Sai Jaswanth Kumar Kunku
Senior Data Engineer
Pleasanton, CA
jaswanth0331645@gmail.com
www.linkedin.com/in/jaswanth45
+1(240)-986-2599

SUMMARY

Senior ML/Data Engineer with 9+ years of experience building scalable **batch and streaming** data pipelines across retail, finance and healthcare. Skilled in Spark, Kafka, Flink, Snowflake, Airflow, and **modern lakehouse/warehouse architectures** with strong SQL, **data modeling, and optimization** expertise. Designed end-to-end ingestion, transformation, governance, and analytics ecosystems using big-data frameworks, MDM/ETL tools, and cloud-native services. Delivered reliable, compliant, and high-performance data platforms supporting real-time processing, BI reporting, and enterprise-scale analytics.

TECHNICAL SKILLS

Programming & Scripting Languages: Python, SQL, KQL
Databases: DB2, Oracle, MS SQL Server, MongoDB, Cassandra
Data Warehousing: Snowflake, Amazon Redshift, Google BigQuery, Azure Synapse Analytics
Big Data: Spark, Kafka, Flink, Airflow, Databricks
ETL/ELT and Orchestration: Informatica, IBM DataStage, DBT, Fivetran, Azure Data Factory
Data Visualization & Monitoring: Tableau, Power BI, Alteryx, Splunk, Elasticsearch
Cloud Platforms: Amazon Web Services (AWS), Microsoft Azure, Google Cloud Platform (GCP)
CI/CD & IaC: Git, GitHub Actions, Docker, Kubernetes, Terraform, Databricks Asset Bundles
GenAI & ML: LLMs, Prompt engineering, RAG, Text embeddings, Vector databases
SDLC & Testing: Agile, Waterfall, A/B testing, unit testing
IDEs & Productivity Tools: Visual Studio Code, Github Copilot, MS Office, Lucid Chart, JIRA, Confluence

CERTIFICATIONS

Databricks Certified Data Engineer Professional
Microsoft Fabric Data Engineer (DP-700)
IBM Big Data Engineer

PROFESSIONAL EXPERIENCE

Albertsons Companies — **Pleasanton, California**
Data Engineer — **Mar 2024 – Present**

Responsibilities:

- Designed and **maintained** scalable **Databricks** ETL and ELT pipelines supporting **10+ critical ML models** in business growth area across **retail, supply chain, and merchandising domains.**
- **Translated Snowflake queries/stored procedures** from **Snowflake** to **GCP BigQuery** for **100+ tables and views** without affecting business logic.
- Resolved **data type mismatches**, and optimized **joins, partitioning, and clustering**, reducing query runtimes by **20 to 40 percent**.
- Tuned PySpark and SQL jobs using **shuffle optimization, caching, and broadcast joins**, improving **pipeline throughput,** and stabilizing **large-scale batch executions**.
- **Optimized BigQuery storage usage** by auditing and decommissioning more than **2700 obsolete tables**, reclaiming **1.1 Petabytes,** and reducing **monthly cloud costs** by **12000$**.
- Developed Python **automation scripts** to extract and track **model-related tables and views**, streamlining **data migration tracking** and reducing manual effort during model cutovers.
- Upgraded **Databricks Workflows** and **Job API** from version 2.0 to 2.1, enabling parallel task execution and reducing workflow runtime by 30 to 50%.

- Led Azure to Google Cloud Databricks migrations for ML models, including **notebook refactoring, data validation, and production deployment**, achieving zero post deployment rollbacks.
- Built and maintained **CI/CD** pipelines using **GitHub Actions and Databricks asset bundles**, automating deployments and reducing **manual release efforts**.
- **Troubleshot and resolved production issues** such as **timeouts, schema drift, and intermittent failures**, reducing recurring job failures by more than 60 percent.
- **Authored runbooks, validation playbooks**, and migration documentation improving **on call readiness** and accelerating **root cause analysis**.
- Built interactive monitoring dashboards in **Grafana** by integrating **KQL**-based queries, enabling real-time visibility into pipeline health, job runtimes, and system metrics
- Developed a **text to SQL** proof of concept using BigQuery metadata enrichment and large language models, improving query generation quality for business users.
- Contributed to **Databricks to Dataproc migration** by adapting PySpark jobs and building **Cloud Composer, Vertex AI** workflows for orchestration and scheduling.
- Implemented real-time model serving POC using **Mosaic AI Model Serving** and **Google Cloud Run** to benchmark latency, autoscaling, and reduce infrastructure costs and achieved **<15** RPS.

Environment: Python, Snowflake, BigQuery, CloudRun, Vertex AI, Databricks, Delta Lake, Docker, AKS, Azure SQL Grafana, Cloud Composer

Amazon **Hilliard, OH**
Data Engineer **Jun 2022 – Feb 2024**

Responsibilities:

- Orchestrated **end-to-end data pipelines** by systematically collecting, loading, and transforming sales, product, customers and other supply chain data from enterprise databases into Delta Lakes for **BI** and **Machine Learning**.
- Utilized **AWS DMS** to implement capture data capture in **Aurora DB** tables and stream the new data into **Kinesis Data Streams.**
- Created optimized buckets with dynamic **partitioning**, and **compression** mechanisms and developed **Lambda step functions** to seamlessly ingest data from **Kinesis Firehose** into the **S3 raw bucket zone**.
- Created **Glue ETL jobs** and **Glue Crawlers** in **Glue Studio** to perform batch processing and ingest the transformed data into the **S3 intermediate zone** for ad-hoc analysis using **AWS Quick Sight**.
- Leveraged **Amazon EMR** to perform distributed **data transformations** and aggregations using **Hive** and **Spark**, uncovering valuable insights from the stored data.
- Created **Spark SQL** scripts in **Apache Spark** for robust data cleansing, transformation, and filtering operations on Hive tables to export data in **Parquet** format with snappy compression to S3 for efficient read/write processing.
- Developed **Hive** User-Defined Functions (**UDFs**) written **HQL queries** to create **external** tables on **S3** with **partitions** and **bucketing** techniques for ad-hoc analysis.
- Conducted **meetings** with business stakeholders, clients, and director-level members for requirement gatherings and presented various solutions using best practices.
- Performed data transformations using **DBT** to develop data marts for the Machine learning team to create models and improve product recommendations.
- Instantiated, created, and maintained **containerization** mechanism with continuous integration and deployment of DevOps pipelines by applying automation applications using **GIT**, **Terraform**.
- Monitored data pipelines and resolved issues evolved with data throughput using **CloudWatch logs** and **Cloud Trail log**s.
- **Optimized data pipelines** and ensured data quality and integrity throughout the migration process, leveraging tools like AWS EMR and Spark SQL for efficient data movement and computation.

Environment: Python, Amazon EMR, Amazon DMS, DynamoDB, CI/CD, Linux, Amazon Kinesis, AWS QuickSight, AWS Glue, Amazon Redshift, Spark SQL, Git

Citizen's Bank **Fairfax, VA**
Data Engineer **Dec 2021 – May 2022**

Responsibilities:

- Developed scalable ELT pipelines using **Medallion Architecture** (Bronze, Silver, Gold) in **Microsoft Fabric**, migrating accounts, transactions, and mortgage data from enterprise warehouses into OneLake.
- Designed **logical and physical data models** using Erwin, enabling seamless migration from DB2 to **Fabric Lakehouse and Warehouse**, ensuring data consistency and integrity.
- Built **data ingestio**n pipelines using **Data Factory** in Fabric, creating parameterized pipelines and linked services to dynamically load on-prem DB2 data into **OneLake**.
- Implemented **incremental and bulk data loading strategies** using COPY INTO and Fabric-native ingestion patterns and built curated **gold layer data marts** for analytics and reporting teams.
- Leveraged **Fabric Real-Time Analytics (KQL)** to monitor pipeline performance, analyze EventHub streams, and proactively detect data inconsistencies and failures.
- Developed **event-driven ingestion pipelines** using **Event house** and Fabric pipelines to process streaming data into Delta tables within the Lakehouse.
- Built and optimized **PySpark notebooks in Fabric** for large-scale transformations, data quality validation, type casting, and enrichment to create optimized silver and gold datasets.
- Used Delta Lake tables in **Fabric Lakehouse** for efficient storage, ACID compliance, and faster downstream querying.
- Implemented **Auto Loader–like incremental ingestion patterns** for processing new files in OneLake with minimal latency.
- Optimized performance by leveraging **partitioning strategies, file sizing, and query optimization techniques** within Fabric Warehouse and Lakehouse.
- Collaborated with cross-functional teams to build **data APIs and integration layers**, enabling seamless data access across internal systems.
- Deployed and promoted pipelines across environments using **CI/CD and Azure DevOps** ensuring consistent and reliable releases.

Environment: Python, CI/CD, Snowflake, Cosmos DB, ADF, Power BI, Microsoft Fabric, Apache Spark, KQL

Tata Consultancy Services **Hyderabad, India**
Big Data Engineer **Apr 2019 – Aug 2021**

Responsibilities:

- Offered valuable analytical support for the Claims, Ancillary, and Medical Management departments, contributing to data-driven decision-making and process improvements.
- Loaded data from microservices to **Google Cloud Storage** using **IBM DataStage** and performed structural modifications using **Spark** in **Google Data Proc**.
- Implemented change data capture pipelines (**CDC**) using **Google Pub/Sub** with **Debezium Connector** to process data from **MSSQL** using **Google Cloud functions.**
- Configured **Pub/Sub** topics with **Spark Streaming API** in **Google Compute Engine** to fetch near real-time data from multiple sources, such as weblogs into Delta Lake tables for timely analysis and actionable insights.
- Ingested billions of **claim records** into the spark cluster and applied transformations and aggregations, resulting in a 30% reduction in data processing time.
- Designed and implemented **Snowflake** database schemas architecture, tables, and views to accommodate structured and semi-structured data.
- Designed and implemented **custom visualizations and dashboards** using **Looker** advanced visualization capabilities, providing stakeholders with intuitive and actionable insights.
- Developed and implemented **risk adjustment models** for Medicare data, ensuring accurate and compliant reporting.
- Collaborated closely with **Quality Control** Teams to develop comprehensive Test Plans and Test Cases, ensuring rigorous testing of systems and validating data accuracy.
- Used **Matplotlib** and **Seaborn** in Python to visualize the data and perform **feature engineering** to detect outliers and perform normalization.
- Written **Spark** applications in **Scala** to interact with the MSSQL database using Spark SQL Context to access Hive tables.

- Developed supervised ML models with hyperparameters using **Spark MLlib** to identify different kinds of fraud in **Medicare claims**.
- Developed dashboards and visualizations to help business users analyze data and provide insight to business using **SQL Server Reporting Services (SSRS**) and **Power BI**.

Environment: Python, Scala, BigQuery, Spark, Dataproc, Power BI, IBM DataStage, SQL Server, Google Dataflow

Brainy n Bright Inc. **Hyderabad, India**
Data Engineer **Nov 2017 - Mar 2019**

Responsibilities:

- Developed **SSIS** packages and maintained **SQL server agents** to perform **initial loads** and **full loads** into cloud storage.
- Written Map reduce code to transform unstructured data into Structured Data into Hive Tables Using Java for querying.
- Loaded the data from CSV, JSON, XML data sources into **HDFS** using **Sqoop** and integrated them into Hive tables.
- Used **Pandas, NumPy, Seaborn, SciPy, Matplotlib, Scikit-learn, and NLTK** in **Python** for implementing various machine learning algorithms.
- Used **Spark-SQL** to Load JSON data and create Schema RDD, loaded it into Hive Tables, and handled structured data using Spark SQL.
- Designed and implemented data pipelines using **ThoughtSpot** capabilities to process and transform raw data into actionable insights, optimizing for speed and accuracy.
- Executed **scheduled tasks** for weekly and monthly data updates while adeptly managing and manipulating the data for efficient database management.
- Designed dashboards utilizing **SSAS** and **SSRS** to construct matrixes as well as tabular reports using reporting services.

Environment: Python, CI/CD, Oracle, Hadoop, MongoDB, SSIS, SSAS, SSRS, Spark, MySQL, MongoDB

LogicMatter Inc. **Hyderabad, India**
ETL Developer **Sept 2016 - Nov 2017**

Responsibilities:

- Extracted, transformed, and loaded **Salesforce** data (Opportunities, Accounts, Leads, Contacts, Tasks, Events, etc. using **Informatica PowerCenter,** enabling seamless migration between Salesforce, Teradata, Oracle, and SQL systems.
- Built ETL and Data Quality mappings, performed deep data profiling, and automated validation/deduplication to improve SFDC data accuracy and reliability.
- Designed **logical and physical data models** using ER diagrams, class diagrams, and Erwin Data Modeler to map source–target schemas and support smooth migration and integration projects.
- Developed complex SQL (materialized views, CTEs, nested queries, case logic) and **created Power BI reports** with KPIs and advanced DAX for ad-hoc analytics.
- Prepared **business requirement documentation** and delivered end-user training to ensure successful adoption of new data workflows and reporting solutions.

Environment: MS Excel, AWS, Oracle, DB2 , Power BI, ERWIN Data Modeler, Informatica, UNIX, Salesforce

EDUCATION

George Mason University, Masters in Data Analytics Engineering **Fairfax, VA**
GITAM University, Bachelor of Technology in Computer Science **Visakhapatnam, India**