SANJANA VENKATESH

Los Angeles, CA | (213) 913-1568 | sv51119@usc.edu | LinkedIn | GitHub | Portfolio

EDUCATION

University of Southern California   Los Angeles, CA
Master of Science in Analytics GPA: 3.8/4.00   Aug 2023-May 2025
Relevant Coursework: Data Analysis, Software Engineering, Data Warehousing & ETL/ELT, Cloud Analytics

BMS College of Engineering   Bengaluru, India
Bachelor of Technology in Computer Science; GPA: 3.7/4.00   Jul 2017 - Jul 2021

WORK EXPERIENCE

University of Southern California   Los Angeles, CA
Data Analyst Research Assistant   May 2025 – Present
- Orchestrated ML pipelines on Vertex AI and Snowflake, added CI/CD & Git to automate ops, cutting training time **by 2hrs**
- Crafted Tableau EDA and engineered features; optimized BigQuery (partitioning, clustering) to **reduce query cost 38%**

SAG-AFTRA Health Plan   Burbank, CA
Data Analyst Intern   May 2024 – Aug 2024
- Managed **AWS** resources for Big Data Analytics, optimising **EC2 Instances**, reducing operational cost by **23k annually**
- Automated ingestion pipelines from ThoughtSpot APIs into Snowflake using **REST APIs**, streamlining ETL operations and **reducing latency by 42%**

Quantiphi, Inc   Bengaluru, India
Data Engineer   Feb 2021 – Jul 2023
- Modernized a legacy Hadoop estate on Google Cloud: **migrated Hadoop/MapReduce jobs to Dataproc,** tuned Spark (executor/EFM parameters, SSD/shuffle balancing), yielding **$10M/year infrastructure savings**
- Directed an 8-engineer program to migrate **1,500+** tables to **BigQuery**; hardened data quality with SQL checks and DVT, **achieving ≥99% accuracy** and audit-ready reporting compliance
- Productionized ETL/ELT pipelines with Apache Airflow from GCS to BigQuery, **reducing manual ops by 16 hours/day** and consistently meeting daily refresh SLAs
- Operationalized a governed metrics layer: authored LookML and launched energy **KPI** dashboards that drove **$1.2M annual cost savings**; presented results to client VPs, enabling cross-functional adoption
- Implemented a real-time pipeline to process **150M** raw records from 30+ sources using **Kafka and Pyspark**

PROJECTS

Recommendation Engine | Link   Aug 2025 – Sep 2025
- **Engineered** a RAG recommender with **LangChain**, **OpenAI embeddings**, **LanceDB** to index titles/genres/synopses and return **top-N semantic recs** from NL queries and increased user engagement by **25%**
- **Packaged** an end-to-end, reproducible **pipeline** to **speed LLM-for-RecSys experiments** and enable plug-and-play datasets

Retail Data Analytics | Link   Mar 2024 – Jul 2024
- Automated ETL with **Apache Airflow + Python** to ingest GCS to BigQuery, slashing data latency from **2 hours to 12 minutes** and enabling reliable daily refresh SLAs
- Built the analytics layer: **dbt** star schema with data quality tests, partitioning & clustering for performance; delivered **Power BI** dashboards with DAX for Top Products /Markets, **reducing manual reporting time by 48%**

Intelligent Fab Monitoring & Maintenance System | Link   Jan 2024 – Mar 2024
- Built a real-time geospatial analytics platform tracking **2M+** devices ; delivered KPI dashboards, interactive data visualization, and event-driven alerts, eliminating manual monitoring
- Engineered ETL/ELT for 100K+ RTLS logs using Python performed data cleaning, geofencing, feature engineering and time-series modeling; trained XGBoost for RUL prediction (**82% accuracy**). **1st place among 20+ USC capstone teams**

TECHNICAL SKILLS

Languages: Python; Advanced SQL (BigQuery, Postgres); Bash; PySpark; **Certification**: Associate Cloud Engineer (Google Cloud)
Big Data & Cloud Technologies: GCP (BigQuery, GCS, Dataproc, Vertex AI); AWS (S3, Redshift), Snowflake
ETL and Data Pipelines: Apache Airflow, dbt, Kafka, Informatica, Google Cloud Composer, Git, Azure DevOps, CI/CD
AI & Analytics: Power BI, Tableau, ThoughtSpot, Looker, XGBoost; Terraform; LangChain, embeddings, RAG,Root cause analysis