**Jaya Sai**

**jaysai1100.us@gmail.com**

**704-350-5687**

**Senior Data Engineer**

**Professional Summary:**

Experienced and results-driven **Senior Data Engineer** with over 9 years of expertise in designing, developing, and optimizing large-scale data systems for various industries, including banking, insurance, healthcare, and retail. Proficient in **ETL pipeline development**, **cloud computing** (AWS, GCP, Azure), **real-time data streaming**, and building high-performance data models for analytical and transactional workloads. Skilled in technologies such as **Apache** **Spark**, **PySpark**, **Kafka**, **Snowflake**, and **Redshift**, with a strong track record of improving data processing efficiency, reducing operational costs, and enhancing data quality.

Demonstrated expertise in integrating complex data sources, automating workflows, and ensuring compliance with industry regulations such as **GDPR**, **HIPAA**, and **PCI-DSS**. Adept at leveraging **machine** **learning** tools like **Scikit-learn** for predictive analytics and collaborating cross-functionally with business analysts and data scientists to provide actionable insights. Experienced Senior Data Engineer with expertise in designing scalable data solutions on cloud platforms including AWS, GCP, and Azure. Skilled in developing and optimizing machine learning pipelines using Scikit-learn and TensorFlow for predictive analytics and business insights. Proven leader in mentoring junior engineers, optimizing data pipelines, and implementing cloud-based solutions for improved scalability, performance, and data security. Passionate about building reliable data systems that enhance decision-making and support business growth.

**Technical Skills:**

| **Category** | **Skills** |
| --- | --- |
| **Data Engineering** | ETL Development, Data Warehousing, Data Modeling, Real-time Data Streaming, Batch Processing, Data Pipeline Optimization |
| **Big Data Technologies** | Apache Spark, PySpark, Hadoop, Apache Kafka, Snowflake, Amazon Redshift |
| **Cloud Platforms** | AWS (Redshift, Glue, Lambda, S3), GCP (BigQuery, Dataflow, Pub/Sub, Vertex AI, Cloud Storage, Cloud Functions), Azure (Azure Data Lake, SQL Database, Azure Databricks) |
| **Databases** | SQL (MySQL, PostgreSQL, SQL Server), NoSQL (MongoDB, Cassandra), Cloud Databases (Snowflake, Redshift, BigQuery) |
| **Programming Languages** | Python, SQL, Java, Scala |
| **Data Integration** | Apache NiFi, Talend, Informatica, Apache Airflow, DBT |
| **Data Governance & Security** | Data Privacy (GDPR, HIPAA), Data Security, Compliance, Data Quality Management |
| **Machine Learning** | Scikit-learn, TensorFlow (for basic ML models), GCP Vertex AI, AutoML, TensorFlow, Scikit-learn, MLflow, Model Training, Hyperparameter Tuning, Deployment of ML models on Cloud , Predictive Analytics |
| **DevOps & Automation** | CI/CD (Jenkins, GitLab), Docker, Kubernetes, Terraform |
| **Data Visualization** | Tableau, Power BI, Looker, Data Studio |
| **Version Control & Collaboration** | Git, GitHub, Bitbucket, Jira, Confluence |
| **Scripting & Automation** | Shell Scripting, Python Scripting, SQL Scripting, **Scope Scripting** |
| **Operating Systems** | Linux, Windows |
| **Business Intelligence** | Data Warehousing, OLAP Cubes, Data Lake Implementation, Reporting Solutions |
| **Other Tools** | Apache Airflow, Apache Flume, Talend, Informatica, DBT, HDFS |

**Senior Data Engineer**

**US BANK – Charlotte, NC** | **May 2023 – Present**

* Designed, developed, and optimized **ETL pipelines** using **Apache Spark, PySpark, and SQL** to process large-scale financial datasets, improving processing efficiency by **30%**.
* Developed data ingestion frameworks for structured and semi-structured data from diverse sources such as **Kafka**, **S3**, and **RDBMS**, ensuring seamless data flow across systems.
* Engineered high-performance data models for analytical and transactional workloads in **Snowflake**, **Redshift**, and **Databricks**, optimizing query performance and reducing latency.
* Implemented **data lake** and **data warehouse** solutions supporting both **real-time** and **batch processing**, ensuring data integrity, consistency, and security, while reducing storage costs by **20%**.
* Built **CI/CD pipelines** for automated deployment of data workflows using **Apache Airflow**, **Jenkins**, and **Terraform**, streamlining deployment processes and reducing manual errors.
* **Developed and maintained data engineering pipelines supporting OSI and Cosmic Supporting Services** as part of managed support operations, ensuring consistent data flow and availability across environments.
* **Provided Tier 2 support for incident management involving OSI and Cosmic Supporting Services**, collaborating with cross-functional teams to resolve escalated issues and maintain service uptime.
* Developed and maintained Scope Scripts to automate data processing tasks, enabling dynamic workflow execution and reducing manual intervention by 30%.
* Integrated **cloud-based solutions** (AWS/GCP/Azure) for scalable data processing, leveraging **S3, Lambda, Glue**, and **EMR** to enhance data processing efficiency.
* Collaborated with data science teams to deploy ML models using Vertex AI and AutoML for fraud detection and predictive analytics
* Architected scalable data pipelines on Azure using Azure Data Factory for orchestrating ETL workflows across multiple environments.
* Used Azure Synapse Analytics to build and manage enterprise data warehouse solutions, enabling faster and unified data access.
* Designed custom Scope Scripts for automated ETL orchestration across AWS and GCP environments, improving pipeline efficiency and error handling.
* Developed and optimized **SQL queries**, stored procedures, and views for high-performance data retrieval in **PostgreSQL**, **Oracle**, and **SQL Server**.
* Designed and optimized data pipelines on GCP using BigQuery, Dataflow, and Cloud Functions for scalable ETL processing
* Led the migration of legacy ETL jobs from **Informatica/Talend** to **PySpark**, reducing processing time by **30%** and improving scalability.
* Ensured compliance with banking regulations (e.g., **SOX, GDPR, PCI-DSS**) by implementing **data governance frameworks**, **encryption**, and **access controls**.
* Spearheaded the implementation of **real-time data streaming** solutions using **Apache Kafka** and **Kinesis**, reducing transaction processing latency and enhancing fraud detection capabilities.
* Developed and maintained **data quality frameworks** with automated checks to ensure the accuracy and completeness of financial transactions and account data, reducing errors by **25%**.
* Collaborated with **data science teams** to provide clean, structured datasets for machine learning models used in predictive analytics and fraud detection.
* Integrated Azure Blob Storage and Azure Data Lake for storage and retrieval of structured and semi-structured data.
* Designed robust data transformation pipelines using Azure Databricks, enhancing performance for machine learning workflows.
* Optimized data pipelines on **cloud platforms**, reducing cloud storage and processing costs by **20%** while maintaining high performance.
* Led the **data integration strategy** for onboarding new banking services and third-party data sources, accelerating the time-to-market for new products.
* Leveraged Azure Key Vault for secure management of credentials and secrets within CI/CD pipelines.
* Used Azure Event Hubs for real-time ingestion of transactional data, integrating with Kafka for scalable streaming.
* Developed a suite of automated **data monitoring tools** to proactively detect and resolve performance degradation, improving pipeline reliability.
* Worked with the **data security team** to implement **end-to-end encryption** and **data masking** to safeguard sensitive customer data and ensure regulatory compliance.
* Designed and implemented **data partitioning strategies** for optimizing large-scale query performance and reporting in financial data warehouses.
* Mentored junior data engineers, fostering collaboration and improving team efficiency and code quality.
* Implemented Azure Functions for event-driven data processing tasks, reducing operational overhead.
* Built monitoring dashboards in Azure Monitor and Log Analytics to proactively manage pipeline performance.
* Worked closely with **business analysts**, **data scientists**, and **stakeholders** to implement data-driven solutions that enhanced financial insights and decision-making.

**Technologies Used:** Python, SQL, PySpark, Java**,** Apache Spark, Apache Kafka, Apache Airflow, AWS Glue, EMR, Kinesis**,** AWS (S3, Lambda, Redshift), GCP, Azure**,** Snowflake, Redshift, Databricks**,** PostgreSQL, Oracle, SQL Server, MongoDB**,** Data Lakes, Data Warehouses, Star Schema, Snowflake Schema**,** Git, Jenkins, Terraform, Apache Atlas, Data Masking, Encryption, PCI-DSS, SOX, GDPR Compliance**,** Datadog, Splunk,Pandas, Scikit-learn (for feature engineering)**,** Apache Atlas, AWS Glue**,** Custom Data Validation Tools.

**Senior Data Engineer**

**USAA – San Antonio, TX** | **October 2021 – April 2023**

* Designed and optimized **ETL pipelines** with **Apache Spark**, **PySpark**, and **SQL**, processing large-scale datasets from **Claims Data**, **Policy Data**, and **Customer Information**, cutting data processing time by **40%**.
* Built hybrid data processing architecture leveraging both AWS and Azure to **handle cross-cloud data integrations**.
* Developed automated workflows in **Azure Data Factory** to extract and transform insurance policy and claims data.
* Configured **Azure SQL Database** for transactional storage needs and reporting use cases, improving query latency.
* Developed **data ingestion frameworks** to integrate structured and semi-structured data from **APIs**, **RDBMS**, and **flat files (CSV, JSON)** into a **centralized data lake** and **data warehouse**.
* Created high-performance **data models** in **Snowflake** and **Redshift** for insurance claims analysis, fraud detection, and risk modeling, ensuring fast and efficient query performance for business reporting.
* **Contributed to the deployment and configuration of service components integrated with OSI and Cosmic Supporting Services**, aligning with M365-specific security protocols and organizational SLAs.
* Led migration efforts from legacy data systems to a modern **cloud-based data architecture** leveraging **AWS** (S3, Lambda, Glue) and **Azure**, improving scalability and reducing operational overhead by **25%**.
* Integrated **real-time data streaming** solutions using **Apache Kafka** and **Kinesis** for ingesting **transaction data** and **customer behavior insights**, improving insurance claims processing speed and personalization of services.
* Implemented ML-based predictive analytics models using Scikit-learn and TensorFlow on GCP for fraud detection and risk assessment.
* Developed automated **data quality frameworks**, enhancing the accuracy and completeness of **policy** and **claims data** by **20%** through validation and reconciliation processes.
* Orchestrated complex data workflows using **Apache Airflow**, ensuring error-free and timely execution of data pipelines supporting various business units.
* Designed Scope Scripts to automate metadata management, data quality checks, and compliance validation across cloud-based data pipelines.
* Optimized data pipelines leveraging GCP services such as BigQuery and Cloud Composer (Apache Airflow on GCP).
* Enforced **GDPR**, **HIPAA**, and **PCI-DSS** compliance by implementing **encryption**, **data masking**, and **access controls** for sensitive insurance data.
* Partnered with **data scientists** to build pipelines supporting **predictive models** for fraud detection and customer risk scoring, leveraging **Python** and **Scikit-learn** for feature engineering and model development.
* Implemented access controls using **Azure RBAC (Role-Based Access Control)** to enhance security and compliance.
* Used **Azure DevOps for managing source control**, continuous integration, and release management of ETL pipelines.
* Created Power BI dashboards sourcing data directly from Azure Data Lake, improving decision-making for underwriting.
* Created optimized **SQL queries** and **stored procedures** to support claims data reporting, cutting report generation time by **35%** and improving stakeholder access to insights.
* Developed and maintained **data monitoring tools** that proactively detect issues in data pipelines, ensuring **99.8% uptime** for insurance operations.
* Collaborated with **business analysts** to define and translate business requirements into effective data solutions, helping drive key insurance metrics and decisions.
* Enhanced **data governance** with **metadata management** solutions using **AWS Glue** and **Apache Atlas**, improving data lineage, traceability, and operational oversight.
* Played a significant role in **cloud cost optimization**, reducing **storage** and **compute costs** by **25%** through efficient strategies in **AWS** and **Azure**.
* **Mentored junior data engineers**, providing guidance on data pipeline development, best practices, and contributing to team productivity and code quality.
* Participated in **Agile sprints**, working cross-functionally with **product managers**, **data scientists**, and **data analysts** to support new insurance product features and enhancements.

**Technologies Used:** Python, SQL, PySpark, Java, Apache Spark, Apache Kafka, Apache Airflow, AWS Glue, Kinesis, AWS (S3, Lambda, Redshift), Azure, Snowflake, Redshift, PostgreSQL, Oracle, SQL Server, MongoDB, Data Lakes, Data Warehouses, Star Schema, Snowflake Schema, Git, Jenkins, Terraform, Data Masking, Encryption, PCI-DSS, GDPR, HIPAA, AWS Glue, Apache Atlas, Datadog, Splunk, AWS CloudWatch, Scikit-learn (for predictive models)

**Big Data Engineer**

**Stryker – Cary, NC** | **November 2019 – August 2021**

* **Designed and implemented** scalable ETL pipelines using **Apache Spark**, **PySpark**, and **Hadoop** to process and transform large volumes of structured and unstructured healthcare data, such as patient records, clinical data, and medical imaging data, reducing processing time by **35%**.
* Developed and optimized **data ingestion frameworks** for ingesting data from **multiple sources** including **FHIR (Fast Healthcare Interoperability Resources)** APIs, **electronic health records (EHR)**, medical devices, and **IoT sensors** into a **centralized data lake** built on **Hadoop HDFS** and **AWS S3**.
* Designed and maintained a **data warehouse solution** using **Snowflake** and **Amazon Redshift** to support data analytics for clinical decision-making, patient risk scoring, and predictive analytics for patient outcomes.
* Integrated real-time **data streaming solutions** using **Apache Kafka** and **AWS Kinesis** to ingest live data from patient monitoring systems and medical devices, enabling real-time alerting for patient conditions and improving response times by **40%**.
* Implemented **machine learning pipelines** using **PySpark** and **MLlib**, enabling predictive modeling for patient risk stratification, readmission prediction, and fraud detection, leading to **a 30% reduction in readmission rates**.
* Migrated legacy **ETL processes to Azure Data Factory**, reducing operational maintenance and improving scalability.
* Integrated Azure Machine Learning service with Databricks for streamlined model training and deployment.
* Leveraged Azure Storage Queues to coordinate asynchronous data processing between healthcare systems.
* **Optimized SQL queries** and created **stored procedures** to perform large-scale analytical queries on healthcare data, improving query performance by **25%** and ensuring reliable access to key health metrics.
* Led the migration of legacy data processing systems to a **cloud-based architecture** on **AWS**, leveraging services like **EMR**, **Glue**, and **Redshift**, improving processing speed and scalability while reducing operational costs by **20%**.
* Developed a **data quality framework** that automatically validated incoming patient data to ensure data integrity, reducing errors in clinical datasets by **15%** and improving the quality of patient information used in clinical decision-making.
* Implemented **data security measures** to comply with **HIPAA** and **GDPR**, including **encryption**, **data masking**, and **access control** policies to protect sensitive healthcare data.
* Worked closely with healthcare data scientists and business analysts to identify key business requirements, translating them into actionable data solutions to improve patient care and operational efficiency.
* Built custom logging mechanisms using **Azure Application Insights** to monitor pipeline behavior and detect anomalies.
* Participated in disaster recovery planning using Azure Backup and **geo-redundant storage solutions**.
* Conducted cost analysis and optimization across Azure resources using Azure Cost Management and Billing.
* Collaborated with cross-functional teams including **data scientists**, **product managers**, and **clinical experts** to support the development of new healthcare products and services.
* Led and mentored a team of junior data engineers, sharing knowledge on **Big Data technologies**, best practices, and improving team efficiency and code quality.

**Technologies Used:** Python, SQL, PySpark, Java, Apache Spark, Hadoop (HDFS), Apache Kafka,Apache Airflow, AWS (S3, EMR, Redshift, Glue), Azure, Snowflake, Redshift, Apache Spark, PySpark, PostgreSQL, MySQL, MongoDB, MLlib (Apache Spark), Scikit-learn, HIPAA Compliance, GDPR, Data Masking, End-to-End Encryption, Datadog, Data Lakes, Data Warehouses, Star Schema, Git, Jenkins, Terraform

**Data Analyst**

**Crayon Data – Chennai, IND** | **July 2017 – June 2019**

* **Analyzed large datasets** using **Excel**, **SQL**, and **Python** to derive actionable insights for business decision-making in the retail sector, improving sales forecasting accuracy by **20%**.
* **Developed and maintained automated reports** to track key performance indicators (KPIs) like customer acquisition, churn rates, and product performance, reducing manual reporting time by **40%**.
* **Performed data cleansing** and validation using **Python (Pandas)**, ensuring high data quality and integrity by identifying and rectifying inconsistencies in customer and sales data.
* Created and implemented **data visualization dashboards** using **Tableau** and **Power BI**, allowing the marketing and sales teams to quickly assess performance trends and make data-driven decisions.
* Collaborated with **business analysts** and **stakeholders** to identify business requirements and translated them into data models for deeper insights on customer segmentation and purchasing behaviors.
* Conducted **market basket analysis** to identify product associations and cross-sell opportunities, which resulted in a **10% increase in sales** for bundled product offerings.
* Developed and tested **SQL queries** to extract data from relational databases (SQL Server and MySQL), reducing report generation time by **30%** and improving data retrieval efficiency.
* Utilized **Excel (VLOOKUP, Pivot Tables, and Macros)** to perform detailed statistical analysis and trend forecasting, helping the product team optimize inventory management and reduce stock-outs.
* Conducted **ad-hoc analyses** to support business initiatives, including customer behavior analysis and pricing strategy optimization.
* **Collaborated with cross-functional teams** to support data-driven initiatives such as new product launches, marketing campaigns, and loyalty programs, ensuring alignment with business goals.
* Assisted in the creation of **predictive models** using **Excel** and **R** for sales predictions and trend analysis, helping the company prepare for seasonal demand fluctuations.

**Technologies Used:** Python, SQL, R, Tableau, Power BI, Excel (Pivot Tables, VLOOKUP, Macros), Pandas (Python), SQL Server, MySQL, R (for predictive modeling and trend analysis), Git

**Jr Data Analyst**

**Data Sutram – Mumbai, IND** | **September 2015 – June 2017**

* **Assisted senior analysts** in processing and analyzing large datasets from various business domains, including sales, customer demographics, and product performance, contributing to business insights that supported operational improvements.
* Developed and maintained **data entry templates** and standard reports using **Excel** (Pivot Tables, VLOOKUP, Charts), which streamlined data collection and reporting processes, reducing manual work by **25%**.
* **Performed data validation and cleansing** using SQL queries to identify and correct inconsistencies and inaccuracies in raw data, ensuring high-quality data for further analysis.
* Supported the **data visualization team** in creating simple **dashboards and reports** in **Excel** and **Tableau**, allowing the management team to easily track sales performance and product trends.
* **Extracted and processed transactional data** from relational databases (SQL Server) and spreadsheets, providing support for ongoing analysis of customer behavior and market trends.
* Contributed to creating automated reports for **daily sales metrics** and **customer feedback**, reducing the turnaround time for key business performance updates.
* Assisted in **customer segmentation analysis** using **Excel** to identify high-value customers and trends in purchasing behavior, helping the marketing team target promotional campaigns more effectively.
* Supported **inventory data analysis** to identify stock levels and demand forecasts, helping reduce overstocking and stockouts in retail locations.
* Conducted basic **trend analysis** on customer and sales data, providing insights into seasonal buying patterns and assisting senior analysts with preparing forecasts.
* **Collaborated with business units** to gather requirements for ad-hoc reports, ensuring that the data solutions met business needs for decision-making.
* Helped create **SQL-based reports** for the team, learning the ins and outs of **SQL queries** to retrieve key performance data for weekly business reviews.
* Worked with other data analysts to help maintain the **data repository**, ensuring data was properly stored, updated, and backed up for future analysis.

**Technologies Used:** SQL, Excel, Tableau, SQL Server, Excel (Pivot Tables, VLOOKUP, Charts), Basic Statistical Functions, Git (for team collaboration on projects)