**Senior Data Engineer**

**Sahruday**

**Phone:** +1 (269)-364-1924

**Email:** nagulapelly9@gmail.com

**PROFESSIONAL SUMMARY**

* Around 9+ years of experience in developing, deploying, and managing big data applications, ETL architecture, and cloud computing platforms like Amazon Web Services (AWS) and Microsoft Azure. Extensive expertise in designing and optimizing ETL pipelines using Ab Initio and Snowflake, ensuring efficient data integration, transformation, and cloud-based data warehousing. Proven ability to deliver results at a global level with deep knowledge in Retail, Healthcare, and Finance Domains.
* Strong background in Azure Data Factory (ADF), Azure Synapse Analytics, and Apache NiFi for data integration, movement, and transformation.
* Designed and optimized ETL pipelines using Ab Initio, improving data extraction, transformation, and loading processes for financial and healthcare datasets.
* Developed high-performance SQL queries for data modeling, analytics, and reporting, optimizing data retrieval and reducing query execution times.
* Hands-on experience in Apache NiFi for data movement and workflow automation, ensuring seamless data processing.
* Strong knowledge of the Software Development Lifecycle (SDLC), Agile methodologies, and DevOps practices for seamless project execution.
* Extensively worked on AWS Cloud services like EC2, VPC, IAM, RDS, EMR, EKS, Auto-scaling, S3, Cloud Front, Glacier, Elastic Beanstalk, Lambda, Elastic Cache, Cloud Watch, Cloud Formation, RedShift, DynamoDB, SNS, SQS, SES, Kinesis, IAM.
* Hands on experience in Azure Analytics Services-Azure Data Lake store (ADLS), Azure Data Lake Analytics (ADLA), Azure SQL DW, Azure SQL DB, Azure Data Factory (ADF), Azure Data Bricks (ADB), Azure Cloud, Azure Devops, Azure Synapse Analytics, Azure Cosmos NOSQL DB, etc. and built Azure environments by deploying Azure IaaS Virtual Machines (VMs) and cloud services (PaaS).
* Expertise in using various Hadoop infrastructures such as Map Reduce, Pig, Hive, Zookeeper, Sqoop, Oozie, Flume, Drill and spark for data storage and analysis, Scala, Python, SQL and NoSQL databases such as HBase, MongoDB, Cassandra.
* Proficiency in AWS Console and CLI, effectively leveraging AWS services such as RDS (PostgreSQL, MySQL, Aurora), CloudWatch, Redshift, and Snowflake to architect robust, scalable, and high-performance database solutions.
* Efficient in writing Infrastructure as a code (Iac) in Terraform, Azure resource management (ARM), AWS cloud formation. Created reusable Terraform modules in both Azure and AWS cloud platforms.
* Excellent work base experience on OLAP application development using AbInitio, Sybase, Oracle and Teradata Procedure’s.
* Implemented MDM configuration for Party, Vendor, Product and customer and Expertise in developing MDM using Cadis/ Markit EDM.
* Experience in using highly scalable parallel processing infrastructure using DataStage parallel Extender and experienced in Mapping Server/Parallel Jobs in DataStage to populate tables in Data warehouse and Data marts.
* Hands on experience in setting up workflow using Apache Airflow and Oozie workflow engine for managing and scheduling Hadoop jobs.
* Designed and integrated adobe Web portal with adobe campaign module to provide a streamline and automated campaign execution system to business users and Experienced with Adobe Campaign in creating, executing, tracking and managing cross channel campaigns.
* Worked on Designing of the logical and physical data model creating custom objects, defining relationships as per application requirements.
* Used Power BI to create self-service BI capabilities and use tabular models. Performed time series analysis using DAX expression in Power BI and Worked with both live and import data into Power BI for creating reports.
* Experience in developing custom UDFs for Pig and Hive to incorporate methods and functionality of Python/Java into Pig Latin and HQL (HiveQL) and Used UDFs from Piggybank UDF Repository.
* Expertise in working with Hive optimization techniques like Partitioning, Bucketing, vectorizations and Map side-joins, Bucket-Map Join, skew joins, and creating Indexes.
* Strong working experience with SQL and NoSQL databases, data modeling and data pipelines. Involved in end-to-end development and automation of ETL pipelines using SQL and Python.
* Worked on reading and writing multiple data formats like JSON, ORC, Parquet on HDFS using PySpark.

**TECHNICAL SKILLS**

|  |  |
| --- | --- |
| **Big Data Technologies** | AB Initio, MapReduce, Hive, Tez, Python, PySpark, Scala, Kafka, Spark streaming, Oozie, Sqoop. |
| **Hadoop Distribution** | Cloudera, Horton Works |
| **Cloud Eco-system** | AWS S3, EC2, EMR, Lambda, cloud formation, Redshift, Glue, EKS, Azure Data Factory(ADF), Azure Data Lake, Azure Databricks, Azure SQL Data Warehouse. |
| **Reporting Tools** | Tableau, Power BI, Splunk, IBM Cognos. |
| **Operating System** | Linux, Unix, windows, Cent OS |
| **Big data Technologies** | Teradata, Hadoop, HDFS, Map Reduce, Hive, Pig, SQOOP, HBase, Oozie, Flume, Zookeeper, PySpark, Spark, YARN, snowflake, Spark Streaming and Kafka |
| **Languages** | Java, Scala, Python, SQL, T-SQL, Spark SQL, Snowflake SQL, Pig Latin, HiveQL, Shell Scripting, IBM DataStage scripting language. |
| **Development Environments** | IntelliJ, PyCharm, VS Code, Eclipse, Azure Data Studio, IBM DataStage Designer, Apache Zeppelin, Databricks Notebook, Jupyter Notebook. |
| **Development Methodologies** | Agile/Scrum, Waterfall, Kanban, JIRA (Project Management). |
| **Certifications** | Snowflakes Certification. |

**PROFESSIONAL EXPERIENCE**

**AAA Insurance | Dallas, TX April 2022 - Present**

**Senior Data Engineer**

**Responsibilities:**

* Demonstrated working knowledge of Azure cloud components, such as Databricks, Data Lake, Blob Storage, Data Factory, Storage Explorer, SQL DB and SQL DWH.
* Developed ETL pipelines in Ab Initio, optimizing data extraction, transformation, and loading processes for large-scale data integration.
* Implemented data movement and transformation solutions using Apache NiFi, ensuring efficient and reliable data workflows.
* Designed and optimized SQL queries to extract and process financial and healthcare datasets, improving data accessibility and reporting accuracy.
* Led the migration of legacy ETL processes to Snowflake, reducing query execution times by 40% through performance tuning and indexing.
* Automated data ingestion and validation using Azure Data Factory (ADF) and Synapse Analytics, improving efficiency in financial data processing.
* Utilized Databricks on Azure to analyze data from Azure data storages, leveraging Spark cluster capabilities to derive valuable insights and worked with Databricks lake house architeure and used the features such as Unity Catalog, DLT, DB-SQL and DBR-Workflows and Delta tables .
* Performed migrating SQL Database to Azure Data Lake, Azure data lake Analytics, Azure SQL Database, Databricks and Azure SQL Data warehouse and Controlling, granting database access and migrating on premise databases to Azure Data Lake store using Azure Data factory (ADF) and in Databricks notebooks.
* Maintain RDMS database, ensuring space needs are addressed in a timely manner and designed, programmed, implemented and monitored the master client Index system using RDMS and MQ to maintain a common client index for the legacy SAVERR system and the TIERS system. Also Provided database support for RDMS.
* Involved in creating End-to-End data pipeline within distributed environment using the big data tools, Spark framework and Tableau for data visualization.
* Install and configure Apache Airflow for S3 bucket and Snowflake data warehouse and create dags to run the Airflow
* Managed ingestion workflows into Azure Services (Azure Data Lake, Azure Storage, Azure SQL) and facilitated data processing in Azure Databricks, ensuring compliance with regulations and security standards.
* Use of Salesforce emailing APIs to automate email services.
* Implemented Jenkins jobs to create Azure infrastructure from GitHub repositories containing Terraform code and created on-premises active directory authentication using automation with ansible play books.
* Developed Spark applications using Spark-SQL in Databricks for data extraction, transformation, and aggregation from multiple file formats for analyzing & transforming the data to uncover insights into the customer usage patterns.
* Developed dashboards and visualizations to help business users analyze data as well as providing data insight to upper management with a focus on Microsoft products like SQL Server Reporting Services (SSRS) and Power BI.
* Created Spark clusters and configured high concurrency clusters using Azure Databricks to speed up the preparation of high-quality data.
* Developed Spark Streaming applications to process real-time data from various sources, such as Kafka and Azure Event Hubs.
* Built streaming ETL pipelines using Spark Streaming to extract data from various sources, transform it in real-time, and load it into a data warehouse like Azure Synapse Analytics.
* Worked extensively with Azure BLOB and Data Lake storage, efficiently loading data into Azure SQL Synapse analytics (DW) and Good Knowledge of Data Build Tool (DBT) with Snowflake.
* Migrated batch processing jobs from Teradata to Snowflake, leveraging Snowflake’s time travel and zero-copy cloning features.
* Utilized tools like Azure Databricks or HDInsight to scale out the Spark Streaming cluster as per requirements.
* Developed Spark APIs to import data into HDFS from Teradata and created Hive tables.
* Designed and optimized Teradata SQL queries, reducing query execution time by 40% using indexing and partitioning strategies.
* Developed and managed ETL pipelines for processing large insurance datasets, utilizing Teradata BTEQ, FastLoad, and MultiLoad.
* Migrated data from Teradata to Snowflake, leveraging Teradata Parallel Transporter (TPT) and Snowpipe for efficient data movement.
* Created partitioned and bucketed Hive tables in Parquet File Formats with Snappy compression and Loaded data into Parquet Hive tables from Avro Hive tables.

**Environment:** Azure Cloud, Hadoop, HDFS, Yarn, MapReduce, Hive, Sqoop, Oozie, Kafka, Spark SQL, Spark Streaming, Eclipse, Informatica, Oracle, CI/CD, PL/SQL UNIX Shell Scripting, Cloudera Snowflakes.

**MetLife | New York City, NY Sept 2020 - March 2022**

**Senior Data Engineer**

**Responsibilities:**

* Developed PySpark Data Ingestion framework to ingest source claims data into HIVE tables by performing Data cleansing, Aggregations, and applying De-dup logic to identify updated and latest records.
* Improved ETL processes using Ab Initio, resulting in a 25% reduction in data processing time.
* Integrated financial data from multiple sources into Azure Synapse and Snowflake, ensuring compliance with industry regulations.
* Designed data models and optimized Snowflake schemas, ensuring efficient query performance for financial risk analysis.
* Created NiFi data flows to triggers spark jobs and uses an email processor to get notifications if there are any failures.
* Extracted, transformed, and loaded data from source systems to Azure Data Storage services using a combination of Azure Data Factory, Databricks, PySpark, Spark, SQL, and U-SQL Azure Data Lake Analytics.
* Designed and implemented **ETL pipelines** using **Azure Data Factory**, automating data workflows from on-premises and cloud sources (SQL Server, Azure Blob Storage, Azure SQL Database).Developed data ingestion pipelines on Azure HDInsight Spark cluster using Azure Data Factory and Spark SQL.
* Prepared an ETL framework with the help of Sqoop, pig and hive to be able to frequently bring in data from the source and make it available for consumption.
* Built ETL pipelines using Teradata utilities (FastLoad, MultiLoad, TPump) to handle high-volume data ingestion.
* Used Scala and Kafka to create data pipelines for structuring, processing, and transforming given data.
* Developed stored procedures and macros in Teradata SQL, automating data transformation processes.
* Apache Spark data processing project to handle data from various RDBMS and Streaming sources.
* Developed Spark application using Spark Core, Spark SQL and Spark Streaming API’s.
* Used Oozie to create big data workflows for ingesting data from various sources to Hadoop.
* Developed Spark jobs using Scala for better data processing and used Spark SQL for querying.
* Leverage Spark features such as In-Memory processing, Distributed Cache, Broadcast, Accumulators, Map side Joins to implement data preprocessing pipelines with minimal latency.
* Designed and Developed Real Time Stream Processing Application using Spark, Kafka, Scala, and Hive to perform Streaming ETL and apply Machine Learning.
* Created pipelines in Azure Data Factory to facilitate the extraction, transformation, and loading of data from Azure SQL, Blob storage, and Azure SQL Data Warehouse.
* Ingest the Data into Cargill data lake from different sources and did some transformations in data lake with Spark-Scala as per the business requirements.
* Developed CICD pipelines and developed require Docker images for the pipelines.
* Worked extensively with Azure BLOB and Data Lake storage, efficiently loading data into Azure SQL Synapse analytics (DW) and Good Knowledge of Data Build Tool (DBT) with Snowflake.
* Improved the performance of the pipelines further using Apache Spark and Scala with batch and stream processing of the data based on the requirement.
* Automated the data flow and data validations on the input and output data to simplify the testing process using Shell Scripting and SQL.
* Developed complex Teradata SQL scripts to aggregate and analyze insurance claim data for regulatory compliance.

**Environment:** Azure, S3, RDS, Redshift, Lambda, Boto3, DynamoDB, Apache Spark, HBase, Apache Kafka, HIVE, SQOOP, Map Reduce, Snowflake, Scala, Apache, Pig, Java, SSRS, Tableau.

**Cigna | Bloomfield, Connecticut March 2018 - Aug 2020**

**Data Engineer**

**Responsibilities:**

* Created new inputs for AWS CloudWatch Logs, Config Rules, and Billing in Splunk Add on for AWS.
* Integrated Big Data Spark jobs with EMR and glue to create ETL jobs for around 450 GB of data daily.
* Developed Spark applications using Scala and Spark-SQL for data extraction, transformation, and aggregation from multiple data file formats to uncover insights into the customer usage patterns.
* Designed and developed ETL pipelines and dashboards using Step Function, Lambda, Glue and Quick Sight.
* Designed and developed high-quality integration solutions by using Denodo virtualization tool (read data from multiple sources including Oracle, Hadoop, MySQL).
* Converting HL7 V.2 messages to XML format to facilitate proper mapping onto the CDA document templates.
* Worked on HL7 inbound and outbound interfaces through Mirth connect and designed HL7 interfaces from various vendor specifications for internal and external labs.
* Responsible for developing web applications using LINQ, LAMDA, WCF functions.
* Architected a cloud-based data warehousing solution leveraging Redshift on AWS, optimizing for scalability, security, and performance.
* Knowledge on Implementing hierarchy, relationships type, packages and profile for hierarchy management in MDM Hub implementations.
* Extracted the data from other data sources into HDFS using Sqoop.
* Handled importing data from various data sources, performed transformations using Hive, MapReduce, loaded data into HDFS and Extracted the data from MySQL into HDFS using Sqoop.
* Implemented automation for deployments by using YAML scripts for massive builds and releases.
* Implemented Data classification algorithms using MapReduce design patterns.
* Collaborated with stakeholders to discern specific requirements, implementing tailored data models and structures in Redshift.
* Developed and fine-tuned Spark jobs for data transformations, aggregations, and machine learning tasks on extensive datasets on AWS EMR, ensuring accuracy and reliability in analytics.
* Orchestrated seamless transitions from on-premises databases to AWS RDS and S3, utilizing AWS DMS for efficient data transfer while ensuring minimal downtime and preserving data integrity.
* Migrated ETL pipelines from on-premise Teradata to Snowflake on AWS.
* Integrated Power BI with diverse data sources, ensuring real-time or scheduled updates and customized reports to match reporting requirements and organizational preferences using Athena in AWS
* Implemented data lineage and metadata management solutions to monitor data flow and transformations effectively, maintaining audit trails and compliance records AWS.

**Environment**: Snowflakes, HDFS, Spark, Scala, PySpark, ADF, Kafka, AWS, Pig, Map Reduce, Hive, Spark SQL, RedShift, AWS, Hadoop, Oozie, SQOOP, Maven, Zookeeper, Power BI.

**Abbott Laboratories | Hyd, India May 2014 - Nov 2017**

**Data Engineer**

**Responsibilities**

* Worked on GIT to maintain source code in Git and GitHub repositories.
* Designed and developed scalable and cost-effective architecture in AWS Big Data services for data life cycle of collection, ingestion, storage, processing, and visualization.
* Processed HDFS data and created external tables using Hive and developed scripts to ingest and repair tables that can be reused across the project and Batch processing with Java/ Spark SQL.
* Development of RESTful WS web services using Java and Spring boot.
* Worked on building centralized Data Lake on AWS Cloud utilizing primary services like S3, EMR, Redshift, and Athena.
* Developed PySpark-based pipelines using spark data frame operations to load data to EDL using EMR for jobs execution & AWS S3 as a storage layer.
* Developed ETL jobs using Spark -Scala to migrate data from Oracle to new MySQL tables.
* Rigorously used Spark -Scala (RRD’s, Data frames, Spark SQL) and Spark - Cassandra -Connector APIs for various tasks (Data migration, Business report generation etc.)
* Designed, developed, and deployed ETL pipelines using AWS services like, Lambda, Glue, EMR, Step Function, CloudWatch events, SNS, Redshift, S3, IAM, etc.
* Developed Spark Streaming application for real-time sales analytics.
* Analyzed the source data and handled efficiently by modifying the data types. Used excel sheet, flat files, CSV files to generate PowerBI ad-hoc reports.
* Utilized Amazon AWS to generate Ubuntu replications for use in testing of altered PHP and JavaScript code blocks.
* Analyzed the SQL scripts and designed the solution to implement using PySpark
* Extensively worked on creating combiners, Partitioning, distributed cache to improve the performance of MapReduce jobs.

**Environment:** Hadoop, Hive, AWS, Spark, PySpark, Sqoop, Spark SQL, Shell script, Cassandra, YAML, ETL.