Saroj Kapali

Senior Data Engineer/Scientist

Dallas, Tx | sarojkapali2024@gmail.com | (832) 280-7615

Professional Summary

Senior Data Engineer/Scientist with six years of experience in designing, developing, and optimizing scalable data pipelines, end-to-end **ETL processes**, and **cloud-based solutions** across **AWS, Azure, and GCP**. Expertise in **Infrastructure as a Service (IaaS)**, **Platform as a Service (PaaS)**, and **Software as a Service (SaaS)** for building and deploying cloud-native data solutions. Skilled in **Apache Spark (PySpark, Spark-SQL, Spark Streaming), Snowflake, Databricks, and BigQuery** for high-performance data processing and analytics. Proficient in **workflow orchestration (Airflow, dbt, Informatica), CI/CD automation (Jenkins, Terraform, CloudFormation)**, and **API development (Python, Java, R)**. Strong background in data modeling **(Star & Snowflake schemas), metadata management (Azure Purview, Looker), and BI tools (Power BI, Tableau, Looker, SAP)**. Adept at collaborating with cross-functional teams and leveraging machine learning **(TensorFlow, PyTorch, Spark ML, advanced statistics and mathematics)** to drive AI-driven insights. Proven ability to optimize **big data processing, migration to cloud technologies, and distributed computing** while delivering high-impact data solutions in Agile environments.

Professional Experience

Diverge Health, Glencoe, IL **March 2022 – Current**

Senior Data Scientist/Engineer

Responsibilities

- Collaborated with stakeholders to gather business requirements and design **scalable data solutions**, enhancing data-driven decision-making for **customer analytics and operational efficiency**.
- Led a **cross-functional team** of **data engineers, data scientists, machine learning engineers, and data analysts**, developing end-to-end **data science workflows** for predictive analytics, customer segmentation, and anomaly detection.
- Developed Spark applications using **Spark-SQL** in **Azure Databricks**, processing high-volume transactional data from billing and customer systems to enable real-time insights into customer behavior.
- Designed and implemented data pipelines using **Azure Data Factory**, reducing ETL execution time by 50% for financial data processing workflows.
- Developed predictive models using **machine learning** techniques, such as **Random Forest, Gradient Boosting, and Neural Networks,** to identify at-risk patient populations, improving early intervention effectiveness.
- Conducted **exploratory data analysis (EDA)** and applied statistical methods to extract actionable insights, leading to enhanced healthcare service delivery.
- Built **NLP** pipelines for clinical text analysis to automate extraction of medical terminologies, reducing manual effort by 40%.
- Implemented and evaluated **A/B testing** frameworks to assess new healthcare interventions, increasing patient engagement rates by 25%.
- Created interactive dashboards using **Tableau and Power BI** to visualize healthcare data, facilitating strategic decisions by senior management.
- Led anomaly detection projects employing unsupervised learning algorithms, significantly improving accuracy in detecting healthcare fraud.
- Conducted feature engineering and selection on large datasets, boosting predictive model performance by 20%.
- Utilized Deep Learning techniques such as **CNNs and RNNs** for medical image and time-series analysis, enhancing diagnostic precision.
- Collaborated cross-functionally with clinicians and technical teams to design and validate clinical predictive tools, achieving improved patient outcome predictions.
- Managed and maintained comprehensive documentation of model development lifecycle, ensuring transparency and regulatory compliance within healthcare analytics.
- Built real-time data streaming pipelines using **Spark Streaming** and **Kafka** on **Azure HDInsight**, enabling event-driven data architecture for fraud detection and network anomaly monitoring.
- Optimized **Snowflake OLAP** and **OLTP data models** on **Azure**, improving query performance by 45% for customer analytics dashboards, enabling faster insights into customer churn and engagement.
- Migrated legacy **MapReduce** programs to **Spark** using **Scala** and **PySpark**, reducing processing time by 60% for network log analysis workflows.
- Automated machine learning model retraining using **Apache Airflow** on **Azure**, streamlining the deployment of predictive models for customer lifetime value (CLV) and revenue forecasting.

- Integrated **TensorFlow** and **Scikit-learn** models into data pipelines, using team collaboration on **Azure ML** for predictive maintenance of network equipment.
- Created interactive dashboards and reports using **Power BI**, enabling business users to analyze customer behavior, operational KPIs, and financial performance in real time.
- Led the migration of legacy reports from **OBIEE** to **Power BI**, improving reporting performance by 35% and enabling self-service analytics for finance and marketing teams.
- Built real-time data streaming pipelines using **Spark Streaming** and **Kafka** on **AWS MSK**, achieving sub-second latency for customer engagement analytics.
- Optimized **Redshift** data models on **AWS**, improving query performance by 45% for sales and revenue tracking dashboards.
- Automated machine learning model retraining using **AWS Step Functions**, streamlining predictive model deployment for network capacity planning.
- Implemented **CI/CD workflows** using **Terraform** and **AWS CloudFormation** for **AWS**, and **Azure DevOps** for **Azure, integrating version control to enable automated and seamless deployments** for **data science applications and machine learning models.**
- Developed AI-powered financial analysis pipelines using **PL/SQL** stored procedures on **Azure SQL Database** and **AWS RDS**, improving revenue forecasting accuracy by 30%.

American Airlines, Dallas, TX **June 2020 – February 2022**
Data Engineer
Responsibilities

- Collaborated with **stakeholders** to gather business requirements and design scalable data solutions, enhancing data-driven decision-making for **flight operations, customer analytics, and revenue management.**
- Developed Spark applications using **Spark-SQL** in **AWS EMR**, processing high-volume transactional data from booking and customer systems, enabling real-time insights into customer behavior and flight performance.
- Designed and implemented data pipelines using **GCP Dataflow**, reducing **ETL execution time** by **48%** for flight operations and revenue **data processing workflows**.
- Built real-time data streaming pipelines using **Spark Streaming** and **Kafka** on **AWS MSK**, enabling event-driven architectures for flight delay prediction and baggage tracking.
- Optimized **BigQuery** data models on **GCP(Google Cloud Platform)**, improving query performance by **42%** for customer analytics dashboards, enabling faster insights into customer satisfaction and loyalty.
- Migrated legacy **MapReduce** programs to **Spark** using **Scala** and **PySpark**, reducing processing time by **55%** for flight log analysis and maintenance scheduling **workflows**.
- Automated machine learning model retraining using **Apache Airflow** on **GCP**, streamlining the deployment of predictive models for flight demand forecasting and dynamic pricing.
- Integrated **TensorFlow** and **Scikit-learn** models into data pipelines, deploying them on **AWS SageMaker** for predictive maintenance of aircraft systems.
- Created interactive dashboards and reports using **Tableau** implementing advanced data analytics, enabling business users to analyze flight performance, **operational KPIs**, and revenue trends in real time.
- Led the migration of legacy reports from **OBIEE** to **Tableau**, improving reporting performance and enabling self-service analytics for operations and finance teams.
- Designed and maintained automated data pipelines using **Azure Data Factory**, reducing **ETL** execution time for crew scheduling and payroll data.
- Built real-time data streaming pipelines using **Spark Streaming** and **Kafka** on **GCP Pub/Sub**, achieving sub-second latency for customer engagement analytics and personalized offers.
- Optimized **Redshift** data models on **AWS**, improving query performance for revenue tracking and sales dashboards.
- Automated machine learning model retraining using **AWS Step Functions**, **streamlining predictive model deployment** for route optimization and fuel efficiency.
- Implemented CI/CD workflows using **Terraform** and **AWS CloudFormation** for **AWS**, **GCP Deployment Manager** for **GCP**, and **Azure DevOps** for **Azure**, ensuring automated deployments for cloud-based data solutions.
- Managed infrastructure using **Kubernetes** on **GCP GKE** and **AWS EKS**, enabling containerized deployment of data applications.
- Developed AI-powered revenue analysis pipelines using **PL/SQL** stored procedures on **AWS RDS** and **GCP Cloud SQL**, improving revenue forecasting accuracy by **32%.**

- Designed and executed data governance protocols with **GCP** and **AWS**, enhancing oversight on **500+ data assets**, achieving 100% compliance during audits, and streamlining **metadata management** to boost team **productivity by 25%.**

Western Union, Milwaukee, WI **October 2018 – April 2020**
Data Engineer
Responsibilities:

- Gathered and analyzed business requirements to design **scalable data solutions** for processing over **500,000 daily transactions**, ensuring compliance with financial regulations.
- Developed and maintained ETL pipelines using **Apache Spark**, **PySpark**, and **Scala**, optimizing data transformation and reducing processing time by **40%**.
- Built and optimized real-time fraud detection pipelines utilizing **Kafka**, **Spark Streaming**, and **Apache Flink**, reducing false positives by **3 million cases annually**.
- Designed and implemented **AWS Data Pipelines** with **AWS Lambda**, **API Gateway**, **S3**, and **DynamoDB**, automating fraud detection for **200M+ monthly transactions**.
- Integrated **SAP ERP** and **Salesforce** data pipelines into **AWS Lambda** and **Databricks**, enabling real-time financial reporting across **150+ global business units**.
- Developed and automated **AML (Anti-Money Laundering) monitoring systems** using **Apache Airflow**, **Snowflake**, and **SQL**, ensuring compliance with **200+ global banking partners**.
- Optimized **data warehouse schemas** using **Star Schema** and **Snowflake Schema**, reducing **query execution times from 45 minutes to under 5 minutes** for financial analysts.
- Created interactive **self-service financial analytics dashboards** using **Looker** and **Power BI**, allowing **1,500+ finance professionals** to monitor key revenue and risk metrics.
- Automated **currency exchange rate reconciliation** across **150+ global remittance corridors**, improving forex accuracy and reducing pricing discrepancies by **20%**.
- Led the **migration of business intelligence reports** from **OBIEE** to **Power BI**, improving report generation speeds from **10 minutes to 2 minutes** for **5,000+ daily users**.
- Enhanced **ETL workflows** for real-time data ingestion and regulatory compliance reporting, ensuring **100% adherence to SOX, PCI DSS, and GDPR standards**.

Technical Skills

Programming & Data Manipulation: Python, Java, Scala, R, SQL (T-SQL, PL/SQL), C++, C#, Bash, Shell Scripting, ASP.NET, Pandas, NumPy, Matplotlib, JSON, XML

Machine Learning & Analytics: TensorFlow, PyTorch, Scikit-learn, Spark ML, BigQuery ML, Predictive Analytics, NLP, Neural Networks, Random Forest, Gradient Boosting

Data Engineering & Big Data: Apache Spark, PySpark, Hadoop, Hive, HDFS, MapReduce, Apache Flink, Apache Airflow, dbt, Informatica, ETL Pipelines, Dimensional Modeling

Cloud & Infrastructure: AWS (EC2, S3, EMR, Glue, Redshift, Step Functions, CloudWatch, MSK, Kinesis), Azure (Data Factory, Synapse, Data Lake, DevOps), GCP (BigQuery, GCS, Dataproc, Composer, Pub/Sub), Snowflake, Databricks, Docker, Kubernetes, Terraform, Jenkins, Git, Prometheus, Datadog, CI/CD

Visualization, Tools & Management: Power BI, Tableau, Looker, Amazon Quicksight, Google Data Studio, Grafana, Excel, Google Analytics, SAP, Radius, RESTful APIs, Azure Purview, AtScale, Jira, ServiceNow, Agile (SCRUM), SOX Compliance

Education

University of Houston
Bachelor of Science in Computer Science, Minor in Mathematics