SAI SRUJAN KUMAR MAJETI

Boston, MA | (+1) 857-891-6994 | srujankumar.majeti@gmail.com
LinkedIn: https://www.linkedin.com/in/srujanmajeti | GitHub: https://github.com/srujanmajeti/

SKILLS

Languages	Python, Java, R, SQL (PostgreSQL, SQL Server), NoSQL (DynamoDB, Cassandra)
Distributed Tools	Hadoop, Spark, Sqoop, Kafka, Hive, AWS, Airflow
Libraries	Pandas, NumPy, SciPy, scikit-learn, Beautiful Soup, matplotlib, seaborn, Keras, boto3
Tools	Git, Linux, Postman, Tableau

EXPERIENCE

Data Engineering Fellow, Insight Data Science — May 2020 – Present

- Built a tool *Personal Broker* that helps retail investors to visualize the real-time happenings of the stock market deployed in AWS
- Leveraged Spark to process the data from Yahoo API and ingest into DynamoDB and to compute metrics
- Calculated the change percent of the stock price and alerted user if the change percent is above the threshold set for his personal portfolio
- Automated data pipeline using Airflow and created interactive visualizations using Python Dash in the frontend

Data Analyst Intern, All the Way Live Analytics — Sep 2019 - Dec 2019

- Architected models to predict difference of home and away team scores in NBA by performing feature engineering
- Developed an API to scrape data in JSON format from different webpages and created pre-processing scripts to clean the data
- Hosted the model in AWS using EC2 and Kinesis by creating a streaming service to ingest the data for stream processing

Big Data Developer, Accenture — Aug 2016 - Jul 2018

- Created a Data Lake in Hadoop by unifying data from different source systems (Teradata, SQL Server, and Oracle DB) to HDFS using Sqoop
- Performed claim processing for a health insurance client using Spark Batch Processing on more than 2 TB of customers claim data and ingested the results into Hive
- Developed shell scripts to automate creation of HiveQL statements that are compatible with the source system
- Improved efficiency by creating Hive external and internal tables using partitioning and bucketing techniques with ORC format
- Scheduled CRON jobs to implement Change Data Capture technique to capture changing data for multiple systems
- Automated ETL processes by creating a framework that reduced the manual data load effort by almost 60 percent and built this as an asset for future data loads

PROJECTS

Million Songs Data Warehouse on AWS Redshift

- Created a Data Warehouse for a subset of Million songs data and stored on the staging area on AWS S3 using boto3
- Loaded JSON formatted files from S3 to AWS Redshift to create Star schema tables in aiming to align with OLAP and BI processes
- Configured DAG workflows using Airflow to maintain and schedule sequential jobs for the pipeline

VIACOM Ad Targeting

- Engineered demographics-based features to estimate cost per mile (CPM) rate in random forest algorithm
- Developed Key Performance Indicators using k-means clustering to monitor market gaps and hidden patterns
- Utilized Tableau to make clear and concise visual representations of page level data

EDUCATION

Northeastern University — May 2020

Master of Science in Data Analytics (Specializing in Statistical Modeling)

- Parallel Data Processing, Data Management and Big data, Machine Learning, Probability and Statistics

Vellore Institute of Technology — May 2016

Bachelors in Electronics and Communication Engineering

- Object Oriented Design, Data Structures & Algorithms, Computer Networks, Database Management

ACTIVITIES

Graduate Teaching Assistant, Northeastern University — Jan 2020 – May 2020

- Instructed the courses ‘Probability and Statistics’ and ‘Intermediate Analytics’ using R
- Helped graduate students in lab sessions that emphasize hands-on machine learning algorithms and problem solving