Sri Omkar D

sri.omkar.d@gmail.com | **+1 951-545-2146** | **linkedin.com/in/sri-omkar-58r4r4r8**

Professional Summary

- Data Engineer with **7+ years of experience** building and supporting batch and near-real-time data pipelines across **AWS and Azure**, using **Python, SQL, PySpark, Apache Spark, Databricks, AWS Glue, Airflow, S3, ADLS Gen2, Redshift, Snowflake, Teradata, and SQL Server**.
- Hands-on experience supporting **equipment rental and fleet operations data pipelines**, working with data from **RentalMan, AS400, Oracle OLTP, Teradata, and internal equipment monitoring applications** for rental operations, fleet utilization, billing, and business reporting.
- Experienced with **Databricks, PySpark, Spark SQL, AWS S3, AWS Kinesis, and Databricks Streaming** for processing structured, semi-structured, and near-real-time operational data.
- Built and supported **Azure-based data engineering pipelines** for retail supply chain and client ERP datasets using **ADLS Gen2, Azure Databricks, PySpark, FastAPI, SQLAlchemy, Snowflake, Exasol, Terraform, Docker, Azure Key Vault, Azure Monitor, and GitHub Actions**.
- Worked on **AWS data lake and banking compliance pipelines** involving legacy **Ab Initio** workflows, PySpark transformations, AWS Glue Data Catalog, Airflow orchestration, Redshift reporting, Git, and Jenkins.
- Strong background in traditional ETL and analytics workflows using **SQL Server, Sqoop, AWS EMR, HiveQL, AWS Glue, PySpark, Amazon Redshift, Athena, Tableau, and Power BI** for sales, subscription, revenue, and customer usage reporting.
- Skilled in **SQL development, data validation, source-to-target reconciliation, pipeline troubleshooting, job monitoring, dimensional modeling, query tuning, partitioning, joins, aggregations, window functions, and performance optimization** across cloud and legacy data platforms.
- Comfortable working with **business users, analysts, data teams, and project stakeholders** to gather requirements, support production data loads, document workflows, and deliver reporting-ready datasets for **Qlik Sense, Tableau, and Power BI**.

Technical Skills

- **Cloud Platforms:** AWS, Microsoft Azure
- **AWS Services:** S3, EMR, Athena, Kinesis, AWS Glue, Glue Data Catalog, IAM, CloudWatch
- **Azure Services:** ADLS Gen2, Azure Databricks, Azure Key Vault, Azure Monitor, Log Analytics, Azure RBAC
- **Big Data & Processing:** Apache Spark, PySpark, Spark SQL, Databricks, Databricks Streaming, HiveQL, Sqoop
- **Programming & Scripting:** Python, SQL
- **Data Warehousing & Databases:** Amazon Redshift, Snowflake, Exasol, Teradata, Microsoft SQL Server, Oracle OLTP, AS400
- **Orchestration & DevOps:** Apache Airflow, SQL Server Agent, Docker, Terraform, Git, GitHub Actions, Jenkins, CI/CD
- **Data Modeling & Data Quality:** Star Schema, Snowflake Schema, Fact/Dimension Tables, SCD Type 2, Materialized Views, CDC, Data Masking, RBAC, Source-to-Target Reconciliation, Data Validation
- **Python Libraries & Frameworks:** FastAPI, SQLAlchemy, Pandas, NumPy, SciPy, Matplotlib, Scikit-learn

- **BI & Reporting:** Qlik Sense, Tableau, Power BI
- **Tools & Collaboration:** Jira, Confluence, Agile/Scrum
- **Spark Optimization**: Spark UI, AQE, Broadcast Joins, Data Salting, Partition Tuning, Spark Catalyst
- **File Formats**: Parquet, JSON, XML, CSV
- **Data Quality**: Reconciliation, Duplicate Checks, Row Count Validation, Variance Checks

Nortek Consulting INC (Herc Rentals Inc.) **Florida**
Data Engineer **Nov 2025 – Present**

- Developed **Python**-based ingestion workflows to route high-volume operational data from **RentalMan**, **AS400**, **Oracle OLTP**, and **Teradata-based systems** into an **AWS S3** data lake for rental, inventory, billing, and branch operations reporting.
- Built near-real-time equipment monitoring pipelines using **AWS Kinesis** and **Databricks Streaming** to process semi-structured **JSON/XML** telemetry payloads, including **GPS location**, **engine metrics**, **fuel usage**, and fleet activity for maintenance and utilization reporting.
- Created and maintained **PySpark** and **Spark SQL** pipelines in **Databricks** to clean, normalize, and join datasets across rental contracts, fleet inventory, billing, customer operations, and equipment usage.
- Reduced daily batch runtimes by approximately **25–30%** by tuning **PySpark** partition strategies, optimizing **Spark SQL** joins, and replacing unnecessary full-refresh cycles with scheduled incremental load patterns.
- Decreased pipeline rerun effort by approximately **30–40%** by breaking long **Databricks** transformation chains into modular, restartable stages with intermediate curated outputs for rental, billing, and fleet reporting loads.
- Added **data validation** and **source-to-target reconciliation** checks across pipeline stages, including record counts, duplicate checks, null handling, and source-to-target comparisons to improve data reliability before business reporting.
- Built SQL-based curated views and reporting layers using **Teradata-based datasets**, **fact/dimension modeling**, **materialized views**, and partition-aware **SQL** design to support **Qlik Sense** dashboards for fleet utilization, rental performance, and financial reporting.
- Orchestrated scheduled workflows using **Apache Airflow** and **SQL Server Agent**, including dependency handling, retry logic, monitoring, and troubleshooting for failed loads.

Quess Corp (Blue Yonder India Pvt Ltd.) **India**
Data Engineer **June 2022 - July 2023**

- Reduced batch transformation runtimes by approximately **25–30%** by consolidating iterative column casting and enrichment logic into unified **PySpark DataFrame** operations, reducing **Spark Catalyst** planning overhead.
- Optimized execution plans for large-scale supply chain datasets, including inventory, POS, orders, and product movement, by eliminating unnecessary repartitioning, tuning partition boundaries, and optimizing **Spark SQL** joins.
- Developed distributed **Azure Databricks** pipelines to clean, deduplicate, normalize, and join high-volume retail supply chain data for downstream retail optimization and analytics use cases.

- Built programmatic data ingestion frameworks using **Python**, **FastAPI**, and **SQLAlchemy** to validate, standardize, and route client ERP datasets from APIs, flat files, databases, and source systems into **Azure Data Lake Storage Gen2 (ADLS Gen2)**.

- Maintained secure and repeatable **Azure** data environments using **Terraform**, **Docker**, **Azure Key Vault**, and **Azure RBAC**, supporting controlled access, credential management, and reduced unnecessary compute usage.

- Implemented pipeline monitoring and logging using **Azure Monitor** and **Log Analytics** to track job health, runtime performance, failed loads, and operational issues across ingestion and transformation workflows.

- Supported **SQL** tuning, query profiling, and data model changes within **Snowflake** and **Exasol**, improving reporting performance for supply chain analytics workloads.
- Managed **Apache Airflow** DAG dependencies and scheduled batch workflows, including retries, dependency handling, SLA monitoring, and alerting for **Databricks** processing jobs.

Trigent Software Pvt Ltd. (Accenture Solutions Pvt Ltd.) **India**
Data Engineer **Jan 2020 - June 2022**

- Reduced runtime by approximately **35–40%** on selected high-volume financial pipelines by using **Spark UI** to diagnose data skew and applying **data salting**, **broadcast joins**, **Adaptive Query Execution (AQE)**, and caching strategies.

- Optimized incremental load performance by replacing costly full-table merge patterns with partition-aware **anti-join** and append logic for insert-only datasets.

- Built and maintained distributed **PySpark** and **Spark SQL** transformation jobs to perform filtering, complex joins, aggregations, deduplication, enrichment, and **source-to-target validation** for regulatory compliance reporting.

- Supported modernization of legacy **Ab Initio** workflows by developing **Python**-based ingestion services and processing components, moving high-volume banking datasets into an **AWS S3** data lake organized across raw, cleansed, and curated layers.

- Registered and structured data lake assets in **AWS Glue Data Catalog**, enabling metadata discovery across **Spark**, **SQL**, and **Amazon Redshift** query workflows.

- Modeled and loaded audit-ready financial datasets into **Amazon Redshift**, creating optimized **SQL** reporting tables and materialized views to support **Financial Crime** and Business Intelligence teams.

- Performed source-to-target data quality validation, reconciliation, duplicate checks, and variance checks across large financial datasets to support compliance reporting accuracy and SLA expectations.

- Managed, monitored, and automated batch DAGs in **Apache Airflow**, implementing retry logic, log tracking, dependency management, and failure alerting for compliance loads.

- Maintained CI/CD deployment workflows using **Git** and **Jenkins**, supporting safe, version-controlled code migration across DEV, QA, and PROD environments.

Experis IT – (Thomson Reuters) **India**
ETL Developer **June 2016 - Sep 2019**

- Built and supported large-scale data extraction workflows from legacy **SQL Server** databases into an **AWS S3** data lake using **Apache Sqoop** on **AWS EMR**.

- Authored **HiveQL** scripts to process and transform large flat files and relational datasets within the Hadoop ecosystem for downstream analytical consumption.

- Reduced selected batch write delays by approximately **20–25%** by tuning **Parquet** output write patterns, partition strategies, and S3 load behavior for **AWS EMR** and **Spark** workloads.

- Developed **AWS Glue** and **PySpark** jobs to clean, aggregate, and summarize sales, subscription, and revenue metrics before loading curated datasets into **Amazon Redshift** and **AWS Athena** for downstream BI reporting.

- Wrote complex **SQL** queries using window functions, subqueries, aggregations, and multi-table joins to support ad-hoc analysis, reporting, data validation, and query tuning for business stakeholders.

- Supported end-to-end batch ingestion workflows to extract sales, subscription, customer usage, and revenue data from on-premise applications into a centralized analytics environment.

- Implemented daily **data validation** checks, including missing record detection, duplicate checks, **source-to-target row count validation**, and revenue variance reviews to improve dashboard accuracy and reporting reliability.

- Monitored pipeline execution, diagnosed failed data loads, and troubleshot ETL bottlenecks to support scheduled reporting timelines for sales and customer usage reporting.

Education:

Masters in Information Technology Management **2024**
California Baptist University**,** Riverside, CA

B. Tech in Computer Science and Engineering **2017**
JNTU-K University, India