**Lakshmi P**

**SR. Data Engineer**

**4707371148**

[**lakshmi.p.23.c@gmail.com**](mailto:lakshmi.p.23.c@gmail.com)

**PROFESSIONAL SUMMARY:**

* Experienced data engineer with 9+ years of hands-on experience building and managing data pipelines and infrastructure.
* Expertise in creating and maintaining data pipelines using **GCP, Azure, and AWS** cloud technologies.
* Experience in designing and implementing data processing systems using **GCP services** such as **BigQuery, Dataflow, Dataproc, and Pub/Sub.**
* Develop data set processes for data modelling, and Data mining. Recommend ways to improve data reliability, efficiency and quality.
* Knowledge of **ETL** methods for data extraction, transformation and loading in corporate-wide ETL Solutions and Data Warehouse tools for reporting and data analysis.
* Experience working with **AWS Data Platform - AWS** CloudFormation, Development Endpoints, AWS **Glue, EMR, Athena, Lambda and Jupyter/Sagemaker Notebooks, Redshift, S3, and EC2 instances.**
* Experience in **Microsoft Azure/Cloud Services like SQL Data Warehouse, Airflow, Azure SQL Server, Azure Databricks, Azure Data Lake, Azure Blob Storage, Azure Data Factory.**
* Hands on working experience with **Databricks** and **snowflake** database.
* Implemented data transformation logic and data cleansing processes using **Java** based frameworks.
* Experienced in Branching, Tagging, and maintaining the version across the Environments using SCM

tools like **Git**, **GitLab**, **GitHub,** and SVN on Linux and Windows platforms.

* Evaluated model output using the uncertainty matrix (Precision, Recall).
* Created IAM Roles and defining Policies and applying to AWS services.
* Performed Exporting and importing of data into simple storage service (S3).
* Good knowledge on Importing volumes, launching EC2, RDS, creating security groups, auto-scaling, load balancers (ELBs) in the defined virtual private connection.
* Knowledgeable in building data pipelines and ETL workflows using tools such as Apache Beam, Apache Airflow, and Cloud Composer.
* Knowledgeable in designing and optimizing data models for performance and scalability using Synapse Analytics and SQL Database.
* Skilled in designing data architectures and optimizing data models for performance and scalability using tools such as Apache HBase and Apache Cassandra.
* Hands on experience in Test-driven development, Software Development Life Cycle (SDLC) methodologies like Agile and Scrum.
* Good analytical, communication skills and ability to work with a team as well as independently with minimal supervision.
* Evaluated model output using the uncertainty matrix (Precision, Recall).
* Research on membership inference attacks to protect the sensitive information used in ML model training.

**TECHNICAL SKILLS:**

|  |  |
| --- | --- |
| **Programming Languages** | JAVA, Python, JavaScript, SQL, PySpark, GCP |
| **Cloud Tools** | **Amazon Web Services** (EC2,EMR,IAM, Lambda, Glue, S3, Athena, Glue Crawler, DMS, Step Functions, State Machines, Glacier, S3 Transfer Acceleration, Code Pipeline, Redshift, S3, CloudWatch, CloudTrail, Dynamo DB)  **Microsoft Azure** (Data factory, Blob Storage, ADLS, Data Bricks, Azure Sql, Synapse Analytics, AZURE HDInsight, Monitor, Active Directory, DevOps)  **GCP** (Big query, DialogFlow, DataProc, Google Vision, Google Collab, Google Cloud Natural Language, Cloud pub/sub, Cloud run) |
| **Datawarehouse** | AWS Redshift, Cloudera, Spark, Star Schema, Snowflake schema, SAS, SSIS, and Splunk |
| **BI/Analytic Tool** | Tableau, Azure ML, SSRS |
| **Big Data tools** | Apache Spark, PySpark, Hive, HDFS, MapReduce, Apache Airflow. |
| **Hadoop Distributions** | Cloudera, MapReduce, Hortonworks |
| **Utilities/Tools** | Eclipse, Tomcat, NetBeans, JUnit, SQL, SVN, Log4j, SOAP UI, ANT, Maven,PyCharm, Visio, Jenkins, Jira, IntelliJ, GIT. |
| **Database** | MySQL, PostgreSQL, MongoDB. |
| **Visualization Tools** | Tableau, Power BI. |

**PROFESSIONAL EXPERIENCE:**

**LTImindtree, OH**

**Sr Data Engineer**

**June 2023 – Jan 2024**

**Responsibilities**:

* Designed a standardized data model in alignment with client use case documents.
* Developed code to retrieve active workflows from the database, focusing on workflow status and scheduled tasks within Cloud Composer.
* Configured email notifications for clients in the event of exceptions, ensuring timely communication.
* Implemented scripts to transfer files between external Google Cloud Storage (GCS) and our GCS landing bucket
* Successfully loaded data from GCS bucket files into BigQuery tables, enhancing data accessibility and analysis.
* Demonstrated proficiency in leveraging Google Cloud Platform (GCP) services, including BigQuery, Cloud Composer, Cloud Run, and Airflow, for efficient data processing and management.
* Utilized BigQuery within GCP to execute data manipulations, including duplicate deletion and staging to final layer data insertion.
* Employed BigQuery for data transformations, conditional processing, and data quality validation, ensuring data accuracy and reliability.
* Developed code for seamless file transfers from Amazon S3 to GCS bucket, streamlining data movement processes.
* Utilized GCP's Airflow service to create Directed Acyclic Graphs (DAGs) with specific scheduling intervals, optimizing workflow automation and execution.
* Performed data profiling and transformation on the raw data using Pig, Python, and Java and developed predictive analytic using Apache Spark Scala APIs.
* Designed and implemented end-to-end data pipelines on GCP to extract, process, and analyze data from various sources, including IoT devices and customer interactions.

**Environment:** Data Analytics, Cloud Storage, Python, Data Studio, BigQuery, Cloud Build, ETL, GitHub, Apache Airflow, Cloud Run, Agile, JIRA, Cloud Composer

**Macmillan, NY**

**Sr Data Engineer (GCP)**

**Sep 2021– May2023**

**Responsibilities**:

* + Designed and implemented end-to-end data pipelines on GCP to extract, process, and analyze data from various sources, including IoT devices and customer interactions.
  + Utilized Google Pub/Sub to efficiently handle high-volume data streams in real-time from IOT devices and customer interactions.
  + Automating the cloud formation using Terraform.
  + Developed real-time data processing pipelines using Google Dataproc, leveraging its powerful parallel processing capabilities to ensure scalability and performance.
  + Stored processed data in Google Cloud Storage, providing reliable and cost-effective storage for large volumes of data.
  + Conducted real-time data analysis using Google BigQuery, leveraging its SQL-based querying capabilities to derive insights and make data-driven decisions.
  + mplemented AWS Code Pipeline and Created Cloud formation JSON templates in Terraform for infrastructure as code
  + Experience in designing and implementing data processing systems using **GCP services** such as **BigQuery, Dataflow, Dataproc, and Pub/Sub , Analytics**
  + Designed and implemented a microservices architecture for a blogging platform using Spring Boot and MongoDB.
  + Developed RESTful APIs for user authentication, blog post creation, and comment management.
  + Utilized Docker for containerization and Kubernetes for managing the deployment and scaling of microservices.
  + Implemented asynchronous communication between services using RabbitMQ for improved performance and scalability.
  + Built a microservices-based system using Java, Spring Boot, and Apache Kafka for real-time stock price updates.
  + Designed and developed RESTful APIs for retrieving stock data and receiving real-time price updates.
  + Implemented event-driven architecture with Kafka to enable real-time data processing and notification of price changes.
  + Built custom dashboards and reports using Google Data Studio, allowing stakeholders to visualize and interact with real-time data in a user-friendly manner.
  + Used Cloud Monitoring to monitor the performance of our applications and infrastructure in real-time.
  + it is imperative that such sensitive information never be stored in Terraform code in an unencrypted format. Instead, all passwords, TLS certificates, SSH keys, and other confidential information should be securely stored in encrypted storage to prevent unauthorized access.
  + Integrated Google Cloud KMS with other GCP services, such as Google Compute Engine and Google Kubernetes Engine, to provide a secure and scalable solution for key management.
  + Coordinated and collaborated with team members for project dependencies following scrum methodology.
  + Designed and implemented end-to-end data pipelines on GCP to extract, process, and analyze data from various sources, including IoT devices and customer interactions.
  + Implemented CI/CD pipelines with the Cloud build to help automate the development and deployment of pipelines.
  + Implemented AWS Code Pipeline and Created Cloud formation JSON templates in Terraform for infrastructure as code

**Environment:** Cloud Pub/Sub, Dataproc, Data Analytics, Cloud storage, Python, Data Studio, Big Query, Cloud Build, Cloud Monitoring, Cloud KMS, Kubernetes Engine, PowerBI.

**Ingram Micro, Irvine, CA.**

**Data Engineer (GCP)**

**Dec 2018 – Aug 2021**

**Responsibilities**:

* Analyze, design, and build Modern data solutions using Azure PaaS service to support visualization of data. Understand current Production state of application and determine the impact of new implementation on existing business processes.
* Implemented solutions for ingesting data from various sources and processing the Data- at-Rest utilizing Big Data technologies such as Hadoop, Map Reduce Frameworks, HBase, and Hive.
* Designed and implemented end-to-end data pipelines on GCP to extract, process, and analyze data from various sources, including IoT devices and customer interactions.
* Experience in designing and implementing data processing systems using **GCP services** such as **BigQuery, Dataflow, Dataproc, and Pub/Sub.**
* Terraform module for the AWS Guard Duty, ADFS IAM, Trusted Advisor for the security in all the AWS OUS which are used for development, staging and production.
* Managed and lead the development effort with the help of a diverse internal and overseas group and design/ architected and implemented complex projects dealing with the considerable data size (TB/ PB) and with high complexity.
* Created Pipelines in ADF using Linked Services/Datasets/Pipeline/ to Extract, Transform, and load data from different sources like Azure SQL, Blob storage, Azure SQL Data warehouse, write-back tool and backwards.
* Automating the cloud formation using Terraform.
* nvolved in using Terraform and Ansible, migrate legacy and monolithic systems to Azure and managing Ubuntu and RHEL virtual servers on Azure by creating AnsibleNodes.
* Performed data profiling and transformation on the raw data using Pig, Python, and Java and developed predictive analytic using Apache Spark Scala APIs.
* Developed Spark applications using Pyspark and Spark-SQL for data extraction, transformation, and aggregation from multiple file formats for analyzing & transforming the data to uncover insights into the customer usage patterns.
* Developed JSON Scripts for deploying the Pipeline in Azure Data Factory (ADF) that process the data using the SQL Activity.
* Experience in designing and implementing data processing systems using **GCP services** such as **BigQuery, Dataflow, Dataproc, and Pub/Sub.**
* Imported millions of structured data from relational databases using Sqoop import to process using Spark and stored the data into HDFS in CSV format and used Data Frame API in Scala for converting the distributed collection of data organized into named columns.
* Exploring DAG's, their dependencies and logs using Airflow pipelines for automation and use Apache Airflow to schedule and run the airflow dags to execute code and involved in scheduling Airflow workflow engine to run multiple Hive and pig jobs using python.
* I have utilized Terraform for cloud resource management, which requires access credentials to communicate with the cloud provider's API.
* Virtualized the servers using Docker for the test environments and dev-environments needs and Utilized Kubernetes and Docker for the runtime environment for the CI/CD system to build, test, and deploy

**Environment:** Databricks, GCP,Apache Spark, Python,Data Analytics, Scala, Azure, Kafka, ADLS Gen2, Cosmos Db, MongoDB, ADF, NIFI.

**Definity First Fremont, CA**

**Data Engineer**

**Feb 2016 – Dec 2018**

**Responsibilities:**

* Created and implemented ETL procedures to transfer data from AWS S3 to Snowflake using Snow pipe.
* Utilized AWS Glue to establish a batch processing pipeline for patient information and carried out data purification, data manipulation, and mapping transformations utilizing PySpark scripting.
* Established a real-time Data Analytics , data pipeline using AWS Kinesis for patient information, extracting it fromS3and storing it to Dynamo DB.
* I have utilized Terraform for cloud resource management, which requires access credentials to communicate with the cloud provider's API.
* Constructed a cost-efficient and internally managed serverless email system utilizing AWS Lambda and Simple Email Services (SES).
* Developed automated CI/CD data pipelines using AWS, Python, Code Pipeline and deployed code modifications into AWS EC2instances using AWS Code Deploy.
* Implemented a trigger for Lambda functions to build a pipeline for the transfer and transformation of small datasets, resulting in cost savings and time efficiency.
* Experience in designing and implementing data processing systems using **GCP services** such as **BigQuery, Dataflow, Dataproc, and Pub/Sub.**
* Created Apache Airflow DAGs (directed acyclic graphs) to schedule the pipeline for transferring data from AWS S3toSnowflake
* Created dashboards and reports for business intelligence purposes using AWS Quick Sight on data obtained from the DynamoDB.
* mplemented AWS Code Pipeline and Created Cloud formation JSON templates in Terraform for infrastructure as code
* Utilized GitHub as a source code management tool for performing tasks such as branching, merging, and tagging, among ,
* Automate provisioning and repetitive tasks using Terraform and Python, Docker container, Service Orchestration.
* Collaborated in an Agile workspace and utilized JIRA to keep track of and create user stories. **Environment:** AWS EC2, S3, Glue, Kinesis, Code Pipeline, Code Deploy, GCP,Splunk Cloud, QuickSight, Athena, ETL, GitHub, Apache Airflow, DAGs, Snowflake, Agile, JIRA.

**Object Arena, Chennai, India**

**SQL Developer Jun 2014 -Jan 2016**

**Responsibilities:**

* Involved in Analysis, Planning and Defining data based on business requirements and provided documentation.
* Extensively worked in analyzing Stored Procedures on Legacy Environment, produce documentation and re-create them on new environment.
* Created more stored procedures using Common Table Expression (CTE) and various types of UDF functions
* Used DDL and DML for writing triggers, stored procedures to check the data entry and payment verification.
* Analyzed the existing application programs and tuned SQL queries using Execution plan, SQL Performance Monitor, SQL Profiler.
* Created DDL scripts to create database schema and database objects.
* Expertise in solving any issues related Microsoft SQL server.
* Build and maintain complex SQL queries using T SQL for data analysis, data mining and data manipulation. Organize analytical data; ensure data quality and aggregate data for strategic reporting and performance optimization on database solutions.
* Different database triggers were created and stored in the database and fired off when contents of database were changed.
* Responsible for creating Databases, Tables, Cluster/Non-Cluster Index, Unique/Check Constraints Views, Stored Procedures, Triggers, Rules
* Extensively used joins and sub-queries for complex queries involving multiple tables from different databases.
* Created and reallocated database objects on appropriate files groups to easy maintenance and improve data access performance.
* Developed Scenarios for unit, integration testing to ensure that all the components work correctly when integrated.

**Environment**: Oracle 9i/10g, Windows XP, SQL\*PLUS, SQL Developer, B Crystal reports, data modeling, SSIS, SQL Server 2006.

**Education Details:**

Bachelors in mechanical engineering LBRCE, India- 2014