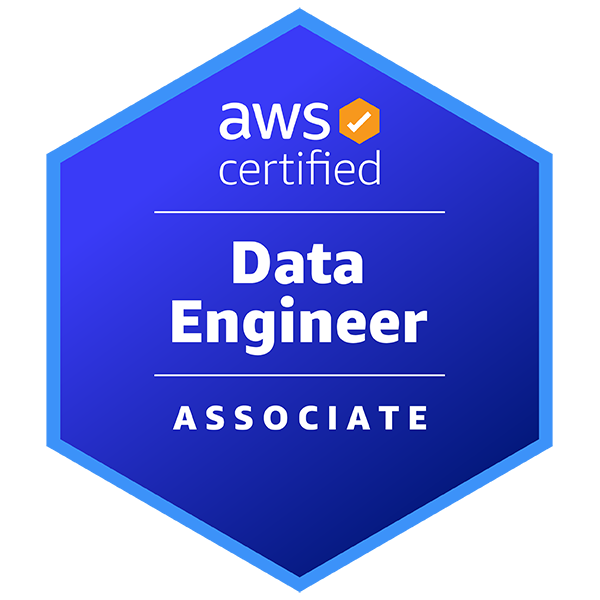
**Sai Kiran**

**Phone:** (717) 631-1265

**Email:** [saikiranb829@gmail.com](mailto:saikiranb829@gmail.com)

**Data Engineer**

**GitHub URL:** [**https://github.com/SaiKiranB369**](https://github.com/SaiKiranB369)

**SUMMARY:**

* **AWS-Focused Data Engineer & Analytics Consultant** with 5+ years of experience delivering end-to-end data solutions across telecom, finance, and enterprise domains, specializing in **AWS Cloud** and **Big Data ecosystems**.
* Proven expertise in **on-prem to AWS migration**, designing secure data lakes, and modernizing legacy Hadoop/ETL pipelines to **S3, EMR, Glue, and Redshift** for scalable, cost-optimized analytics.
* Skilled in **orchestrating ETL/ELT workflows** using AWS Step Functions, Airflow, and Lambda, enabling automated, SLA-driven pipelines for batch and near-real-time analytics.
* Hands-on experience building **interactive dashboards and KPIs** using QuickSight, Tableau, and Power BI, integrating structured, semi-structured, and streaming datasets for operational and executive reporting.
* Strong background in **data modeling** (star/snowflake schemas), **SQL tuning** (Redshift, Athena), and **Spark optimization** for large-scale distributed processing and analytics-ready datasets.
* Implemented **data governance and security frameworks** with Glue Data Catalog, Lake Formation, and IAM/KMS, ensuring compliance with financial regulations and telecom security requirements.
* Delivered **real-time analytics pipelines** with Kinesis, Firehose, Lambda, and EMR for anomaly detection, fraud monitoring, and operational intelligence.
* Adept at collaborating with stakeholders to gather requirements, design AWS architectures, present cloud adoption roadmaps, and deliver **executive-level insights** to accelerate business decision-making.
* Partnered with **data science** and **BI teams** to deliver advanced statistical analysis, cohort segmentation, and predictive trend insights using Redshift, Athena, and PySpark — enabling data-driven business strategies.
* Collaborated cross-functional teams on **DevOps** principles (CI/CD with CodePipeline, Infrastructure as Code with Terraform/CloudFormation), driving a culture of reusability, reliability, and operational excellence.
* Experienced in **Agile delivery, migration planning, and cost optimization** using CloudWatch and Budgets to ensure predictable spend and cloud governance.
* Core strengths: **AWS Data Lake Architecture, Big Data Engineering, On-Prem to Cloud Migration, ETL & Pipeline Automation, Data Modeling & Governance, Real-Time Analytics, Executive Presentations.**

**TECHNICAL SKILLS:**

| **Programming Languages** | Python, SQL, HiveQL, PySpark, R |
| --- | --- |
| **AWS Ecosystem** | EMR, Glue, Athena, DynamoDB, Redshift, Step Functions, Airflow, S3,  IAM, CloudWatch, EC2, Lambda, Snowflake, Databricks. |
| **BI & Visualization tools** | Tableau, Power BI, QuickSight, Excel (Pivot Tables, Power Query, VBA), matplotlib, ggplot2 |
| **Big Data Ecosystem** | Spark, Hive, HDFS, MapReduce, Yarn, Pig, Kafka, Airflow |
| **Scripting Languages** | Shell scripting |
| **IDE** | Intellij, Jupyter Notebook, PyCharm |
| **Databases** | MySQL, Oracle, MongoDB, HBase |
| **Version Control** | GitHub, Bitbucket |
| **Platform** | Windows, Linux |

**EDUCATION:**

**Saint Louis University - Saint Louis, MO Jan 2023 - May 2024**

***Master’s in Computer & Information Sciences (Data Analytics)* *GPA - 3.83 / 4***

**Prasad V Potluri Siddhartha Institute of Technology Sep 2014 - Mar 2018**

***Bachelor’s in Electrical & Electronics Engineering* *GPA - 8.23 / 10***

**CERTIFICATION:**

**AWS Data Engineer - Associate Certification**

**PROFESSIONAL EXPERIENCE:**

**AT & T - Remote Aug 2024 - Present**

**Data Engineer**

**Responsibilities:**

* Collaborated with network operations, product, and engineering stakeholders to gather requirements and design scalable AWS data architectures for telecom analytics (CDRs, network telemetry, OSS/BSS), supporting NOC, capacity planning, and executive reporting.
* Led proof-of-concepts demonstrating near-real-time network monitoring and capacity forecasting using Kinesis Data Streams / Amazon MSK, Lambda, and EMR/Glue to ingest and process streaming call-detail records and telemetry.
* Designed and implemented end-to-end ETL/ELT pipelines on AWS: ingest (Kinesis/MSK, Data Firehose, S3), processing (EMR PySpark, Glue Jobs), cataloging (Glue Data Catalog), storage (S3, Redshift), and serving (Athena, Redshift Spectrum, QuickSight) for low-latency operational analytics.
* Built interactive operational dashboards and KPIs in QuickSight and Tableau to surface network health metrics (call drop rate, latency, throughput), subscriber behavior, service-level compliance, and capacity utilization for NOC and product teams.
* Optimized Redshift and Athena query performance through partitioning, compression, and distribution/sort key strategies for large-scale CDR and telemetry datasets, reducing query latency and cost.
* Designed telecom-specific data models (normalized CDR fact tables and subscriber/network dimension tables) using star/snowflake patterns to enable efficient drill-downs and cross-domain analytics.
* Implemented robust data validation, schema-evolution handling, deduplication, and referential-integrity checks across streaming and batch sources; created quarantine workflows for malformed or out-of-spec records.
* Built real-time anomaly detection and alerting pipelines (Kinesis → Lambda → EMR → CloudWatch + SNS) to detect outages, congestion, and service degradations and notify incident response teams.
* Automated migration of on-prem Hadoop workloads and legacy analytics pipelines to AWS (S3, EMR, Glue, Redshift), improving scalability and reducing operational overhead.
* Orchestrated reliable, auditable ETL workflows using Airflow and AWS Step Functions with retry/rollback logic for telecom batch jobs and SLA-driven reporting.
* Integrated subscriber, OSS/BSS, and operational datasets (S3, RDS/DynamoDB) to enable joined analytics for churn prediction, revenue assurance, roaming analysis, and network optimization use cases.
* Applied data governance, encryption (KMS), RBAC, and data retention policies to meet telecom compliance and security requirements for sensitive subscriber and billing data.
* Supported pre-sales and stakeholder briefings by presenting telecom-specific POCs, architecture diagrams, and ROI analyses to accelerate adoption of AWS analytics solutions.
* Conducted technical workshops and whiteboard sessions with NOC, product, and business teams to translate telecom KPIs into analytical requirements and operational runbooks.
* Documented ETL pipelines, data lineage, and operational runbooks; produced SLA and cost reports to inform capacity planning and spend optimization.
* Delivered executive presentations summarizing network insights, pipeline SLAs, and architecture recommendations that informed operational decisions and reduced incident MTTR.

**Environment:** AWS (S3, Kinesis Data Streams, Amazon MSK, Kinesis Data Firehose, EMR, Glue, Glue Data Catalog, Redshift, Athena, Lambda, Step Functions, CloudWatch, IAM, KMS), QuickSight, Tableau, Python, PySpark, SQL, Git.

**Citi Bank - Chennai, India July 2020 – Dec 2022**

**Data Engineer**

**Responsibilities:**

* Partnered with business, infrastructure, and compliance teams to analyze legacy on-prem data warehouse architecture and design a scalable AWS-based data lakehouse solution for enterprise BI and analytics.
* Led assessment and migration strategy for Hadoop/Hive and on-prem ETL workloads to AWS S3, EMR, Glue, and Redshift, ensuring minimal downtime and data integrity during cutover.
* Designed and implemented secure, scalable ingestion pipelines using AWS Glue Jobs, Glue Crawlers, and EMR PySpark to process structured (Oracle, Teradata) and semi-structured (JSON, logs) data.
* Automated batch ETL workflows using AWS Step Functions, Airflow, and Lambda to orchestrate data ingestion, transformation, and downstream loads with dependency handling and SLA tracking.
* Developed star and snowflake schema models for Redshift to optimize reporting performance for finance, risk, and customer analytics dashboards.
* Tuned Redshift queries with distribution/sort keys, WLM queues, and compression encoding to improve large-scale workload performance and reduce cost.
* Integrated streaming data from Kinesis Data Streams and Firehose for near real-time fraud detection, transaction monitoring, and operational dashboards.
* Built interactive dashboards and executive reports in Tableau and Power BI, surfacing KPIs such as customer transactions, credit risk exposure, and operational SLAs for business leaders.
* Implemented robust data validation, deduplication, and quality checks to guarantee accuracy and consistency of financial datasets before downstream consumption.
* Enabled lineage tracking and governance with Glue Data Catalog and AWS Lake Formation, providing RBAC and auditability for sensitive PII and compliance data.
* Coordinated migration sprints with cross-functional teams, managing dependencies, rollback plans, and post-migration validation to ensure production stability.
* Optimized EMR Spark clusters (autoscaling, spot instances, parallelism tuning) to reduce processing time and operational cost of daily ETL workloads.
* Automated cost monitoring and budget alerts using CloudWatch and Budgets to ensure predictable cloud spend and chargeback reporting.
* Delivered executive-level summaries and architecture recommendations, driving stakeholder alignment and AWS adoption for enterprise analytics modernization.
* Authored detailed runbooks, pipeline documentation, and knowledge-transfer materials to support ongoing operations and governance post-migration.

**Environment:** Python, SQL, PySpark, AWS (S3, EMR, Glue, Glue Crawlers, Glue Data Catalog, Lake Formation, Redshift, Athena, Kinesis Data Streams, Kinesis Firehose, Lambda, Step Functions, CloudWatch, IAM, KMS), Tableau, Power BI, Airflow, Git.

**Three UK - Chennai, India Sep 2018 – June 2020**

**Hadoop Developer**

**Responsibilities:**

* Collaborated with business stakeholders to **gather requirements and design end-to-end Big Data solutions**, ensuring alignment with customer objectives and reporting needs.
* Developed and maintained **Hadoop-based ETL pipelines** for ingestion, transformation, and processing of large-scale structured and semi-structured datasets.
* Built distributed data workflows using **Spark and Hive**, enabling efficient analytics-ready datasets for downstream reporting and BI dashboards.
* Designed Hive table schemas (managed and external) and implemented **partitions and bucketing strategies** to optimize query performance for large datasets.
* Engaged directly with business users to **translate KPIs and reporting needs into technical designs**, ensuring actionable insights and stakeholder satisfaction.
* Performed **data validation, deduplication, and quality checks**, ensuring completeness, accuracy, and reliability of datasets before downstream consumption.
* Leveraged **Sqoop to ingest enterprise data from MySQL and Oracle into HDFS**, enabling a centralized and unified data lake for analytics purposes.
* Created PySpark applications to perform **exploratory data analysis (EDA)** and deliver curated datasets for operational dashboards and metrics reporting.
* Tuned HiveQL and SparkSQL queries for high-volume processing, improving batch execution times and ensuring timely delivery of insights.
* Integrated on-premise Hadoop workloads with internal enterprise systems, supporting **cross-department analytics and reporting requirements**.
* Automated batch data pipelines using shell scripts and cron jobs, enhancing reliability and efficiency of recurring data processes.
* Prepared and delivered **technical briefings and presentations** to business users, explaining Big Data architecture, pipeline designs, and insights derived from analytics.
* Monitored Hadoop job performance using **YARN ResourceManager and Spark UI**, identifying bottlenecks and tuning memory/parallelism for optimized execution.
* Coordinated with business analysts to define metrics, KPIs, and reports, ensuring technical outputs aligned with **strategic business objectives**.
* Participated in Agile sprints, collaborating with teams to deliver incremental improvements in Big Data pipelines and reporting, driving **timely value delivery to stakeholders**.

**Environment:** Python, Java, SQL, Spark, Hadoop, HDFS, Hive, Sqoop, UNIX, Shell scripting, Tableau.

**ACADEMIC EXPERIENCE:**

**Saint Louis University - Saint Louis, MO Jan 2023 – May 2024**

**Professional Graduate Student - *MS in Computer & Information Sciences (Data Analytics)***

**Responsibilities:**

* Collaborated with stakeholders to **gather requirements and design end-to-end analytics solutions**, simulating customer-facing consulting scenarios.
* Developed **cloud-native ETL pipelines on AWS**, integrating structured and semi-structured datasets into analytics-ready models using BigQuery, Cloud SQL, and Cloud Storage.
* Built interactive dashboards and KPIs in Tableau and Excel, **translating analytics insights for both technical and non-technical audiences**.
* Conducted statistical analysis and predictive modeling using R and Python (ANOVA, regression, forecasting) to enable **data-driven decision support**.
* Integrated Python (pandas, matplotlib) and SQL for **data extraction, transformation, and visualization**, reinforcing ETL and reporting expertise.
* Delivered **proof-of-concept dashboards and presentations**, demonstrating cloud analytics pipelines and actionable insights to simulated stakeholders.
* Applied **data modeling techniques (star/snowflake schema, fact/dimension design)** to optimize analytical workflows and BI performance.
* Implemented **data validation, integrity checks, and deduplication**, ensuring reliable and accurate analytics outputs.
* Authored a **Major Research Paper (MRP) on ITIL framework**, proposing enhancements for IT project migrations and structured problem-solving.
* Designed **business process models (BPMN) and UML diagrams** to map workflows, aligning technical solutions with organizational objectives.
* Leveraged GCP services to **handle large-scale datasets**, gaining hands-on experience with cloud-native analytics and scalability principles.
* Applied Agile principles using Jira and Trello for **project tracking, sprint planning, and cross-team collaboration**, simulating enterprise consulting delivery.
* Conducted exploratory data analysis (EDA) to uncover actionable insights, supporting **strategic recommendations and executive-style presentations**.
* Integrated healthcare datasets for analytics projects, combining **statistical modeling with dashboards**, demonstrating real-world problem-solving.
* Developed **self-service analytics capabilities** using Python and SQL, enabling stakeholders to query datasets and generate reports independently.
* Presented technical solutions, dashboards, and analytics outcomes in **executive-style briefings**, simulating customer engagement scenarios.
* Documented analytics workflows, pipeline specifications, and data lineage for **knowledge transfer, reproducibility, and audit readiness**.
* Applied advanced visualization techniques (ggplot2, matplotlib) to effectively **communicate insights and enhance decision-making impact**.

**Environment:** Python, SQL, R, Tableau, Excel (Pivot Tables, Power Query, VBA), pandas, tidyverse, matplotlib, ggplot2, AWS, Jira, Trello, BPMN, UML, PlantUML, Visual Paradigm.