python with knowledge of Data Structures, list comprehensions, file handling, functions, decorators, generators, regular expressions
- Experienced in using Python Functions like lambda and User-defined functions
- Skilled in writing **Complex SQL queries**, including **Joins**, **Window functions**, Sub Queries, Correlated Sub Queries, and Regular Expressions.

TECHNICAL SKILLS

Amazon Cloud Services
- : S3, EC2, Glue, Redshift, EMR, Athena, CloudWatch, Lambda, RDS, DynamoDB, SNS, IAM

Scheduling Tools
- : Airflow

Version Control Software
- : Git, GitHub

Programming Languages
- : SQL, Python, Pyspark

Hadoop Ecosystem
- : Apache Hive, Apache Spark, HDFS, Map Reduce

Databases
- : Amazon RDS, Oracle SQL, MySQL

Data Warehouse
- : Redshift, Snowflake, Hive

IDE Tools & Software's
- : PyCharm, VSCode, Jupyter Notebook, Thonny

OS
- : Windows, Linux

PROFESSIONAL EXPERIENCE

06/2019 – present **AWS Data Engineer** Pune, India
ALLSAFE IT SERVICES PVT LTD

PROJECTS

2022 – present **Health Care & Pharma Domain**
Client : Novexpharma
Description

Novexpharma is a South African pharmaceutical company created in 2007 Based in Cape Town, The firm provides equipment and systems within healthcare and Works on Drug Discovery and Development.

Role :AWS Data Engineer

Responsibilities :

- Gathering business requirements and conducting source data analysis.
- Employing AWS S3 to efficiently store and oversee substantial data volumes.
- Creating and automating data transformation and ETL processes through AWS Glue.
- Adapting existing business logic, if necessary, post-validation of data between source and target.
- Utilizing Amazon EMR to test and develop, capitalizing on its scalable and managed capabilities for big data processing.
- Utilizing PySpark code extensively for performing transformations within ETL scripts.
- Architecting and implementing automated solutions for common test cases (such as null checks, duplicate checks, filling missing values, and schema validation) using PySpark.
- Managing project tasks through JIRA and overseeing source code with GIT.
- Leveraging Amazon Redshift for data storage and analysis purposes.
- Orchestrating and developing Dags in Airflow for streamlined data pipeline management and automation.
- Applying Python data structures (strings, lists, tuples, sets, and dictionaries) for efficient data organization and manipulation.
- Implementing Python exception handling to effectively manage errors and exceptional scenarios in code.
- Crafting data pipelines: Extracting data from diverse sources, transforming and refining it, and loading it into the data warehouse.

03/2021 – 01/2022 **Retail Domain**
Client: Ackermans
Description

Ackermans is a South African chain of clothing retail stores. Founded in 1916 in Wynberg, Cape Town, Ackermans has over 700 stores across Southern Africa, including in Namibia, Botswana, Lesotho, eSwatini and Zambia, and is headquartered in Kuilsrivier near Cape Town.

Role :Data Engineer

Responsibilities :

- Developed and upheld Python code: Crafted clean, efficient, and meticulously documented code to realize project necessities.
- Engaged in the analysis of project requirements.
- Enhanced existing PySpark code for enhanced generality across the platform.
- Made use of AWS S3 to store and oversee extensive data volumes.
- Constructed SQL queries for extracting data from databases.
- Assumed responsibility for pulling, pushing, and committing code via GitHub.

- Participated in knowledge-sharing endeavors with the team.
- Arranged extracted data within the Oracle database to facilitate subsequent querying and analysis.
- Participated in daily meetings.

06/2019 – 01/2021

Airline Domain
Client: Air Alps
Description

Air Alps was an independent Austrian regional airline based in Innsbruck Austria. The carrier operated a domestic network in Italy in close cooperation with Alitalia. In addition the airline also operated charter and ad-hoc charter services.

Role: Data Engineer

Responsibilities :

- Engagement in analyzing project requirements.
- Enhanced existing PySpark code to achieve greater generality across the entire platform.
- Employed AWS S3 for the storage and management of extensive data volumes.
- Leveraged Athena to query and analyze data stored within Amazon S3 using standard SQL queries.
- Formulated SQL queries to extract data from the database.
- Devised and implemented scripts for common test cases (such as null checks, duplicate checks, filling missing values, and schema validation) utilizing PySpark.
- Scripted and optimized procedures to efficiently transform data on the AWS EMR cluster.
- Python Functions: Created reusable code blocks for specific tasks (including lambda, user-defined functions, and recursion).
- Fashioned Python utilities with PySpark to establish connections between source and target systems for the data processing pipeline.
- Assumed responsibility for pulling, pushing, and committing code through GitHub.

EDUCATION

Bachelors Mumbai, India
University Of Mumbai

CERTIFICATES

AWS Cloud Quest: Cloud Practitioner
Issued by Amazon Web Services Training and Certification