Name: Sahithi
Email: sahithi.kjfs@gmail.com
Ph#: 405-237-4698

Professional Summary:

- Having over **5+ years** of experience as **Data Engineer** with strong technical expertise, business experience, and communication skills to drive high - impact business outcomes.
- Experience with **Spark Architecture** including **Spark Core, Spark SQL, Spark Streaming** and **Spark MLlib.**
- Experience in **Spark** using **Scala** for loading data from **local file systems, HDFS, Amazon S3, Relational** and **NoSQL** databases using **Spark SQL,** import data into **RDD, DataFrames** and ingesting the data from a range of sources using **Spark Streaming.**
- Experience building distributed high-performance systems using **Spark and Scala**.
- Experienced in working with spark eco system using **Spark SQL and Scala queries** on different formats like **Text file, CSV file.**
- Hands-on experience in developing and deploying enterprise-based applications using major **Hadoop ecosystem** components like **MapReduce, YARN, Hive, HBase, Flume, Sqoop, Spark MLlib, Spark GraphX, Spark SQL, Kafka.**
- Experience in **Extraction, Transformation, and Loading (ETL)** data from various sources into Data Warehouses and data processing like collecting, aggregating, and moving data from various sources.
- Experienced in ETL concepts, building ETL solutions, and Data modeling. Worked on architecting the ETL transformation layers and writing spark jobs to do the processing.
- Experience in **Tableau** data visualization using Cross tabs, Heat maps, Box and Whisker charts, Scatter Plots, Geographic Map, Pie Charts and Bar Charts and Density Chart.
- Good knowledge of creating visualizations, interactive dashboards, reports, and data stories using **Tableau and Power BI.**
- Experience in writing **Map-Reduce** Jobs in **Python** for processing large sets of structured, semi-structured and unstructured data sets and stores them in **HDFS.**
- Experience in writing data quality checks using **Python, PySpark**.
- Experience in implementing Azure **data** solutions, provisioning storage account, Azure **Data** Factory, SQL server, SQL Databases, SQL **Data** warehouse, Azure **Data** Bricks and Azure Cosmos DB.
- Experience in migrating on premise to Windows Azure using Azure Site Recovery and Azure backups.
- Experienced with Dimensional modelling, **Data** migration, **Data** cleansing, **Data** profiling, and ETL Processes features for **data** warehouses.
- Expertise in building CI/CD on AWS environment using AWS Code Commit, Code Build, Code Deploy and Code Pipeline and experience in using AWS CloudFormation, API Gateway, and AWS Lambda in automation and securing the infrastructure on AWS.
- Experience working with **GitHub/Git** source and version control systems.
- Experienced **dimensional modeling (Star schema, Snowflake schema), transactional modeling,** and **SCD (Slowly changing dimension).**
- Experience in the application of various data sources like **Oracle SE2, SQL Server, Flat Files**, and Unstructured files into a data warehouse.
- Hands on experience in **SQL** and **NOSQL** database such as **Snowflake, HBase, Cassandra** and **MongoDB.**
- Experience in analytical, presentation, communication, problem solving with the ability to work independently as well as in a team and had the ability to follow the best practices and principles defined for the team.

Technical Skills:

Databases	Snowflake, AWS RDS, Teradata, Oracle, MySQL, Microsoft SQL, Postgre SQL.
NoSQL Databases	MongoDB, Hadoop HBase and Apache Cassandra.
Programming Languages	Python, SQL, Scala, MATLAB.
Cloud Technologies	AWS, Docker
Data Formats	CSV, JSON
Querying Languages	SQL, NO SQL, PostgreSQL, MySQL, Microsoft SQL
Integration Tools	Jenkins

Scalable Data Tools	Hadoop, Hive, Apache Spark, Map Reduce, Sqoop.
Operating Systems	Red Hat Linux, Unix, Windows, macOS.
Reporting & Visualization	Tableau, Matplotlib.

Professional Experience:

Client: Verizon Irving TX **May 2023 to Present**
Role: Data Engineer
Responsibilities:

- Worked with the business users to gather, define business requirements and analyze the possible technical solutions.
- Developed **Spark** scripts by using **Scala** shell commands as per the requirements for cleaning and testing.
- Developed **scala** scripts, **UDF's** using both **data frames/SQL and RDD** in **Spark** for data aggregation, queries and writing back into S3 bucket.
- Developed **Tableau** data visualization using Cross tabs, Heat maps, Box and Whisker charts, Scatter Plots, Geographic Map, Pie Charts and Bar Charts and Density Chart.
- Developed highly complex **Python** and **Scala** code, which is maintainable, easy to use, and satisfies application requirements, data processing and analytics using inbuilt libraries.
- Used **Spark** streaming to divide streaming data into batches as an input to Spark engine for batch processing.
- Wrote **Spark** applications for data validation, cleansing, transformation, and custom aggregation and used **Spark engine, Spark SQL** for data analysis and provided to the data scientists for further analysis.
- Worked on developing a **Pyspark** script to encrypt the raw data by using Hashing algorithms concepts on client-specified columns.
- Developed **Python-based API (RESTful Web Service)** to track revenue and perform revenue analysis.
- Performed **ETL** testing activities like running the Jobs, Extracting the data using necessary queries from database transform, and uploading into the Data warehouse servers.
- Used **ETL** to implement the Slowly Changing Transformation, to maintain Historically Data in the Data warehouse.
- Visualized the data with the help of box plots and scatter plots to understand the distribution of data using **Tableau.**
- Build **Data** pipelines using **Python, Apache Airflow for ETL** related jobs inserting **data** into Oracle.
- Creating job flow using **Airflow in python** and automating the jobs. **Airflow** will have separate stack for developing DAGs on and will run jobs on EMR or EC2 Cluster.
- Written queries in **MySQL and Native SQL.**
- Integrated Azure **Data** Factory with Blob Storage to move **data** through DataBricks for processing and then to Azure **Data** Lake Storage and Azure SQL **data** warehouse.
- Pipelines were created in Azure **Data** Factory utilizing Linked Services to extract, transform, and load **data** from many sources such as Azure SQL **Data** warehouse, write-back tool, and backwards.
- Working on **data** management disciplines including **data** integration, modeling and other areas directly relevant to business intelligence/business analytics development.
- Performed Real time event processing of data from multiple servers in the organization using **Apache Storm** by integrating with **apache Kafka.**
- Created automated pipelines in **AWS Code Pipeline** to deploy **Docker** containers in **AWS ECS** using **S3.**
- Used **AWS services** to implement data lake and data warehouse solutions **(AWS Lambdas, AWS Glue, AWS S3, Event bridge, SQS, SNS, RDS, DynamoDB, AWS Kinesis, AWS Athena, etc.).**
- Worked on Snowflake Schemas and Data Warehousing and processed batch and streaming data load pipeline using Snow Pipe and Matillion from data lake Confidential AWS S3 bucket.
- Developed **AWS Cloud** Formation templates and set up Auto scaling for EC2 instances.
- Worked on infrastructure with **Docker containerization.**
- Unit tested the data between Redshift and **Snowflake.**
- Used **SQL** queries and other tools to perform data analysis and profiling.
- Involved **in Agile** development methodology active member in scrum meetings.
- Actively participating in the code reviews, meetings and solving any technical issues.

Environment: Spark, Scala, Pyspark, Python, AWS, Docker, Restful, HDFS, Tableau, Snowflake, Apache Airflow, Apache Spark, Azure, Power BI, ETL, Agile and SQL.

Client: Charter Communications Stamford CT **Jan 2022 to April 2023**
Role: Data Engineer
Responsibilities:

- Worked as Data Engineer to review business requirement and compose source to target data mapping documents.
- Developed **spark** applications for performing large scale transformations and denormalization of relational datasets.
- Developed **Spark scripts** by using **python** commands as per the requirement.
- Implemented advanced procedures like text analytics and processing using the in-memory computing capabilities like **Apache Spark** written in **Scala.**
- Developed **Spark** programs with **Scala** and applied principles of functional programming to do batch processing.
- Worked on **PySpark APIs** for data transformations.
- Responsible for helping the team in daily ongoing issues as per the business needs & requirements.
- Designed, developed and implemented **ETL pipelines** using **python API (PySpark)** of **Apache Spark** on **AWS EMR.**
- Administered user, user groups, and scheduled instances for reports in **Tableau.**
- Monitoring of **Tableau Servers** for its high availability to users.
- Configured, assembled, and dispatched new information pipelines underway utilizing **Apache Spark.**
- Designed an **Apache Airflow** Data Pipeline to automate data ingestion and retrieval.
- Extract, transform, and load **data** from source systems to Azure **Data** Storage services using a combination of Azure **Data** Factory and **Data** Lake Analytics.
- **Data** ingestion to one or more Azure services (Azure **Data** Lake, Azure Storage, and Azure SQL) and processing the **data** in Azure **Data** bricks.
- Wrote **AWS Lambda** functions in **python for AWS's Lambda** which invokes python scripts to perform various transformations and analytics on large data sets in EMR clusters.
- Involved in automation and provisioning services on **AWS**.
- Worked with **Docker**, by using **Docker-compose**.
- Performed analysis on the unused user navigation data by loading into **HDFS** and writing **MapReduce** jobs.
- Creating Reports in Looker based on **Snowflake Connections**.
- Responsible for data services and data movement infrastructures, worked with **ETL** concepts, building **ETL** solutions and **Data modeling.**
- Performed analysis on the unused user navigation data by loading into **HDFS** and writing **MapReduce** jobs.
- Design and Develop ETL Processes in **AWS** Glue to migrate Campaign data from external sources like S3, ORC/Parquet/Text Files into **AWS Redshift.**
- Created and modified several database objects such as Tables, Views, Indexes, Constraints, Stored procedures, Packages, Functions and Triggers using **SQL and PL/SQL.**
- Involved in **Agile** methodologies, daily scrum meetings, spring planning.
- Actively participated and provided feedback in a constructive and insightful manner during weekly Iterative review meetings to track the progress for each iterative cycle and figure out the issues.

Environment: Spark Spark, Scala, Pyspark, Python, AWS, Docker, Tableau, Azure, Apache Spark, Apache Airflow, Snowflake, ETL, Agile and SQL.

Client: Emigrant Bank, NYC NY **Nov 2020 to Aug 2021**
Role: Data Engineer
Responsibilities:

- Involved in Requirement gathering phase to gather the requirements from the business users to continuously accommodate changing user requirements.
- Developed **Spark** scripts by using **Python** in **PySpark** shell command in development.
- Worked on migrating **Map Reduce** programs into **Spark transformations** using **Spark and Scala.**
- Implemented **Spark** Scripts using **Scala, Spark SQL** to access hive tables into **spark** for faster processing of data.
- Responsible for Writing the Data Quality checks, based on the existing source code, using **Python and PySpark dataframe work** in **Databricks** platform (which Improved process time).
- Used **Spark streaming** to receive real time data from the Kafka and store the stream data to HDFS using **Scala** and databases such as **HBase**.
- Responsible for extracting the data and loading the data using the **Python.**

- Created report schedules on **Tableau server.**
- **ETL** Restarting capability for a date or date range or from point of failure or from beginning
- Validated the **Tableau insight** center reports to make sure all the data is populated as per requirements.
- Develop stored procedures/views in **Snowflake** and use in Talend for loading Dimensions and Facts.
- Involved in converting Hive/SQL quries into Spark transformation using **Spark RDDs**.
- Responsible for creating on-demand tables on S3 files using Lambda Functions and AWS Glue using **Python and PySpark.**
- Worked on infrastructure deployment on **AWS** using **EC2 (Virtual Cloud Servers), RDS (Configured Relational Database Service), VPC (Virtual Private Cloud), managed the Network and its Security, Route 53, Cloud Formation, Direct Connect, AWS S3, AWS Ropeworks (operations automation), IAM, Glacier (Cloud storage) & Amazon CloudWatch Monitoring Management.**
- Implemented **AWS** Elastic Container Service (ECS) scheduler to automate application deployment in the cloud using **Docker** Automation techniques.
- Created database maintenance planner for the performance of **SQL Server**, which covers Database integrity checks and re-indexing.
- Involved in **Agile** methodologies, daily scrum meetings, spring planning.
- Participated in meetings, reviews, and user group discussions as well as communicating with stakeholders and business groups.

Environment: Spark, Scala, Pyspark, Python, AWS, Docker, Restful, HDFS, Tableau, Snowflake, ETL, Agile and SQL.

Client: Virtusa Hyd **July 2018 to Oct 2020**
Role: Data Engineer.
Responsibilities:

- Collaborated with Business Analysts, SMEs across departments to gather business requirements, and identify workable items for further development.
- Developed **Spark streaming** application to pull data from cloud to Hive table.
- Efficiently handled periodic exporting of SQL data into Elastic search.
- Developed **ETL jobs** to automating the several students' data from enrollment step to graduation process and service fee calculation for weekly finance payment to Confidential.
- Develop **Python, PySpark** to Transform, and Load data across on premise and cloud platform.
- Create **ETL scripts** for the ad-hoc requests, requests to retrieve data from analytic sites.
- Implement One time Data Migration of Multistate level data from **SQL server to Snowflake** by using **Python and SnowSQL.**
- Collaborated with business owners of products for understanding business needs and automated business processes and data storytelling in **Tableau.**
- Implemented Data pipelines for big data processing using Spark transformations and **Python API** and clusters in **AWS.**
- Implemented a Continuous Delivery pipeline with **Docker.**
- Implemented **SQL, PL/SQL** stored procedures.
- Worked on **Agile** Methodology.
- Performed extensive performance tuning and improvement of ETL jobs and processes.
- Providing 24/7 On-call Production Support for various applications.

Environment: Spark, Scala, Pyspark, Python, AWS, Docker, Restful, HDFS, Tableau, Snowflake, ETL, Agile and SQL.

References: Will be provided upon request.