Daniel Vaa

Senior Data Engineer

Gilbert, AZ, 85295 | +1 (305) 402-4958 | daniel.h.lessor@gmail.com | https://www.linkedin.com/in/danny-v-767413329

❖ SUMMARY

Senior Data engineer with **11+** years of experience designing and optimizing data pipelines, ETL processes, and scalable systems. I'm passionate about transforming complex data into actionable insights and working across teams to solve real-world challenges.

❖ SKILLS

Programming	Python, R, Scala, Java, Javascript, Typescript, PHP, C#, C++, Go
Frameworks	Django, Flask, Java Spring Boot, Node.js, Express, Nest, React, Next.js, Angular, Vue, Nuxt.js, Laravel, ASP.Net, .NET Core, GraphQL, FastAPI, Rest APIs, React Native, Swift, Ionic, Splunk
Database	MySQL, PostgreSQL, MongoDB, T-SQL, NoSQL, DynamoDB, RabbitMQ
Tools	Git, Github, Gitlab, Bitbucket, NPM, Yarn, PNPM, Webpack
Cloud Services	AWS, GCP, Azure
Automated Testing	Jest, Mocha, Chai, Cypress, Enzyme, Playwright, Unittest, Pytest, Selenium, Puppeteer

❖ EDUCATION

- **Western Governors University, Millcreek, UT** – *Master`s degree in Data Science*
  Apr 2014 - Aug 2017
- **Western Governors University, Millcreek, UT** – *Bachelor`s degree in Computer Science*
  Apr 2012 - Sep 2014
- **College of Central Florida, Ocala, FL** – *Associate's degree*
  Apr 2010 - Aug 2012

❖ CERTIFICATIONS

- ✓ AWS Certified: Solutions Architect – Associate
- ✓ AWS Certified: SysOps Administrator – Associate
- ✓ Microsoft Certified: Azure Network Engineer – Associate
- ✓ Microsoft Certified: DevOps Engineer – Expert
- ✓ Data Science with Python
- ✓ Machine Learning in Python

❖ EXPERIENCE

State of Wisconsin, Madison, WI – *Senior Data Engineer*
Apr 2017 – Sep 2024

- **[Clients: Personify Health, Alpine, Khealth, Contextualize, Shine Software]**
- Implemented scalable big data pipelines using Scala and Python on Azure Databricks to handle real-time streaming data, enabling near-instantaneous analytics and built complex data science models in R & Python to provide insights into customer segmentation.
- Designed web applications using Django/Flask, integrating data pipelines with backend services and implemented business logic in Python with Django/Flask to automate workflows and optimized query performance using Django ORM, SQLAlchemy in Flask.
- Automated data transformation workflows using Python and mentored junior data engineers in Python and Scala coding practices.
- Implemented distributed services & microservices with Go and integrated Go services with Kafka for the real-time data streaming.
- Architected end-to-end Databricks solutions with Kafka integration to enable data streaming and designed ETL pipelines to extract, transform, and load data from the disparate sources including relational database, NoSQL, RESTful APIs, and cloud data storage.
- Implemented Alteryx workflows for automating ETL processes and developed complex data pipelines using the Alteryx Designer to streamline data integration from multiple sources and optimized existing Alteryx workflows, reducing data processing time by 40%.
- Integrated spatial data pipelines using QGIS & ArcGIS to process geospatial datasets and implement geospatial data warehousing solutions, combining PostGIS and ArcGIS to analyze the large volume of spatial data and migrated the legacy geospatial systems.
- Developed ontology models to structure complex datasets, enabling efficient data categorization, retrieval, and semantic querying.
- Created hierarchical taxonomies to organize unstructured & structured data, and applied industry standards & domain knowledge.
- Designed healthcare APIs for real-time data integration across various systems, ensuring compliance with HL7 standards & FHIR protocols and implemented encryption techniques with HIPAA & HITECH regulations and conducted HITRUST gap assessments.
- Developed data pipelines to support AI/ML models and integrated the machine learning workflows into a data engineering pipeline.
- Built Power BI dashboards for deep insights into business performance, using Azure Synapse, Data Lake, & Azure SQL Database.
- Implemented Power BI dashboards with complex DAX calculations, providing real-time analytics and actionable business insights.

- Integrated Looker into data ecosystem and designed over 50 Looker explores and dashboards to streamline business intelligence.
- Automated data processing workflows using VBA and Bash (Smash) scripts and implemented Apache Hive queries and optimized Hive table for querying on distributed dataset and developed interactive dashboard in Tableau to visualize real-time data analytics.
- Created data model (tabular & multidimensional) optimized for performance and used Databricks Delta Lake for data consolidation.
- Implemented advanced data models for HR, Payroll, and Financial datasets using Workday's Prism Analytics, and optimized ETL pipelines to extract, transform, & load Workday data into data warehouses using tools such as Talend, Apache Airflow, & Python.
- Designed EDI systems that facilitated secure data exchange, ensuring compliance such as EDIFACT, X12, and HL7, implemented of Interoperability solutions across multi-platform environments and utilized ETL pipelines to extract, transform, and load EDI data.
- Managed Linux infrastructure for the data engineering environment and automating routine tasks using Bash scripts and Python.
- Migrated ETL workflows from the traditional on-premise system to Databricks Lakehouse platform and implemented Unity Catalog, ensuring centralized governance and fine-grained, adhering to industry best practices and compliance standards (GDPR, HIPAA).
- Implemented ClickHouse as primary OLAP database solution and optimized Data Pipelines by replacing traditional ETL processes with ELT using ClickHouse and designed materialized view & data partitioning strategy in ClickHouse to optimize query efficiency.
- Designed complex T-SQL queries to support large-scale data manipulation and built MS SQL Server Integration Services (SSIS) packages and collaborated with product team to integrate the external data source into the analytics pipelines using SDK libraries.
- Migrated legacy data systems into Quickbase, and implemented custom Quickbase API solutions automated reporting workflows.
- Enhanced data visualization by integrating SSRS, SSIS, SSAS, and full Microsoft BI stack to create reporting & analytics solutions.
- Implemented Azure DataLake Storage (ADLS) solution to manage petabyte data, optimizing data access with fine-grained security policy using Azure RBAC & ACLs and built TimeTravel capabilities with Databricks Delta Lake to enable snapshot data versioning.
- Engineered custom Datatables to streamline data integration processes, using Spark SQL and integrated AutoLoader for real-time ingestion pipelines in Databricks and utilized JupyterLab with hands-on experience in RStudio for the data statistical modeling in R.
- Developed complex ETL solutions using IBM DataStage & optimized DataStage jobs for multi-terabyte financial data warehouse.
- Designed data pipelines using PySpark to process large datasets (TBs/PBs) on distributed clusters, reducing the ETL/ELT runtime.
- Automated deployment and orchestration of PySpark jobs using Apache Airflow and CI/CD pipeline and enabled model training by preparing feature set from raw data using PySpark DataFrames & Spark SQL and Implemented advanced Spark tuning technique.
- Optimized data pipelines using AutoSys for job scheduling, integrating with ETL technologies like Apache Airflow and Informatica.
- Automated data pipelines using Apache Airflow for scheduled ETL processes and integrated Apache NiFi as data ingestion layer.
- Integrated Openlink's ETRM platform to automate energy trading and ingested trading data from sources into the Openlink system.
- Designed Oracle & PostgreSQL databases for enterprise applications and conducted performance tuning for relational databases (PostgreSQL, Oracle, MySQL) and automated routine database maintenance tasks using shell scripts & database automation tool.
- Implemented MySQL databases supporting large-scale data-driven applications and migrated data from legacy systems to MySQL & PostgreSQL and improved MySQL query performance and integrated MySQL with NoSQL and AWS RDS & Google Cloud SQL.
- Architected Microsoft SQL Server database environment, implemented database migration to SQL Server 2022 for system stability
- Optimized SAP HANA data models and ETL workflows, and designed SAP HANA data integration solutions for real-time analytics.
- Built SAP HANA database structures to support data warehousing need and managed the complex SAP HANA stored procedures.
- Used tidyverse packages (dplyr, ggplot2, tidyr) to clean, transform, and analyze complex datasets, reducing data preparation time.
- Managed NVIDIA DGX A100 systems to support data processing and machine learning workflows, reducing model training times
- Implemented data solution using MongoDB & CockroachDB and optimized multi-region replication strategy for Apache Cassandra.
- Designed Redis as in-memory data structure store, used Redis Pub/Sub mechanism to enable messaging & data synchronization.
- Architected ETL processes using PETL to transform the large volumes of raw data from disparate sources into normalized formats.
- Implemented large Splunk infrastructure to monitor and manage multi-terabyte data environments and developed dashboards and custom searches using SPL to provide real-time insights, used Splunk to ingest data from cloud services, databases, applications.
- Architected distributed data processing solution using Hadoop (HDFS, MapReduce, and YARN) and developed real-time analytics pipelines using Elasticsearch, Logstash, and Kibana (Elastic Stack) and deployed Docker containerized microservices architecture.
- Designed data models and machine learning pipelines using Python libraries (Pandas, Scikit-learn) to predict customer behavior.
- Optimized large ETL pipelines using Pandas and Dask and utilized Dask for distributed computing, reducing data processing time.
- Designed ETL/ELT solutions with Matillion and adopted Azure Pipelines for CI/CD practices and implemented robust MFT solution, ensuring compliance with data transfer standards (e.g., GDPR) and drove the integration of HashiCorp Vault for managing secrets.
- Developed ETL pipelines using Visual Studio integrated with MS SQL Server Data Tools (SSDT) for big data warehousing project.
- Implemented Ray to parallelize ML model training and automated the workflow orchestration with the Prefect, Dagster and Airflow.
- Spearheaded database design for large applications and used VBA to automate repetitive data tasks and optimized data pipelines using Bash (Smash) scripts and built interactive Tableau dashboards to track KPIs and used Apache Hive to process big datasets.
- Built automated Unix/Shell scripts for managing large data pipelines and created custom Shell scripts to automate job scheduling.
- Migrated to Snowflake from an on-premise data warehouse and optimized ETL pipelines in Snowflake and implemented RBAC & data masking and integrated Snowflake with various third-party tools (Qlik, Fivetran, Airflow, Looker) for data ingestion & reporting.
- Implemented data pipelines using Matillion to automate ETL processes and managed data warehousing solutions on Snowflake.
- Built ETL pipelines using Talend Data Integration for large datasets and enhanced data quality through Talend Data Quality tools.
- Implemented Ataccama Data Quality (DQ) Platform to automate data profiling, cleansing, and validation processes, and designed data quality rules & scorecards within Ataccama, and integrated various Ataccama Master Data Management (MDM) capabilities.
- Orchestrated the complex data pipeline construction using Azure Data Factory (ADF) & Databricks, handling high-volume dataset.
- Led end-to-end data integration, transformation, and loading (ETL) pipelines using Azure Synapse for large-scale data processing.

- Implemented data pipelines in Azure Data Factory (ADF) to automate & schedule workflows, integrating data sources (on-premise, cloud) and optimized Azure SQL database for large-scale transactional data and developed T-SQL scripts for data transformation.
- Automated ETL processes for ingesting CSV, JSON, and HTML files into the Azure SQL databases, applying the data validation.
- Designed Apache DataFusion pipelines to orchestrate data ingestion from multiple data sources, streamlining the ETL processes.
- Integrated Ruby microservices with AWS to automate ingestion & validation and developed data-centric APIs using Ruby on Rails.
- Implemented Microservices for data pipeline platform and optimized data structures & algorithms to enhance data retrieval speed and developed multi-threaded solutions for parallel data ingestion and migrated legacy monolithic applications to the microservice architecture and implemented algorithms for data deduplication, ETL transformations, & data aggregation, improving performance.
- Designed ETL pipelines using Apache Airflow to process big datasets and integrated Apache NiFi for data ingestion & processing.
- Integrated Salesforce Data Cloud with data warehouse, enabling customer views and built automation solutions using Salesforce Flow Builder, streamlining data transformation processes and generated the customer insights with Salesforce Calculated Insights.
- Implemented Oozie workflows to automate ETL processes and built Sqoop jobs to transfer large-scale data between Hadoop and relational database like MySQL & Oracle and integrated Apache Flume to stream real-time log data from various sources to HDFS.
- Integrated C++ services with Python data pipelines, enabling ETL operations across distributed systems & memory management.
- Used SAS for data manipulation & ETL process, and built data models in MATLAB, integrating with business intelligence platform.
- Designed GPU ML models with CUDA and TensorFlow, and migrated the legacy CPU processes to GPU processing using CUDA.
- Architected Apache Iceberg tables on distributed data lake environment and optimized partitioning strategies in Iceberg to improve query performance and integrated Apache Iceberg with Apache Spark for incremental data processing & time travel functionality.
- Developed CI/CD pipelines for data pipelines using dbt (Data Build Tool) across the data lake and adopted dbt for data modeling.
- Built Kubernetes clusters to manage streaming data workloads and automated Docker image builds and Kubernetes deployments.
- Implemented FSTP across data pipelines and deployed automated FSTP solutions between on-premise and cloud environments.
- Integrated Generative AI (Gen AI) models into core platform and deployed Vertex AI pipelines for training & deploying ML models.
- Developed LLM for natural language understanding in chatbots and integrated OpenAI's GPT models for report generation system.
- Designed data pipeline to support LMS, using ML algorithms to provide personalized learning path & predictive recommendations.
- Built deep learning models using TensorFlow and PyTorch for various predictive tasks, such as image recognition, NLP tasks, and time-series forecasting, and architecture deep learning pipelines, implementing the model training using Kubernetes and Docker.
- Led DataOps initiatives to streamline data integration and processing workflows, using Pentaho, Boomi, and Jitterbit, reducing ETL processing times, managed large EHR & clinical data migrations using Informatica PowerCenter tools & BTEQ for EDC integration.
- Implemented real-time data processing pipelines using Apache Flink on AWS EMR and deployed Data Lake architectures on AWS S3, using Apache Hudi for efficient upserts, data versioning & optimized resource management and job scheduling on AWS EMR.
- Built DevSecOps processes using AWS EFS and AWS EBS for storage management and implemented automated data validation workflows with PowerShell and Apache Flink, and created interactive dashboards and reports using QlikView and Microsoft Fabric.
- Reduced manual data handling by implementing automated data capture solution with UiPath Studio & Microsoft Power Automate.
- Implemented GitOps pipelines for automating infrastructure management and established automated deployment workflow using Git, Jenkins, and ArgoCD, and automated GitOps model and integrated Helm charts & Kubernetes manifests with Git repositories.
- Built data models in Azure Analysis Services and managed CI/CD pipelines in Azure DevOps, automating deployment processes.
- Architected large Parquet data lakes on AWS S3, and enhanced data ingestion pipelines with Apache Parquet, and implemented cloud infrastructure using Terraform on AWS, leading automation efforts for serverless services like Lambda, DynamoDB, and S3 and designed data models for use in Looker and Tableau, improved data pipelines feeding into BI tools by optimizing SQL queries.
- Developed data strategy for e-commerce platform, and implemented data architecture using AWS such as S3, Redshift, and EMR.
- Optimized HBase architecture for petabyte data warehouse & migrated relational databases to HBase reducing data storage costs.
- Optimized message queue performance by configuring AWS SQS and designed data processing pipelines using Amazon EMR.
- Deployed data pipelines using AWS services like AWS Lambda, Glue, and EMR, and migrated legacy data infrastructure to AWS, using Amazon S3, Redshift, & RDS and integrated real-time data streaming solution with Kinesis & Amazon MSK (Apache Kafka).
- Established monitoring solutions using AWS CloudWatch across all critical AWS services, including AWS EC2, Lambda, and RDS.
- Migrated on-premises data infrastructure to Google Cloud Platform (GCP) and implemented data pipelines using Google Cloud Dataflow (GCP) & Apache Beam for batch data processing, optimized GCP BigQuery architecture, improving query performance.
- Designed serverless data architecture using Cloud Functions, BigQuery, Pub/Sub and optimized data lake architecture on GCP by using Cloud Storage, BigQuery, Dataflow and implemented Cloud IAM, Cloud KMS, & VPC Service Controls to ensure compliance with GDPR & CCPA and designed disaster recovery strategies on GCP, ensuring the system redundancy across multiple regions.

Google, Madison, WI – *Python Engineer*

Jul 2014 - Mar 2017

- Designed real-time recommendation engine with Python &Spark and built R-based predictive models for customer churn analysis.
- Implemented Scala-based data pipelines for processing transaction data in AWS, cutting down the batch processing time by 50%.
- Built web services using Django/Flask to support ML pipelines and managed large data using Django ORM or Flask-SQLAlchemy.
- Coordinated cross-functional teams to integrate data from multiple sources into the central data lake using Python and AWS Glue.
- Designed data architecture integrating Kafka streaming with Databricks for real-time analytics and built ETL pipelines to automate the ingestion, transformation, and loading of structured and unstructured data from various on-premise and cloud-based systems.
- Developed Tableau visualizations for various departments and built Apache Hive table for the fast processing of terabytes of data.
- Managed large Hadoop cluster and optimized MapReduce jobs for financial data processing and built the monitoring system using Elastic Stack (Elasticsearch, Logstash, Kibana) and implemented Java data processing pipelines, improving data ingestion rates.

- Developed Java applications for processing large datasets and built Elasticsearch search engine and managed Docker containers for various data processing applications, optimized performance by implementing Hadoop-Hive integrations for query optimization.
- Designed spatial ETL workflows using FME, ArcGIS, and QGIS, and worked on ArcGIS ModelBuilder to automate repetitive tasks.
- Developed healthcare data pipelines using APIs, meeting data exchange needs of EHR systems following HL7 standards, ensured data governance, addressing HIPAA and HITECH regulatory requirements while supporting ongoing HITRUST compliance efforts.
- Built data engineering pipelines for AI/ML projects and developed data pipelines for ML applications using Apache Spark & Kafka.
- Developed dashboard using Power BI to visualize key performance metrics, using advanced DAX calculations and optimized data models (tabular and multidimensional) supporting OLAP solutions, improving decision-making for the finance and operations team.
- Directed implementation of Looker as primary BI tool, reducing reliance on Excel-based report and cutting manual reporting efforts.
- Developed Linux big data environments, managing Hadoop & Spark clusters for large-scale data processing and maintained Linux Interoperability solutions, reduced operational costs by improving system monitoring with Linux tools like Nagios and Prometheus.
- Implemented complex T-SQL scripts for data validation, auditing, and transformation across diverse datasets and integrated SSRS for automated reporting system and used SDKs & APIs for integration of external data source into existing reporting infrastructure.
- Deployed DGX A100 systems for deep learning & AI model training and used DGX A100 to streamline data-intensive workflows.
- Developed tabular models using Microsoft SQL Server Analysis Services (SSAS) to optimize business reporting and implemented data transformations using Python and Spark to standardize the data from the disparate systems into the unified data warehouse.
- Developed Quickbase applications that streamlined data integration across team and optimized the ETL pipelines with Quickbase.
- Embedded Power BI solutions across various enterprise systems, allowing for real-time insights & better decision-making process.
- Collaborated with cross-functional teams to integrate Workday data into existing enterprise systems, ensuring seamless data flows using Workday API, Workday Studio, and Workday Report Writer and improved Workday reporting and the analytics performance.
- Debugged T-SQL queries supporting mission-critical data flows and reporting systems, including stored procedures, CTES, UDFs, & RBAC and optimized SSIS package for complex ETL workflow and deployed SSRS & SSAS solution to unify reporting standard.
- Migrated legacy ETL processes to Talend and integrated system like AWS Redshift, Salesforce, & SAP Hana across organization.
- Built interactive visualizations and reports with ggplot2 and other tidyverse tools, transforming the raw data into actionable insights.
- Led the automation of daily and monthly data pipelines using SAS macros and MATLAB scripts, and optimized the MATLAB code.
- Implemented Salesforce Data Cloud, ensuring integration with legacy systems & cloud environments, and optimized data pipelines for Salesforce to enable analytics, using Flow Builder to automate workflows and implemented Calculated Insights in Salesforce.
- Designed cloud data warehouse solutions on Azure Synapse and developed complex Azure Data Factory (ADF) pipelines for data extraction, transformation, and implemented data partitioning on Azure SQL and migrated legacy database to Azure SQL solution.
- Utilized ingestion framework for structured & semi-structured data (CSV, JSON, HTML) using ADF pipelines and Azure Functions.
- Migrated on-premise data lakes to Azure DataLake Storage (ADLS) and developed TimeTravel feature using Delta Lake on Spark, and optimized Datatables for efficient data storage & retrieval and implemented AutoLoader to automate streaming data ingestion.
- Optimized PySpark jobs for processing large volumes of data and built data ingestion frameworks on Apache Spark and Hadoop.
- Built PySpark batch jobs on AWS EMR and integrated with Amazon S3 and streamlined data checks using PySpark & Delta Lake.
- Developed ETL workflows using Apache Flink and Apache Hudi for a large-scale Data Lake and used AWS EMR and S3 for cost-effective data processing and built a robust data governance framework within the Data Lake environment, ensuring compliance.
- Implemented Splunk solutions for enterprise-level log management and managed Splunk data onboarding and optimized custom alerts and reports in Splunk to detect anomalies in real time and developed the Splunk apps to automate data ingestion workflows.
- Architected MariaDB databases for transactional web platform and designed ETL/ELT workflows to streamline data ingestion from multiple sources (CSV, API, third-party database) into MariaDB & PostgreSQL environment and optimized database performance.
- Deployed Matillion for cloud ETL workflows and integrated Azure Pipelines with data engineering ecosystem to automate testing & deployment of IaC and implemented MFT workflows to automate secure transfers and enhanced data security by integrating Vault.
- Utilized advanced CUDA libraries such as Thrust, cuBLAS, and cuDNN to optimize real-time data ingestion and processing tasks.
- Used Visual Studio for debugging stored procedures and integrated C# libraries within Visual Studio to automate data cleansing.
- Implemented Azure DataLake Storage (ADLS) to centralize enterprise data and deployed TimeTravel solutions within Delta Lake, and designed custom Datatables for large datasets and used JupyterLab and RStudio for scalable notebook-based development.
- Migrated from Oracle to PostgreSQL for mission-critical applications and managed MySQL databases and tuned PostgreSQL and Oracle database performance by optimizing query plans and implemented PL/SQL stored procedures for data transformation logic.
- Migrated to MS SQL Server 2016 and automated routine data extraction processes using complex T-SQL scripts & SSIS package.
- Adopted Apache Cassandra for large data warehousing project and deployed CockroachDB to implement distributed transactions.
- Optimized existing MongoDB clusters by configuring shard keys and acted as subject matter expert for advanced NoSQL systems.
- Integrated clinical and survey data using CMIX, Survey Monkey, and Qualtrics for health research and optimized SSIS workflows for ETL processes, and customized Case Report Forms (CRFs) and automated data entry tasks with RPA tools (UiPath, Power Automate), and used Trillium and DMBOK standards to ensure data governance, quality, and compliance with industry regulations.
- Migrated on-premise big data infrastructure to Dockerized Hadoop clusters on AWS and optimized Java MapReduce jobs for batch processing of large datasets and integrated Kafka with Hadoop for real-time data streaming, increasing the data pipeline efficiency.
- Developed data engineering pipelines with Pandas for data cleansing and transitioned the team to Dask for distributed computing.
- Integrated Ray to parallelize hyperparameter tuning for ML models and optimized processes with Python's multiprocessing library.
- Implemented Shell scripts for data validation and optimized Unix/Shell scripting workflows for Hadoop clusters and integrated Shell scripts with version control systems (Git) and used Unix/Shell scripting to automate file processing tasks, integrate third-party APIs.
- Used Snowflake's Time Travel and automated data ingestion pipelines using Snowpipe from cloud sources (AWS S3, Azure Blob).
- Built ETL processes using Talend Big Data Integration and integrated Talend with the real-time data sources like Kafka and Spark.

- Migrated from legacy ETL tool to Apache DataFusion and implemented DataFusion plugin to accommodate business requirement.
- Designed complex AutoSys job flows and integrated AutoSys with cloud data warehouses like AWS Redshift & Google BigQuery.
- Developed data pipeline workflows using Apache Airflow and used Apache NiFi to implement enterprise data ingestion framework.
- Migrated historical data from legacy storage formats to Apache Hudi and integrated AWS SNS to enable data processing workflow.
- Engineered data models using Azure Analysis Services and managed Azure DevOps pipelines for version control (Git) and testing.
- Architected cloud-based microservices platform, using Kubernetes and Docker for containerization and enhanced data processing pipelines by designing and optimizing custom algorithms for sorting, filtering, and data transformation, improving efficiency by 25%.
- Utilized data structures (heaps, graphs, hash maps) and developed multi-threading strategies to enable concurrent processing in distributed systems and streamlined inter-service communication by implementing asynchronous messaging patterns using Kafka.
- Designed cloud-based ETL workflow using AWS Glue, AWS Lambda, & Step Functions and drove adoption of Amazon Redshift, managed secure migration of on-premise databases to AWS RDS (PostgreSQL), ensuring compliance with industry standards like GDPR, HIPAA by employing AWS IAM, KMS, & Security Groups and developed APIs using AWS API Gateway and AWS Lambda.
- Developed custom CloudWatch dashboard to visualize real-time performance metric and set up CloudWatch alarms & notification.
- Integrated SQS with AWS Lambda and S3 and tuned EMR clusters for optimal performance by optimizing Spark, Hive, and Presto.
- Deployed cloud-based data solutions using EC2, S3, and Redshift to support large-scale data processing and implemented AWS Glue to create ETL infrastructure and architected VPC & subnet designs and managed AWS Elastic Beanstalk for the deployment.
- Implemented Apache Iceberg on cloud-based data lake and managed data pipelines utilizing Iceberg's ACID transaction support.
- Integrated dbt into the ETL framework and created dbt models for business metrics, establishing clear and auditable data lineage.
- Collaborated closely with DevOps team to ensure seamless deployment of C++ microservices within the Kubernetes environment.
- Transitioned to GCP data warehouse solution using BigQuery and Data Fusion and integrated streaming data processing pipelines using Pub/Sub, Dataflow, & Bigtable and implemented Cloud IAM policy and optimized the large batch processing using Dataproc.
- Architected cloud-native data platform on GCP using BigQuery, Cloud SQL, Cloud Storage and deployed machine learning models using the AI Platform & Cloud ML Engine and established GCP monitoring and logging using Cloud Monitoring and Cloud Logging.
- Engineered data pipelines using Ruby and Shell scripting to automate and built Ruby on Rails platform for real-time data reporting.
- Built automation workflows in Microsoft Excel and used Microsoft Word and Microsoft Outlook to manage communication pipelines.
- Implemented database design and automated data transformation and reporting tasks using VBA and wrote Bash (Smash) scripts.
- Developed front-end project, using HTML5, CSS3, JavaScript, jQuery and built client-side logic using JavaScript for asynchronous API calls (AJAX) to retrieve large dataset and improved performance of legacy system by optimizing HTML, CSS, JavaScript code.
- Designed ML pipeline using Vertex AI for real-time data analytics & deployed Generative AI model for automated content creation.
- Utilized GitOps strategy to automate deployment of cloud infrastructure and used GitOps principle to manage Kubernetes clusters.
- Implemented LLM solutions for semantic search and integrated OpenAI's API & built predictive analytics models for LMS platform.
- Designed data pipelines using Parquet format for large data processing, and created IaC templates using Terraform for managing cloud resources on Google Cloud Platform (GCP), and automated data processing workflows in Terraform to handle infrastructure changes, and integrated data pipelines with Power BI and Tableau and optimized dashboard performance through advanced SQL.
- Worked with multiple Agile development teams, applying Scrum methodology to streamline data engineering processes and used Jira to manage backlogs and integrated Confluence as the central knowledge repository and adopted Agile SAFe methodologies.