RABBANI MOHAMMAD

Contact: (646) 416-6932; **Email:** rabbanimoh93@gmail.com

Lead Data Engineer/Architect

AWS, Azure, GCP | Hadoop | Kafka | Snowflake | ETL

PROFILE SUMMARY

- **Lead Data Engineer/Architect** with **10+ years** of experience in Big Data development. Demonstrates a **strong ability to leverage leading cloud platforms (AWS, Azure, GCP)** to design, develop, and deploy scalable, high-performance data processing pipelines.
- Expert in designing and deploying enterprise-grade **Generative AI solutions**, leveraging **LLMs**, Retrieval-Augmented Generation (**RAG**), LangChain, Hugging Face Transformers, **NLP, OCR**, and vector databases (pgvector, FAISS, OpenSearch) to extract insights from unstructured biomedical and clinical datasets.
- **Big Data Development & Cloud Platform Expertise:**
  Expertise in Big Data tools such as Apache Spark, Kafka, Hadoop, Flink, Redshift, and Snowflake, with a focus on data ingestion, ETL processes, and real-time analytics. Strong experience with distributed computing, **data lake,** and **data warehouse solutions**, as well as advanced analytics techniques including machine learning and predictive modeling.
- **AWS Cloud Solutions & Data Security:**
  Expert in utilizing AWS services such as Redshift, Kinesis, Glue, EMR, Lambda, and CloudFormation for building and optimizing data solutions. Skilled in implementing robust security and compliance measures through AWS IAM, CloudTrail, and CloudWatch to ensure regulatory adherence. Proficient in automating workflows using AWS Step Functions and managing infrastructure as code.
- **Hadoop Ecosystem & Data Transformation:**
  In-depth experience with the Hadoop ecosystem (Cloudera, Hortonworks, AWS EMR, Azure HDInsight, GCP DataProc) for data ingestion, extraction, and transformation. Hands-on expertise with AWS Glue, Lambda, and Azure Databricks for seamless ETL and data integration.
- **Google Cloud Platform (GCP) & Event-Driven Architecture:**
  Proficient in GCP services such as DataProc, Dataprep, Pub/Sub, Cloud Composer, and Cloud Storage for building sophisticated data workflows. Skilled in implementing scalable event-driven messaging solutions with GCP Pub/Sub and utilizing Google Cloud Audit Logging for compliance tracking. Experienced in using Terraform for GCP resource management and CI/CD pipeline automation.
- **Azure Cloud Services & Data Solutions:**
  Expertise in Azure Data Lake, SynapseDB, Data Factory, Databricks, and HDInsight for big data processing and data management. Well-versed in using Azure Functions for serverless computing and optimizing data storage and retrieval with Azure Storage solutions.
- **Data Security & Compliance:**
  Strong focus on data security, access control, and compliance through tools like AWS IAM, CloudTrail, Google Cloud Audit Logging, and encryption methods. Expertise in implementing secure data governance frameworks to ensure data privacy and integrity.
- **Advanced Analytics, Machine Learning & AI Initiatives:**
  Skilled in executing data analytics, machine learning, and AI-driven initiatives to derive actionable insights for strategic decision-making. Proficient in deploying models for predictive analytics, fraud detection, and customer insights.
- **Data Pipeline Optimization & Performance Tuning:**
  Proven experience in building and managing highly efficient data pipelines across AWS, Azure, and GCP. Expertise in Spark performance tuning in platforms such as Databricks, AWS Glue, EMR, and on-premises environments to ensure optimal performance for large-scale data processing.
- **CI/CD Pipelines & DevOps Practices:**
  Hands-on experience with continuous integration and deployment (CI/CD) using tools such as Jenkins, Azure DevOps, and CodePipeline. Proficient in containerization technologies (Kubernetes, Docker) and version control systems (GitHub) to streamline the deployment of big data solutions.
- **Data Formats & Querying:**
  Experienced with diverse data formats including JSON, XML, and Avro, and proficient in SQL dialects such as HiveQL and BigQuery SQL for querying large datasets and supporting advanced data analytics.
- **Agile & Collaborative Project Delivery:**
  Active contributor in Agile/Scrum teams, managing Sprint Planning, Backlog Management, and Requirements Gathering.

Strong communicator with stakeholders and project managers to ensure alignment of project goals with business needs and successful delivery of data solutions.

TECHNICAL SKILLS

Cloud Platforms:
- **Amazon Web Services (AWS):** Amazon S3, AWS Glue, Amazon EMR, AWS Lambda, AWS Kinesis, Amazon MSK, AWS Redshift, and AWS SNS.
- **Google Cloud Platform (GCP):** BigQuery, Google Cloud Functions, and Google Cloud Storage.
- **Microsoft Azure:** Azure Data Factory, Azure Data Lake Storage (ADLS), and Azure Synapse Analytics.

Gen AI Tools:
- **Cloud & AI Platforms:** GCP (BigQuery ML, Vertex AI, Dataflow, Dataproc), AWS (SageMaker, Bedrock)
- **Machine Learning & NLP:** PyTorch, TensorFlow, Hugging Face Transformers, LangChain, OpenAI GPT, Claude, LLaMA, OCR, Vector Databases (Pinecone, FAISS)
- **AI Operations:** RAG, LLMOps, MLOps, AI Model Deployment, AutoML, MLflow

Big Data Technologies:
- **Apache Hadoop Ecosystem:** Hadoop Distributed File System (HDFS), Apache Hive, Apache HBase, Apache Pig, and Apache Flume.
- **Apache Spark:** Spark Core, Spark SQL, Spark Streaming, and Spark Structured Streaming.
- **Apache Kafka:** Kafka for real-time data streaming and messaging.
- **Apache NiFi:** Experienced in data ingestion and flow management.

Databases:
- **NoSQL Databases:** Cassandra, HBase, DynamoDB, MongoDB, and AWS Redshift.
- **Relational Databases:** MySQL, Oracle, PL/SQL, NoSQL, MongoDB, and Amazon RDS.
- **Data Warehousing:** Snowflake and Teradata

Programming and Scripting Languages:
- **Python:** Experienced in data processing and automation.
- **Scala:** Proficient in functional programming and big data processing.
- **Java:** Skilled in object-oriented programming and application development.
- **SQL:** Experienced in database querying and management.
- **Bash:** Proficient in shell scripting for automation.

ETL and Data Integration Tools:
- **Apache Airflow:** Experienced in workflow orchestration and scheduling.
- **Apache NiFi:** Skilled in data ingestion and flow management.
- **DBT (Data Build Tool):** Experienced in data transformation and modeling.
- **Pentaho:** Skilled in data integration and analytics.
- **SQL Server Integration Services (SSIS):** Experienced in data extraction, transformation, and loading.

Data Visualization Tools:
- **Tableau:** Proficient in creating interactive dashboards and reports.
- **Power BI:** Skilled in business analytics and data visualization.
- **Amazon QuickSight:** Experienced in cloud-based data visualization.
- **Looker:** Proficient in data exploration and visualization.
- **Kibana:** Skilled in log and time-series data visualization.

Cluster Security and Management:
- **Kerberos:** Experienced in network authentication.
- **Apache Ranger:** Skilled in data security and access control.
- **Identity and Access Management (IAM):** Proficient in managing user permissions and roles.
- **Virtual Private Cloud (VPC):** Experienced in network isolation and security.

CI/CD and Version Control:
- **Jenkins:** Proficient in continuous integration and delivery.
- **GitHub:** Experienced in source code management and collaboration.
- **GitLab:** Skilled in version control and DevOps practices.

Project Management and Methodologies:
- **Agile and Scrum:** Experienced in iterative development and project management.
- **DevOps:** Skilled in collaboration between development and operations teams.
- **Continuous Integration (CI) and Continuous Deployment (CD):** Proficient in automating code integration and deployment.
- **Test-Driven Development (TDD):** Experienced in writing tests before code implementation.
- **Unit Testing and Functional Testing:** Skilled in ensuring code quality and functionality.
- **Design Thinking:** Proficient in user-centered design and problem-solving.

PROFESSIONAL EXPERIENCE

Gen AI/Data Architect
BNY Mellon, New York City, NY | July 2023 – Present

As a Lead AI/Data Engineer, I oversaw the design and deployment of scalable cloud architecture using AWS services, including S3, Redshift, EMR, and Glue. I optimized data processing workflows, integrated various data sources, and implemented robust security measures, driving performance and compliance across big data platforms. Also, I architected and deployed scalable AI/ML solutions leveraging AWS cloud services, focusing on Generative AI, NLP, and large-scale data processing, including the design and implementation of Retrieval-Augmented Generation (RAG) architectures, optimized LLM performance through prompt engineering and fine-tuning, and built AI-driven data processing workflows for real-time analytics.

- ∞ Administered a comprehensive suite of AWS services, including S3, Athena, Glue, EMR, Kinesis, Redshift, IAM, VPC, EC2, ELB, RDS, CodeDeploy, ASG, and CloudWatch, to architect scalable and high-performance cloud solutions.
- ∞ Orchestrated the deployment of Amazon EC2 instances, customizing configurations for diverse applications and Linux distributions, while managing Amazon S3 storage for optimized data accessibility.
- ∞ Enhanced data processing efficiency and scalability through Amazon EMR and EC2, aligning system performance with evolving business demands.
- ∞ Provided strategic guidance on AWS architecture, implementing robust security protocols with IAM and Amazon Macie to protect sensitive data and ensure compliance.
- ∞ Designed scalable workflows using AWS Glue, EMR, and Kinesis for efficient data cleansing, streaming, and analytics, alongside automated scheduling with Apache Airflow and custom Python scripts.
- ∞ Leveraged Snowflake and Redshift for advanced data warehousing, improving data accessibility and analytics to support informed decision-making.
- ∞ Integrated Hive and MySQL with Apache Spark and Python to streamline data extraction, transformation, and loading, enabling seamless data flow and processing.
- ∞ Refined Scala and NiFi implementations to optimize performance, scalability, and resource utilization in handling large-scale data streams.
- ∞ Implemented AWS CloudFormation to automate resource provisioning, ensuring consistent deployments across environments, while utilizing Step Functions and Kinesis for event-driven data pipelines.
- ∞ Applied rigorous error handling, logging, and data validation to maintain pipeline integrity and deliver reliable ETL operations.
- ∞ Explored on-premises infrastructure migration to AWS, integrating fully managed services like Kafka and Glue for seamless data streaming and transformation.
- ∞ Continuously monitored and tuned workflows, enabling real-time analytics and ensuring optimal cloud resource utilization.
- ∞ Worked on NiFi flow optimization, including configuring processors, managing flow files, and ensuring high-throughput and fault tolerance in complex data pipelines.

Gen AI (LLM, NLP) & ML:

- ∞ Designed and implemented Retrieval-Augmented Generation (RAG) architectures for enterprise-grade Generative AI applications.
- ∞ Developed custom large language models (LLMs) based solutions, fine-tuning models using Hugging Face, LangChain, and OpenAI APIs for domain-specific tasks.
- ∞ Conducted AI model behavior analysis, evaluating outputs for relevance, consistency, and bias, and refining LLM responses for optimal user experience.
- ∞ Designed prompt engineering strategies, optimizing LLM outputs and building reusable prompt libraries for various enterprise applications.
- ∞ Built scalable AI pipelines integrating AWS SageMaker, Lambda, Bedrock, and Step Functions to deploy and serve LLM models in production.
- ∞ Researched and implemented NLP techniques, including named entity recognition (NER), sentiment analysis, and document summarization.
- ∞ Developed validation frameworks and automated evaluation metrics to assess AI model performance and ensure continuous improvement.
- ∞ Ensured ethical AI practices, maintaining inclusivity, bias mitigation, and data privacy compliance in AI applications.
- ∞ Integrated AI workflows with AWS services such as S3, Glue, EMR, Redshift, Kinesis, IAM, CloudWatch, and Apache Spark for large-scale data ingestion and transformation.
- ∞ Collaborated with cross-functional teams, including data scientists, ML engineers, and business stakeholders, to align AI strategies with enterprise goals.

Tech Stack: AWS, S3, Athena, Glue, EMR, Kinesis, Redshift, IAM, VPC, EC2, ELB, RDS, CodeDeploy, ASG, CloudWatch, Snowflake, Hive, MySQL, Apache Spark, Apache NiFi, Apache Airflow, Python, Scala, AWS CloudFormation, Step Functions, Kafka, Amazon Macie, data warehousing, data streaming, ETL, big data, security compliance, real-time analytics, resource optimization, Linux, data validation, fault tolerance, scalable workflows, cloud architecture, on-premises migration.

LLMs & NLP: OpenAI GPT, Claude, OCR, LLaMA, LangChain, Hugging Face, Vector Databases (Pinecone, FAISS)
AI/ML: AWS Bedrock, SageMaker, TensorFlow, PyTorch, MLflow, AutoML

Senior AI/Data Architect
Champion Energy, Houston, TX | Mar 2021 – Jun 2023

As a Senior AI/Data Architect at Champion Energy, I spearheaded the design and deployment of high-performance AI-driven data pipelines and cloud-based solutions on GCP. I integrated advanced Natural Language Processing (NLP) and Machine Learning (ML) techniques with real-time data ingestion and large-scale ETL workflows to enhance analytics-driven decision-making. This included real-time data ingestion, migration to BigQuery, implementing machine learning capabilities with BigQuery ML, and optimizing large-scale ETL workflows using Dataproc and Terraform for seamless scalability and analytics-driven decision-making.

- ∞ Engineered and deployed high-performance data ingestion pipelines utilizing Google Cloud Dataflow at Champion Energy, enabling seamless processing of both streaming and batch data with low latency and robust fault tolerance.
- ∞ Conducted extensive data migrations to Google BigQuery, following industry best practices to establish reliable real-time data streams and improve analytics capabilities.
- ∞ Championed the integration of Google BigQuery ML to facilitate in-warehouse machine learning operations, including model creation, training, and deployment on large datasets.
- ∞ Designed and implemented Snowflake data warehouses, creating dynamic data models and pipelines for real-time processing across multiple data sources.
- ∞ Managed and optimized Hadoop clusters using Google Cloud Dataproc, ensuring efficient data processing workflows and reducing operational costs with preemptible VMs and autoscaling.
- ∞ Built scalable ETL (Extract, Transform, Load) pipelines using Scala and PySpark to efficiently process and manage massive datasets.
- ∞ Developed dynamic Terraform templates to streamline infrastructure provisioning within GCP for agile deployment cycles.
- ∞ Orchestrated event-driven architectures using Google Cloud Functions, minimizing operational complexity by executing serverless functions triggered by cloud events.
- ∞ Enhanced data governance and processing efficiency by combining Google Cloud Data Prep and Dataflow, while enforcing robust security protocols using Google Cloud IAM.
- ∞ Leveraged Google Kubernetes Engine (GKE) to deploy containerized applications, utilizing Docker for efficient resource scaling and management.
- ∞ Partnered with data scientists and analysts to build comprehensive data pipelines and conduct deep analysis within the Google Cloud ecosystem.
- ∞ Monitored and managed Kafka topic and partition traffic to resolve bottlenecks and maintain consistent data flow.
- ∞ Secured Kafka clusters by implementing advanced encryption, authentication, and authorization mechanisms.
- ∞ Utilized Databricks to create efficient ETL pipelines, ensuring streamlined workflows for large-scale data processing.
- ∞ Automated data conversion and ETL pipeline generation with Python scripts, enhancing scalability and performance with Spark, YARN, and GCP Dataflow.
- ∞ Integrated Apache NiFi with diverse data sources and destinations, including databases, cloud storage, messaging systems, and APIs, to build robust data pipelines.
- ∞ Improved HiveQL queries and Python scripts for performance, optimizing resource usage and reducing execution times.
- ∞ Designed and optimized PostgreSQL databases, focusing on schema design, indexing, and performance tuning for large-scale queries.
- ∞ Configured GCP firewall rules to efficiently manage traffic for VM instances, improving content delivery via GCP Cloud CDN.
- ∞ Employed various GCP services, including Compute Engine, Cloud Load Balancing, Cloud Storage, Cloud SQL, Stackdriver Monitoring, and Deployment Manager, to architect end-to-end solutions.
- ∞ Led the migration of enterprise-grade applications to GCP, ensuring scalability, reliability, and seamless integration with existing on-premises systems.
- ∞ Applied natural language processing tools to analyze data, extract insights, and generate visualizations like word clouds.
- ∞ Mentored engineers on ETL best practices, Snowflake, and Python, while managing and optimizing ETL pipelines using Apache Spark and Python on GCP.
- ∞ Designed and developed scalable big data applications using Scala, with frameworks like Apache Spark and Akka.
- ∞ Fostered innovation by leveraging GCP Cloud Source Repositories for collaborative development and source code management.
- ∞ Collaborated with data scientists and analysts on projects such as fraud detection, risk assessment, and customer segmentation, contributing to business recovery strategies and implementation.

Gen AI (LLM, NLP) & ML:

- ∞ Developed and fine-tuned domain-specific Large Language Models (LLMs) using PyTorch and TensorFlow, optimizing model performance for biomedical text processing and retrieval.
- ∞ Implemented Retrieval-Augmented Generation (RAG) pipelines using LangChain, pgvector, and FAISS to enhance contextual understanding and response accuracy in clinical NLP applications.
- ∞ Engineered scalable NLP workflows leveraging Apache Spark and Airflow to process terabyte-scale biomedical literature and regulatory documents in real-time.
- ∞ Led MLOps initiatives by implementing CI/CD pipelines for NLP model training, evaluation, and deployment using GitHub Actions, Docker, and Kubernetes.
- ∞ Integrated multi-modal AI models that combine text, image, and genomic data for enhanced biomedical insights, utilizing GCP AI services and TensorFlow Extended (TFX).

∞ Built interactive NLP and OCR-powered dashboards using FastAPI, Streamlit, and Dash, enabling real-time querying, document processing, and visualization of AI-driven insights for researchers, with seamless integration of LLMs, Transformers, and vector databases for enhanced search and retrieval.

Tech Stack: GCP, Google BigQuery, Dataflow, BigQuery ML, Dataproc, Snowflake, Terraform, PySpark, Scala, Google Cloud Functions, Google Kubernetes Engine (GKE), Docker, Data Prep, IAM, Kafka, Apache NiFi, Databricks, ETL, Spark, YARN, PostgreSQL, HiveQL, Python, Compute Engine, Cloud Load Balancing, Cloud Storage, Cloud SQL, Stackdriver Monitoring, Deployment Manager, Cloud CDN, natural language processing, machine learning, event-driven architecture, data governance, fault tolerance, real-time data streaming, container orchestration, infrastructure as code, fraud detection, customer segmentation, risk assessment, business recovery strategies.

Cloud & AI Platforms: GCP (BigQuery ML, Vertex AI, Dataflow, Dataproc), AWS SageMaker
Machine Learning & NLP: PyTorch, TensorFlow, Hugging Face Transformers, OCR, LangChain, FAISS, Pinecone
LLM & AI Operations: RAG, MLOps, LLMOps, Retrieval-Augmented Generation, AI Model Deployment

Sr. Azure Cloud Data Engineer
Carrefour, Philadelphia, PA | Nov 2019 - Feb 2021

As a Sr. Azure Cloud Data Engineer at Carrefour, I managed a comprehensive data migration initiative to Azure, utilizing Azure HDInsight, Databricks, and Data Factory. I enhanced data processing efficiency by applying Hive partitioning and RDD caching techniques. During the migration, I transitioned data from Oracle and SQL Server to Azure storage solutions, leveraging Stream Analytics for real-time processing. I implemented stringent data security protocols and closely monitored migration for peak performance. Through optimized Python and Scala code and the automation of ETL workflows, I transformed the data infrastructure and delivered actionable insights via advanced data visualization tools.

∞ Migrated data from Oracle and SQL Server to Azure Blob Storage and Azure HDInsight Hive using Azure Data Factory, ensuring smooth and seamless data transfer.
∞ Designed and implemented Hive partitioning strategies within Azure HDInsight to optimize data separation and enhance processing efficiency, following best practices.
∞ Improved processing performance by caching RDDs in Azure Databricks, enabling efficient operations on large, distributed datasets.
∞ Utilized Azure Stream Analytics to facilitate real-time data processing during the migration, maintaining continuous data flow and consistency.
∞ Imported and processed large volumes of data (terabytes) into Spark RDDs for analysis, effectively integrating Azure Blob Storage for seamless data access.
∞ Ensured strong data security by leveraging Azure Active Directory and Key Vault for access control and encryption throughout the migration process.
∞ Mapped data to optimal Azure storage solutions like Blob Storage, Data Lake Storage, and Azure SQL Data Warehouse based on specific storage and analytical requirements.
∞ Used Azure Monitor to track migration progress, optimize performance, and automate repetitive tasks with Logic Apps for greater efficiency.
∞ Conducted thorough testing to verify data integrity, performance, and scalability across both RDBMS (MySQL, MS SQL Server) and NoSQL databases.
∞ Managed job scheduling and file systems on Azure Linux Virtual Machines with UNIX shell scripting to enhance operational workflows.
∞ Migrated legacy MapReduce jobs to PySpark in Azure HDInsight, boosting processing speed and scalability for large datasets.
∞ Continuously fine-tuned MySQL and NoSQL databases, identifying performance enhancements to improve system efficiency.
∞ Generated data frames from multiple sources such as RDDs, JSON datasets, and databases within Azure Databricks to simplify data analysis.
∞ Developed data visualization dashboards using Tableau and Power BI to present complex datasets as actionable insights for business stakeholders.
∞ Wrote optimized and maintainable Python and Scala code within Azure Databricks, utilizing built-in libraries for specific data processing needs.
∞ Automated ETL workflows with UNIX shell scripts, ensuring efficient scheduling, error handling, file management, and data transfer via Azure Blob Storage.

Tech Stack: Azure, Azure HDInsight, Databricks, Azure Data Factory, Hive, Spark, RDD caching, Stream Analytics, Azure Blob Storage, Azure SQL Data Warehouse, Azure Data Lake Storage, Azure Active Directory, Key Vault, Azure Monitor, Logic Apps, MySQL, NoSQL, PySpark, UNIX shell scripting, Linux, Azure Virtual Machines, Tableau, Power BI, ETL automation, Python, Scala, data migration, data security, data integrity, real-time data processing, job scheduling, data visualization, performance optimization, data transformation, cloud migration, legacy system modernization.

Sr. Data Engineer
Brystol Myers Squibb, New York, NY| May 2017 – Oct 2019

As a Senior Data Engineer at Bristol Myers Squibb, I led the migration of legacy data processing systems to a modern cloud-based platform, utilizing AWS services like S3 and EMR. I developed and executed a comprehensive strategy to modernize ETL workflows, ensuring seamless integration, enhanced performance, and streamlined operations within the cloud ecosystem.

- ∞ Led the migration of legacy scripts and orchestration to modernized SQL scripts and workflow orchestration, modernizing infrastructure using AWS services.
- ∞ Developed and executed a detailed migration strategy, including analysis, design, code migration, orchestration setup, and CI/CD integration, utilizing cloud-native tools and AWS services.
- ∞ Migrated core logic and orchestration code to a cloud-based platform, utilizing AWS EMR for optimized data processing and workflow management.
- ∞ Collaborated with cross-functional teams to define requirements and design migration strategies, aligning with business goals and leveraging AWS services such as Amazon S3, AWS EMR, and third-party tools.
- ∞ Performed development testing using SQL editors, job runs, compute resources, and data catalogs, ensuring compatibility with AWS services.
- ∞ Configured cloud environments within the AWS ecosystem and integrated them with native tools for automated deployment and consistent environment management.
- ∞ Managed code reviews and approval processes using collaborative features, ensuring adherence to coding standards and AWS integration.
- ∞ Coordinated thorough testing of migrated solutions using custom tools, resolving discrepancies and ensuring AWS compatibility.
- ∞ Automated ETL tasks with Python, boosting processing efficiency and minimizing manual intervention.
- ∞ Managed changes to job schedules using Shell and Python scripts, reflecting the migration and ensuring seamless AWS integration.
- ∞ Executed parallel runs of migrated processes alongside legacy systems to ensure consistency and accuracy within the AWS environment.
- ∞ Planned and executed the final cutover from legacy systems to the new platform, ensuring minimal disruption and leveraging AWS services for smooth transition.
- ∞ Conducted environment cleanup activities using development tools, Git, and AWS CLI, focusing on optimization within the AWS ecosystem.
- ∞ Designed and optimized algorithms for data processing using cloud-based tools and PySpark, ensuring full compatibility with AWS services.
- ∞ Conducted post-migration analysis to assess the performance of migrated pipelines, focusing on AWS metrics and overall efficiency.
- ∞ Identified further optimization opportunities based on stakeholder feedback, prioritizing improvements that enhance AWS integration and operational efficiency.

Tech Stack: AWS, S3, AWS EMR, Python, SQL, AWS CLI, AWS integration, ETL, cloud migration, CI/CD, workflow orchestration, shell scripting, AWS services, environment management, AWS metrics, data processing, PySpark, job scheduling tools, Git, development tools, code migration, data validation, performance optimization, legacy systems, cloud ecosystem, migration strategy, automated deployment, testing, data pipelines.

Hadoop Data Administrator
Nissan, Franklin, Tennessee| Jan 2015 – Apr 2017

As a Hadoop Data Administrator at Nissan, I managed and optimized the Hadoop ecosystem, ensuring seamless data processing and storage. I implemented efficient data workflows, leveraged HDFS for large-scale data management, and maintained the stability and performance of the Hadoop cluster, driving improved data accessibility and insights for the organization.

- ∞ Leveraged Pig, Python, and Oracle to conduct in-depth data profiling and transformation, ensuring raw datasets were prepared for analysis.
- ∞ Designed and optimized Hive external tables for efficient data storage, loading, and querying using HQL, facilitating quick and reliable data retrieval.
- ∞ Utilized Sqoop to transfer large datasets between relational databases and HDFS, and integrated Flume for real-time streaming of server logs into the Hadoop ecosystem.
- ∞ Enhanced legacy Hadoop algorithms by integrating Spark technologies, such as Spark Context, DataFrames, Spark SQL, Paired RDDs, and Spark YARN, significantly boosting performance.
- ∞ Developed and managed ETL pipelines to extract raw data, transform it into structured formats, and load it into target data stores, ensuring accuracy, consistency, and availability for reporting and analysis.
- ∞ Engineered big data ingestion and processing solutions using HBase, Hive, and MapReduce for diverse data sources.
- ∞ Created high-performance Spark code in Scala and Spark SQL to accelerate data processing and improve testing efficiency.
- ∞ Efficiently imported millions of structured records from relational databases into HDFS using Sqoop, ensuring data was stored in CSV format for Spark-based processing.

Tech Stack: Hadoop, HDFS, Pig, Python, Oracle, Hive, HQL, Sqoop, Flume, Spark, Spark Context, DataFrames, Spark SQL, Paired RDDs, Spark YARN, HBase, MapReduce, ETL pipelines, big data ingestion, data transformation, Scala, data profiling, data storage, data retrieval, real-time streaming, legacy algorithms, structured data, relational databases, data processing, data consistency, reporting, analysis, performance optimization, CSV, data accessibility, data stores.

Data Analyst
Zoom, San Jose, CA | Jan 2014 – Dec 2014

As a Data Analyst at Zoom Communications, I conducted in-depth analyses of large datasets to derive insights that guided key business decisions. I worked with cross-functional teams to establish KPIs, optimized marketing strategies through A/B testing, and created clear, actionable reports and dashboards that facilitated data-driven decision-making across the organization.

- ∞ Performed thorough analyses of large datasets to derive actionable insights that informed strategic business decisions.
- ∞ Worked closely with cross-functional teams to identify key performance indicators (KPIs) and assess the effectiveness of marketing campaigns.
- ∞ Simplified complex data sets into clear, visually engaging reports and dashboards, ensuring efficient communication of insights to stakeholders across all levels.
- ∞ Designed and managed dashboards to track business metrics, facilitating quick decision-making and performance monitoring.
- ∞ Led A/B testing and statistical analysis to evaluate the impact of various strategies, driving optimization efforts.
- ∞ Applied SQL and Python to process data, automate reporting workflows, and improve data collection efficiency.
- ∞ Kept up to date with industry developments and emerging analytical tools, constantly seeking innovative approaches to enhance data analysis capabilities.
- ∞ Implemented predictive modeling techniques to forecast trends and optimize business strategies, enhancing decision-making processes.
- ∞ Collaborated with data engineering teams to ensure the integrity, accuracy, and consistency of data across multiple systems, enabling reliable analysis and reporting.

EDUCATION

University of New Haven, Connecticut, USA
Master of Science – Computer Science
Bachelor of Technology – Information Technology