LAKSHMI G Ph: +1 (913) 912-3349
Senior Data Engineer Email: lakshmisreegsl@gmail.com

Professional Summary:

- Accomplished Senior Data Engineer **over 7 years'** experience in IT industry. Skilled in **Business Requirement Analysis, Data Analysis, Data Warehousing, Database Management, ETL Development, and Data Modeling**.
- Well versed in leveraging key **AWS** services like **AWS S3** for storage, **EC2** for scalable compute, **DynamoDB** for NoSQL storage, **EMR** for big data processing, and **AWS Glue** for ETL workflows.
- Experience with **Snowflake** cloud data warehouse and **AWS S3** bucket for integrating data from multiple source system which include loading nested JSON formatted data into **snowflake** table.
- Expertise in utilizing **AWS Cloud Formation**, **API Gateway**, and **AWS Lambda** for automation and ensuring the security of AWS infrastructure.
- Good knowledge in migrating SQL databases to **Azure Data Lake**, **Azure Data Lake Analytics, Azure SQL Database, Data Bricks,** and **Azure SQL Data Warehouse.**
- Proficient in controlling and granting database access and executing migrations of on-premises databases to **Azure Data Lake Store** using **Azure Data Factory**.
- Well versed in utilizing **Tableau, Power BI**, and **Flask** for impactful dashboards, coupled with advanced proficiency in **Microsoft Excel**, including formulas, pivot tables, charts, and **DAX** commands.
- Strong experience in the creation and performance tuning of business reports using tools like **OBIEE, Power BI, Tableau** and SQL server tools like **SSIS** and **SSRS**.
- Experience in Big Data analytics and data manipulation, utilizing the Hadoop ecosystem tools such as **Spark, Kafka, HDFS, MapReduce, YARN/MRv2, Pig, Hive, HBase, Flume, Sqoop, Oozie**, and integration of **Spark** with **Cassandra, Avro, and Zookeeper**.
- Have extensive experience in IT data analytics projects, hands on experience in migrating on premise ETs to **Google Cloud Platform (GCP**) using cloud native tools such as **BIG query**, **Cloud Data Proc**, **Google Cloud Storage**, **Composer**.
- Good knowledge in **Data Modeling** Techniques using **Star Schema, Snowflake Schema**, Fact and Dimension tables, RDBMS, Physical and Logical data modeling for **Data Warehouse** and **Data Mart.**
- Expertise in extracting, transforming, and loading (**ETL**) data from spreadsheets, database tables and other sources using **DataStage**, **Informatica, SQL Server Integration Services (SSIS) and SQL Server Reporting Service (SSRS)** for managers and executives.
- Possess a strong proficiency in the creation and maintenance of physical data models, showcasing expertise across diverse database platforms including **Oracle, Teradata, Netezza, DB2, MongoDB, HBase, and SQL Server**.
- Proficient in writing, implementing, and testing of triggers, procedures and functions in **PL/SQL**, **Oracle,** and good command over programming languages like **Python**, **Shell Scripting**.
- Well versed in managing source code repositories like **Git, GitHub** and **Bitbucket**. Experience in configuring Continuous Integration (**CI**) and Continuous Delivery (**CD**) using **JENKINS**.
- Experienced in utilizing **Agile methodologies** such as **SCRUM** and **Waterfall methodologies** throughout the Software Development Lifecycle (**SDLC**).

Technical Skills:

- **Big Data Ecosystem**: Hadoop Map Reduce, Kafka, Spark, Impala, HDFS, Hive, Pig, HBase, Flume, Storm, Sqoop, Oozie
- **Programming Languages:** Python, R, Scala, SQL, HiveQL, PL/SQL, LINUX, shell

Scripting, Java, T-SQL

- **AWS :** S3, EC2, EMR, Lambda, Glue, Athena, RDS, Cloud Formation, Cloud Watch
- **Azure:** Data Lake, Data Factory, Databrick, SQL Database, SQL data Warehouse
- **GCP:** Cloud Storage, Big Query, Composer, Cloud Data Proc, Cloud SQL
- **Databases :** Snowflake, MySQL, Teradata, Oracle, MS SQL SERVER, PostgreSQL, DB2
- **NoSQL Databases :** HBase, Cassandra, Mongo DB, DynamoDB and Cosmos DB
- **ETL/BI :** Tableau, Power BI, Informatica, OBIEE, SSIS, SSRS, SSAS, QlikView, Erwin, Matillion, MS Excel
- **Machine Learning :** Linear Regression, Logistic Regression, Decision Tree, Random Forest, SVM, Naïve Bayes, PCA, LDA, K-Means, KNN, Neural Network

Work Experience:

Client: Equifax, Saint Louis, MO **Jan 2022 - Present**
Role: Senior Data Engineer

Project Description: Equifax is a leading global data analytics company specializing in consumer credit reporting and insights. I am a part of identity verification and authentication team which involves working with wide range of data sources such as databases and provide robust solutions and generate reports to enable organizations to effectively confirm the identities of individuals, safeguard against identity theft, and adhere to regulatory requirements with confidence.

Responsibilities:

- Participated in requirement gathering with business users to understand and document the business requirements in alignment to the financial goals of the company.
- Developed **ETL** workflows using **Apache Airflow** and **AWS Glue**, extracting data from diverse sources, transforming it, and loading it into the Hadoop Distributed File System (**HDFS**) within the **AWS EMR** cluster, ensuring data availability and reliability for identity verification process.
- Integrated **AWS Dynamo DB** using **AWS Lambda** to store the values of items and backup the Dynamo DB streams.
- Developed data ingestion pipelines using **Apache Kafka** and **AWS Kinesis** to stream real-time identity data into the Hadoop system, ensuring timely updates and enabling immediate verification processes.
- Leveraged **Apache Hive** and **AWS Athena** to build and optimize data tables and views, enabling efficient ad-hoc querying and data exploration for identity verification and fraud detection purposes.
- Worked with **AWS Cloud Formation** to focus on high availability, fault tolerance, and auto-scaling while designing and deploying multi-tier applications using **EC2, Route53, S3, RDS, Dynamo DB, SNS, SQS, and IAM**.
- Implemented data backup and disaster recovery strategies using AWS services such as **S3** and **Backup**, ensuring data availability and minimizing downtime in the event of system failures or data loss.
- Set up scalability for application servers using command line interface for Setting up and administering DNS system in **AWS using Route53.**
- Visualize and manipulate the data using various machine learning libraries like **NumPy, SciPy and Pandas** in Python scripts for the perfect analysis of data.

- Developed scalable data processing workflows using **Python** and **Scala** on **Apache Spark**, leveraging big data technologies like **Hadoop** and **Spark Streaming** to handle large volumes of identity data in real-time, enabling efficient authentication processes.
- Implemented data replication and synchronization between the Hadoop cluster and **AWS Redshift** using **AWS Glue Data Catalog** and **AWS Database Migration Service (DMS)**, ensuring consistent and up-to-date identity data for reporting and analytics in **Tableau**.
- Developed **Spark** jobs on **Databricks** to perform tasks like data cleansing, data validation, standardization, and then applied transformations as per the use cases.
- Developed and maintained **Tableau** data connectors to AWS data sources, enabling seamless integration with **AWS** data sources and services.
- Conducted performance tuning and optimization of **Tableau** dashboards and reports, resulting in improved query response times and reduced data processing costs.
- Worked with **AWS Elasticsearch** to build full-text search capabilities on large datasets, enabling users to easily search and retrieve data for use in **Tableau** visualizations and reports.
- Implemented **Git** as the version control system and configured **Jenkins** pipelines to automate continuous integration, testing, and deployment processes, enabling efficient code collaboration, version control, and seamless delivery of the process at Equifax, in conjunction with AWS services.
- Worked on software development life cycle (SDLC) in **Agile** environment.

Environment: CloudFormation, EC2, Route53, RDS, DynamoDB, SNS, SQS, S3, AWS Backup, AWS Kinesis, Apache Airflow, Kafka, Hive, Scala, Athena, Glue, Redshift, Git, Jenkins, Tableau, Agile

Infosys Limited, India **Apr 2020 – Sep 2021**
Role: Big Data Engineer

Project Description: Infosys Limited, based in India, is a global consulting and technology services company. As a Big Data Engineer at Infosys, I specialize in leveraging the power of Amazon Web Services (AWS) to design and develop scalable data solutions. The projects focus is to better understand and assist the clients and their customers.

Responsibilities:

- Responsible for maintaining quality reference data in source by performing operations such as cleaning, transformation and ensuring Integrity in a relational environment by working closely with the stakeholders & solution architect.
- Worked on **Snowflake Schemas** and Data Warehousing, developed and implemented efficient batch and streaming data pipelines utilizing **Snow Pipe** for data ingestion and extraction from a secure **AWS S3** bucket within the data lake using **Matillion**.
- Utilized **Snowflake's** built-in features for query optimization and performance tuning, such as materialized views and clustering keys, to improve query response times and reduce costs.
- Leveraged **Python** and **Snowflake's** Python connector to automate data extraction, transformation, and loading processes, improving efficiency, and reducing manual effort.
- Developed robust data models specifically tailored for data-intensive **AWS Lambda** applications, enabling complex analysis, generating analytical reports, and ensuring end-to-end traceability, lineage, and accurate definition of key business elements sourced from **Aurora**.
- Collaborated with Data engineers and operation team to implement **ETL** process, wrote and optimized SQL queries to perform data extraction to fit the analytical requirements.

- Utilized **Tableau** and integrated with **AWS** data sources, such as **AWS S3** and **AWS RDS**, to create real-time, automated reporting solutions and create interactive visualizations and dashboards, providing stakeholders with actionable insights and facilitating data-driven decision-making.
- Implemented data archiving and retention strategies for long-term storage and compliance purposes using **AWS Glacier**.
- Implemented **AWS EC2**, Key Pairs, Security Groups, Auto Scaling, **ELB**, **SQS**, and **SNS** using AWS API and exposed as the Restful Web services. Used **EC2** as virtual servers to host **Git, Jenkins,** and configuration management tool like **Ansible**.
- Import customer information data from Oracle database into **HDFS** for data processing along with minor by using cleansing **Spark** Scala.
- Utilized **AWS CloudWatch** and other monitoring tools to monitor the health and performance of data pipelines and infrastructure.
- Developed **Hive** queries for the analysts by loading and transforming large sets of structured, semi structured data using hive.

Environment: Snowflake, Snowflake Schema, Snow pipe, S3, Matillion, Hive, EC2, lambda, Tableau, Cloud Watch, Spark, Scala, Oracle, Git, Jenkins, Glacier, RDS, EMR, SQL

Infosys Limited, India **Aug 2018 – Mar 2020**
Role: Big Data Developer

Project Description: The project was about the application on Customer Services System which was designed to support the application process for the needs of clients. This application is an online workflow system to monitor and drive their big data initiatives. The business of this division includes designing and implementing scalable data architectures to handle the vast volume of data generated and creating reports from the database for internal and external use.

Responsibilities:

- Experienced with real-time data processing with **Azure Synapse Analytics**. Developed and managed **SSIS** packages to extract, transform, and load data from various source systems into **Azure Data Storage** services.
- Possess in-depth expertise in developing pipeline jobs, scheduling triggers, mapping data flows using **Azure Data Factory (V2**) and storing credentials using Key Vaults.
- Extract Transform and Load data from Sources Systems to **Azure Data Storage** services using a combination of **Azure Data Factory, T-SQL, Azure Data Lake Analytics, Data Ingestion to Azure Services -** (Azure Data Lake, Azure Storage, Azure SQL, Azure DW) and processing the data in In **Azure Databricks**.
- Writing complex SQL Queries, Stored Procedures, Triggers, Views, Cursors and User Defined Functions to implement the business logic.
- Used **Apache Spark pool** and **Synapse Pipelines** in **Azure Synapse Analytics** to access and transfer data at large.
- Built analytics dashboards and embedded reports in a dedicated SQL pool to share business insights with internal teams within the organization and used **Azure Analysis Services** to perform business analysis.
- Built a continuous integration (**CI**) and continuous deployment (**CD**) pipeline for accelerating application development and development lifecycle.

- Evaluated and worked on **Azure Data Factory** as an **ETL** tool to process business-critical data into aggregated tables in **Hive Cloud**. Deployed and Development in Bigdata applications like **Spark**, and **Hive** in **Azure** cloud.
- Created **DAX** queries to generate computed columns in **Power BI**.
- Used **Power BI**, **Power Pivot** to develop data analysis on the data and used **Power View** and **Power Map** to visualize reports.
- Designed and developed **FLINK** pipelines to consume streaming data from Data Lakes and applied business logic to message and transform and serialize raw data.
- Processed Large amounts of Data using Azure Lake Analytics and stored it in the Data Lake store.
- Scheduled **Airflow DAGs** to run multiple Hive and Pig jobs, which independently run with time and data availability.
- Created notifications, alerts, to reports in **Power BI** service.
- Experienced using **SQL** and python to find Data Patterns and Data anomalies in **python Jupyter notebook**.

Environment: Azure Synapse Analytics, Azure Data Factory, Azure Data Storage, T-SQL, Azure Data Lake, Azure Databricks, Apache Spark pool, Synapse pipelines, ETL, Hive Cloud, SQL, Python, Agile

Client: Star Health Insurance, India **Feb 2017 – Jul 2018**
Role: Hadoop Developer

Project Description: Star Health Insurance advances the healthcare system by helping healthcare organizations reduce costs and improve health outcomes. Worked on an application to identify and correct improper healthcare billings and payments data based on encounter type. The project involved ingesting the legacy and the data into the new platform using the Hadoop applications.

Responsibilities:

- Collaborated with cross-functional teams to gather requirements and design data solutions aligned with business needs.
- Worked in **Azure** environment for development and deployment of Custom Hadoop Applications.
- Tested loading the raw data, populate staging tables and store the refined data in partitioned tables in the enterprise data warehouse.
- Manage enterprise Data Warehouse operation, big data advanced predictive application development using **Cloudera (**CDH)& **Hortonworks (**HDP**)**.
- Developed and delivered various **Power BI** reports, predictive trend analysis using reports like **Power Pivot, Power View** to analyze efficiency performance and to predict future level activities.
- Troubleshooting issues in the execution of **MapReduce** jobs by inspecting and reviewing log files.
- Developed **PIG** scripts to generate **MapReduce jobs** and performed **ETL** procedures on the data in **HDFS** and exported data using **Sqoop** for generating reports.
- Build data pipelines in airflow in **GCP** for **ETL** related jobs using different airflow operators.
- Experience in moving data between **GCP** and **Azure** using Azure Data Factory.
- Develop framework for converting existing power center mappings and to **PySpark** (python and spark) jobs.
- Used **Spark** streaming to receive real time data from the **Kafka** and store the stream data to **HDFS** using **Scala** and NoSQL databases such as **HBase** and **Cassandra**.
- Experience in **Agile** methodologies and tools likes **JIRA**.

Environment: Agile methodology, Sqoop, PIG, MapReduce, Spark, HBase, Zookeeper, Oozie, Cloudera, Horton works HDP, MapReduce, GCP, ETL, Python, Spark, Kafka, Scala, Cassandra

Client: ICICI Bank, India **Feb 2016 – Jan 2017**
Role: Data Analyst

Project Description: In ICICI Bank as a data analyst, it focused on evaluating customer data to improve business performance. I found patterns and trends to enhance decision-making processes by utilizing data from numerous sources, such as transaction records and consumer demographics.

Responsibilities:

- Manage multiple reporting requirements, including requests for ad-hoc analyses, and provide data-driven insights in an accurate manner to the executives by maintaining 100% client uptime.
- Automated data import and transformation processes using power query editor to save time and increase efficiency; saving 90% of time invested in report generation.
- Involved in creating the external and internal tables in **Azure SQL Datawarehouse** and created stored procedures to move the data from external to internal tables.
- Created **Python notebooks** on **Azure Databricks** for processing the datasets and loading them into **Azure SQL databases**.
- Performed verification and validation for accuracy of data in the monthly and quarterly reports.
- Responsible for the design, development, and administration of complex **T-SQL queries** (DDL / DML), Views& functions for transactional and analytical data structures.
- Design and develop business intelligence dashboards, analytical reports, and data visualizations using **Power BI** by creating multiple measures using **DAX** expressions for user groups likes sales, operations as per the business requirements.
- Experience in combining data from different sources by using power query.
- Used **Power BI Power Pivot** to develop data analysis prototype and used **Power View** and **Power Map** to visualize reports.
- Using a query editor in **Power BI** performed certain operations like fetching data from different file.
- Data conversions and data loads from various databases and file structures and compared the data for the purpose of analysis.
- Involved in **Normalization and De-Normalization** of existing tables for faster query retrieval.
- Expertise in data transformations such as adding calculated columns, manage relationships, create different measures, merge queries, append queries, replace values, split column, group by, Date & Time Column, etc.

Environment: MS SQL Server 2016/ 2018, SQL, Power BI, Excel, Word, Microsoft Teams, Oracle database, Azure Data Lake, Azure Databricks, Azure SQL DAX, Azure SQL Datawarehouse, ETL

Education Qualifications:

The University of Central Missouri, Warrensburg, MO **GPA 3.7/4**
Master of Science, Computer Science

Sagi Rama Krishnam Raju Engineering College, Andhra Pradesh, India **CGPA 8.05/10**
Bachelor of Technology, Computer Science and Engineering