

**TEK09930** (Cloud - AWS - 5 + Year) MSc

Highlights

Data Engineer

I have total more than 5+ years of experience, and my areas of expertise include SQL, Python, Hadoop, Sqoop, Hive, and Spark/PySpark. I am skilled in data sharing and reporting because my areas of expertise are data transformation, data ingestion, and ETL/ELT operations. I have a solid foundation in data processing, agile methodologies, and providing reliable solutions through debugging. I am able to develop scalable ETL/ELT processes because of my in-depth knowledge of data structures including the snowflake schema, star schema, and medallion architecture. I guarantee easy data processing, integration, and management in the Big Data landscape by being knowledgeable about Azure services like Azure Data Lake Storage, Azure Databricks, Delta Lake, Azure Synapse Analytics, Azure Functions and like AWS EC2, EMR, RDS, Redshift, Glue, S3, Lambda, Athena as well as version control systems like Git, GitHub, and Bitbucket. My ability to give complete data engineering solutions is further enhanced by my experience with Databricks Notebooks, PySpark for distributed data processing, and Azure DevOps for CI/CD pipelines.

Skills

Primary Skills

- AWS Glue
- Redshift
- PySpark
- Snowflake
- Python

Secondary Skills

- AWS ECS
- AWS Lambda
- SQL

Other Skills

SKILLS

- Azure Data Factory
- Azure Databricks
- Azure Blob Storage
- ADLS Gen2
- Azure DevOps
- Azure Synapse
- Databricks Notebook
- Delta Lake
- Data Transformation
- ETL / ELT
- Data Warehouse
- Spark / PySpark
- AWS EC2
- AWS Lambda
- AWS EMR
- AWS S3
- AWS Glue
- Amazon Athena
- Python
- MYSQL
- Oracle
- Hadoop
- Hive
- Git / GitHub / Bitbucket
- Jira
- Agile Methodology
- Scrum

Projects

Project 1. - IT Industry

Roles & Responsibilities:

- Experience as an Assistant Professor of Physics.
- IT experience, specializing in Azure Data Factory, Databricks, Python, SQL, and Azure Synapse Analytics for ETL processes.
- Orchestrated ETL processes and transportation projects, ensuring seamless data extraction, transformation, and loading using Azure Databricks and Medallion architecture.
- Spearheaded the development of efficient ETL pipelines on Azure cloud platforms, ensuring data accuracy and reliability.
- Adept in data extraction logic with Azure Data Factory and Databricks SQL, seamlessly integrating data from various sources, including RDBMS.
- Proficient in Azure Databricks architecture, optimizing and processing large datasets from diverse sources, and data warehousing.
- Specialized in Spark integration with Azure for seamless data operations also Optimized Spark workflows for better performance and efficiency.
- Proficient in MySQL and SQL for database management having strong command of joins, subqueries, and window functions.
- Strong background in data governance and ensuring data quality.
- Implemented CI/CD tools for continuous integration and deployment also Implemented CI/CD pipelines using Azure DevOps, Git, GitHub, and Bitbucket.
- Adhered to Agile methodologies and Scrum practices for project management.

Project 2. Health Insurance Data Analysis - IT Industry (22 months)

Role: Data Engineer

Responsibilities -

- Analyzed healthcare data for client revenue growth using on-premises SQL server.
- Used Azure Data Lake Gen2 for data storage and transformation.
- Created and managed 50+ tables connected via a mount point.
- Designed control table in Azure SQL Server to track data ingestion.
- Implemented For-Each activity pipeline with Copy, Notebook, and Stored Procedure activities.
- Followed Medallion Architecture: Raw, Bronze, Silver, Gold zones.
- Loaded and cleansed raw data before moving to bronze zone.
- Applied business logic and aggregations in silver zone.
- Loaded final data into gold zone for client analysis.

Project 3. Payroll And Salary Data Analysis - IT Industry (17 months)

Role: Data Engineer

Responsibilities -

- Led secure data transfer of payroll and salary data to AWS S3 using AWS DMS.
- Developed AWS Lambda functions and CLI scripts for importing CSV, TSV, Parquet, and ORC data formats.
- Managed ingestion of diverse file types into AWS Redshift Data Warehouse.
- Optimized Redshift tables with Partitioning and Clustering for enhanced query performance.
- Utilized AWS Glue for scalable data processing and PySpark scripts for reliability.
- Conducted thorough validation to ensure data integrity.
- Demonstrated proficiency in AWS ecosystem tools like DMS, Lambda, Redshift, Glue, and PySpark.
- Delivered enhanced data analysis platform for valuable client insights.