VENKAYYA J

Dallas, TX, 75232| 913-586-6083 | venkayyaj938@gmail.com

PROFILE SUMMARY

- 4 years of Software development experience with expertise in Data Engineering using Big data tools and Application development using **Java, Scala and Python**.
- Experience on working with various tools in Hadoop ecosystem like **MapReduce, Hive, Yarn, HDFS, Kafka, Sqoop, Flume, Oozie, HBase, Impala etc.,**
- Experience on working with **Spark** for performing large scale data processing, data cleansing, data aggregations etc.,
- Experience on utilizing **Spark RDD, Spark Dataframes, Spark SQL and Spark Streaming Api's** extensively.
- Proficient **SQL** experience in querying, data extraction/transformations .
- Experience working with **NoSQL** databases like **MongoDB.**
- Strong experience working with various file formats like **Avro, Parquet, Oracle, Json, Csv** etc.
- Strong experience building various data models in **Hive** and writing a hive script for various data analysis requirements.
- Experience on building real time streaming pipelines using **Kafka** and **Spark Structured streaming**.
- Worked on building various data ingestion pipelines to pull data from various sources like **S3 buckets**, FTP servers and Rest Applications.
- Implemented robust security measures within **Snowflake.**
- Strong in **core Java** concepts including Object-Oriented Design (OOD) and Java components like Collections Framework, Exception handling.
- Effective collaboration within cross-functional teams, providing expertise in **Snowflake** to support data related projects.
- Extensive experience in data modeling and designing **Snowflake schemas**, ensuring effective organization and accessibility of data for analytical purposes.
- Day to-day responsibility includes developing **ETL** Pipelines in and out of data warehouse, develop major regulatory and financial reports using advanced SQL queries in **snowflake**.
- Successfully integrated **Snowflake** with various data sources, ensuring seamless data flow and compatibility with different applications and platforms.
- Experience working with **GitHub, Jenkins and Maven**.
- Strong knowledge building data lakes in **AWS Cloud** utilizing services like S3, EMR, Athena, Redshift, Glue, Metastore, Lambda Functions etc.,
- Strong experience developing end to end **data pipelines**, automating, and maintaining different data pipelines in production.

TECHNICAL SKILLS

Hadoop/Big Data Technologies:	HDFS, Hive, Spark, Spark SQL
Hadoop Distributions:	Cloudera and AWS EMR
Programming Languages:	Java, Python
Operating Systems:	Linux, Unix and Windows
Databases:	SQL Server, MySQL

Build Tools: Maven, Jenkins
Version Control: Git, SVN, CNS

ORGANIZATIONAL EXPERIENCE

Client: Cerner Corporation, Remote **Jan 2023 to Present**
Job Title: Junior Data Engineer

Responsibilities:
• Worked on migrating datasets and **ETL** workloads from On-prem (MapR Cluster) to **AWS** Cloud.
• Built series of **Spark** Applications and **Hive scripts** to produce various analytical datasets needed for digital marketing teams.
• Bulk loading from the External stage (**AWS s3**), internal stage to Snowflake cloud using the **COPY** command.
• Loading the data into **Snowflake** tables from the internal stage using **snowsql**.
• Used **SNOW PIPE** for continuous data ingestion from the S3 bucket.
• Experience in **Splunk** reporting system.
• Creating a **RESTFUL** web service using elastic search services and creating queries in elastic search.
• Worked extensively on building and automating **data ingestion pipelines** and moving terabytes of data from existing data warehouses to cloud.
• Experience in data aggregation and search using **Elastic search**.
• In-depth knowledge of **Snowflake architecture**, including virtual warehouses, storage, and compute resources, enabling efficient data storage and processing.
• Extensive experience in data modeling and designing **Snowflake schemas**, ensuring effective organization and accessibility of data for analytical purposes.
• Worked extensively on fine tuning **spark** applications and providing production support to various pipelines running in production.
• Experience in designing and data modeling in **Elastic search**.
• Used import and export from the internal stage(Snowflake) from the external stage(**AWS S3**).
• Worked closely with business teams and data science teams and ensured all the requirements are translated accurately into our data pipelines.
• Worked on full spectrum of data engineering pipelines: **data ingestion, data transformations and data analysis/consumption.**
• Implement data aggregations scripts using elastic search.
• Worked on automating the infrastructure setup, launching and termination **EMR clusters** etc.,
• Created **Hive** external tables on top of datasets loaded in S3 buckets and created various **hive scripts** to produce series of aggregated datasets for downstream analysis.
• Successfully integrated **Snowflake** with various data sources, ensuring seamless data flow and compatibility with different applications and platforms.
• Build real time streaming pipeline utilizing **Kafka, Spark Streaming and Redshift**.
• Worked on creating **Kafka** producers using Kafka Java Producer Api for connecting to external Rest live stream application and producing messages to Kafka topic.
• Responsible for deploying final workflows to production and maintaining and supporting production pipelines.

.

Experian, TX **May 2019 to July 2021**
Job Title: Junior Data Engineer

Responsibilities:
• Application design for integration with **REST API's, Merchant UI** and custom python libraries.
• Worked on migrating datasets and **ETL workloads** from On-prem (MapR Cluster) to AWS Cloud.
• Built series of **Spark Applications and Hive scripts** to produce various analytical datasets needed for digital marketing teams.
• Experience in **Splunk** reporting system.
• Worked extensively on fine tuning **spark applications** and providing **production support** to various pipelines running in production.
• Worked on performance tuning of Spark applications to reduce job execution times.
• Expertise in maximizing performance and resource usage when establishing and configuring **EMR clusters** for distributed data processing jobs.
• In-depth knowledge of **Snowflake architecture**, including virtual warehouses, storage, and compute resources, enabling efficient data storage and processing.
• Responsible for **deploying** final workflows to production and maintaining and supporting production pipelines.
• Extensive experience in data modeling and **designing Snowflake schemas**, ensuring effective organization and accessibility of data for analytical purposes.
• Familiarity with **AWS Lambda Functions** for serverless data processing and automation tasks, enabling cost-effective and scalable solutions.
• Experience in utilizing **AWS Data Pipeline** for orchestrating and scheduling data processing workflows, ensuring timely and efficient data movement and transformation.
• Worked closely with business teams and data science teams and ensured all the requirements are translated accurately into our data pipelines.
• Worked on full spectrum of data engineering pipelines: data ingestion, data transformations and data analysis/consumption.
• Experience in utilizing **AWS Data Pipeline** for orchestrating and scheduling data processing workflows, ensuring timely and efficient data movement and transformation.
• Worked on automating the infrastructure setup, launching and termination EMR clusters etc.,
• Created **Hive** external tables on top of datasets loaded in S3 buckets and created various hive scripts to produce series of aggregated datasets for downstream analysis.

EDUCATION

Bachelor of Technology in Electronics and communication Engineering **July 2016 – Dec 2020**

- Jawaharlal Nehru Technological University, Kakinada

Masters in computer science

- University of Central Missouri **Aug 2021 - Dec 2022**