Mansi Vyas

Software Engineer

Phone: +1 (619) 918-6961 | Email: mvyas1685@gmail.com |Location:CA

SUMMARY

- Software Engineer with expertise in Data Engineering and ML engineering with 3+ years of experience including large data sets of Structured and Unstructured data, Data Acquisition, DataValidation, Predictive modeling, Statistical modeling, Data modeling, and Data Visualization.
- Ability to develop enterprise-level solutions using batch processing and streaming frameworks (Spark Streaming, Apache Kafka & Apache Flink).
- Proficient in Big Data using Hadoop and Spark framework and related technologies such as Hadoop, MapReduce, Hive, Pig, BigQuery, HDFS, Spark, Sqoop, and Zookeeper.
- Strong knowledge of RDBMS concepts, Data Modeling (Facts and Dimensions, Star/Snowflake schemas), Data Migration, Data Cleansing, and ETL Processes.
- Extensively used Azure Databricks for data validations and analysis on Cosmos structured streams.

TECHNICAL SKILLS

Programming Language: Java, Python,Scala, SQL
Full Stack Frameworks: Node.js, REST APIs , Vue.js,React , Apollo , GraphQL
Big Data Ecosystem: Hadoop, MapReduce, Hive, Pig, DynamoDB, BigQuery, HDFS, Spark
Machine Learning: Linear Regression, Logistic Regression, Decision Tree, K mean, Naïve Bayes, Random Forest
Cloud Technologies: AWS (EC2, S3 Bucket, Redshift, Lambda, IAM), GCP, Azure
Packages: NumPy, Pandas, Matplotlib, SciPy, Scikit-learn, Seaborn, TensorFlow, Pytorch, Kafka, PySpark, spaCy
Reporting Tools: Tableau, Power BI, SSRS
Database: MS SQL Server, PostgreSQL, MongoDB, MySQL

EXPERIENCE

Data Science Lab, San Diego State University | Data Engineer Student Assistant **August 2023 - Present**

- Developed data pipelines to extract, transform, and load large volumes of medical conversations, clinical guidelines, and electronic health records into BigQuery for training data.
- Created Docker containers and Kubernetes jobs to orchestrate and deploy large healthcare language models for inference. Optimized for efficiency and high availability.
- Developed a React, Node.js full stack application that invoked conversational model APIs to enable interactive visual search of medical recordsand images through natural dialogue.

Neo4j, San Mateo, CA | Consulting Data Engineer **May 2022 – Aug 2022**

- Performed data modeling of unstructured file systems into graph databases by extracting key entities as nodes and relationships as edges.
- Created robust data pipelines for Amazon Prime users, harnessed unsupervised learning to deliver visualization reports for data-driven decision
- Worked on Pandas, NumPy, Seaborn, matplotlib, Scikit-learn, SciPy, and NLTK in Python for developing various machine learning algorithms.
- Established ETL workflows using Apache Spark and Python, resulting in a 30% reduction in data processing time and improved data accuracy.
- Develped end-to-end application using GRAND stack, built connectivity between neo4j db and Apollo server to React using GraphQL.

Adons Softech, India | Data Engineer **Jun 2020 – July 2021**

- Collaborated with data engineers and operation team to implement ETL process, wrote and optimized SQL queries to perform data extraction to fit the analytical requirements.
- Achieved 90% customer monthly retention by predicting the likelihood of returning customers using a Random Forest, XG Boost algorithms.
- Improved Spark Streaming programs to process near real-time data from Kafka, and process data with stateless and state-full transformations.
- Managed the integration of AWS services with on-premise resources, creating a hybrid cloud environment that increased flexibility and reduced operational costs by 20%.

Accenture, India | Software Developer **Nov 2018 – May 2020**

- Designed complex, PL/SQL queries, stored procedures, and devised ETL pipelines using Apache Kafka.
- Extracted data from SQL Server Database, copied into HDFS File system and used Hadoop tools such as Hive and Pig Latin to retrieve the data required for building models.
- Built real-time data pipelines loading high-velocity time-series data from disparate sources into Hadoop using Flume and transferring batch and incremental data to relational databases using Sqoop optimized for performance.
- Created scripts to load data to Hive from HDFS, was ingested data into the Data Warehouse using various data-loading techniques.
- Conducted Data blending and data preparation in SQL for Tableau consumption and publishing data sources to the Tableau server.

RELEVANT PROJECT

Cryptocurrency Market Sentiment Analysis for Investment Recommendations

- Aggregate real-time data from multiple sources. Implement NLP models to process and extract relevant information from textual data predict market movement and recommend cryptocurrencies to invest.

EDUCATION

Master of Science in Computer Science | San Diego State University, San Diego, CA **Dec 2023**
BE in Information Technology | Pune University, India **June 2018**