SAI YESASWY MYLAVARAPU

4+ years of professional and research experience in Data analytics, problem solving and Big data technologies.

Email: saiyesaswy@gmail.com | Phone: 980-938-9891 | 3464 Fieldstone Dr, Columbus IN 47201

SUMMARY

- Data Engineer with 4+ years of experience in building data pipelines for ingesting & transforming data.
- Hands-on experience in writing complex SQL queries to extract, transform and load (ETL) data from databases.
- Good knowledge of Big data applications and implementation of end-to-end streaming solutions using Spark.
- Knowledge of design & data modeling for OLTP & OLAP databases with problem solving and analytical skills.
- Strong hands-on experience in data cleaning and exploration using various libraries in Python and Scala.

TECHNICAL SKILLS

Programming:	Python, Scala, Java, R
BI/Analytics Tools:	Tableau, Zoho Reports, D3.js, Shiny, Plotly, MS Excel, WEKA
Big Data/Cloud:	HDFS, MapReduce, HIVE, Apache Spark, AWS, Databricks, Microsoft Azure
Databases:	Snowflake, Teradata, Microsoft SQL server, Redshift
Other Skills:	Natural Language processing, Airflow, Docker, GIT, FLASK, NoSQL databases

EDUCATION

UNIVERSITY OF NORTH CAROLINA AT CHARLOTTE, Charlotte, NC — 3.9 GPA
Master of Science in Computer Science (Data Science & Management) — Spring'17 – May'18

JAWAHARLAL NEHRU TECHNOLOGICAL UNIVERSITY, Hyderabad, India — 3.5 GPA
Bachelor of Technology in Electrical & Electronics Engineering — August'10 – May'14

EXPERIENCE

DATA ENGINEER — Columbus, IN
Cummins — May'20 – Present

- **Environment**: Apache SPARK, Databricks, Microsoft Azure, Scala, SQL, Python, HIVE
- **Data engineering frameworks:** Part of the Advanced Analytics and Artificial Intelligence team, worked on developing two frameworks for data ingestion and transformation.
- Played a key role in migrating the frameworks' environment to reflect the latest Databricks runtime version 7.3 LTS
- **Delta lake:** Worked on migrating hive tables from parquet to delta format in the Azure data lake Gen2 environment, which brought a significant improvement in the overall query performance for the team.
- **Structured Streaming:** Implemented an end-to-end structured streaming solution for a product, which replaced an existing batch data pipeline with an almost real-time pipeline from the raw layer to feature layer.
- **Databricks migration:** Migrated shell script projects running on Azure HDInsight to Databricks using Spark Scala.
- **Databricks Workspace setup:** Worked on setting up an NPIP Databricks workspace for product teams.

DATA ENGINEER — Richmond, VA
Capital One — August'18 – May'20

- **Environment**: Snowflake, Databricks, Apache Spark, Python, Redshift, AWS tools
- **Scripting**: Wrote production level complex SQL queries involving aggregate and window functions for pulling large sets of time sensitive data from OLAP environments like Teradata, Snowflake, Redshift.
- **ETL Pipelines**: Developed ETL pipelines to ingest transactional data, transform it by applying data munging techniques and move the data using a real time processing pipeline into data warehouse for analysis.
- **Schema validation**: Developed scripts to assign schema, validate data types and remove duplicates on large amounts of transactional data and created analytical version of data for analysis using Databricks.
- **Ingestion**: Collaborate with Product team to understand the requirements to ingest new tables and work with Data Architects to understand standards and security concerns for implementing new ingestion methods.
- **Metrics**: Derived actionable insights and metrics by analyzing the customers' response to various communications.
- **Dashboarding**: Presented analytical results to business teams by building charts and interactive visualizations in dashboards by leveraging BI tools like Tableau and AWS Quicksight.

DATA SCIENTIST (INTERN) Charlotte, NC

Catalyst January'18 – May'18

- Responsible for ingestion of customer data from various data sources into HDFS and build internal and external HIVE table schemas using performance techniques like partitioning & bucketing.
- Built regression models on sales data & applied techniques like regularization & dimension reduction. Created interactive dashboards to track supply & demand in Tableau & Zoho reports.

DATA SCIENTIST Hyderabad, India

Ikvox Software May'16 – December'16

- Wrote complex SQL queries to extract, transform and load (ETL) data from relational and NoSQL databases.
- Used qualitative and quantitative user feedback data such as user rating, Net promoter score and daily sales to predict the customer churn rate using logistic regression, decision trees, SVM and KNN.
- Performed sentiment analysis on user reviews by extracting features from huge corpus of data using statistical packages in R and Python. Designed dashboard to visualize the feedback metrics to the stakeholders.

SYSTEMS ENGINEER Hyderabad, India

Infosys Ltd August'14 – May'16

- Understood requirements from Stakeholders and create documentation with required set of instructions.
- Prepare project timeline and schedule for all deliverables throughout the Software Development Lifecycle.
- Used statistical techniques to deduct key UX metrics of products that shape customer satisfaction.
- Used Numpy & Pandas libraries in Python extensively by writing functions to bring raw data into structured format.
- Responsible for writing complex SQL and PL/SQL statements – Stored procedures, functions and triggers.

PROJECTS

SPAM/HAM Filter with Sentiment Analysis (Web application using Flask) **Python, NLTK, FLASK**

- Built a Naïve bayes classifier on the Enron email corpus, to determine whether the given text is spam or not.
- Data from the corpus is tokenized and word features are extracted before building the model.
- The model is serialized using Pickle and created a web application using HTML & CSS with Flask as the backend.

Movie lens dataset analysis **MapReduce, HIVE**

- Analyzed the Movie lens 1M dataset. This dataset has one million ratings from 6040 users for 3952 movies.
- Wrote Mapper & Reducer functions to understand the relationships between ratings and users' profiles.
- Performed similar operations by writing Hive Query language (HQL) queries on the ratings data.

Sentiment Analysis & Automatic emotion detection on Twitter data **Scala, Python, Spark MLLib, SVM**

- Classified tweets collected using twitter streaming API into 8 different emotions by building a classifier.
- Built Decision tree and SVM One-Vs-All classifiers on processed tweets using Scala and Spark MLLib.
- The model performance is determined by executing on Spark cluster using various evaluation metrics.