**Rohith Dhane  
Sr. Data Engineer**

**Email :** [**rohith.dhane69@gmail.com**](mailto:rohith.dhane69@gmail.com) **|| Contact .No: +1 (667) 577-0261**

**Professional Summary:**

* Around 12 years of hands-on experience, I've refined my expertise in crafting and executing comprehensive data solutions and architectures, with a specialization in Data Engineering.
* Extensive experience in Data Warehousing and Decision Support Systems, with a robust track record in managing full lifecycle projects.
* Skilled in utilizing GCP services such as Compute Engine, Cloud Load Balancing, Cloud Storage, Cloud Dataproc, Cloud Functions, Cloud Pub/Sub, Cloud Shell, Cloud SQL, BigQuery, Cloud Data Fusion, Cloud Dataflow, Stack Driver Monitoring, And Cloud Deployment Manager.
* Hands-on experience in data analysis using HiveQL, custom MapReduce programs (Java/Python), and complex HiveQL queries for data extraction, along with developing Hive User Defined Functions (UDFs) as necessary. Proven track record of utilizing Snowflake database and Python for development and optimization tasks, ensuring high proficiency and efficiency in database-related coding and implementation.
* Profound understanding and practical application of Hadoop/Big Data technologies, encompassing storage, querying, processing, and analysis.
* Engaged in software development on cloud computing platforms, including Amazon Web Services (AWS), Azure, and Google Cloud Platform (GCP).
* Deep knowledge of Hadoop's architecture and ecosystem, including key components like HDFS, MapReduce, and various ecosystem tools.
* Proficient in Python and Apache Beam for data validation and processing in Google Cloud Dataflow.
* Implemented custom-built input adapters using Spark, Hive, and Sqoop to ingest data for analytics from various sources (Snowflake, MS SQL, MongoDB) into HDFS. Imported data from web servers and Teradata using Sqoop, Flume, and Spark Streaming API.
* Skilled in installing, configuring, and utilizing Hadoop ecosystem components like Hadoop MapReduce, HDFS, Hive, and Sqoop.
* Expertise in data analysis using HiveQL, HBase, and custom MapReduce programs.
* Proficient in importing/exporting data using Sqoop between HDFS and Relational Database Systems like Teradata, Oracle, and SQL Server.
* Experienced in designing and implementing migration strategies for traditional systems on Azure, utilizing services like Azure SQL Database and Azure Data Factory.
* Hands-on experience in GCP, particularly in BigQuery, Cloud Dataflow, and Data Proc.
* Developed complex data mappings and executed Proof of Concepts (POC) for transitioning MapReduce jobs into Spark transformations.
* Skilled in developing Apache Spark jobs using Python and Spark SQL for efficient data processing.
* Strong understanding of statistics and experience in developing machine learning models, including usage of Databricks.
* Proficient in Python, Scala, and core computer science concepts like data structures and algorithms.
* Developed applications for data processing tasks using various databases including Teradata, Snowflake, and Postgres.
* Built ETL pipelines, visualizations, and analytics-based solutions using AWS, Azure Databricks, and open-source frameworks.
* Well-versed in Hadoop distributions like Cloudera, Hortonworks, and AWS EMR.
* Deep understanding of AWS services and cloud architectures across AWS, Azure, and GCP.
* Implemented schedulers using tools like Oozie, Airflow, and shell scripts.

**Technical Skills:**

|  |  |
| --- | --- |
| Big Data Technologies | HDFS, YARN, MapReduce, Hive, Pig, Impala, Sqoop, Storm, Flume, Spark, Apache Kafka, Zookeeper, Ambari, Oozie, MongoDB, Cassandra, Mahout, Puppet, Avro, Parquet, Snappy, Falcon. |
| NO SQL Databases | Postgres, HBase, Cassandra, MongoDB, Amazon DynamoDB, Redis |
| Hadoop Distributions | Cloudera (CDH3, CDH4, and CDH5), Hortonworks, MapR, and Apache. |
| Lanuages | Scala, Python, R, XML, XHTML, HTML, AJAX, CSS, SQL, PL/SQL, HiveQL, Unix, Shell Scripting |
| Source Code Control | GitHub, CVS, SVN, ClearCase |
| Cloud Computing Tools | Amazon AWS, (S3, EMR, EC2, Lambda, VPC, Route 53, Cloud Watch, CloudFront), Microsoft Azure, GCP |
| Databases | Teradata Snowflake, Microsoft SQL Server, MySQL, DB2 |
| DB languages | MySQL, PL/SQL, PostgreSQL & Oracle |
| Build Tools | Jenkins, Maven, Ant, Log4j |
| Business Intelligence Tools | Tableau, Power BI |
| Development Tools | Eclipse, IntelliJ, Microsoft SQL Studio, Toad, NetBeans |
| ETL Tools | Talend, Pentaho, Informatica, Ab Initio, SSIS |
| Development Methodologies | Agile, Scrum, Waterfall, V model, Spiral, UML |

**Professional Experience:**

**Client: CareFirst, MD Nov 2023 to Present**

**Role: SR Data Engineer**

**Responsibilities:**

* Migrating on-premise ETLs to GCP using cloud native tools such as BigQuery, Cloud DataProc, Google Cloud Storage, Composer.
* Process and load inbound and outbound Data from Google pub/sub topic to Bigquery using cloud Dataflow with Python.
* Working with Cloud Data Catalog and other Google Cloud APIs for monitoring, query, and billing related analysis for BigQuery usage.
* Used Cloud Functions to support the data migration from BigQuery to the downstream applications.
* Creating BigQuery authorized views for row-level security to expose the data to other teams.
* Collaborated with data analysts and business stakeholders to understand their SQL query requirements and provide tailored solutions to meet their needs.
* Implemented and maintained data governance policies and procedures to ensure the integrity, security, and compliance of SQL databases within GCP.
* Designed and implemented NoSQL data models in Google Cloud Bigtable, tailoring structures to specific application requirements.
* Proficient in writing complex SQL queries for data analysis, manipulation, and reporting tasks within GCP environments.
* Performed Data analysis and Data profiling using complex SQL on various sources systems including Oracle and Teradata.
* Integrated and optimized data processing workflows using Google Cloud Dataflow for scalable, parallelized data transformations and analysis across various data sources and formats.
* Managed workflow orchestration and task scheduling using Cloud Composer, ensuring seamless execution of Airflow DAGs for Spark job submissions and other data processing tasks.
* Designed and implemented data pipelines leveraging GCP Data Fusion for streamlined orchestration, transformation, and ingestion of structured and unstructured data from diverse sources.
* Conducted proof-of-concepts using PySpark, comparing performance with Hive and SQL/Teradata on Yarn clusters, informing technology decisions and optimizations.
* Integrated Teradata Warehouse into EMR clusters, handling error resolution and performance tuning in Teradata queries and utilities, ensuring smooth data integration.
* Led the design, development, and implementation of new features within the existing framework
* Led the development of machine learning models for predictive analytics, enhancing decision-making processes and driving business growth.
* Managed end-to-end ETL pipelines, ensuring seamless data integration across various systems and platforms.
* Orchestrated data streaming pipelines using Kafka, enabling real-time data processing and analysis for timely insights.
* Developed complex PL/SQL scripts to optimize data retrieval and manipulation, improving system performance.
* Implemented data engineering best practices, including CI/CD pipelines and automated testing frameworks, to ensure the reliability and scalability of data pipelines.
* Utilized Git for version control and Jira for project management, ensuring transparency and efficiency throughout the development lifecycle.
* Played a key role in troubleshooting and resolving data-related issues, maintaining high availability and reliability of systems.
* Developed efficient standalone scripts in PySpark, incorporating transformations in line with defined business logic, optimizing data workflows.
* Leveraged expertise in Python and Object-oriented concepts to tackle complex problems, ensuring robust and scalable solutions for data processing tasks.

**Environment:** GCP, BigQuery, GCS Bucket, G-Cloud Functions, SSIS, Cloud Dataflow, Cloud Data Fusion, Cloud Shell, Cloud Composer, Gsutil, Dataproc, Snowflake, VM Instances, Airflow, Jenkins, Jira, Git, Gitlab, Cloud SQL, MySQL, Postgres, DBeaver, Scala, Spark, Hive, Spark-SQL.

**CBRE, Dallas, TX Aug 2021 to Oct 2023**

**Role: Sr. Data engineer**

**Responsibilities:**

* Developed Spark programs to parse raw data, populate staging tables, and store refined data in partitioned tables within the Enterprise Data Warehouse.
* Wrote SQL scripts for data mismatch analysis and managed history data loading during data migration from Teradata SQL to Snowflake.
* Leveraged cloud and GPU computing technologies for automated machine learning and analytics pipelines such as GCP.
* Worked on POC to check various cloud offerings including Google Cloud Platform (GCP).
* Compared Self hosted Hadoop with respect to GCs Data Proc and explored Big Table (managed HBase) use cases, performance evolution.
* Built data pipelines in airflow in CP for ETL related jobs using different airflow operators.
* Designed various Jenkins jobs to continuously integrate the processes and executed CI/CD pipeline using Jenkins.
* Performed extensive exploratory data analysis using Teradata to improve the quality of the datasets.
* Successfully migrated an entire Oracle database to BigQuery and utilized Power BI for reporting. Constructed data pipelines in Google Cloud Platform's Apache Airflow for ETL tasks, leveraging various Airflow operators.
* Worked on analyzing the data using PySpark, Hive, bases on ETL mappings.
* Experienced in Implementing Continuous Delivery pipeline with Maven, Ant, Jenkins, and GCP.
* Experienced in migrating legacy systems into GCP technologies, storing data files in Google Cloud S3 Buckets daily basis, Using DataProc, Big Query to develop and maintain GCP cloud base
* Developed both streaming and batch processing applications using PySpark to ingest data from diverse sources into the HDFS Data Lake.
* Utilized Cloud Shell SDK in GCP to configure services like Data Proc, Storage, and BigQuery.
* Got involved in migrating on prem Hadoop system to using GCP (Google Cloud Platform).
* Migrating an entire oracle database to BigQuery and using of power bi for reporting. Build data pipelines in airflow in GCP for ETL related jobs using different airflow operators. Experienced in GCP DataProc, GCS, Cloud functions, BigQuery.
* Experienced in Google cloud components, Google container builders and CP client libraries and cloud SDK'S.
* Optimized SQL queries and data pipelines in Looker to improve query performance and reduce processing time, enhancing overall system efficiency.
* Developed shell scripts to parameterize Hive actions within Oozie workflows and schedule jobs. Loaded massive amounts of data into HDFS and Cassandra using Apache Kafka.
* Worked with NoSQL databases like HBase, integrating them with PySpark for real-time streaming data processing and persistence.
* Proficient in GCP services such as Dataproc, Google Cloud Storage (GCS), Cloud Functions, and BigQuery.
* Designed and developed data pipelines for integrated data analytics, utilizing Hive, Spark, Sqoop, and MySQL.

**Environment:** GCP, GCS, Pub/Sub, Airflow, Data Proc, Airflow, Looker, Hadoop, Spark, Sqoop, Data Warehouse, Kafka, Cloud Functions, Big Query, Pyspark, Spark SQL, No SQL.

**Merck Pharma, Branchburg, NJ Oct 2019 to July 2021**

**Role: Data Engineer**

**Responsibilities:**

* Engineered data pipelines in Apache Airflow on Google Cloud Platform (GCP) to efficiently handle ETL tasks using a diverse set of Airflow operators.
* Explored Spark to optimize performance and refine existing algorithms on Hadoop, leveraging Spark's robust features including Spark Context, Spark SQL, DataFrames, and Spark YARN.
* Utilized Spark Streaming to seamlessly ingest data into an in-house ingestion platform.
* Developed RESTful APIs using Python, employing Flask and Django frameworks, and seamlessly integrated various data sources such as Java, JDBC, RDBMS, Shell Scripting, Spreadsheets, and Text files.
* Provided technical support and troubleshooting for Python-based implementations in Athena and Teradata, resolving issues promptly to maintain continuous operation of real-time data analytics pipelines.
* Designing and implementing data processing systems on GCP using services such as Big Query, Dataflow, and Datapost.
* Build data pipelines in airflow in GCP for ETL-related jobs using different airflow operators. Experience in GCP Data proc, GCS, Cloud functions, Big Query.
* Building and managing data warehouses and data lakes on GCP, ensuring data integrity and security.
* Integrated Python scripts with other data sources and analytics tools to enable comprehensive real-time data analysis and reporting capabilities across Athena and Teradata platforms.
* Integrating external data sources and APIs into GCP data solutions, ensuring data quality and consistency.
* Building data transformation pipelines using GCP services like Dataflow or Apache Beam to cleanse, normalize, and enrich data.
* Assisted with migration and upgrade of Jira instances, ensuring a seamless transition with minimal disruption to projects. Facilitated communication between development, operations, and business teams through Jira's collaboration features.
* Crafted PySpark scripts to automate the processing and transfer of files to third-party vendors. Established robust data pipelines in GCP using Apache Airflow, leveraging diverse Airflow operators to streamline ETL-related tasks.
* Demonstrated proficiency in GCP services such as Dataproc, Google Cloud Storage (GCS), Cloud Functions, and BigQuery.
* Managed continuous data transfers using Snowpipe and crafted SnowSQL queries for thorough data analysis.
* Operated within a Scrum/Agile environment, utilizing effective project management tools such as JIRA.
* Implemented Spark using Scala and SparkSQL to expedite data processing and streamline testing procedures.
* Translated complex Hive/SQL queries into Spark transformations using Spark RDDs in tandem with Scala.
* Developed Python scripts to efficiently extract data from datasets and store them in HDFS using PySpark.
* Designed Python solutions to retrieve data from HBase and effectively implemented them using PySpark.

**Environment:** GCP, GCP Dataproc, Apache Beam, Airflow, Hadoop, Teradata, Teradata Spark, EMR, S3, Hive, Python, Airflow, Teradata, Hive, Spark SQL, SQL.

**Dollar Tree, Chesapeake, VA Dec 2016 to Sep 2019**

**Role: Data Engineer**

**Responsibilities:**

* Strategically designed and configured relational servers and databases on the Azure Cloud, meticulously assessing both current and future business requirements.
* Played a pivotal role in the seamless migration of data from on-premises SQL servers to Cloud databases, specifically Azure Synapse Analytics (DW) and Azure SQL DB.
* Demonstrated extensive proficiency in creating pipeline jobs, scheduling triggers, and mapping data flows using Azure Data Factory (V2), ensuring secure storage of credentials in Key Vaults.
* Skillfully created elastic pool databases and scheduled elastic jobs to execute T-SQL procedures, optimizing resource allocation.
* Developed ETL jobs to efficiently load, serve, and transport data into buckets, facilitating the transfer of S3 data to the Data Warehouse.
* Leveraged Kusto Explorer for log analytics and enhanced query response times, crafting alerts using Kusto query language.
* Successfully designed tabular models within Azure Analysis Services to fulfill specific business reporting requirements.
* Efficiently worked with Azure BLOB and Data Lake storage, seamlessly loading data into Azure SQL Synapse Analytics (DW).
* Addressed complex business queries involving multiple tables from different databases by crafting both correlated and non-correlated sub-queries.
* Designed and implemented business intelligence solutions using SQL Server Data Tools 2015 and 2017 versions, effectively loading data into both SQL and Azure Cloud databases.
* Conducted comprehensive analysis of data quality and enforced business rules at all stages of the data extraction, transformation, and loading process.
* Managed the testing and migration of ETL Informatica workflows and mappings, ensuring smooth transitions.
* Oversaw validation and verification of software across various testing phases including Functional Testing, System Integration Testing, End to End Testing, Regression Testing, and more.
* Developed batch processing pipelines for data processing using Python and Airflow, while also scheduling Spark jobs using Airflow for streamlined operations.
* Actively contributed to writing, testing, and running MapReduce pipelines using Apache Crunch for efficient data processing.
* Effectively managed Hadoop log files, analyzed SQL scripts, and designed solutions for processes using Spark.
* Crafted insightful reports in TABLEAU for data visualization and thoroughly tested native Drill, Impala, and Spark connectors for data exploration.
* Developed diverse Python scripts for vulnerability assessment, including SQL injection, permission checks, and performance analysis.
* Proficiently orchestrated the import of data from various sources into HDFS using Sqoop, executed transformations using Hive and MapReduce, and subsequently loaded data into HDFS for further processing.

**Environment:** Microsoft SQL Server 2012, 2016, SSDT-2012 &2015, Azure Synapse Analytics, Azure Data Lake & BLOB, Azure SQL, Azure data factory, Azure analysis services, BIDS.

**Concentrix Inc, Hyderabad, India Nov 2013 to Oct 2016**

**Role: Data Analyst**

**Responsibilities:**

* Applied a range of data transformations, encompassing Lookup, Aggregate, Sort, Multicasting, Conditional Split, Derived Column, and more.
* Developed Mappings, Sessions, and Workflows to efficiently extract, validate, and transform data in compliance with business rules using Informatica.
* Tailored target tables based on reporting team requirements and formulated Extraction, Transformation, and Loading (ETL) processes utilizing Talend.
* Utilized Netezza SQL scripts to seamlessly transfer data between Netezza tables.
* Scheduled Talend Jobs using Job Conductor, a scheduling tool within the Talend ecosystem.
* Took charge of querying, stored procedure creation, crafting complex queries, and leveraging T-SQL joins to address varied reporting operations and handle ad-hoc data requests.
* Prioritized performance monitoring and optimized indexes using tools like Performance Monitor, SQL Profiler, Database Tuning Advisor, and Index Tuning Wizard.
* Acted as the primary contact for resolving locking, blocking, and performance-related issues. Authored scripts and devised indexing strategies for migrating data to Amazon Redshift from SQL Server and MySQL databases.
* Utilized AWS Data Pipeline to configure seamless data loads from S3 into Redshift.
* Employed JSON schema to define table and column mappings from S3 data to Redshift, devising indexing and data distribution strategies optimized for sub-second query response.
* Possessed hands-on experience with Dell Boomi Connectors, covering FTP, Mail, Database, Salesforce, Web Services Listener, HTTP Client, Web Services SOAP Client, Success Factors, and Trading Partner.
* Developed Database/Flat-file/JSON/XML profiles, Boomi Mappings, and Processes utilizing various connectors and logic shapes between application profiles and different Trading partners within Dell Boomi.

**Environment:** Amazon Redshift, AWS Data Pipeline, Talend Platform for Big Data MS SQL Server 2008R2/2012, Oracle 10g/9i, Dell Boomi, Netezza Mako 7.2, S3, SQL Server Reporting Services (SSRS), SQL Server Integration Services (SSIS), Share Point, TFS, MS Project, MS Access and Informatica.