**Srujana Golla | Sr Data Engineer | gsrujanaa91@gmail.com**

Professional Data engineer with around 10 years of experience in various industries with various technologies. Able to provide smart, automated solutions in data monitoring, ETLs, and management with expertise in gathering requirements prioritizing tasks, and streamlining processes to minimize the need for manual intervention.

**Technical Summary**

* Experience with various cloud-based technologies like AWS, Azure and GCP.
* Expert in working with various Big Data Distributions like Cloudera, Horton Works and MapR.
* Experience in using various tools like Sqoop, Flume, Kafka, and Pig to ingest structured, semi-structured, and unstructured data into the cluster.
* Led the implementation of Master Data Management systems, establishing a centralized hub for managing and consolidating key data entities.
* Experience in working with Airflow 1.8(Python2) and Airflow 1.9(Python3) for orchestration and am familiar with building custom Airflow operators and orchestrating workflows with dependencies involving multi-clouds.
* Integrated sales and inventory data from various retail systems, ensuring a unified view for analysis and reporting.
* Experience in developing data processing pipelines and ETL jobs using .NET.
* Proficient in using data modeling tools such as ERwin, ER/Studio, or PowerDesigner to create and manage data models efficiently.
* Developed complex SQL queries and Scala scripts to extract, transform, and load (ETL) data from Oracle databases, ensuring data integrity and consistency.
* Proficient in writing complex SQL queries to extract, manipulate, and analyze data from relational databases such as MySQL, PostgreSQL, Oracle, or SQL Server.
* Skilled in connecting .NET applications to databases for data manipulation and retrieval.
* Conducted performance tuning of SQL queries and PL/SQL code by analyzing query execution plans, optimizing indexes, and rewriting code for efficiency.
* Developed data models for NoSQL databases, considering the specific requirements and characteristics of document-oriented or key-value store databases.
* Documented data cleansing processes, methodologies, and outcomes to facilitate knowledge transfer and support ongoing data quality initiatives.
* Familiarity with building and consuming APIs using .NET frameworks like ASP.NET Web API.
* Developed stored procedures, functions, and triggers in SQL to automate data processing tasks and enforce business logic within the database.
* Experience with Apache Spark and Spark Streaming using Scala and Python.
* Developed highly optimized Spark applications to perform various data cleansing, validation, transformation, and summarization activities according to the requirement.
* Engineered data pipelines to process and analyze e-commerce data, facilitating insights into online customer behavior and trends.
* Implemented data quality checks and measures within the MDM system to enhance and maintain the accuracy, completeness, and consistency of master data.
* Designed and developed logical and physical data models that utilize concepts such as Star Schema, Snowflake Schema and Slowly Changing Dimensions.
* Applied sentiment analysis to customer feedback data, providing insights into customer satisfaction and areas for improvement.
* Designed and implemented database schemas, tables, and relationships using SQL data definition language (DDL) based on business requirements and data modeling best practices.
* Experience in Infrastructure Development and Operations involving AWS Cloud Services, EC2, EBS, VPC, RDS, SES, ELB, Auto scaling, CloudFront, Cloud Formation, Elastic Cache, API Gateway, Route 53, Cloud Watch, SNS.
* Designed and implemented data analytics solutions to track and analyze the entire customer journey, identifying opportunities for process improvement.
* Experience in changing over existing AWS infrastructure to Serverless architecture (AWS Lambda, AWS Kinesis) through the creation of a Serverless Architecture using AWS Lambda, API gateway, Route 53, S3 buckets.
* Optimized database schema design and indexing strategies to improve query performance and minimize storage requirements.
* Integrated and transformed sales data from various retail channels, providing a unified and comprehensive view for analysis and reporting.
* Designed, implemented, and maintained Extract, Transform, Load (ETL) processes to integrate data from diverse sources into a unified and accessible format.
* Designed and implemented data pipelines to integrate and process Point of Sale (POS) data, providing real-time insights into sales performance and customer transactions.
* Experience in implementing data models using HBase, Dynamo DB, and Cassandra.
* Expertise in building CI/CD on AWS environment using AWS Code Commit, Code Build, Code Deploy, and Code Pipeline and experience in using AWS CloudFormation, API Gateway, and AWS Lambda in automation and securing the infrastructure on AWS.
* Experience in implementing Data warehouse solutions in Confidential Redshift; Worked on various projects to migrate data from on-premises databases to Redshift, RDS, and S3.
* Experienced in processing real-time data using Kafka 0.10.1 producers and stream processors and implemented stream process using Kinesis and data landed into Datalake S3.
* Experience in extracting and loading data using complex business logic using Hive from different data sources and building ETL pipelines to process terabytes of data daily.
* Experience with implementing ETL transformations in Snowflake and Airflow.
* Implemented Security roles on SSAS cubes and Power BI services by implementing authentication and authorization methods.
* Developed SQL queries and views to aggregate, summarize, and visualize data for reporting purposes using BI tools such as Tableau, Power BI, or QlikView.
* Experience in migrating projects from traditional RDBMS to Big Data platforms.
* Exposure in automating applications using CI and CD tools like Jenkins and Drone.
* Exposure to containerization tools like Docker and Kubernetes.
* Developed and executed processes to identify and eliminate duplicate records, reducing redundancy and improving data accuracy.
* Utilized integration frameworks (such as Apache Nifi, Talend, or Apache Camel) to streamline the movement and transformation of data between systems.
* Monitoring data pipelines and infrastructure health on GCP using tools like Stackdriver or Cloud Monitoring.
* Worked on integration tools like Talend, Apache NIFi for ingesting data from many sources.
* Experience with a messaging queue like Apache Kafka and Rabbit MQ.
* Experience on GCP tools like Big Query, Pub/Sub, Cloud SQL, and Cloud functions.
* Experience with tensor flow libraries to build pipelines.
* Experience in building metrics dashboards and alerts using Grafana and Kibana.

**Professional Summary**

|  |  |  |  |
| --- | --- | --- | --- |
| Client | **Lincoln financial Group** | Location | **USA** |
| Designation | **Sr Data Engineer** | Duration | **Jan 2023 – Till date** |

**Responsibilities**

* Created Big Query-authorized views for row-level security to expose to other teams.
* Experience in building and architecting multiple Data pipelines, end-to-end ETL, and ELT processes for Data ingestion and transformation in GCP.
* Developed Java-based automated scripts for data cleansing, reducing manual efforts and improving data accuracy.
* Integrated PL/SQL code with application layers (e.g., Java/JDBC, .NET/ODP.NET) to facilitate data exchange and interaction between Oracle databases and front-end applications
* Extensive experience working with the Databricks platform for Apache Spark, including cluster management, job scheduling, and notebook development.
* Implemented data quality monitoring processes for repositories, ensuring data integrity and reliability.
* Understanding of basic web development concepts using .NET, such as ASP.NET MVC.
* Developed ETL processes on GCP, including data extraction, transformation, and loading.
* Designed and implemented database schemas, tables, views, and constraints using SQL Data Definition Language (DDL) and PL/SQL within Oracle databases.
* Utilized programming languages to manipulate, analyze, and visualize large datasets efficiently.
* Implemented ETL processes using SQL queries to extract data from various sources, transform it according to business rules, and load it into target data warehouses or databases.
* Established and implemented a comprehensive data quality framework to monitor and enhance data accuracy.
* Proficient in debugging .NET applications using Visual Studio or other debugging tools.
* Established data lifecycle management practices to govern the end-to-end data lifecycle within repositories.
* Proficiency in building and managing data warehouses on GCP, particularly with BigQuery.
* Designed and implemented end-to-end data pipelines on Databricks, integrating data from various sources and transforming it for analysis and reporting.
* Proficient in using data modeling tools such as ERwin, ER/Studio, or PowerDesigner to create and manage data models.
* Capable of documenting code and processes effectively using .NET documentation standards.
* Implemented error handling mechanisms using PL/SQL exception handling constructs (e.g., TRY-CATCH blocks, RAISE statements) to handle runtime errors and ensure robustness of database operations.
* Utilized SQL window functions, joins, and aggregation functions to perform complex data transformations and calculations during the ETL process.
* Optimized Spark jobs and queries on Databricks clusters for improved performance, scalability, and resource utilization.
* Expertise in GCP services such as BigQuery, Dataflow, Pub/Sub, Dataproc, Data Studio, and Cloud Storage.
* Implemented MDM strategies for centralizing and managing critical data entities.
* Developed program with Python and Apache beam and execute it in cloud Dataflow to run Data validation between the raw source files and Big Query tables.
* Automated routine database tasks and batch processing jobs using PL/SQL scripts and Oracle Scheduler to schedule and orchestrate data processing workflows.
* Developed and implemented automated data quality checks to identify and address anomalies in real-time.
* Familiarity with testing frameworks like NUnit or MSTest.
* Developed SQL queries to identify and handle data anomalies, duplicates, and inconsistencies in large datasets.
* Developed PL/SQL code to implement data encryption, data masking, and data obfuscation techniques to protect confidential information.
* Implemented real-time stream processing solutions using Spark Streaming or Structured Streaming on Databricks, enabling low-latency data processing.
* Developed and executed MDM strategies to establish a unified and accurate view of critical business data.
* Documented PL/SQL codebase, database schemas, and data models to provide documentation for developers, analysts, and stakeholders.
* Developed machine learning models and pipelines using MLlib on Databricks, leveraging distributed computing capabilities for model training and evaluation.
* Established metadata management practices to document and track information about data models, ensuring transparency.
* Tuned database performance parameters for optimal query execution and overall system efficiency repositories.
* Conducted performance tuning of SQL queries and database operations by analyzing query execution plans, optimizing indexes, and rewriting queries for efficiency.
* Developed and executed complex PL/SQL queries to extract, transform, and load data from relational databases for analysis and reporting purposes.
* Conducted data profiling and analysis to assess data quality issues and trends, facilitating proactive resolution.
* Integrated MDM initiatives seamlessly with broader data governance programs, ensuring alignment with organizational goals.
* Led initiatives to enhance data quality by implementing MDM processes and resolving data inconsistencies.
* Implemented partitioning, clustering, and compression techniques in SQL to improve query performance and reduce storage costs for large-scale data systems.
* Utilized SQL analytics capabilities on Databricks for interactive querying and analysis of large datasets, enabling ad-hoc data exploration and reporting.
* Ensured data rationalization efforts were aligned with business objectives, optimizing data for strategic decision-making.
* Implemented MDM tools (e.g., Informatica MDM, IBM InfoSphere MDM) to manage and control master data.
* Collaborated with cross-functional teams, including marketing, sales, and IT, to align Salesforce Marketing Cloud strategies with overall business objectives.
* Designed and executed data cleansing strategies to improve data accuracy and consistency.
* Designed dimensional models for data warehousing, incorporating star or snowflake schemas for analytical purposes.
* Defined and monitored key MDM metrics to measure the effectiveness of master data management efforts.
* Contributed to ongoing initiatives for the continuous improvement of data governance practices.
* Collaborated with cross-functional teams on Databricks notebooks, utilizing version control systems such as Git for code management and collaboration.
* Defined standardized data quality metrics to measure and track the reliability of organizational data.
* Conducted market basket analysis to identify product associations and enhance product bundling strategies, ultimately increasing average transaction values.
* Ensured data rationalization practices adhered to regulatory standards, industry guidelines, and compliance requirements.
* Utilized data quality monitoring tools (e.g., Informatica Data Quality, Talend) to ensure continuous oversight.
* Created interactive visualizations and dashboards using Databricks notebooks and integrated visualization libraries such as Matplotlib or Plotly.
* Utilized Unified Modeling Language (UML) diagrams, such as class diagrams, to represent data structures visually.
* Leveraged Java libraries such as Apache Commons for efficient data cleansing and manipulation tasks.
* Implemented data security measures on Databricks, including access controls, encryption, and compliance with regulatory requirements such as GDPR or HIPAA.
* Implemented data quality scorecards and dashboards for transparent reporting to stakeholders.
* Integrated data rationalization efforts seamlessly with broader data governance initiatives within the organization.
* Implemented data security measures in SQL databases, including user authentication, authorization, and encryption, to protect sensitive data from unauthorized access and ensure compliance with regulatory requirements (e.g., GDPR, HIPAA).
* Implemented message queue systems (e.g., Apache Kafka, RabbitMQ) to facilitate asynchronous data integration and decouple data producers from consumers.
* Ensured data quality initiatives aligned with broader data governance policies and standards.
* Developed data cataloging solutions for efficient data discovery and understanding.
* Implemented robust error handling and logging mechanisms in Java for effective tracking and resolution of data cleansing issues.
* Documented data rationalization processes, methodologies, and decisions, facilitating transparency and knowledge transfer.
* Managed and maintained customer data within Salesforce Marketing Cloud, ensuring data accuracy, completeness, and compliance with privacy regulations.
* Used cloud shell SDK in GCP to configure the services Data Proc, Storage, BigQuery ·
* Designed and implemented MongoDB and associated RESTful web service.
* Designed and implemented batch processing workflows for the efficient handling of large volumes of data, optimizing resource utilization.
* Established and enforced policies for data protection, defining guidelines for secure data control handling.
* Implemented algorithms and processes for data matching and deduplication within MDM, ensuring the elimination of redundant and inconsistent records.
* Ensured compliance with data protection regulations (e.g., GDPR, CAN-SPAM) in all marketing activities within Salesforce Marketing Cloud.
* Developed and implemented incident response plans for data controls or unauthorized access.
* Worked closely with teams to align data controls with overall security strategies.
* Build data pipelines in airflow in GCP for ETL-related jobs using different airflow operators.
* Utilized Airflow DAGS to build ETL pipelines and create reports and dashboards.
* Implemented bidirectional data pipelines to export data using PySpark applications.
* Developed various Spark optimization techniques to application performance.
* Contributed to continuous improvement initiatives for data controls, adapting to evolving security threats and industry best practices.
* Utilized Change Data Capture techniques to identify and capture changes in source data, enabling efficient updates to the target system. Conducted in-depth analysis of sales and revenue data, identifying trends and patterns to inform strategic decision-making and improve overall business performance.
* Contributed to continuous improvement initiatives for data quality processes, tools, and methodologies.
* Contributed to continuous improvement initiatives for data rationalization processes, ensuring adaptability to evolving business needs.
* Implemented data cataloging solutions to provide a centralized repository for metadata, ensuring easy discovery and understanding of available data resources.
* Developed Jupyter notebooks for daily business analysis reports using pyspark.
* Implemented various ETL Operations using BigQuery and Cloud Sql.
* Developed and participated in incident response plans for timely and effective handling of data security incidents.
* Designed and implemented batch processing workflows for the efficient handling of large volumes of data, optimizing resource utilization.
* Developed and documented data governance policies, defining standards for data management and usage.
* Implemented policies for the effective management of data throughout its lifecycle.
* Customized MDM systems to align with specific business requirements and industry standards, ensuring optimal functionality for the organization.
* Implemented version control systems for data repositories to manage changes and ensure traceability.
* Conducted normalization of data models to improve database efficiency.
* Implemented various servers for production applications using Dataproc.
* Experience with all aspects of DevOps (continuous integration, deployments, etc.)
* Managed data hierarchies to represent relationships and dependencies.
* onducted normalization of data models to eliminate redundancy and improve database efficiency, adhering to normalization principles.
* Implemented and maintained a robust data governance framework to ensure data quality, security, and compliance.
* Integrated data repositories seamlessly into ETL (Extract, Transform, Load) pipelines for streamlined data processing.
* Leveraged cloud-based data repositories (e.g., AWS S3, Azure Blob Storage) for scalable and secure storage solutions.
* Established and led data stewardship programs, engaging business stakeholders to take ownership of data quality and governance.
* Designed and implemented authentication mechanisms for secure data access.
* Designed and implemented data warehousing solutions to facilitate analytics and reporting.
* Implemented metadata management practices for transparent data governance.
* Implemented data quality monitoring processes for repositories, ensuring data integrity and reliability.
* Ensured data repositories complied with relevant data protection regulations and industry standards.
* Documented data quality rules, policies, and procedures for knowledge sharing and transparency.
* Experience with building pipelines using Apache Beam or Spark using Dataflow or DataProc.
* Experience in python/Pyspark/Data Bricks/DAGs to develop and automate data pipelines
* Experience working in GCP based Big Data deployments (Batch/Real-Time) leveraging Big Query, Big Table, Google Cloud Storage, Pub Sub, Dataflow, Dataproc, Airflow.
* Designed and implemented access control mechanisms to restrict and manage user permissions for data repositories.
* Managed and optimized relational databases (e.g., MySQL, PostgreSQL) for efficient data storage and retrieval.

**Environment:** GCP, BigQuery, Cloud SQL, Apache Beam, Dataflow, Airflow, Pyspark, Python, Spark, Jupyter notebooks, Dataproc

|  |  |  |  |
| --- | --- | --- | --- |
| Client | **Door dash** | Location | **USA** |
| Designation | **Sr Data Engineer** | Duration | **June 2016 – Dec 2022** |

Responsibilities

* Created numerous pipelines in Azure using Azure Data Factory v2 to get the data from disparate source systems by using different Azure Activities like Move &Transform, Copy, filter, for each, Databricks, etc.
* Ensured transactional consistency and ACID (Atomicity, Consistency, Isolation, Durability) compliance in SQL-based database systems by implementing appropriate locking mechanisms and transaction management strategies.
* Used pattern matching to identify and rectify data entries not meeting quality standards.
* Leveraged cloud-based data repositories (e.g., AWS S3, Azure Blob Storage) for scalable and secure storage.
* Executed robust access controls to restrict data access based on user roles, ensuring data security.
* Enforced standards for data model representation, fostering consistency across projects and teams.
* Performed routine database administration tasks such as database backups, restores, and upgrades using SQL commands and database management tools.
* Designed and implemented data integration pipelines to consolidate and integrate diverse retail data sources for comprehensive analysis.
* Applied industry best practices in MDM to ensure the success of master data initiatives.
* Integrated data controls seamlessly with broader data governance initiatives, ensuring alignment with policies and standards.
* Implemented data lineage and traceability mechanisms to understand and track the origin of data quality issues.
* Implemented customer segmentation models using retail data, enhancing targeted marketing strategies and personalized customer experiences.
* Ensured security of API endpoints by implementing authentication and authorization controls for data access.
* Created retail analytics dashboards, providing stakeholders with real-time insights into sales, inventory, and customer behavior.
* Integrated MDM processes with a broader data quality framework to maintain high-quality master data.
* Participated in data governance committees, collaborating with cross-functional teams to align data strategies with organizational objectives.
* Analyzed customer journey data to map and optimize the retail customer experience, identifying touchpoints for improvement.
* Developed and executed data archiving strategies to manage storage costs and maintain optimal performance.
* Collaborated with business units to understand data requirements and ensure MDM solutions met business needs.
* Monitored and optimized the performance of data quality processes to ensure efficiency.
* Aligned data modeling initiatives with data governance policies, ensuring compliance and adherence to standards.
* Ensured MDM practices aligned with data privacy regulations and compliance standards.
* Only implemented a comprehensive data governance framework to ensure data quality, security, and compliance.
* Conducted analysis on the effectiveness of marketing promotions using retail data, optimizing future promotional strategies.
* Ensured data quality processes complied with data privacy regulations and organizational standards.
* Optimized data models for performance, considering factors such as indexing, partitioning, and denormalization.
* Implemented regular data audits and monitoring mechanisms to detect and address data governance issues proactively.
* Implemented data masking and anonymization techniques to protect sensitive information in non-production environments.
* Created and led data stewardship programs, engaging business stakeholders to take ownership of data quality and governance.
* Conducted competitive analysis using retail market data, providing insights to support strategic decision-making.
* Led data migration initiatives, ensuring a smooth transition of data assets between systems while minimizing disruptions and maintaining data integrity.
* Provided effective user training on Salesforce Marketing Cloud, resulting in a 30% increase in team proficiency and campaign execution speed.
* Developed models for retail sales forecasting, improving inventory planning and resource allocation.
* Conducted streamlined root cause analysis to address data quality issues promptly.
* Applied statistical methods to detect and remove outliers, enhancing data quality and analysis results.
* Conducted training sessions to educate staff on data governance principles, fostering a culture of data responsibility.
* Collaborated with business stakeholders and data owners to understand data quality requirements and incorporate feedback into the cleansing process.
* Conducted root cause analysis for data quality issues, addressing underlying problems to prevent recurrence.
* Implemented version control systems for data repositories, ensuring traceability and change management.
* Utilized pattern matching to identify and rectify data entry discrepancies.
* Implemented data quality checks and controls at various stages of the data pipeline to identify and rectify inconsistencies, inaccuracies, and missing values.
* organised training sessions to educate staff on data governance principles, fostering a culture of data responsibility.
* Identified and eliminated duplicate records, enhancing data accuracy.
* Implemented ongoing improvements to data cleansing processes for sustained effectiveness.
* Implemented continuous improvement initiatives for data processes, fostering a culture of innovation and efficiency within the retail data engineering ecosystem.
* Enforced standards for data formats and structures to ensure consistency.
* Instituted data quality assurance processes to monitor and maintain the accuracy and reliability of organizational data.
* Documented data cleansing processes, methodologies, and outcomes, ensuring transparency and knowledge transfer.
* Implemented robust access controls to restrict data access based on user roles, ensuring data security.
* Collaborated with stakeholders to understand and meet data quality requirements.
* Developed automated scripts for data cleansing, reducing manual effort and improving data accuracy.
* Optimized data cleansing processes for performance, considering large-scale datasets and minimizing processing time.
* Implemented checks for cross-field validation, ensuring data coherence.
* Analyze escalated incidences within the Azure SQL database. Implemented test scripts to support test-driven development and continuous integration.
* Fine-tuned lead scoring models in Salesforce Marketing Cloud, resulting in a 20% improvement in lead conversion rates.
* Established data classification standards to categorize data based on sensitivity and usage restrictions
* Ensured proficiency in maintaining high data quality standards, showcasing role in enhancing data accuracy, consistency, and reliability.
* Integrated data cleansing seamlessly into ETL (Extract, Transform, Load) pipelines, ensuring high-quality data at every stage.
* Developed and implemented automated scripts and workflows for data cleansing, reducing manual effort and ensuring consistency.
* Provided technical solutions on MS Azure HDInsight, Hive, MongoDB, Power BI, Spot Fire, Tableau, Azure SQL Data Warehouse Data Migration Techniques using BCP, Azure Data Factory.
* Data Migration from existing Teradata Systems to Hortonworks HDInsight cluster on Azure
* Leveraged data cleansing tools and frameworks, such as Apache Nifi or Talend, to streamline and automate cleansing processes.
* Used Azure Data Factory as an orchestration tool for integrating data from upstream Systems.
* Employed data segmentation strategies within Salesforce Marketing Cloud to target specific audience segments, ensuring relevant and personalized messaging. Automated jobs using different triggers (Event, Scheduled, and Tumbling) in ADF.
* 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.
* Implemented Security roles on SSAS cubes and Power BI services by implementing authentication and authorization methods.
* Implemented ongoing improvements to data cleansing processes for sustained effectiveness.
* Developed different reports using Microsoft BI tools like SSIS and SSRS and developed Power BI dashboards that would help businesses operate smoothly.
* Creating Spark clusters and configuring high concurrency clusters using Azure Databricks to speed up the preparation of high-quality data.
* Instituted data quality assurance processes to monitor and maintain the accuracy and reliability of organizational data.
* Contributed to ongoing initiatives for the continuous improvement of data governance processes and practices.
* Implemented regular data audits and monitoring mechanisms to detect and address data governance issues proactively.
* Designed and implemented algorithms for inventory forecasting and demand planning, optimizing stock levels to meet customer demand and reduce excess inventory costs.
* Responsible for ingesting data from various source systems (RDBMS, Flat files, Bigdata) into Azure (Blob Storage) using the framework model.
* Ensured compliance with data protection regulations (e.g., GDPR, HIPAA) and industry-specific standards.
* Data ingestion to one or more Azure services (Azure Data Lake, Azure Storage, Azure SQL, Azure DW) and processing the data in Azure Databricks.
* Collaborated closely with retail operations teams to understand business requirements and translate them into effective data engineering solutions, ensuring alignment with organizational goals.
* Documented SQL queries, database schemas, and data processes to facilitate knowledge sharing and collaboration among team members.
* Applied sentiment analysis to customer feedback data, providing insights into customer satisfaction levels and uncovering opportunities for service and product improvements.
* Expert working with Azure BLOB and Data Lake storage and loading data into Azure SQL Synapse analytics (DW).
* Planning and design the solution to implement, using Azure Files Share.
* Propose architectures considering cost/spend in Azure and develop recommendations to right-size data infrastructure.
* Worked on Azure Templates for quick deployments and advanced PowerShell scripting.
* Implemented HA with Azure Classic and Azure Resource Manager deployment models.
* Implementing measures to ensure data quality, including validation and cleansing procedures.
* Using the Azure SDK, automated the testing of data pipelines and implemented a data quality check framework in Python.
* Implemented data-driven customer segmentation strategies, enabling personalized marketing campaigns, targeted promotions, and improved customer engagement.
* Experienced in Gitlab CI and Jenkins for CI and End-to-End automation for all build and CD.
* Implemented automated alerting and monitoring systems on top of our applications, which are built using tools like Grafana and Kibana. So, we would receive automated alerts in case of any production failures to report any on-call rotation person in the team.
* Extensive experience with Azure Delta Lake, Azure Databricks, Azure Data Factory, Azure HD Insights, and Azure functions.
* Established robust error handling and logging mechanisms within data pipelines to capture and address issues promptly, minimizing data discrepancies.
* Cleaning and preprocessing raw data to ensure accuracy, completeness, and consistency.
* Worked closely with data scientists for building predictive models using PySpark.
* Expertise in Architectural blueprints and detailed documentation. Create a bill of materials, including required cloud services and tools.

Environment: Azure Databricks, Azure Datalake Gen 1, Azure data factory, Databricks Delta Lake, Snowflake, Spark, HIVE, Kafka, JMS, Oracle, Teradata, Python, Tableau, Power Bi

|  |  |  |  |
| --- | --- | --- | --- |
| Client | **CBSI India Pvt. Ltd** | Location | **INDIA** |
| Designation | **Sr Data Engineer** | Duration | **Jan 2014 – Oct 2015** |

Responsbilities

* Involved in Data Extraction, aggregations, and consolidation of data within AWS Glue using PySpark.
* Created data pipeline and ETL process to load Data from AWS S3 to RDS and redshift.
* Developed and maintained conceptual, logical, and physical data models to guide database design and implementation.
* Designed EMR/Hadoop/Spark process to convert files to parquet and query on s3 using Athena, SQS, and Redshift Spectrum.
* Proficient in using data modeling tools such as ERwin, ER/Studio, or PowerDesigner to create and manage data models.
* Experience on AWS services like S3 bucket, lambda, AWS Glue, Cognito, Step Function, RedShift, and Cloud Watch using CDK.
* Used EMR to create spark clusters and EC2 instances and imported data stored In S3.
* Implemented Data Quality framework using AWS Athena, Snowflake, Airflow, and Python.
* Created ETL jobs on AWS glue to load data from different sources, transformations involving data cleaning, data imputation, and data mapping and storing the results into S3 buckets and stored data was later queried using AWS Athena.
* Worked on AWS Data Pipeline to configure data loads from S3 to Redshift.
* Used AWS Redshift, S3, Spectrum, and Athena services to query large amounts of data stored on S3 to create a Virtual Data Lake without having to go through the ETL process.
* Implemented AWS Elastic Container Service (ECS) scheduler to automate application deployment in the cloud using Docker Automation techniques.
* Used AWS Athena to query the data on top s3 bucket (files in inbound) which helps the business for data analysis.
* Enforced standards for data model representation, fostering consistency across projects and teams.
* Setting up of CI/CD pipeline using continuous integration tools such as Cloud Bees Jenkins and automated the entire AWS EC2, VPC, S3, SNS, RedShift, EMR based infrastructure using Terraform, Chef, Python, Shell, Bash scripts and managing security groups on AWS and custom monitoring using Cloud Watch.
* Monitor Resources and Applications using AWS Cloud Watch, including creating alarms to monitor metrics such as EBS, EC2, ELB, RDS, S3, and SNS, and configured notifications for the alarms generated based on events defined
* Identifying and investigating the root causes of data anomalies or discrepancies.
* Automate AWS infrastructure through infrastructure as code by writing various Terraform modules and scripts by creating AWS IAM users, groups, roles, policies, custom policies, AWS Glue, Crawlers, Redshift clusters, snapshots of clusters, EC2, and S3 buckets.
* Implemented version control for data artifacts, ensuring proper tracking and management of changes to data structures over time.
* Worked extensively on AWS Components such as Airflow, Elastic Map Reduce (EMR), Athena, and Snowflake.
* Implemented access controls and permissions to restrict data access based on user roles, ensuring data confidentiality and privacy.
* Written PostgreSQL sequels according to the business transformations to place the data into respective tables.
* Conducting exploratory data analysis (EDA) to identify patterns, trends, and outliers.
* Created the Splunk dashboard for run time errors directed from AWS cloud using kinesis.
* Used Autosys and MoveIT tools to place the file from on premise to AWS-S3 buckets.
* Created Redshift clusters on AWS for quick accessibility for reporting needs.
* Designed and deployed a Spark cluster and different Big Data analytic tools including Spark, Kafka streaming, AWS, and HBase with Cloudera Distribution.
* ptimized data models for performance, considering factors such as indexing, partitioning, and denormalization.
* Responsible for real-time heterogeneous replication from MySQL to Oracle DB and MySQL to AWS Redshift using the Tungsten replicator tool while performance tuning and maintaining both databases.

Environment: PySpark, Spark RDD, S3. JSON, Parquet, ETL, Hive, UNIX, HDFS, Power BI, SQL Server. DevOps, Kerberos, Scala, Python, AWS, Kafka, Redshift, Airflow, Snowflake, Cassandra, Map Reduce (EMR), AWS Athena, Linux, Ubuntu, EC2, MongoDB, JavaScript, Ajax, CAT, AWS Glue, Crawlers, Cloudera, Jenkins. data bricks.

|  |  |  |  |
| --- | --- | --- | --- |
| Client | **Financial Technologies Ltd.,** | Location | **INDIA** |
| Designation | **DE/DA** | Duration | **June 2012 – Dec 2013** |

Responsbilities

* Experience in developing SFTP, NAS integrations to ingest data into HDFS using Python.
* Implemented messaging queues and routes using microservices camel in java.
* Developed batch ingestion jobs from Teradata to HDFS and Hive using Sqoop.
* Developed data lake platform using multiple tools like Kafka, Sqoop, Hive, Spark and Oozie.
* Implemented end to end job automation in Hadoop using Apache Oozie.
* Developed transnationals system updates in HBase for data lake implementations.
* Developed end to end ETL operations in optimized way using Hive and Spark.
* Documenting data analysis processes, methodologies, and results for future reference.
* Expert in handling complex data issues and memory optimizations and tuning in Spark.
* Implemented multiple data pipelines using Apache Spark using python and Scala.
* Developed real time streaming application to ingest Json messages using Apache Kafka.
* Implemented Data security features in data exposed through API endpoints.
* Reated visually compelling charts, graphs, and dashboards to communicate complex data findings to both technical and non-technical stakeholders.
* Expert in writing in many features using scripting languages like Bash, Shell and Python.
* Worked on implementing CRUD operations in HBase for multiple applications.
* Handling the Tickets/Service Calls raised by End-users and providing them faster resolution.
* Implemented multiple Change Request includes of new developments.

Environment: Hortonworks, Spark, Hive, Oozie, Bash, Shell, Python, HBase, Teradata, Sqoop, Scala, data lake, SFTP, microservices, Java, Kafka, API, SQL.