# Lakshmi Sudini DATAENGINEER

**Email**: lakshmisudini1@gmail.com | Contact: +18562593875

**SHORT SUMMARY:**
Experienced **Data Engineer** with **10+** years of expertise in **designing, developing, and deploying scalable data pipelines** and distributed systems. Proficient in cloud platforms **(AWS, Azure, GCP)** and **big data technologies (Spark, Hadoop, Kafka**, **Snowflake**). Skilled in **ETL/ELT** processes, data modeling, and data integration using tools like **Azure Data Factory, AWS Glue, SSIS, and DBT**. Strong background in **SQL, Python,** and **Scala** for **data transformation, machine learning,** and **real-time analytics**. Adept at leading **data migration** projects, optimizing data workflows, and delivering data-driven solutions. Experienced in **data visualization** tools like **Power BI, Tableau** to create interactive dashboards and reports. **Certified** in Azure Data Engineering (DP-203) and AWS Data Engineering.

**PROFESSIONAL OVERVIEW:**

* **10+ years of experience** as a Data Engineer, specializing in designing, developing, and deploying large-scale distributed systems and data pipelines.
* Expertise in **cloud platforms (AWS, Azure, GCP)**, leveraging services like **EC2, S3, Lambda, Glue, Redshift, Azure Data Factory, Synapse Analytics, and BigQuery** for scalable data solutions.
* Proficient in building and optimizing **ETL/ELT pipelines** using tools like **SSIS, Azure Data Factory, AWS Glue, DBT, and Python**.
* Strong experience in **data modeling**, including **Star Schema, Snowflake Schema**, and **Kimball Methodology**, for OLAP and OLTP systems.
* Skilled in **SQL, Python, and Scala** for data extraction, transformation, and loading (ETL), as well as performance tuning of queries and workflows.
* Proficient in writing **SQL queries, stored procedures, and Python scripts** for data extraction, transformation, and loading (ETL) processes.
* Hands-on experience with **big data technologies** like **Hadoop, HDFS, Spark, Kafka, Hive, Sqoop, and MapReduce** for batch and real-time data processing.
* Expertise in **real-time analytics** using **Spark Streaming, Kafka, and Azure Event Hubs** for high-throughput data pipelines.
* Proficient in **data warehousing** and migration projects, including on-premises to **Snowflake**, optimizing storage and query performance.
* Led data migration efforts to move on-premises data and workloads to **Snowflake**, ensuring minimal downtime and optimizing data storage and query performance.
* Integrated **ThoughtSpot** with **Snowflake** and other data warehouses to enable self-service analytics for business users.
* Designed and developed interactive dashboards and reports in **ThoughtSpot** to support data-driven decision- making.
* Strong knowledge of **NoSQL databases** like **Cassandra, MongoDB, and HBase** for scalable big data solutions.
* Experience in **data visualization** tools like **Power BI, Tableau, and ThoughtSpot** for creating interactive dashboards and reports.
* Skilled in **performance tuning** of **Spark jobs, Hive queries, and SSIS packages** to ensure optimal data processing efficiency.
* Proficient in **data pipeline development**, integrating data from **relational databases, flat files, APIs, and cloud storage**
* Manager and **SaaS, PaaS and IaaS** concepts of Cloud Computing and Implementation Worked with
* Extensive work on ETL processes consisting of data transformation, data sourcing, mapping, conversion and loading data from heterogeneous systems like flat files, **Excel, Oracle, Teradata, MSSQL Server.**
* Experience of building ETL production pipelines using **Informatica Power Centre, SSIS, SSAS, SSRS**.
* Expertise in **machine learning** using **PySpark MLlib, Scikit-Learn, and TensorFlow** for predictive modeling and data-driven decision-making.
* Experience with **data streaming platforms** like **Apache Kafka** and **AWS Kinesis** for real-time data ingestion and processing.
* Proficient in **orchestrating workflows** using **Apache Airflow, Oozie, and Azure Synapse Pipelines** for scheduling and managing data jobs.
* Strong knowledge of **data governance** and **compliance standards** like **GDPR and HIPAA** for secure data handling.
* Experience in **API integration** and deployment, including **REST, SOAP**, and tools like **MuleSoft** for seamless data exchange.
* Proficient in **version control tools** like **Git, Bitbucket**, and project management tools like **Jira, Confluence, and Rally**.
* Expertise in **data migration** and **cloud infrastructure setup** using **AWS CloudFormation, VPC, and Azure Resource Manager**.
* Hands-on experience with **Databricks Delta Lake** and **Snowflake** for building scalable and efficient data lakes and warehouses.
* Strong experience in project management, stakeholder communication, and creating documentation, data flow diagrams, and standard operating procedures.
* Hands-on experience with data visualization tools like **Power BI, Tableau, and QlikView**.

**CERTIFICATIONS:**

* + Microsoft Certified: Azure Data Engineer Associate (DP-203)
	+ Amazon Certified: AWS Data Engineer Associate
	+ Microsoft Certified: Azure Data Fundamentals (DP-900

**TECHNICAL SKILLS:**

|  |  |
| --- | --- |
| **Databases** | PostgreSQL, AlloyDB, Snowflake, MySQL, MongoDB, BigQuery, Oracle, SQL Server. ADLS, HDFS. |
| **Languages** | SQL, PL/SQL, Spark SQL, Shell Scripting, Python (Pandas, NumPy, Scikit Learn, OOPS, Functional Programming), Java, Pytest, Scala, Advanced Excel |
| **Cloud Technology** | **Azure Services:** Data Factory, Data Lake, ADLS Gen2, Data ricks, Logic Apps, Functional App, Event hubs, Azure Synapse Analytics.**GCP Services**: GCP (BigQuery, Dataflow, Pub/Sub, Cloud Storage, Dataproc, Cloud Composer, Cloud Functions, Cloud SQL, AlloyDB),**AWS Services:** (EC2, ECS, Lambda, RDS, S3 etc.), AWS SageMaker, Kinesis, Athena,Glue,EMR,Cloud Watch, DMS, Aurora, Apache Airflow,  |
| **Machine Learning** | Scikit-Learn, PySpark MLlib, TensorFlow, Keras, Regression,Classification, Clustering, Feature Engineering, Model Evaluation, AzureML, |
| **ETL Tools** | SSIS, Azure Data Factory, AWS Glue, Informatica PowerCenter, Talend, SnapLogic, DBT (Data Build Tool) |
| **Big Data Technologies** | Apache Spark (PySpark, Spark SQL), Hadoop (HDFS, Hive, Sqoop), Kafka, Flink, MapReduce, YARN, Kafka, HBase, Oozie, Zookeeper |
| **Data Modelling** | Star Schema, Snowflake Schema, Kimball Methodology, OLAP/OLTP systems. |
| **Data Visualization** | Power BI, Tableau, Looker, ThoughtSpot, QlikView, Qlik Sense |
| **IDE & Utilities** | SQL Developer, Dai Query using Hive, Visual Studio, CI/CD, SAS JupyterNotebook, PyCharm, Databricks, Dbeaver |
| **Project Management & Versioning Tools** | Jira, Bitbucket, Jenkins, Star Team, Git |
| **SDLC** | Waterfall, Agile (Scrum) |
| **Operating Systems** | Windows, Linux, UNIX |

**E****DUCATION:**

## Bachelor’s Degree in Computer Science July 2008 - June 2012

JNTUH, Hyderabad, India.

**WORK EXPERIENCE:**

# State Of Michigan, MI Sep 2022 - Present

**Sr. Data Engineer**

The State of Michigan’s Department of Health and Human Services (MDHHS) is responsible for public health, welfare,

and social services, serving millions of residents. The department relies on robust data systems to support healthcare programs, public safety and community services.

Supported MDHHS by modernizing data systems to improve public health reporting and decision-making. Migrated legacy systems to cloud platforms, designed data pipelines, and created dashboards to track key metrics, enabling better insights and operational efficiency. Collaborated with cross-functional teams to ensure data accuracy and compliance

with regulatory standards.

**Responsibilities:**

* Created high-level and low-level design documentation, ensuring alignment with standards and technical guidelines.
* Developed and maintained data models in **Azure Synapse**, optimizing query performance and scalability for reporting and analytics.
* Designed and implemented **ETL pipelines** in Azure Synapse using SQL pools, Linked Services, and Datasets for seamless data integration.
* Configured and managed **Azure Synapse Analytics** workspaces, ensuring cost efficiency and optimal performance.
* Migrated data ingress solutions to **AWS SFTP with S3** and **API Gateway**, enabling event-driven processing and inline data validation.
* Designed anomaly detection algorithms to identify data inconsistencies, reducing processing bottlenecks by **30%**.
* Maintained and enhanced an **SSO web application** for monitoring and administering data workflows.
* Built data ingestion pipelines for customer data from **FTP Servers** and **S3 buckets**, migrating traditional ETL jobs to **PySpark** and **Hive** on a cloud data lake.
* Developed **Java UDFs** for Hive queries and designed **Flink pipelines** to process streaming data from **Kafka**.
* Created and maintained logical/physical data models and metadata repositories using **AWS Glue** and **Crawlers**.
* Optimized **Spark jobs**, reducing processing times by **30%** for efficient large-scale data processing.
* Utilized **AWS Redshift** for ETL processes, creating tables and extracting data using **T-SQL** for business users.
* Designed and managed data integration pipelines using **SnapLogic iPaaS** for seamless data flow across systems.
* Led one-time data migration from **SQL Server** to **Snowflake** using **Python** and **Snow SQL**.
* Worked on **Azure DevOps CI/CD pipelines** for code deployment and created scalable data models in **DynamoDB**.
* Designed and managed **QlikView** and **Qlik Sense** dashboards, reports, and storytelling for business insights.
* Applied **DBT (Data Build Tool)** for modular data modeling, testing, and documentation of data pipelines.
* Expertise in **Databricks** for real-time data processing, streaming, and machine learning.
* Created on-demand tables on **S3** using **AWS Lambda** and **Glue** with **Python** and **PySpark**.
* Configured **Oracle HCM Cloud** for performance management, including process flows and evaluation templates.
* Developed **Scala scripts** and **UDFs** in **Spark** for data aggregation and integration with **RDBMS** via **Sqoop**.
* Experienced in using monitoring tools and performance diagnostics to identify and resolve performance bottlenecks in **Cassandra** clusters.
* Processed data from **S3** to **Snowflake** using **Tasks, Streams, Pipes**, and stored procedures.
* Built complex **stored procedures, SSIS packages, triggers, and views** for application development.
* Processed data using **Spark** before ingesting into **HBase**, creating batch and real-time jobs in **Scala**.
* Implemented cloud technologies like **IAM, EC2, S3, Lambda, Glue, and EMR** for scalable solutions.
* Monitored and optimized data pipelines using **Airflow** and custom logging for **Databricks** jobs.
* Automated scheduling and processes using **Shell** and **Python** scripting.
* Deployed and managed containerized applications using **Kubernetes**, converting VM-based apps to microservices.
* Collaborated with business users to deliver reporting solutions using **Power BI, Tableau**, and **ThoughtSpot**.
* Orchestrated **ETL processes** on **Cloudera Hadoop** using **Oozie** with **MapReduce, Pig, Hive**, and **Sqoop**.

**Environment**: SQL, Python, AWS EMR, Lambda, Redshift, Glue, IAM, Airflow,SSIS, DBT, Azure Data Factory, Azure Synapse, Scala, Flink, Spark, Pyspark, Hive, HDFS, MongoDB, QlikView, Kubernetes, Snowflake, GraphDB, Qlik Replicate, Snap Logic, OpenShift, SharePoint, Sqoop, Oozie, Kafka, HBase, Scala, MapReduce, ThoughtSpot, PowerBI, Tableau.

# Bank Of New York, Woodland Park, NJ Oct 2020 - Aug 2022 Data Engineer

BNY Mellon is a global investments company specializing in asset management, wealth management, and investment services. The organization relies on accurate and timely financial data to support regulatory reporting and business operations.

Transformed financial data systems by migrating legacy data to cloud platforms, enabling faster and more accurate reporting. Automated data workflows and built predictive models to enhance decision-making and support regulatory compliance.

Collaborated with business teams to deliver data solutions that improved accessibility and usability.

## Responsibilities:

* Extract Transform and Load data from Sources 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 in In **Azure Databricks**.
* Responsible for estimating the cluster size, monitoring, and troubleshooting of the **Spark data bricks cluster**.
* Implemented advanced statistical techniques using **PySpark** and **SQL** for predictive modeling, improving data- driven decision-making processes.
* Developed and migrated existing **Python code to Scala.**
* Built the data pipeline using Azure service like Data factory to load the data from Legacy **SQL** server to

## Azure Data Base using Data Factories, API Gateway Services, SSIS Packages, Talend Jobs, custom.Netand Python codes.

* Developed PL/SQL triggers and master tables for automatic creation of primary keys.
* Proficient in integrating **Azure Cosmos DB** with other Azure services such as Azure Functions, Azure Logic Apps, and Azure Event Grid for seamless data processing and workflows.
* 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.

## led the integration of Qlik Replicate with Azure Delta Lake, facilitating seamless and immediate data ingestion into the Delta Lake storage.

* Day to-day responsibility includes developing **ETL Pipelines in and out of data warehouse, develop major regulatory and financial reports using advanced SQL queries in snowflake.**
* Implement One time Data Migration of Multistate level data from **SQL server to Snowflake by using Python and Snow SQL.**
* Used various sources to pull data into Power BI such as **SQL Server. Ex& Oracle, SQL Azure** etc.
* Propose **architectures** considering cost/spend in **Azure and develop recommendations to right-size data infrastructure.**
* Automated resulting scripts and workflow using **Apache Airflow** and **shell scripting** to ensure daily execution in production.
* Ingested huge volume and variety of data from disparate source systems into **Azure Data Lake Gen2** using

## Azure Data Factory V2.

* Collaborate with application architects and **DevOps. Identify and implement best practices, tools, and standards.**
* Perform Data Cleaning, features scaling, features engineering using pandas and NumPy packages in python. **Environment**: SQL, Python, Scala , SSIS , Azure SQL Database, Azure data factory, ADLS Gen 2, Cosmos DB, Azure Analysis Service, Azure SQL Data warehouse, Qlik replicate, Airflow, PySpark, Kafka, Snowflake, Power BI

# CVS, New York, NY Mar 2019 - Sep 2020

**Sr.Data Engineer**

CVS Health is a leading healthcare company providing pharmacy services, health insurance, and retail health solutions. The organization relies on data-driven insights to optimize customer experiences and operational efficiency.

Modernized CVS’s data infrastructure by migrating on-premises systems to the cloud, improving scalability and reducing costs. Built data pipelines to analyze customer and operational data, enabling real-time insights for business optimization. Developed tools and dashboards to visualize key metrics, helping stakeholders make data-driven decisions..

## Responsibilities:

* Developed **Spark Applications by using Python** and implemented Apache Spark to process and handle data from various RDBMS and Streaming sources.
* Loaded data into **S3 buckets using AWS Lambda Functions, AWS Glue and PySpark and filtered data stored in S3 buckets using Elasticsearch and loaded data** into Hive external tables.
* Maintained and operated Hadoop cluster on AWS EMR.
* Migrated an existing on-premises application to **AWS and used AWS services like EC2 and S3 for small data sets**

## processing and storage.

* Scheduled **Spark/Scala jobs using Oozie workflow in Hadoop Cluster** and generated detailed design documentation for source-to-target transformations.
* Loaded data into S3 buckets using **AWS Lambda Functions, AWS Glue and PySpark** and filtered data
* stored in S3 buckets using Elasticsearch and loaded data into Hive external tables**. Maintained and operated Hadoop cluster on AWS EMR.**
* Architected and implemented end-to-end data pipelines using **Cloud Dataflow and BigQuery for real- time analytics processing,** reducing data processing time by 30%.
* Designed and deployed scalable data storage solutions using **Cloud Storage, Cloud Spanner, and Bigtable**

for multi-terabyte datasets.

* Led the migration of legacy data warehouse to **BigQuery**, improving query performance by 40% and reducing infrastructure.
* Developed real-time data streaming pipelines using **Cloud Pub/Sub and Cloud Functions** to process data from IoT devices.
* Used **Spark Streaming APIs** to perform transformations and actions on the fly for building common learner data model which gets the data from Kafka in real time and push it to Cassandra.
* Scheduled **Spark/Scala jobs** using Oozie workflow in Hadoop Cluster and generated detailed design documentation for source-to-target transformations.
* Developed Kafka consumer’s API in Python for consuming data from Kafka topics.
* Worked with Spark for improving performance and optimization of the existing algorithms in Hadoop using

## Spark Context, Spark-SQL, Spark MLlib, Data Frame, Pair RDD's and Spark YARN.

* Performed tuning of Spark Applications to set batch interval time and correct level of parallelism and memory tuning.
* Created live **real-time Processing and core jobs using Spark** Streaming with Kafka as a data pipe-line system.

## Worked on Amazon Redshift for shifting all Data warehouses into one Data warehouse.

* Designed columnar **families in Cassandra and Ingested data from RDBMS**, performed data transformations, and then exported the transformed data to Cassandra as per the business requirement.
* Designed, developed, deployed, and **maintained MongoDB.**
* Monitored and managed Snowflake's performance, scaling resources as needed to handle increased data load and query complexity.
* Worked extensively on Hadoop Components such as **HDFS, Job Tracker, Task Tracker, Name Node, Data Node, YARN, Spark and Map Reduce programming.**
* Worked extensively with Sqoop for importing and exporting the data from **HDFS to Relational Database systems (RDBMS) and vice-versa.**
* Written **Map reduce Jobs using PySpark, NumPy and** used Jenkins for continuous integration.
* Created HBase tables to load large sets of structured, semi-structured and unstructured data coming from UNIX, NoSQL, and a variety of portfolios.
* Generated reports using **Power BI and Tableau** based on client specification.
* Extracting the data from **the S3 using AWS Athena via SQL.**
* Implement solutions following security best practices and regulatory standards like GDPR, HIPAA ensuring ongoing data protection and compliance.

**Environment:** Python**,** Spark, Spark-Streaming, Spark SQL, AWS EMR, S3, EC2, MapR, HDFS, Hive, PIG, Apache Kafka, Sqoop, Python, Scala, PySpark, shell scripting, Linux, MySQL, NoSQL, Jenkins, Eclipse, Oracle, Git, Tableau, Power BI and Agile Methodologies, GCP,Google Big Query,Pub sub, Cloud Functions.

# Equifax, Alpharetta, GA Nov 2017 - Feb 2019 Data Engineer

Equifax is a global data, analytics, and technology company specializing in consumer credit reporting and risk management. The organization relies on accurate and secure data systems to support its services.

Developed a centralized data platform to store and analyze consumer data, improving credit reporting accuracy. Built data pipelines and predictive models to support fraud detection, customer segmentation, and regulatory compliance. Automated data quality checks and monitoring processes, reducing errors and improving compliance.

## Responsibilities:

* Designed and setup **Enterprise Data Lake** to provide support for various uses cases including Analytics, processing, storing and Reporting of voluminous, rapidly changing data.
* Enhanced data transformation processes using Snowflake’s cloud-native features, integrating diverse

datasets to support analytics and business intelligence.

* 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 &amp; solution architect.
* Designed and developed Security Framework to provide fine grained access to objects in **AWS S3 using AWS Lambda, DynamoDB.**
* Performed end- to-end Architecture; **implementation assessment of various AWS services like Amazon EMR, Redshift, S3.**
* Built robust data pipelines on **AWS** cloud environment using **Snowflake** for data storage and **DBT**

for transformation and modeling.

* Implemented data governance policies, ensuring security, compliance, and proper usage of sensitive information within the Snowflake environment.
* Designed efficient **Dimensional Models** to support business intelligence and reporting functions,improving query performance and user experience.
* Configured **Snow pipe** to pull the data from S3 buckets into Snowflakes table and stored incoming data in the Snowflakes staging area.
* Implemented the machine learning algorithms using python to predict the quantity a user might want to order for a specific item so we can automatically suggest **using kinesis firehose and S3 data lake.**
* Used **AWS EMR** to transform and move large amounts of data into and out of other AWS data stores and databases, such as **Amazon Simple Storage Service (Amazon S3) and Amazon DynamoDB.**
* Used Spark SQL for Scala &amp; amp, Python interface that automatically converts RDD case classes to schema RDD.
* Creating Lambda functions with **Boto3** to deregister unused AMIs in all application regions to reduce the cost for EC2 resources.
* Developed reusable framework to be leveraged for future migrations that automates **ETL from RDBMS systems to the Data Lake utilizing Spark Data Sources and Hive data objects.**
* Conducted Data blending, Data preparation using Pandas and SQL for Tableau consumption and publishing data sources to Tableau server.
* Developed Kibana Dashboards based on the Log stash data and Integrated different source and target systems into Elasticsearch for near real time log analysis of monitoring End to End transactions.
* Implemented **AWS** Step Functions to automate and orchestrate the **Amazon SageMaker** related tasks such as publishing data to **S3, training ML model** and deploying it for prediction.
* Automated nightly build to run quality control using **Python** with **BOTO3** library to make sure pipeline does not fail which reduces the effort by 70%.
* Worked on **AWS Services like AWS SNS** to send out automated emails and messages using **BOTO3** after the nightly run.
* Implement solutions following security best practices and regulatory standards like **GDPR, HIPAA,** and others, ensuring ongoing data protection and compliance.

**Environment:** AWS EMR, S3, RDS, Redshift, Lambda, Boto3, DynamoDB, Amazon SageMaker, Apache Spark, HBase, Apache Kafka, HIVE, SQOOP, Map Reduce, Snowflake, Python, Tableau

# Polaris, Hyderabad, India Oct 2012 - Nov 2015

**Data Analyst**

Polaris is a global leader in digital transformation and technology solutions, serving industries such as banking, insurance, and healthcare. The organization focuses on delivering innovative solutions to improve business

operations.

Designed and implemented data models and reporting systems to improve business intelligence. Developed ETL processes to integrate data from multiple sources, enabling better analysis and decision-making for stakeholders. Created tools and dashboards to visualize key metrics, helping leadership track performance and identify trends.

## Responsibilities:

* Developed stored procedures in **MS SQL** to fetch the data from different servers using **FTP** and processed these files to update the tables.
* Responsible for Designing Logical and Physical data modeling for various data sources on

**Confidential Redshift.**

* Performed **logical data modeling, physical Data modeling (including reverse engineering)** using the **Erwin**

**Data modeling tool.**

* FCreated dimensional model for the reporting system by identifying required dimensions and facts using Erwin.
* Designed and Developed ETL jobs to extract data from Salesforce replica and load it in data mart in Redshift.
* Involved in performance tuning, stored procedures, views, triggers, cursors, pivot, unpivot functions, CTE's
* Extensively used Erwin for Data modeling. Created Staging and Target Models for the Enterprise Data Warehouse.
* Involved in Normalization / De normalization techniques for optimum performance in relational and dimensional database environments.
* Resolved the data type inconsistencies between the source systems and the target system using the Mapping Documents and analyzing the database using SQL queries.
* Worked on ETL testing, and used SSIS tester automated tool for unit and integration testing.
* Designed and created SSIS/ETL framework from ground up.
* Created new **Tables, Sequences, Views, Procedure, Cursors and Triggers** for database development.
* Creating reports using **SQL Reporting Services (SSRS)** for customized and ad-hoc Queries
* Coordinated with clients directly to get data from different databases.
* Worked on **MS SQL Server,** including **SSRS, SSIS, and T-SQL.**
* Designed and developed schema data models.
* Documented business workflows for stakeholder review.

**Environment:** ER Studio, SQL Server 2008, SSIS, Oracle, Business Objects XI, Rational Rose, Data stage, MS Visio, SQL, Crystal Reports 9