Sree lasya Manga Puram
Data Engineer
Email: Sreelasya199@gmail.com |**Ph#:** +1 4692625320 |
LinkedIn : linkedin.com/in/sreelasya-mangapuram-583b0a307

Professional Summary:

- Around **5+ years** of experience in Data Engineering, Data Pipeline Design, Development and Implementation as a **Data Engineer/Data Developer and Data Modeler.**
- Experience in **Software Development Life Cycle (SDLC)** including Requirements Analysis, Design Specification and Testing as per Cycle in both Waterfall and Agile methodologies
- Experience with **Spark Core, Spark SQL, Spark MLlib, Spark GraphX and Spark streaming f**or processing and transforming complex data using in-memory computing capabilities written in **Scala.** Worked with **Spark** to improve efficiency of existing algorithms using **Spark Context, Spark SQL, Spark MLlib, Data Frame, Pair RDD's and Spark YARN.**
- Experience with designing, implementing, or operating IT systems on leading commercial **Cloud platforms**, including **AWS, Azure, and GCP.**
- Experience in **Python and Scala, user-defined functions (UDF)** for **Hive and Pig** using **Python**.
- Hands-on experience with **Hadoop architecture** and various components such as **Hadoop File System HDFS, Job Tracker, Task Tracker, Name Node, Data Node and Hadoop MapReduce programming.**
- Experience in design and development of all **data warehousing** components e.g. source system data analysis, **ETL strategy, data staging and data migration strategy**.
- Experience in **Extraction, Transformation and Loading (ETL)** data from various sources into Data Warehouses, as well as data processing like collecting, aggregating and moving data from various sources.
- in **Tableau Desktop** for data visualization, Reporting and Analysis; Cross Map, Scatter Plots, Geographic Map, Pie Charts and Bar Charts, Page Trials and Density Chart.
- Experience structural modifications using **Map-Reduce, Hive** and analyze data using visualization/reporting tools **(Tableau).**
- Design and implement large scale distributed solutions in **AWS** and **GCP** clouds
- Experience in creating **Power BI** Dashboards (Power View, Power Query, Power Pivot, and Power Maps).
- Hands on experience in setting up workflow using **Apache Airflow.**
- Experienced in fact dimensional modeling **(Star schema, Snowflake schema), transactional modeling** and **SCD (Slowly changing dimension).**
- Experience working on creating and running **Docker** images with multiple micro - services.
- Experienced in using **Pig scripts** to do transformations, event joins filters and pre-aggregations before storing the data into **HDFS.**
- Experience with ETL tools such as **Pentaho Kettle, Informatica, Talend, Open Refine.**
- Strong understanding of data-warehousing and data lake concepts
- Hands-on experience developing cloud native applications on platforms like **Cloud Foundry, Kubernetes, DC/OS, Heroku, AWS, GCP, Azure, etc.**
- Hands on experience working **Amazon Web Services (AWS)** using **Elastic Map Reduce (EMR), Redshift, and EC2** for **data processing.**
- Experience in developing **JSON** Scripts for deploying the Pipeline in Azure Data Factory (ADF) that process the data using the Cosmos Activity.
- Experience with **IaaS** with preference given to **GCP and AWS**
- Hands on experience in **SQL and NOSQL** database such as **Snowflake, HBase, Cassandra and MongoDB.**
- Extensive experience in **agile** software development methodology.
- Team Player as well as able to work independently with minimum supervision, innovative & efficient, good in debugging and strong desire to keep pace with latest technologies.
- Excellent **Communication and presentation skills** along with good experience in communicating and working with various stake holders.

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, Pig, Map Reduce, Sqoop.
Operating Systems	Red Hat Linux, Unix, Windows, macOS.
Reporting & Visualization	Tableau, Matplotlib.

Professional Experience:

Client: Paychex, Rochester, NY **Jan 2023 – Till Date**

Role: Data Engineer

Responsibilities:

- Developed **Scala** based **Spark** applications for performing data cleansing, event enrichment, data aggregation, de-normalization and data preparation needed for machine learning and reporting teams to consume.
- Developed **spark** applications in **python (PySpark)** on distributed environment to load huge number of **CSV files** with different schema in to Hive ORC tables.
- Developed **Scala** functional programs for streaming data and gathered **JSON and XML** data.
- Developed Simple to complex **MapReduce** Jobs using **Hive and Pig.**
- Developed **PIG** scripts to transform the raw data into intelligent data as specified by business users.
- Utilized **Spark, Scala, Python** for querying, preparing from big data sources.
- Wrote pre-processing queries in **python** for internal **spark** jobs
- Demonstrated experience in delivering data and analytic solutions leveraging AWS, Azure or similar cloud data lake
- Developed **ETL** workflows, sessions for initial full loading and incremental loading.
- Developed **ETL mappings** to Extract data from staging area and perform required transformations as per business requirements and load into **ODS and Data Warehouse tables.**
- **Extract, transform, and load (ETL)** data from multiple federated data sources (**JSON**, **relational database**, etc.) with Data Frames in **Spark**.
- Design and Develop **ETL Processes** in AWS Glue to migrate Campaign data from external sources like **S3, Parquet/Text Files into AWS Redshift.**
- Experience with integrations in a cloud environment **(AWS, GCP, Azure)**
- Created **Tableau** scorecards, dashboards using stack bars, bar graphs, scattered plots, geographical maps, Gantt charts using the functionality.
- Created scripts to read **CSV, JSON and parquet files** from S3 buckets in **Python** and load into **AWS S3, DynamoDB and Snowflake.**
- Involved in creating dashboards and reports in **Tableau** and maintaining user and server activities
- Deployed web embedded **power BI** dashboards refreshed using gateways by using workspace and data source.
- Used **Power BI, Power Pivot** to develop data analysis prototype, and used **Power View and Power Map** to visualize reports.
- Develop complex big data ingestion jobs in Talend for relational, big data, streaming, IOT, flat file, JSON, API, and many other data sources
- Performed transformations, cleaning and filtering on imported data using **Hive, Map Reduce**, and loaded final data into **HDFS.**
- Created **airflow DAGs** to sync files from box, analyze data quality, and alert for missing files.
- Experience working in cloud architecture including **AWS** and **Azure** environments
- Created scripts to read **CSV, JSON and parquet files** from S3 buckets in **Python** and load into **AWS S3, DynamoDB and Snowflake.**
- Build a high-quality Data Lakes and Data Warehousing team and design the team to scale. Build cross functional relationships with Data analysts, Product owners and Engineers to understand data needs and deliver on those needs
- Implemented **AWS** Elastic Container Service (ECS) scheduler to automate application deployment in the cloud using **Docker** Automation techniques.
- Analyzed the **SQL scripts** and designed the solution to implement using **PySpark.**
- Extracted files from **MongoDB** through **Sqoop** and placed in **HDFS** and processed.
- Use **SQL** queries and other tools to perform data analysis and profiling.
- Followed agile methodology and involved in daily **SCRUM** meetings, sprint planning, showcases and retrospective.

Environment: Spark, Scala, AWS, ETL, Hadoop, Python, Snowflake, Tableau, Data Lake, HDFS, Hive, Tableau, MapReduce, PySpark, Pig, Tableau, Teradata, Docker, JSON, XML, Apache Kafka, SQL, PL/SQL, Agile and Windows.

Client: Comcast, Philadelphia, PA **May 2022 – Dec 2022**
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 **Python and Scala** shell commands as per the requirement.
- Wrote **Spark** jobs with **RDD's, Pair RDDs, Transformations and actions, data frames** for data transformations from relational sets.
- Experience with ETL/ELT tools and design, specifically Informatica or Talend Open Studio 6.x (Talend Big Data Integration preferred)
- Designed and Developed **Scala** workflows for data pull from cloud-based systems and applying transformations on it.
- Developed highly complex **Python** and **Scala** code, which is maintainable, easy to use, and satisfies application requirements, data processing and analytics using inbuilt libraries.
- Provide guidance on **AWS & GCP** best practices to internal customers and external vendors
- Extract transfer and load data source system to cloud GCP data storage system using a combination of **Airflow.**
- Developed **PySpark** script to merge static and dynamic files and cleanse the data.
- Developed **ETL** framework using Spark and Hive (including daily runs, error handling, and logging) to useful data.
- Developed **ETL** technical specs, Visio for **ETL process flow and ETL load plan, ETL execution plan, Test cases, Test scripts etc.**
- Responsible for ensuring that service issues within **AWS & GCP** are resolved in a timely manner
- Analyze Finance data models, create and optimize data ingestion processes for the Data Lake through ETL technologies like data replication and scripting with **Greenplum, Talend, etc.**
- 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.
- Worked on **Tableau** activities, Multidimensional database and writing SQL queries.
- Hands on porting the existing on-premise **Hive** code migration to GCP.
- Used **Microsoft Power BI, Power Query** to extract data from external sources and modify data to certain format as required in Excel and created **SSIS** packages to load excel sheets from PC to database.
- Created basic reports using confidential files as source to fetch the data in **Power BI.** Designed and developed **Power BI** graphical and visualization solutions with business requirement documents and plans for creating interactive dashboards.
- Written **Pig** Scripts for sorting, joining, filtering and grouping data.
- Experience in ETL and Big Data Technologies like **Talend, Hadoop, Greenplum, HVR, HIVE etc.**
- Extracted data from **Teradata** database and loaded into Data warehouse using **spark.**
- Implemented a Continuous Delivery pipeline with **Docker** and **GitHub.**
- 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.
- Performed analysis on the unused user navigation data by loading into **HDFS** and writing **MapReduce** jobs.
- 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.**
- Used **SQL** queries and other tools to perform data analysis and profiling.
- 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, Scala, GCP, ETL, Hadoop, Python, Snowflake, HDFS, Hive, MapReduce, PySpark, Pig, Docker, GitHub, Apache Spark, Teradata, JSON, PostgreSQL, MongoDB, SQL, Agile and Windows.

Client: Western Union, Milwaukee, WI **Jan 2020 – Dec 2021**
Role: Data Engineer
Responsibilities:

- Worked with the business users to gather, define business requirements and analyze the possible technical solutions.
- Ensure highly reliable information delivery for 360i's leading clients by leveraging a variety of data sources using cloud-based data services such as AWS and Azure for ELT
- Developed **Spark scripts** by using **Python and Scala** shell commands as per the requirement.
- Developed **ETL** framework using Spark and Hive (including daily runs, error handling, and logging) to useful data.
- Developed **PIG** scripts for the analysis of semi structured data.
- Used **Pig** as **ETL** tool to do transformations, event joins, filters and some pre-aggregations before storing the data onto **HDFS.**
- Experience in data transformation including (but not limited to) such tools as **Informatica, Talend, Boomi**
- Demonstrated experience in delivering data and analytic solutions leveraging **AWS, Azure** or similar cloud **data lake**
- Imported the data from different sources like **AWS S3**, Local file system into **Spark RDD.**
- Involved in converting **Hive/SQL** queries into **Spark** Transformations using **Spark RDDs and Scala.**
- Used **Hive** to analyze the Partitioned and Bucketed data and compute various metrics for reporting.
- Used **Kafka** to load data into **HDFS** and move data back to S3 after data processing
- Understanding of real-time streaming technologies such as **Apache Kafka, Azure EventHub, Spark Streaming, Apache Storm, Apache Flink etc.**
- Worked on migrating **MapReduce** programs into Spark transformations using **Scala.**
- Used **ETL** to implement the Slowly Changing Transformation, to maintain Historically Data in Data warehouse.
- Designing and implementing data warehouses and data marts using components of Kimball Methodology, like Data Warehouse Bus, Conformed Facts & Dimensions, Slowly Changing Dimensions, Surrogate Keys, Star Schema, Snowflake Schema, etc.
- Used **AWS S3** to store large amount of data in identical/similar repository.
- Utilized **Spark** SQL API in **PySpark** to extract and load data and perform **SQL queries.**
- Worked on Ingestion, Parsing and loading the data from **CSV and JSON** files using **Hive and Spark.**
- Written **pig script** to load processed data from **HDFS** into **MongoDB**.
- Extensively involved in writing SQL queries (sub queries and join conditions) for building and testing ETL processes.
- Actively participating in the code reviews, meetings and solving any technical issues.

Environment: Spark, Scala, ETL, Python, AWS, HDFS, Hive, Kafka, Pig, CSV, JSON, PySpark, SQL, Agile and Windows.

Client: Xoom Works Inc, India **Jun 2018 – Dec 2019**
Role: Data Engineer
Responsibilities:

- Gathering business requirements, business analysis and design various data products.
- Developed **Spark** scripts by using **Python** shell commands as per the requirement.
- Implemented Data pipelines for big data processing using **Spark** transformations and **Python API** and clusters in **AWS.**
- Designed and Developed **Spark** workflows using **Scala** for data pull from **AWS S3 bucket** and **Snowflake** applying transformations on it.
- Developed **Spark code** in **Python and Spark SQL** environment for faster testing **and** processing of data and Loading the data into **Spark RDD** and doing In-memory computation to generate the output response with less memory usage.
- Built key business metrics, Visualizations, dashboards, reports with **Tableau.**
- **Extract, transform, and load (ETL)** data from multiple federated data sources (**JSON**, **relational database**, etc.) with Data Frames in **Spark**.
- Created **PIG Latin** Scripts to sort, group, join and filter the enterprise wise data.
- Created **Hive** tables to store the processed results in a tabular format.
- Analyzed the **SQL scripts** and designed the solution to implement using **PySpark.**
- Performed transformations, cleaning and filtering on imported data using **Hive, Map Reduce**, and loaded final data into **HDFS.**
- 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.

Environment: Scala, Spark, Python, PySpark, ETL, HDFS, Pig, AWS, MapReduce, Hive, XML, CSV, JSON, Kafka, SQL.