# Lakshmi Sudini DATAENGINEER

**Email**: [lakshmisudini1@gmail.com](mailto:lakshmisudini1@gmail.com) | Contact: +18562593875

**PROFESSIONAL SUMMARY:**

* Around **10 years** of professional experience as a Data Engineer Specializing in designing and deploying scalable data ingestion pipelines using **Google Cloud Dataflow** and **Azure Data Factory**, with expertise in advanced functionalities such as **copy activity**, **lookup activity**, and seamless integration with **Google Cloud Storage (GCS)** and **Azure Data Lake Storage Gen2 (ADLS Gen2)** for optimized data storage and management.
* **Led the migration of on-premises data infrastructure to the cloud** using **Google Cloud Dataflow** and **Azure Data Factory**, ensuring minimal downtime and maximum data integrity during the transition.
* **Designed and implemented robust data pipelines** for efficient data transfer and storage in **GCS** and **ADLS Gen2**, significantly improving scalability, reliability, and cost-efficiency for enterprise-level operations.
* **Proficient in building high-performance data processing systems** using **Databricks**, **Google Dataproc**, and **PySpark API**, enabling efficient handling of both batch and real-time data for rapid and actionable insights.
* **Expert in designing and deploying ETL pipelines** that integrate **Teradata** with cloud platforms such as **Google BigQuery**, **Azure Synapse**, and **Snowflake**, ensuring seamless data migration and processing across hybrid environments.
* **Optimized serverless functions** using **Google Cloud Functions** and **Azure Functions** for ETL processes, incorporating **Pub/Sub** and **EventHub** for real-time data streaming and event-driven architectures.
* **Seasoned data professional** with a strong focus on **ETL pipeline management**, leveraging advanced compression techniques to enhance storage efficiency and ensure high-speed data transfers.
* **Skilled in advanced data warehousing techniques**, including **data cleansing**, **Slowly Changing Dimensions (SCD)**, **surrogate key assignment**, and **Change Data Capture (CDC)** within **Snowflake** and **BigQuery**, ensuring data accuracy and integrity.
* **Utilized BigQuery and Snowflake** for efficient data storage, retrieval, and analytics, integrating **GCS** and **ADLS Gen2** for comprehensive data management solutions.
* **Configured and automated data workflows** in **Google Cloud Dataflow**, **Azure Data Factory**, and **Snowflake** using **Python**, aligning with **Medallion Architecture** principles for scalable and maintainable data pipelines.
* **Proficient in Teradata** for high-performance data analytics, employing advanced techniques such as **indexing**, **partitioning**, and **query optimization** to maximize data retrieval efficiency.
* **Expert in building scalable data ingestion pipelines** using **Apache Kafka**, **Apache NiFi**, and **Apache Flume**, with a focus on real-time data processing through **Pub/Sub** and **EventHub**.
* **Experienced in developing and maintaining ETL/ELT pipelines** using **Apache Spark**, **Apache Beam**, **Google Cloud Dataflow**, and **Apache Airflow**, integrating **Event Queues** for streamlined event-driven data operations.
* **Implemented robust data quality checks** and cleansing processes to ensure data accuracy and integrity across pipelines, while applying compression techniques to optimize storage and processing efficiency.
* **Designed and optimized data models** using **Apache Hive**, **Apache HBase**, and **Dataplex**, ensuring efficient data organization and retrieval for analytics and reporting.
* **Extensive experience with big data technologies** such as **Hadoop**, **HDFS**, **MapReduce**, **Hive**, **Python**, **PySpark**, and **Google Dataproc**, enabling scalable and efficient data processing.
* Proficient in managing data science projects life cycle including data acquisition, cleaning, engineering, feature scaling, engineering, and modelling **(Regression Models, Classification and Clustering, Decision Trees, Naive Bayes, Random Forest, Gradient Boosting, SVM, KNN, Neural Networks).**
* **Proficient in query optimization** and indexing strategies to enhance data retrieval performance, with hands-on experience in **Vertica**, **Teradata**, and **BigQuery** for high-performance analytics.
* **Skilled in SQL programming**, including **DDL** and **DML**, for data manipulation and retrieval, with expertise in connecting diverse data sources using **linked services** in **Google Cloud Dataflow** and **Azure Data Factory**.
* **Automated data loading processes** for staging, intermediate, and core tables in **Snowflake** and **BigQuery** using **Matillion**, ensuring efficient and timely data availability.
* **Strong experience in data integration** across **Hadoop**, **Google Cloud Dataflow**, and **RDBMS environments**, utilizing file formats such as **Avro**, **Parquet**, **JSON**, **ORC**, and **text** for efficient data loading, parsing, and transformation.
* **Proficient in managing Hadoop distributions** such as **Cloudera** and **Hortonworks**, leveraging **Event Queues** for streamlined data processing and management.
* **Designed Hive external tables** with **static and dynamic partitioning**, **bucketing**, and **indexing** to enhance query performance in **Vertica**, **Teradata**, and **BigQuery**.
* **Leveraged Apache Spark** to optimize existing algorithms in **Hadoop** and **Google Dataproc**, utilizing **Spark SQL**, **DataFrames**, and **RDDs** for improved performance and efficiency.
* **Hands-on experience in ETL processes** using **Power Query Editor**, **Data Modeling**, and creating interactive **Reports/Dashboards** in **Power BI** and **Looker Studio**.
* **In-depth knowledge of the Software Development Life Cycle (SDLC)**, with expertise in **requirements analysis**, **design**, and **development** phases.
* Hands-on experience with data visualization tools like **Power BI, Tableau, and QlikView**.
* **Experienced in creating advanced visualizations** such as **tree maps**, **funnel charts**, and **custom visuals** in **Tableau** and **Looker Studio** for interactive data analysis and reporting.
* **Proficient in implementing CI/CD pipelines** for data workflows using **Jenkins**, **Google Cloud Build**, and **Cloud Deploy**, ensuring seamless automation and deployment in cloud and hybrid environments.
* **Skilled in Agile methodologies**, fostering cross-functional collaboration to enhance project delivery efficiency and adaptability in dynamic environments.
* Strong experience in project management, stakeholder communication, and creating documentation, data flow diagrams, and standard operating procedures.
* Experience with **Salesforce object and data structures**, **APIs (REST, SOAP)**, and integration tools like MuleSoft.

**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), Pytest, Scala, Advanced Excel |
| **Cloud Technology** | **GCP Services**: GCP (BigQuery, Dataflow, Pub/Sub, Cloud Storage, Dataproc, Cloud Composer, Cloud Functions, Cloud SQL, AlloyDB),Cloud Key Management Service (Cloud KMS), Identity and Access Management (IAM)  **Azure Services:** Azure Data Factory, Azure Data Lake, ADLS Gen2, Azure Data ricks, Logic Apps, Functional App, Azure Event hubs, Key Vault, Azure Active Directory, Azure Synapse Analytics.  **AWS Services:** (EC2, ECS, Lambda, DynamoDB, RDS, S3, SQS, CI/CD Tools, etc.), AWS SageMaker, AWS Kinesis, AWS Athena, AWS Glue, AWS EMR, AWS CloudWatch, AWS DMS, AWS Aurora, Apache Airflow, ETL, ELT, Adobe Analytics, MuleSoft, Snowflake. |
| **Machine Learning** | Scikit-Learn, PySpark MLlib, TensorFlow, Keras, Regression,  Classification, Clustering, Feature Engineering, Model Evaluation, AzureML, AWS Sagemaker |
| **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, ThoughtSpot, QlikView, Qlik Sense |
| **IDE & Utilities** | SQL Developer, Dai Query using Hive, Visual Studio, CI/CD, SAS Jupyter  Notebook, PyCharm, Databricks, Dbeaver |
| **Project Management & Versioning Tools** | Jira, Bitbucket, Jenkins, Star Team, Git |
| **SDLC** | Waterfall, Agile (Scrum) |
| **Scheduling Tools** | Autosys, Tivoli, Control-M, Tidal, AWS Cloud Services (CloudWatch, AWS Batch, AWS Lambda, AWS SWF), Apache Airflow, Google Cloud Composer, Data Fusion |
| **Operating Systems** | Windows, Linux, UNIX |
| **Tools & Utilities** | Atlassian Jira, FileZilla, WinSCP, MS-Visio, MS Office, Lucid Chart |

**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:**

* **Designed and implemented a robust healthcare data analytics platform** using **GCP services** such as **BigQuery**, **Dataflow**, and **Pub/Sub** to process and analyze large-scale patient data, enabling real-time insights for healthcare providers and policymakers.
* **Migrated on-premises healthcare data** to **GCP**, leveraging **Cloud Storage** and **BigQuery** for scalable data warehousing, reducing infrastructure costs by **30%** and improving query performance by **40%**.
* **Developed and optimized ETL pipelines** using **Cloud Dataflow** and **Apache Beam** to ingest, transform, and load data from diverse sources, including **SQL databases**, **CSV files**, and **REST APIs**, into **BigQuery** for analytics and reporting.
* **Built real-time data streaming solutions** using **Pub/Sub** and **Dataflow** to process **IoT data** from medical devices, enabling real-time monitoring of patient vitals and improving healthcare outcomes.
* **Automated data workflows** using **Cloud Composer** (Apache Airflow) and **Cloud Functions**, reducing manual effort by **50%** and ensuring seamless data integration across systems.
* **Implemented data quality checks** and **data cleansing techniques** using **Dataproc** and **Cloud Dataflow**, ensuring the accuracy and integrity of healthcare data throughout the pipeline.
* **Designed and implemented STAR schema data models** in **BigQuery** to support efficient querying and reporting for healthcare analytics, improving query performance by **30%**.
* **Integrated PySpark with Dataproc** for large-scale data processing and transformation, enabling efficient handling of **batch** and **real-time data** for healthcare analytics.
* **Utilized Google Secret Manager** for secure management of secrets and certificates, ensuring compliance with **HIPAA** and **GDPR** regulations.
* **Developed interactive dashboards** in **Looker** and **Power BI** to provide real-time insights into patient outcomes, healthcare utilization, and operational efficiency, enabling data-driven decision-making.
* **Collaborated with data scientists** to deploy **machine learning models** on **Vertex AI**, integrating predictive analytics into healthcare workflows for early disease detection and personalized treatment plans.
* **Orchestrated a large-scale data migration** from legacy systems to **GCP**, utilizing **Snowflake** for efficient data warehousing and storage solutions, ensuring minimal downtime and data loss.
* **Implemented Snowpipe** for continuous data ingestion into **Snowflake**, automating the loading of real-time data from various sources into the data warehouse.
* **Optimized PySpark jobs** for performance by leveraging techniques like **partitioning** and **caching**, reducing processing times by **25%** and improving system efficiency.
* **Designed and deployed scalable data pipelines** using **Cloud Dataflow** and **Dataproc**, ensuring high availability and reliability for critical healthcare applications.
* **Utilized JDBC connectors** to establish and manage secure data access from on-premises **SQL Servers** to **Dataproc**, ensuring seamless data integration and processing.
* **Implemented Dataplex** to establish comprehensive data governance and cataloging, ensuring data discovery, classification, and lineage tracking across the organization’s data landscape.
* **Collaborated with cross-functional teams**, including data analysts, data scientists, and business stakeholders, to understand reporting requirements and deliver customized reports driving strategic business decisions.
* **Integrated GitHub repositories** with **GCP services** for enhanced collaboration and automated deployment workflows within the **GCP ecosystem**.
* **Utilized JIRA** for project reporting and created sub-tasks for development, QA, and partner validation, ensuring timely delivery of data engineering solutions.
* **Conducted performance tuning** and **capacity planning exercises** to ensure scalability and efficiency of data infrastructure, reducing operational costs by **20%**.
* **Developed complex SQL queries** and **data models** in **BigQuery** to integrate big data processing and analytics capabilities, enabling seamless data exploration and insights generation.
* **Automated CI/CD pipelines** using **Cloud Build** and **Jenkins** for deploying data workflows, improving code quality and project management efficiency.
* **Implemented data security measures**, including **IAM roles**, **VPC**, and **encryption**, to ensure compliance with **HIPAA** and **GDPR** regulations.
* **Designed and implemented real-time data processing solutions** using **Pub/Sub** and **Spark Streaming**, facilitating the ingestion, transformation, and analysis of high-volume streaming data.
* **Developed data processing workflows** using **Dataproc** and **Spark** for distributed data processing and transformation tasks, ensuring fast and reliable insights for healthcare analytics.
* **Utilized Cloud Monitoring** for operational insights and analytics, enabling proactive identification and resolution of performance bottlenecks.
* **Integrated Snowflake with GCP cloud services** to establish secure and efficient data warehousing solutions, enabling insightful reports for strategic analysis.
* **Designed and implemented scalable ETL processes** using **Databricks Delta**, **Spark**, and **Python** to process large volumes of healthcare data, ensuring high performance and scalability..
* **Optimized Qlik Replicate performance** by tuning replication settings and monitoring task statuses.
* **Extensive Shell/Python Scripting experience** for scheduling and process Automation.
* **Experienced in using monitoring tools** and performance diagnostics to identify and resolve performance bottlenecks in **Cassandra clusters**.
* **Scheduled, deployed, and managed container replicas** onto a node cluster using **Kubernetes** and converted **VM-based applications** to **microservices** and deployed as a container managed by **Kubernetes**.
* **Collaborated with business users** to understand reporting requirements and delivered solutions using **BI tools** like **Power BI** and **Tableau**, with a focus on transitioning to **ThoughtSpot** for advanced analytics.
* **Designed, developed, built, tested, and maintained Tableau Analytics** and functional reports based on user requirements.
* **Plan and orchestrate ETL processes** on **Cloudera Hadoop** by integrating **Oozie** with **Map-Reduce**, **Pig**, **Hive**, and **Sqoop** and developing **Oozie workflows**.

**Environment**: GCP (BigQuery, Dataflow, Pub/Sub, Cloud Storage, Dataproc, Cloud Composer, Cloud Functions), Snowflake, Apache Spark (PySpark, Spark SQL), Python, SQL, Looker, Power BI,Tableau,Vertex AI, Dataplex, Cloud Build, Jenkins, JIRA, GitHub, IAM, VPC, HIPAA, GDPR, Sqoop, Oozie, Kafka, HBase, Scala

# 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