**Senior Data Engineer**

**Phone:**

**E-mail:**

**LinkedIn:**

**PROFESSIONAL SUMMARY :**

Results-driven Data Engineer with 10+ years of experience in designing and implementing scalable, cloud-native data pipelines across AWS, Azure, and GCP. Expertise in Big Data Analytics, Data Warehousing, Real-Time Streaming, AI/ML-driven Data Solutions, and Data Quality frameworks. Proven ability to transform complex datasets into actionable insights, leveraging cutting-edge cloud technologies.

* Designed and optimized cloud-native data pipelines across AWS, Azure, and GCP, ensuring high-performance data ingestion, transformation, and analytics.
* Developed and fine-tuned ETL workflows using Apache Spark (PySpark, Scala), Airflow, AWS Glue, Azure Data Factory, and GCP Dataflow, enhancing data processing efficiency.
* Engineered real-time data streaming architectures with Apache Kafka, AWS Kinesis, Azure Event Hubs, and Spark Streaming, enabling low-latency event-driven processing.
* Built and optimized data warehousing solutions on Amazon Redshift, Azure Synapse, and Google BigQuery, implementing Dimensional Data Modeling (Star & Snowflake Schema).
* Managed multi-cloud data lake architectures using AWS S3, Azure Data Lake, and GCP Cloud Storage, ensuring scalable and cost-effective storage solutions.
* Designed high-performance SQL and NoSQL databases using PostgreSQL, MySQL, SQL Server, DynamoDB, Cosmos DB, and MongoDB, optimizing queries and transaction throughput.
* Automated workflow orchestration with Apache Airflow, AWS Step Functions, and GCP Workflows, ensuring fault-tolerant and scalable data pipelines.
* Implemented AI/ML-driven data processing with Azure ML, AWS SageMaker, Google Vertex AI, TensorFlow, and PyTorch, integrating predictive analytics into pipelines.
* Enhanced distributed data processing performance by optimizing Spark jobs, SQL queries, and ETL pipelines, reducing execution time and cloud costs.
* Secured data pipelines with IAM policies, encryption (AWS KMS, Azure Key Vault, GCP KMS), and GDPR/CCPA compliance, ensuring regulatory adherence.

| **Category** | **Technologies & Tools** |
| --- | --- |
| **Hadoop Ecosystem** | HDFS, YARN, MapReduce, Hive, Sqoop, Spark (1.x/3.x), PySpark, Scala, Pig, Storm, Zookeeper, Oozie, Kafka, Flume, T-SQL |
| **Programming Languages** | Python (3.x), Java (8/11+), Scala, PySpark, Shell Scripting, T-SQL, SQL, R, SAS, MATLAB |
| **Big Data Platforms** | Azure Data Lake, Azure Synapse Analytics, Azure Databricks, AWS Redshift, Google BigQuery, Snowflake, Teradata, IBM Netezza (Banking Risk Modeling), Cloudera (Energy Analytics), Palantir Foundry (Oil & Gas) |
| **Operating Systems** | Linux (RHEL, Ubuntu, CentOS), Windows, UNIX |
| **Databases** | Azure SQL Database, Azure Cosmos DB, Oracle (10g-19c), MySQL, PostgreSQL, MongoDB, Cassandra, DB2, Netezza, SQL Server, Amazon Aurora, HDFS-based storage (genomic data), SCADA databases, SAP HANA (Banking, Oil & Gas), Sybase (Financial) |
| **Development Methods** | Agile (Scrum, SAFe), Waterfall |
| **IDEs & Tools** | PyCharm, IntelliJ, VS Code, Ambari |
| **Data Visualization** | Tableau, Power BI, Looker, Grafana, Splunk (Oil & Gas monitoring), Bloomberg Terminal (Finance), QlikView, IBM Cognos, SAP BusinessObjects |
| **Azure Cloud Technologies** | Azure Data Factory, Azure Data Lake, Azure Databricks, Azure SQL Database, Azure Synapse Analytics, Active Directory, Application Insights, Azure Monitoring, Azure Search, Key Vault |
| **AWS Cloud Technologies** | AWS Glue, AWS Redshift, Amazon RDS, Amazon Aurora, AWS Kinesis, AWS Lambda, Amazon S3, AWS EMR (Spark, Hive, Presto), AWS SageMaker, AWS Step Functions, AWS Lake Formation |
| **GCP Cloud Technologies** | Google BigQuery, Google Cloud Dataflow, Google Cloud Pub/Sub, Google Cloud Storage (GCS), Google Dataproc, Google AI Platform, Google Data Studio, Google Vertex AI |
| **Streaming & Messaging** | Apache Kafka, Azure Event Hubs, AWS Kinesis, RabbitMQ, ActiveMQ, Google Pub/Sub, IBM MQ, Azure IoT Hub, TIBCO EMS (Banking & Finance) |
| **Data Science & Analytics** | Azure ML, Databricks MLFlow, AWS SageMaker, TensorFlow, PyTorch, SAS, R, Hadoop Mahout, IBM Watson, MATLAB, Predictive Maintenance Models (Oil & Gas) |
| **ETL & Data Integration** | Azure Data Factory, Talend, Informatica PowerCenter, Apache NiFi, SSIS, Apache Beam, Fivetran, IBM DataStage, MuleSoft, Trifacta (Oil & Gas Data Prep), Informatica BDM (Banking & Finance) |

**PROFESSIONAL EXPERIENCE :**

**Client: Western Alliance Bank Jun 2023 to Present**

**Role: Sr. Data Engineer**

**Responsibilities:**

* Designed and implemented scalable ETL pipelines for ingesting and processing banking transactions, loan applications, mortgage data, and credit card transactions using Azure Data Factory, Google Cloud Dataflow, AWS Glue, and Apache Airflow.
* Developed real-time fraud detection and risk assessment workflows in Databricks (Azure/GCP/AWS), leveraging Spark SQL and Snowflake to analyze transaction patterns, detect anomalies, and prevent financial fraud.
* Engineered ETL workflows to integrate core banking data, loan approvals, credit risk models, and investment portfolios, ensuring seamless data movement across financial institutions and regulatory bodies using Databricks Delta Lake, Snowflake Streams, and Change Data Capture (CDC).
* Optimized high-volume banking data processing using partitioned tables, Hive scripts, and Delta Lake, enabling faster credit risk modeling and real-time trade settlements across multi-cloud environments (Azure, AWS, and GCP).
* Configured and optimized Apache Spark and Snowpark jobs for processing loan applications, wealth management accounts, investment funds, and anti-money laundering analytics, reducing latency in risk assessments.
* Built and managed Databricks clusters on Azure, AWS, and GCP to support high-frequency financial transactions, regulatory compliance tracking, and portfolio risk analysis.
* Developed complex SQL and PL/SQL scripts for bank reconciliation, mortgage calculations, interest rate adjustments, and financial audits, ensuring accurate reporting and compliance across multi-cloud data warehouses.
* Created scalable ETL pipelines for ingesting structured and semi-structured financial data (JSON, XML, XBRL, FIX Protocol) into Snowflake, BigQuery, Redshift, and Databricks Delta Lake, supporting core banking, risk management, and customer analytics.
* Collaborated with financial analysts, credit risk officers, and compliance teams to define banking data requirements, ensuring adherence to multi-cloud regulatory frameworks such as SOX, Basel III, IFRS, and AML guidelines.
* Designed and managed NoSQL and distributed storage solutions, including HBase, Bigtable, DynamoDB, and Firestore, to store multi-format financial datasets, including loan repayment schedules, customer credit scores, trade execution records, and regulatory filings.

**Environment: Databricks (Azure/AWS/GCP), Snowflake, Apache Spark, Delta Lake, Snowpark, Apache Airflow, Azure Data Factory, AWS Glue, Google Cloud Dataflow, Azure Synapse Analytics, BigQuery, Redshift, Snowpipe, Kafka, Pub/Sub, Kinesis, Cloud Storage (Azure Data Lake, S3, GCS), Vertex AI, Azure ML, AWS Sagemaker.**

**Client:** **Stellantis, Auburn Hills, MI. Aug 2021 to May 2023**

**Role: Sr. Data Engineer**

**Responsibilities:**

* Collaborated with automotive engineers, supply chain analysts, and compliance teams to gather and define ETL strategies for vehicle production, supply chain logistics, IoT sensor data, and regulatory compliance (ISO 26262, GDPR, NHTSA).
* Designed and implemented scalable ETL pipelines for ingesting vehicle telematics, manufacturing process data, warranty claims, and supplier quality reports into Amazon S3, Azure Data Lake, Google Cloud Storage (GCS), Snowflake, and Redshift, using AWS Glue, Azure Data Factory, Google Dataflow, and Apache Airflow.
* Developed data workflows on Databricks and Snowflake, leveraging Spark SQL, Snowpark, and BigQuery, ensuring accurate and efficient processing of vehicle diagnostics, predictive maintenance, and parts inventory.
* Engineered ETL pipelines for integrating production line data, supplier performance metrics, and automotive recalls, ensuring seamless data movement across OEMs, Tier 1/2 suppliers, and dealerships using AWS Step Functions, Google Composer, and Azure Data Factory.
* Optimized performance for large-scale automobile manufacturing datasets using partitioned and clustered tables in Snowflake and BigQuery, integrating Hive scripts on Dataproc for production analytics.
* Configured and tuned Spark and Snowflake Snowpark jobs for real-time monitoring of factory sensor data, autonomous vehicle telemetry, and fleet performance analytics.
* Built and managed Databricks clusters on Azure, AWS, and GCP to support high-volume processing of IoT sensor logs, production cycle times, and emissions compliance data.
* Developed complex SQL, PL/pgSQL, and Snowflake procedures for manufacturing defect analysis, supplier scorecards, and vehicle lifecycle tracking, ensuring accurate reporting and compliance with automotive regulations.
* Created and optimized ETL workflows to process structured and semi-structured automotive data (CAN bus logs, IoT telemetry, JSON, XML), ensuring secure and standardized integration into manufacturing execution systems (MES) using Snowflake, AWS Glue, and Cloud Functions.
* Designed and managed NoSQL and distributed storage solutions including Bigtable, DynamoDB, and Firestore to store and retrieve multi-format datasets, including vehicle diagnostics, production cycle logs, and real-time fleet tracking data.

**Environment:** **Databricks (Azure/AWS/GCP), Snowflake, Apache Spark, Snowpark, Delta Lake, Apache Airflow, AWS Glue, Azure Data Factory, Google Cloud Dataflow, BigQuery, Redshift, Snowpipe, Kafka, Pub/Sub, Kinesis, Cloud Storage (Azure Data Lake, S3, GCS), Vertex AI, AWS Sagemaker, Azure ML.**

**Client: XOOM Energy, Troy, MI Apr 2020 – Jul 2021**

**Role: Data Engineer**

**Responsibilities:**

* Transformed raw exploration, drilling, and production data into optimized sequence formats such as Avro and Parquet, improving data processing efficiency and seismic data transfer across oilfield networks.
* Designed and implemented end-to-end data pipelines on AWS S3, Azure Data Lake, AWS Glue, and Azure Synapse Analytics to enable efficient data ingestion, reservoir modeling, and real-time analytics for oil and gas operations.
* Applied normalization and de-normalization techniques to optimize performance in relational and dimensional models for well logs, production forecasts, and reservoir simulations.
* Tuned SQL queries in AWS Redshift and Azure Synapse Analytics and optimized geospatial data processing using AWS Glue, Azure Data Factory, and Azure Databricks.
* Leveraged AWS EMR and Azure Databricks with Spark to optimize seismic data processing, drilling efficiency analysis, and predictive maintenance algorithms using Spark SQL, DataFrames, and RDDs.
* Developed Python scripts with PySpark for real-time sensor data processing, drilling logs, and production monitoring on AWS Glue and Azure Databricks.
* Migrated HiveQL queries into Spark transformations using RDDs and Scala, improving processing efficiency for well performance metrics, oil production forecasting, and equipment monitoring on AWS EMR and Azure Synapse Analytics.
* Developed APIs for generating XML documents from oilfield databases, integrating XML and XSL transformations for regulatory compliance using AWS Lambda and Azure Functions.
* Created custom User Defined Functions (UDFs) and User Defined Aggregated Functions (UDAFs) in AWS Redshift and Azure Synapse Analytics to extend query capabilities for reservoir analysis and predictive maintenance.
* Built ETL workflows using Spark on AWS Glue and Azure Data Factory to clean, transform, and map large-scale geological and production datasets, ensuring high data accuracy for asset management.
* Developed Kafka custom encoders for real-time streaming of IoT sensor data from drilling rigs, ensuring secure ingestion into Amazon Kinesis, AWS MSK (Managed Kafka), Azure Event Hubs, and Kafka partitions.
* Automated data ingestion workflows using AWS Kinesis and Azure Event Hubs, ensuring secure and real-time streaming of pipeline flow rates, equipment sensor readings, and refinery process data.

**Environment**: Python, HDFS, Spark, Flume, Kafka, Zookeeper, Pig, Hive, HQL, HBase, Scala, ETL, Linux Red Hat, RESTful, SOAP, Log4J, Cloud Formation, Axis 1.2, JAX-WS, SOA.

**Client: State Of WI Aug 2018 – Mar 2020**

**Role: Data Engineer**

**Responsibilities:**

* Designed and implemented scalable ETL pipelines for ingesting and processing policyholder data, claims processing, underwriting models, actuarial risk assessments, and fraud detection into multi-cloud storage (Azure Data Lake, Google Cloud Storage, AWS S3) and data warehouses (Snowflake, BigQuery, Redshift) using Azure Data Factory, Google Cloud Data Fusion, AWS Glue, and Apache Airflow for orchestration.
* Developed data workflows on Databricks, Dataproc, and Snowflake Streams, leveraging Spark SQL, Snowpark, and BigQuery to process insurance claims, policy renewals, fraud analytics, and risk modeling efficiently.
* Engineered ETL pipelines for integrating claims adjudication, reinsurance calculations, portfolio risk analysis, and policy premium modeling, ensuring seamless data movement between insurance carriers, brokers, regulatory agencies, and third-party data providers using Cloud Composer, Dataflow, and Snowpipe for real-time ingestion.
* Optimized performance for large-scale insurance and reinsurance datasets by leveraging partitioned and clustered tables in Snowflake, BigQuery, and Redshift, while integrating Hive scripts on Dataproc to support actuarial computations and regulatory reporting (NAIC, IFRS 17, Solvency II).
* Configured and optimized Apache Spark and Hive jobs on multi-cloud platforms (Azure Databricks, Google Dataproc, AWS EMR) to enable low-latency processing for real-time claims fraud detection, policy underwriting, loss ratio analysis, and risk exposure assessments.
* Built and managed high-performance data processing clusters on Databricks, Dataproc, and AWS EMR, integrated with Snowflake, to support high-volume policy management, customer risk profiling, reinsurance settlements, and financial audits.
* Developed complex SQL, PL/pgSQL, and Snowflake procedural scripts for data transformation, stored procedures, and materialized views, supporting claims payouts, premium calculations, reserve estimates, and actuarial model validations.
* Created scalable ETL workflows to process structured and semi-structured insurance data formats (ACORD, JSON, XML, EDI, IFRS 17, XBRL), ensuring standardized data integration into policy administration systems, claims management platforms, and regulatory compliance frameworks using multi-cloud tools and Snowflake ingestion features.
* Collaborated with actuarial teams, underwriters, claims adjusters, and compliance officers to define data-driven risk models, ensuring adherence to regulatory requirements such as IFRS 17, Solvency II, NAIC, AML, and GDPR.

**Environment**: Azure Data Factory, Azure Databricks, Google Cloud Data Fusion, Cloud Dataflow, Google Dataproc, AWS Glue, Snowflake, BigQuery, Redshift, Cloud Storage (Azure, GCS, AWS S3), Apache Spark, Hive, Kafka, Pub/Sub, Terraform, Cloud Composer, Azure DevOps, Vertex AI, AWS SageMaker.

**Client: Razorpay, India. Apr 2015 - Sep 2017**

**Role: Big Data Engineer**

**Responsibilities:**

* Developed scalable big data solutions, integrating structured and unstructured data from enterprise applications, microservices, and SaaS platforms across multi-cloud and on-premise environments.
* Designed and implemented data pipelines for processing high-volume transactional data using Apache Spark, Hive, and Snowflake, ensuring efficient transformations, validation, and optimized storage for analytics and reporting.
* Built real-time data streaming solutions using Apache Kafka, Spark Streaming, and Flink, enabling low-latency processing of IoT telemetry, financial transactions, and log analytics.
* Optimized cloud and on-premise big data environments, implementing partitioning, bucketing, and indexing in Hive, HBase, and Snowflake to improve query performance and data retrieval speeds.
* Developed data lake architectures using AWS S3, Azure Data Lake, and Google Cloud Storage, ensuring scalable, cost-effective data storage for analytical workloads.
* Engineered ETL pipelines for log ingestion using Kafka, Flume, and Airflow, transforming application logs, telemetry data, and system metrics into structured datasets for real-time monitoring and insights.
* Automated ETL testing and data validation frameworks using PySpark, dbt, and Great Expectations, ensuring data integrity, completeness, and consistency across development, staging, and production environments.
* Utilized Spark SQL, Databricks, and Hive for data cleansing, enrichment, and transformation, supporting business intelligence, machine learning, and predictive analytics applications.
* Implemented workflow orchestration using Apache Airflow, Prefect, and Dagster, ensuring automated and timely execution of data pipelines for critical business functions.
* Developed API-based data integrations using Spark, REST APIs, and GraphQL, enabling seamless data exchange between enterprise SaaS applications, CRM platforms, and cloud-based data warehouses.
* Designed and implemented data warehouse and data lakehouse solutions using Snowflake, Redshift, and BigQuery, optimizing data storage and retrieval for scalable analytics and AI-driven insights.

**Environment**: Apache Spark, PySpark, Databricks, Apache Hive, Snowflake, Redshift, BigQuery, Apache Kafka, Apache Flume, Apache Airflow, AWS S3, Azure Data Lake, Google Cloud Storage, Cassandra, SQL Server, Oracle, PostgreSQL, DB2, Hadoop, Apache Iceberg, Delta Lake, Terraform, Python, Java.

**Client: Freshworks, India. Jun 2013 - Mar 2015**

**Role: Hadoop Engineer**

**Responsibilities:**

* Developed scalable ETL solutions using Hadoop ecosystem tools, integrating structured and unstructured data from diverse enterprise applications, microservices, and SaaS platforms.
* Designed and implemented ETL workflows for processing high-volume transactional data using Apache Spark, Hive, and Sqoop, ensuring efficient data transformations, validation, and storage for large-scale analytics and reporting.
* Collaborated with software developers, business analysts, and product owners to understand data processing needs, ensuring seamless integration between application databases and Hadoop-based data lakes.
* Developed real-time ETL processing pipelines using Apache Kafka, Spark Streaming, and Flume, enabling real-time data ingestion and processing between software applications and distributed databases.
* Optimized Hadoop-based database performance by implementing partitioning, bucketing, and indexing techniques in Hive and HBase, improving data retrieval speeds for large-scale analytics.
* Integrated ETL processes with cloud-based data storage solutions, leveraging HDFS, AWS S3, and Google Cloud Storage to support scalable data pipelines.
* Designed ETL pipelines for log data ingestion using Flume and Kafka, transforming application logs and system metrics into structured datasets for real-time monitoring and debugging.
* Implemented automated ETL testing frameworks in Hadoop, ensuring data accuracy, consistency, and completeness across development, testing, and production environments.
* Utilized Apache Pig, Spark SQL, and Hive for data cleansing, transformation, and enrichment, ensuring high-quality input for business intelligence applications.
* Implemented workflow orchestration using Apache Oozie and Airflow, ensuring timely execution of ETL jobs for critical enterprise functions.

**Environment:** Hadoop, Apache Spark, PySpark, Apache Hive, HDFS, Apache HBase, Apache Kafka, Apache Flume, Apache Sqoop, Apache Oozie, Apache Zookeeper, MapReduce, Apache Pig, Airflow, AWS S3, Google Cloud Storage, Cassandra, SQL Server, Oracle, DB2, Python, Java.