Vamshikrishna Sunnam

Data Engineer

(330) 431-0747 | sunnamvamshikrishna236@gmail.com | LinkedIn

Professional Summary:

- 3+ years of professional experience in information technology as **Data Engineer** with an expert hand in the areas of Database Development, ETL Development, Data modelling, Data Visualization, Report Development and Big Data Technologies.
- Result-oriented and highly skilled professional with experience in **AWS**, **Snowflake** and **Big Data technologies.**
- Strong experience in writing scripts using **Python API, PySpark API** and **Spark API** for analyzing data.
- Extensive experience in working with **HDFS, Sqoop, PySpark, Hive, MapReduce**, and **HBase** for **big data** processing and analytics.
- Expertise in **Snowflake** to create and Maintain Tables and views. **Python** Libraries **PySpark, PyTest, NumPy** and **Pandas**.
- Utilized **AWS S3** for scalable and cost-effective data storage and retrieval.
- Adept in integrating **AWS SNS** and **SQS** for real-time event processing and messaging.
- Skilled in utilizing **AWS** services such as **CloudWatch**, **Kinesis**, **Route53** for effective monitoring, data streaming, DNS management, and network access control in cloud environments.
- Proficient in managing user access and permissions to **AWS** resources using **IAM**.
- Experienced in utilizing **AWS Glue** for **ETL** workflows, enabling efficient data extraction, transformation, and loading.
- Strong knowledge of **AWS** CloudWatch for monitoring and managing **AWS** resources, setting up alarms, and collecting metrics.
- Proficient in creating effective data pipelines and performing complex data manipulation tasks using **Snow SQL**.
- Implemented Data pipelines using **Pandas** Data Frame, **Spark Data Frame, RDDs**.
- Skilled in designing roles, views, and implementing performance tuning techniques to enhance **Snowflake** system performance.
- Intensive experience in **AWS** Cloud Environment, **Hadoop ecosystem** in **designing**, **deploying**, and **operating** highly available, **scalable**, and **fault**-**tolerant** systems.
- Proficient in utilizing virtual warehouses, caching, and **Snow pipe** for real-time data ingestion and processing in **Snowflake**.
- Experienced in integrating data from diverse sources, including loading nested JSON-formatted data into Snowflake tables, using the **AWS S3** bucket and the **Snowflake** cloud **data warehouse.**
- Highly proficient in **Snowflake** scripting to automate **ETL** processes, data transformations, and data pipelines.
- Proficient in developing and optimizing **Spark** and **Spark**-**Streaming** applications for real-time data processing and analytics.
- Strong database development skills in **Teradata**, **Oracle**, **SQL Server**, including the development of stored procedures, triggers, and cursors.
- Proficient in version control systems like **Git**, **GitLab**, and **VSS** for code repository management and collaboration.

TECHNICAL SKILLS:

Big Data Technologies	HDFS, MapReduce, Spark, Yarn, Hive, Pig, HBase, Sqoop, Flume, Kafka, Oozie, Zookeeper, Nifi, Impala
Hadoop Technologies	Apache Hadoop, Cloudera CDH4/CDH5.
Programming Languages	Python, Java, C, C++, HTML, JavaScript, SQL
Frameworks	Django REST framework, MVC, Hortonworks
Databases	T-SQL, PL-SQL, Oracle RAC, data guard, TOAD, SQL Server, PostgreSQL, Snowflake, HBase
Visualization/ Reporting	Tableau, matplotlib and Power BI
Cloud	AWS, Azure, GCP
Versioning Tools	SVN, Git, GitHub
Development Methodologies	Agile, Waterfall
Tools	PyCharm, Eclipse, Visual Studio, ADF, Turbine, SQL*Plus, SQL Developer, TOAD, SQL, SQL Server Management Studio, Terraform, Eclipse, Postman, Docker.

Professional Experience:

Client: Centene Corp, St. Louis, Missouri **June 2023 – Till date**
Data Engineer

Project Description: Centene Corp seeks to optimize its healthcare provider network to improve access to quality care, enhance member satisfaction, and reduce healthcare costs. The objective of this project is to develop a real-time provider network optimization platform that leverages data analytics and machine learning to dynamically match members with the most appropriate and cost-effective providers based on their healthcare needs and preferences.

Responsibilities:

- Utilized Spark, SQL, and Python on **Databricks** platform to design and implement data ingestion and storage solutions using **AWS S3, Redshift**, and **Glue**.
- Developed a **Python Script** to load the **CSV** files into the **S3** buckets.
- Created and configured workflows for data processing and **ETL pipelines**.
- Developed data processing pipelines using **Hadoop**, including **HDFS**, **Sqoop**, **Hive**, **MapReduce**, and **Spark**.
- Integrated **Hadoop** into traditional **ETL**, accelerating the extraction, transformation, and loading of massive structured and unstructured data.
- Implemented **Spark Streaming** for real-time data processing and analytics.
- Implemented scheduling and job automation using IBM Tivoli, Control-M, **Oozie**, and **Airflow**.
- Designed and developed database solutions using **Teradata**, **Oracle**, and **SQL Server.**
- Utilized **Scala** to deploy comprehensive data engineering applications, integrating AWS S3, Redshift, and Glue to facilitate data extraction, transformation, and loading processes, ensuring the delivery of scalable and maintainable solutions.
- Managed various **Snowflake** table types and optimized warehouses for performance.
- Developed complex **Snow SQL** queries and partitioning techniques for efficient data retrieval.
- Worked extensively with importing metadata into **Hive** and migrated existing tables and applications to work on **Hive** and **AWS** cloud and making the data available in **Athena** and **Snowflake**.
- Implemented **AWS Athena** for ad-hoc data analysis and querying on **S3 data**.

- Integrated **AWS SNS** and **SQS** for real-time event processing and messaging.
- Data Extraction, aggregation, and consolidation of Adobe data within **AWS Glue** using **PySpark**.
- Utilized **AWS CloudWatch** for monitoring and managing resources, setting up alarms, and collecting metrics.
- Designed and implemented data streaming solutions using **AWS Kinesis** for real-time data processing.
- Successfully deployed Data Analytics and Engineering resources on AWS using **Terraform**, configuring workflows for data processing and ETL pipelines, and integrating **AWS Athena** for ad-hoc data analysis and querying on S3 data.
- Optimized DNS configurations and routing using **AWS Route53** for efficient application and service deployment.
- Loaded the transformed data into **AWS RedShift data warehousing** to analyze the data.
- Designed and implemented **Snowflake** stages to efficiently load data from various sources into **Snowflake** tables.
- Built Apache **Airflow** with **AWS** to analyze multi-stage machine learning processes with **Amazon SageMaker** tasks.
- Designed and developed Security Framework to provide fine-grained access to objects in **AWS S3** using **AWS Lambda.**
- Used **AWS EMR** to move large data (Big Data) into other platforms such as **AWS** data stores, **Amazon S3 and Amazon Dynamo DB.**
- Configured **multi-cluster warehouses** and defined access privileges for security.
- Implemented caching mechanisms, **Snowpipe** for real-time data ingestion, and time travel for historical data tracking.
- Used **regular expressions** and **Snowflake scripting** for automation of pipelines and transformations.
- Utilized **Git, GitLab**, and **VSS** for code repository management and collaboration.
- Worked on **Apache NIFI** to decompress and move **JSON** files from local to **HDFS**.
- Moving data from **Teradata** to a **Hadoop** cluster Using TDCH/Fast export and **Apache NIFI.**
- Implemented **data governance** frameworks on AWS, including **access control policies** and **encryption mechanisms**, to ensure the security and privacy of customer data in compliance with regulatory standards.

Environment: Python, Hadoop (Cloudera Stack), Spark, AWS, EC2, EMR, Hive, Kafka, HBase, HDFS, Pig, Sqoop, Oracle, SQL Workbench, Tableau, Kibana, Spark SQL, Spark Streaming, Scala, Informatica, Jenkins, Docker, Teradata, JIRA, Hue, ETL, AWS S3, AWS Glue, Redshift, SNS, SQS, Athena, Kinesis, Route53, IAM, GIT, Grafana.

Client: Meluha Technologies, Hyderabad, IN **Jan 2021 – Dec 2022**
Data Engineer

Responsibilities:

- Design and implement scalable, fault-tolerant big data solutions using **Hadoop** and related technologies such as **HDFS, MapReduce, Yarn, Hive, Pig,** and **Spark**.
- Configure and manage **Hadoop** clusters using tools such as **Cloudera Manager**, **Ambari**, or **Hortonworks** Data Platform
- Develop and maintain data pipelines using tools like **Apache NiFi, Apache Kafka,** and **Apache Storm**.
- Build and maintain data warehousing solutions using **Hive** and **Impala.**

- Optimize and improve the performance of **Hadoop** clusters by tuning parameters and implementing best practices.
- Collaborate with data scientists, data analysts, and other team members to support data-driven decision-making.
- Experience with big data processing and analysis frameworks such as **Apache Spark, Storm,** and **Flink.**
- Experience with data integration and migration tools such as **Apache NiFi, Apache Kafka**, and **Sqoop**.
- Experience with cluster management and orchestration tools such as **Cloudera Manager**, **Ambari**, and **Hortonworks** Data Platform.
- Work with different data sources like **HDFS**, **Hive** and **Teradata** for **Spark** to process the data.
- Use **Kafka** a publish-subscribe messaging system by creating topics using consumers and producers to ingest data into the application for **Spark** to process the data and Configure **Zookeeper** to coordinate and support the distributed applications as it offers high throughput and availability with low latency.
- Configure Nginx to serve the static content of the web pages reducing the load on the web server for the static content.
- Write **SQL queries** to perform CRUD operations on **PostgreSQL** to save, store, update, and delete rows in tables using Play Slick.
- Create and update **Jenkins** jobs to develop pipelines to deploy the application in different environments like develop, **QA** and Production.

Environment: Spark, SQL, Scala, Jenkins, Kafka, HBase, HDFS, Hive, Teradata, NiFi, Storm, Flink, Sqoop, MapReduce, Yarn, Zookeeper, Pig.

Education:

Masters in Business Analytics | Kent State University, OH — May 2024

Bachelors in Bachelor of Business Administration | Gitam University, IN — May 2022