BABY SUSHMA

510-578-8162 | vunnambabysushma2000@gmail.com | LinkedIn

SUMMARY

Data Engineer with 3+ years of experience building large-scale Spark-based data pipelines using Java, PySpark, and SQL. Strong background in batch and streaming systems, ETL/ELT design, data quality frameworks, and performance optimization across cloud platforms (Azure/AWS). Proven ability to deliver reliable, high-throughput pipelines processing 100M+ records daily and supporting enterprise analytics and ML workloads.

WORK EXPERIENCE

Walmart Global Tech **Dec 2024 - Present**

Data Engineer *Sunnyvale, CA*

- Built a real-time marketplace data pipeline to ingest and process product metadata, pricing discounts, updates, deletions, and new offer-based item additions, enabling accurate downstream pricing and catalog analytics. Developed Java-based Kafka consumers using CCM profiles, implementing robust event handling, schema validation, and state management for high-throughput streaming workloads.
- Designed and implemented a large-scale batch data processing framework consisting of 7 interconnected Spark pipelines, handling 600M+ records per run for enterprise analytics workloads.
- Owned and scaled enterprise Spark pipelines (Java, Scala, SQL) processing over 100 million records per day, which improved data reliability and speed for supply chain, inventory, and customer analytics, enabling teams to make faster and more informed business decisions.
- Designed and implemented high-throughput batch and near–real-time ingestion systems handling 50M+ events/day, enabling near-real-time decision-making for merchandising and operations teams.
- Delivered 80% improvement in Spark job performance by tuning Adaptive Query Execution (AQE), optimizing partitioning strategies, caching, and broadcast joins at scale.
- Engineered 20+ reusable, production-grade Spark UDFs powering pricing, replenishment, forecasting, and personalization use cases with 99.95% validated data accuracy.
- Built metadata-driven ETL workflows on the Harmony platform, reducing pipeline configuration and onboarding time by 60% across analytics teams.
- Developed automated data quality and reliability frameworks (schema enforcement, validation checks, anomaly detection), maintaining 99.9% pipeline reliability in production.
- Implemented end-to-end observability using Grafana and Prometheus, which improved SLA compliance by providing real-time monitoring and alerts, and reduced the time to detect and resolve data issues by enabling faster root cause analysis.

UCode Technologies LLC **Jun 2024 - Nov 2024**

Software Trainee (Intern) *Remote*

- Developed backend data processing services using Python (Flask/FastAPI) supporting analytics and automation workflows.
- Built ETL pipelines integrating MySQL, internal APIs, and preprocessing modules, reducing manual data preparation.
- Implemented robust validation, logging, and exception handling ensuring reliable, analysis-ready datasets.
- Supported CI/CD automation for backend deployment pipelines, improving release reliability.
- Contributed to Agile sprint execution, documentation, and cross-team collaboration.

Accenture Solutions Pvt Ltd **Feb 2021 - Jun 2022**

Associate Software Engineer *Bangalore, India*

- Designed and implemented ETL pipelines using Python, SQL, and Apache Spark to process enterprise client data from multiple sources including SAP systems, databases, and flat files.
- Built automated data ingestion workflows handling 10M+ daily records from heterogeneous sources (Oracle, SQL Server, CSV files), ensuring data quality and consistency across downstream systems.
- Developed data validation frameworks and quality checks, implementing schema validation, null checks, and business rule enforcement, reducing data quality issues.
- Created data transformation logic using PySpark and SQL for cleansing, aggregation, and standardization of client datasets, supporting analytics and reporting requirements.
- Optimized pipeline performance through partitioning strategies, caching mechanisms, and query optimization, improving processing speed.
- Collaborated with business analysts and client stakeholders to understand data requirements and translate them into technical specifications for pipeline development.
- Implemented monitoring and alerting mechanisms for data pipelines using custom Python scripts and logging frameworks, ensuring pipeline uptime.
- Supported data migration projects from legacy systems to modern data platforms, handling data mapping, transformation, and validation processes.
- Participated in Agile development cycles, code reviews, and documentation of data engineering processes and best practices.

TECHNICAL SKILLS

- **Big Data & Distributed Systems:** Apache Spark (PySpark, Scala, Java), Kafka, Hadoop/HDFS, Hive, Streaming Pipelines, Micro-Batch Processing, Harmony Platform (ETL/ELT, UDFs, UI Workflows), Metadata-Driven Orchestration
- **Databases & Storage:** Apache Cassandra, Azure SQL, MS SQL Server, PostgreSQL, MySQL, MongoDB, Delta Lake, Data Warehouses, Data Lakes
- **Programming Languages:** Java (8/17), Python, SQL, Scala, R, Bash/Shell
- **Frameworks & Libraries:** Flask, FastAPI, Spring Boot, Flask-Assets, Pandas, NumPy, Matplotlib, PyTorch (basics), MLflow, Maven, Lombok
- **Data Formats:** Parquet, JSON, ORC, CSV
- **Cloud Platforms:** Azure, GCP (GCS basics), AWS (Bedrock, EC2, S3)
- **Data Orchestration & ETL:** Airflow, Harmony Platform, Job Scheduling, Workflow Automation, Dependency Management, Data Quality/Validation, Schema Enforcement
- **Monitoring, Logging & DevOps:** Grafana, Prometheus, Splunk, Docker, Git, CI/CD (Concord, Maven, GitHub/GitLab)
- **Data Science & Decision Science:** Regression, Classification, Time-Series Forecasting, Clustering, Feature Engineering, Statistical Analysis, A/B Testing, Experiment Design, ML Pipelines, Model Evaluation
- **Supporting (AI/LLM):** Prompt Engineering, Claude, Titan, Sonnet, AWS Bedrock, Responsible AI (Foundational)

PROJECTS

End-to-End ETL Data Pipeline for Enterprise Analytics
George Mason University

- Built a scalable ETL pipeline using PySpark to ingest, clean, and aggregate 50M+ structured & semi-structured records.
- Designed Airflow DAGs for ingestion, validation, deduplication, schema checks, and automated notifications.
- Improved PySpark job performance by 45% using optimized partitioning, window functions, and UDFs.
- Built a reusable data quality framework including null checks, referential integrity, anomaly detection, and profiling.
- Modeled curated datasets into star & snowflake schemas supporting dashboards, KPI tracking, and ML feature pipelines.

Generative AI Chatbot with AWS (Supporting AI Project) **Jan 2024 - May 2024**
https://github.com/babysushma/GMU-MSSC-Chatbot
George Mason University *Fairfax*

- Developed a multi-model AI chatbot benchmarking Claude, Titan, and Sonnet on accuracy, reasoning quality, and summarization.
- Built structured prompt engineering templates (zero-shot, multi-shot, persona prompts) to evaluate LLM consistency.
- Designed evaluation metrics and comparison pipelines identifying Claude as the top-performing model.
- Presented findings to the CENTRL ITS team at George Mason University, supporting internal AI evaluation and adoption.
- Served as a supporting AI project complementing core data engineering & analytics curriculum.

EDUCATION

George Mason University – Fairfax, USA | *MS, Data Analytics Engineering* (GPA: 3.70 / 4.00)

SASTRA University – Tanjavur, India | *B.Tech, Electrical & Electronics Engineering* (GPA: 7.0 / 10)