Sai Kiran Reddy

(Data Engineer)
***Ph.:** +1(412) 451-5660*
***Email:** Kirande2097@gmail.com*
***LinkedIn Id:** https://www.linkedin.com/in/sai-kiran-reddyoptum/*

PROFESSIONAL SUMMARY:

- Around 5+ years of technical experience as a Data Engineer in business need of clients, developing effective and efficient solutions.
- Extensive expertise and understanding of **Hadoop** ecosystem products such as **HDFS, MapReduce, YARN, Spark, Kafka, Hive, Sqoop, Pig, Impala, HBase.**
- Strong understanding/knowledge of Hadoop architecture and concepts such as **HDFS, MapReduce**, Job Tracker, Task Tracker, Name Node, and Data Node.
- Working knowledge of Cloudera, Azure HDInsight, and the AWS cloud.
- Converted **SQL** queries into Spark Transformations utilising **Spark RDDs, Scala**, and performed map-side joins on RDDs.
- Worked on **AWS Redshift** and RDS for model and data implementation on RDS and **Redshift.**
- End-to-end writing expertise Processing of Data analysis jobs utilising **MapReduce, Spark, and Hive.**
- Proficient in designing and implementing **ETL processes** using Snowflake, ensuring efficient data ingestion, transformation, and storage across large datasets.
- Extensive experience implementing production-ready Spark applications with Spark components like **as Spark SQL, MLlib, Spark Streaming**, and Graph X.
- Proficient in **Apache Flink** for real-time stream processing, building scalable ETL pipelines to handle high-throughput event streams.
- Skilled in creating high-performance data intake pipelines using **Azure Data Factory** and **Azure Databricks**, integrating data from multiple sources to support analytics-based solutions targeting consumer subscribers directly.
- Strong familiarity with Amazon cloud web services such as **EMR, Redshift, DynamoDB, Lambda, Athena, S3, RDS**, and **CloudWatc**h for effective large data processing.
- Create **ETL** processes in **AWS Glue** to move Campaign data from external sources like as **S3, ORC/Parquet/Text** Files into **AWS Redshift**.
- Experience extracting files from **MongoDB** using **Sqoop,** storing them in **HDFS**, and then processing them.
- Worked with several ingestion services to handle batch and real-time data using **Spark** streaming, **Kafka** Confluent, Storm, Flume, and **Sqoop**.
- Skilled in executing data migration strategies to Snowflake, ensuring seamless integration and improved performance for analytics projects.
- Extensive experience in data migration projects utilizing **SQL, SQL Azure, Azure Storage, Azure Data Factory, SSIS**, and **PowerShell**, focusing on seamless data transfers and infrastructure optimization.
- Hands-on expertise with **NoSQL** databases such as **HBase, Cassandra**, and **MongoDB**, as well as their interaction with **Hadoop** and Kubernetes clusters.
- Hands-on experience with **Apache Airflow** and the **Oozie** workflow engine for managing and scheduling **Hadoop** tasks.
- Experience integrating **Apache Flink** with **Kafka** and other streaming platforms to perform near real-time data transformations and analytics.
- Constructed excellent database architectures by migrating data from on-premises **SQL** databases to **Azure Synapse Analytics** using **Azure Data Factory**, emphasizing security, efficiency, and compliance.
- Designing tableau dashboards with high data volumes from a **SQL server** data source.

- Python skills include **NumPy, SciPy, Pandas, Scikit-learn, Matplotlib**, and **TensorFlow.**
- Extensive expertise with relational databases such as **MySQL, MS SQL,** and Oracle in creating stored procedures and complicated **SQL** queries.
- Integrated **Power BI** with **Azure Data Lake Storage** and **Azure SQL** Database to deliver seamless end-to-end data analytics solutions, enhancing data visualization and business intelligence capabilities.
- Design and implement data models within **Cosmos DB** to meet application requirements.
- Working knowledge of **RESTful** web services and **RESTful API** development.
- Participated actively in all scrum rituals, including Sprint Planning, Daily Scrum, Sprint Review, and Retrospective meetings, as well as assisting the Product Owner in designing and prioritising user stories.
- Working with **UNIX/LINUX** environments and developing shell scripts is a must.
- Worked on several stages of the Software Development Life Cycle, such as development, component integration, performance testing, deployment, and support maintenance.

TECHNICAL SKILLS:

Programming & Scripting	Python, Scala, Java, SAS, R, SQL, MATLAB, HiveQL, PowerShell and BASH Scripting
Cloud Technologies	Azure Cloud. Amazon AWS
Big Data Ecosystem	HDFS, YARN, MapReduce, Sqoop, Hive, Oozie, Pig, Spark, Kafka, Nifi, Connect, Airflow, Stream Sets, Kafka connect, confluent, DBT
MySQL Libraries	Scikit-learn, SciPy, Pandas, and NumPy, TensorFlow,
Frameworks	Django, Flask, NodeJS, ReactJS, Angular
IDE Tools	Eclipse, PyCharm, Visual Studio Code
CI/CD/Build Tools	Jenkins, Maven, Ant
Version Control	Git, SVN
BI Reporting Tools	Power BI
SQL Databases and ORM	Oracle, MySQL, Teradata, Postgres, Django ORM, SQL Alchemy
NoSQL Database	HBase, Cassandra, PostgreSQL Dynamo DB, Cosmos, MongoDB
Operating Systems	Ubuntu, Mac OS-X, CentOS, Windows.

Professional Experience:

Charles Schwab- Westlake, TX
Data Engineer
January 2023 to Present
Responsibilities:

- Managed data transfer between AWS compute and storage services utilizing AWS Data Pipeline to streamline data workflows.
- Developed **Scala** scripts and UDFs in Spark for data aggregation, querying, and storage into **S3** buckets, leveraging both Data Frames/SQL and RDDs for efficient data handling.
- Executed **ETL** processes with ingested data by developing and running Scala programs in **Apache Spark**, optimizing data transformation and loading.
- Configured AWS Data Pipeline for efficient data transfers from **S3** to Redshift, enhancing data warehousing processes.
- using **Apache Flink** to handle high-throughput event streams, enabling real-time analytics and decision-making for mission-critical business processes.
- Designed and implemented data ingestion processes into **Cosmos DB** from multiple sources, streamlining data integration and accessibility.
- Extracted, aggregated, and merged Adobe data within **AWS Glue** using PySpark, demonstrating proficiency in complex data processing.
- Integrated Unix-based systems with databases, proficient in executing **SQL** queries via command line and managing connections for effective data retrieval and storage.

- Utilized MongoDB for real-time data ingestion and analytics to support time-sensitive data processing tasks, enabling rapid decision-making for business operations.
- Developed scripts and indexing strategies for migrating from **SQL** Server and **MySQL** to Confidential Redshift, optimizing data transition and storage efficiency.
- Worked with **PL/SQL** to streamline ETL processes, ensuring efficient data management and workflow optimization.
- **Flink** to perform low-latency transformations, aggregations, and windowed operations on streaming data from sources like **Apache Kafka** and **AWS Kinesis**.
- Created a solid **ETL** pipeline for migrating data from an on-premises Cloudera cluster to **AWS EC2/EMR,** and architected a log processing solution with **PySpark** into an **AWS S3 Data Lake**
- Developed automated ingestion scripts in **Python** and **Scala**, integrating data from **API**s, **AWS S3**, Teradata, and **Snowflake**, to streamline and enhance data acquisition process.
- Created and executed **Spark** routines in **Scala** for efficient data retrieval from **AWS S3** buckets and complex transformations in **Snowflake**.
- Designed and developed the core data pipeline infrastructure using **Python**, laying the foundation for robust data processing.
- Deployed and managed containerized ETL pipelines on Kubernetes, enabling scalable and efficient data processing workflows.
- Developed **Python** scripts for data integration, proficiently reading CSV, JSON, and Parquet files from AWS S3 buckets and efficiently loading them into AWS S3, DynamoDB, and Snowflake.
- Performed time series analysis and handling of date-time data using Pandas' robust time series functionality.
- Implemented **AWS Lambda** functions for event-driven script execution in response to Amazon **DynamoDB** table changes, S3 bucket events, and HTTP requests via Amazon **API** Gateway, optimizing system responsiveness and scalability.
- Integrated Apache Flink with Kafka for real-time event ingestion and processing, enabling the company to react to live events and perform near real-time analytics.
- Implemented complex mathematical functions and statistical operations using NumPy.
- Developed a custom **Scala** utility function for efficient data transfer between **AWS S3** and **Snowflake**, streamlining the data integration process and enhancing data workflow efficiency.
- Managed **Snowflake** schemas and data warehousing, expertly handling both batch and streaming data pipelines from a confidential **AWS S3** data lake using Snowpipe and Matillion.
- Used **Apache Airflow** for orchestrating data workflows across **AWS S3** buckets and Snowflake data warehouses, with expertise in creating and executing DAGs to automate and streamline processes.
- Designed and executed DAGs on EC2 instances using Email, Bash, and Spark Livy operators within **Apache Airflow**, enhancing automation and operational efficiency in data processing workflows.

Environment: Agile Scrum, MapReduce, Snowflake, Pig, Spark, Scala, Hive, Kafka, Python, Airflow, JSON, GCP, Parquet, CSV, Code cloud, AWS.

Weaver - TX
Data Engineer
July 2018- December 2021
Responsibilities:

- Exploring with Spark improving the performance and optimization of the existing algorithms in Hadoop using Spark context, Spark-SQL**, Data Frame**, pair RDD's, **Spark YARN**.
- Recreating existing application logic and functionality in the Azure Data Lake, Data Factory, SQL Database and SQL data warehouse environment.
- Designed and implemented by configuring Topics in the new Kafka cluster in all environments.
- Strong experience working with Spark Data frames, **Spark SQL** and Spark Structured Streaming APIs using Scala. Implemented batch processing of jobs using Spark Scala API's.

- Proficient in utilizing Snowflake features such as **SnowSQL** and **SNOWPIPE** for seamless and continuous data ingestion for analytical purposes.
- Strong experience working with Spark Data frames, Spark SQL and Spark Structured Streaming APIs using Scala. Implemented batch processing of jobs using Spark Scala API's.
- Used Scala collection framework to store and process the complex consumer information. Based on the offer's setup for each client, the requests were post processed and given offers.
- Developed interactive dashboards and reports for business stakeholders.
- Conducted performance tuning and optimization activities on **Snowflake** to enhance query performance, minimize latency, and improve overall system efficiency.
- Developed real time data streaming solutions using Spark Structured Streaming with Scala applications to consume the JSON messages from Kafka topics.
- Extensively worked with automation tools like **Jenkins**, **Artifactory**, **SonarQube** for continuous
- integration and continuous delivery **(CI/CD)** and to implement the End-to-End Automation.
- Involved in creating Hive Tables, loading the data from the **cornerstone** tool.
- Developed and managed data integration workflows to load data from various sources into **Snowflake** using tools like **Snow pipe**, **ETL/ELT** frameworks, and third-party connectors
- Developed data pipeline using EventHub's, PySpark, and Azure SQL database to ingest.
- customer events data and financial histories into Azure cluster for analysis.
- Involved in **converting Hive/SQ**L queries into **Spark transformations** using **Spark RDDs**, by using **Python**.
- Optimized data processing workflows for performance and cost-efficiency using **Databricks**.
- Ensured compliance with **data governance** policies and industry standards.
- Wrote Automation Script to auto start and stop the services in **Azure cloud** for cost saving.
- Automated resulting scripts and workflow using **Apache Airflow** and shell scripting to ensure
- daily execution in production.
- Have knowledge on partition of **Kafka messages** and setting up the replication factors in **Kafka Cluster.**
- Developing Spark applications using **Spark - SQL** in **Databricks** for data extraction, Participating in migration **Scala code** into **Microservices**.
- Responsible to develop the Code and Unit Test and move the code to **UAT** and **PROD**.
- Used Amex Internal Framework **Event engine** to trigger the jobs and monitor the jobs.
- Optimized complex **SQL** queries and materialized views for improved performance and reduced compute costs.
- Automated data workflows using **Databricks** Jobs and integrated with orchestration tools like Apache Airflow for scheduling and monitoring.
- Used **Snowflake's query** profiling and performance monitoring tools to identify and resolve bottlenecks.
- Worked Extensively on Talend Admin Console and Schedule Jobs in Job Conductor.
- Data sources are extracted, transformed, and loaded to generate CSV data files with **Python programming** and **SQL queries**.

Environment: Hadoop, Snowflake, Scala, Data Bricks, data bricks, HDFS, Talend, Azure, Azure Data bricks, Pig, Sqoop, HBase, Shell Scripting, Maven, Jenkins, Ubuntu, Mark Logic, MDM, Linux