Abdullah Durrani

m.abdullahdurrani7@gmail.com LinkedIn

WORK EXPERIENCE

Data Engineer *Mar-2022 to Present*

[Blutech Consulting]

- Developed and maintained ETL pipeline using Databricks, Azure Data Lake, Pyspark, and Azure Data Factory to efficiently extract, transform data into a Lakehouse Architecture. Implemented SCD2 to handle slowly changing dimensions and handled incremental data for efficient data processing.
- Collaborated on a Cloudera-based project using Pyspark, Oozie, and HDFS to identify and resolve gaps in the ETL pipeline.
- Developed a dynamic data ingestion logic that enabled seamless ingestion of data into tables, reducing manual effort and increasing efficiency.
- Implemented a resumption logic that enabled scripts to resume processing from the point of failure, reducing manual intervention.
- Designed and implemented custom dependency logic that enabled scripts to execute only when all dependencies were met, reducing the risk of errors or data loss.
- Implemented an audit logging mechanism for ETL pipeline to provide visibility into data processing activities and support troubleshooting efforts. Investigated and resolved issues related to data discrepancies in production dashboards, identifying root causes and proposing effective solutions.
- Worked on a project using AWS Glue, S3, and Apache Hudi to develop a data transformation script that automated the transformation of all files in an S3 location. Developed a logic to handle the change data capture (CDC) of data.

Tools and Technology

- **Data processing and transformation**: Apache Spark, PySpark, AWS Glue, Databricks, Azure Data Factory, Apache Airflow.
- **Data storage and management**: AWS S3, Azure Data Lake, AWS Redshift, Google BigQuery, PostgreSQL.
- **Cloud platforms:** Amazon Web Services (AWS), Microsoft Azure, Google Cloud Platform (GCP).
- **Programming Languages:** Python, SQL, C#.
- **Version Control:** Git, GitHub.
- **Other tools:** Jupyter Notebook, Docker, Apache Oozie.

EDUCATION

National University of Modern Languages *Islamabad*

Bachelor Of Computer Science *2019-2023*

PROJECTS

- **Reddit ETL pipeline**

- Designed and developed an ETL pipeline to derive insights from Reddit, utilizing its API as a primary data source.
- Created a data lake using MinIO.
- Utilized Apache Spark to transform and cleanse nested JSON data, creating high-quality, normalized data that was subsequently stored as Parquet files in the MinIO bucket.
- Created a robust data warehousing solution by loading the cleaned data into Postgres and implementing dimensional modeling to optimize data querying and analysis.
  Orchestrated the entire ETL workflow using Apache Airflow, ensuring a smooth and efficient data processing experience.

CERTIFICATIONS

- Databricks Certified Associate Developer for Apache Spark
- Databricks Certified Data Engineer Associate
- Databricks Accredited Lakehouse Platform Fundamentals
- AWS Partner Accreditation (Technical)
- MTA Python Certified