Sai Vamsi Dudigam

+1(216) 333-7191 | saivamsidudigam@gmail.com | LinkedIn

Professional Summary

As an experienced Data Engineer with a strong foundation in Data Science and Data Analysis, I have a proven track record in building and optimizing scalable data pipelines and ETL processes across AWS, Azure, and GCP. I have advanced expertise in Snowflake, Databricks, and data warehousing, leveraging Python, Scala, and PowerShell to automate workflows and ensure seamless data integration. With a deep understanding of big data technologies like Spark, Hadoop, and Hive, I design high-performance architectures that enable real-time reporting and data-driven decision-making, streamlining operations and delivering actionable insights at scale.

Work Experience

JP Morgan and Chase – Data Engineer; Remote **February 2024 – Current**

Project Description:

As a Data Engineer at JP Morgan Chase, I collaborated with cross-functional teams to deliver data solutions that align with business objectives. Utilizing Azure and GCP, I developed scalable data pipelines and optimized workflows to drive impactful outcomes. My ability to communicate complex technical concepts to non-technical stakeholders ensured the successful implementation of high-performance data solutions.

Responsibilities:

- In this role, my responsibilities included analyzing and evaluating business rules, data sources, and data volumes to create comprehensive estimation, planning, and execution strategies. I designed and implemented ETL processes using Azure Data Factory and Google Cloud Platform tools, ensuring efficient data ingestion and transformation to meet business requirements across both cloud environments.
- Managed large datasets by developing data pipelines in Azure Databricks and GCP Dataproc using PySpark, Spark SQL, Scala, and Python. Performed data transformations and ingestion, optimized performance using Spark's in-memory capabilities, partitions, and broadcasting, and orchestrated workflows using Azure Automation Accounts, GCP Cloud Composer, and Tidal Scheduler.
- Designed and optimized scalable data pipelines using SSIS, YARN, and Spark, integrating with Azure, AWS, and GCP for efficient data transformation and reporting.
- Created efficient data models and schemas for analytics and reporting in platforms like Snowflake, Google BigQuery, and Amazon Redshift. Developed complex SQL queries and stored procedures, optimizing query performance using advanced techniques such as query plan analysis and optimization.
- Implemented data quality checks and cleansing processes using Azure Data Quality Services, GCP Dataflow, and other data profiling tools. Automated data ingestion, transformation, and quality checks using scripting languages like Bash and PowerShell to enhance pipeline performance and integration across Azure and GCP environments

Environment: SQL, SSIS, SSAS, Azure, PowerShell, Power BI MapReduce, HBase, JSON, Spark, Hive, Pig, Hadoop YARN, Spark Core, Spark SQL, Scala, Python, Java, Hive, Sqoop, Impala, Oracle, Google BigQuery, Cloud Dataflow.

Perfexion – Data Associate; Hyderabad, India **May 2021 – June 2022**

Project Description:

The comprehensive customer profile integrates data from diverse touchpoints, including physical stores, online platforms, and social media, using an AWS-powered data lake for seamless data ingestion and storage. This centralized architecture links customer data (CBBID) across systems, enabling teams to drive personalized marketing, increase customer loyalty, and enhance share of wallet. Leveraging AWS services like Redshift and S3, the solution empowers data-driven insights and optimizes customer engagement strategies. This results in improved customer lifetime value and business growth through scalable, cost-effective solutions.

Responsibilities and Contribution:

- Performed data transformations, cleaning, and filtering on imported data using Spark DataFrame API, Hive, and MapReduce, and loaded the final data into Hive. Developed Spark scripts with Python Shell commands as per requirements and used AWS EMR for distributed processing. Utilized Pig as an ETL tool to perform transformations, event joins, filters, and pre-aggregations.

- Led the migration of the existing application to AWS, leveraging AWS services such as EC2 and S3 for large dataset processing. Worked extensively with AWS Elastic MapReduce (EMR) to manage distributed data processing, improving overall workflow efficiency.
- Gained hands-on experience with AWS services like EC2, S3, and AWS Data Pipeline to optimize large-scale data processing and storage. Automated data ingestion and transformation workflows and integrated data into Amazon Redshift and AWS Data Warehouse for business analytics.
- Worked with NoSQL databases like HBase and integrated data from multiple sources into AWS Data Warehouse solutions. Leveraged AWS Databricks for scalable data processing, machine learning, and real-time data analytics, ensuring high data availability and performance across the system.

Environment: AWS, AWS Databricks, AWS Data Warehouse, NoSQL, AWS Lambda, EC2, S3, EMR, MapReduce, Hive, Pig, Spark, Python, PySpark, Linux, Hadoop, Shell Scripting, PL/SQL, Agile Methodologies.

HDLC Info Tech – Data Science Intern; Chennai, India **April 2020 – April 2021**

Project Description:

The project involved building a comprehensive data privacy and transformation framework on Snowflake Cloud to ensure regulatory compliance with stringent privacy laws. It focused on handling large-scale data platforms, securing sensitive personal information, and automating data governance processes. The project also required the development of high-performance data pipelines integrated with Apache Spark, enabling scalable, real-time analytics. By leveraging Snowflake's native data-sharing capabilities, the framework streamlined data access across multiple teams while maintaining privacy and security standards.

Responsibilities and Contribution:

- Architected and implemented robust data transformation pipelines on Snowflake Cloud, optimizing data processing to meet client expectations for scalability and accuracy, leveraging Snowflake's native data-sharing features to streamline access across multiple teams.
- Configured, monitored, and automated Snowflake tasks and services, ensuring seamless data flow between Cloud Storage and Snowflake, and developed advanced data pipelines integrating Snowflake with Apache Spark for high-performance ETL processes.
- Designed and deployed scalable data models and clustering algorithms using Snowflake and AI-driven frameworks, automating job scheduling with Airflow for timely data processing and accurate reporting.
- Utilized Snowflake's data warehousing capabilities to convert raw data into actionable insights, building frameworks for aggregating large datasets and developing dynamic dashboards for real-time business analytics.

Education

- Kent State University, USA | MS in computer science **August 2022 – December 2023**
- Hindustan Institute of Technology and Science, India | Bachelors in CS **June 2018 – May 2022**

Technical Skills

I am skilled in programming languages such as **C, C++, Java, R, SQL, PL/SQL, Python, Apache Pig, HiveQL, Scala, Shell Scripting, MATLAB,** and **Machine Learning**. Experienced with Big Data technologies including **HDFS, MapReduce, Hive, Pig, Sqoop, Flume, Oozie, Kafka, Cassandra, Pyspark, Apache Spark, Spark Streaming, HBase, Impala, Zookeeper,** and **AWS**. Proficient in cloud platforms like **Azure, AWS, GCP, and Snowflake.** Familiar with web technologies (**HTML, CSS, JavaScript, XML**), and various operating systems (**UNIX, LINUX, UBUNTU, CENTOS**). Experienced with application servers like **WebLogic, WebSphere Application Server, WebSphere Portal Server, JBOSS**. Knowledgeable in automation tools (**SBT, Ant, Maven**), **version control** (**GIT**), and **databases** (**Oracle, SQL Server, PL/SQL, MySQL, MS Access, HBase, MongoDB, Teradata**). Skilled in data visualization tools (**Power BI, Tableau) and design tools (Photoshop, UNITY**).

Certifications

- **Microsoft Certified: Azure Fundamentals(AZ-900)**
- **Microsoft Certified: Azure Developer Associate(AZ-204)**
- **AWS Certified Data Analytics – Specialty**
- **Google Professional Data Engineer**