**Aditya K**

**Mobile: 682-256-7419**

**Email: aditya.dtaeng@gmail.com**

[**https://www.linkedin.com/in/aditya-k-239ab0191/**](https://www.linkedin.com/in/aditya-k-239ab0191/)

**Senior Data Engineer**

PROFESSIONAL SUMMARY

10+ years of experience in Information Technology with a strong background in Database development, Data warehousing, Design principles, ETL process, Data Visualization.

5+ years of experience in cloud computing technologies which include Amazon Web Services (AWS), Microsoft Azure, Google Cloud Platform (GCP).

Strong experience in various ETL tools like Informatica Power Center, IDMC/IICS, DataStage, Dell Boomi, MS SSIS with various databases, including SQL Server, DB2, Oracle, MySQL, and cloud databases like AWS Redshift and Google BigQuery.

Experience in GCP Dataproc, GCS, Cloud functions, BigQuery.

Proficient in using Informatica Data Quality (IDQ) for profiling, cleansing, standardizing, and enriching data to ensure high data quality and integrity across various business processes.

Extensive experience with T-SQL in constructing Triggers, Tables, implementing stored Procedures, Functions, Views, User Profiles, Data Dictionaries and Data Integrity.

Knowledge of designing Dimensional models for Data Mart and Staging databases.

Experience in implementing different types of Replication models like Snapshot, Transactional and Merge Replication.

Extensive experience in **Master Data Management (MDM),** ensuring data consistency, quality, and governance across enterprise systems using Azure Data Services.

Experience in Big Data / Hadoop Ecosystem technologies which includes Map Reduce, Spark, Cloudera Manager, Cassandra, Zookeeper, Sqoop, Oozie, GCP, Amazon web services, S3.

* Extensive Knowledge in designing functional and detailed design documents for data warehouse development.
* Experience in creating and modifying UNIX shell, Perl & Python scripts.
* Built application metrics and alerting dashboards using Kibana.
* Built kibana dashboard for real time logs aggregations and analysis.
* Hands on experience in writing, testing unit testing and debugging stored procedures, triggers and troubleshooting database issues.
* Good knowledge of different business domains like Insurance, e-Commerce, Health Care, Automobile industries.
* Experience in analyzing existing code and identifying potential issues.

Strong experience in Performance tuning, Query Optimization.

Experience in Data mining, finding correlations or patterns among fields in Database.

Worked in Agile methodology and Kanban method for faster development.

Good exposure of Electronic Data Interchange (EDI) transaction files and HIPAA standards.

Experience in creating Ad-hoc reports involving complex datasets and query database for BI.

Understanding of Relational (ROLAP), Multidimensional (MOLAP) modeling, data warehousing concepts, star and schema, database design methodologies and Meta data management.

Experience in working with Stored Programs, Packages, Cursors, Triggers, Database Link, Snapshot, Roles, Privileges, Tables, Constraints, Views, Indexes, Sequence, and Synonyms Dynamic SQL and SQL\*Loader in distributed environment.

Experience in understanding Fact Tables, Dimension Tables, Summary Tables.

Scheduled the ETL, SQL jobs in scheduling tools.

Skilled in developing Test Plans, Creating and Executing Test Cases.

Experience with 24x7-production support.

Capable of forming and maintaining positive and productive working relationships in internal/external and team environments.

Experience with DevOps tools like GIT, Kubernetes, Docker.

Excellent Analytical, Written and Communication skills.

Technical Skills

|  |  |
| --- | --- |
| **ETL**  | Informatica Power Center, IDQ, DataStage, Dell Boomi, SSIS, Alteryx |
| **Cloud Technologies** | AWS, GCP, MS Azure, Salesforce. |
| **Google Cloud Platform:** | GCP Cloud Storage, Big Query, Composer, Cloud Data proc, Cloud SQL, Cloud Functions, Cloud Pub/Sub, Dataflow etc. |
| **Databases** | Oracle, DB2, Teradata, SQL Server, Red Shift,MongoDB, Netezza, Cassandra, Snowflake, DBT, Matillion, PostgreSQL. |
| **Reporting Tools** | Business Objects, Tableau, SSRS, Cognos, Alteryx. |
| **Database Modeling** | SAP Sybase, Erwin, Snowflake, Sigma. |
| **Master Data Management (MDM):** | Azure Purview, Profisee, Informatica MDM, Data Governance, Data Quality, Golden Record Management |
| **Big Data Technologies**  | Hadoop, HDFS, Hive, MapReduce, Pig, Sqoop, Flume, Oozie, Hadoop distribution, And HBase, Spark, Kafka, Apache Flink. |
| **Languages**  | T-SQL, COBOL, PL/SQL, JAVA, C, Python |
| **Web Tools**  | HTML, XML, JavaScript, Servlets, EJB |
| **Scripting/ Web Languages** | JavaScript, HTML5, CSS3, XML, SQL, Shell |
| **Software Life Cycles** | SDLC, Waterfall and Agile models |
| **Operating System** | Windows NT/2000/2003/7, UNIX, Linux. |
| **Analytical Skills**  | Data Analysis, SQL, Python, Data visualization (Tableau, Power BI), Sigma. |
| **Utilities /Tools** | Eclipse, Intellij, Tomcat, NetBeans, Junit, SQL, SNV, Log4j, SOAP UI, ANT, Maven, Automation and MR-Unit. |
| **Testing Skills** | System & Unit Testing, Acceptance testing, Regression Testing, Load Testing |
| **Apps & Web servers**  | Apache Webserver & Tomcat. |
| **Office Tools** | MS-Office, MS-Project and Risk Analysis tools, Visio |

educational Information

Bachelor's Degree & Passed out Year: 2013 JNTU

Master’s degree & Passed out Year: 2019, CSE, Cumberland university

professional experience

**Amazon, Dallas, TX March 2022 – Till date**

**Sr AWS Data Engineer**

Customer Segmentation project focusing on enhancing customer relationships and operational efficiency. Implemented machine learning models for customer segmentation, credit risk assessment, and fraud detection, resulting in significant improvements in marketing ROI, reduced loan defaults, and lower fraud incidents. Automated key data processes introduced a recommendation system to increase user engagement and utilized analytics and NLP to refine web interfaces and product features.

* Led the design and implementation of a customer segmentation project using **AWS S3** for data storage, **Python**, and **Pandas** for data manipulation, applying **K-means clustering** in **Scikit**-learn to segment customers, enhancing marketing strategies.
* Designed and implemented scalable Data Lake solutions like AWS S3 to consolidate structured and unstructured data from multiple sources.
* Designed and optimized Snowflake data warehouses to support large-scale analytics and reporting, ensuring cost-efficiency, scalability, and performance.
* Developed a credit risk assessment model using **AWS Sage Maker** for model training and deployment, applying logistic regression in **Python** and **R** to reduce loan defaults by 18%.
* Engineered a real-time fraud detection system using **AWS Kinesis** for stream processing and data ingestion, and **Apache Spark** for analytics, reducing fraudulent transactions by 22%..
* Built real-time and batch data pipelines using tools like Spark, Kafka and AWS Glue to ingest data into Data Lakes.
* Used Informatica IICS to prep and shape data before feeding it into Tableau dashboards, streamlining the data lifecycle from ingestion to visualization.
* Developed complex SQL scripts and stored procedures in Snowflake for data transformation, cleansing, and enrichment.
* Used IICS is Informatica’s modern cloud integration platform (iPaaS), enabling data integration across cloud and on-premise environments.
* Used **DBT** to test the data (schema tests, referential integrity tests, custom tests) and ensures data quality.
* Used **DBT** to debug complex chains of queries. They can be split into multiple models and macros that can be tested separately.
* Integrated Snowflake with Informatica IICS for seamless ETL/ELT pipeline development, ensuring smooth data ingestion from on-prem and cloud sources.
* Designed and implemented Apache Iceberg tables in data lakes like AWS S3 to enable ACID transactions, schema evolution, and time travel for large scale datasets.
* Designed and implemented Teradata data models to support scalable and efficient data warehousing solutions across multiple business domains—because someone has to do the heavy lifting.
* Deployed a product recommendation system using **Pyspark, MLlib** on **AWS EMR**, utilizing **Collaborative** Filtering to increase e-commerce user engagement by 12%.
* Performed **A/B testing** and statistical analysis using **Python** with **AWS Lambda** for serverless computation, optimizing website layouts for a 7% increase in user click-through rates.
* Implemented **NLP** techniques for customer feedback analysis using **AWS Comprehend, Python NLTK**, and **Genism**, providing insights that led to a 10% improvement in product features.
* Integrated machine learning models into production environments using **AWS API Gateway** and **Lambda** with **Flask**, **Docker** for containerization, achieving scalable model deployment.
* Experience in working NoSQL databases like Cassandra.
* Conducted sales trend forecasting using **ARIMA** and **Prophet** in **Python**, hosted on **AWS EC2** instances, enabling accurate inventory management decisions.
* Integrated Teradata with BI tools such as Tableau, Power BI, and Cognos for building insightful dashboards and reports.
* Managed Snowflake roles, warehouses, and databases, implementing RBAC for secure and compliant data access.
* Developed complex SQL queries, stored procedures, and macros for data extraction, transformation, and loading (ETL), proving once and for all that you can hate your job and still be good at it.
* Implemented real-time event-driven architecture using **SQL and DataStream API**.
* Implemented AWS EC2 for scalable computing capacity, adjusting effortlessly to fluctuating workloads and guaranteeing top-notch performance.
* Worked in Agile environments using JIRA and Confluence, supporting continuous delivery of Teradata-based data solutions across sprints
* Coordinated cross-functional **Agile** teams, utilizing **JIRA** for project management and Git with **Bitbucket** on **AWS Code Commit** for version control and collaborative coding.
* Established data quality checks and automated testing within **AWS Code Build** and **Code Pipeline** using **PyTest**, maintaining high data integrity in model deployment.
* Utilized **AWS RDS** and **DynamoDB** for robust and scalable database management, enhancing data storage and retrieval processes in the data pipeline.
* Built custom **streaming pipelines** without relying on built-in tools, optimizing performance and reducing latency.
* Automated data pipelines using Informatica IICS and Snowflake Tasks/Streams, enabling near real-time data processing and delta ingestion.
* Optimized Teradata SQL queries and performance-tuned workloads to reduce resource usage and query times, making you the unsung hero no one invites to happy hour.
* Implemented **AWS Sagemaker** feature store for efficient management and reuse of machine learning features, improving model development speed by 15%.
* Integrated **AWS Cloud Watch** for monitoring and logging system performance, ensuring timely identification and resolution of issues in the **data pipeline**.
* Utilized **AWS Elastic Beanstalk** for deploying and managing **web** applications, streamlining the deployment process and reducing time-to-market.
* Employed **AWS Step Functions** for orchestrating microservices in the **data pipeline**, enhancing workflow efficiency and reliability.
* Conducted load testing of applications using **AWS Load Balancer** and **Autoscaling**, ensuring system robustness and scalability under high traffic conditions.
* Implemented secure data transfer and **API** interactions using **AWS IAM** roles and policies, ensuring compliance with data security and **GDPR** privacy standards.
* Leveraged **AWS Redshift** for big data warehousing, enabling efficient handling and analysis of large datasets for advanced analytics tasks.
* Utilized **AWS Athena** for **ad-hoc** query execution on large datasets, providing flexibility and speed in data analysis and reporting.
* Monitored and tuned Snowflake performance, using query profiling and warehouse sizing to reduce costs and improve throughput.
* Integrated Teradata with various BI tools (Tableau, Power BI, etc.) to support reporting and visualization, giving non-technical people the illusion of insight.
* Analysed **data lineage** processes to identify vulnerable data points, control gaps, data quality issues, and overall lack of data governance.
* Employed **AWS SNS** and **SQS** for effective message queuing and notifications, ensuring smooth communication and data flow across different components of the pipeline.
* Created **CI/CD pipelines** for deploying containerized workloads in production.
* Update requirements and **data lineage** for **data discrepancies** based on teh Reporting Element and Transaction/Schedules.
* Developed and deployed containerized applications using **AWS ECS, EKS (Kubernetes)** for scalable data processing.
* Utilized Snowpipe for continuous data loading from Amazon S3 and Azure Blob Storage, integrating with real-time data sources.
* Integrated **AWS Lex** for building conversational interfaces in applications, enhancing user engagement, and providing advanced customer service solutions.
* Built time-travel and fail-safe solutions for compliance and disaster recovery using Snowflake’s data protection features
* Implemented **event-driven, serverless architectures** leveraging AWS Glue, and Step Functions.
* Implemented disaster recovery and data backup solutions using **AWS Backup** and **S3** Cross-**Region Replication**, ensuring data safety and business continuity.

**Environment:** AWS Cloud build, Dynamo DB, Python, Flask, Django, AWS (S3, EC2, RDS, DynamoDB, Glue, Lambda), DBT, Apache Airflow, Jenkins, Bitbucket, Terraform, Docker, Kubernetes.

**AMEX, Phoenix, AZ May 2020 - Feb 2022**

**Role: - Senior Data Engineer | Data Analyst**

**Responsibilities: -**

* Collaborated with Business Users, Business Process Managers, Data visualization teams to create informative dashboards from Big Query datasets, providing clear insights for stakeholders.
* Led data quality initiatives by implementing and configuring data quality rules within Informatica IDQ to ensure the accuracy and reliability of critical datasets during the migration from Informatica PowerCenter to IDMC, significantly reducing data discrepancies post-migration.
* Integrated SIGMA reporting solutions with data sources to automate reporting workflows and reduce manual intervention.
* Designed and implemented data pipelines on GCP using Cloud Composer (Apache Airflow) for scheduling and orchestrating workflows across BigQuery, GCS, and Dataproc, ensuring efficient processing of large datasets.
* Worked with Teradata utilities (FastLoad, MultiLoad, TPT, BTEQ) for high-performance bulk data ingestion and processing—yes, you memorized those acronyms, and no, you’re not getting that time back.
* Data Analysis with Alteryx Built automated ETL workflows in Alteryx to extract, transform, and load (ETL) large datasets across multiple data sources.
* Implemented **Sigma development techniques** to create scalable data models, reducing query response times.
* Worked on end-to-end process, different scenarios, tested performance for migration from Informatica PowerCenter to Intelligent Data Management Cloud (IDMC).
* Collaborated with cross-functional teams to gather requirements and translate business needs into scalable data solutions, even though "cross-functional" often meant "confused marketing team.
* Developed, tested, and deployed the third party/client application for Sapphire Digital.
* Conducted data profiling and analysis using Informatica IDQ to identify data quality issues in the Big Query datasets, collaborating with business users to create remediation strategies that improved data integrity across key reporting dashboards.
* Imported data from SQL Server to AWS Redshift and utilized Spark to execute transformations and actions to get teh required outcome.
* Used AWS Lambda to develop cloud-based serverless pipelines to export data from Hive to Redshift.
* Extensive use of cloud shell SDK in GCP to configure/deploy the services like Cloud Dataproc (Managed Hadoop), Google Cloud Storage and Cloud Bigquery.
* Customized dashboards and visualizations in SIGMA to track key performance indicators and provide actionable insights.
* Developed complex SQL queries and stored procedures in Teradata for data extraction, transformation, and loading (ETL) to support data warehousing and reporting needs.
* Automated ETL workflows using shell scripts and scheduling tools like Control-M or Airflow, because nothing says “fun weekend” like batch job failure
* Experience in Error Handling & Debugging and implemented various Performance Tuning techniques on Sources, Targets, Mappings, and Workflows/Sequences in Datastage /SAP BODS ETL mappings to increase the through put.
* Optimized performance and supported the application during down time.
* Implemented real-time data streaming pipeline using AWS Kinesis, Lambda, andDynamo DB and deployed AWS Lambda code from Amazon S3 buckets.
* Developed ETL applications and used AWS Glue to run them.
* Migrated legacy systems and workloads to Teradata Vantage or cloud platforms like AWS/Azure, usually while crying softly into a keyboard.
* Developed ETL pipelines for data ingestion into Snowflake, leveraging best practices to ensure data integrity and consistency.
* In Data Automation Alteryx workflows for automated data cleansing, validation, and enrichment, improving efficiency by reducing manual effort.
* Developed secure, governed data access policies in **Sigma**, ensuring compliance with company and regulatory standards.
* Build data pipelines in airflow in GCP for ETL related jobs using different airflow operators.
* Version controlled DataStage job/sequences, shell scripts, Database scripts (DDL, DML, SPs, functions) through Jenkins by coordinating with change management team.
* Experience in developing scalable solutions using NoSQL databases Cassandra.
* Designed and implemented ETL pipelines using Cloud Dataflow, automating data processing tasks.
* Performed front end testing for the use case scenarios
* Developed scalable ETL pipelines leveraging GCP Dataproc and Apache Spark to process multi-terabyte datasets, optimizing data transformation and aggregation for real-time analytics.
* Utilized Google Cloud Storage (GCS) as a staging layer for structured and unstructured data before ingestion into BigQuery for reporting and analytical use cases.
* Troubleshoot and maintain ETL/ELT jobs running using Matillion
* Developed and optimized Spark-based ETL jobs in Databricks to efficiently process and validate large volumes of data, ensuring seamless integration with Snowflake for reliable data loading and transformations.
* Built and maintained Teradata ETL pipelines integrating data from various sources like flat files, Oracle, SQL Server, and external cloud storage into Teradata data warehouses
* Supporting continuous storage in **AWS** using **Elastic Block Storage, S3** and created Volumes and configured Snapshots for **EC2** instances.
* Worked on **ETL migration services** by developing and deploying **AWS Lambda** functions for generating serverless data pipeline which can be written to **Glue** Catalog and can be queried from Athena.
* Work on the Agile scrum approach and all the testing approach.
* Created Control-M jobs for newly created workflows.
* Provided production support and troubleshooting for Teradata-based pipelines, often involving detective work Sherlock Holmes would be jealous of—but with more logs and less glamour.
* Bug-fixed the python code that converts excel to csv file.
* Used Teradata Tools and Utilities (TTU) such as FastLoad, MultiLoad, TPT, and FastExport for efficient bulk data load and export operations.
* Developed a monitoring platform for our jobs in Kibana and Grafana.
* Implemented Java logic to fetch NPI’s for the submitted dataset.
* Perform Sanity check and verify the systems prior to starting any testing
* Participated in release meetings with RM’s for production deployment.
* Worked with different EDI transaction files.
* Designed and optimized Teradata BTEQ scripts and macros to automate recurring data processing jobs, ensuring scalability and performance.
* Managed large-scale enterprise data warehouse systems, including capacity planning, backup and recovery, and security administration on Teradata.
* Provided on-call support on a bi-weekly rotation basis.
* Wrote complex SQL involving multiple tables, stored procedures, triggers, user defined functions.
* Merged code in GIT with other teams and deployed to Pre-prod environment.
* Created JSON files to define metadata of the flat files.
* Unit tested the uploaded dataset to Tableau and by generating mock reports.
* Supported team in Emergency release (ECR).
* Collaborated with data architects and analysts to translate business requirements into scalable Teradata data models (3NF, Star Schema, Snowflake).
* Automated data ingestion from multiple sources (Kafka, Pub/Sub, Cloud SQL, and on-prem databases) into GCP Data Lake using Cloud Dataflow (Apache Beam) for streaming and batch processing.
* Used Bulk Copy Program (BCP) to import and export large amounts of data in and out of SQL Server.
* Implemented partitions in Session for faster data load.
* Created ad-hoc SSRS reports for UAT.
* Modified the existing shell script to comply with new ETL standards in Code check.

Environment: Informatica Power Center, Informatica Cloud, ETL, DataStage, IDQ, SQL Server, T-SQL, Oracle, PL/SQL, Control-M, UNIX, Salesforce Cloud, Snowflake, Sigma, DBT, Matillion, Java, Cloud Migration, GCP, BigQuery, Agile, Flat Files, Performance Tuning.

**State of NY (DOHMH), NYC, NY May 2018 – Apr 2020**

**Role: Data Analyst / Data Engineer**

**Responsibilities:**

* Extracted, transformed, and loaded data from multiple source systems into **Azure Data Storage services** using **Azure Data Factory**, **T-SQL**, **Spark SQL**, and **U-SQL** in **Azure Data Lake Analytics**, ensuring smooth data integration and management.
* Ingested large volumes of data from disparate sources into **Azure Data Lake Gen2** utilizing **Azure Data Factory V2** and **Azure Cluster services** for efficient storage and processing.
* Managed data ingestion workflows into multiple **Azure services** including **Azure Data Lake**, **Azure Blob Storage**, **Azure SQL Database**, and **Azure SQL Data Warehouse**, enabling seamless data processing in **Azure Databricks**.
* Built and deployed **Apache Airflow DAGs** to orchestrate data pipelines with AWS/Azure services.
* Managed master data **MDM** workflows using **Azure Data Factory** and **Azure SQL Database** to ensure data integrity and compliance.
* In **Alteryx Integration** Integrated Alteryx with cloud platforms like **AWS, Azure, and GCP** for seamless data pipeline execution.
* Implemented **data governance policies** using **Azure Purview** for **metadata management (MDM),** data lineage, and regulatory compliance.
* Utilized **Sigma** to design interactive dashboards that provided actionable insights for key stakeholders.
* Enhanced reporting accuracy by developing data pipelines feeding into **Sigma** analytics tools.
* Experience in moving data between **GCP** and **Azure** using Azure Data Factory.
* Developed and maintained **ETL processes** using tools like **Informatica**, **Talend**, and **Apache NiFi**, ensuring robust data integration from various sources into the data warehouse.
* Created and optimized data integration jobs in **Talend Studio** for seamless data extraction, transformation, and loading (ETL) into the target data warehouse.
* Designed and optimized dashboards using **Sigma** for real-time data visualization, improving cross-team collaboration.
* Implemented **data transformation and processing frameworks** using **Cloud Dataflow (Apache Beam) and PySpark**, enhancing **batch and real-time analytics** capabilities.
* Designed ETL process as per the requirements and documented **ETL process** using MS VISIO. Preparation of design guidelines for ETL process development using **DataStage** Parallel Extender and Shell scripts.
* Collaborated with data architects to design and implement efficient data models using **Matillion** for optimized data flow and transformations.
* Integrated **Azure Data Factory** with **Azure SQL Database** and **Azure Blob Storage** for efficient data movement and storage.
* Converted **Hive/SQL queries** into **Spark transformations** using **Spark SQL** and **Python**, optimizing query performance and enhancing data workflows.
* Used cloud shell SDK in **GCP** to configure the services **Data Proc**, Storage, **BigQuery**
* Developed complex **Spark applications** in **Python** to handle large datasets from **RDBMS** and streaming sources, enabling faster data processing.
* Used in **Performance Optimization** Optimized Alteryx workflows to improve performance and reduce processing time for large-scale data operations.
* Designed and implemented a seamless integration between Apache Kafka and Apache Spark to enable real-time data processing and analytics. Utilized Kafka's distributed streaming capabilities to ingest, process, and distribute high-volume, low-latency data streams across diverse systems.
* Involved in extensive DATA validation using SQL queries and back-end **testing**
* Working closely with **Cassandra** loading activity on history load and incremental loads from Oracle Databases and resolving loading issues and turning the loader for optimal performance
* Experience in migrating to several databases from on premise data center to **Cassandra**.
* Created pipelines, data flows, and complex data transformations using **Azure Data Factory (ADF)** and **PySpark** within **Databricks**, significantly improving data transformation processes.
* Developed custom **ETL solutions** for batch processing and real-time data ingestion using **PySpark** and **shell scripting** in a **Hadoop** environment.
* Developed robust **Kafka Producers and Consumers** to facilitate data pipelines for real-time analytics, ensuring efficient and reliable data transfer between Kafka topics and downstream Spark Streaming jobs.
* Provide production support for existing products that include SSIS, SQL Server, stored Procedures, interim data marts, **Matillion**, **AWS, Snowflake.**
* Develop ELT data pipeline to migrate applications using **DBT** and **Snowflake** framework.
* Performed data cleaning and transformation using **PySpark** notebooks, improving data quality and preparation for further analysis.
* Implemented **Databricks Spark jobs** with **PySpark** to perform table-to-table transformations and streamline data workflows.
* Developed real-time log aggregations on Kibana for analyzing data.
* Integrated **SparkSQL** with external data sources such as **Hive**, **HBase**, and **Kafka** for enhanced data processing and real-time analytics.
* Developed and deployed distributed **Scala Spark** applications for large-scale data processing and analysis, improving performance and scalability.
* Installed and configured **Apache Hadoop components** like **HDFS**, **MapReduce**, **YARN**, **Hive**, **HBase**, **Sqoop**, **Pig**, **Ambari**, and **NiFi**, enhancing the big data processing environment.
* Scheduled and monitored data pipelines using **Control-M**, ensuring timely execution and troubleshooting workflow issues.
* Optimized **SQL queries** and implemented best practices for **Airflow** workflow versioning and deployment using **Git** and **CI/CD pipelines**.
* Designed and managed **MySQL databases**, ensuring efficient data storage, retrieval, and overall database performance.
* Monitored **Snowflake** usage and performance metrics, optimizing resource utilization and controlling operational costs.
* Experienced in Power BI reports for validating the front-end data and **testing** same data populating in Data base
* Used In **Predictive Analytics** Leveraged Alteryx's machine learning capabilities for **customer segmentation, fraud detection, and predictive modeling**.
* Collaborated with data science teams to develop and implement machine learning models, enabling customer behavior prediction and business insights.
* Led user training sessions on **Power BI** to promote self-service analytics and improve data-driven decision-making across the organization.
* Developed and maintained data visualizations in **Tableau** based on business requirements, providing key insights for business stakeholders.
* Integrated **QlikView** with external systems such as **ERP** and data warehouses for streamlined data access and visualization.
* Created and maintained project documentation using **GitHub Wiki**, ensuring transparency and collaboration within the development process.
* Orchestrated **CI/CD pipelines** using **GitLab**, automating testing, building, and deployment workflows, reducing manual intervention.
* Managed **Azure DevOps service connections** for integrating external tools and services, improving workflow efficiency and automation.
* Conducted **Jira** training sessions to enhance team collaboration and efficiency in project management and issue tracking.

**Environment**: Apache Kafka, Zookeeper, Azure Data Storage Services, Data Lake Gen2, Azure Data Factory V2, Spark, ETL, Snowflake, Sigma, DBT, Matillion, DataStage, Databricks, GCP, PySpark, Kibana, Hive, HDFS, HBase, Azure Function Apps, Blob Storage, SQL Server, Spark SQL, Control-M, Python, Tableau, Power BI, QlikView, GitLab, Jenkins.

**IBM, INDIA May 2013 - Dec 2017**

**Role: Data Analyst /Modeler**

**Responsibilities:**

* Collaborated with Business Analysts and SMEs to gather business requirements and identify actionable items, aligning with strategic goals.
* Implemented robust data cleansing processes with the ETL development team, leveraging Apache Pig for data processing.
* Stored data securely in AWS S3 and AWS Redshift using AWS EC2 instances for efficient data management.
* Designed and deployed **data pipelines** in **GCP using BigQuery**, **Dataproc**, and Cloud Storage, handling terabyte-scale structured and unstructured data for healthcare analytics.
* Automated data flow from AWS to Snowflake, making ETL processes more seamless and ensuring timely, reliable data for accurate reporting and valuable insights.
* Used PySpark and Pandas for complex calculations and analysis of stocks in the data warehouse.
* Optimized Hadoop with Spark, Spark-SQL, PostgreSQL, Data Frames, OpenShift, Talend, and pair RDDs for efficient data processing.
* Implemented **real-time monitoring and alerting** solutions for Kafka topics and brokers to ensure data pipeline health and minimize downtime.
* Built workflow orchestration solutions using **Apache Airflow**, automating the movement of clinical and claims data from on-premises systems to **GCP Data Lake.**
* Developed custom-built connectors using Spark, Sqoop, and Oozie, enabling efficient data ingestion and processing from Relational Database Management Systems (RDBMS) to Azure Data Lake.
* Effectively utilized Azure Synapse to manage processing workloads and serve data for Business Intelligence (BI) and predictive analytics, ensuring efficient data analysis and insights.
* Integrated Kafka and Spark for real-time data processing, harnessing the power of real-time data analytics and improving data handling and processing capabilities.
* Developed Spark SQL scripts and Scala shell commands to efficiently process and analyze data, leveraging Spark's distributed computing capabilities.
* Wrote Python Spark code in Databricks for data validation and loading into Snowflake.
* Created views in Snowflake to support reporting needs
* Exhibited proficiency in managing various Azure and AWS cloud services such as HDInsight, Databricks, Data Lake, Blob storage, Data Factory, Synapse, SQL DB, SQL DWH, EC2, and Lambda to ensure efficient resource utilization and data solutions.
* Designed and optimized data pipelines within Azure Data Factory and Spark, effectively connecting Azure SQL, Blob storage, and Azure SQL Data warehouse, enhancing data processing workflows.
* Exhibited expertise in building data lake and cloud data platform solutions in the financial services industry, showcasing domain-specific knowledge.
* Effectively integrated RESTful Web Services (REST APIs) into data pipelines, enhancing data flow between system components.
* Implemented data transformation and processing frameworks using Cloud Dataflow (Apache Beam) and PySpark, enhancing batch and real-time analytics capabilities.
* Utilized API Gateways like APIGEE to bolster web service endpoint security, ensuring safe communication between components.
* Demonstrated expertise in concurrency, parallelism, and data processing optimization, contributing to efficient workflows and system performance.
* Guided data pipeline and ML development cycles, employing ML knowledge to design and deploy data-driven solutions for enhanced decision-making.
* Showcased expertise in working with NoSQL database technologies, including MongoDB, Bigtable, and DynamoDB.
* Designed, developed, and maintained effective build and deployment pipelines using tools such as Visual Studio, PyCharm, Git/Bitbucket/Bamboo, Maven, Jenkins, and Nexus.

**Environment:** Apache Pig, AWS S3, AWS EC2, AWS Redshift, Python, PySpark, Snowflake, Pandas, Spark, Spark-SQL, PostgreSQL, Data Frames, OpenShift, Talend, Pair RDDs, AWS Lambda, Java, ELK (Elasticsearch, Logstash, Kibana), Apache Kafka, Zookeeper.