SURYA TEJA DAVID

david.suryatej@ufl.edu | 352-745-5854 | linkedin.com/in/suryatejadavid | GitHub | AWS Certified Data Engineer

EDUCATION

University of Florida, *Gainesville* Jan 2023 – Dec 2024
Master of Science in Information Systems and Operations Management (Data Science)

TECHNICAL SKILLS

Programming Languages	Python, R, SQL, Unix/Linux
Cloud Computing	AWS (Kinesis, Lambda, Redshift, Glue, SNS, CloudWatch), GCP (DataProc, BigQuery, Pub/Sub)
Databases (SQL/NoSQL)	MySQL, PostgreSQL, PL/SQL, Oracle, AmazonRDS, Aurora, MongoDB, Cassandra, DynamoDB
Big Data Technologies	Kafka, Spark, Hadoop, Snowflake, Databricks, Airflow, Informatica
SRE/ BI tools	Git, CI/CD, Docker, Kubernetes, Tableau, PowerBI, QuickSight, Kibana
Machine Learning	NumPy, Pandas, TensorFlow, PyTorch, OpenCV

WORK EXPERIENCE

Data Engineer, *University of Florida – Health* *May 2023 – Present*

- Orchestrated the ETL process of image files using **Airflow**, **PySpark**, and **AWS data lake**, reducing preprocessing time by 50%, allowing researchers to analyze 2X more medical images per week and accelerate key experiments in Type-1 Diabetes research.
- Improved database performance by tuning and indexing databases, enhancing **SQL** query efficiency as patient data size increased, leading to faster data retrieval and analysis.
- Developed **Python** scripts to extract key features from images, perform EDA, and store metadata in relational databases, cutting manual effort by 40% and enabling data visualizations in **Tableau** offering clearer insights into immune cell patterns.
- Implemented transfer learning with **U-NET** and **CNN** models for autoimmune cell detection and improved classification accuracy by 8% through data augmentation and fine-tuning techniques, leading to more accurate identification of immune cell activity.

Data Engineer*, ServiceNow* *Jan 2022 – Jan 2023*

- Led the development of an **ELT** pipeline using **Apache Kafka**, Snow APIs, and JDBC to unify logs in a Parquet-format data lake in **S3**, resulting in a 50% increase in data accessibility for the Performance Analytics team.
- Developed **Apache Spark** transformations for time-based aggregations and windowing, loading data into **Redshift**, and lowering query response times by 40% for analytics teams.
- Built an **ETL** pipeline with **AWS Kinesis**, **Apache Spark** on **EMR,** and Snowflake to process 500K daily events from Event Management systems, performing data preprocessing for error rate and service latency analytics.
- Optimized **Snowflake** queries with partitioning, clustering, and materialized views, reducing query run times by 40% and enhancing the AIOps team's ability to identify downtimes and track incident trends.
- Orchestrated workflows with **AWS Step Functions**, optimizing data flow for downstream time-series analytics and cutting pipeline run-time by 2 hours weekly.
- Designed real-time **Tableau** dashboards to quickly identify high-priority incidents, improving resource allocation efficiency by 30%.

Associate Data Engineer, *Apple (TCS)* *Jun 2019 – Dec 2021*

- Captured and streamed traffic data for Apple internal websites into **Cassandra** and **DynamoDB**, enabling near real-time querying and proactive monitoring through **CloudWatch** dashboards and **SNS** alerts for performance and traffic anomalies.
- Implemented large-scale computing frameworks such as **Apache Spark** and **AWS Glue** for data processing tasks like aggregation and filtering of log data, resulting in 2X better query performance.
- Revamped the data pipeline with **Airflow** and **Docker**, using **CI/CD** to automate deployments, reduce deployment time by 30%, and ensure consistent environments with enhanced scalability.
- Led a batch processing pipeline project for 200K daily requests utilizing **Hive**, **Hadoop(MapReduce)**, and **Spark**, addressing traffic skewness and identifying efficient load-balancing algorithms.

PROJECTS

Amazon Black Friday Sales Insights: Real-time Projection and Near Real-time Analytics | GitHub

- Engineered an **ETL** pipeline using **Python**, **Kinesis**, **DynamoDB, EventBridge**, **Lambda**, and Athena for real-time data processing.
- Solved delays in data transformation by integrating **CDC** with **DynamoDB Stream,** Kinesis**,** and **Firehose**, decreasing processing time by 40% and facilitating real-time analytics through **Athena**.

Media Content Analytics Pipeline: Scalable Batch Processing with AWS | GitHub

- Engineered a scalable batch processing system for media data analysis using AWS services (**S3, Glue, Redshift, EventBridge**, and **Athena**), addressed inefficiencies in data processing, and elevated data integrity through data quality checks and **SNS** notifications.

GCP Incremental Analysis Pipeline: Databricks and Pyspark | GitHub

- Executed an **Apache Spark** solution on **GCP** with **Databricks**, automating file processing into daily and historical tables via **Delta Lake** and optimized data staging with **upsert**, reducing **ETL** time by 38%.