Saikumar Gunda

Data Engineer

901-632-7334 | saikumargunda30@gmail.com | linkedin.com/in/saikumargunda/

SUMMARY

Experienced Data Engineer with 3+ years of hands-on experience designing, building, and optimizing high-performance, cloud-native data solutions across healthcare and enterprise IT domains. Specializes in developing real-time ETL pipelines, big data architectures, and predictive analytics leveraging Apache Spark, Kafka, Airflow, and AWS/Azure platforms. Expertise in Python, PySpark, SQL, and Power BI, with a strong emphasis on ensuring data quality, enforcing regulatory compliance (HIPAA), and enhancing operational efficiency. Proven track record of driving data strategy, automating workflows, and enabling scalable insights through modern data engineering practices and machine learning integration.

SKILLS

- **Data Engineering & Workflow Orchestration:** Apache Spark (Scala, PySpark), Apache Kafka, Apache NiFi, Apache Flink, Apache Airflow, Control-M, AWS Glue, Microsoft SSIS, Talend, Informatica PowerCenter, Apache Beam, StreamSets
- **Big Data & Storage Technologies:** Hadoop HDFS, Amazon S3, Delta Lake, Apache HBase, Apache Cassandra, Azure Data Lake Storage, Snowflake, Teradata, Oracle, MySQL, PostgreSQL, Google Cloud Storage, Apache Parquet, Apache ORC
- **Cloud Platforms & Services:** AWS (EMR, SageMaker, Kinesis, QuickSight, Lambda), Microsoft Azure (Data Factory, Logic Apps, Azure Databricks, Synapse Analytics, Azure Functions), Google Cloud Platform (BigQuery, Dataflow, Pub/Sub)
- **Programming & Query Languages:** Python, Scala, SQL (T-SQL, PL/SQL), Shell Scripting, Java
- **Containerization, Infrastructure & CI/CD:** Docker, Kubernetes, Terraform, AWS CloudFormation, Azure ARM Templates, Git, Jenkins, GitLab CI/CD
- **Analytics, BI & Machine Learning:** MLflow, SageMaker, Scikit-learn, TensorFlow, Tableau, Power BI, Grafana, Kibana, ELK Stack
- **Data Governance & Security:** HIPAA Compliance, Apache Atlas, AWS IAM, PII Masking, Data Encryption, Data Quality Management, Collibra

EXPERIENCE

CVS Health, USA — Jun 2024 – Present

Data Engineer

- Engineered scalable ML pipelines using Apache Spark and Databricks to predict 30-day patient readmission risks, integrating EHR and claims data from Amazon S3 and Redshift.
- Developed real-time data ingestion workflows with Apache NiFi and Kafka, enabling seamless processing of structured and semi-structured healthcare datasets for predictive modeling.
- Implemented real-time prescription monitoring by streaming transactional and IoT data from pharmacies using Apache Kafka and AWS Kinesis, detecting anomalies indicative of fraud or supply chain issues.
- Leveraged Apache Flink and Spark Streaming to process high-velocity pharmacy data, facilitating real-time analytics and alert generation for potential drug abuse and compliance breaches.
- Constructed a unified Member 360 data platform by integrating pharmacy, insurance, and retail data through ETL pipelines using Apache NiFi and Talend, ensuring data quality and regulatory compliance.
- Orchestrated data workflows with Apache Airflow and Azure Data Factory, transforming and loading data into Snowflake and Azure Synapse to support business intelligence and personalization engines.
- Established data governance frameworks utilizing Collibra and Apache Atlas, enhancing data lineage tracking, metadata management, and compliance with HIPAA and other privacy regulations.

Tata Consultancy Services, India — Jul 2021 – Jul 2022

ETL Developer

- Designed and implemented end-to-end ETL workflows using Apache NiFi, Talend, SSIS, and Azure Data Factory to onboard and standardize healthcare provider and insurance enrollment data.
- Extracted data from diverse formats and sources including CSV, XML, JSON, Parquet, and APIs, applying complex transformation logic and business rules to ensure data consistency and quality.
- Loaded cleansed and enriched data into MySQL, PostgreSQL, Azure Data Lake, and Synapse Analytics, enabling accurate provider directory services and eligibility tracking.
- Developed and maintained data validation, deduplication, and enrichment processes using Python, SQL, and T-SQL to support compliance with industry data standards.
- Orchestrated and monitored complex workflows with Apache Airflow and Azure Data Factory pipelines, ensuring timely and reliable data delivery for billing and claims processing.
- Collaborated with cross-functional teams to align data pipeline outputs with regulatory, credentialing, and reporting requirements, improving operational efficiency and data governance.
- Built scalable, reusable data integration frameworks supporting high-volume, multi-source data ingestion across cloud and on-premise environments, optimizing performance and maintainability.

Vmware, India — Aug 2020 – May 2021

Data Engineer

- Engineered scalable ETL pipelines using Apache Spark (PySpark/Scala) and Apache Flink to process real-time telemetry data from VMware vCenter and ESXi hosts for capacity planning and resource optimization.
- Designed and implemented robust data ingestion pipelines leveraging Apache Kafka and vRealize Operations APIs to stream CPU, memory, and disk metrics from thousands of VMware VMs.
- Optimized performance data aggregation and forecasting on AWS EMR and HDFS, storing time-series metrics in Amazon S3 and Apache HBase for high-throughput analytics.
- Automated end-to-end data workflows with Apache Airflow, enabling scheduled extraction, transformation, and loading of infrastructure telemetry across VMware Cloud on AWS.
- Delivered actionable insights through real-time dashboards in Grafana and interactive visualizations in Tableau, enhancing VM performance monitoring and predictive alerting.

EDUCATION

University of Memphis, TN, USA — **May 2024**

Masters in Data Science

JNTUH, Hyderabad, India — **Jun 2021**

Bachelorsin Electronics & Communication Engineering

CERTIFICATION

- **Microsoft Certified: Azure Data Engineer Associate**
- **Google Certified: Associate Cloud Engineer**
- **Microsoft Certified: Power BI Data Analyst**