|  |  |
| --- | --- |
| **Vinay kumar** **Senior Data Engineer****Email: karumanchivinaykumar01@gmail.com** | Professional Data Engineer Certification badge image. Certification. Advanced level. Issued by Google Cloud**A logo for a company  AI-generated content may be incorrect.A blue and white logo  AI-generated content may be incorrect.** |

**Professional Summary**

* Over 12+ years of professional IT experience, with 8+ years specialized in Data Engineering, and 6+ years in Data Warehousing across cloud-native and hybrid ecosystems.
* **Certified Data Engineer** with expertise across **Databricks, Microsoft Fabric, and Google Cloud**, specializing in Lakehouse architectures, real-time and batch data pipelines, and seamless integration with platforms like Power BI, Synapse, OneLake, BigQuery, Dataflow, Pub/Sub, and Vertex AI to enable advanced analytics and machine learning initiatives.
* Results-driven and highly adaptable Data Engineering professional with strong command over **big data architecture**, data lakes, ETL/ELT development including distributed data scripting with **Scope Script (Cosmos)** and event-based pipelines using **Azure EventHub**.
* Real-time stream processing using platforms like **Databricks**, **Apache Spark, Snowflake, Azure, and AWS**.
* Proven experience in **financial and healthcare data domains**, with a focus on **compliance, governance, and performance.**
* Deep expertise in implementing scalable, secure, and high-performing data pipelines using Databricks Workflows, **Airflow**, Azure Data Factory, AWS Glue, and **Delta Live Tables**, enabling smooth orchestration of batch and streaming workloads.
* Built real-time ingestion pipelines using **GCP Pub/Sub** and **Dataflow** for event-driven processing.
* Experienced in **Perl, Shell scripting**, and **Unix-based ETL troubleshooting**, Led ETL issue resolution in distributed Linux environments and worked on **Telecom data platforms.**
* Proficient in designing and managing data lakehouse architectures leveraging Delta Lake, **Unity Catalog**, Apache Iceberg, and Hudi, enhancing metadata management, **governance**, and **cost-efficient querying**.
* Adept in crafting robust **data models**, both relational (3NF) and dimensional (star/snowflake schemas) for analytical workloads using tools like Snowflake, **Synapse**, **Redshift**, and **BigQuery**.
* Hands-on experience with **data ingestion** frameworks (e.g., **Kafka, NiFi, AWS DMS, Talend**) and real-time event streaming for IoT and mobile data capture, ensuring **low-latency analytics**.
* Experienced in leading healthcare data engineering initiatives using **Matillion**, **DBT**, and **Snowflake**, with strong **focus on HIPAA compliance, ELT performance**, and **data governance**.
* Strong background in **NoSQL** technologies including **MongoDB, Redis, Neo4j, and Cassandra**, enabling schema-less, high-volume, and **graph data use cases**.
* Extensive use of cloud services from **AWS (S3, Glue, Redshift, Lambda, EMR), Azure (ADF, ADLS, Synapse, AKS),** and **GCP (BigQuery, Cloud Functions, Storage**) for seamless data operations and warehouse scaling.
* Demonstrated expertise in building and maintaining **CI/CD pipelines using GitHub, Jenkins, Azure DevOps, and Terraform** to automate deployment, infrastructure provisioning, and testing in **Agile/Scrum environments.**
* Strong advocate of **data governance and quality frameworks**, implementing tools like **Apache Griffin**, **DataBuck,** Great Expectations, and to automate validations, enforce compliance, and **track lineage**.
* **Configured and maintained Apache NiFi clusters**, including setup of custom processors, flow templates, and parameter contexts to support scalable ingestion from MongoDB, SQL Server, and real-time Kafka sources.
* Tuned **NiFi flows** for latency reduction and throughput enhancement, implementing backpressure, prioritization, and provenance **tracking for secure and reliable data delivery**.
* **Delivered business-impacting analytics** via **Power BI, Tableau, and Looker**, helping cross-functional teams derive insights, **monitor KPIs, and power executive dashboards**.
* Implemented **Dremio** as a semantic layer to enable fast, **self-service SQL analytics** on S3-**based Iceberg tables**, integrating with **Power BI dashboards** and reducing query times by 60% and demonstrated **success working in multi-cloud and cross-functional team environments**, collaborating closely with data scientists, software engineers, and business analysts to deploy production-grade data products.
* Known for maintaining a **strong** **documentation** culture, process standardization, and knowledge transfer through tools like **JIRA, Confluence, and Lucidchart.**
* Passionate about **continuous learning and mentorship**, currently pursuing **advanced training in AI/ML integration** with data platforms and **cloud cost optimization best practices**.
* Highly skilled in advanced performance tuning techniques, including **query optimization**, **index strategies**, **partitioning**, and **caching**, significantly reducing processing times and optimizing cost efficiency in platforms like **Snowflake**, **Redshift**, **Synapse**, and **BigQuery**.
* Proven capability in **migrating legacy data systems** into modern, scalable cloud solutions, managing complex migrations seamlessly with **minimal downtime** and disruption using AWS DMS, Azure Migration Services, and **custom scripting**.
* Expert in developing and deploying resilient data systems leveraging **container orchestration** technologies (Docker, Kubernetes, Helm) and **infrastructure as code (Terraform, CloudFormation, Ansible)** to provide robust and scalable environments that can grow dynamically with enterprise demands.
* Proficient **in advanced data transformation** and analytics using **Snowflake scripting, window functions, regular expressions, and JSON/XML parsing**, empowering business users with deeper, real-time analytical capabilities and accurate **predictive modeling inputs**.
* Adept at **orchestrating data pipeline monitoring and logging** using sophisticated tools such as Prometheus, Grafana, ELK Stack (Elasticsearch, Logstash, Kibana), **CloudWatch**, and **Datadog**, ensuring proactive identification and resolution of performance issues.
* Experienced in creating and managing robust security frameworks, **including IAM roles, data encryption (AES-256, TLS), audit logging, OAuth2.0, and compliance management (GDPR, CCPA),** ensuring full protection of sensitive data and regulatory adherence.
* Exposure to **SAS**, **Business Objects**, and **Crystal Reports** for legacy analytics and reporting.
* Experience working with analytical tools like **OLAP cubes, SAS reports**, and **legacy B**I platforms in data warehouse migration projects.
* Deep understanding of **AI/ML integration** within **data pipelines**, proficient in deploying predictive models and automation through frameworks and tools like **MLflow, SageMaker, TensorFlow, Azure ML**, and Feature Stores, enhancing business decision-making through data-driven intelligence.

**Technical skills**

| **Category** | **Skills** |
| --- | --- |
| **Data Processing and Analysis** | Apache Spark, Databricks, Apache Kafka, Matillion, DBT, Snowflake, Hadoop, Databricks SQL, Dremio, Wherescape (Red, 3D), AWS EMR, AVRO, JSON |
| **Cloud Technologies** | Amazon Web Services (AWS), Microsoft Azure (Cosmos DB, **Scope Script/U-SQL**, EventHub, ADF, Synapse), Google Cloud Platform (GCP), Snowflake, Databricks, Cloudera, IBM Cloud, Oracle Cloud Infrastructure (OCI) |
| **Data Storage and Management** | Azure Data Lake Storage (ADLS), Azure Blob Storage, AWS S3, GCP Cloud Storage, Azure Synapse, Unity Catalog, Delta Lake, Iceberg, Hudi, Hive Metastore, Auto Loader |
| **Data Warehousing** | Azure Synapse Analytics, BigQuery, AWS Redshift, Snowflake, Teradata |
| **Data Orchestration** | IBM DataStage, Apache Airflow, Azure Data Factory, AWS Glue, AWS Lambda, gePrefect, Luigi, Oozie, Control-M, Dagster, ETL Pipelines, Delta Live Table Pipelines, Databricks Workflows |
| **Visualization** | Tableau, Power BI, Databricks SQL Dashboardsterr |
| **Programming Languages** | Python, PySpark, Spark SQL, SQL, T-SQL, Scala, C, Perl, Java, JavaScript, Shell Scripting, R, Node.JS, C# |
| **NoSQL Databases** | MongoDB, Neo4j, Redis, Cassandra, DynamoDB, HBase |
| **Databases** | MySQL, PostgreSQL, SQL Server, Oracle, Snowflake, IBM DB2 |
| **Version Control & Methodology** | Git, GitHub, GitLab, Bitbucket, Jenkins, Azure DevOps, Terraform, YAML, Ansible, Docker, Kubernetes, Azure DevOps, Terraform |
| **Monitoring & Logging** | Prometheus, Grafana, ELK Stack (Elasticsearch, Logstash, Kibana), CloudWatch, Datadog, Splunk, Kusto (KQL) |
| **Security & Compliance** | IAM, OAuth2.0, KMS, Data Encryption (AES256, TLS), Audit Logging, VPCs, Network Isolation, Azure Purview, Access Management |
| **Documentation & Collaboration** | Confluence, Jira, Erwin Data Modeler, Notion, Lucidchart, Draw.io |
| **Software Development & Methodology** | Agile, Scrum, Kanban, SDLC, Test-Driven Development (TDD), DataOps, DevSecOps |
| **AI/ML & Generative AI**  | ChatGPT (OpenAI), Amazon Bedrock, Hugging Face Transformers, LangChain, Prompt Engineering, Retrieval-Augmented Generation (RAG), LLM API Integration |

**Work Experience**

**Client: Select Health, Remote** **Jul 2023 – Present**

**Role: Data Engineer**

**Responsibilities:**

* Participated in the strategic design and implementation of comprehensive **end-to-end data pipelines, integrating diverse data extraction methods, including robust API integrations**, sophisticated **web scraping techniques**, and database connectivity through technologies such as **PyODBC, Python, and T-SQL**, ensuring seamless, reliable, and **timely data processing**.
* Developed and optimized Snowflake data pipelines and built low-code solutions using Power Apps and Power Automate to streamline data integration, automated approvals, and reporting workflows for business operations.
* **Orchestrated ETL workflows** using **Matillion**, streamlining data ingestion, transformation, and load processes across cloud environments while ensuring consistency and reusability of pipeline components.
* Enabled data-driven insights by supporting the development and maintenance of Tableau dashboards, integrating curated Snowflake datasets for real-time healthcare analytics and operational reporting.
* **Architected** and continuously **optimized high-performance, scalable data workflows**, carefully designed to meet **dynamic business intelligence** and analytics requirements, significantly **enhancing data-driven decision-making** capabilities across the organization.
* Designed metadata-driven ETL workflows using **Wherescape Red**, standardizing healthcare data transformation and improving governance, auditability, and delivery speed of payer data pipelines into Snowflake.
* Recommended and implemented **data cleansing strategies** to resolve data inconsistencies in migrated EDI X12 datasets (837/835/270), improving reliability of payer-provider reporting systems.
* **Actively collaborated with cross-functional teams**, including data scientists, software engineers, business analysts, and business intelligence professionals, **facilitating agile data-driven solutions** aligned directly with **strategic business objectives**, ensuring clarity in data outcomes and **maximizing business impact**.
* **Translated complex business requirements** into actionable, scalable technical solutions, effectively bridging the **communication gap between technical and business teams**, resulting in smoother project execution, increased alignment, and **faster project delivery cycles**.
* **Designed and implemented robust data models and schemas**, adhering strictly to best practices in relational (3NF normalization) and dimensional (Kimball methodologies) modeling, ensuring optimal database performance, usability, and maintainability, while enhancing query efficiency and scalability.
* Worked on **healthcare-specific EDI data** processing pipelines **(X12 formats like 837, 835, 270/271**) and payer-provider integrations, implementing **Lambda and Step Functions** for scalable pre-/post-processing and **compliance validation**.
* Tuned Databricks Runtime configurations for cost-performance optimization in HIPAA-compliant batch pipelines, adjusting cluster policies and workload profiles to reduce compute costs while ensuring SLA adherence.
* Effectively managed, monitored, and optimized complex data pipelines by **employing leading data orchestration tools such as Apache Airflow, Prefect, DBT, and SSIS**, significantly enhancing ETL/ELT operational efficiencies and **enabling reliable automation, job scheduling, and rapid troubleshooting**.
* Implemented SLA-driven monitoring and ServiceNow-based incident management frameworks, ensuring proactive identification and rapid resolution of data pipeline failures and compliance breaches.
* Spearheaded continuous improvement initiatives within the data engineering department, routinely evaluating, adopting, and promoting **industry-leading practices** and technologies to continually enhance **pipeline robustness, data quality, team productivity**, and overall operational excellence.
* Cultivated and championed a strong culture of **clean coding standards**, rigorous **peer-review processes**, adherence to best practices, and comprehensive documentation to promote maintainability, readability, and long-term scalability of **data engineering solutions**, resulting in **decreased technical debt** and higher team efficiency.
* **Actively mentored junior** and **mid-level data engineers**, providing guidance on best practices, career development, and **technical skill enhancement**, thereby fostering an environment of learning, innovation, and **professional growth** across the engineering team.
* Developed detailed and thorough **documentation**, including technical manuals, **data flow diagrams**, **architecture diagrams, and process documentation**, ensuring clear **knowledge transfer**, streamlined onboarding of new team members, and **sustained operational clarity**.

**Technologies Used:** Snowflake, Python, T-SQL, PyODBC, Matillion, QlikSense, Azure (Data Factory, Synapse, Data Lake), AWS (Glue, S3, EMR, Redshift), Docker, Kubernetes, Apache Airflow, Prefect, DBT, SSIS, Microsoft Fabric, REST API Integrations, CI/CD (Jenkins, GitHub), Agile methodologies.

**Environment:** AWS (S3, Redshift, EMR, SNS, SQS, Glue, CloudWatch, Kinesis, Route53, IAM), Azure, Aethna, Sqoop, MySQL, HDFS, Apache Spark, Hive, Cloudera, Kafka, Zookeeper, Oozie, PySpark, Ambari, JIRA, IBM Tivoli, Control-M, Teradata, Oracle, SQL, Matillion, QlikSense

**Client: Frontier Communications, Irving, TX**

**Role: Data Engineer**

**Responsibilities:**

* **Participated in a strategic migration from Snowflake and Teradata to Databricks**, coordinating seamless pipeline transitions with DBT, and ensuring full compatibility of existing transformations while minimizing downtime through careful planning and validation.
* Acted as **Data Migration Engineer** in moving large datasets from **on-prem Teradata and SQL Server** to **AWS S3 and Databricks Delta Lake**, leveraging **PySpark** and **SQL-based orchestration.**
* Collaborated with infrastructure and DevOps teams to **operationalize migration pipelines**, applying **data transformation, cleansing**, and quality validation prior to cloud ingestion.
* Developed and tested data movement logic to ensure **minimal downtime** and complete fidelity during migration.
* Integrated and governed metadata using **IBM Knowledge Catalog** and custom lineage validation processes to enforce data governance and compliance in regulated environments.
* Developed and maintained Tableau dashboards for operational KPIs and service delivery metrics, optimizing user experience and reducing data access latency for business teams.
* Built real-time **log analytics pipelines** using **Apache Kafka, Kafka Connect**, and **Apache Druid/ElasticSearch**, enabling predictive monitoring for edge devices and retail-like usage metrics. Applied **Redis** for fast in-memory caching and operational telemetry.
* Built fault-tolerant Matillion pipelines and automated SLA compliance monitoring integrated with ServiceNow for efficient production support and incident resolution workflows.
* Wrote and optimized complex **T-SQL queries**, including **window functions**, **recursive CTE**s, and subqueries to support analytics and operational reporting.
* Designed and implemented **dimensional data models** (star/snowflake schemas), ensuring scalable reporting and fast querying performance.
* Integrated REST APIs and explored **GraphQL contract-driven ingestion** for structured and nested data from mobile apps, ensuring schema evolution support and consistent joins across source datasets
* Built and **orchestrated ETL pipelines** in **Matillion**, streamlining data ingestion from RDS and file-based sources into **Snowflake** prior to full migration to Databricks, supporting both historical and real-time data loads.
* Collaborated with analysts to design **Looker dashboards and LookML models**, transforming customer usage data into actionable KPIs, supporting executive decision-making for high-volume consumer transactions.
* **Developed and optimized SLA-bound real-time ingestion pipelines using Kafka and Spark Structured Streaming, ensuring low-latency processing of telecom events and logs. Integrated BigQuery for scalable analytics and reporting across terabytes of near real-time data**.
* **Refactored DBT scripts for Databricks compatibility**, ensuring schema alignment, data type consistency, and optimal performance of all transformation logic within the new environment.
* **Implemented Databricks Medallion Architecture (Bronze, Silver, Gold)** to structure data pipelines, improve governance, and enhance lineage and maintainability across raw, processed, and aggregated datasets.
* **Replaced legacy Airflow with Databricks Workflows and Asset Bundles to package and deploy jobs, notebooks, and configurations across dev/stage/prod environments—ensuring reliable, version-controlled releases and improved collaboration with GitHubConducted rigorous testing and reconciliation**, validating data consistency and integrity through reconciliation testing, audit logs, and post-migration verification reports.
* **Configured and maintained critical AWS infrastructure**, including EC2, RDS, and S3, while setting up CI/CD pipelines via GitHub Actions to automate testing, version control, and deployment workflows.
* **Enabled scalable metadata management** by updating configuration tables within Databricks, supporting dynamic onboarding of new datasets and evolving transformation needs.
* **Integrated Operational Data Mart (ODM) principles** into the Databricks Lakehouse, streamlining data accessibility and responsiveness for critical business analytics.
* **Oversaw clean data egress from Databricks to Snowflake**, managing daily truncation-load tasks to ensure high accuracy and timeliness of analytics-ready datasets.
* **Enabled seamless Power BI integration with Snowflake**, optimizing data models to support consistent and performance-driven dashboards for business stakeholders.
* **Authored detailed documentation**, including architectural diagrams, lineage maps, and operational guides to support knowledge transfer, onboarding, and ongoing sustainability.
* **Mentored junior engineers**, fostering a culture of collaboration, innovation, and technical excellence across the data engineering team.

**Technologies Used:** PostgreSQL, Amazon RDS, Amazon S3, Teradata, Snowflake, Databricks (Delta Lake, PySpark, Spark SQL, Workflows, Unity Catalog), DBT, Matillion, QlikSense, Medallion Architecture, Apache Airflow, Power BI, Git, AWS EC2, GitHub Actions, CI/CD, Agile Methodology.

**Environment:** Sqoop, MySQL, HDFS, Apache Spark (Scala/PySpark), Hive, Hadoop, Cloudera, Kafka, MapReduce, Zookeeper, Oozie, Ambari, Python, Data Pipelines, RDBMS, JIRA, Matillion, QlikSense

**Client: Equality Health, Remote**

**Role: Databricks Developer**

**Responsibilities:**

* **Collaborated cross-functionally with Product Managers, Data Scientists, Analysts, and ML Engineers** to translate complex requirements into **scalable data engineering solutions**, significantly improving decision-making and operational efficiency.
* **Designed and optimized high-throughput data pipelines** for ingesting, transforming, and curating massive datasets from **global advertising platforms**, leveraging technologies like **Apache Spark, Flink, Kafka**, and **Databricks** to ensure high availability, low latency, and real-time analytics capability.
* Led data migration initiatives from **on-prem SQL Server** to **Databricks on AWS/GCP**, utilizing **PySpark notebooks** and **DBT** for transformation logic, while coordinating with analytics teams to validate migrated tables.
* Configured Delta Table triggers in Databricks to implement audit logging and downstream CDC (Change Data Capture) mechanisms, improving traceability and compliance with HIPAA requirements..
* **Implemented robust infrastructure solutions** using **Hadoop, Snowflake, BigQuery, and Databricks**, supporting petabyte-scale data processing and analytical workloads with minimal downtime and maximum reliability.
* **Architected and deployed advanced data modeling and visualization strategies** using **Tableau, Power BI**, and **Databricks SQL Analytics**, enabling clear and actionable dashboards for campaign performance and key metrics.
* **Established Matillion and Informatica-based data validation and match-merge frameworks, improving data reliability and aligning healthcare datasets to industry regulatory standards**.
* **Proactively enhanced data pipeline performance** by implementing **query optimizations, partitioning, indexing, and caching strategies**, leading to reduced processing time, cost savings, and better system scalability.
* **Authored and maintained detailed technical documentation and governance standards**, improving organizational data literacy, onboarding efficiency, and long-term knowledge sharing across the engineering team.
* **Participated in roadmap planning, project scoping, and strategic leadership discussions**, helping align business objectives with domain-specific data strategies and resource allocation.
* **Drove data-first business transformation initiatives**, identifying and executing on strategic data opportunities that improved **advertising ROI, campaign effectiveness**, and customer engagement.
* **Used GitLab CI/CD pipelines to automate deployment of PySpark notebooks and testing suites in Databricks, and developed mock data generation tools (JSON, CSV, XML) for simulating patient and campaign records in lower environments.**
* **Applied advanced analytical and problem-solving skills** to address complex data architecture challenges, adopting innovative approaches to support high-performing, data-driven marketing strategies.
* **Stayed ahead of industry trends**, continuously evaluating new tools, emerging frameworks, and best practices to evolve data engineering capabilities and maintain competitive advantage in the ad-tech space.

**Technologies Used:** Apache Hadoop, Apache Spark, Apache Kafka, Apache Flink, Snowflake, Databricks, BigQuery, AWS, GCP, Python, Java, Go, SQL, Scala, Tableau, Power BI, Databricks SQL Analytics, Airflow, Kubernetes, Docker, Prometheus, Grafana, Terraform, JIRA, Confluence, Data Governance Tools.

**Environment**: AWS, AWS S3, redshift, EMR, SNS, SQS, Athena, glue, cloudwatch, kenisis, route53, IAM, Sqoop, MYSQL, HDFS, Apache Spark, Hive, Cloudera, Kafka, Zookeeper, Oozie, PySpark, Ambari, JIRA, IBM Tivoli, control-m, OOZIE, airflow, Teradata, oracle, SQL

**Client: CooperSurgical, Livingston, NJ**

**Role: Data Engineer**

**Responsibilities:**

* **Designed and optimized end-to-end cloud-native data solutions** using Snowflake, Azure Data Factory, Synapse Analytics, ADLS, and Azure SQL—automating workflows and improving data throughput, query speed, and operational efficiency by up to 45%.
* Developed and maintained SCOPE scripts in **Azure Cosmos environment** for high-performance distributed query processing, enabling efficient analysis of large-scale datasets across EDL.
* Built near-real-time ingestion pipelines using **Azure EventHub and Azure Data Factory** to support streaming analytics and timely alerting mechanisms.
* **Managed and scaled hybrid cloud infrastructure** across Azure and AWS, leveraging native services for secure, cost-effective, and high-performing data environments—ensuring continuous monitoring, reliability, and infrastructure optimization.
* **Developed and maintained secure, high-performance RESTful APIs** to streamline data exchange between internal systems and third-party platforms—implementing thorough testing, monitoring, and documentation to ensure reliability and robustness.
* **Implemented proactive data validation and monitoring frameworks**, reducing downstream errors by 35% and establishing high data quality standards across the entire pipeline.
* Applied data cleansing and masking policies during **migration of legacy health records** from **on-prem SQL** to **Azure Data Lake, leveraging ADF** and **EventHub f**or batch and streaming pipelines, aligning with regulatory and business SLAs.
* **Automated CI/CD pipelines using Azure DevOps**, integrating testing, version control, and continuous deployment to accelerate release cycles and maintain high software quality across environments.
* **Collaborated cross-functionally with product, analytics, and engineering teams**, identifying and applying best-in-class data engineering practices and emerging technologies to drive innovation and competitive advantage.
* **Provided mentorship and technical leadership** to junior engineers and analysts—promoting a collaborative, growth-oriented environment focused on clean code, continuous learning, and process excellence.
* **Optimized cloud costs through strategic resource management**, monitoring, and capacity planning—reducing unnecessary spend and maximizing ROI on Azure and AWS services.
* **Established and enforced secure data governance frameworks**, ensuring compliance with privacy regulations and internal policies via regular audits, security assessments, and proactive risk mitigation.

**Technologies Used:** Snowflake, Azure Data Factory, Azure Synapse Analytics, Azure SQL Database, Azure Data Lake Storage (ADLS), Azure Pipelines, Azure Stack, AWS, Python, SQL, RESTful API Development, Power BI, CI/CD Automation, Data Warehousing, Data Quality Management, Data Visualization, Cloud Security, Data Governance, Agile Methodologies, Algorithmic Optimization.

**Environment:** Hybrid cloud environment (Azure + AWS), cloud-native architecture, CI/CD pipelines with Azure DevOps, secure and compliant data governance frameworks, high-performance RESTful API systems, Agile development methodology, cost-optimized and scalable cloud infrastructure, cross-functional team collaboration (product, analytics, engineering teams).

**Client: Nationwide, Columbus, OH**

**Role: AWS Data Engineer**

**Responsibilities:**

* Demonstrated extensive proficiency in designing, developing, and managing sophisticated **ETL pipelines** using **AWS Glue** and **AWS Data Pipeline**, effectively supporting diverse analytical and operational business needs. Streamlined complex data integration processes, facilitating rapid, reliable ingestion and transformation across numerous applications.
* Built and maintained secure, high-performance Snowflake pipelines for financial reporting and regulatory risk analytics in banking and insurance domains, ensuring data governance, compliance (SOX, GDPR), and high-availability reporting for executive dashboards.
* Integrated **Matillion** into the ETL workflow during **AWS-Snowflake** hybrid architecture implementation, enabling **seamless orchestration** and efficient transformation logic across **AWS Glue and Snowflake**.
* Enabled business teams to derive actionable insights through Power BI and Tableau dashboards, integrating Power Platform services to automate refresh cycles, approval workflows, and financial KPI tracking for faster executive reporting.
* **Designed scalable, cross-platform data workflows** bridging AWS services and **Snowflake**, ensuring data validation and transformation standards were consistently met in BI and marketing reporting layers.
* Expertly engineered and maintained robust data validation frameworks leveraging advanced capabilities of **Spark SQL** and **AWS Glue DataBrew**, significantly improving data accuracy, consistency, and reliability. Implemented proactive monitoring and alerting mechanisms to swiftly detect and rectify data inconsistencies or anomalies.
* Automated AWS operations using **boto3 scripts** to manage S3 lifecycle policies, trigger Lambda functions, and control AWS Glue jobs, streamlining ETL pipeline performance for insurance reporting workloads.
* Successfully led integration projects involving **Adobe Analytics** and multiple third-party marketing and analytics platforms, expertly utilizing **AWS Glue** and **AWS Data Pipeline** to automate data synchronization, streamline reporting processes, and enhance data-driven decision-making capabilities for marketing teams.
* Architected and implemented robust **ETL data pipelines** specifically designed for ingestion of large-scale data from **RDBMS** systems, including **MySQL databases**. These pipelines optimized data processing workflows, reduced latency, and ensured timely data availability within the **AWS environment**.
* Adhered stringently to industry-leading **data governance practices** utilizing **AWS Data Catalog** for metadata management, version control, and comprehensive lineage tracking. This meticulous governance approach significantly improved data transparency, compliance, security, and traceability across the entire data lifecycle.
* Exhibited strong proficiency in **data modeling**, **statistical analysis**, and **visualization capabilities** leveraging powerful AWS services including **Amazon EMR** for big data processing, and **Amazon QuickSight** for interactive dashboards and analytics. Facilitated deeper insights, rapid analytics turnaround, and actionable visualizations for business stakeholders.
* Optimized **Tableau** visualization performance significantly by implementing advanced **caching techniques** using **AWS Data Lake House Engine**, resulting in drastically reduced dashboard load times and enhanced user experience, driving greater adoption and user satisfaction.
* Effectively optimized resource-intensive **SQL queries**, meticulously following AWS and database-specific performance best practices. Achieved substantial improvements in query execution speed, resource utilization, and overall database performance, thus delivering faster and more efficient data-driven decision-making capabilities.
* Authored complex **SQL queries** for data validation, reliability testing, and creation of robust **data warehouses**. Ensured integrity and dependability of analytical results, maintained high standards of data quality and accuracy through rigorous validation processes.
* Successfully developed and maintained comprehensive **AWS Glue workflows**, ensuring uninterrupted, scheduled execution of ETL jobs and data processing tasks. Significantly streamlined data operations, reduced manual interventions, and optimized resource management.

**Technologies Used:** AWS Glue, AWS Data Pipeline, AWS Lambda, AWS Step Functions, Amazon Redshift, **Snowflake**, **Matillion**, **QlikSense**, AWS Glue DataBrew, Amazon EMR, Amazon QuickSight, AWS Data Catalog, AWS Lake Formation, Teradata, SQL Server, MySQL, Python, SQL, Apache Spark (Spark SQL), Terraform, Apache Presto, Apache Drill, Tableau, Adobe Analytics, Bitbucket, Jenkins, CI/CD.

**Environment:** Cloud-native AWS environment (serverless and managed services), large-scale enterprise data warehousing (Amazon Redshift, **Snowflake**), big data processing frameworks (EMR, Spark SQL), ETL orchestration and automation (AWS Glue Workflows, Data Pipeline, **Matillion**), real-time and batch data processing, strong focus on data governance and metadata management (AWS Data Catalog, Lake Formation), Agile development practices with CI/CD pipelines (Bitbucket, Jenkins), integration with third-party analytics tools (**QlikSense**, Adobe Analytics, Tableau).

**Client: Kellogg, Battle Creek, MI**

**Role: ETL Developer**

**Responsibilities:**

* Designed and implemented highly scalable **data pipelines** tailored for both **real-time streaming** and **batch processing** workloads, utilizing advanced frameworks such as **Apache Spark**, **PySpark**, and **AWS Glue**. Successfully managed the ingestion, transformation, and integration of **structured**, **semi-structured**, and **unstructured data** across distributed systems, significantly reducing latency and ensuring reliable, accurate data availability for analytics.
* Led the comprehensive development and optimization of robust **ETL workflows** leveraging tools such as **Apache NiFi** and **Talend** to automate and streamline data extraction processes from diverse, enterprise-grade sources, including **Oracle**, **PostgreSQL**, and **MongoDB**, thus enabling near-real-time analytics and faster operational reporting capabilities.
* Engineered, maintained, and optimized complex **data warehousing solutions** within **AWS Redshift**, strategically applying **data modeling techniques**, **database normalization**, and **query optimization methodologies**. This substantially improved query performance, reduced processing time, and ensured highly efficient analytical query execution.
* Integrated and expertly orchestrated **cloud-native services**, including **Amazon S3**, **Amazon RDS**, **Redshift**, and **AWS Lambda**, to deliver cost-effective, scalable, and high-performance data solutions supporting extensive analytical workloads and **machine learning pipelines**. Facilitated seamless data integration, operational efficiency, and rapid scalability across cloud infrastructures.
* Developed and maintained **reusable, modular Python and PySpark codebases**, actively employing best practices such as **automated unit testing**, **clean coding standards**, and extensive documentation. Leveraged **CI/CD practices** using **GitHub** and **Jenkins pipelines**, streamlining code deployment processes and ensuring seamless integration into production environments.
* Worked extensively with stakeholders to establish and enforce standardized **data integration** processes and guidelines, facilitating efficient, high-throughput data ingestion from **IoT devices**, **log-based sources**, and **external APIs**. Expertly utilized technologies such as **Apache Kafka** and **Apache NiFi** for **real-time streaming** and **event-driven workflows**, enabling near-instantaneous insights and analytics.
* Documented detailed **technical designs**, **data flows**, and **architectural patterns** comprehensively, fostering knowledge transfer, transparency, and continuous improvement within engineering teams. Produced and maintained clear, detailed documentation, significantly enhancing onboarding efficiency, operational clarity, and overall team productivity.
* Provided **technical mentorship and leadership** to junior data engineering team members, promoting a culture of excellence, rigorous testing, professional growth, and best practices adherence. Delivered actionable feedback and hands-on training sessions, ensuring continuous team development and high-quality outcomes.

**Technologies** **Used**: Python, PySpark, Apache Spark, Apache NiFi, Talend, SQL, Oracle, PostgreSQL, MySQL, MongoDB, Cassandra, AWS (Glue, S3, Redshift, RDS, Lambda), GitHub, Jenkins, Data Modeling (Relational & Dimensional), Query Optimization, Data Warehousing, ETL, Streaming & Batch Processing, Kafka, CI/CD, Database Performance Tuning, Agile methodologies.

**Environment**: SQL Server 2008/2012 Enterprise Edition, SSRS, SSIS, T-SQL, Windows Server 2003, Performance Point Server 2007, Oracle 10g, visual Studio 2010.

**Education**

* Bachelors in Computer Science Engineering in 2013 from India.