SAI SUMANTH MUVVA

muvvasaisumanth1001@gmail.com | +1 (812) 837-6858 | United States |
Open to Relocation | 3+ years' of experience | linkedin.com/in/sumanth-muvva | github.com/sumanthmuvva

EDUCATION

Master's - Computer Science, (Indiana University-Bloomington, USA) - *GPA (3.8/4.0)* **Aug 2022 - May 2024**
Bachelor's - Computer Science, (KL University, India) - *CGPA (9.02/10.0)* **Aug 2015 - May 2019**

PROFESSIONAL EXPERIENCE

Data Engineer (Part-time) — Jan 2024 - May 2024

Indiana University

- Optimized a real-time streaming data ingestion pipeline using **PySpark** and Databricks **Autoloader**, achieving a **90% reduction in API calls** to AWS S3 objects.
- Conducted **performance tuning** and optimization of **Hadoop** clusters, achieving a 30% improvement in processing efficiency by analyzing and optimizing **Spark** job and cluster resource configurations.
- Performed data analysis on large, structured datasets (parquet file format) stored in S3 object storage using optimized SQL queries in Athena.

Data Engineer II — Aug 2021 - Jul 2022

UnitedHealth Group

- Migrated complex data pipelines EtLT (ELT) handling 180 million records from **Teradata** to Azure **Synapse Analytics**, implementing **Data-Lakehouse (Medallion)** architecture.
- Innovated a **Reconciliation Framework** with **Presto** and Python to reconcile large tables, saving $11K annually in DAG run costs.
- Designed and implemented logical (LDM) and **physical database models** (PDM) for Teradata data warehouse tables using **ERwin**, effectively managing over 200 million records per table.
- Automated **data validation** for 8 ETL projects by developing **SQL scripts** for each layer, and orchestrating execution with **shell scripting** and Apache Airflow.
- **Mentored 6 junior data engineers** and spearheaded the development of data acquisition and data integration layers for 3 financial batches.
- Implemented Role-Based Access Controls **(RBAC)** within **ADLS Gen2**, enforcing **data governance** and **security** policies to handle sensitive PII/PHI healthcare data.

Data Engineer I — Jan 2020 - Aug 2021

UnitedHealth Group

- Developed and **fine-tuned** 900+ **SQL** queries and proposed strategic indexing and partitioning in Teradata leading to a 40% improvement in data access speeds in **Enterprise Data Warehouse** (EDW).
- Collaborated with a 27-member **cross-functional team** to build an event-driven framework that incorporates **data quality checks** and **notifications**, improving data integrity across layers.
- Designed 7 data ingestion pipelines in **Azure Databricks** for storing **structured and unstructured** healthcare financial data into **Delta Lake** tables, ensuring ACID compliance and schema enforcement.
- Implemented **distributed SQL** queries using **Presto** on an **Apache Spark** cluster to validate data between PROD and HADR environments, achieving sub-second response times for ad-hoc queries.
- Built **Semantic views** in Teradata using **SQL**, enabling business users to efficiently access and analyze aggregated financial data.
- Orchestrated 600+ **DataStage jobs** from IBM Scheduler to a multi-environment, open-source **Apache Airflow** following **Agile** best practices.
- Applied **SCD** (Type 1 & 2) and Change Data Capture **(CDC)** techniques, improving incremental data accuracy within financial data pipelines.

Software Engineer (ETL) — Jul 2019 - Jan 2020

UnitedHealth Group

- Developed **complex SQL logics** for calculating financial metrics, improving reporting accuracy within 5 downstream applications.
- Built a **batch monitoring dashboard** with automated alerts that tracks 80+ financial batches in **Python**, helping on-call members in quick resolution of issues.
- Engineered an **event-driven**, **file monitoring system** with Docker, Kubernetes, and **Kafka**, handling 10,000+ file events hourly to trigger airflow jobs with a 40% reduction in latency.

TECHNICAL SKILLS

ETL & Big Data	: **Teradata**, IBM DataStage, SSIS, SQL Server, **Databricks**, ERwin, Apache **Spark**, Airflow, Kafka, Hadoop, Hive.
Programming & Databases	: Python, **PySpark**, **SQL**, Presto, Linux, Oracle, **PostgreSQL**, MySQL, NoSQL (DynamoDB, MongoDB).
Visualization & Reporting	: Microsoft Power BI, Tableau, Grafana.
Cloud Platforms	: **AWS** (EC2, S3, Glue, EMR, **Redshift**), Azure Data Factory, **Azure Synapse Analytics,** Snowflake.
DevOps & Others	: Git, Jenkins, CI/CD pipelines, REST APIs, Terraform, Agile, Azure DevOps, JIRA.
Certifications	: **AWS Certified** - Solutions Architect Associate (SAA-C03).

PROJECTS

Kaggle Ecosystem Analytics (Business Intelligence) - *[Python, Azure Databricks, Azure Synapse, Power BI]* **Feb 2024 - Apr 2024**

- Architected a scalable data analytics platform, ingesting 3 GB raw Kaggle data into Azure Synapse with Databricks as intermediate layer for data cleaning and transformation.
- Engineered interactive Power BI dashboards via Direct Query and Import modes from Synapse, providing immediate insights into user behavior and competition trends.

Reddit Data Engineering Pipeline (AWS) - *[Reddit API, Apache Airflow, Celery, PostgreSQL, AWS]* **Jan 2023 - Apr 2023**

- Engineered a comprehensive data pipeline integrating Reddit API with AWS services (S3, Glue, Athena, Redshift) streamlining data extraction, transformation, and loading processes.
- Implemented task automation and distributed task management using Airflow and Celery, significantly reducing manual intervention.
- Developed and executed ad hoc SQL queries on Athena and **Redshift**, enabling advanced data analytics and visualization.