Sr. Data Engineer

Name:Rabiya

Email:rabiya1729@gmail.com

Phone:(419)971-8966

Professional Summary

- Seasoned Data Engineer with over 9 years of extensive experience in designing, implementing, and optimizing data solutions across various cloud platforms including GCP, AWS, and Azure.
- Proficient in leveraging GCP services such as BigQuery, DataPrep, and DataFlow to build scalable and efficient data processing pipelines.
- Skilled in cloud storage technologies like GCS Bucket and Cloud Storage for managing and storing large volumes of data securely.
- Experienced in working with Snowflake, Cloud SQL, and other relational databases for data warehousing and analytics purposes.
- Expertise in handling complex data integration tasks using tools like Matillion and performing machine learning tasks with BQ-ML.
- Strong background in data visualization and reporting using DataStudio, ensuring actionable insights for stakeholders.
- Proficient in programming languages including Python and shell scripting for automation and data manipulation tasks.
- Experienced in implementing data governance best practices and ensuring compliance with IAM security protocols.
- Well-versed in setting up VPC configurations and managing networking protocols for secure data transmission.
- Skilled in deploying and managing containerized applications using Docker and Kubernetes for scalable data solutions.
- Proficient in DevOps practices with hands-on experience in Jenkins, Ansible, and Git for continuous integration and deployment.
- Experienced in working with relational databases such as MySQL, MS-SQL, Oracle, and DB2, ensuring data integrity and performance.
- Familiar with Azure services including Azure Data Lake, Azure Data Factory, and Azure SQL for building data pipelines and analytics solutions.
- Proficient in data processing frameworks like Hadoop, Spark, and Kafka, ensuring efficient data processing and analysis at scale.
- Strong problem-solving skills with a focus on optimizing data workflows and improving overall system performance.
- Excellent communication and collaboration skills with the ability to work effectively in cross-functional

teams.

- Proven track record of delivering high-quality data solutions within specified timelines and budget constraints.
- Dedicated to staying updated with the latest trends and technologies in the field of data engineering to drive innovation and continuous improvement.

TECHNICAL SKILLS:

Category	Skills
Cloud Platforms	GCP, AWS, Azure
GCP	GCP BigQuery, GCP DataPrep, GCP Dataproc, Gcs Bucket, G-Cloud Function, GCP DataFlow, Gsutil, IAM Security, Service Data Transfer, Federated Queries, VPC Configuration, Data Catalog
AWS	EC2, S3, EBS, ELB, RDS, SNS, SQS, VPC, Redshift, CloudFormation, CloudWatch, ELK Stack, AWS Glue
Azure	Azure Datalake, Azure Datafactory, Azure AD, Azure Service Bus, Azure SQL, Cosmos DB, Log Analytics, AKS, Event Hub, Service Bus, Key Vault, App Insights, Azure VM creation, ACR, Azure Function App, Azure WebApp, Azure SQL MI, Azure DevOps
Databases	Snowflake, Cloud SQL, MySQL, MS-SQL, Oracle, DB2
Data Integration	Matillion, DataStage, QUALITYSTAGE, SSIS, SSAS, SSRS, Azure Data Factory, Kafka, Sqoop
Programming	Python, Shell Scripting, PowerShell, Java, Scala, Unix/Linux Shell Scripting
Big Data Technologies	Hadoop, MapReduce, Hive, HBase, Oozie, Airflow, Spark SQL, Spark DataFrames, pyspark, Kafka, Zookeeper
Containerization	Docker, Kubernetes, AWS ECS, Azure Kubernetes Service (AKS)
Version Control	GIT, SVN
CI/CD Tools	Jenkins, Bamboo, Azure DevOps
Monitoring Tools	Nagios, Splunk
Application Servers	JBOSS, WebLogic, WebSphere
Data Visualization	DataStudio, Tableau, Power BI, QlikView

PROFESSIONAL EXPERIENCE:

ROLE: Sr. GCP Data Engineer

Mayo Clinic, Rochester MN **February 2022 to Present**

Responsibilities:

- Implemented data processing pipelines on GCP DataFlow to efficiently handle large-scale data ingestion and processing tasks, ensuring optimal performance and scalability.
- Utilized GCP BigQuery for real-time analytics and to execute complex SQL queries on large datasets, facilitating data-driven decision-making processes within the organization.
- Designed and developed data models on BigQuery ML (BQ-ML) for predictive analytics, enabling proactive identification of potential health risks and opportunities for improvement.
- Managed and optimized GCP DataPrep workflows for data cleansing, transformation, and enrichment, ensuring data quality and integrity across various data sources.
- Worked on Data Migration on data sources within the legacy system, including patient records, appointment schedules, clinical notes, lab results, and billing information.
- Created a data mapping document that outlines how each data field in the legacy system corresponds to fields in the new system.
- Used BigPanda to parse incoming events to identify duplicates or updates to existing alerts, which can then be merged or discarded.
- Big Panda quickly drill down and isolate the root cause of incidents and outages.
- Configured and maintained GCP Dataproc clusters to process and analyze healthcare data efficiently, leveraging Apache Spark and Hadoop ecosystem technologies.
- Implemented and maintained GCS buckets for storing and managing structured and unstructured data, adhering to best practices for data security and access control.
- Developed and deployed G-Cloud Functions for serverless data processing tasks, automating routine data operations and improving overall workflow efficiency.
- Optimized data storage and retrieval processes using gsutil, ensuring fast and reliable data access for analytics and reporting purposes.
- Integrated and managed Snowflake data warehouse on GCP, enabling seamless data sharing and collaboration across different departments and stakeholders.
- Configured and administered Cloud SQL instances for relational database management, ensuring high availability, scalability, and data integrity.
- Implemented data replication and synchronization processes using Service Data Transfer, ensuring data consistency and reliability across distributed environments.
- Developed and deployed Matillion ETL jobs for data integration and transformation tasks, streamlining the data pipeline from source to destination.
- Created interactive and insightful dashboards using DataStudio, enabling stakeholders to visualize and

explore data trends and patterns effortlessly.

- Performed database administration tasks on MySQL, MS-SQL, Oracle, and DB2 databases, including installation, configuration, backup, and recovery.
- Designed and created interactive Tableau dashboards to visualize healthcare data, including patient demographics, clinical outcomes, and financial performance.
- Integrated data from multiple sources such as Electronic Health Records (EHR), patient management systems, and financial databases.
- Collaborated with the financial team to implement and maintain Maestro Financiero for financial data management and reporting purposes.
- Executed federated queries across multiple data sources using Cloud Data Catalog, ensuring comprehensive data discovery and accessibility for analytics.
- Implemented and enforced IAM security policies and access controls, ensuring data privacy and compliance with regulatory requirements.
- Configured VPC settings and VPN connections for secure network communication within the GCP environment, minimizing the risk of unauthorized access.
- Developed and deployed Pub/Sub messaging solutions for real-time data streaming and event-driven architecture, facilitating timely data processing and analysis.
- Designed and developed SSIS, SSAS, and SSRS packages for data integration, analysis, and reporting on Microsoft SQL Server platforms.
- Utilized DataStage and QualityStage for data integration, cleansing, and quality assurance tasks, ensuring data consistency and accuracy.
- Automated routine tasks using Python and Shell scripts, enhancing operational efficiency and reducing manual effort in data management processes.

Environment: GCP, GCP Big Query, GCP DataPrep, GCP Dataproc, Gcs Bucket, G-Cloud Function, GCP DataFlow, Gsutil, Snowflake, Cloud SQL, Cloud Storage, MySQL, MS-SQL, ORACLE, DB2, MAESTRO FINANCIERO, Federated Queries, IAM Security, Service Data Transfer, Matillion, BQ-ML, DataStudio, python, shell scripts, Federated Queries, VPC Configuration, Data Catalog. VPN Google-Client, Pub Sub, SSIS, SSAS, SSRS, DATASTAGE, QUALITYSTAGE.

ROLE: Senior AWS Data Engineer

Broadridge, Lake Success, NY **November 2019 to January 2022**

Responsibilities:

- Deployed and managed data processing pipelines on AWS infrastructure, leveraging services such as EC2, S3, and EBS for efficient data storage, computation, and scalability.

- Utilized AWS Redshift for large-scale data warehousing and analytics, optimizing queries and performance to meet stringent financial reporting requirements.
- Migrated the data sources within the legacy system to a cloud-based data warehouse (such as Amazon Redshift, Google BigQuery, or Snowflake) to leverage advanced analytics and improve decision-making processes.
- Chose and performed migration approach (big bang, phased, parallel) based on data criticality, downtime tolerance, and business continuity needs.
- Configured and maintained AWS ELB (Elastic Load Balancer) to ensure high availability and fault tolerance for data processing applications and services.
- Managed relational databases on AWS RDS (Relational Database Service), including MySQL and PostgreSQL, ensuring data consistency, integrity, and performance.
- Implemented event-driven architectures using AWS SNS (Simple Notification Service) and SQS (Simple Queue Service), enabling real-time data processing and analysis.
- Designed and implemented VPC (Virtual Private Cloud) configurations to establish secure network boundaries and control access to data resources.
- Automated infrastructure provisioning and management using AWS CloudFormation, streamlining deployment processes and improving scalability.
- Monitored and optimized system performance and resource utilization using AWS CloudWatch, identifying and resolving bottlenecks and inefficiencies.
- Implemented centralized log management and analysis using the ELK stack (Elasticsearch, Logstash, and Kibana), providing insights into system behavior and performance.
- Implemented AWS Glue for ETL (Extract, Transform, Load) tasks, automating data preparation and integration processes to support analytical workloads.
- Utilized Jenkins for continuous integration and deployment (CI/CD) of data engineering pipelines, ensuring reliable and efficient delivery of data solutions.
- Managed infrastructure configuration and deployment using Ansible, ensuring consistency and reliability across multiple environments.
- Developed and maintained Python and Shell scripts for automation of data processing tasks and system administration activities.
- Utilized PowerShell scripting for Windows environment automation and management, ensuring seamless integration with AWS services.
- Collaborated with development teams using GIT for version control, ensuring code integrity and facilitating collaborative development workflows.
- Implemented microservices architecture using AWS ECS (Elastic Container Service) and Kubernetes, enabling modular and scalable data processing solutions.
- Utilized Jira for project management and issue tracking, ensuring timely resolution of data engineering tasks and deliverables.

- Deployed and managed JBoss application server environments, ensuring reliability and performance of data processing applications.
- Implemented Bamboo for continuous integration and deployment (CI/CD) pipelines, automating the build, test, and deployment process.
- Utilized Docker for containerization of data engineering applications, enabling consistent deployment across different environments.
- Managed WebLogic and WebSphere application server environments, ensuring availability and performance of mission-critical data processing applications.
- Utilized Maven for dependency management and build automation, ensuring consistency and reproducibility of data engineering workflows.
- Managed and maintained Unix/Linux server environments, ensuring security, stability, and performance of data processing systems.
- Implemented monitoring and alerting solutions using Nagios, ensuring proactive identification and resolution of system issues.
- Utilized Splunk for log analysis and troubleshooting, providing insights into system performance and behavior to support data engineering operations.

Environment: AWS (EC2, S3, EBS, ELB, RDS, SNS, SQS, VPC, Redshift, Cloud formation, CloudWatch, ELK Stack), Jenkins, Ansible, Python, Shell Scripting, PowerShell, GIT, Microservice, Jira, JBOSS, Bamboo, Kubernetes, Docker, Web Logic, Maven, Web sphere, Unix/Linux, Nagios, Splunk, AWS Glue.

ROLE: Azure Data Engineer

Amway Corp ADA, MI **November 2017 to October 2019**

Responsibilities:

- Architected and deployed scalable data solutions on Azure, leveraging Azure Data Bricks and Azure Data Factory to orchestrate complex ETL processes, resulting in a 40% reduction in data processing times.
- Worked on migration of data from On-prem SQL server to Cloud databases (Azure Synapse Analytics (DW) & Azure SQL DB).
- Performed Data Ingestion one or more Azure Services - (Azure Data Lake, Azure Storage, Azure SQL, Azure DW) and processing the data in In Azure Databricks.
- Pipelines were created in Azure Data Factory utilizing Linked Services/Datasets/Pipeline/ to extract, transform, and load data from many sources such as Azure SQL, Blob storage, Azure SQL Data warehouse, write-back tool, and backwards.
- Architected and implemented scalable data warehousing solutions using Snowflake, optimizing storage and compute resources

- Developed and optimized data models for Snowflake, ensuring efficient storage, retrieval, and analysis of large datasets, contributing to a 30% improvement in query performance.
- Used Azure ML to build, test and deploy predictive analytics solutions based on data.
- Developed Spark applications with Azure Data Factory and Spark-SQL for data extraction, transformation, and aggregation from different file formats to analyze and transform the data to uncover insights into customer usage patterns.
- Developed advanced Python scripts for data extraction, transformation, and loading (ETL), enhancing data quality and enabling sophisticated data analysis to inform strategic business decisions.
- Collaborating with BI developers to establish and maintain data connections using OBIEE adapters.
- Utilized SSRS, SSIS and Cognos to create detailed and dynamic reports for business stakeholders, ensuring data accuracy and actionable insights.
- Streamlined report generation processes, reducing delivery time by 30% through the implementation of automated scheduling and data extraction techniques.
- Created and integrated test automation frameworks within Apache Airflow using tools like pytest and Airflow's own testing utilities, ensuring the accuracy and integrity of data pipelines.
- Integrated Airflow with CI/CD pipelines using GitLab CI, automating the deployment and testing of data workflows, leading to more efficient and reliable data processing.
- Conducted end-to-end testing of Airflow workflows, leveraging automated tests to validate data transformations, ensure data quality, and maintain system performance.
- Using Terraform templates, setting up the Azure infrastructure (IaaS) for Azure Databricks deployment like workspace, clusters, Databricks local groups, private endpoints, workflows, and Key Vault etc.
- Successfully completed a proof of concept for Azure implementation, with the larger goal of migrating on-premises servers and data to the cloud.
- Worked on pivot tables and slicers in excel. And I use excel frequently for any data related tasks.
- Worked on other tools like power BI and Tableau for data visualization.

Environment: Azure, Red Hat Linux,Azure Datalake, Azure Datafactory, Jenkins, Ansible, Shell Scripting, Azure AD, Azure Service Bus, Azure SQL, Cosmos DB, Log Analytics, AKS, Event Hub, Service Bus, Key Vault, App Insights, Azure VM creation, ACR, Azure Function App, Azure WebApp, Azure SQL, and Azure SQL MI, SSH, YAML, WebLogic, Python, Azure DevOps, Git, Maven, Jira

ROLE: Data Analyst

Pennant Technologies **August 2014 to July 2017**

Responsibilities:

- Analyzed large datasets stored in HDFS (Hadoop Distributed File System) using MapReduce and Spark SQL to extract valuable insights and trends for software performance optimization.

- Performed data extraction, transformation, and loading (ETL) tasks using Sqoop and Apache Kafka, ensuring seamless integration of data from various sources into analytical platforms.
- Utilized Hive for data warehousing and SQL-like querying of structured data, facilitating efficient data analysis and reporting processes for software development metrics.
- Designed and implemented data pipelines using Apache Oozie and Apache Airflow, automating data workflows and scheduling data processing tasks for timely insights delivery.
- Conducted exploratory data analysis (EDA) using Spark DataFrames and pyspark, identifying patterns and anomalies in software performance and user behavior.
- Developed and executed complex SQL queries to extract and manipulate data from relational databases, ensuring data accuracy and integrity for analysis purposes.
- Utilized HBase for real-time access to large datasets, enabling rapid retrieval and analysis of software logs and performance metrics.
- Implemented data ingestion processes from external systems into HDFS using Sqoop, ensuring efficient data transfer and synchronization.
- Utilized Apache Kafka and Zookeeper for real-time data streaming and event processing, enabling proactive monitoring and analysis of software events and alerts.
- Performed data visualization and reporting using tools such as Tableau and Power BI, communicating insights and findings to stakeholders effectively.
- Developed scripts in Linux Shell and Java for automation of data processing tasks and system monitoring, streamlining data analysis workflows.
- Collaborated with software development teams to identify key performance indicators (KPIs) and metrics, providing data-driven insights to support decision-making.
- Utilized Scala programming language for data manipulation and analysis tasks in Spark environment, leveraging functional programming capabilities for efficient data processing.
- Utilized Amazon AWS S3 for scalable and cost-effective storage of large datasets, ensuring data availability and durability for analysis purposes.
- Performed data parsing and manipulation of semi-structured data formats such as JSON, enabling integration of data from diverse sources into analytical pipelines.

Environment: HDFS, MapReduce, Hive, Sqoop, HBase, Oozie, Airflow, Sqoop, Kafka, Zookeeper, Spark SQL, Spark Data frames, pyspark, Scala, Amazon AWS S3, Java, JSON, SQL and Linux Shell Scripting.