Dheeraj T
Sr. Data Engineer
+1 845-320-3996
dheeraj.n9490@gmail.com

Professional Summary:

- Over 9 years of experience specializing in **GCP** and **AWS** cloud platforms, Hadoop ecosystems, and **ETL** pipeline development for scalable and efficient data processing solutions.
- Proficient in utilizing **AWS** services such as **S3**, **Glue**, **Redshift**, **Lambda**, DynamoDB, and **CloudWatch**.
- Led the design and implementation of scalable **ETL/ELT** pipelines for batch and real-time data processing.
- Managed data migrations from on-premises Oracle databases to cloud platforms like **GCP BigQuery** and **AWS Redshift**.
- Demonstrated expertise in distributed computing frameworks, including **Apache Spark**, **PySpark**, and **Spark SQL**.
- Adept in programming with **Python**, **Scala**, **Spark**, and **SQL** for data processing and analytics.
- Delivered real-time data streaming solutions using Google **Pub/Sub** and **Kafka**.
- Built scalable data lake and data warehouse architectures using **BigQuery** and **AWS Redshift**. Skilled in optimizing data warehousing and analytics using BigQuery.
- Orchestrated robust data workflows with tools like **Apache Airflow** and **AWS Glue** for seamless integration.
- Implemented row-level security in **BigQuery** Authorized Views and managed **IAM** policies to enforce secure access.
- Ensured compliance with data governance standards, including **GDPR** and **HIPAA** requirements.
- Delivered real-time and interactive data visualizations using **Tableau**, **Google Data Studio, Power BI**.
- Automated **CI/CD** pipelines using tools like **Jenkins**, **Git**, **Bitbucket**, and **Terraform** for streamlined deployments.
- Deployed scalable and resilient solutions using container orchestration platforms like **GKE**.
- Managed large-scale data storage and processing with HDFS, Teradata, and cloud-based data stores.
- Migrated and integrated datasets using tools such as **Sqoop**, **JDBC** for comprehensive analytics.
- Optimized query performance and cost efficiency in **GCP** and **AWS** environments with advanced tuning techniques.
- Developed and monitored scalable frameworks for efficient data pipeline execution and job tracking.
- Supported predictive analytics and machine learning workflows in Hadoop and cloud-based environments.
- Collaborated with cross-functional teams to deliver high-quality, business-critical data solutions.
- Documented **ETL** frameworks and data pipelines to ensure maintainability, scalability, and operational efficiency.

Skills and Tools:

Cloud Platforms	Google Cloud Platform (GCP), AWS (S3, Redshift, Glue, Lambda)
ETL/ELT Development	Apache Airflow, AWS Glue, Dataflow, Dataproc, Talend, Informatica PowerCenter
Programming Languages	Python, Scala, SQL, Java
Big Data Frameworks	Apache Spark, PySpark, Hadoop (HDFS, Hive, Pig, HBase, MapReduce)
Data Warehousing	Google BigQuery, AWS Redshift, Teradata
Real-Time Streaming	Google Pub/Sub, Kafka
Visualization Tools	Tableau, Google Data Studio, Power BI
CI/CD & DevOps	Jenkins, Git, Bitbucket, Terraform
Version Control	Git, SVN
Workflow Orchestration	Apache Airflow, Oozie
Database Management	Oracle, MySQL, DynamoDB, SQL Server, PostgreSQL
Data Governance	GDPR Compliance, HIPAA, BigQuery Authorized Views, IAM Policies
Build & Automation	Maven, Terraform
Machine Learning	Vertex AI
Other Tools	Sqoop, Flume, JDBC, Cloudera Manager

Professional Experience:

Client: Strive Health, Denver, CO **Oct 2022 - Present**
Job Role: Senior Data Engineer

Responsibilities:

- Migrated on-premises Oracle databases to **BigQuery** while ensuring data integrity and performance optimization.
- Designed and implemented a system to handle batch data transfers using Transfer Service and real-time data streaming with Google Pub/Sub, ensuring smooth and timely data ingestion into GCS.
- Developed incremental and full-load strategies for efficient data ingestion and processing across heterogeneous data sources.
- Utilized **GCP DataProc** for processing large-scale batch data using **Spark** and **Presto**.
- Implemented data pipelines using Google Cloud Dataflow to process and transform streaming data enabling real-time analytics and scalable data workflows.
- Leveraged SQL, Python (Pandas, NumPy), and PySpark to build data validation, transformation, and manipulation workflows, ensuring adherence to business logic.
- Loaded transformed and validated data into BigQuery for seamless integration of advanced analytics.
- Designed and implemented ETL pipelines using **Apache Airflow** and Composer for ingesting data from clinical trials, drug manufacturing, and supply chain systems into Google Cloud Storage and **BigQuery**.

- Configured and optimized **GCS** and **BigQuery** services using Cloud Shell SDK for seamless data storage and querying.
- Created **BigQuery** Authorized Views to enforce row-level security and facilitate secure data sharing with other teams.
- Configured alert-based monitoring for workflows using Airflow and **GCP** services like **Cloud Functions.**
- Conducted query performance tuning and optimized workflows based on usage statistics and logs.
- Developed custom reporting frameworks using **BigQuery** and connected them with visualization tools like Tableau to generate real-time insights into clinical trial progress and supply chain metrics.
- Created ad hoc reports and automated extract processes to meet the dynamic needs of stakeholders.
- Implemented secure access mechanisms by managing metadata and lineage with Google Data Catalog to maintain data governance and compliance standards.
- Optimized cloud costs through query analysis and usage monitoring via Google APIs and **BigQuery** billing insights.
- Built robust frameworks for data pipeline execution and job monitoring.
- Implemented security protocols and row-level permissions using **BigQuery** and **IAM** policies to comply with the strict regulatory standards of the pharmaceutical industry.
- Implemented exception handling and error reporting mechanisms in ETL pipelines to ensure data quality.
- Coordinated with cross-functional teams, including data scientists, business analysts, and QA teams, to deliver high-quality solutions.
- Documented ETL frameworks, pipelines, and processes to ensure scalability and ease of maintenance.
- Supported regulatory compliance efforts by ensuring the data platform adhered to **GDPR** and **HIPAA** requirements.

Environment: GCP (BigQuery, Dataproc, DataFlow, Data Transfer Service, GCS, Cloud Functions, Cloud Shell SDK, Data Catalog), Apache Airflow, Python (Pandas, NumPy), PySpark, SQL, Tableau, BigQuery, Oracle, Google Pub/Sub, Google APIs, Google Data Catalog.

Client: Deckers Brands, Goleta, CA **Feb 2020 – Sep 2022**
Job Role: Data Engineer

Responsibilities:

- Designed and implemented data integration pipelines using **Google Cloud Platform (GCP)** tools like Cloud Pub/Sub for seamless orchestration across diverse data sources.
- Developed high-performance ETL pipelines using **Python** Scripting, **Scala**, and **PySpark**, ensuring scalability and efficiency in data processing.
- Created workflows by using **Google Transfer Service** to ingest and integrate data from Teradata tables into **Google Cloud Storage** (GCS) and facilitating comprehensive analytics and reporting.
- Utilized **Sqoop** and **JDBC** connectors for efficient data ingestion and synchronization between on-premises and cloud-based systems.
- Unified disparate datasets using **Apache Spark-SQL** to enable advanced analytics and data transformation capabilities.
- Developed **Dataflow** pipelines to process and transform raw data from Google Cloud Storage (GCS) and seamlessly load it into **BigQuery** for efficient analytics and reporting.
- Implemented serverless data processing using **GKE**, ensuring scalability and cost-effective solutions for data integration.
- Utilized **Terraform** for managing **Google Kubernetes Engine (GKE)** clusters and resources.

- Optimized **BigQuery** data warehousing solutions, reducing query execution times by 40% and enhancing reporting performance.
- Developed interactive dashboards in **Google data studio** to visualize key business metrics and deliver insights to stakeholders.
- Automated daily ad-hoc reports and data extracts from **BigQuery**, supporting enterprise-level analytics and decision-making.
- Collaborated with the Data Science team to implement advanced analytics models using **Vertex AI** supporting predictive analytics and machine learning initiatives.
- Utilized version control tools such as **Git** for collaborative pipeline development, ensuring code consistency and version management.
- Automated data engineering workflows using Jenkins and **CI/CD** pipelines, enabling streamlined deployments and continuous integration.
- Deployed and monitored data pipelines, integrating services like Pub/Sub, **DataFlow**, and **BigQuery** for seamless operation.
- Implemented **Apache Airflow** for orchestrating complex **ETL** workflows.

Environment: Google Cloud Platform (GCP), Cloud Pub/Sub, Python, Scala, PySpark, Teradata, BigQuery, Apache Spark-SQL, Google DataFlow, Google Cloud Storage (GCS), GKE, Terraform, Google Data Studio, Vertex AI, Git, Sqoop, JDBC, Jenkins, CI/CD, Airflow.

Client: Comerica Bank, Dallas, TX **Jul 2019 – Jan 2020**
Job role: Data Engineer

Responsibilities:

- Designed and maintained scalable, secure, and cost-efficient S3-based data lakes for structured and unstructured data.
- Built and optimized ETL pipelines using **AWS Glue** to extract, transform, and load data into the target system **Redshift**, while leveraging **Kafka** for real-time data streaming to ensure low-latency data availability and seamless integration.
- Utilized **Python** and **Scala** to develop custom ETL scripts for data cleansing, validation, and transformation.
- Leveraged **Apache Spark** and **Spark SQL** to process large-scale datasets, ensuring efficient transformations and analytics.
- Implemented partitioning, compression, and parallelism techniques for performance optimization.
- Developed and scheduled data workflows using **Apache Airflow** to automate end-to-end processes.
- Integrated Airflow with upstream and downstream systems.
- Set up automated **CI/CD** pipelines in **GitLab** and **Jenkins** for code deployment, testing, and version control.
- Utilized Kafka for real-time data ingestion and processing, ensuring low-latency data availability for downstream systems.
- Designed fault-tolerant **Kafka** consumer applications to handle high-throughput data streams.
- Developed complex SQL queries in **Athena** and **Spark SQL** to enable business-critical data analytics.
- Created **Power BI** dashboards and reports to visualize key metrics and insights for stakeholders.
- Collaborated with business teams to gather requirements and ensure alignment with reporting goals.
- Monitored data workflows and health system using **AWS CloudWatch** and custom logging mechanisms.

- Diagnosed and resolved performance bottlenecks in Spark jobs, Glue ETL scripts, and Athena queries.
- Tuned AWS Glue jobs, Spark applications, and data queries for performance efficiency and cost reduction.
- Implemented best practices for data partitioning, indexing, and compression.
- Ensured data pipelines adhered to organizational security policies and regulatory compliance standards (e.g., SOC 2, GDPR).
- Managed data lifecycle policies, versioning, and access permissions using **AWS IAM** and bucket policies.
- Maintained detailed audit trails for data transformations and system operations using CloudTrail and Glue logs.
- Worked closely with data analysts, scientists, and business stakeholders to understand requirements and deliver solutions.
- Collaborated with DevOps teams to integrate data workflows with existing infrastructure.

Environment: AWS, S3, Glue, Athena, Python, Scala, Spark SQL, Airflow, Kafka, Redshift, Power BI, Gitlab, Jenkins, CI/CD pipelines.

Client: Unipoint IT Solutions Pvt Ltd, Hyderabad, India **Jan 2017 – Dec 2018**
Job role: ETL Developer

Responsibilities:

- Collaborated with users to gather business requirements and translated them into detailed technical specifications and design documents.
- Designed and implemented ETL workflows using tools such as **Informatica PowerCenter**, **Microsoft SSIS**, and **Talend**.
- Developed complex ETL mappings with Informatica PowerCenter to extract, transform, and load data.
- Created and optimized SSIS packages, incorporating transformations like lookups, merge joins, fuzzy lookups, and derived columns.
- Developed logging mechanisms at the package and task levels to track records processed by ETL workflows.
- Managed metadata, implemented data lineage, and supported governance practices using **Talend Metadata Manager**.
- Designed and implemented complex SQL objects such as stored procedures, triggers, functions, indexes, tables, and views.
- Conducted query optimization and applied advanced SQL techniques for performance improvement.
- Developed reports and data extracts from SQL Server using Excel pivot functionality to meet user requirements.
- Wrote unit testing code with frameworks like Unit Test and **PyTest**, integrating tests into build processes.
- Ensured high-quality ETL and database processes through rigorous testing and validation.
- Hands-on experience with ETL tools (Informatica PowerCenter, SSIS, Talend), **SQL Server**, and **Python** testing frameworks.

Environments: Informatica PowerCenter, Microsoft SQL Server Integration Services (SSIS), Talend, SQL Server, Excel (Pivot Tables and Data Extraction), Talend Metadata Manager, Python, Unit Test, PyTest, SQL Server Reporting Services (SSRS), Dataflow, and Java.

Amigos Software Solutions, Hyderabad, India **July 2015 – Jan 2017**
Job role: Hadoop developer

Responsibilities:

- Installed, configured, and maintained Apache Hadoop clusters and key components of the Hadoop ecosystem, including **Hive**, **Pig**, **HBase**, **Sqoop**, **Flume**, **Oozie**, and **Zookeeper**.
- Set up and managed a six-node CDH4 Hadoop cluster on CentOS.
- Imported and exported data between **HDFS**, **Hive**, and various **RDBMS** using Sqoop for seamless data integration.
- Defined job flows using Oozie to orchestrate multiple MapReduce and Pig jobs.
- Imported log files using **Flume** into **HDFS** and loaded them into Hive tables for querying and analysis.
- Scheduled downtime for non-production servers to optimize **AWS** pricing and reduce operational costs.
- Loaded data from **UNIX** file systems into **HDFS** for processing.
- Created Hive tables and implemented advanced features such as Partitioning, Dynamic Partitioning, and Bucketing to enhance data access performance.
- Developed MapReduce programs in Java for data extraction, transformation, and aggregation across various file formats, including XML, JSON, CSV, and compressed formats.
- Designed, developed, tuned, and maintained NoSQL databases, including data operations using **HBase Shell**, **HBase Client API**, **Pig**, and **Sqoop**.
- Experienced in writing the Python-based MapReduce programs using the Hadoop streaming API.
- Developed and executed unit test cases for Hadoop MapReduce jobs using MRUnit to ensure data processing accuracy.
- Continuously monitored and managed the Hadoop cluster through **Cloudera Manager** and Web UI for optimal performance.
- Collaborated with application teams to perform operating system and Hadoop updates, patches, and version upgrades as needed.
- Utilized **Maven** for building automation and SVN for version control.

Environments: Hadoop, MapReduce, HDFS, HBase, Hive, Impala, Pig, Java, SQL, Sqoop, Flume, Oozie, Unix, JavaScript, Maven, Eclipse.