|  |
| --- |
| **Syam Chandu Koya**  **Lead Data Engineer** |

**Logo, company name

Description automatically generated**

**Professional Summary**

* Total 11+ years hands on experience with building product ionized data ingestion & processing pipelines using Java, Spark, Scala etc also experienced in designing & implementing production grade data warehousing solutions on large scale data technologies.
* Strong experience in migrating other databases to Snowflake.
* Worked with domain experts, engineers, & other data scientists to develop, implement, & improve upon existing systems.
* Experienced in analysing data using HiveQL.
* Well versed with Big data on AWS cloud services i.e. EC2, S3, Glue, Anthena, DynamoDB and RedShift
* Participated in design meetings for creation of the Data Model & provide guidance on best data architecture practices.
* Experienced with Snowflake Multi Cluster Warehouses.
* Hands on experience in configuring and deploying Applications using different web/application servers such as Terraform and CircleCI
* Understanding of Snowflake cloud technology.
* Experienced with Snowflake cloud data warehouse & AWS S3 bucket for integrating data from multiple source system which include loading nested JSON formatted data into snowflake table.
* Professional knowledge of AWS Redshift.
* Strong experience on AWS-EMR and Azure Cluster setup, Spark installation, HDFS and MapReduce Architecture. Along with that having a good knowledge on Spark, Scala and Hadoop distributions like Apache Hadoop, Cloudera and Azure.
* Strong experience in All Hadoop and Spark ecosystems include Hive, Pig, Sqoop, Flume, Kafka, Cassandra, SparkSQL, Spark Streaming, SparkR and Flink.
* Hands in Terraform in managing resource scheduling, disposable environments and multitier application
* Experience working on creating and running Docker images with multiple micro - services .
* Experienced in building Snowpipe.
* Extensive knowledge of Database Technologies (DBT) and best practices in data modeling.
* Experienced in using Snowflake Clone & Time Travel.
* Experienced in various data ingestion patterns to Hadoop.
* Utilized Kubernetes and Docker for the runtime environment for the CI/CD system to build, test, and deploy.
* Participated in the development improvement & maintenance of snowflake database applications.
* Experienced in various methodologies like Waterfall & Agile.
* Writing Unit Test cases and submitted Unit test results as per the quality process for Snowflake, Ab initio and Teradata changes.
* Experienced in Microsoft Power BI, Power BI Pro, Power Query, Power Pivot, Power BI Service, Power BI Desktop, Microsoft Excel Pivot Table, and data models to gather information from multiple sources and deliver information to the end user.
* Extensive experience in developing complex stored Procedures/BTEQ Queries.
* In-depth understanding of Data Warehouse/ODS, ETL concept & modeling structure principles.
* Built Logical & Physical data model for snowflake as per the changes required.
* Defined roles, privileges required to access different database objects.
* In-depth knowledge of Snowflake Database, Schema & Table structures.
* Defined virtual warehouse sizing for Snowflake for different type of workloads.
* Worked with cloud architect to set up the environment.
* Coding for Stored Procedures/ Triggers.
* Designed batch cycle procedures on major projects using scripting & Control.
* Developed SQL queries SnowSQL.
* Experienced with using Microsoft Power BI for the creation of dashboards, standard reports and ad-hoc reports.
* Hands on experience in setting up workflow using Apache Airflow and Oozie workflow engine for managing and scheduling Hadoop jobs.
* Worked on connecting tableau to various other data sources like Oracle, Teradata, SAP Hana, Microsoft Access, & SQL Server & had experience in establishing respective driver connections.
* Worked on developing complex SQL scripts for Teradata database for creating BI layer on DW for tableau reporting.
* Extensively worked with Microsoft Business Intelligence Services such as SSIS, SSRS, SSAS.
* Worked in the conversion of SQL server functions into the Teradata Store procedure for confirming the data.
* Develop transformation logic using snowpipeline.
* Optimized & fine-tuned queries.
* Delivered all API integration efforts for the Automic platform and WEB interface
* Documented and delivered upgrade procedure for the Automic application cluster
* Worked on performance tuning of Big Data workloads.
* Have good Knowledge in ETL & hands on experience in ETL.
* Operationalized data ingestion, data transformation & data visualization for enterprise use.
* Mentored & trained junior team members & ensured coding standard is followed across the project.
* Helped talent acquisition team in hiring quality engineers.
* Experienced in real time streaming frameworks like Apache Storm.
* Worked on Cloudera & Hortonworks distribution.
* Have progressive experience in the field of Big Data Technologies, Software Programming & Developing, which also includes Design, Integration, Maintenance.
* Hands-on experienced with Snowflake utilities, SnowSQL, SnowPipe, Big Data model techniques using Python / Java.
* Worked on ETL pipelines in & out of data warehouse using combination of Python & Snowflakes SnowSQL and writing SQL queries against Snowflake.

**Technical Skills**

|  |  |
| --- | --- |
| **Cloud Technologies:** | Snowflake, SnowSQL, Snowpipe AWS |
| **Spark, Hive:** | LLAP, Beeline, Hdfs, MapReduce, Pig, Sqoop, HBase, Oozie, Flume |
| **Reporting System:** | Splunk |
| **Hadoop Distributions:** | Cloudera, Hortonworks |
| **Programming Languages:** | Scala, Python, Perl, Shell Scripting |
| **Data Warehousing:** | Snowflake, Redshift, Teradata |
| **DBMS:** | Oracle, SQL Server, MySQL, Db2, Database Technologies (DBT). |
| **Operating System:** | Windows, Linux, Solaris, Centos, OS X |
| **IDES:** | Eclipse, Netbeans |
| **Servers:** | Apache Tomcat |

**Professional Experience**

**Centene Corporation St. Louis, MO August 2021 – Till Date**

**Lead Data Engineer**

**Responsibilities:**

* Worked on SnowSQL & Snowpipe.
* Converted Talend Joblets to support the snowflake functionality.
* Created Snowpipe for continuous data load.
* Used COPY to bulk load the data.
* Created data sharing between two snowflake accounts.
* Was responsible for creating on-demand tables on S3 files using Lambda Functions and AWS Glue using Python and PySpark.
* Created internal & external stage & transformed data during load.
* Redesigned the Views in snowflake to increase the performance.
* Unit tested the data between Redshift & Snowflake.
* Developed data warehouse model in snowflake for over 100 datasets using whereScape.
* Created Reports in Looker based on Snowflake Connections.
* Experienced in working with AWS, Azure & Google data services.
* Automated resulting scripts and workflow using Apache Airflow and shell scripting to ensure daily execution in production.
* Extensively used Terraform in AWS Virtual Private Cloud to automatically setup and modify settings by interfacing with control layer.
* Developed Proof of Concepts (POC) for various projects using Power BI.
* Developed business logic using DAX to provide custom functionality in Power BI Applications.
* Scheduled refreshes through the Power BI gateway to refresh Dashboard at regular intervals and appropriate dependencies.
* Implemented AWS Elastic Container Service (ECS) scheduler to automate application deployment in the cloud using Docker Automation techniques.
* Support developers with complex scripting in Automic and developed multiple scripts for KPI reporting
* Install Spark, and integrate with other Big data ecosystems like Hive, HBase.
* Integrate Kafka with Spark and get the social media data through Twitter API.
* Expertise in Terraform for multi cloud deployment using single configuration.
* Collaborate with Other Analysis team (R & Python & Tableau) to analyze the data.
* Integrate HiveQL, JSON, CSV data and run SparkSQL on the top of the different datasets.
* Created monitors, alarms, notifications and logs for Lambda functions, Glue Jobs, EC2 hosts using Cloudwatch
* Used AWS Glue for the data transformation, validate and data cleansing.
* Validation of Looker report with Redshift database.
* Worked on cloud deployments using maven, docker and Jenkins.
* Constructing the enhancements in Ab Initio, UNIX and Informix.
* Created and Documented upgrade procedure for the Automic cluster
* Used Toad to verify the counts and results of the graphs and Tuning of Ab Initio graphs for better performance.
* Developed and maintained data pipelines using Database Technologies (DBT) to transform raw data into actionable insights.
* Designed and launched CI/CD framework with CircleCI which led to more reliable code and cleaner data. Interface with stakeholders to assess needs, development requirements, and timelines.
* Implemented version control and CI/CD pipelines for DBT projects using tools like GitLab CI/CD, enabling automated testing and deployment.
* Good working knowledge of any ETL tool (Informatica or SSIS).
* Created Talend Mappings to populate the data into dimensions & fact tables.
* Wrote ETL jobs to read from web APIs using REST & HTTP calls & loaded into HDFS using java & Talend.
* Used Talend big data components like Hadoop & S3 Buckets & AWS Services for redshift.
* Validated data from SQL Server to Snowflake to make sure it has a match.
* Consulting on Snowflake Data Platform Solution Architecture, Design, Development, & deployment focused to bring the data driven culture across the enterprise.
* Building solutions once and for all with no band-aid approach.
* Implemented Change Data Capture technology in Talend to load deltas to a Data Warehouse.
* Developed stored procedures/views in Snowflake & use in Talend for loading Dimensions & Facts.
* Designed, developed, tested, implemented & supported Data Warehousing ETL using Talend.
* Excellent knowledge on RDBMS topics, ability to write complex SQL, PL/SQL.

**Environment:** Snowflake, Redshift, SQL server, AWS, Azure, Talend, DBT, Jenkins & SQL.

**Dell Technologies Round Rock, TX Fed 2020 to July 2021**

**Data Engineer**

**Responsibilities:**

* Created data pipeline for several events of ingestion, aggregation & load consumer response data from AWS S3 bucket into Hive external tables.
* Worked on Oracle Databases, RedShift & Snowflakes.
* Core Technologies: End to end data platform with Snowflake, Matillion, Power BI, Qubole, Databricks, Tableau, Looker, Python.
* Implemented a 'server less' architecture using API Gateway, Lambda, & Dynamo DB & deployed AWS Lambda code from Amazon S3 buckets. Created a Lambda Deployment function & configured it to receive events from your S3 bucket.
* Install and configure Apache Airflow for S3 bucket and Snowflake data warehouse and created dags to run the Airflow.
* Created conceptual, logical & physical models for OLTP, Data Warehouse Data Vault & Data Mart Star/Snowflake schema implementations.
* Developed Spark Applications by using Scala, Java and Implemented Apache Spark data processing project to handle data from various RDBMS and Streaming sources.
* Converted the Boulder Tables from Excel to Power BI with better Graphs and visualizations for the understanding and improvement of business.
* Used Terraform in managing resource scheduling, disposable environments and multitier application.
* Collected the Plants Data in Excel sheets from different Regions around the World and combined them in Power BI to know the Loss and Profit of the Product.
* Performed COPQ and DPPM analysis in Power BI and helped Global Quality Leaders to take the Business-driven decisions.
* (DevOps) Automated CI/CD workflow with CircleCI, sbt, Docker, Terraform.
* Utilized Kubernetes and Docker for the runtime environment for the CI/CD system to build, test, and deploy.
* Worked with the Spark for improving performance and optimization of the existing algorithms in Hadoop using Spark Context, Spark - SQL, Spark MLlib, Data Frame, Pair RDD's, Spark YARN.
* Experience in implementing Spark RDD's in Scala.
* Configured Spark streaming to get ongoing information from the Kafka and store the stream information to HDFS.
* Designed and implemented data models using Database Technologies (DBT) methodologies, optimizing for performance and efficiency.
* Designed and implemented modular DBT models to streamline the data transformation process, allowing for easy reuse and scalability across multiple projects.
* Working experience with Kimball Methodology & Data Vault Modeling.
* Defined virtual warehouse sizing for Snowflake for different type of workloads.
* Major challenges of the system were to integrate many systems & accesses them which are spread across South America, creating a process to involve third party vendors & suppliers, creating authorization for various department users with different roles.
* Experienced in Building a Talend job outside of a Talend studio as well as on TAC server.
* CSM certified, Worked in Scaled Agile (SAFE) environment as System/DBT QA, hands on experience with Rally, JIRA.
* Setup full CI/CD pipelines so that each commit a developer makes will go through standard process of software lifecycle & gets tested well enough before it can make it to the production.
* Validated the data load process for Hadoop using the HiveQL Query’s.
* Evaluated Snowflake Design considerations for any change in the application.
* Build the Logical & Physical data model for snowflake as per the changes required.
* Defined roles, privileges required to access different database objects.
* Defined virtual warehouse sizing for Snowflake for different type of workloads.
* Designed & coded required Database structures & components.
* Built Logical & Physical data model for snowflake as per the changes required.
* Developed the Pysprk code for AWS Glue jobs & for EMR.
* Worked on scalable distributed data system using Hadoop ecosystem in AWS EMR, MapR distribution.
* Used more components in Talend & Few to be mentioned: tjava, toracle, txmlMap, tdelimited files, tlogrow, tlogback components etc. in many of my Jobs Designs.
* Worked on Joblets (reusable code) & Java routines in Talend.
* Implemented Talend POC to extract data from Salesforce API as an XML Object & .csv files & load data into SQL Server Database.
* Worked with cloud architect to set up the environment.
* Educated developers on how to commit their work & how they can make use of the CI/CD pipelines that are in place.

**Environment:**Snowflake, SQL server, AWS, & SQL, DBT.

**Vanguard Valley Forge, PA February 2016 to December 2019**

**Data Engineer**

**Responsibilities:**

* Developed workflow in SSIS to automate the tasks of loading the data into HDFS & processing using hive.
* Developed alerts & timed reports also developed and managed Splunk applications.
* Involved in various Transformation & data cleansing activities using various Control flow & data flow tasks in SSIS packages during data migration.
* Applied various data transformations like Lookup, Aggregate, Sort, Multicasting, Conditional Split, Derived column etc.
* Worked with multiple data sources.
* Developed Mappings, Sessions, & Workflows to extract, validate, & transform data according to the business rules using Informatica.
* Worked with Various HDFS file formats like Avro, Sequence File, & various compression formats like snappy, Gzip.
* Worked on data ingestion from Oracle to hive.
* Conducted performance tuning and optimization of DBT models and queries, resulting in significant improvements in query execution times and resource utilization.
* Used Terraform for infrastructure to spin up EC2, EMR, Lambda as per requirement and enable
* Utilized Kubernetes and Docker for the runtime environment for the CI/CD system to build, test, and deploy.
* Implemented DBT tests and assertions to validate data integrity and accuracy throughout the transformation process, ensuring reliable and trustworthy insights.
* Involved in fixing various issues related to data quality, data availability & data stability.
* Worked in determining various strategies related to data security.
* Performed performance monitoring & optimizing Indexes tasks by using, Performance Monitor, SQL Profiler, Database Tuning Advisor & Index tuning wizard.
* Worked on Hue interface for Loading the data into HDFS & querying the data.
* Designed & Created Hive external tables using shared Meta-store instead of derby with partitioning, dynamic partitioning, & buckets.
* Wrote scripts & indexing strategy for a migration to Confidential Redshift from SQL Server & MySQL databases.
* Used spark-SQL to create Schema RDD & loaded it into Hive Tables & handled structured data using Spark SQL.
* Worked on AWS Data Pipeline to configure data loads from S3 to into Redshift.
* Used JSON schema to define table & column mapping from S3 data to Redshift.
* Involved in converting Hive/SQL queries into Spark transformation using Spark RDDs.
* Used Avro, Parquet & ORC data formats to store in to HDFS.

**Environment:** AWS, Redshift, Hive, Spark, MS SQL Integration Services, MS SQL Reporting Services 2008, DBT.

**Client: Sankalpa Soft Solutions, India** **July2013 – Dec2015 Role: ETL Developer**

**Responsibilities**:

* Involved in gathering Business Requirements, Business Analysis, Design and Development, Testing, And Implementation of business rules.
* Involved in the Complete Software Development Lifecycle Experience (SDLC) from Business Analysis to Development,
* Testing, Deployment, and Documentation.
* Created tables, views in Teradata, according to the requirements.
* Used Teradata utilities FastLoad, MultiLoad, TPT to load data.
* Wrote numerous BTEQ scripts to run complex queries on the Teradata database.
* Created proper Primary Index taking into consideration of both planned access of data and even distribution of data.
* Documented DBT models, transformations, and data lineage to provide transparency and facilitate knowledge sharing among stakeholders.
* across all the available AMPS. Responsible for Performance tuning at various levels during development.
* Implemented complex data flows through data flow tasks, executed sql tasks, sequence containers, and for each loop
* containers.
* Assisted in staging Historical and Incremental Loads.
* Worked extensively on Aggregated/Summarized data.
* Performed tuning and optimization of complex SQL queries using Teradata Explain.
* Used Power Centre Workflow Manager to create workflows, sessions, and also used various tasks like command, event
* wait, event raise, email.
* Worked extensively in the Development of large Projects with complete END to END participation in all areas of the
* Software Development Life Cycle and maintained documentation.

Environment: Teradata 12, Oracle 9.i/10g, Teradata Visual Explain, BTEQ, DBT, Teradata SQL Assistant, Fast Load, Multi Load, Fast Export, UNIX Shell Scripting, ETL

**Education**

Bachelor’s in mechanical engineering from JNTUK.  
Masters in computer science from Colorado technical university 2017  
Masters in Information Technology and Cybersecurity New England college 2022

**Certifications:**

1. AWS Certified Data Analytics-Specialty (Security).
2. Snowflake DB.