VINEETHA SHAKAMURI
DATA ENGINEER

Phone: 248-602-9094 | **Email Id**: shakamurivineetha039@gmail.com

Professional Summary

Data engineer with four years of experience in designing, implementing, and optimizing data solutions to support business objectives. Proficient in leveraging a diverse set of technologies and tools to build scalable and efficient data pipelines, enabling data-driven decision-making and business insights. Skilled in managing and processing large volumes of structured and unstructured data, with a strong focus on data quality, reliability, and performance. Demonstrated expertise in cloud platforms such as Azure, AWS, and Snowflake, as well as proficiency in technologies like Python (NumPy, Pandas, Matplotlib, SciPy, Scikit-learn, Seaborn), SQL, Apache Airflow, Azure Synapse and Scala. Proven ability to collaborate effectively with cross-functional teams to deliver impactful data solutions and drive organizational success.

Education

Masters in Computer Science **Jan 2022 – April 2023**
Oakland University

Work Experience

Data Engineer **Jul 2023- Current**
Change Healthcare, GA

- Interacted and gathered requirements for Designing and developing common architecture for storing Retail data within Enterprise and building Data Lake in Azure cloud.
- Developed Spark applications for data extraction, transformation and aggregation from multiple systems and stored on Azure Data Lake Storage using Azure Databricks notebooks.
- Developed applications using PySpark to integrate data coming from other sources like ftp, csv files processed using Azure Databricks and written into Snowflake.
- Written Unzip and decode functions using Spark with Scala and parsing the xml files into Azure blog storage.
- Developed PySpark scripts from source system like Azure Event Hub to ingest data in reload, append, and merge mode into Delta tables in Databricks.
- Design and Develop ETL Processes in DataBricks to migrate Campaign data from external sources like Azure DataLake, and gen2 in ORC/Parquet/Text Files.
- Optimized PySpark applications on Databricks, which yielded a significant amount of cost reduction.
- Created Pipelines in Azure Data Factory to copy parquet files from ADLS Gen2 location to Azure Synapse Analytics Data Warehouse.
- Worked on replacing existing Hive scripts with Spark Data-Frame transformation and actions for faster analysis of the data.
- Developed PySpark scripts to Reduce costs of organization by 30% by migrating customers data in SQL Server to Hadoop.
- Experience in handling JSON datasets and writing custom Python functions to parse through JSON data using Spark.
- Used Spark for interactive queries, processing of streaming data, and integration with popular NoSQL databases for huge volumes of data.
- Responsible for loading Data pipelines from web servers using Kafka and Spark Streaming API
- Used Airflow for orchestration and scheduling of the ingestion scripts.
- Generate weekly based reports and ops reports, customer goals reports, mobile scan and pay goals and usage in data by using power BI.

Data Engineer **May 2018 - Nov 2021**
Kingston Info Solution Services, India

- Worked as a Data Engineer with Big Data and Hadoop ecosystem components.
- Involved in converting Hive/SQL queries into Spark transformations using Scala.
- Created Spark data frames using Spark SQL and prepared data for data analytics by storing it in AWS S3.
- Responsible for loading data from Kafka into HBase using REST API.
- Developed batch scripts to fetch the data from AWS S3 storage and perform required transformations in Scala using the Spark framework.
- Used Spark streaming APIs to perform transformations and actions on the fly for building a common learner data model which gets the data from Kafka in near real-time and persists it to the HBase.

- Created Sqoop scripts to import and export customer profile data from RDBMS to S3 buckets.
- Developed various enrichment applications in Spark using Scala for cleansing and enrichment of clickstream data with customer profile lookups.
- Troubleshooting Spark applications for improved error tolerance and reliability.
- Used Spark Data frame and Spark API to implement batch processing of Jobs.
- Well-versed with Pandas data frames and Spark data frames.
- Used Apache Kafka and Spark Streaming to get the data from adobe live stream rest API connections.
- Automated creation and termination of AWS EMR clusters.
- Worked in real-time data streaming data using AWS Kinesis, EMR, and AWS Glue.
- Worked on fine-tuning and performance enhancements of various spark applications and hive scripts.
- Used various concepts in spark like broadcast variables, caching, and dynamic allocation to design more scalable spark applications.
- Imported Hundreds of structured data from relational databases using Sqoop import to process using Spark and stored the data into HDFS in CSV format.
- Created data partitions on large data sets in S3 and DDL on partitioned data.
- Improving Efficiency by modifying existing Data pipelines on Matilion ETL tool to load the data into AWS Redshift.
- Identify source systems, their connectivity, related tables, and fields and ensure data suitability for mapping, preparing unit test cases, and providing support to the testing team to fix defects.
- Defined HBase tables to store various data formats of incoming data from different portfolios.
- Developed the verification and control process for daily data loading.

Technical Skills

Programming Languages: Python, Scala, SQL, Shell Scripting.
Big Data Ecosystem: HDFS, Pig, Hive, Oozie, Sqoop, Kafka, Spark Streaming, Spark SQL, Dataframe.
Cloud Ecosystem: Azure (Azure Data Factory, Azure Data Bricks, Azure Synapse, ADLS Gen2), AWS (EC2, EMR, Lambda, Kinesis, Glue, Redshift and S3).
Databases: MySQL Server, Oracle DB, HiveQL, Spark SQL, HBase, Redshift, Snowflake.
Orchestration/Tools: Airflow, Oozie, Jira, Git, Matilion, Postman, Power BI, Docker, Jenkins.
Streaming: Spark Streaming, Kafka.
IDE: DataBricks, Pycharm, Anaconda, IntelliJ.