Employee_ID: 1380

Data Engineer

TECHNICAL SKILLS

Databases: HiveQL, MySQL, AWS RDS, SQL Server, MongoDB.

ETL/ELT Tools: AWS Glue, Azure Databricks, Azure Data Factory, AWS EMR, Azure HDInsight, DBT, Informatica.

Data Analytics Tools: Power BI, Tableau, Amazon QuickSight.

Data Warehousing Tools: Snowflake, Azure Synapse Analytics, AWS Redshift, Hive.

Cloud Computing: AWS, Microsoft Azure.

Web Frameworks and technologies: Django, REST API, HTML, CSS.

Data Storage tools: AWS S3, Azure Blob storage, Azure Data Lake storage.

Data Streaming: AWS Kinesis, Flink.

Workflow Scheduling tools: Airflow, Oozie, SAP BODS, AWS Step functions.

Source Controls: GitHub, AWS Code Commit, GitLab.

Languages: Python, SQL, Scala, JavaScript.

CI/CD: Azure DevOps.

Tracking and documentation Tools: JIRA, Confluence, Rally.

SUMMARY STATEMENT

6+ years experienced Data Engineer with a strong background in developing resilient and scalable data engineering solutions within AWS and Azure cloud environments. Worked extensively in ETL/ELT process to load data from source to destination using AWS Glue, Azure cloud integration services (ADF and Databricks) and Snowflake native capabilities.

EXPERIENCE

PROJECTS

#1 Republic Services - US Ecology Data Migration – Data Engineer

- Development of incremental and full load Informatica IICS pipelines to ingest data from source systems (SQL Server, SharePoint) to staging schema in Snowflake staging layer.
- Create ServiceNow tickets to deploy table DDL and IICS job exports to STG and QA environments.
- Create full load and incremental DBT models to apply column naming standards and transformation logics like de-duplication, handling null and negative values in the ODS layer.
- Create DBT test cases for not null and uniqueness in YML files.
- Create GitHub pull requests to merge DBT models to develop, master, and release branches.
- Work collaboratively with QA team to resolve defects to ensure data quality.
- Development of DBT macros for purging full load staging tables as post-hook runs and implementing soft delete flags.
- Development of CORE Fact and dimension DBT models for Billing and Retail dashboard.
- Conceptual and logical Data modeling for Invoiced and Not invoiced Work Orders, Retail Orders, and Receipts.
- Development of Aggregate tables and views for flash invoiced and uninvoiced reporting and retail dashboards.

#2 Oracle to Snowflake Data Migration – Snowflake/AWS Glue Developer

- Designed and implemented end-to-end data migration pipelines from Oracle to Snowflake using AWS Glue PySpark jobs.
- Created Snowflake staging, integration, and ODS schemas with optimized table

CERTIFICATION

EDUCATION

B. Tech – Electrical and Electronics Engineering

structures, clustering, and micro-partitioning strategies.
- Built external stages, file formats, and Snowflake integration objects for secure data ingestion through S3.
- Developed Glue ETL jobs to extract Oracle data via JDBC, transform datasets, and write Snowflake-ready Parquet files to S3.
- Implemented COPY INTO pipelines to load incremental and full datasets from S3 into Snowflake staging tables.
- Created transformation layers using advanced Snowflake SQL, including joins, aggregations, CDC logic, surrogate key generation, and data quality rules.
- Built Streams and Tasks to automate incremental processing and maintain near-real-time updates in target tables.
- Used Time Travel, Zero-Copy Cloning, and Query Profiling to debug issues, optimize performance, and ensure data consistency.
- Implemented column-level masking, RBAC roles, warehouses, and resource monitors to secure and optimize Snowflake environments.
- Developed DBT models on Snowflake for business logic, standardization, and automated testing.
- Conducted end-to-end validation between Oracle and Snowflake using reconciliation scripts and row-level comparisons.
- Prepared HLD/LLD documentation and supported CI/CD deployments for Snowflake objects and Glue pipelines.

#3 Enterprise Data Lake Integration – ETL/Snowflake Developer

- Designed and developed end-to-end data ingestion pipelines using AWS Glue PySpark to load data from SQL Server and REST APIs into Snowflake landing layers and S3 for archival data.
- Built Snowflake schemas, staging tables, ODS models, file formats, and pipes, ensuring scalable and high-performance data loading.
- Implemented complex SQL transformations, including data normalization, surrogate key generation, CDC logic, and SCD-based updates.
- Created Snowflake external stages on S3, enabling optimized ingestion from Glue jobs and downstream workflows.
- Developed reusable AWS Glue transformation modules that produced Snowflake-ready structures (column formatting, partitioning, validation layers).
- Implemented column-level Snowflake data masking policies to ensure compliance with data security requirements.
- Coordinated deployments of Snowflake DDL, role grants, and Glue ETL pipelines through ServiceNow and CI/CD processes.
- Performed extensive data quality checks within Snowflake, using custom SQL tests and reconciliation logic to validate Glue-processed data.

#4 Real-Time Analytics Dashboard using AWS Kinesis

- Implemented real-time analytics to process high-velocity beat logs from an e-commerce application.
- Created IAM roles and permissions for AWS services including Kinesis, DynamoDB, API Gateway, Glue, and QuickSight.
- Provisioned an API Gateway endpoint to receive POST requests from the e-commerce application for real-time data ingestion.
- Configured the API Gateway trigger to process incoming logs and write raw data into S3 storage.
- Recorded audit information such as source, record counts, previews, and timestamps into DynamoDB.
- Streamed processed events into an AWS Kinesis input data stream for real-time processing.
- Developed a Flink application in Kinesis Data Analytics to read the input stream and create streaming tables.
- Performed real-time transformations and tumbling window aggregations for user activities, product status, and brand insights.
- Published transformed results into a Kinesis output data stream.
- Configured Kinesis Firehose to batch data based on time and memory thresholds and deliver it to S3.
- Organized S3 folders for curated datasets such as user activities, product status, and brand insights with proper partitioning.
- Used AWS Glue Crawlers to catalog the curated S3 folders in the Glue Data Catalog.
- Queried curated datasets through Amazon Athena for analysis and reporting.
- Built near real-time dashboards in QuickSight using Athena as the data source.
- Ensured end-to-end reliability and performance across API Gateway, DynamoDB, Kinesis Streams, Firehose, and Flink pipelines.

#5 HCL Tech: Snowflake Central ingestion – ETL/Snowflake Developer

- Creation of file format and storage integration object for AWS S3 external stage.
- Data ingestion from multiple sources like REST API, SAP HANA.
- Developed Snowflake external tables and views with transformation and flattening queries.
- Development and deployment of Glue PySpark jobs through CI/CD to ingest historical data from S3.
- CDC and delta data ingestion from SAP HANA through streams and tasks.
- Developed Snow pipe to continuously load CSV data into staging tables from S3.
- Enabled role-based data masking policies on certain columns to ensure data privacy and confidentiality.
- Efficiently used time-travel and zero copy cloning to debug and solve column

data type mismatches and data inconsistencies.
- Created stored procedures to insert master ingestion load statistics table for audit purposes.
- Manual ingestion of parquet data from web scraping sources like Recruitment, Media, Blogs, Patent, and YouTube by building SQL queries in Snowflake worksheet.
- Development of composite keys in Market Analytics tables as a part of snowflake COPY INTO statement to merge delta records.
- Development of CDC modules for Snowpark stored procedure.
- Development of technical flow documentation, HLD and LLD documentation.

#6 HCL Tech: DAP Analytics Dashboard – Azure Databricks/PySpark Developer

- Developed and maintained data processing pipelines on the Databricks platform using PySpark, optimizing for performance and scalability.
- Implemented ETL processes to extract, transform, and load data from various sources such as SAP HANA, API, and flat files from Azure Blob storage.
- Developed SQL queries and Spark SQL scripts for data manipulation and analysis.
- Automated data pipeline orchestration and scheduling using Apache Airflow and Databricks Jobs.
- Implemented Airflow monitoring and logging solutions to track performance and troubleshooting of issues.
- Optimized Spark jobs for performance and scalability, leveraging techniques such as partitioning and caching.
- Generation of parquet files for archival purposes in Azure blob storage.
- Created interactive visualizations and dashboards using Databricks notebooks and visualization libraries to communicate insights to stakeholders.

#7 HCL Tech: AWS EMR Migration and Transformation - Big Data Developer

- Deployed Hadoop components including Spark, YARN, Hive, Hue, Oozie, Tez, Sqoop, Drill, and Zookeeper on AWS EMR.
- Developed IAM users and roles with respective policies for EMFT (SFTP) to access AWS resources like S3.
- Development of Bash scripts/utilities to automate remediation of PySpark jobs, python/shell scripts, and Hive tables.
- Implementation of SCD Type-2 in SAP Concur solution.
- Created partitioning and bucketing in Hive tables and usage of Tez engine and CBO for improved Query performance.
- Handled migration and re-architecture of applications: ServiceNow, AIOPS, Supply, SAP Fieldglass, and Product management.
- Creation of Glue data catalogs to consume parquet files on S3 from

Tableau using Amazon Athena connector.

- Apache Drill setup in AWS, configuration of S3 storage plugins in Drill UI explorer to access HBase tables through Drill views.
- Hadoop administrative tasks: Changing HDFS file permissions and ownerships, creation of users, groups, and IAM policies.
- Collaborated with SAP Basis team and worked on Remote Sources in SAP HANA to validate schemas.
- Remediation of ETL jobs on SAP BODS and Wherescape Red.
- Installation of Amazon Hive ODBC in Wherescape Red and DSN configuration for loading and consolidation jobs.
- Established SDA Remote source for AWS in SAP HANA to consume Hive data through virtual tables.
- Workflow scheduling using CRON expressions in Oozie coordinator using Hue UI.
- MIT Kerberos configuration and authentication in Tableau server to consume data from Hive Data Warehouse.
- Implemented data warehouse for archival solutions on S3 using Amazon Redshift to support analytics and reporting needs.
- Creation of external Hive tables on HBase tables with respective column mappings and storage handlers.
- Development of PySpark script leveraging Spark-SQL API to convert Hive-HBase data to Parquet compressed format.

#8 Integra Software Services: Azure Data Ingestion and transformation – Azure ETL/ELT Developer

- Development of ELT pipeline in Azure Data Factory to ingest raw data from On- premises SQL server into Azure Data Lake Storage in parquet format.
- Developed data transformation logic using Azure Databricks notebook for processing and enriching large volumes of data.
- Optimized ETL workflows for performance and scalability by leveraging Azure Databricks clusters, parallel processing, and partitioning techniques.
- Provision of serverless SQL pool in Synapse Analytics for data warehouse schemas on transformed data.
- Created views on Data Lake storage for Power BI and Tableau reporting.
- Integrated Azure Data Factory with Azure DevOps for continuous integration and continuous deployment (CI/CD), automating build, test, and deployment processes.
- Performed POC on interactive dashboards and reports using Power BI to visualize insights from Azure Synapse Analytics.

#9 Integra Software Services: Azure HDInsight migration – Big Data Developer

- Coordinated with Platform team on Azure HDInsight provisioning, including cluster sizing, network configuration, and integration with Azure services.
- Creation of secrets in Azure Key Vault service to store credentials for external systems.
- Conducted performance tuning and optimization of Apache Spark jobs to improve processing efficiency and reduce latency.
- Implemented Spark SQL queries and PySpark scripts for data analysis. Created Oozie schedules and coordinator jobs for PySpark jobs.
- Analysis of YARN resource manager for performance and root queuing.
- Creation of Hive external tables with partitions to load staging data using Hive queries.
- Documentation of architecture diagrams, deployment procedures, and technical specifications for the data analytics platform.