**Gopi Goud Mareedu**

**Sr. Data Engineer**

**Professional Summary:**

* **Senior Data Engineer** with Around **11 years of experience** in designing, developing, and managing **cloud-based data warehousing, big data solutions, ETL/ELT pipelines, and business intelligence applications** across various industries.
* Expertise in **AWS (S3, Redshift, DynamoDB, EMR, Lambda, Kinesis, RDS, ECS, CloudFormation)** and **Azure (Data Factory, Synapse Analytics, Data Lake, Cosmos DB, Databricks)** for scalable and high-performance data solutions.
* Hands-on experience in **Hadoop Ecosystem, Cloudera, Apache Spark (PySpark/Scala), Hive, Kafka, and HDFS**, optimizing large-scale data processing.
* Strong database management skills across **SQL (PostgreSQL, MySQL, MSSQL Server, Oracle)** and **NoSQL (Cassandra, MongoDB, DynamoDB)** with expertise in query optimization, indexing, and data modeling.
* Experienced in designing and automating **ETL & Data Pipelines** using Apache Airflow, DBT, Informatica PowerCenter, Talend, and DataStage.
* Led **data platform modernization initiatives** using **Snowflake, Databricks, and DBT**, enhancing data processing, analytics, and reporting for enterprise-scale solutions.
* Deep knowledge of **data modeling & warehousing**, including fact/dimension modeling, star and snowflake schemas, and data lake architectures.
* Skilled in **Python (Pandas, NumPy, SciPy, SQLAlchemy), Shell Scripting, Spark RDD, DataFrame API, and Spark Streaming** for data processing and workflow automation.
* Proficient in **DevOps & CI/CD** using Terraform, Ansible, Docker, Kubernetes, and Git-based pipelines for seamless deployment and infrastructure automation.
* Strong understanding of **data governance & security**, including IAM roles, encryption, GDPR/CCPA compliance, and cloud security best practices.
* Expertise in **business intelligence & data visualization** with Tableau, Power BI, QlikView, and SAS for creating impactful reports and dashboards.
* Optimized **SQL queries** and **Spark jobs** for complex data transformations in **Parquet** and **Delta** formats, leading to improved job execution time and resource utilization.
* Skilled in **software development & API engineering**, developing REST APIs with Django & Flask for data integration and transformation.
* Experienced in implementing **master data management (MDM)** solutions using **Semarchy xDM**, ensuring data consistency, integrity, and governance across enterprise systems.
* Experience working in **Agile & SDLC methodologies**, leading data-driven projects from design to deployment.
* Hands-on expertise in **AI-driven data solutions**, integrating AI-powered analytics into data pipelines for predictive insights and automation.
* Strong problem-solving skills in **performance optimization & scalability**, including Spark tuning, query optimization, and efficient data processing.
* Passionate about **architecting robust data solutions, streamlining workflows, and leveraging AI to drive business impact**.

**Technical Skills:**

|  |  |
| --- | --- |
| **Category** | **Skills & Tools** |
| Programming Languages | Python, SQL, Scala, Java, R, Shell Scripting |
| Cloud & Infrastructure | AWS (EC2, S3, Lambda, Redshift, DynamoDB, CloudFormation, Kinesis, EMR), Azure (Data Factory, Synapse Analytics, Data Lake, Cosmos DB), GCP (BigQuery, Dataflow) |
| Data Engineering & Big Data | Apache Spark (PySpark, Scala), Hadoop, Kafka, Airflow, NiFi, Delta Lake, Snowflake, Data Lake Architecture |
| Databases (SQL & NoSQL) | PostgreSQL, MySQL, SQL Server, Oracle, MongoDB, DynamoDB, Cassandra, Cosmos DB, AWS Redshift |
| ETL & Data Pipelines | Apache Airflow, DBT, Informatica PowerCenter, Talend, DataStage, AWS Glue, AWS Lambda, AWS Kinesis |
| Data Modeling & Warehousing | Fact & Dimension Modeling, Star & Snowflake Schemas, Data Vault, Data Governance, Collibra |
| Machine Learning & AI | TensorFlow, Scikit-learn, Basic Classification, Regression Models, Data Preprocessing, Feature Engineering, Model Evaluation, Feature Engineering |
| DevOps & CI/CD | Jenkins, GitLab CI/CD, Terraform, CloudFormation, Ansible, Docker, Kubernetes |
| Security & Compliance | IAM Roles, OAuth, JWT, RBAC, Data Encryption, GDPR/CCPA, AWS CloudTrail, KMS |
| Scripting & Automation | Python Automation, Shell Scripting, AWS Lambda Functions |
| Data Science & Analytics | R, NumPy, Pandas, SciPy, Matplotlib, Power BI, Tableau, AWS QuickSight |
| Version Control | Git, GitHub, Bitbucket |

**PROFESSIONAL EXPERIENCE:**

**First Citizens Bank,** **Raleigh, NC Jan 2023- To date**

**Sr. Data Engineer**

**Key Responsibilities:**

* Designed and implemented a **scalable data architecture** on **AWS**, using **Amazon S3, AWS Glue, and Redshift** to build a centralized data lake, optimizing storage and analytics for large, diverse datasets.
* **Developed automated ETL pipelines** using **AWS Glue, Apache NiFi, PySpark, and Lambda**, which reduced manual processing by **70%**, significantly improving the efficiency of data ingestion, transformation, and movement.
* Led data processing efforts with **PySpark** and **Databricks**, ensuring large-scale transformations were optimized for business intelligence and real-time analytics.
* **Built real-time streaming data solutions** using **AWS Kinesis** and **PySpark Streaming**, reducing latency and enabling quick, actionable insights to drive business decisions.
* Integrated **chatbot functionality** with **AWS Lex** and **Python**, automating routine tasks and improving customer interaction efficiency.
* Optimized data warehouses on **AWS Redshift and Snowflake**, improving **query performance** and ensuring scalability for large-scale analytics.
* **Implemented cost-saving measures** through **AWS best practices** like **reserved instances** and **auto-scaling**, reducing infrastructure costs by **20%** while maintaining scalability.
* Enabled **self-service analytics** by integrating **AWS Athena** and **QuickSight**, empowering business teams to access and interpret data independently, leading to better decision-making.
* **Ensured data security and compliance** by setting up **AWS IAM, CloudTrail**, and **KMS**, safeguarding sensitive data and aligning with industry standards.
* Developed and managed **CI/CD pipelines** with **Jenkins, Docker, and Kubernetes**, enabling faster and more reliable deployment cycles.
* Developed **batch and real-time data pipelines using PySpark and AWS Glue**, ensuring smooth data ingestion from multiple sources and enabling high-performance processing.
* **Enhanced data governance** practices with **Collibra**, improving data lineage and governance accuracy by **50%** and minimizing manual oversight.
* **Mentored junior engineers**, providing guidance on cloud-based data architectures, helping them adopt best practices in data engineering and development.
* Deployed multi-terabyte data warehouses on **AWS Redshift**, optimizing storage and retrieval for large-scale analytics and business intelligence operations.
* Automated **customer segmentation models** using **Python** and **AWS Lambda**, improving marketing campaign targeting and increasing customer engagement.
* Worked closely with data architects to design **Star and Snowflake schemas**, ensuring that the data warehouse supported efficient business intelligence and reporting needs.
* **Optimized Kafka infrastructure** using **Kafka Manager** and **Prometheus**, improving monitoring and performance, and ensuring minimal latency in real-time data streaming.
* Led integration projects for **Epic EMR** and **Workday**, ensuring seamless data transfer between systems for more accurate and timely operational insights.
* Built **interactive dashboards** in **Tableau** and **QuickSight**, providing leadership with key performance insights for improved decision-making.
* Used **Terraform** to automate cloud resource provisioning, improving infrastructure management and reducing deployment time by **30%**.
* Implemented **metadata management** tools to enhance data discovery and improve data quality, enabling better traceability and auditability across data systems.

**Environment:** AWS, S3, Glue, Lambda, Kinesis, Redshift, Athena, CloudTrail, IAM, CloudFormation, QuickSight, Spark, Spark SQL, Databricks, Kinesis Firehose, Python, Scikit-Learn, Jenkins, Docker, Kubernetes, Terraform, PostgreSQL, MongoDB, DynamoDB, RDS, Collibra, NiFi, PySpark, Snowflake, Tableau.

**South West Airines, Dallas, TX Aug 2021 – Dec 2022**

**Sr Data Engineer**

**Responsibilities:**

* Led the design and development of **ETL/ELT pipelines** using using PySpark, **AWS Glue**, **AWS Lambda**, and **Amazon Kinesis**, enabling efficient data ingestion and transformation while ensuring scalability and low-latency processing for real-time decision-making.
* Architected secure **data storage solutions** with **Amazon S3**, **Amazon Redshift**, and **Amazon RDS**, optimizing cost and query performance through data partitioning, compression, and indexing to meet business needs.
* Ensured full **compliance with HIPAA** and other security regulations, implementing strong **IAM controls**, data encryption, and secure **VPC architectures** to protect sensitive healthcare data.
* Migrated legacy systems to modern platforms like **Snowflake** and **OBIEE**, improving reporting accuracy and system performance, resulting in faster and more reliable business insights.
* Developed automated data workflows using **Python**, **AWS Glue**, and **Lambda**, streamlining data processing and eliminating manual intervention for faster, error-free pipeline execution.
* Utilized **Databricks** and **Delta Lake** to build real-time data analytics frameworks, enhancing the processing and performance of high-volume datasets in a cost-efficient manner.
* Created reusable **PySpark UDFs (User Defined Functions)** to perform complex business transformations and standardize data processing across multiple use cases
* Automated cloud infrastructure with **AWS CloudFormation** and **Terraform**, ensuring consistent, reliable deployments across all environments, supporting Southwest's scalability needs.
* Collaborated with cross-functional teams to implement a **data governance framework**, ensuring **data quality**, integrity, and regulatory compliance to enhance trust across the organization.
* Optimized data pipelines and SQL queries, reducing processing time by 40%, enabling faster insights and improving the efficiency of key business operations.
* Spearheaded the implementation of a comprehensive **data governance framework** using **AWS Glue Catalog** and **Apache Atlas**, ensuring HIPAA compliance, enhancing data quality, and improving data discoverability across the organization.
* Architected and optimized **data models** within **Snowflake** and **Databricks**, enabling real-time analytics on large-scale datasets, improving decision-making speed, and reducing processing times by 40%.
* Empowered business teams by creating intuitive **dashboards** and reports in **AWS QuickSight** and **Tableau**, promoting **self-service analytics** and enabling data-driven decisions.
* Managed and optimized **Kafka clusters** for real-time data processing, ensuring high availability and minimal data disruption with automated monitoring and alerting systems.
* Architected and maintained a high-performance **data warehouse** in **Confidential Redshift**, managing millions of records daily and ensuring rapid data availability for mission-critical applications.
* Applied optimizations to **AWS Glue** and **Spark** applications, reducing runtime by 30% and improving the speed of data processing for large-scale analytics.
* Managed and optimized Kafka clusters for real-time data processing, ensuring high availability and minimal data disruption with PySpark-based monitoring solutions.
* Led cost optimization initiatives with **AWS Cost Explorer** and **Terraform**, automating multi-region infrastructure deployments and achieving a 30% reduction in AWS data processing costs while improving provisioning efficiency.
* Led the migration of on-premise infrastructure to **AWS**, utilizing **AWS Data Migration Services** to cut infrastructure costs by 40% and enhance scalability and flexibility.
* Designed dynamic **dashboards** using **Tableau**, **Power BI**, and **AWS QuickSight**, providing real-time insights for senior leadership and improving decision-making speed.
* Designed robust disaster recovery and backup strategies, incorporating **PySpark-based data replication techniques** to ensure data resilience and minimize risks.

**Environment:** AWS Glue, AWS Lambda, Amazon Kinesis, Amazon S3, Amazon Redshift, AWS CloudFormation, Terraform, Python, AWS SageMaker, Databricks, Delta Lake, Tableau, Power BI, Snowflake, Kafka, OBIEE, HIPAA, AWS QuickSight, Spark, Hadoop, AWS Data Migration Services.

**State of PA, PA**  **Aug 2018 – July 2021**

**Hadoop Spark Developer**

**Responsibilities:**

* **Ensured data integrity** by identifying outliers and inconsistencies, performing necessary **data transformations** to maintain **quality** and **integration** in a fast-paced **government data processing environment**.
* Developed and optimized **data pipelines** using **Sqoop**, **Spark**, and **Hive** to efficiently ingest, transform, and analyze **operational data** for a variety of state departments.
* Used **Spark SQL** with **Scala** to create and manipulate **data frames**, applying complex **transformations** to meet the analytical needs of business users.
* Built custom **multi-threaded Java** ingestion jobs and leveraged **Sqoop** for seamless **data extraction** from **FTP servers** and **data warehouses**, ensuring smooth data flows into Hadoop.
* Implemented **real-time data streaming** using **Spark** and **Kafka**, enabling **instant data processing** for critical, time-sensitive decision-making processes.
* **Troubleshot and enhanced Spark applications**, improving **error tolerance** and **system reliability** in a highly dynamic, **data-driven environment**.
* **Optimized Spark applications**, improving **processing times** and boosting overall **pipeline efficiency**, ensuring data is processed swiftly for urgent governmental needs.
* Enhanced **Hive** and **Pig** functionalities by creating custom **User Defined Functions (UDFs)**, **Table-Generating Functions (UDTFs)**, and **Aggregating Functions (UDAFs)** using **Python**, extending the capabilities of the state’s data processing systems.
* Developed **Kafka producers** to stream data from **external REST APIs** into **Kafka topics**, ensuring **consistent data flow** across multiple systems.
* Built **Spark Streaming** applications to consume data from **Kafka topics** and write processed data to **HBase**, enabling real-time data availability for reporting and analysis.
* Leveraged **Spark's in-memory processing** capabilities, utilizing **broadcast variables** and **optimized joins** to handle **large datasets**, improving both speed and efficiency.
* Managed high-volume data streams with **Kafka**, processing **thousands of megabytes per second**, ensuring low-latency **real-time data processing** for state operations.
* Wrote **Pig Latin scripts** to **sort**, **group**, **join**, and **filter** enterprise-wide data, enabling seamless data processing across multiple state systems.
* Utilized **Spark API** over **Cloudera Hadoop YARN** to perform **analytics** on **Hive data**, improving data accessibility and performance in the state’s data analytics infrastructure.
* Worked extensively with **Sqoop** to import data from **Oracle** into Hadoop, ensuring smooth migration and integration of critical government data systems.
* Created and managed **Hive tables**, implemented **dynamic partitions** and **bucketing**, and used **Hive QL** to enable scalable, efficient data analysis for the state’s data repository.
* Used **Maven** to build **JAR files** for **MapReduce programs**, facilitating efficient deployment to the cluster in a **high-performance, big data environment**.
* Led the **migration of data** from **legacy RDBMS** systems to **HDFS** using **Sqoop**, modernizing the state's data storage architecture and improving long-term data accessibility.
* Continuously optimized system performance, enhancing **operational efficiency** by identifying bottlenecks and implementing improvements in a **government sector environment**.
* Utilized **Spark SQL** to process data from various file formats such as **JSON**, **text**, and **Parquet**, ensuring flexibility and compatibility in the state’s data processing ecosystem.
* Led **proof-of-concept (POC)** projects with **Apache Spark** using **Scala**, evaluating its capabilities for large-scale data processing and driving **innovation** in government data initiatives.

**Environment**: Sqoop, Spark, Hive, Spark SQL, Scala, Kafka, HBase, Java, Pig Latin, Python, Apache Spark Streaming, Cloudera Hadoop YARN, Maven, Oracle, HDFS, MapReduce, JSON, Parquet, REST APIs, Data Transformation, Data Quality Assurance, Multi-threading, Data Integration, Real-time Data Processing, High-Volume Data Handling, Performance Optimization, UDFs, UDTFs, UDAFs.

**Dollar General- Goodlettsville, TN Jan 2017 – July 2018**

**Data Engineer**

**Responsibilities:**

* Developed and implemented **scalable ETL pipelines** using PySpark on **Azure Databricks**, transforming large datasets and integrating data from cloud and on-premises sources, improving **data availability** for business intelligence.
* Led the **design and optimization of data workflows** in **Azure Data Factory**, automating the **extraction, transformation, and loading (ETL)** of transactional data from **POS systems, inventory databases**, and **external supply chain platforms**, enhancing operational efficiency.
* Spearheaded the **migration of legacy data workflows** to Azure, designing and implementing cloud-based infrastructures using **Azure Synapse Analytics** and **Azure SQL Data Warehouse**, ensuring scalability for growing retail data volumes.
* Implemented **real-time data processing** using **Apache Kafka** and **Azure Spark Streaming**, streamlining the flow of **sales transactions** and **inventory data**, leading to improved demand forecasting and supply chain efficiency.
* Developed **advanced Spark applications** using **Scala and PySpark**, performing large-scale **data transformations** and **denormalization** of relational datasets, providing **actionable insights** for various business teams.
* Leveraged **Apache Kafka** to build **real-time data pipelines**, consuming data from multiple sources and **reducing data latency**, enabling faster and more accurate decision-making in retail operations.
* **Optimized HBase tables** and **Hive partitions** to improve **query performance**, ensuring **efficient storage** and fast retrieval of structured, semi-structured, and unstructured data for analytics.
* Utilized **AWS Glue and PySpark** for **data transformation** and **cataloging**, scheduling **workflows and crawlers** to ensure the timely delivery of **clean, transformed datasets** for analytics teams.
* Migrated **batch processing workflows** from **Oozie to Apache Airflow**, automating the scheduling and execution of **incremental data loads**, reducing delays and improving data pipeline efficiency.
* Implemented **Pig Latin scripts** for data transformations, leveraging **Hadoop ecosystem tools** such as **Sqoop, Flume, and Hive** for **seamless data integration** across distributed storage systems.
* Created **Spark SQL queries** on **Hive tables**, ensuring structured data was **readily available** for downstream analytics and reporting.
* Built and maintained **interactive Power BI dashboards**, integrating data from multiple sources to provide key insights on **sales trends, inventory health, and customer engagement**, supporting data-driven decision-making.
* Led **data capacity planning** and **node forecasting**, ensuring the **Hadoop cluster environment** (including **HBase and Hive**) was optimized for **high-volume data processing** and storage.
* Enhanced **data pipeline performance** using **compression mechanisms** and optimized **MapReduce jobs**, ensuring **efficient use of HDFS and Hadoop cluster resources**.
* Configured and **maintained Cloudera Hadoop Distribution**, supporting **large-scale data processing**, improving the speed and reliability of data ingestion and processing workflows.

**Environment:** Azure Databricks, PySpark, Spark SQL, Scala, Hadoop (HDFS, MapReduce, YARN), Cloudera Hadoop Distribution, Azure Data Factory, Azure Synapse Analytics, Cosmos DB, AWS Glue, AWS S3, AWS Redshift, Apache Kafka, Apache Spark Streaming, Apache Airflow, Oozie, HBase, Hive, Pig, Sqoop, Flume, DynamoDB, Power BI, Azure DevOps, Kubernetes (AKS), SQL Server, MySQL, Oracle, Snowflake, Python, Scala.

**Omya Health Care, NJ Aug 2014 – Dec 2016**

**Big Data Developer**

**Responsibilities:**

* Led the design and implementation of large-scale **Hadoop data pipelines** to process and analyze healthcare data, integrating **HDFS**, **Pig**, **Hive**, and **HBase** for efficient data storage and retrieval.
* Migrated legacy data processing systems from **Hadoop** to **Apache Spark**, optimizing healthcare workflows and reducing processing times, resulting in faster decision-making.
* Developed and managed **Oozie workflows** to automate and schedule **ETL** processes, integrating **Hive**, **Pig**, and **Sqoop** jobs to ensure seamless data processing.
* Configured **Flume** for efficient log file collection from web servers into **HDFS/Hive**, implementing features like multiplexing, replication, and interceptors for optimized data flow.
* Automated job scheduling and management on the Hadoop cluster using **Shell scripting**, ensuring consistent and error-free execution.
* Coordinated and monitored cluster health and performance using **Zookeeper** to ensure high availability and stability of the Hadoop ecosystem.
* Worked with **Hive** to filter partitioned data based on various year ranges and formats, leveraging advanced **Hive functions** to optimize query performance.
* Designed and optimized **Pig** scripts for large-scale data ingestion, sorting, filtering, and joining, improving data processing efficiency.
* Developed **HBase** tables for random read/write operations via **MapReduce**, facilitating high-speed data access.
* Built end-to-end data pipelines from ingestion to reporting, ensuring integration with analytics platforms and streamlined data processing.
* Contributed to predictive analytics solutions by integrating **Apache Spark**, **SQL/HiveQL**, **JavaScript**, and **High Charts** for visual reporting and decision-making.
* Implemented data cleansing, transformation, and reduction techniques using **MapReduce** jobs to enhance data quality for downstream analytics.
* Created and optimized **Hive tables** and queries to ensure performance efficiency for reporting and analytical needs.
* Utilized **Cloudera Manager** for monitoring and maintaining the performance and health of the Hadoop cluster, ensuring data availability and processing accuracy.

**Environment**: HDFS, CDH, Apache Spark, Hive, HBase, Flume, Pig, Sqoop, SQL, Java, Python, Cloudera Manager, Oracle, Tableau, Apache Kafka, AWS, Healthcare Data Standards (HL7, FHIR)