Shashank Nampalli

Tampa, Florida | 470-455-7979 | svarma.de@gmail.com | LinkedIn

Summary

Senior Data Engineer with 10+ years of experience designing and operating large-scale, production-grade data pipelines and platforms. Skilled in Python, SQL and Apache Spark on cloud-native architectures (GCP and AWS), with a track record of optimizing ETL workflows, reducing latency, and ensuring reliability for high-volume workloads. Proven ability to collaborate with cross-functional teams to build scalable data infrastructures supporting analytics and ML use cases.

Experience

Walmart (via Virtuoso Info Systems) **Aug 2022 - Present**
Senior Data / ML Engineer *Sunnyvale, CA (Remote)*

- Designed and owned distributed Spark pipelines on Dataproc processing 30+ TB/day across multi-tenant workloads, reducing runtime by ~25% while meeting strict SLAs and establishing robust data ingestion pipelines.
- Architected a configuration-driven, multi-tenant recommendation platform supporting 20+ strategies (BAB, SI_VAV, RFU, Trending, Deals, Easy Reorder), reducing per-strategy development effort by ~60%.
- Built Spark-based ML training and batch inference pipelines in PySpark, integrating feature engineering and producing versioned prediction outputs for downstream serving layers.
- Implemented inline feature engineering to ensure training–scoring parity, reducing feature drift without dependence on an external feature store.
- Designed and productionized semantic similarity pipelines generating text and image embeddings using Sentence Transformers and CLIP, persisting vectors in Vespa DB for low-latency similarity search.
- Developed incremental and delta-processing frameworks eliminating full recomputation, reducing compute cost by ~35% while meeting sub-hour SLAs.
- Optimized BigQuery datasets using partitioning, clustering, and materialized views, cutting query cost by ~40% on multi-TB analytics workloads.
- Orchestrated end-to-end workflows with Airflow for data preparation, model execution, offline evaluation, artifact publishing, and recovery/rollback, improving pipeline reliability and introducing data observability for proactive monitoring.
- Implemented offline model validation metrics and versioned artifacts stored in GCS, enabling auditable runs and safe rollback.
- Generated batch prediction outputs in Parquet on GCS for efficient downstream consumption and loading into serving layers (e.g., Azure Cosmos DB).
- Established testing, data quality, and CI/CD standards (pytest, SonarQube, Concord), improving reliability and reducing production defects.
- Delivered an LLM-powered enhancement using Gemini that generated complementary queries, retrieved candidate items via internal search APIs, and reranked results with engagement signals, which increased query relevance and boosted user click-through rates
- Implemented LLMOps-style controls (prompt/config management, batching, concurrency limits, retries/fallbacks, and output validation) for reliable, cost-aware production runs.
- Developed and deployed Random Forest machine learning models in Python to accurately predict item repurchasability, enabling personalized reorder recommendations that increased reorder rates and drove substantial revenue growth by optimizing the shopping experience and promoting relevant items.
- Collaborated with product, data-science, and engineering teams to gather requirements and prioritize the AI project backlog, defining clear success metrics that enabled on-time delivery of two AI initiatives

Verizon (via Virtuoso Info Systems) **Jan 2019 - Aug 2022**
Senior Data Engineer *Tampa, Florida*

- Led end-to-end migration of Revenue Recognition data warehouse from Teradata to BigQuery on Google Cloud Platform, supporting analytics for 115M+ customers.
- Re-architected ETL and ELT pipelines handling ~3 TB per day (400–600M records/day) using cloud-native architectures, while meeting sub-2-hour revenue close SLAs.
- Reduced revenue close reporting latency from 12 hours to 4 hours through pipeline optimization and platform modernization.
- Built automated reconciliation and audit frameworks across data layers, improving data lineage and reducing defects by ~90% and manual validation time to under one hour.
- Optimized SQL execution using partitioning and clustering, reducing job runtime by ~35% and query costs by ~30%.
- Migrated AWS EMR MapReduce workloads to Apache Spark, improving performance and reducing compute costs.
- Served as technical lead mentoring 5–7 engineers and reducing production incidents to fewer than one per month.

IBM (via Server Management Systems) **Feb 2018 - Jan 2019**
Data Engineer *Bentonville, AR*

- Built and automated enterprise ETL pipelines that processed millions of records using SQL, Python, and Google Cloud Platform services such as BigQuery, ensuring reliable data flow for downstream analytics
- Delivered multi-platform data integration solutions generating over $12.5M in annual cost savings.

- Developed and maintained Teradata ingestion and transformation workflows with Apache Spark on EMR, supporting analytics teams and improving data availability for reporting

Iconix Labs Inc. **Jun 2017 - Feb 2018**
ETL Developer *Flower Mound, TX*

- Installed, configured, and maintained Oracle databases, including RAC environments on Linux, enabling reliable data access for users and decreasing database-related incidents
- Optimized database performance, backup, and upgrade procedures using indexing, partitioning, and automated scripts, which led to faster query response times and reduced system downtime
- Supported ETL workflows and enterprise data integration pipelines using Apache Spark, improving data accuracy and reducing delivery time by 30%, and collaborating with cross-functional teams to identify and resolve data issues for enhanced reporting quality.

Suchir Softech **Jun 2014 - Dec 2015**
Systems Administrator *India*

- Installed and maintained Linux servers for web, mail, and application services using package managers and configuration scripts, ensuring continuous availability for business operations
- Performed system administration, monitoring, patching, and troubleshooting with Nagios and Ansible, improving incident response time and maintaining system stability

Education

Southern Arkansas University **2017**
Master of Science, Information Science *Magnolia, AR*

Jawaharlal Nehru Technological University **2014**
Bachelor of Technology, Mechanical Engineering *India*

Skills

- **Cloud & Platforms**: GCP, BigQuery, Dataproc, Dataflow, GCS, Secret Manager, AWS, EC2, S3, EMR, Azure Cosmos DB, Vespa DB, Cloud-native Architectures
- **Data Engineering & Processing**: Apache Spark, PySpark, Spark SQL, Batch ETL, Data Modeling, Data warehousing, Schema Evolution, Teradata, Oracle, Data Ingestion Pipelines, ETL Pipelines, Feature Stores, Data Observability, Data Lineage, Data Quality
- **Machine Learning, MLOps & LLMOps**: Spark-based ML Pipelines, Batch Training & Inference, Feature Engineering-, Embedding-Based Similarity, SentenceTransformers, CLIP, FAISS, AutoFAISS, Offline Evaluation, Model Artifact Versioning, Prompt and Config-Driven LLM Pipelines, Reliability Controls
- **Programming & Tooling**: Python, SQL, Scala, Airflow, CI/CD, Git, SonarQube, Informatica, Concord
- **Analytics & Visualization**: Big Query SQL, Tableau, Looker