# 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, developing, deploying, and supporting large-scale distributed systems.
* Extensive expertise in building scalable cloud infrastructure on **AWS and Azure**, leveraging services such as **Cloud Formation, EC2, VPC, Elastic Load Balancing, Auto Scaling, Route 53, RDS, S3, Lambda, SNS , SQS, Glue, Athena, and Redshift Azure Synapse Analytics, and Azure Data Factory** to optimize data workflows and system performance.
* Strong knowledge and experience in implementing **ETL/ELT pipelines** using **SSIS, Azure Data Factory, AWS Glue, and Python** for data integration, transformation, and loading.
* Proven ability to design, develop, and maintain complex **ETL/ELT** workflows using tools like **SSIS, Azure Data Factory, and DBT (Data Build Tool)** for data integration tasks, ensuring efficient and scalable data pipelines.
* Hands-on experience with **SQL Server Integration Services (SSIS)** for creating, troubleshooting, and optimizing SSIS packages and SQL queries for better performance.
* Expertise in **data modeling, process design, and data pipeline architecture**, including dashboard/report development and end-to-end ETL processes.
* Proficient in writing **SQL queries, stored procedures, and Python scripts** for data extraction, transformation, and loading (ETL) processes.
* Strong experience in **data pipeline development**, integrating data from various sources like relational databases, flat files, APIs, and cloud storage.
* Skilled in **performance tuning** of ETL workflows, SQL queries, and SSIS packages to ensure optimal data processing and system performance.
* Experience in **leading small to large data engineering projects**, managing risks, and ensuring timely delivery of technical solutions.
* 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.
* Proven ability to design, develop, test, operate, and maintain enterprise-level API service architectures to support seamless integration and deployment across on-premises and cloud environments.
* Ability to build machine learning solutions using **PySpark for large datasets on Hadoop ecosystem**.
* Build the infrastructure required for integration and deployment of API services both on-prem and cloud.
* 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).**
* Expertise in **Azure Data Factory, Glue, Azure Synapse Analytics, Amazon Redshift,** and **Databricks** for batch and real-time data processing. Skilled in data streaming platforms like **Apache Kafka, Azure Event Hubs.**
* Hands-on exposure with big data technologies, including **Hadoop, HDFS, Spark, Kafka, Pig, Hive, Sqoop, HBase, Oozie, Zookeeper, Cloudera Manager, and Hortonworks**, for efficient data processing and analytics. Experienced in building **PySpark, Spark Java, and Scala applications for batch and stream processing, involving Transformations, Actions, Spark SQL queries on RDDs, DataFrames, and Datasets**.
* Strong experience writing, troubleshooting and optimizing Spark scripts using Python, Scala.
* Experienced in using Kafka as a distributed publisher-subscriber messaging system.
* Strong knowledge on performance tuning of Hive queries and troubleshooting various issues related to Joins, memory exceptions in Hive.
* Exceptionally good understanding of partitioning, bucketing concepts in Hive and designed both Managed and External tables in Hive.
* Experience in importing and exporting data between **HDFS** and Relational Databases using Sqoop.
* Experience in real time analytics with **Spark Streaming, Kafka** and implementation of batch processing using **Hadoop, Map Reduce, Pig and Hive**.
* Experienced in building highly scalable Big-data solutions using **NoSQL** column-oriented databases like
* **Cassandra, MongoDB** and HBase by integrating them with Hadoop Cluster.
* Hands on experience in creating real time data streaming solutions using **Apache Spark Core, Spark SQL, and Data Frames.**
* 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**.
* Proficient at writing **MapReduce jobs and UDF’s** to gather, analyse, transform, and deliver the data as per business requirements and optimize the existing algorithms for best results.
* Strong expertise in data warehousing concepts, including **Star Schema, Snowflake Schema, and Kimball Methodology**, with sound knowledge of normalization and denormalization techniques for **OLAP and OLTP** systems
* Good experience with version control tools like **Bitbucket, GitHub, and Git**, and project management tools like
* **Jira, Confluence, and Rally**. Proficient in scheduling tools like **Oozie and Apache Airflow**

# Proficient in utilizing Azure Synapse Pipelines for orchestrating and managing complex data integration and transformation workflows.

* Data pipelining using frameworks based on **Kafka and Spark and Kafka producer APIs to send live- stream data into various Kafka topics.**
* Hands on experience in setting up workflow using **Apache Airflow** engine for managing and scheduling Hadoop jobs.
* Designed and implemented **scalable ETL processes using Databricks Delta, Spark, and Python to process large volumes of data.**
* Developed SQS message queues and integrated them with AWS Lambda functions to trigger automatic processing of incoming messages.
* Monitored **DynamoDB** performance metrics and optimized table configurations for improved throughput and latency.
* 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**.
* 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** | **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.  **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)  **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:**

* Creation of high-level and low-level design documentation for various modules. Review your design to make sure it meets standards, patterns, and company guidelines. Verification of design specifications against proof of concept and technical considerations.
* Developed and maintained data models in **Azure Synapse** to support reporting and analytics, optimizing query performance and scalability.
* Designed and implemented ETL pipelines in Azure Synapse using dedicated **SQL pools, Linked Services, and Datasets** for seamless data integration.
* Configured and managed **Azure Synapse Analytics workspaces,** ensuring optimal performance and cost efficiency.
* Migrated existing data ingress solutions to a cloud-based solution using **AWS SFTP with S3 and API Gateway**, enabling event-driven processing and inline data validation.
* Designed and implemented anomaly detection algorithms to identify data inconsistencies, missing data, and unrealistic data scenarios, reducing data processing bottlenecks by 30%.
* Maintained and enhanced an **SSO** web application for configuration, monitoring, and administration of data processing workflows.
* Involved in creating data ingestion pipelines for collecting customer data and subjects’ data from various external sources like **FTP** Servers and **S3** buckets in migrating existing traditional **ETL** jobs to **Pyspark** and Hive Jobs on new cloud data lake.
* Developed user-defined functions in **Java** when needed for hive queries.
* Designed and developed **Flink** pipelines to consume streaming data from Kafka and applied business logic to massage and transform and serialize raw data.
* Performs data analysis and design, and creates and maintains large, complex logical and **physical data models, and**

## metadata repositories using AWS GLUE and Crawlers.

* Optimized Spark jobs for performance tuning, reducing processing times by 30%, ensuring efficient data processing at scale.
* Using AWS Redshift, Extracted, transformed and loaded data from various heterogeneous data sources and destinations Created Tables and extracted data using T-SQL for business users whenever required.
* Utilized **SnapLogic iPaaS** to design and manage data integration and ETL pipelines, enabling seamless data flow across multiple systems.
* Implement One time Data Migration of Multistate level data from SQL server to snowflake by Using Python and **Snow SQL.**
* Solid understanding of graph database concepts including nodes, edges, and properties, with practical experience in handling complex relationships and interconnected data.
* Worked **on Azure Devops CI/CD pipelines** for deployment for code.
* Created efficient and scalable data models in **DynamoDB** to support high-volume transactional applications.
* Designed, developed, and managed **QlikView**, Qlik Sense Apps including Dashboard, Reports, Storytelling.
* Familiar with **DBT (Data Build Tool)** concepts, including modular data modeling, testing, and documentation, to build maintainable and testable data pipelines.
* Worked on Dimensional Data modelling in **Star and Snowflake schemas** and **Slowly Changing Dimensions (SCD)**.
* Expertise in using **Databricks** for real-time data **processing, streaming, and machine learning.**
* Responsible for creating on-demand tables on S3 files using Lambda Functions and AWS Glue using Python and

## PySpark.

* Configure the **Oracle HCM** Cloud Performance Management module to incorporate interim evaluations for various qualifying groups, process flows, review procedures, and multiple performance templates.
* Experience in Oracle supplied packages, Dynamic SQL, Records and **PL/SQL** Tables.
* The deployment model uses Atlassian development repository tools, Jenkins as the build engine, while execution deployments to container orchestration tools ranged over time from OpenShift on EC2, AWS.
* Designed a reusable **REST API** framework for seamless data consumption and insertion into **MongoDB**.
* Developed **Scala scripts**, UDF's using both **Data frames/SQL and RDD/MapReduce** in Spark for **Data Aggregation**, queries and writing data back into RDBMS through Sqoop.
* Process Location and Segments data from s3 to Snowflake by using Tasks, Streams, Pipes, and stored procedures.
* Worked in creating complex stored procedures, **SSIS** packages, triggers, cursors, tables, and views and other SQL joins and statements for applications.
* Use Spark to process the data before ingesting the data into the HBase. Both Batch and real-time spark jobs were created using **Scala**.
* Performed implementation of cloud-specific technologies including IAM, Amazon Cloud Services like **Elastic Compute Cloud (EC2), ElastiCache, Simple Storage Services (S3), Cloud Formation, Virtual Private Cloud (VPC), Route 53, Lambda, Glue, EMR**.
* Monitored and optimized performance of data pipelines using **Airflow**'s built-in tools and custom logging for

Databricks jobs.

* 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 application 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**: 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