Sowmya Marripeddi

## Sr. Data Engineer

 [My Portfolio](https://v0-portfolio-rho-pink.vercel.app/)

 +(838)-300-3088

 [sowmyamarripeddi12@gmail.com](mailto:sowmyamarripeddi12@gmail.com)  [linkedin.com/in/sowmya-marripeddi-dataengineer](https://www.linkedin.com/in/sowmya-marripeddi-dataengineer)



# PROFESSIONAL SUMMARY

* Over 10+ years of experience architecting, developing, and maintaining scalable Big Data and Cloud solutions across retail, logistics, and enterprise domains, specializing in designing solutions that meet stringent regulatory requirements and drive data-informed

decision-making.

* Designed and implemented end-to-end data processing frameworks using Hadoop, Spark, Hive, and HBase, and built cloud-native data lake architectures on Azure Data Lake Storage Gen2, enabling centralized access and analytics on multi-terabyte datasets.
* Engineered robust ETL pipelines in Azure Data Factory to automate data ingestion from diverse sources (MongoDB, MS SQL, Azure Blob, ADLS Gen2), achieving 95% reduction in manual processing and improving data accessibility and quality.
* Architected real-time streaming platforms using Apache Kafka and Spark Structured Streaming, processing millions of events per second for behavioral analytics, fraud detection, and real-time decision-making, enabling proactive risk management and personalized customer experiences.
* Developed custom Spark applications in Databricks for event enrichment, de-duplication, and aggregation, optimizing batch jobs to run 3x faster and enhancing data processing efficiency and analytical capabilities.
* Deployed Azure Synapse Analytics to build cloud-based data warehouses, integrating data from various systems to power executive dashboards and providing actionable insights for stakeholders.
* Trained and deployed machine learning models using Spark MLlib and TensorFlow in Databricks for customer segmentation and product recommendation engines, enabling sub-second personalization and enhanced customer engagement.
* Administered and optimized Azure Databricks clusters, configuring job parameters, autoscaling, and cost controls, reducing compute spend by 30% and maximizing resource utilization and minimizing operational costs.
* Created data pipelines to process semi-structured and unstructured data (JSON, AVRO, Parquet, ORC), collaborating with ML and BI teams for advanced data visualization, facilitating deeper data analysis and improved decision support.
* Delivered CI/CD automation using Jenkins and Maven to ensure safe, fast, and repeatable deployments of Spark jobs, REST APIs, and Azure services, and engineered metadata-driven ingestion logic for enterprise-level compliance and auditability, reducing deployment risks and ensuring data integrity.
* Orchestrated end-to-end data pipelines using Azure Data Factory, integrating diverse data sources (MongoDB, MS SQL, Azure Blob, ADLS Gen2), ensuring seamless data flow for real-time analytics and reporting, and driving data-driven insights across the organization.
* Proficient in architecting and managing scalable cloud data solutions using Snowflake on GCP, enabling cost-effective, high-performance analytics, secure data sharing, and seamless integration across modern data platforms.
* Architected and implemented a cloud-native data mesh architecture, decentralizing data ownership and enabling domain-driven data access for real-time analytics and self-service BI, driving agile decision-making and innovation.
* Implemented robust data security measures, including data encryption at rest and in transit using Azure Key Vault and Azure Active Directory, and applied Zero Trust security principles, ensuring the highest levels of data protection and compliance with industry standards.

# SKILLS

## Programming Languages:

Python (PySpark, Pandas, NumPy, Scikit-learn, TensorFlow/PyTorch), Scala, SQL (Spark SQL, T-SQL, PL/SQL, SnowSQL), R, Shell, Java, Go

## Cloud & Big Data:

Azure (ADLS Gen2, Synapse, Databricks, Data Factory, Event Hubs, Purview, Functions, Cosmos DB, ML, AKS, Key Vault, HDInsight), GCP (BigQuery, Dataflow, Composer, GCS, Pub/Sub, Cloud Functions), Snowflake, Hadoop (HDFS, YARN), Spark (Core, Streaming, SQL, MLlib), Hive, Kafka, Delta Lake, Airflow

## Databases & Warehousing:

SQL (Oracle, SQL Server, MySQL, PostgreSQL, Snowflake), NoSQL (Cassandra, MongoDB, Cosmos DB, HBase), DW (Snowflake, Synapse, BigQuery), Redis

## Data Engineering & DevOps:

ETL (ADF, Airflow, Composer), Data Governance (Purview, Alation), CI/CD (Jenkins, Azure DevOps, GitHub Actions), IaC (Terraform, ARM), Docker, Kubernetes, Observability (Prometheus, Grafana, ELK), Data Mesh, MLOps

## ML & Analytics:

ML (Regression, Classification, Clustering, DL), Visualization (Tableau, Power BI, Looker), Statistical Modeling, MLflow, Feature Engineering, A/B Testing

# EDUCATION

**JNTU Kakinada** *June 2009 – Aug 2013*

Bachelor’s Degree in Computer Science and Engineering Kakinada, IN

**Relevant Coursework:** Data Structures and Algorithms, Cloud Computing, Distributed Systems, Database Systems, Data Warehousing, Big Data Analytics, Machine Learning, Statistical Modeling, Data Mining, ETL Design, Data Governance.

# WORK EXPERIENCE

**Azure Data Engineer** *Sep 2023 – Present*

*Citizens Bank | Jersey City, NJ*

At ***Citizens Bank***, I spearheaded the design and implementation of a cloud-native data mesh architecture utilizing Azure services, enabling decentralized data ownership and facilitating real-time, data-driven decision-making across diverse business units. This role involved building scalable ETL pipelines, ensuring robust data quality, and implementing comprehensive data governance frameworks to comply with stringent regulatory requirements.

* Architected a cloud-native data lake on Azure Data Lake Storage Gen2, consolidating 10+ data sources and improving analytics accessibility by 40%, directly supporting business intelligence and regulatory reporting requirements effectively.
* Engineered metadata-driven ETL pipelines in Azure Data Factory, reducing deployment time by 35% and enabling rapid adaptation to evolving data integration needs across multiple business domains.
* Implemented real-time ingestion with Kafka and Azure Event Hubs, enabling instant fraud detection and reducing loss exposure by 25% through actionable alerts and automated response mechanisms.
* Developed scalable data transformations in Azure Databricks (Spark) with Delta Lake, enabling precise customer segmentation and boosting campaign ROI by 30%, significantly enhancing marketing effectiveness.
* Designed and deployed Azure Synapse Analytics dedicated SQL pools, optimizing query performance and delivering actionable insights for BI and ML, accelerating executive decision-making processes consistently.
* Automated data archival and retention with Azure Blob Storage and Lifecycle Management, cutting storage costs by 30% and ensuring strict compliance with evolving data governance policies and standards.
* Built robust CI/CD pipelines using Azure DevOps and Terraform, reducing release cycles by 40% and ensuring consistent, error-free deployments across multiple environments and teams.
* Established enterprise data governance with Azure Purview, ensuring end-to-end lineage, metadata management, and GDPR/CCPA compliance, increasing audit-readiness and regulatory adherence significantly.
* Optimized Azure SQL Database for high-concurrency workloads via advanced indexing, query tuning, and partitioning, maintaining 99.99% uptime for critical digital banking services with minimal latency.
* Developed Power BI dashboards for real-time KPI visualization, empowering stakeholders with self-service analytics and improving operational responsiveness and decision-making efficiency substantially.
* Automated data validation and quality checks with Azure Functions (Python), reducing manual errors by 15% and enhancing trust in enterprise data assets across multiple business units.
* Established a federated computational governance model across all data domains, leveraging Azure Policy and centralized templates, which standardized data quality, security, and interoperability, reducing cross-domain integration issues by 40% and ensuring consistent regulatory compliance across the enterprise.
* Implemented advanced data observability and cataloging solutions using Azure Purview and custom monitoring dashboards, providing end-to-end lineage, real-time data health metrics, and automated SLA tracking, which improved data discoverability and reduced incident resolution times by 35%.
* Established a federated computational governance model across all data domains, leveraging Azure Policy and centralized templates, which standardized data quality, security, and interoperability, reducing cross-domain integration issues by 40% and ensuring consistent regulatory compliance across the enterprise.
* Integrated third-party financial data into ADLS Gen2 using Data Factory and API connectors, enriching customer profiles and improving cross-sell/up-sell analytics by 15%, driving revenue growth.
* Enforced Zero Trust security by implementing encryption (at rest/in transit) with Azure Key Vault and Active Directory, ensuring compliance with industry best practices and regulatory requirements.
* Leveraged Azure Monitor and Log Analytics for proactive system health monitoring, reducing downtime by 20% and increasing reliability of mission-critical pipelines and services significantly.
* Applied advanced ML algorithms (linear regression, PCA, K-means, KNN) using Databricks and Python to drive risk modeling and fraud detection innovation, supporting advanced data science initiatives effectively.
* **Environment:** Azure (ADLS Gen2, Synapse, Databricks, AKS, Data Factory, Event Hubs, Stream Analytics, DevOps, Purview, Functions, Cosmos DB, Machine Learning, AI Services), Python (Pandas, NumPy, Scikit-learn, TensorFlow/PyTorch), SQL, Kafka, Data Governance (Data Mesh), MLOps, Real-time Processing, Cloud-Native, IaC, Data Security (Zero Trust), Power BI/Tableau.

**Azure Data Engineer** *Aug 2021 – Sep 2023*

*Johnson & Johnson | Santa Clara, CA*

At ***Johnson & Johnson***, I led data engineering initiatives across pharmaceutical supply chains, clinical trial analytics, and public health data systems. My work focused on building secure, scalable, and intelligent data platforms to improve patient outcomes, enhance production reliability, and accelerate drug and device approval pipelines, in alignment with J&J’s mission to deliver life-saving products efficiently and safely.

* Engineered real-time streaming pipelines with Kafka and Spark Streaming to monitor pharmaceutical supply chain events, enabling proactive quality control, predictive analytics, and minimizing delivery delays of critical medications globally.
* Architected a centralized data lake using Azure Data Lake Storage Gen2, consolidating global clinical trial datasets and reducing research cycle times for new drug development by 25% through unified access.
* Designed and deployed Azure Synapse Analytics solutions to analyze patient health records, driving J&J’s personalized medicine initiative, improving clinical decision-making, and supporting advanced research teams.
* Enforced data governance and HIPAA compliance using Azure Key Vault, IAM roles, and automated security policies to protect sensitive patient data and enhance trust among stakeholders and regulators.
* Developed PySpark pipelines to ingest and process manufacturing quality control data, enabling anomaly detection, regulatory compliance, and improving product quality across global production sites.
* Integrated pharmacovigilance data pipelines using Azure Data Factory and Cosmos DB to monitor adverse drug events, accelerating safety reporting and reducing compliance risk for regulatory authorities.
* Built automated workflows using Azure HDInsight for ingesting IoT data from medical devices, enabling near real-time monitoring of device performance, patient adherence, and clinical trial integrity.
* Constructed event-driven data pipelines via Azure Event Hubs and Service Bus to analyze medical device usage, contributing to post-market surveillance, performance analytics, and timely product recalls.
* Created highly optimized ETL pipelines for genomic data processing using Databricks and Spark SQL, supporting R&D teams in gene-targeted therapies and precision medicine initiatives worldwide.
* Integrated Tableau dashboards with Hadoop and Spark sources to visualize trial outcomes and drug efficacy metrics, streamlining FDA submission reviews and accelerating regulatory approval cycles.
* Implemented secure NoSQL stores via Azure Cosmos DB to support dynamic, real-time access to patient data across healthcare systems, improving care coordination and patient safety.
* Automated adverse event report processing using Azure Functions (Python), enhancing the speed, accuracy, and auditability of critical drug safety workflows and compliance submissions.
* Used Azure Machine Learning to build predictive models for early detection of chronic illnesses, helping J&J teams pilot preventative care programs with healthcare partners and improve patient outcomes.
* Developed scalable data warehouse models (Star and Snowflake schemas) to consolidate pharmaceutical sales, marketing, and prescription data, enabling improved forecasting and strategic planning.
* Designed IaC pipelines with Azure Resource Manager (ARM) templates to provision production-grade cloud infrastructure, ensuring enterprise security, compliance, and rapid scaling for new projects.
* **Environment:** Azure (ADLS Gen2, Synapse Analytics, Databricks, Cosmos DB, Functions, Event Hubs, Service Bus, Machine Learning, IAM, Key Vault, Blob Storage, HDInsight, PostgreSQL, ARM, Event Grid), Python (PySpark, Pandas, NumPy), Scala, SQL (Spark SQL,

T-SQL), Kafka, Cassandra, Hadoop (Hive, HDFS), Talend, Tableau, Linux, Git, Jenkins, YAML.

**ETL Developer** *Aug 2020 – Aug 2021*

*First Republic Bank | San Francisco, CA*

At ***First Republic Bank***, I developed and optimized enterprise-grade ETL pipelines to support regulatory compliance, financial analytics, and customer data integration across the bank’s digital ecosystem. This work was critical in enabling the institution to adapt to surging digital banking activity during the COVID-19 period, ensuring robust reporting, faster loan processing, and data integrity across banking systems.

* Designed and implemented highly-optimized ETL pipelines using Azure Data Factory, orchestrating data ingestion from on-prem SQL Server to ADLS Gen2, increasing financial data availability for risk management by 30%, and enabling near real-time analytics for compliance teams.
* Developed advanced SQL queries and complex stored procedures to transform, cleanse, and validate customer account data, ensuring integrity and accuracy for FDIC audits, SOX compliance, and regulatory reporting across all lines of business.
* Engineered comprehensive data quality checkpoints within ETL workflows using Databricks and Spark SQL, automating anomaly detection and enhancing accuracy for loan servicing, treasury, and financial statement data.
* Automated large-scale loan data migration using Azure Data Factory and scheduling triggers, supporting surges in loan applications during COVID-19 and improving processing speed and operational scalability by 40%.
* Created detailed data lineage documentation for all ETL pipelines, supporting traceability, auditability, and adherence to evolving federal banking regulations and internal governance standards.
* Built robust ETL solutions to integrate third-party financial data sources into ADLS Gen2, enriching datasets for predictive analytics, credit scoring models, and customer risk profiling initiatives.
* Implemented advanced data cleansing and normalization logic to prepare CRM datasets for targeted marketing, improving campaign precision, segmentation accuracy, and customer engagement rates.
* Designed automated ETL processes for FDIC report generation, ensuring compliance and timely submissions during high volatility in the financial sector (2020–2021), reducing manual workload for risk teams.
* Tuned ETL pipeline performance using indexing, partitioning, and SQL optimization techniques, maintaining stability and consistent throughput under increasing transactional loads from digital banking services.
* Developed automated workflows to produce daily financial summaries and KPI dashboards for regional branch leaders, enhancing decision-making speed, accuracy, and operational transparency across the organization.
* Created serverless ETL logic using Azure Functions and event-driven triggers for fast, scalable transformations and real-time validation of incoming financial datasets, reducing latency and manual intervention.
* Defined comprehensive data mappings and orchestrated transformation layers for legacy-to-Azure data migration, ensuring business continuity, data integrity, and accurate cross-platform mapping during digital modernization.
* Implemented archival logic in Azure Blob Storage, automating retention policies, reducing storage costs, and ensuring compliance for historical loan and customer datasets in accordance with regulatory mandates.
* Developed unified ETL solutions for integrating data from diverse financial instruments, including securities and derivatives, improving portfolio risk visibility, exposure tracking, and regulatory transparency for trading operations.
* **Environment:** Azure (Data Factory, Data Lake Storage Gen2, Synapse Analytics, SQL Database, Functions, Blob Storage, Analysis Services, Azure DevOps), SQL Server, Spark SQL, Databricks, SSIS, Oracle SQL Developer, Power BI, Windows Server, Linux (for scripting/data processing), Git.

**Data Engineer** *Feb 2018 – Aug 2020*

*Kaiser Permanente | Oakland, CA*

At ***Kaiser Permanente***, I designed and implemented real-time, cloud-native data solutions across clinical, supply chain, and operational domains. My work was pivotal in enabling predictive healthcare delivery, automating patient engagement, and supporting data governance across a large-scale healthcare enterprise, with strict adherence to HIPAA compliance and operational resiliency.

* Architected robust real-time data pipelines using Kafka and Azure Event Hubs to process clinical events, supporting instant decision-making in emergency care workflows and reducing response times by 20% for critical interventions.
* Built a unified data processing layer in Azure Databricks (Spark) to enrich patient records by integrating lab results, prescriptions, and imaging data, enabling holistic profile generation and advanced population health analytics.
* Developed serverless, event-driven architecture using Azure Functions and Cosmos DB to automate care coordination and appointment reminders, improving patient outreach, reducing readmissions, and supporting value-based care delivery.
* Engineered inventory tracking workflows by merging point-of-care usage data with supply chain analytics via Azure Synapse, improving stock optimization with ML-backed forecasting models and reducing supply shortages by 15%.
* Implemented a centralized data lake and analytics hub using ADLS Gen2, Databricks, and Synapse Analytics to enable outcome-based clinical analysis, performance reporting, and regulatory submissions for enterprise stakeholders.
* Built predictive machine learning models in Azure ML for patient risk scoring and chronic disease intervention planning, supporting value-based care initiatives and reducing hospital readmission rates across the network.
* Containerized Spark-based pipelines using Docker and deployed on AKS, achieving high resilience, seamless scaling during periods of elevated healthcare data flow, and minimizing downtime for mission-critical workloads.
* Applied robust data governance frameworks using Azure AD, Azure Policy, and Purview to ensure compliance with HIPAA, GDPR, and internal auditing mandates, automating access controls and audit trail management.
* Created real-time Power BI dashboards for care teams and administrators, visualizing population health trends, individual care metrics, and supporting proactive medical interventions at the point of care.
* Leveraged serverless design patterns and auto-scaling infrastructure to reduce compute costs while maintaining real-time SLAs in critical clinical processing pipelines, optimizing resource utilization and operational efficiency.
* Implemented GitOps-based CI/CD automation for data service deployment, reducing manual errors, improving release frequency, and ensuring consistent, reliable updates to healthcare data platforms.
* Integrated FHIR-compliant APIs for secure data exchange between EHR systems, supporting interoperability, and enabling seamless patient information transfer across multiple hospital departments and partners.
* Automated data quality monitoring and anomaly detection using Azure Monitor and custom Python scripts, ensuring high data integrity, timely alerts, and compliance with regulatory reporting standards.
* **Environment:** Azure (Event Hubs, Databricks, Azure ML, ADLS Gen2, Functions, Cosmos DB, Synapse Analytics, AKS, Azure AD, Azure Policy, Azure Purview), Kafka, Spark (Structured Streaming, MLlib), Python, SQL, Docker, Git, Power BI, FHIR APIs, IoT.

**Big Data Engineer** *Jan 2017 – Feb 2018*

*Guild Mortgage | San Diego, CA*

At ***Guild Mortgage***, I was responsible for modernizing the data infrastructure and building scalable big data pipelines to support risk analytics, loan portfolio analysis, and regulatory compliance. The focus was on transitioning from legacy Hadoop workflows to modern Spark-based frameworks while integrating real-time data streams and enhancing reporting speed for business intelligence.

* Architected and productionized Sqoop-based ingestion pipelines to extract large-scale mortgage data from Oracle and PostgreSQL into HDFS, enabling scalable analytics for loan performance and risk scoring models used by the credit risk team and auditors.
* Migrated legacy MapReduce jobs to Spark (Python), significantly improving ETL execution times, reducing latency in regulatory and financial reporting cycles, and enabling faster compliance submissions to federal regulators and internal stakeholders.
* Developed robust batch processing jobs using Spark on Cloudera, improving loan analytics performance 10x over MapReduce and enabling near real-time loan trend insights for business intelligence and executive dashboards.
* Built Spark DataFrame-based applications for advanced Hive analytics, delivering actionable metrics used for forecasting delinquency, servicing trends, and mortgage application funnel performance, directly impacting portfolio management strategies.
* Integrated Kafka-based real-time pipelines to stream mortgage application data, facilitating faster credit decisions, fraud detection workflows, and improving customer experience with instant loan status notifications and alerts.
* Designed and executed automated migration pipelines from Hive data lake to Amazon S3, enhancing data durability, enabling cloud-based analytics, and supporting disaster recovery and business continuity planning.
* Loaded and modeled structured data in Snowflake from S3, supporting interactive BI dashboards and compliance reporting used across finance, risk, and executive teams for strategic decision-making.
* Wrote complex HiveQL queries over partitioned and bucketed datasets to generate detailed reports for loan servicing, credit risk, default probability tracking, and investor performance summaries.
* Created custom Spark UDFs and UDAFs to implement domain-specific logic, extending analytical capabilities for mortgage processing, validation rules, and automating exception handling for edge-case scenarios.
* Tuned Spark performance by optimizing shuffle partitions, leveraging map-side joins, and using Zookeeper for concurrent Hive table access, ensuring stability and high throughput during peak processing windows.
* **Environment:** Linux, Apache Hadoop Framework (HDFS, YARN), Hive, HBase, Scala, Spark, Sqoop, Pig, Hadoop (MAPR 5.0), MapReduce, Informatica PowerCenter, Python, Microsoft SQL Server, Cassandra, Jira, UNIX Shell Scripting, Kafka, Snowflake, Cloudera.

**Data Warehouse Developer** *Sep 2013 – Oct 2015*

*Value Labs | Hyderabad, India*

At ***Value Labs***, I focused on building high-performance data warehouse solutions and ETL frameworks for enterprise-scale applications. My work empowered teams with optimized data access, improved reporting workflows, and consistent data integration across traditional and modern platforms, significantly enhancing decision-making processes across departments.

* Orchestrated development of advanced stored procedures, triggers, and custom functions in SQL Server, dramatically improving query efficiency and core application responsiveness for mission-critical business operations and analytics workloads.
* Executed complex SQL Server performance tuning by analyzing execution plans, adding strategic indexes, and restructuring queries to deliver superior system throughput and minimize latency for large-scale reporting.
* Designed robust ETL frameworks using SSIS, transforming and loading data from SQL Server, Access, and Excel sources into structured warehouse models, ensuring analytics readiness and data consistency across business units.
* Automated cross-platform data pipelines with Azure Data Factory, Informatica, and SSIS, ensuring consistent, real-time data synchronization between heterogeneous systems and supporting seamless business process integration.
* Authored detailed entity-relationship diagrams and data lineage documentation, enabling full traceability of upstream/downstream data dependencies for compliance, auditing, and regulatory reporting requirements.
* Designed and implemented dimensional data models including SCD Types I and II, star and snowflake schemas, and surrogate key generation logic to support scalable, high-performance OLAP analytics and reporting.
* Integrated advanced ETL error handling using checkpoints, breakpoints, logging, and precedence constraints, ensuring robust failover, data validation workflows, and rapid root cause analysis for data quality incidents.
* Created and maintained multidimensional cubes in SSAS with partitioning, aggregations, and KPIs to support rapid OLAP querying, enabling business users to perform deep-dive analytics with sub-second response times.
* Automated workflow orchestration for complex ETL pipelines using Apache Airflow and Oozie, enabling scheduled and event-driven data jobs, improving operational efficiency and reducing manual intervention across the warehouse stack.
* Built dynamic SSRS reports on SSAS cubes, delivering dashboards with drill-down, drill-through, and cascading parameters for actionable business intelligence, executive reporting, and real-time operational monitoring.
* **Environment:** MS SQL Server, Visual Studio, SSIS, SSAS, SharePoint, MS Access, Team Foundation Server, Git, Apache Airflow, Apache Oozie, Informatica.