## Lead GCP Data Engineer

* Highly motivateda IT professional with 9.5+ ysaearsaa of experience as a Cloud Engineer, specializing in Google Cloud Platform (GsCP)a, AWS, On-Preamise, and Data Analytics.
* Google Cloud certified Professional Daata Enginee7r.9
* Proven expertise in designing and developing daeta-intensive applications using GCP, AWS, Hadoop, Big Data Analytics, Data Warehousing, danad Data Visualization.
* Highly experienced Senior GCP Data Enginweer with a proven track record of seamlessly migrating on-premise ETLs to GCP using cloud-native technologies like Composer, BigQuery, Cloud Dataproc, and Cloud Fdunctions.
* Expert in ingesting various databases (Oracle, DB2, Teradata, MySQL, PostgreSQL,

MongoDB, etc.) to BigQuery and building efficient data pipelines using Apache Airflow for ETL tasks, driving seamless data integration and enabling advanced analytics capabilities on Google Cloud Platform.

* Successfully automated data ingestion, transformation, and validation processes, resulting in improved system efficiency and significant cost reductions.
* Demonstrated proficiency in Apache Beam and stream/batch data processing concepts, enabling efficient data transformations.
* Experienced in monitoring and alerting solutions using GCP Stackdriver, ensuring data pipeline stability.
* Developed comprehensive data quality assurance frameworks and integrated machine learning models with BigQuery and Cloud ML for advanced analytics.
* Experienced in orchestrating data injection from diverse sources into the BigQuery warehouse, resulting in enhanced data insights and improved accessibility to vital information sources.
* Expert in conducting root cause analysis for GCP services, identifying and addressing issues within the cloud infrastructure to ensure optimal performance and reliability.
* Expert in implementing and optimizing serverless computing on GCP, leveraging services like Cloud Functionsand Cloud Run to build scalable and cost-effective applications.
* Leveraged GCP's Stackdriver Logging and Monitoring tools to trace events, logs, and metrics, enabling swift identification of root causes.
* Played a key role in resolving critical incidents by meticulously analyzing logs from GCP services, pinpointing the root causes, and implementing effective solutions.
* Created comprehensive documentation outlining best practices for log analysis, root cause identification, and incident response, enhancing team knowledge and capabilities.
* Expert in implementing robust security measures on AWS, utilizing IAM, VPC, and AWS WAF to establish granular access controls, network isolation, and protection against potential cyber threats, maintaining the integrity and confidentiality of critical data.
* Expert in managing serverless applications on AWS, utilizing services like AWS Lambda, API Gateway, and DynamoDB, resulting in improved scalability, reduced operational overhead, and enhanced user experiences.
* Hands-on experience with Amazon EC2, S3, RDS, VPC, IAM, Elastic Load Balancing, Auto Scaling, CloudWatch, SNS, SES, SQS, Lambda, and EMR.
* Expertise in Informatica, enabling seamless data movement and transformation across diverse sources and targets.
* Instantiated, created, and maintained CI/CD (continuous integration & deployment) pipelines and apply automation to environments and applications.
* Seasoned DevOps Engineer with hands-on experience in building and automating CI/CD pipelines using tools like Jenkins, GitLab CI/CD, and AWS CodePipeline, accelerating software delivery and minimizing time-to-market.
* Expertise in creating, debugging, scheduling, and monitoring jobs using Airflow.
* Practical knowledge of data modeling concepts, including Star-Schema Modeling, Snowflake Schema Modeling, Fact, and Dimension tables.
* Expertise in the Data Warehousing tool known as Informatica, enabling me to proficiently design, execute, and manage complex data integration workflows, facilitating seamless data movement and transformation across diverse sources and targets.
* Hands on Bash scripting, Shell scripting experience and building data pipelines on Unix/ Linux systems.
* Collaborative team player with a commitment to continuous learning, actively sharing knowledge and best practices with colleagues to drive organizational growth.

Authorized to work in the US for any employer

# Work Experience

## Lead GCP Data Engineer

Kaiser Permanente

to Present

Responsibilities

* Successfully led ingesting data from various databases (MySQL, DB2, Oracle, Teradata,

PostgreSQL, MongoDB, etc.) to BigQuery, enabling scalable data storage and efficient querying on Google Cloud Platform (GCP).

* Developed a comprehensive data quality assurance framework for the batch processing environment, implementing data validation and error handling mechanisms to ensure high data integrity and accuracy.
* Leveraged GCP services, including Dataproc, GCS, Cloud Functions, and BigQuery to optimize data processing and analysis, achieving improved system efficiency and cost reductions.
* Developed data pipelines using Airflow to ingest data from various file based sources such as

(FTP, SFTP, API, Main Frame) in GCP for ETL-related jobs, utilizing various airflow operators to streamline data processing workflows.

* Leveraged the power of Airflow's scheduling capabilities to automate data pipelines, ensuring timely data updates and delivery.
* Implemented data encryption and access controls to ensure the security and privacy of sensitive data in transit and at rest within BigQuery.
* Integrated machine learning models with BigQuery and Cloud ML to enable advanced analytics and predictive insights, contributing to data-driven decision-making processes.
* Conducted performance tuning on BigQuery queries to optimize execution times and reduce query costs, resulting in faster data retrieval and lower expenses.
* Utilized PySpark in-memory processing capabilities to accelerate data computations and enhance data querying performance, leading to faster data transformation and analysis.
* Created custom monitoring and alerting solutions using Stackdriver and Cloud Monitoring to proactively identify and resolve data processing issues.
* Performed data validation and quality checks during the ETL process to ensure data consistency and integrity, identifying and resolving anomalies and discrepancies.
* Contributed to the development of a data catalog for organizing and discovering datasets, enhancing data discovery and reuse across the organization.
* Implemented serverless Cloud Functions to automate lightweight data processing tasks, enabling efficient routine data operations.
* Collaborated with data engineers and data scientists to design and implement complex data transformation logic.
* Coordinated with cross-functional teams, including database administrators and data analysts, to plan and execute a seamless migration strategy.
* Demonstrated commitment to continuous learning and staying up-to-date with the latest advancements in GCP services and data analytics technologies, actively sharing knowledge and best practices with team members to contribute to the overall growth of the organization.
* Led proof-of-concept projects for exploring new technologies and approaches, evaluating their potential impact on the existing data infrastructure, and providing recommendations for adoption.
* Conducted workshops and knowledge-sharing sessions for stakeholders to increase awareness and understanding of GCP capabilities and data-driven decision-making.
* Collaborated with external vendors and partners to leverage their expertise and solutions for specific data analytics requirements, fostering a network of valuable industry relationships.

Environment: Google Cloud Platform (GCP), BigQuery, Cloud Dataproc, Cloud Storage, Cloud Functions, Cloud ML, Data Catalog, Apache Spark, Airflow, CI/CD Pipelines (Continuous Integration and Deployment)

## Sr GCP Data Engineer

Healthfirst

April 2018 to January 2021

Responsibilities:

* Engineered ETL pipelines using Google Cloud Dataflow to efficiently ingest and transform data from diverse sources.
* Conducted data transformation and cleansing using SQL queries within the Dataflow environment.
* Leveraged GCP's native tools such as BigQuery, Dataproc, and Dataflow for ETL jobs, aligning technology choices with GCP's strengths.
* Designed and implemented data models and schemas using Google BigQuery to support efficient data processing and analysis.
* Developed real-time data streaming pipelines utilizing Pub/Sub and Google Cloud Dataflow, seamlessly integrating with GCP's for scalable storage and processing capabilities.
* Enabled seamless real-time data ingestion and processing for timely analytical insights.
* Conducted performance tuning for ETL workflows within Google Cloud Dataflow to enhance data processing speed and minimize latency.
* Optimized queries using BigQuery’s capabilities to improve overall ETL pipeline performance.
* Implemented data quality checks and data cleansing procedures within GCP using SQL queries and native data preparation tools.
* Utilized GCP's data profiling capabilities to identify and rectify data anomalies.
* Optimized cloud-based data warehouse infrastructure using GCP's partitioning and caching functionalities.
* Maintained comprehensive documentation of ETL processes, data pipelines, and data models within GCP's documentation tools.
* Utilized appropriate data structures to ensure data integrity and optimize query performance.
* Collaborated closely with data science teams within GCP's environment to understand data requirements and consumption patterns.
* Ensured compliance with industry regulations and adhered to GCP best practices for data protection.
* Implemented data access controls, encryption, and audit trails using GCP's built-in security features such as (Identity and Access Management (IAM), VPC, Audit logging, Cloud Key

Management).

* Collaborated with stakeholders to ensure accurate data acquisition and integration.
* Shared knowledge and best practices using GCP's collaboration features to foster a collaborative work environment.

Environment: Google Cloud Dataflow, Google BigQuery, Google Cloud Storage, Google Cloud Dataproc, Google Cloud Composer, Apache Kafka, SQL, Git, Continuous Integration/Continuous Deployment (CI/CD) Pipelines

## AWS Data Engineer

COGNIZANT TECHNOLOGY SOLUTIONS - Chennai, Tamil Nadu

December 2016 to March 2018

Responsibilities:

* Worked with Data Science team running Machine Learning models on Spark EMR cluster and delivered the data needs as per business requirements.
* Automated the process of transforming and ingesting terabytes of monthly data in Parquet format using Kinesis, S3, Lambda and Airflow.
* Loaded data into S3 buckets using AWS Glue and PySpark.
* Utilized Spark’s in memory capabilities to handle large datasets on S3 Data Lake.
* Developed Spark jobs on Databricks to perform tasks like data cleansing, data validation, standardization, and then applied transformations as per the use cases.
* Migrated Java analytical applications into Scala. Used Scala where performance and logic is critical.
* Created workflows using Airflow to automate the process of extracting weblogs into S3 Datalake .
* Involved in developing batch and stream processing applications that require functional pipelining using Spark Scala and Streaming API.
* Involved in extracting and enriching multiple Cassandra tables using joins in SparkSQL . Also converted Hive queries into Spark transformations.
* Hands-on experience on API design and development using Spring Boot for Data movement across different systems.
* Fetched live data from Oracle database using Spark Streaming and Amazon Kinesis using the feed from API Gateway REST service.
* Performed ETL operations using Python, SparkSQL, S3 and Redshift on terabytes of data to obtain customer insights.
* Performed interactive Analytics like cleansing, validation and quality checks on data stored in S3 buckets using AWS Athena.
* Wrote Python scripts to automate ETL pipeline and DAG workflows using Airflow. Manage communication between multiple services by distributing tasks on celery workers.
* Integrated applications using Apache tomcat servers on EC2 instances and automated data pipelines into AWS using Jenkins, git, maven and artifactory.
* Wrote unit tests, worked along with DevOps team in Installing libraries, Jenkins agents and productionized ETL jobs and microservices.
* Developed a custom-built Rest API to support real time customer analytics for data scientists and applications.
* Managed and deployed configurations for the entire datacenter infrastructure using Terraform.
* Experience with analytical reporting and facilitating data for Quicksight and Tableau dashboards.
* Used Git for version control and Jira for project management, tracking issues and bugs. Environment: AWS, EC2, S3, Athena, Lambda, Glue, Elasticsearch, RDS, DynamoDB, Redshift, ECS, Hadoop, Hive, Spark, Databricks, Python, Java, Scala, SQL, Sqoop, Kafka, Airflow, HBase, Oracle, Cassandra, MLlib, Quicksight, Tableau, Maven, Git, Jira.

## Hadoop Data Engineer

COGNIZANT TECHNOLOGY SOLUTIONS - Chennai, Tamil Nadu

September 2015 to December 2016

Responsibilities:

* Created Hive tables, loaded data, and wrote Hive queries running internally in MapReduce.
* Implemented Partitioning, Dynamic Partitions, and Buckets in HIVE for efficient data organization.
* Developed efficient data import and export procedures that enable smooth transfers of data to Hadoop Distributed File System (HDFS) using Sqoop.
* Improved Hive performance through query optimization, appropriate join strategies, and vectorization.
* Designed and implemented efficient data models and schemas in Hive to enhance data retrieval and query performance.
* Leveraged partitioning and bucketing strategies in Hive for better data management and faster processing.
* Performed complex data processing using Pig scripts, handling semi-structured and unstructured data in the Hadoop ecosystem.
* Built and tested Proof of Concepts for streaming applications with Kafka, enabling real-time data ingestion from multiple sources.
* Ensured data quality and accuracy by implementing data validation checks in Hive and Pig scripts.
* Performed data testing and validation to identify and rectify data anomalies.
* Collaborated with cross-functional teams, including data analysts, and business stakeholders to understand requirements and deliver effective data solutions.
* Designed and managed managed/external tables in HIVE as per project requirements.
* Worked on POC to migrate MapReduce jobs into Spark RDD transformations using Python.
* Managed code versioning using Git and participated in code reviews for quality assurance.
* Maintained detailed documentation of Hive and Pig scripts, Sqoop jobs, and Spark transformations, ensuring ease of maintenance and knowledge sharing.
* Shared knowledge and insights with the team, contributing to a more informed and skilled workforce. Environment: Apache Hadoop, HDFS, Hive, Pig, Sqoop, Spark, DB2, Apache Kafka, Git,

Python SQL.

**Data Analyst** LatentView Analytics August 2013 to July 2015

Responsibilities:

* Proficient in SQL, MYSQL, PL/SQL, Informatica, and Data Warehousing.
* Acted as a subject matter expert in Informatica development and provided support in troubleshooting ETL-related issues and incidents.
* Validated workflows and ensured data loading into target tables using SQL.
* Implemented data quality checks and conducted data validation throughout the ETL process.
* Proficient in Oracle SQL and experienced in Informatica Development, adept at crafting complex SQL queries for data retrieval and manipulation, as well as designing and implementing data integration workflows using the Informatica platform.
* Designed and developed ETL workflows to extract data from various sources, transform it, and load it into target tables.
* Created data mapping documents that define how data is transformed from source to target during the ETL process.
* Implemented complex transformation logic using Oracle SQL and Informatica Development, tailoring solutions to meet specific client requirements for efficient data processing and integration.
* Conducted performance tuning of Informatica mappings and workflows to improve data processing efficiency and reduce execution time. Optimized SQL queries for better database performance.
* Developed comprehensive test cases based on client requirements to validate data accuracy and completeness.
* Performed data validation during ETL execution and ensured the data meets the desired quality standards.
* Collaborated with the team to define data models and schema structures that support analytical reporting and data analysis.
* Implemented data governance practices to ensure data integrity, security, and compliance.
* Implemented access controls and data masking techniques to protect sensitive information.
* Scheduled and monitored Informatica ETL jobs to ensure timely data processing and loading.
* Performed job monitoring, error handling, and troubleshooting to ensure smooth execution.
* Worked closely with business users to gather and understand data requirements.
* Communicated with stakeholders to ensure the successful delivery of data solutions aligned with business needs.
* Utilized version control tools like Git for managing code changes and collaborated with the team to deploy ETL workflows across different environments.
* Implemented data quality checks and error handling mechanisms to identify and rectify data issues during the ETL process. Ensured data accuracy and consistency in the target system.
* Maintained detailed documentation of ETL workflows, mappings, and transformations. Followed change management processes to track and manage modifications. Environment: Oracle SQL, Informatica, Informatica PowerCenter, SQL Developer, Data Warehousing

# Education

## Bachelor of Technology in Electronics and Communication Engineering

SRM University - Chennai, Tamil Nadu

# Skills

* Cloud Functions
* IAM
* Apache sqoop
* Mac Tools : Apache Airflow
* Docker
* Google Cloud Certified Professional Data Engineer
* Firestore
* Java
* Cloud Run Big Data Technologies : Apache Spark
* Athena Operating Systems : Windows
* Apache kafka Programming Languages : C
* AWS SageMaker
* APP Engine
* Tableau
* Firebase
* GCP Cloud Storage
* MS Office
* Visual Studio Code
* AWS S3 (Amazon Simple Storage Service)
* DB2
* Vertica
* Cloud SQL
* PostgreSQL
* Apache Hbase
* AWS Redshift
* Python
* JavaScript
* SQL
* MySQL
* Cloud Dataproc
* Apache Hadoop
* Scala
* Cloud ML
* Unix
* Git (Version Control)
* AWS DynamoDB
* VPC
* DNS
* NoSQL
* C++
* SQLite
* Cloud Pub/Sub
* SQL Server Integration Services (SSIS)
* Kubernetes
* PL/SQL
* Apache HIVE
* Visual Basics
* MongoDB Databases : Oracle
* Apache Pig
* BigQuery
* Apache Spark SQL
* Terraform
* Advanced SQL
* AWS Lamda
* Informatica PowerCenter
* Microsoft SQL Server
* Tableau
* Linux
* Gold Badge for SQL (HackerRank)
* Google Cloud Platform
* AWS
* Cloud architecture
* Hadoop

# Certifications and Licenses

## Google Cloud Certified Professional Data Engineer

October 2023 to October 2025

## Gold Badge for SQL(Hackerrank)

Present