**PRAGNYA M PH: 571-305-1632**

**Sr. Data Engineer Email:** **pr9409@gmail.com**

**Professional Summary:**

* Over 8 years of experience as a Data Engineer, including Analysis, Design, Development, and Implementation.
* Proficient in conducting Joint Application Development (JAD) sessions for requirements gathering with SMEs, stakeholders, and project team members.
* Knowledge of NoSQL databases like HBase, MongoDB, Cassandra, and Big Query.
* Skilled in Data Migration, Data Conversion, Data Quality, Data Integration, and Metadata Management. Experienced in configuration management.
* Strong experience in migrating databases to Snowflake cloud data warehouse.
* Proficient in developing scalable systems using Hadoop technologies in various environments.
* Proficiency in SQL, PL/SQL, and Python coding.
* Hands-on experience in AWS Redshift migration, AWS Lambda functions, and API Gateway for data submission.
* Extensive experience with cloud provisioning tools like Terraform and CloudFormation. Skilled in building and optimizing AWS data pipelines and architectures.
* Strong development skills with Azure Data Lake, Azure Data Factory, SQL Data Warehouse, Azure Blob, and Azure Storage Explorer.
* Experience in developing complex mappings, reusable transformations, sessions, and workflows using ETL tools.
* Familiarity with agile project methodologies and ceremonies.
* Hands-on experience with CI build-automation tools like Maven and Jenkins.
* Solid understanding of architecting and designing large-scale data and analytics solutions on Snowflake.
* Experience in performance tuning on Oracle databases using explain plans and SQL query optimization.
* Expertise in writing Hadoop Jobs for data analysis using MapReduce, Hive, Kafka, and Splunk.
* Strong experience in Data Migration, Data Cleansing, Transformation, Integration, and Data Import/Export.
* Experience in importing and exporting large volumes of data between HDFS and Relational Database Systems using Sqoop.
* Familiarity with Business Intelligence/Data Warehousing solutions, including ETL, OLAP, Data marts, and tools like Informatica, Tableau, and Power BI.
* Proficient in Development, Testing, Implementation, Documentation, and Production support.

**Professional Experience:**

**Sr. Data Engineer**

**Cummins Columbus, Indiana July 2022 to Present**

**Responsibilities**

* Function as Data Engineer, assisted in leading the plan, building, and running states within the Enterprise Analytics Team.
* Lead architecture and design of data processing, warehousing and analytics initiatives.
* Engaged in solving and supporting real business issues with your Hadoop distributed File systems and Open Source framework knowledge.
* Responsible for data governance rules and standards to maintain the consistency of the business element names in the different data layers.
* Involved in various phases of development analyzed and developed the system going through Agile Scrum methodology.
* Performed detailed analysis of business problems and technical environments and use this data in designing the solution and maintaining data architecture.
* Designed and implemented a recommendation system which leveraged Google Analytics data and the machine learning models and utilized Collaborative filtering techniques to recommend facilities for different patients.
* Built the data pipelines that will enable faster, better, data-informed decision-making within the business.
* Developed Spark scripts by using python and bash Shell commands as per the requirement.
* Worked on POC to check various cloud offerings including Google Cloud Platform (GCP).
* Developed a POC for project migration from on prem Hadoop MapR system to GCP.
* Compared Self hosted Hadoop with respect to GCPs Data Proc, and explored Big Table (managed HBase) use cases, performance evolution.
* Used Terraform to reliably version and create infrastructure on GCP.
* Created resources, using GCP Terraform modules, and automated infrastructure management.
* Build a program with Python and apache beam and execute it in cloud Data flow to run Data validation between raw source file and big query tables.
* Used Rest API with Python to ingest Data from and some other site to Big Query.
* Loaded and transformed large sets of structured, semi structured and unstructured data using Hadoop/Big Data concepts.
* Performed Data transformations in Hive and used partitions, buckets for performance improvements.
* Optimized Hive queries to extract the customer information from HDFS.
* Involved in scheduling Oozie workflow engine to run multiple Hive jobs.
* Continuously monitor and manage data pipeline (CI/CD) performance alongside applications from a single console with GCP.
* Write a Python program to maintain raw file archival in GCS bucket.
* Designed Google Compute Cloud Data Flow jobs that move data within a 200 PB data lake.
* Implemented scripts that load Google Big Query data and run queries to export data.
* Implemented business logic by writing UDFs and configuring CRON Jobs.
* Providing 24/7 On-call Production Support for various applications.

**Environment**: Hadoop 3.3, Spark 3.1, Python, Terraform, GCP, Data Lake, GCS, HBase, Oozie, Hive, CI/CD, Big Query, Hive, Rest API, Agile Methodology.

**Sr. Data Engineer**

**UBS Weehawken, NJ January 2020 to June 2022**

**Responsibilities**

* Performed data engineering responsibilities using agile software engineering practices.
* Used Agile (SCRUM) methodologies for Software Development.
* Participated in Code Reviews, Enhancement discussion, maintenance of existing pipelines & systems, testing and bug-fix activities on-going basis.
* Worked closely with the business analysts to convert the Business Requirements into Technical Requirements and prepared low- and high-level documentation.
* Migrated data warehouses to Snowflake Data warehouse.
* Configured Hadoop Ecosystems to read data transactions from HDFS and Hive.
* Defined virtual warehouse sizing for Snowflake for different type of workloads.
* Reviewing the explain plan for the SQLs in Redshift.
* Imported data from RDBMS to HDFS and Hive using Sqoop on regular basis.
* Worked on Snowflake environment to remove redundancy and load real-time data from various data sources into HDFS using Kafka.
* Imported data from AWS S3 into Spark RDD, Performed transformations and actions on RDD's.
* Written POCs in Python to analyze the data quickly before applying big data solutions to process at a scale.
* Responsible for data governance rules and standards to maintain the consistency of the business element names in the different data layers.
* Migrated Matillion pipelines and Looker reports from Amazon Redshift to Snowflake data warehouse.
* Utilized Matillion ETL solution to develop pipeline that extract and transform data from multiple sources and loading to Snowflake.
* Worked on Ingesting data by going through cleansing and transformations and leveraging AWS Lambda, AWS Glue and Step Functions.
* Integrated and automated data workloads to Snowflake Warehouse.
* Created tables in snowflake DB, loading and analyzing data using Spark-Scala scripts.
* Wrote complex Hive queries to extract data from heterogeneous sources (Data Lake) and persist the data into HDFS.
* Developed ETL pipelines in and out of data warehouse using combination of Python and Snowflake’s Snow SQL.
* Ingested data into HDFS using SQOOP and scheduled an incremental load to HDFS.
* Worked on Kafka REST API to collect and load the data on Hadoop file system and also used Sqoop to load the data from relational databases.
* Extract Real time feed using Kafka and Spark Streaming and convert it to RDD and process data in the form of Data Frame and save the data as Parquet format in HDFS.
* Used Spark-Streaming APIs to perform necessary transformations and actions on the data got from Kafka and Persists into HDFS.
* Created external tables with partitions using Hive, AWS Athena and Redshift.
* Developed the Pyspark code for AWS Glue jobs and for EMR.
* Created S3 buckets also managing policies for S3 buckets and Utilized S3 bucket and Glacier for storage and backup on AWS.
* Develop Python and SQL used in the transformation process in Matillion.
* Used AWS S3 Buckets to store the file and injected the files into Snowflake tables using Snow Pipe and run deltas using Data pipelines.
* Worked on complex SNOW SQL and Python Queries in Snowflake.
* Create, modify and execute DDL in table AWS Redshift tables to load data.
* Performance tuning the tables in Redshift.
* Developed logging system in Python that keeps tracks of the activities in the ETL solutions developed in Python.
* Troubleshoot and maintain ETL/ELT jobs running using Matillion.
* Collaborated and assisted in building Power BI dashboards.

**Environment**: Apache Hadoop 3.3, Spark 3.3, AWS, Glue, S3, Matillion, EMR, Snowflake, SNOW SQL, AWS Redshift, Snow Pipe, Sqoop, Pyspark, Kafka, REST API, ETL & Agile/Scrum

**Data Engineer**

**Chevron Corporation, Santa Rosa, NM June 2017 to December 2019**

**Responsibilities**

* As Data Engineer I have taken initiative and ownership to provide business solutions on time.
* Created High level technical design documents and Application design documents as per the requirements and delivered clear, well-communicated and complete design documents.
* Implemented Azure Data Factory (ADF) extensively for ingesting data from different source systems like relational and unstructured data to meet business functional requirements.
* Design and developed Batch processing and real-time processing solutions using ADF, Databricks clusters and stream Analytics.
* 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.
* Maintain and provide support for optimal pipelines, data flows and complex data transformations and manipulations using ADF and PySpark with Databricks.
* Automated jobs using different triggers like Events, Schedules and Tumbling in ADF.
* Created, provisioned different Databricks clusters, notebooks, jobs and auto scaling.
* Created several Databricks Spark jobs with Pyspark to perform several tables to table operations.
* Performed data flow transformation using the data flow activity.
* Implemented Azure, self-hosted integration runtime in ADF.
* Developed spark programming code in Python Databricks workbooks.
* Improved performance by optimizing computing time to process the streaming data by optimizing the cluster run time.
* Developed Power BI reports using power query as a feed from SQL Server & different Data sources.
* Ensure data accuracy, integrity, and reliability of both in back-end Power BI reports.
* Created Linked services to connect the external resources to ADF.
* Wrote complex SQL queries including joins, correlated sub queries, scalar sub queries, views, Stored Procedures and Triggers
* Ensured the developed solutions are formally documented and signed off by business.
* Worked with team members to resolve any technical issue, Troubleshooting, Project Risk &Issue identification, and management.
* Worked on the cost estimation, billing, and implementation of services on the cloud.

**Environment**: Azure Cloud, Azure Data Factory (ADF v2), Azure functions Apps, Azure Data Lake, BLOB Storage, SQL server, Windows remote desktop, AZURE Power Shell, Data bricks, Python, Azure SQL Server, Azure Data Warehouse.

**Data Analyst**

**Careator Technologies Pvt Ltd Hyderabad, India December 2015 to March 2017**

**Responsibilities**

* Worked as Data Analyst, analyzed data quality issues against source system and prepared a data quality document confirming all the source data quality.
* Worked with the application and Business Analyst team to develop requirements.
* Involved in Data profiling and performed Data Analysis based on the requirements, which helped in catching many Sourcing Issues upfront.
* Used SQL tools to run SQL queries and validate the data loaded into the target tables.
* Created tables, views, sequences, indexes, constraints and generated SQL scripts for implementing physical data model.
* Used AWS services EC2 and S3 for small data sets processing and storage.
* Designed and developed architecture for data services ecosystem spanning Relational, NoSQL, and Big Data technologies.
* Involved in extensive data validation by writing several complex SQL queries.
* Involved in back-end testing and worked with data quality issues.
* Created dimensional model for reporting system by identifying required dimensions and facts using Erwin.
* Worked and extracted data from various database sources like XML and Flat files into the Data Stage.
* Consumed XML messages using Kafka and processed the xml file using Spark streaming to capture UI updates.
* Evaluated data profiling, cleansing, integration and extraction tools.
* Involved in extensive Data validation by writing SQL queries and Involved in back-end testing and worked with data quality issues.
* Responsible for the data architecture design delivery, data model development, review, approval and Data warehouse implementation.
* Involved in designing and developing Data Models and Data Marts that support the Business Intelligence Data Warehouse.
* Involved in extensive data validation by writing several complex SQL queries and Involved in back-end testing and worked with data quality issues.
* Created SQL tables with referential integrity, constraints and developed queries using SQL, SQL\*PLUS and PL/SQL.
* Performed GAP analysis of current state to desired state and document requirements to control the gaps identified.
* Managed all indexing, debugging and query optimization techniques for performance tuning using T-SQL.
* Generated comprehensive analytical reports by running SQL queries against current databases to conduct data analysis.
* Wrote T-SQL statements for retrieval of data and involved in performance tuning of T-SQL queries and Stored Procedures.
* Developed pre-processing job using Spark Data frames to flatten Json documents to flat file.
* Developed Dashboard reports on Tableau.

**Environment**: SQL, PL/SQL, SAS, Microsoft Excel 2010, T-SQL, Pivot Tables, VLOOK up, triggers, Stored Procedures, Tableau.

**Database Engineer**

**Hudda Infotech Private Limited Hyderabad, India July 2014 to November 2015**

**Responsibilities:**

* Experienced with PLSQL programming for developing applications on databases. Developed PLSQL procedures, functions, triggers, and packages as per application need.
* Automated database health checks and daily operations using Python programming and Unix Shell scripting.
* Developed UNIX Shell scripts for generating HTML email alerts for various performance and monitoring metrics of databases.
* Developed bash scripts to identify database growth periodically and took necessary actions by contacting application developers for either archiving data or purging unwanted data from production environment.
* As a part of performance management, identified long running jobs and user application queries and accordingly took necessary actions. Identified server and memory level bottlenecks. Helped application developers in resolving their database performance related issues.
* Experienced with identifying worst performing SQLs and fixed them with the help of development Teams. Periodically run automated scripts that shows long running queries’ report and thereby understand the performance issues in the databases. Discussed the same with the development teams and tool necessary actions such as tuning queries, changing queries, indexing tables, or purging unwanted data from tables.
* Worked on application developers’ change requests to roll out the development changes in production.
* Performed database installation, upgrade, PSU, CPU, Bugfix patching. Worked on managing disaster recovery (DR) databases and carried out DR drills.

**Environment**: PL/SQL, Python, Unix, HTML