Sai Charan Asapu
Email: saiasapu23@gmail.com
Ph#: 779-910-3426
LinkedIn : www.linkedin.com/in/asaicharan016
Chicago, IL

Professional Summary:

- **Senior Data Engineer** with **7+ years** of professional work experience in the IT industry with Enterprise Application development involving business analysis, development, maintenance & support.
- Experience in **Software Development Life Cycle (SDLC)**, having a thorough understanding of various phases like Requirements Analysis, Design, Development, and Testing.
- Experience in **Spark ecosystem, core, SQL,** and **Streaming modules.**
- Hands-on experience in using **Spark tools** like **RDD transformations, Spark core, Spark MLlib, Spark Streaming, and Spark SQL.**
- Extensive experience developing **Spark applications** using both Scala and PySpark for batch and streaming data processing.
- Experience in developing **Kafka producers and Kafka consumers** for streaming millions of events per minute on streaming data using **PySpark, Python & Spark Streaming.**
- Experience in developing applications using **MapReduce and Hive.**
- 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 Graph X, Spark SQL, and Kafka.**
- Experience in the design and development of multiple **Power BI** Dashboards and managing data privacy and security in **Power BI.**
- Experience with **Tableau** that is used as a reporting tool.
- Extensive experience with **Informatica (ETL Tool)** for Data Extraction, Transformation, and Loading.
- Experience in **Snowflake** data warehouse, developed data extraction queries, and automatic **ETL** for data loading from Data Lake.
- Experience in developing **Data Cleaning, Data Validation,** and reconciliation scripts to validate the data before and after the data processing.
- Experience with **Azure technologies** - **Azure Data Lake, Azure Data Factory, HD Insights**.
- Experience in migrating SQL database **to Azure Data Lake, Azure Data Lake Analytics, Azure SQL Database, Databricks,** and **Azure SQL Data Warehouse**, controlling, granting database access, and migrating on-premises databases to **Azure Data Lake store** using **Azure Data Factory.**
- Experienced in fact **dimensional modeling (Star schema, Snowflake schema), transactional modeling, and SCD (Slowly Changing Dimensions).**
- Experience in implementing and orchestrating data pipelines using **Airflow.**
- Experience in source code & build management with **Git** & Enterprise **GitHub** with **Jenkins, Artifactory.**
- Experience with best practices of Web services development and Integration **(REST and SOAP).**
- Experience working with **NoSQL** databases such as **MongoDB, Cassandra, and HBase.**
- Experience in writing Complex **SQL queries, PL/SQL, Views, Stored procedure, Triggers, etc.**
- Experienced in **Agile Scrum, Waterfall, and Test-Driven Development (TDD)** methodologies.
- Good analytical, communication, and problem-solving skills, and with a passion for continuous learning.

Technical Skills:

Languages	Python, SQL, Scala
Big Data	Apache Spark (PySpark, Spark SQL), Hadoop, Hive, HDFS
Streaming	Apache Kafka, Spark Streaming
ETL & Orchestration	Azure Data Factory (ADF), Data Stage, Apache Airflow,
Cloud	• Azure (ADLS, Databricks, Azure SQL, Synapse, Blob Storage, • AWS (S3, EC2)
Data Warehousing & Modeling	Snowflake, Star Schema, Dimensional Modeling, SCD
Databases	SQL Server, Oracle, MongoDB, Cassandra
Visualization	Power BI, Tableau
DevOps & Tools	Docker, Kubernetes, Git, Jenkins, Jira
Monitoring	Prometheus, Grafana
Methodologies	Agile (Scrum), SDLC

Professional Experience:

Client: Jefferson Bank, San Antonio, TX. || Duration: Jul 2024 – Till Data
Role: Senior Data Engineer
Responsibilities:

- Involved in requirements gathering, analysis, design, development, change management, and deployment.
- Developed **Spark** applications using **Spark SQL** in Databricks for data extraction.
- Developed Spark and Spark SQL applications in Databricks to extract, transform, aggregate, and optimize large-scale datasets from multiple structured and semi-structured sources.
- Worked on migrating **MapReduce** programs into **Spark** transformations, initially done using **Python (PySpark).**
- Designed and developed **Hive** data transformation scripts to work against structured data from various data points and created a baseline.
- Worked on a direct query using **Power BI** to compare legacy data with current data, generated reports, and created dashboards.
- Worked with **Tableau** for generating reports and creating **Tableau dashboards, pie charts, and heat maps** according to the business requirements.
- Worked on architecting the **ETL** transformation layers and writing Spark jobs to do the processing.
- Developed **Star and Snowflake schemas** based on a dimensional model, growing the data warehouse
- Worked on **ER Modeling, Dimensional Modeling (StarSchema, Snowflake Schema), Data warehousing,** and **OLAP tools.**
- Design solutions for various system components using **Microsoft Azure.**
- Extracted, transformed, and loaded data from Source Systems to **Azure Data Storage** services using a combination of **Azure Data Factory, T-SQL, Spark SQL, and U-SQL Azure Data Lake Analytics.** Data Ingestion to one or more **Azure Services - (Azure Data Lake, Azure Storage, Azure SQL, Azure DW)** and processing the data within **Azure Databricks.**
- Implemented **Azure Data Factory (ADF)** extensively for ingesting data from different source systems, like relational and unstructured data.
- Created pipelines in **Azure** using **ADF** to get the data from different source systems and transform the data by using many activities.
- Responsible for setting up and configure monitoring and metric gathering system around **Prometheus and Grafana.**
- Implemented **Apache Airflow** for authoring, scheduling, and monitoring Data Pipelines.
- Involved in creating, debugging, and monitoring jobs using **Airflow.**
- Used **Apache Kafka** to aggregate web log data from multiple servers and make them available in Downstream systems for Data analysis and engineering type of roles.
- Use **Git** commands extensively for code check-in.
- Used **Jira** for bug tracking.
- Developed a **NoSQL** database by using **CRUD**, Indexing, Replication, and Sharing in **MongoDB.**
- Developed **SQL scripts** for creating tables, Sequences, Triggers, views, and materialized views.
- Followed **Agile** Methodology and **Scrum** to deliver the product with cross-functional skills.
- Actively participating in the code reviews, meetings, and solving any technical issues.

Environment: Spark, Python, PySpark, MapReduce, Hive, HDFS, HBase, ETL, Power BI, Tableau, Star Schema, Snowflake Schema, Azure, Apache Airflow, Kafka, Docker, Kubernetes, Git, Jira, NoSQL, MongoDB, SQL, Agile and Windows.

Client: Country Financial, Bloomington, IL. || Duration: Dec 2023 – Jun 2024
Role: Senior Data Engineer
Responsibilities:

- Participated in a requirement gathering session with business users and sponsors to understand and document the business requirements.
- Developed **Spark** jobs to clean data obtained from various feeds to make it suitable for ingestion into **Hive** tables for analysis.
- Developed Custom Input Formats in **Spark jobs** to handle custom file formats.
- Developed multiple **MapReduce** jobs in Java for data cleaning and preprocessing.
- Used **PySpark** jobs to run on a Kubernetes Cluster for faster data processing.
- Worked on data pre-processing and cleaning the data to perform feature engineering and performed data imputation techniques for the missing values in the dataset using **Python.**
- Developed visualizations and dashboards using **Power BI.**
- Worked on creating filters, parameters, and calculated sets for preparing dashboards and worksheets in **Tableau.**
- Implemented Slowly Changing Transformation to maintain historical data in the data warehouse.
- Written **Hive jobs** to parse the logs and structure them in tabular format to facilitate effective querying on the log data.

- Built and managed **ADF pipelines, datasets, and linked services** for batch and incremental data loads.
- Configured **Azure cloud services** for endpoint deployment.
- Design & implement migration strategies for traditional systems on **Azure (Lift and shift/Azure Migrate, other third-party tools)** worked on **Azure suite: Azure SQL Database, Azure Data Lake (ADLS), Azure Data Factory (ADF) V2, Azure SQL Data Warehouse, Azure Service Bus, Azure Key Vault, Azure Analysis Service (AAS), Azure Blob Storage, Azure Search, Azure App Service, Azure data** Platform Services.
- Designed the **Data Marts** in dimensional data modeling using **star** and **snowflake schemas.**
- Used **Kafka** functionalities like distribution, partition, and replication for the commit log service for messaging systems by maintaining feeds.
- Automated and validated the data using **Apache Airflow.**
- Performed installation and managed **Grafana** to visualize the metrics collected by **Prometheus.**
- Used **Git** to check in and check out code changes.
- Used **Jira** for bug tracking to check in and check out code changes.
- Used **SQL** queries and other tools to perform data analysis and profiling.
- Involved in **Agile** methodologies, daily scrum meetings, and sprint planning.
- Actively participated and provided feedback constructively and insightfully during weekly Iterative review meetings to track the progress for each iterative cycle and figure out the issues.

Environment: Spark, Scala, Python, PySpark, MapReduce, Hive, HDFS, HBase, ETL, Power BI, Tableau, Star Schema, Snowflake Schema, Azure, Apache Airflow, Kafka, RestFul, Docker, Kubernetes, Git, Jira, NoSQL, MongoDB, SQL, Agile and Windows.

Client: Tenet Healthcare, Dallas, TX. || Duration: Oct 2022 – Nov 2023
Role: Data Engineer
Responsibilities:

- Interacted with clients to gather business and system requirements, which involved documentation of processes based on the user requirements.
- Developed multiple **Spark** jobs in **Scala & Python** for data cleaning and preprocessing.
- Used **Spark** and **Spark SQL** to read the parquet data and create the tables in **Hive** using the **Scala API.**
- Developed and supported **MapReduce** Programs running on the cluster.
- Developed **PySpark** scripts to encrypt raw data using hashing algorithms on client-specified columns.
- Performed **Data Analysis, Data Migration, Data Cleansing, Transformation, Integration, Data Import, and Data Export** through **Python.**
- Created **Tableau** reports with complex calculations and worked on Ad-hoc reporting using **Power BI.**
- Performing ETL testing activities like running the Jobs, extracting the data using necessary queries from the database transform, and uploading into the Data warehouse servers.
- Implemented a Continuous Delivery pipeline with **Docker, GitHub, and AWS.**
- Migrated an existing on-premises data to AWS S3. Used AWS services like EC2 and S3 for data set processing and storage.
- Worked on **Snowflake Schemas and Data Warehousing, and** processed batch and streaming data load pipeline using **SnowPipe** and Matillion from secure Azure Data Lake and AWS S3 buckets.
- Involved in loading data from rest endpoints to **Kafka Producers** and transferring the data to **Kafka Brokers.**
- Utilized **Jira** as a project management methodology and **Git** for version control to build the program.
- Data sources are extracted, transformed, and loaded to generate **CSV data files** with **Python** programming and **SQL queries.**
- Worked on SQL queries in dimensional data warehouses and relational data warehouses. Performed Data Analysis and Data Profiling using Complex **SQL** queries on various systems.
- Followed **agile** methodology for the entire project.
- Participated in the status meetings and status updates to the management team.

Environment: Spark, Scala, PySpark, Python, MapReduce, Tableau, ETL, Power BI, Docker, AWS, GitHub, Snowflake, Star Schema, Kafka, SQL, Agile, and Windows.

Company: Ebix Software India Pvt. Ltd., India. || Duration: Jan 2018 – Oct 2021
Role: Data Engineer
Responsibilities:

- As a Data Engineer, worked with Business Analysts to gather requirements and design a reliable and scalable data pipeline.
- Developed **Spark** scripts by using **Scala** and **Python** shell commands as per the requirement.
- Used **Scala** to convert **Hive** / **SQL** queries into RDD transformations in **Apache Spark**.
- Developed **MapReduce** jobs for cleaning, accessing, and validating the data.
- Developed data visualizations in **Tableau** to display day to day accuracy of the model with newly incoming Data.

- Performed **data gathering, cleaning, and wrangling** using **Python.**
- Used custom-developed **PySpark** scripts to pre-process, transform data, and map to tables inside the CIF (Non-corporate Information Factory) data warehouse.
- Involved in designing **ETL** processes and developing source-to-target mappings.
- Integrated data quality plans as a part of **ETL** processes.
- Created **DataStage jobs** using different stages like **Transformer, Aggregator, Sort, Join, Merge, Lookup, Data Set, Funnel, Remove Duplicates, Copy, Modify, Filter, Change Data Capture, Change Apply, Sample, Surrogate Key, Column Generator, Row Generator, Etc.**
- Extensively involved in writing SQL queries (sub-queries and join conditions) for building and testing ETL processes.
- Participated in the status meetings and status updates to the management team.

Environment: Spark, Scala, MapReduce, Python, PySpark, ETL, Tableau, SQL, Agile, and Windows.