**Professional Summary:**

- Senior Data Engineer with 8+ years of experience building and operating data platforms across healthcare, banking, retail, and marketing domains in regulated and high-volume environments.
- Experienced in end-to-end data platform delivery, covering ingestion, processing, storage, modeling, orchestration, and analytics enablement across AWS and Google Cloud.
- Led multiple on-prem to cloud migrations, modernizing legacy Teradata, Oracle, SQL Server, and SAS workloads into scalable cloud-native architectures.
- Designed batch, streaming, and incremental ingestion pipelines, selecting appropriate patterns based on data freshness, volume, and downstream usage needs.
- Built distributed processing pipelines using Spark-based frameworks to cleanse, standardize, and enrich large datasets for analytics and operational reporting.
- Implemented layered data lake and warehouse designs, separating raw, refined, and curated data to improve governance, cost control, and reliability.
- Applied ELT modeling practices on cloud warehouses to produce analytics-ready datasets with validation and freshness checks for reporting teams.
- Supported near real-time and event-driven data flows for operational and customer-facing analytics without over-engineering streaming solutions.
- Orchestrated production pipelines using managed schedulers and workflow tools, ensuring predictable execution, dependency control, and recovery handling.
- Embedded data quality, validation, and reconciliation checks into pipelines to improve trust in analytics outputs and reduce downstream issues.
- Optimized pipeline performance and cloud costs through resource tuning, query optimization, and workload right-sizing across compute and storage layers.
- Worked closely with analytics, business, and product teams to align dataset structures with reporting, operational, and regulatory requirements.
- Prepared feature-ready datasets for analytics and experimentation while maintaining clear ownership boundaries with data science teams.
- Strong focus on security and compliance, applying least-privilege access, encryption, and audit-friendly practices in healthcare and financial environments.

**TECHNICAL SKILLS:**

| *Category* | **Tools & Technologies** |
|---|---|
| *Programming & Query Languages* | Python, SQL, Spark SQL, PySpark, Scala, Shell Scripting, SAS |
| *Data Ingestion* | AWS Glue, AWS Lambda, Google Cloud Functions, Sqoop, Secure File Transfers, Boto3 |
| *Streaming & Event Processing* | Apache Kafka, Amazon Kinesis, Amazon SQS, Amazon SNS, Google Pub/Sub |
| *Distributed Data Processing* | Apache Spark, Spark Streaming, Dataproc (GCP), Amazon EMR, Databricks |
| *Cloud Data Storage* | Amazon S3, Google Cloud Storage (GCS), HDFS |
| *Data Warehousing & Analytics Stores* | BigQuery, Amazon Redshift, Hive (LLAP), Snowflake |
| *ELT & Data Modeling* | dbt (BigQuery), Dataform (BigQuery), Delta Lake |
| *Workflow Orchestration* | Cloud Composer (Apache Airflow), Apache Airflow (EC2), AWS Glue Workflows |
| *Monitoring & Observability* | Google Cloud Monitoring, Google Cloud Logging |
| *Security & Governance* | AWS IAM, Google Cloud IAM, AWS KMS, CMEK (GCP), Kerberos |
| *DevOps & CI/CD* | Docker, GitHub, Git, Bitbucket, Bamboo, Cloud Build, Artifact Registry |
| *Infrastructure & Provisioning* | Terraform, AWS EC2 |
| *Analytics & Visualization* | Looker, Power BI, SAS Visual Analytics |
| *Data Formats & Architecture Patterns* | Avro, Parquet, JSON, Medallion Architecture, Layered Data Lake Design |
| *Databases & Operational Stores* | Oracle, SQL Server, Teradata, PostgreSQL, Amazon DynamoDB, Cassandra |
| *Operating Systems* | Linux (Ubuntu, Red Hat) |

**WORK EXPERIENCE:**

**Client: Tivity Health**, Chandler, AZ **Dec 2023 – Present**

**Role: Senior Data Engineer**

**Responsibilities:**

- Led the migration of healthcare data from on-premise Teradata and Oracle systems into Google Cloud, establishing a scalable and compliant cloud data foundation for enterprise analytics and reporting.
- Built batch and near-time data ingestion pipelines using PySpark jobs on Dataproc and Pub/Sub to support scalable and reliable data processing in the cloud.
- Implemented incremental ingestion pipelines using containerized Cloud Functions to detect and process data changes, enabling lightweight, event-driven data updates without full reloads.
- Re-engineered legacy SQL-based ETL workflows into standardized PySpark pipelines on Dataproc, improving consistency and maintainability across the data platform.
- Organized data using a layered storage approach in Google Cloud Storage and BigQuery, separating raw, refined, and curated datasets to support governance, cost control, and reliable downstream consumption.
- Published curated datasets to BigQuery using dbt to model analytics-ready tables with data quality and freshness checks for reporting workloads.
- Improved analytics performance and reduced cloud compute costs by optimizing BigQuery query patterns, table partitioning, and pipeline resource usage.
- Delivered consistently updated BigQuery datasets to support member journey analytics, healthcare engagement metrics, and operational reporting for clinical operations and business.
- Orchestrated data pipelines using Cloud Composer to enable reliable scheduling, dependency management, and end-to-end workflow coordination.
- Implemented automated data quality and consistency checks to ensure reliable and accurate datasets for downstream analytics and reporting.
- Prepared feature-ready, curated datasets in BigQuery to support downstream machine learning use cases and model training workflows, while also enabling analytics dashboards and visualizations in Looker for business and operational insights.
- Established centralized monitoring and alerting using Google Cloud Monitoring and Cloud Logging to proactively detect pipeline failures.
- Strengthened platform security by enforcing least-privilege access controls, securing service-to-service authentication, and applying CMEK-based encryption aligned with HIPAA requirements.
- Collaborated cross-functionally with analytics, product, infrastructure, and machine learning teams to align data platform delivery with business and technical requirements.

**Environment:** Python, SQL, cloud functions, Pub/Sub, Incremental Load, Docker, Dataproc, BigQuery, Google Cloud Storage (GCS), dbt, Cloud Composer, cloud monitoring, cloud logging, IAM, CMEK, Looker, Medallion Architecture.

**Client: Equifax**– St. Louis, Missouri **Jun 2022 – Dec 2023**

**Role: Senior Data Engineer**

**Responsibilities:**

- Led the migration of enterprise financial data platforms from on-premises Oracle and Teradata systems and external data sources, including secure file transfers and API-based extracts, into Google Cloud to support large-scale financial data modernization initiatives.
- Designed scalable batch and event-driven ingestion pipelines using Cloud Functions and containerized Apache Beam jobs on Dataflow, supporting consistent execution across environments.
- Ingested financial data from multiple source systems into Google Cloud Storage and BigQuery to support downstream processing and analytics.
- Implemented privacy-aware ingestion workflows to ensure sensitive financial data was securely managed during ingestion and processing in alignment with enterprise compliance requirements.
- Built distributed data transformation and enrichment pipelines by processing batch workloads using Spark on Dataproc and streaming workloads using Apache Beam–based Dataflow.
- Developed reusable transformation frameworks and standardized data schemas to integrate multiple enterprise data sources into consistent, analytics-ready datasets.
- Orchestrated end-to-end financial data pipelines using Cloud Composer to manage workflow dependencies and ensure reliable scheduling across ingestion and transformation processes.

- Automated build and deployment of data pipelines using Cloud Build and Artifact Registry, enabling controlled and repeatable releases across environments.
- Designed and implemented a layered data storage approach on Google Cloud Storage to organize raw and processed data in preparation for downstream analytics consumption.
- Published curated datasets to BigQuery and used Dataform to model analytics-ready tables with consistent transformations and data quality checks for financial reporting.
- Supported reporting and visualization use cases by providing dependable BigQuery datasets consumed by Looker and Power BI dashboards for finance and business stakeholders.
- Prepared feature-ready BigQuery datasets to support predictive analytics and experimentation workflows, collaborating with data science.

**Environment**: Python, SQL, cloud functions, Apache Beam, Google Cloud Dataflow, Docker, CI/CD, Cloud Build, GitHub, Dataproc, BigQuery, Google Cloud Storage (GCS), Dataform, Cloud Composer, Looker, Terraform, Medallion Architecture.

**Client: Costco IT** – Issaquah, Washington **Sep 2021 – Jun 2022**
**Role: Senior Data Engineer**

**Responsibilities:**
- Led ingestion of enterprise retail data from transactional systems, web logs, and operational sources, including order, inventory, and sales activity data.
- Designed scalable batch ingestion pipelines using AWS Glue to ingest retail data from multiple source systems into Amazon S3.
- Implemented incremental and near-real-time ingestion using AWS Lambda with event-based triggers via Amazon SQS and Amazon SNS.
- Built distributed data processing pipelines on Amazon EMR using Apache Spark and Spark SQL to cleanse, transform, and enrich large-scale retail datasets.
- Implemented a centralized retail data lake on Amazon S3 to store raw and processed datasets in an analytics-ready structure.
- Enabled ad-hoc analytics using Amazon Athena and published selected curated datasets to Amazon Redshift to support structured reporting workloads.
- Prepared cleansed and feature-ready retail datasets to support downstream analytics and machine learning experimentation without owning model development.
- Orchestrated and monitored data pipelines using Apache Airflow on EC2, ensuring reliable scheduling, dependency management, and timely data availability.
- Optimized pipeline performance and cost by tuning EMR configurations with EC2 Spot Instances and enforced secure access using AWS IAM.
- Worked closely with retail analytics and business teams to validate data outputs and align dataset structures with reporting and operational use cases.

**Environment:** Python, SQL, AWS EMR, Amazon S3, AWS Glue, AWS Lambda, Amazon Athena, Apache Spark, PySpark, Apache Airflow (EC2), EC2 Spot Instances, IAM, Amazon Redshift, GitHub, Linux (Ubuntu, Red Hat).

**Client: M&T Bank -** Buffalo, NY **Oct 2020 – Sep 2021**
**Senior Data Engineer**

**Responsibilities:**
- Worked on a centralized banking data platform supporting account activity, payment events, and operational reporting by ingesting data from transactional systems, event streams, and internal banking applications into AWS.
- Designed scalable batch and streaming ingestion pipelines using AWS Lambda, Amazon S3, Amazon SQS, Amazon Kinesis, and Boto3 to ingest banking data from multiple source systems with reliable event handling.
- Built data transformation pipelines using Apache Spark on Databricks and Spark SQL to cleanse, normalize, and enrich banking datasets for downstream consumption.

- Implemented Delta Lake on Amazon S3 to manage curated datasets with ACID guarantees, supporting upserts, schema evolution, and reliable handling of late-arriving banking events.
- Implemented a centralized banking data lake on Amazon S3 to store raw and processed datasets in an analytics-ready structure.
- Enabled enterprise analytics by querying partitioned datasets using Amazon Athena and publishing curated datasets to Amazon Redshift for supporting reporting and operational insights for business and risk teams.
- Prepared cleansed and feature-ready datasets to support downstream analytics and machine learning experimentation, without owning model development.
- Orchestrated and maintained reliable data pipelines using AWS Glue, AWS Lambda, and Amazon EC2, ensuring consistent execution and timely data availability.
- Established CI/CD pipelines using GitHub and AWS deployment tooling to automate build, validation, and promotion of Databricks jobs and ingestion workflows across environments.
- Optimized pipeline performance and cost efficiency using AWS Cost Explorer, AWS Trusted Advisor, and efficient resource utilization strategies across banking workloads.
- Ensured secure and compliant data processing by applying AWS IAM access controls, AWS KMS encryption standards, and collaborating with analytics, compliance, and business teams to align platform delivery with enterprise requirements.

**Environment**: AWS S3, AWS Lambda, Boto3, Amazon SQS, Amazon SNS, Amazon Kinesis, AWS Glue, AWS Glue Crawlers, Amazon Athena, Amazon Redshift, Apache Spark, Spark SQL, Databricks, Delta Lake, Amazon DynamoDB, AWS IAM, AWS KMS, AWS Cost Explorer, AWS Trusted Advisor, PostgreSQL, Kafka, EC2, Linux, GitHub, CI/CD Pipelines.

**Client: D4 Insight Pvt Ltd – Bengaluru, India** **Jul 2018 – Sep 2020**
**Big Data Engineer – Retail Analytics & Customer Insights**

**Responsibilities:**
- Migrated legacy healthcare data processing workloads from on-premises relational systems to an AWS-based big data platform, modernizing SQL and SAS pipelines for scalable analytics.
- Designed ingestion workflows using Sqoop, AWS Lambda, and secure file transfer and Amazon S3 as an initial landing zone to ingest clinical, claims, and operational healthcare data into distributed storage systems.
- Built scalable batch and streaming ingestion pipelines using Apache Kafka to capture healthcare events and transactional data from multiple source systems.
- Developed distributed data processing pipelines using Apache Spark, Spark Streaming, and Spark SQL to cleanse, standardize, and enrich large healthcare datasets.
- Implemented basic data validation and reconciliation checks within Spark pipelines to ensure record completeness and consistency before datasets were published for analytics.
- Integrated streaming pipelines by connecting Kafka with Spark Streaming and exporting processed data into HDFS and Hive for downstream analytics availability.
- Implemented a structured healthcare data lake using HDFS, Hive, and cloud-based data lake patterns, leveraging Avro and Parquet formats with schema evolution support.
- Enabled analytics consumption by publishing curated datasets and metadata in Hive and Snowflake, supporting reporting and operational analytics use cases.
- Supported analytics and visualization by delivering prepared datasets for SAS Visual Analytics and Power BI, enabling insights into healthcare utilization and operational trends.
- Prepared feature-ready datasets to support predictive analytics and statistical modeling workflows using SAS and Python, without owning model deployment.
- Automated workflow orchestration and operational execution using Oozie, Zookeeper, and shell scripting, ensuring reliable and repeatable healthcare data processing.
- Ensured secure handling of sensitive healthcare data by applying access controls, encryption standards, and compliance-aligned practices, while collaborating with analytics, data science, and healthcare stakeholders in an Agile environment.

**Environment**: AWS EC2, Amazon S3, AWS Lambda, AWS Glue, CI/CD, Apache Spark, Spark SQL, Spark Streaming, PySpark, Scala, Hive (LLAP), HDFS, Kafka, Sqoop, Oozie, Zookeeper, Snowflake, SAS, Python, Avro, Parquet, Power BI, Shell Scripting, Git, Bitbucket, Bamboo, Linux (Red Hat, Ubuntu).

**Client: Vsplash Techlabs Pvt Ltd, India** **May 2017 – Jul 2018**
**Hadoop Engineer**

**Responsibilities:**

- Supported ingestion of digital marketing and web interaction data from relational databases, flat files, APIs, and streaming sources capturing user activity and campaign engagement events.
- Built batch ingestion pipelines using Sqoop, UNIX Shell Scripting, and SAS to extract marketing and web data from enterprise databases into distributed storage.
- Implemented streaming ingestion using Apache Kafka, Spark Streaming, and Zookeeper to capture and process real-time web and event data at scale.
- Converted legacy Oracle SQL/PL-SQL and SQL Server T-SQL logic into distributed processing workflows using Apache Spark, PySpark, Scala, and Spark SQL.
- Developed large-scale data processing pipelines on Hadoop (Cloudera) to cleanse, aggregate, and enrich digital marketing datasets for downstream consumption.
- Implemented basic data validation and reconciliation checks within Spark jobs to ensure data completeness and consistency before datasets were consumed by analytics teams.
- Stored raw and processed datasets in HDFS, leveraging Avro and Parquet formats with schema evolution to support flexible analytics use cases.
- Enabled analytics and reporting by preparing curated datasets in Hive and Snowflake, supporting campaign performance analysis and user behavior insights.
- Supported advanced analytics by delivering prepared datasets for SAS Visual Analytics and Power BI and assisting data science teams with feature-ready data using SAS Visual Statistics.
- Automated and monitored end-to-end workflows using Oozie, Shell Scripts, and AWS CI/CD pipelines, enforced secure access with Kerberos, JDBC, and role-based controls, and collaborated with marketing and analytics teams to align delivery with campaign goals.

**Environment**: Hadoop (Cloudera), HDFS, Hive (LLAP), Impala, Spark, Spark SQL, Spark Streaming, Scala, PySpark, Python, SAS, SAS Visual Analytics, SAS Visual Statistics, Kafka, Zookeeper, Sqoop, Oozie, Avro, Parquet, UNIX Shell Scripting, Snowflake, Power BI, Git, Bitbucket, Bamboo, SonarQube, DB2, Oracle, SQL Server, Linux.

**EDUCATION:**
Bachelor of Technology in **Electronics and Communication Engineering** at MLR Institute of Technology, India.