

SUJIT TIWARI

Senior Data Engineer | Azure | Databricks | Spark | Lakehouse Architect

Pune, India | +91-9975026110 | sujit.ai913@gmail.com

PROFESSIONAL SUMMARY

Highly accomplished Senior Data Engineer with 11+ years of experience designing, developing, and delivering enterprise-scale data platforms, cloud-native Lakehouse architectures, and high-performing ETL/ELT pipelines across Azure, Snowflake, Hadoop, and Databricks ecosystems. Strong expertise in MS SQL Server, Azure SQL, Azure Data Factory (ADF), Azure Databricks, Delta Lake, Unity Catalog, Apache Spark, DBT Cloud, and Apache Airflow. Extensive experience migrating legacy SQL Server and Hive-based systems to scalable Databricks Lakehouse architectures with strong governance, security, and performance optimization practices. Skilled in advanced SQL development, data warehousing, real-time streaming, distributed processing, data governance, and CI/CD implementation. Proven ability to lead cross-functional teams, deliver high-quality enterprise solutions under aggressive timelines, and collaborate effectively with business stakeholders and engineering teams.

TECHNICAL SKILLS

- Data Engineering & Lakehouse: Azure Databricks, Delta Lake, ADLS Gen2, Azure Data Factory, Azure Synapse Analytics, Microsoft Fabric, Snowflake, DBT Cloud, Apache Airflow
- Databases: MS SQL Server, Azure SQL Database, PostgreSQL, MySQL, Cosmos DB, Hive, HBase, Teradata
- Big Data Technologies: Apache Spark, PySpark, Hadoop, Kafka, Flink, Kafka Connect
- Data Warehousing: Data Marts, Star Schema, Snowflake Schema, Slowly Changing Dimensions (SCD), Enterprise Reporting
- Programming Languages: Python, SQL, Shell Scripting
- Governance & Security: Unity Catalog, Azure Purview, RBAC, Data Masking, Lineage Tracking, Compliance Monitoring
- Real-Time Streaming: Kafka, Event-driven ingestion, Streaming ETL
- DevOps & CI/CD: Git, Jenkins, Docker, Bitbucket, Jira, Confluence
- APIs & Integration: REST APIs, JSON, Web Services
- Monitoring & Reporting: Grafana, Datadog, Power BI, QuickSight

PROFESSIONAL EXPERIENCE

Assistant Consultant | TCS | Sep 2025 – Present

- Leading enterprise-wide cloud data modernization initiatives using Azure Databricks, Delta Lake, and Azure-native services.
- Architecting scalable Lakehouse solutions for enterprise analytics and reporting.
- Designing reusable ingestion frameworks and orchestration pipelines using Airflow and ADF. Driving governance implementation using Unity Catalog and Azure Purview for lineage, RBAC, and compliance monitoring.
- Collaborating with stakeholders to define technical architecture, implementation best practices, and solution roadmaps.
- Participating in rapid prototyping and proof-of-concept initiatives for new data platform capabilities.

Data Engineer Consultant | GlobalLogic | Sep 2023 – Sep 2025

Project: Insurance Financial Data Processing Platform

Client Domain: Insurance & Finance

Tech Stack: Microsoft Fabric, Snowflake, Spark, Python, Airflow

Responsibilities:

- Designed and implemented enterprise-grade ingestion and processing platform for policy updates, claims processing, premium collections, settlements, and payouts.
- Built scalable orchestration pipelines using Apache Airflow handling ~600k financial events/day.
- Developed real-time streaming ingestion solutions integrating Microsoft Fabric and Snowflake.
- Implemented advanced SQL transformations, data quality validation frameworks, and reconciliation logic.
- Designed optimized data models for enterprise reporting and downstream analytics.
- Performed advanced SQL performance tuning, partitioning, indexing, and Spark optimization.
- Collaborated with finance and actuarial teams to define business requirements and functional specifications.
- Ensured data governance, auditability, and compliance standards across ingestion pipelines.

Project: Retail Store Benchmarking & ML Analytics Platform

Client Domain: Retail & Supply Chain

Tech Stack: PySpark, Azure Synapse, Cosmos DB, Power BI, SQL

Responsibilities:

- Architected and developed an enterprise retail analytics platform aggregating sales, inventory, distribution, and customer datasets.
- Designed scalable ETL/ELT pipelines using PySpark and Azure Synapse for processing high-volume retail datasets.
- Developed data marts using Star and Snowflake schemas supporting BI reporting and executive dashboards.
- Implemented ML-driven benchmarking models for inventory optimization, sales forecasting, and store performance analysis.
- Built low-latency data access patterns using Cosmos DB for near real-time analytics.
- Optimized Spark workloads reducing execution time and improving processing efficiency.
- Created REST-based ingestion integrations and JSON data transformations.

Project: Common Data Platform (CDP) – Omnichannel Customer Analytics

Tech Stack: Databricks, Delta Lake, PySpark, Apache Superset

Responsibilities:

- Developed an omnichannel customer analytics platform using Databricks Lakehouse architecture.
- Implemented Delta Lake for scalable, ACID-compliant storage and reliable downstream reporting.
- Designed customer 360-degree data models integrating multi-channel retail datasets.
- Built interactive dashboards and analytical reporting solutions using Apache Superset.
- Participated in rapid POC development and solution architecture discussions.

Data Engineer | IBM | Aug 2022 – Sep 2023

Project: Enterprise Hive to Azure Databricks Migration

Client Domain: Enterprise Analytics

Tech Stack: Hive, ADLS Gen2, Azure Databricks, SQL Server, ADF, PySpark

Responsibilities:

- Led migration of legacy Hive and SQL Server workloads to Azure-native Databricks Lakehouse architecture.
- Designed scalable ingestion frameworks using Azure Data Factory and Databricks notebooks.
- Implemented Delta Lake architecture with Bronze, Silver, and Gold layers.
- Developed PySpark transformations processing structured, semi-structured, and streaming datasets.
- Designed and optimized SQL Server stored procedures, views, triggers, DDL, and DML operations.
- Configured Cosmos DB Cassandra API integration for low-latency real-time query processing.
- Performed advanced query tuning and Spark optimization reducing execution times significantly.
- Built monitoring and alerting frameworks for pipeline reliability and operational support.
- Collaborated with architecture and infrastructure teams to define enterprise migration strategies.

Technical Specialist – Data Engineering | Abzooba (UST Global) | Jul 2019 – Aug 2022

Project: PISCES – Enterprise Data Ingestion & ML Platform

Client Domain: Enterprise Analytics

Tech Stack: Python, PySpark, ADF, Databricks, Snowflake, MSSQL, Synapse

Responsibilities:

- Developed scalable ingestion pipelines using Spark/PySpark framework reducing processing time by 40%.
- Built enterprise ETL/ELT frameworks using Azure Data Factory and Databricks.
- Designed Azure SQL and Synapse-based warehousing solutions supporting enterprise analytics.
- Implemented advanced SQL transformations, stored procedures, views, and optimization logic.
- Tuned long-running Spark jobs improving performance and cluster utilization.
- Implemented automated monitoring, failure analysis, and remediation workflows.
- Conducted unit testing and validation for production-grade enterprise pipelines.
- Managed and monitored scheduled batch workloads ensuring SLA compliance.

Project: Price Parity & Retail Analytics Platform

Client Domain: Retail Analytics

Tech Stack: SQL Server, Hive, SSIS, Python

Responsibilities:

- Designed and implemented batch ingestion pipelines for large-scale retail pricing datasets.
- Developed advanced SQL Server stored procedures, views, and transformation logic.
- Created external tables in Azure SQL Data Warehouse supporting downstream reporting.
- Processed structured, semi-structured, and unstructured retail datasets.
- Conducted SQL query optimization and ETL performance tuning.
- Collaborated with reporting teams to support executive dashboards and analytics use cases.

Project: Real-Time Streaming Reporting Platform

Tech Stack: Kafka, Flink, Python, Airflow, AWS Glue

Responsibilities:

- Developed scalable real-time streaming solution for IoT and operational reporting data.
- Implemented distributed event processing pipelines using Kafka and Flink.
- Built orchestration workflows using Airflow and AWS Glue.
- Enabled near real-time analytics and monitoring capabilities for enterprise reporting.

Associate | Cognizant | Jan 2018 – Nov 2018

Project: IFRS9 Basel-III Risk Data Platform

Client Domain: Banking & Financial Services

Tech Stack: SAS, PySpark, Hadoop, Teradata

Responsibilities:

- Migrated legacy SAS-based credit risk and Basel frameworks to scalable Spark ecosystem.
- Built distributed processing solutions using Hadoop and Spark.
- Reduced model execution runtime by over 60% using Spark parallelization.
- Implemented lineage tracking and audit logging supporting regulatory compliance.
- Re-engineered risk calculation logic improving scalability and maintainability.

Software Engineer | Infosys Ltd | Dec 2014 – Dec 2017

Project: Enterprise Pipeline Optimization & Hadoop Modernization

Tech Stack: Hive, Spark, HDFS, Impala, SQL, Sqoop

Responsibilities:

- Developed scalable Spark workflows for enterprise-scale data processing and analytics.
- Designed Hive tables, HQL queries, and optimized HDFS-based processing pipelines.
- Imported and exported large datasets using Sqoop.
- Improved data processing speed by over 35% through Spark optimization.
- Supported enterprise reporting and analytics use cases using the Hadoop ecosystem.

EDUCATION

Bachelor of Computer Science & Engineering – Rajiv Gandhi Technical University, Bhopal

CERTIFICATIONS

Google Certified Data Engineer

Data Bricks Certified Data Engineer Associate

Data Bricks Certified Data Analyst

Data Bricks Certified Data Engineer Professional