

[Name]

[Email] | [Contact] | [LinkedIn] | [Location]

About Me:

I am a Data Engineer with 3 years of experience specializing in Azure cloud-based data solutions. My expertise in Apache Spark and Databricks allows me to design, develop, and optimize ETL pipelines and data warehouses effectively. I excel in SQL and Python, delivering reliable data solutions that enhance operational efficiency. I am dedicated to leveraging my skills for impactful data-driven decisions

Work Experience

Data Engineer, XXX – Pune, MH Oct 2022 – Present

- Designed, developed, and optimized ETL pipelines using Azure Databricks and Apache Spark (PySpark) for large-scale data processing, improving data processing throughput by 35% and reducing job completion time by 40%.
- Implemented Delta Lake for reliable and performant data lake storage, ensuring data integrity, enabling data versioning, and improving query performance by 25%.
- Orchestrated data workflows using Azure Data Factory, automating data ingestion, transformation, and loading processes, resulting in a 40% reduction in manual intervention and a 15% decrease in operational costs.
- Designed and implemented data warehousing solutions using Azure Synapse Analytics, optimizing data models and query performance, and decreasing query latency by 30%.
- Managed and optimized Azure Blob Storage and Azure Data Lake Storage Gen2 for efficient data storage and retrieval, scaling storage solutions to accommodate a 50% increase in data volume.

Data Analyst, XXX – Pune, MH Apr 2022 – Oct 2022

- Assisted in the development of data pipelines using Azure Data Factory and SQL, contributing to a 15% increase in data pipeline efficiency and a 10% reduction in data errors.
- Performed data extraction, transformation, and loading (ETL) operations from various data sources into Azure SQL Database, processing over 500GB of data monthly and ensuring data accuracy and consistency.
- Developed and maintained SQL queries for data analysis and reporting, improving report generation efficiency by 10% and providing data-driven insights to stakeholders.
- Collaborated with senior engineers on data migration projects, including migrating data to Azure Blob storage.

Projects

Spark ETL Pipeline

- Developed a real-time data streaming pipeline using Spark Streaming on Azure Databricks, processing and analyzing customer clickstream data. Reduced data latency by 30% and enabled near real-time insights for marketing teams, leading to a 10% improvement in targeted marketing campaign effectiveness.

Lakehouse Governance

- Contributed to the migration of on-premises data to Azure Data Lake Storage Gen2 and Azure SQL Database. Successfully migrated 1TB of data with minimal downtime (under 3 hours) and ensured data integrity during the migration process.

Skills

Programming Languages: Python (Pandas, NumPy, Scikit-learn), SQL.

Platforms: Azure (Azure Databricks, Azure Data Factory, Azure Synapse Analytics, Azure Blob Storage, Azure SQL Database, Azure Data Lake Storage Gen2), Git, Linux.

Big Data Technologies: Apache Spark (PySpark, Spark SQL, Spark Streaming, Spark Core), Databricks (Delta Lake, Databricks SQL, Databricks Runtime), Hadoop (Basic).

Education

SPPU– MS in Mathematics & Statistics 2022-2024