Saran Kumar Durgam

Senior Data Engineer | Azure Databricks Certified

(940) 290-2722 sarandurgam9@gmail.com www.linkedin.com/in/saran-kumar-5795981b3

SUMMARY

Experienced **Senior Data Engineer** with 9+ years of industry experience designing and implementing scalable data solutions across multi-cloud environments (**AWS, Azure, GCP**). Proven track record in building robust ETL/ELT pipelines using tools like **Databricks, Airflow, Glue, Informatica**, and **ADF**. Deep expertise in big data ecosystems (**Spark, Hadoop**), modern data warehousing (**Snowflake, Redshift, BigQuery**), and advanced analytics. Experienced in designing scalable data pipelines and medallion architectures using **Azure Fabric**, **Delta Lake**, and **Dataflows**, supporting direct integration with **Power BI** and other enterprise BI platforms. Demonstrated excellence in data governance, statistical analysis (**regression, hypothesis testing**), and security compliance (HIPAA, GDPR), with a strong track record of collaboration in Agile environments. Strong programming skills in **Python, PySpark**, and **SQL**.

CERTIFICATIONS

Microsoft: Microsoft Azure Data Engineer

Databricks: Academy Accreditation - Databricks Fundamentals

Power BI: Power BI Essential Training

SKILLS

Languages: Python, SQL, PySpark, Scala, Shell Scripting
Big Data & Processing: Apache Spark (Core, SQL, Streaming), Hadoop (HDFS, MapReduce, Hive), Apache Flink, Apache Kafka (Kafka Streams, Kafka Connect), Apache Airflow, Apache Beam
ETL Tools: AWS Glue, Informatica PowerCenter, Azure Data Factory (ADF), Talend, DBT (Data Build Tool), StreamSets, SSIS.
Data Warehousing: Snowflake, Amazon Redshift, Google BigQuery, Teradata, Oracle, SQL Server, Delta Lake
Cloud Platforms: AWS (S3, Redshift, Lambda, EC2, RDS, IAM, CloudWatch, Bedrock, MWAA, Glue Catalog), **Azure** (ADF, Databricks, Azure Key Vault, ADLS Gen2, Azure ML, Azure DevOps), **GCP** (Pub/Sub, Dataflow, BigQuery)
Security & Governance: OAuth 2.0, Apache Ranger, AWS IAM, Unity Catalog (Databricks), AWS Glue Data Catalog, Microsoft Purview, Data encryption (at rest and in motion), HIPAA, GDPR, PII/PHI handling
Visualization: Tableau, Power BI, Excel (PivotTables, Macros), Matplotlib, Seaborn
DevOps & Automation: Jenkins, GitHub Actions, Git, Terraform, Docker, Control-M, Autosys, Azure DevOps Pipelines, CI/CD pipelines for AWS Lambda, Glue, and DBT deployments
Machine Learning: TensorFlow, Scikit-learn, Azure ML, MLflow, PySpark Mllib, AWS Bedrock for GenAI/NLP-based summarization and language generation, ML pipeline orchestration using Airflow and Azure ML
Statistical Tools: R, SPSS, SAS, NumPy, SciPy
Methodologies: Agile (Scrum), SDLC, Sprint Planning, UAT Support, Code Reviews, Documentation Best Practices.

EXPERIENCE

Client: The Cigna Group, Plano, TX **January 2024 – Present**

Role: Senior Data Engineer

Project Details:

Cigna's **Intelligent Healthcare Data Platform (IHDP)** was initiated to build a **cloud-native, scalable data infrastructure** for ingesting, processing, and analyzing high-volume healthcare data in compliance with **HIPAA** and **GDPR** regulations. The platform integrated structured and semi-structured data using **AWS Glue**, **Lambda**, and **Redshift**, while leveraging **Snowflake**, **StreamSets**, and **Databricks** for transformation and warehousing. The solution supported **real-time event-driven architectures** using **SQS**, **SNS**, and **API Gateway**, while **GenAI applications** were deployed using **AWS Bedrock** for natural language insights and summarization of patient notes. Key components included **CI/CD with Jenkins**, **containerization with Docker**, and **data governance frameworks** for lineage, validation, and auditability, enabling secure and efficient delivery of advanced healthcare analytics.

Roles and Responsibilities

- **Infrastructure Deployment:** Built and configured AWS components including EC2, RDS, VPC, CloudFormation, IAM, S3, Glacier, and CloudWatch to support secure, scalable cloud operations.

- **Data Integration Leadership:** Ingested and transformed HL7 and FHIR-based healthcare data using AWS Glue and Lambda, integrating with Redshift and Snowflake via schema normalization, PII masking, encryption, and compliance alignment under HIPAA.
- **StreamSets Integration:** Connected legacy systems and APIs to Snowflake through StreamSets pipelines, enabling schema evolution and high-throughput ingestion.
- **ETL & Data Pipeline Development:** Designed scalable ETL pipelines using AWS Glue, Redshift, and Snowflake integrated DBT for transformation modularity and CI/CD.
- **GenAI Implementation:** Integrated AWS Bedrock to deploy GenAI models for real-time summarization, document classification, and language generation tasks within internal tools.
- **API Development:** Built REST APIs using Python and API Gateway for secure real-time communication between microservices and healthcare platforms.
- **Event-Driven Architecture:** Implemented messaging pipelines using Amazon Lambda, enabling decoupled microservices communication and automated alerting.
- **Serverless Workflow Design:** Created Lambda + DynamoDB-based applications to support real-time healthcare data ingestion with minimal infrastructure overhead.
- **Security & Compliance**: Applied end-to-end encryption, masking, and identity policies (IAM, S3 policies) to ensure PII/PHI data protection under HIPAA/GDPR.
- **CI/CD Integration:** Designed Lambda CI/CD pipelines using Jenkins and GitHub Actions; embedded DBT model runs for automated, testable deployments.
- **Workflow Orchestration:** Designed Snowflake workflows leveraging Snowpipe, Streams, and Tasks for automated batch and micro-batch ingestion.
- **Data Governance:** Managed metadata and lineage using AWS Glue Data Catalog and Unity Catalog in Databricks to support governance and traceability.
- **Containerization & Deployment:** Developed Dockerized ETL applications and deployed them to Amazon ECS and EKS for consistent, scalable workload execution.
- **Performance Optimization:** Tuned Redshift and Snowflake queries, optimized PySpark and Glue job configurations, and reduced ETL execution times significantly.
- **Analytics & Visualization:** Delivered real-time reporting by integrating curated Snowflake views and Databricks outputs into Power BI, aligning with medallion-layered architecture and supporting stakeholder access to trusted Gold-layer datasets.
- **Monitoring & Optimization:** Set up CloudWatch alerts, logging dashboards, and performance tuning for data workflows and AWS services.
- **Collaboration:** Worked closely with cloud architects, DevOps, and data science teams to align solutions with business and technical requirements.

Client: Cisco Systems, San Jose, CA **January 2022 - July 2023**

Role: Senior Data Engineer

Project Details:

Cisco's **Cloud Data Acceleration Platform (CDAP)** was developed to modernize and unify data processing across various business units, including **sales performance**, **customer telemetry**, and **product lifecycle analytics**. The initiative focused on migrating legacy on-prem pipelines to a **cloud-native architecture** leveraging **AWS**, **Snowflake**, **Databricks**, and **DBT**. The platform enabled **near real-time ingestion**, **scalable Spark-based ETL/ELT workflows**, and **AI-enhanced analytics**. Key deliverables included implementing **CI/CD for DBT**, orchestrating pipelines with **Airflow**, and improving **query performance, lineage visibility**, and **data reliability** for downstream consumers.

Roles and Responsibilities

- **Data Pipeline Engineering:** Designed and built scalable batch and streaming data pipelines using DBT, PySpark, and AWS Glue, ingesting structured/unstructured data into Snowflake and Redshift.
- **Airflow Orchestration:** Developed DAGs using Apache Airflow to automate daily workflows, data quality checks, and change data capture jobs.
- **Cloud Architecture Optimization:** Tuned EMR and Spark cluster configurations for batch processing, leveraging AWS EC2, S3, and YARN for distributed compute efficiency.
- **API & Integration:** Developed REST APIs using AWS Lambda and API Gateway for real-time ingestion and integration with Hasura-powered GraphQL APIs.
- **Collaboration with DevOps:** Built automated CI/CD pipelines in Azure DevOps and GitHub Actions for DBT and ETL workflows. Added notifications, test validations, and rollback strategies using YAML configurations.
- Developed automated regression suites for ETL pipeline validation and data dashboard testing using Python and Playwright. Integrated test execution into CI/CD pipelines with email/slack reporting.
- **Data Visualization:** Provided enriched datasets to BI teams using Tableau, enabling dynamic dashboards and KPI tracking for business insights.
- **Source System Integration:** Pulled data from Teradata and Oracle using Sqoop and processed in Spark for ingestion into Snowflake.
- **Monitoring & Alerting:** Configured monitoring with AWS CloudWatch and custom Python logging modules to detect pipeline failures and performance degradation.

- **Data Modeling & Tuning:** Developed medallion-style data models (Bronze/Silver/Gold) in Snowflake to standardize transformation layers and feed Power BI and Tableau dashboards, improving reuse and governance across analytics teams.
- **Data Validations:** Developed internal Python applications for data validation, anomaly detection, and scheduled reporting tasks. Scheduled execution via CloudWatch and orchestrated results back to S3 and BI dashboards.
- **Security & Governance:** Managed access control using IAM roles and tags, ensuring data protection and regulatory alignment.
- **Team Collaboration:** Collaborated with cross-functional teams including product owners, QA engineers, and business analysts in Agile sprint cycles. Authored user stories, test scenarios, and acceptance criteria for analytics features.
- Participated in sprint planning, retrospectives, and UAT support to ensure data product delivery matched stakeholder expectations

Client: Axis Bank, Bengaluru, India **October 2020 – January 2022**

Role: Senior Data Analyst

Project Details:

Axis Bank initiated a data modernization initiative to unify and streamline its risk and compliance analytics across business units. The project focused on building a cloud-native Enterprise **Data Lake** using Microsoft Azure, Apache Spark, and **Snowflake** to support real-time data ingestion, regulatory reporting, and machine learning workflows. Legacy ETL systems were replaced with scalable, cloud-based pipelines using **Azure Data Factory (ADF)**, Microsoft Fabric, and Power BI. The solution enabled secure, governed, and performant data access, aligning with RBI guidelines and **GDPR** compliance standards. **Machine learning** capabilities were introduced for **fraud detection**, customer behavior modeling, and risk scoring across retail banking operations.

Roles and Responsibilities

- **Fabric Medallion Architecture**: Designed and implemented Microsoft Fabric pipelines following medallion architecture principles (Bronze, Silver, Gold) for modular, scalable analytics delivery.
- **Data Architecture:** Designed and implemented scalable data models using Star Schema and Snowflake to improve query efficiency and maintainability.
- **ETL Automation:** Developed and optimized end-to-end ETL workflows using Azure Data Factory (ADF), Apache Airflow, Alteryx Designer/Server, and SSIS to orchestrate data movement and transformation.
- **Big Data Processing:** Leveraged Apache Spark (Scala/Spark-SQL) on Azure Databricks and Yarn to analyze large volumes of structured and semi-structured data.
- **Streaming & Real-Time Analytics:** Configured Kafka Streams and Kafka Connect to enable real-time ingestion and event-driven processing across banking operations.
- **ADF & Dataflows**: Built and scheduled Azure Data Factory pipelines and Power BI Dataflows for ingestion, cleansing, and transformation of financial datasets across structured and semi-structured sources
- **Cloud Storage & Security:** Managed Azure Data Lake Storage for secure, scalable cloud storage. Implemented access control and secrets management using Azure Active Directory, Key Vault, and Apache Ranger.
- **Visualization & Insights:** Designed and built advanced Power BI dashboards with drill-through reports, slicers, KPIs, and scheduled refresh automation. Enabled business users to perform self-service analytics with row-level security integration.
- **Data Quality & Governance:** Ensured data lineage and traceability using DBT, enforced quality rules, and monitored system metrics with Grafana for proactive issue resolution.
- **Machine Learning Workflows:** Built ML pipelines using TensorFlow, Azure ML, and Kubeflow for predictive analytics, including fraud scoring and churn prediction.
- **Deployment Automation:** Used Azure DevOps and Terraform to implement CI/CD pipelines, automate infrastructure provisioning, and manage deployment versions.
- **Cross-Platform Integration:** Integrated diverse data sources including SQL Server, Teradata, Cassandra, and flat files. Standardized ingestion logic and data validation using Alteryx workflows.
- **Hive Migration:** Migrated Hive queries to Spark RDD transformations using Scala, boosting performance and reducing latency.
- **Collaboration & Governance:** Streamlined workflows via ServiceNow integration, enabled team collaboration through GitHub, and ensured audit-readiness through stringent governance protocols.
- **Optimization & Cost Efficiency:** Tuned Spark jobs, optimized SQL queries, and analyzed infrastructure resource usage for cost-effective deployment.

Client: Landmark Group, Bengaluru, India **August 2019 – August 2020**

Role: Data Engineer

Project Details:

The **Enterprise Retail Intelligence Hub (ERIH)** was designed to centralize and modernize retail data management across Landmark Group's in-store and e-commerce operations. Built on **Google Cloud Platform (GCP)**, the platform leveraged **Apache Spark**, **Google Dataflow**, **Pub/Sub**, and **Snowflake** to support **real-time data ingestion**, **streaming analytics**, and **machine learning-driven insights**. The system integrated

diverse data sources, including sales transactions, customer behavior, and inventory movement, enabling use cases like **dynamic pricing**, **demand forecasting**, and **promotion optimization**. The solution replaced legacy ETL pipelines with **cloud-native workflows**, improved data governance, and provided **interactive dashboards** through **Tableau** and other BI tools. It also enforced **OAuth 2.0** security and aligned with internal audit and compliance requirements.

Roles and Responsibilities

- **Developed end-to-end ETL pipelines** using **Informatica** and **GCP Dataflow**, automating ingestion and transformation of transactional and user behavior data.
- **Built real-time data ingestion frameworks** using **Google Pub/Sub** and **Apache Flink** to support **streaming analytics** for online and point-of-sale data.
- Designed and modeled **Snowflake data warehouses**, enabling high-performance reporting for **merchandising**, **sales optimization**, and **executive dashboards.**
- Created **batch-processing pipelines** using **Apache Spark** and **PySpark**, optimizing data processing for product catalogs, customer segments, and clickstream logs.
- Integrated **TensorFlow** and **Scikit-learn** into workflows to support **predictive analytics** including **product recommendation engines**, **customer churn models**, and **promotion response prediction.**
- Designed **interactive Tableau dashboards** for **executive-level KPIs**, campaign effectiveness, and customer value analysis.
- Automated data ingestion from external APIs using **Python** and **AWS Lambda**, storing results into **DynamoDB/RDS**. Integrated with REST endpoints and scheduled triggers for near real-time data sync.
- Monitored platform health using **New Relic**, identifying performance bottlenecks and enabling **proactive alerting** and **failure recovery.**
- Implemented **OAuth 2.0** for securing APIs and enforced **data masking and encryption** for **PCI-DSS and GDPR compliance.**
- Used **Jenkins** for building **CI/CD pipelines**, ensuring continuous integration and deployment of analytics applications and ETL components.
- Managed **MongoDB** databases for storing **semi-structured product metadata** and **customer preferences**, allowing rapid iteration and high availability.
- Enforced **data governance policies** across ingestion, transformation, and reporting layers, including **lineage tracking**, **quality validation**, and **role-based access.**
- Collaborated cross-functionally with **marketing, merchandising, finance, and IT teams** to align analytics delivery with business priorities and operational KPIs.

Client: Citibank, Bengaluru, India **January 2016 – July 2019**

Role: **Data Engineer**

Project Details:

The project involved building an enterprise-grade data pipeline and analytics platform to support fraud detection, customer behavior analysis, and regulatory compliance. Technologies such as **Informatica**, **Teradata**, **Hadoop**, and **Power BI** were used to develop scalable ETL workflows, ensure data quality, and deliver operational dashboards to compliance and risk teams.

Roles and Responsibilities

- Developed and maintained **Informatica ETL pipelines** to extract, transform, and load large-scale banking data into **Teradata** for analytics and reporting.
- Designed complex **Informatica mappings and workflows**, applying transformation logic for **balance adjustments**, **transaction categorization**, and **customer segmentation.**
- Wrote advanced **Teradata SQL** and **PL/SQL scripts** utilizing **joins**, **aggregates**, **window functions**, and **set-based logic** for data staging and enrichment.
- Implemented **Slowly Changing Dimensions (SCD) Type 1 and 2** logic in Teradata for maintaining historical accuracy in dimension tables.
- Improved ETL performance by applying **partitioning**, **indexing**, and **collect stats** across staging and reporting layers.
- Created and deployed high-performance **BTEQ**, **FastLoad**, and **MultiLoad** jobs for bulk and incremental updates.
- Designed **Star and Snowflake schema** models supporting account summaries, customer risk scores, and profiling.
- Built **PySpark scripts** to ingest raw transaction files into **Hadoop**, transforming and loading data into **Hive** curated zones.
- Optimized **Hive tables** with **partitioning and indexing** to reduce query latency and improve analytics performance.
- Conducted **data quality checks**, **validation scripts**, and **lineage tracking** to ensure governance and data accuracy.
- Created dashboards in **Power BI** and **Excel** using curated datasets from **Teradata** and **Redshift** for compliance and audit teams.
- Automated ETL execution using **shell scripts** for job orchestration, pre/post load processing, and health monitoring in **UNIX.**
- Tuned **Apache Spark jobs** for efficient distributed data processing, reducing execution time significantly.

- Integrated **Informatica workflows with Control-M** and Unix scripts to handle job scheduling and automated recovery.
- Participated in **UAT cycles** with QA and business teams, resolving discrepancies in transformed data and ensuring reporting accuracy.