**Vikas**

**Senior Data Engineer**

**Phone: (732) 347-6970**

**Professional Summary:**

* 10+ years of experience as Data Engineer with strong expertise in design, development and maintenance of enterprise analytical solutions using Big Data Technologies.
* Strong expertise in ETL/ELT development, data pipeline orchestration, and real-time/batch processing using AWS, GCP and Azure.
* Designed and implemented data lakes, lake houses and warehouses using Snowflake, Redshift, BigQuery and Synapse Analytics.
* Hands-on with Python, PySpark, SQL, Scala and frameworks like Airflow, dbt and Terraform for scalable data transformation.
* Proficient in data modeling techniques including Data Vault 2.0, Star and Snowflake schemas supporting complex master data structures.
* Built secure, high-performance pipelines handling PHI/PII data with compliance to HIPAA, GDPR, and CCPA. Integrated data from structured and unstructured sources across Kafka, MongoDB, Oracle, SQL Server, Hive, and S3.
* Experienced in ML model support pipelines and collaboration with Data Science teams using Vertex AI and SageMaker. Deep experience in MDM integration using tools like Informatica MDM, Talend, and custom Spark-based pipelines.
* Developed reusable and test-driven data components with strong CI/CD exposure using GitHub, Jenkins, and Docker.
* Skilled in data visualization and reporting through Power BI, Tableau and Data Studio.
* Strong cross-functional collaboration with data governance, business analysts, and stakeholders across Agile/Scrum teams. Led the modernization of legacy ETL systems to cloud-native architectures, improving performance, scalability, and cost-efficiency.
* Automated complex data workflows using Apache NiFi, Step Functions, and event-driven orchestration across hybrid cloud environments.
* Worked on data quality and lineage tracking, implementing golden record definitions and survivorship rules across multiple domains.
* Delivered high-impact solutions in claims analytics, policy data integration, revenue forecasting and customer 360 platforms across insurance and life sciences.

|  |  |
| --- | --- |
| Languages | Python, Scala, Java, SQL, PL/SQL, Shell Scripting |
| Databases | Oracle 9i/10g/11g/12, SQL Server 2000/2005, MS SQL, HBase, MongoDB, MySQL, Cassandra, DynamoDB, PostgreSQL |
| Bigdata Technologies | Apache Spark, Scala, Kafka, HDFS, Hive, Pig, MapReduce, Zookeeper, Sqoop, Oozie, Nifi, and Impala |
| ETL / ELT | Azure Data Factory, Informatica PowerCenter/MDM, AWS Glue, Talend, Apache NiFi |
| Development IDE`s | Eclipse, Visual Studio Code, Toad, SQL Developer |
| Logging & Monitoring | Splunk, CloudWatch, Log4J, SLF4J, Zipkins, Grahana |
| Operating Systems | UNIX, Linux, Ubuntu, Windows XP/2000/VISTA |
| Cloud Technologies | AWS (Lambda, EC2, S3, SNS, CloudWatch, CloudFormation, RDS, VPC, Auto Scaling, IAM, AWS Glue, AWS Batch, AWS DMS, Code Build, Code Deploy), Microsoft Azure (Azure Databricks, Azure Data Factory, Azure Data Explorer, Azure HDInsight, ADLS), Google Cloud Platform (Big Query, Compute Engine, Cloud Functions, Cloud DNS, Cloud Storage, Cloud Deployment Manager). |
| Data Warehousing | Snowflake, Amazon Redshift, Google BigQuery, Azure Synapse, Oracle, SQL Server |
| Visualization & BI | Power BI, Tableau, Google Data Studio, Kibana |
| |  | | --- | |  |   Data Modeling | Data Vault 2.0, Star Schema, Snowflake Schema, Dimensional Modeling, Erwin, Toad DM |

**EDUCATION:** Masters in Information Technology - University of the Cumberlands (2017)

**Professional Experience:**

**Client: Nationwide, Ohio, United States January 2024 - Present**

**Role: Senior Data Modeler**

**Responsibilities:**

Project 1: Cloud-Based Enterprise Data Modeling for P&C Insurance Analytics

* Designed and implemented enterprise-grade data models supporting P&C insurance analytics using Data Vault 2.0 methodology in Snowflake.
* Modeled fact and dimension tables to support underwriting claims and policy performance reporting using Star and Snowflake schemas.
* Defined source-to-target mapping documents and logical/physical data models using Erwin and Toad Data Modeler.
* Led the data profiling, cleansing and deduplication activities to standardize and enrich customer policy and claims datasets.
* Integrated row-level security and role-based access controls into Snowflake models to ensure data protection and compliance (GDPR/CCPA).
* Collaborated with business analysts and data governance teams to define golden record rules and master entity relationships.
* Worked closely with ETL developers and data engineers to translate business requirements into scalable, performance-optimized models.
* Performed impact analysis and managed model versioning to support agile data product delivery cycles.
* Provided model documentation and data dictionaries to business and technical stakeholders across multiple LOBs.
* Enabled data lineage tracking by integrating model metadata into governance platforms and reporting tools.

Project 2: Oracle to GCP BigQuery Migration and Modern Data Architecture Design

* Led data modeling efforts to migrate 100+ Oracle tables to GCP BigQuery, involving schema redesign and denormalization for performance.
* Developed hybrid modeling solutions using a combination of 3NF for operational marts and dimensional models for analytics layers.
* Created conceptual, logical, and physical models for customer and policy domains to support machine learning workflows in Vertex AI.
* Translated business requirements into BigQuery-native structures, optimizing partitioning, clustering, and table design.
* Defined data standards, naming conventions, and model lifecycle practices to support CI/CD-based model deployment via Terraform.
* Partnered with data scientists to prepare curated datasets with clearly modeled features for training, testing, and scoring.
* Worked with data stewards and compliance teams to ensure lineage, traceability and PII tagging across all datasets.
* Validated model accuracy and performance through query optimization and benchmarking in BigQuery and Data Studio.
* Assisted in building automated pipelines to extract metadata and lineage from models and push to governance tools.
* Supported ongoing enhancements and schema evolution by designing extensible, loosely coupled models adaptable to future use cases.

**Environment:** AWS (S3, Redshift, Glue, EMR, Lambda), GCP (BigQuery, Dataflow, Dataproc, Pub/Sub, Vertex AI), Snowflake, Airflow, Informatica MDM, PySpark, Terraform, Erwin, Kafka, SQL, Power BI, Oracle, MongoDB, Git, Jenkins, CloudWatch, Stackdriver.

**Client: Hewlett Packard Enterprise July 2020 – Dec 2023**

**Role: Senior Big Data Engineer**

**Responsibilities:**

* Designed & deployed multi‑cloud ETL/ELT pipelines on AWS and GCP, supporting both batch and streaming workloads.
* Orchestrated ML workflows with AWS Step Functions with SageMaker and GCP Dataflow/Vertex AI for automated training and inference.
* Developed distributed Spark jobs (Scala/PySpark) on EMR & Dataproc, processing terabyte‑scale financial and operational data.
* Ingested real‑time feeds via Kafka & Kinesis, landing in S3/GCS and loading into Redshift, BigQuery, and Hive.
* Built Dataflow pipelines from Pub/Sub to BigQuery in Python, integrating REST API sources and validation checks.
* Automated data quality & lineage monitoring with Airflow, Stackdriver, CloudWatch, and custom Python UDFs.
* Modeled star, snowflake, and Data Vault schemas using Erwin; translated business logic into physical & logical models.
* Migrated 100+ Oracle tables to BigQuery, refactoring schemas and aligning Power BI dashboards with new datasets.
* Provisioned & secured cloud resources (EC2, S3, RDS, Glue, Auto Scaling) via Terraform and strict IAM policies. Implemented S3 lifecycle policies & Glacier archives, driving storage‑cost optimization across environments.
* Generated Hadoop data cubes using Hive, Pig, and MapReduce; authored PySpark UDFs for complex aggregations.
* Created Kibana dashboards over Elasticsearch/Logstash for near‑real‑time end‑to‑end log analytics.
* Enhanced Informatica PowerCenter workflows, tuning mappings, parameters, and sessions for faster loads.
* Led Agile delivery (Scrum, TDD, BDD) and enforced Git‑based CI/CD for reproducible, version‑controlled releases.
* Conducted peer code reviews & SQL tuning, cutting ETL runtimes and query latency by up to 40 %.

**Environment:** Hadoop, Apache Spark, Scala, HDFS, Kafka, Hive, GitHub, Google Cloud Platform (GCP), Amazon Web Services (AWS), Python, PySpark, Jenkins, Perl, Agile, Informatica Power Center, UNIX, Snowflake

**Client: AbbVie, North Chicago, Illinois Mar 2018 – June 2020**

**Role: Senior Data Engineer**

**Responsibilities:**

* Migrated on‑prem Hadoop and ETL workloads to AWS, provisioning EC2, S3 and EMR for elastic storage and compute.
* Created reusable Delta views in Azure Databricks, ADF, and Synapse to speed up self‑service reporting and analytics.
* Engineered high‑volume ETL pipelines in AWS Glue, ingesting multi‑format files (JSON, Avro, Parquet, XML) into Redshift and Snowflake.
* Modeled dimensional, star, snowflake and Data Vault schemas in Snowflake to support governed analytics.
* Developed Spark (Scala/PySpark) batch and streaming jobs on EMR, Databricks, and Hortonworks HDP for near‑real‑time processing.
* Integrated Kafka and Kinesis streams, transforming data with PySpark and persisting outputs to HBase, Hive, and BigQuery.
* Provisioned feature‑ready datasets for ML teams, orchestrating workflows with AWS Step Functions, Lambda, and Glue triggers. Tuned Amazon Redshift clusters—partitioning, distribution and query optimization for faster, cost‑efficient reporting.
* Implemented AWS Data Lake architecture, cataloging data with Glue and automating metadata enrichment.
* Automated CI/CD pipelines via AWS CodePipeline, CodeBuild and CodeDeploy for infrastructure and data‑pipeline releases.
* Designed and managed NiFi flows and Sqoop jobs to ingest terabytes from Oracle, Teradata, and DB2 into Hive/HDFS.
* Built Oozie and Airflow workflows to orchestrate incremental loads, Spark transformations, and data‑quality checks.
* Authored custom Hive UDFs, Pig scripts, and bash utilities to optimize partitioning, bucketing, and business‑rule enforcement.
* Constructed ELK dashboards (Elasticsearch, Logstash, Kibana) for near‑real‑time monitoring of Kafka event logs and pipeline health.
* Established end‑to‑end observability with CloudWatch, Stackdriver, and Datadog across EMR, Glue, and Redshift workloads.
* Led Agile delivery with TDD/BDD and peer code reviews, ensuring secure, production‑ready data solutions on schedule.

**Environment:** Python, AWS, EC2, S3, EMR, Redshift, Hadoop, MapReduce, Hive, Pig, Spark, Kafka, Oozie, Nifi, Scala, PySpark, Snowflake, HBase, SQL

**Client: Black Knight, Jacksonville, FL Feb 2017 – Feb 2018**

**Role: Data Engineer**

**Responsibilities:**

* Created Pipelines in ADF using Linked Services/Datasets/Pipeline to Extract, Transform and load data from different sources like Azure SQL, Blob storage, Azure SQL Data warehouse, and write-back tool.
* Engineered end‑to‑end ADF pipelines (Linked Services, Datasets, Activities) to ingest, transform and load data from Azure SQL, Blob Storage and Synapse DW.
* Built a Python‑SDK scheduling framework to trigger ADF jobs, log run metadata to SQL DB and power real‑time ops dashboards.
* Authored Python UDFs with Spark code to flatten nested JSON, perform daily aggregations, and feed reporting tables.
* Developed & deployed Synapse Analytics pipelines via ADF, Databricks notebooks, and Synapse Studio for high‑volume ETL.
* Automated releases with Azure DevOps & ARM templates, enabling version‑controlled CI/CD for Synapse artifacts.
* Integrated Synapse with Blob Storage, Event Hubs, and Key Vault, ensuring secure secrets management and streaming ingest.
* Configured self‑hosted IR on Windows to securely migrate HDFS data into Azure Data Lake with minimal downtime.
* Implemented Git‑based versioning and scheduled triggers in ADF, standardizing pipeline lifecycle management.
* Ingested real‑time data with Kafka supporting Spark Streaming and used Apache NiFi flows to land data in HDFS for further processing.
* Designed Snowflake & ODS schemas using Data Modeler and Erwin, aligning dimensional models with analytics requirements.

**Environment:** Azure Data Factory (ADF), Azure Dataflow, Apache NiFi , Azure Event Hubs, Azure Event Hubs, Azure Stream Analytics, Apache Kafka, Azure SQL, Blob storage, Azure SQL Data Warehouse, Azure Synapse Analytics, Azure Databricks, Snowflake Schema, Python SDK, Spark Structured Streaming, Azure DevOps, Azure Resource Manager templates, Azure GIT, Azure Key Vault, Erwin (Data Modeling tool)

**Client: Lazard, Hyderabad, India. June 2013 - Nov 2015**

**Role: ETL Developer**

**Responsibilities:**

* Developed Advance PL/SQL packages, procedures, triggers, functions, Indexes and Collections to implement business logic using SQL Navigator.
* Authored advanced PL/SQL packages, triggers, functions and indexes to implement complex business rules and accelerate query performance.
* Created server‑side PL/SQL scripts and materialized views for high‑volume data validation, manipulation and remote reporting.
* Optimized Oracle tables through partitioning, compression, and index tuning, cutting I/O latency and boosting throughput.
* Designed end‑to‑end ETL workflows in Informatica PowerCenter—mappings, mapplets, reusable transformations and session control.
* Implemented incremental aggregation and performance tuning across sources, targets, mappings, and sessions to eliminate bottlenecks.
* Extracted and loaded XML and multi‑source data into Oracle using Informatica and custom PL/SQL for cleansing and enrichment.
* Automated batch schedules and backups with UNIX shell scripts; supported Oracle Streams for data replication.
* Conducted rigorous code reviews, defect resolution and SDLC documentation, ensuring release quality and audit readiness.
* Built management analytics reports using parallel queries and Java stored procedures for real‑time decision support.
* Maintained continuous production support and enhancement cycles, swiftly troubleshooting ETL and database issues.

**Environment:** Oracle 10g/11g, SQL Plus, TOAD, SQL Loader, SQL Developer, PL/SQL, Informatica Power Center, Designer, Workflow Manager, Workflow Monitor, Repository Manager, Shell Scripts, UNIX, Windows XP, Splunk, HTML, TOAD, XML.