**Rupika**

Sr. Data Engineer

📧 Email: krupika1708@gmail.com | 📱 Phone: (732) 314-4035

**PROFESSIONAL SUMMARY**

* Experienced Senior Data Engineer with around 9 years of expertise in guiding end-to-end data solutions and implementing robust architectures.
* Proficient in a diverse range of technologies and tools, encompassing cloud platforms such as AWS (EC2, S3, Athena, Data Pipeline, Lambda, Step Functions, IAM, EMR, CloudFormation, Lake Formation, Redshift, Kinesis, CloudWatch, Glue, DynamoDB), Azure (Data Factory, Databricks, ARM Templates, DevOps), and GCP (Google Cloud Storage, BigQuery, Dataflow, Pub/Sub).
* Skilled in programming languages including Python, PySpark, Scala, and data manipulation libraries such as Pandas and NumPy.
* Created Data Quality Scripts using SQL to validate successful data load and quality of the data.
* Proficient in relational database applications, data analysis, PL/SQL, stored procedures, and application migration.
* Adept at ETL processes with tools like DBT, AWS Glue, Azure Data Factory, SSIS, and Informatica, Oracle PL/SQL, SQL, Java, Attaccoma, Data stage.
* Proficient in containerization with Docker and Kubernetes and experienced in version control systems like Git, GitHub, Bitbucket, and continuous integration tools, including Jenkins and Azure DevOps.
* Adept at developing and maintaining scalable, high-performance data extract jobs using Snowflake, ensuring efficient extraction, transformation, and loading of large volumes of data to support organizational objectives.
* Skilled in data warehousing and modeling techniques within Snowflake, leveraging its architecture for effective data storage, retrieval, and analysis, contributing to robust and scalable data solutions.
* Extensive knowledge in infrastructure as code with AWS CloudFormation, ARM Templates, and Terraform.
* Proficient in data visualization tools such as Amazon QuickSight, Power BI, and Tableau.
* Developed intuitive and visually appealing dashboards using Amazon QuickSight
* Well-versed in handling diverse data formats, including JSON, XML, Avro, and Parquet. Experienced in monitoring and logging using Prometheus, Grafana, and CloudWatch.
* Demonstrated expertise in working with various databases, including MySQL, MongoDB, Cassandra, and Oracle.
* Skilled in data warehousing solutions like DBT models, Amazon Redshift, Snowflake, and Google BigQuery.
* Created and maintained scalable, high-performance data extract jobs to ensure efficient data processing and retrieval.
* Worked on Data Migration from Data Stage to AWS Snowflake environment using DBT.
* Familiar with project management tools like Jira and well-versed in Agile and Scrum methodologies.
* Strong problem-solving skills, excellent communication, and a collaborative approach to delivering high-quality data solutions.

**TECHNICAL SKILLS:**

**Cloud:** AWS, Azure, GCP

**Big Data:** Hadoop, Spark, HDFS, PySpark, Hive, Pig, Sqoop

**Databases:** MySQL, MongoDB, Cassandra, Oracle

**Data Warehousing:** Amazon Redshift, Snowflake, Google BigQuery

**Data Processing and Analysis:** Python, PySpark, Scala, Pandas, NumPy

**ETL:** AWS Glue, Azure Data Factory, SSIS, Informatica

**Data Formats:** JSON, XML, Avro, Parquet

**Containerization and Orchestration:** Docker, Kubernetes

**Version Control & CI/CD:** Git, GitHub, Bitbucket, Jenkins, Azure DevOps

**Infrastructure as Code:** AWS CloudFormation, ARM Templates, Terraform

**Visualization and Reporting:** Amazon QuickSight, Power BI, Tableau

**Monitoring and Alerting:** Prometheus, Grafana, CloudWatch

**Collaboration and Project Management:** Jira, Agile, Scrum

**WORK EXPERIENCE:**

**Walmart, Milpitas, CA | Sr. Data Engineer | May 2022 – Present**

**Roles & Responsibilities:**

* Managed the entire Software Development Life Cycle (SDLC) for data engineering projects, ensuring seamless integration and deployment.
* Implemented and optimized AWS services such as EC2 and S3 for efficient data storage and computation.
* Utilized Athena for ad-hoc querying and Data Pipeline for orchestrating complex data workflows.
* Developed serverless applications using Lambda functions and orchestrated workflows with Step Functions.
* Implemented robust IAM policies to ensure secure access control for AWS resources and RESTful APIs.
* Applied expertise in EMR, Hadoop, and Spark to efficiently process and analyze large-scale datasets.
* Developed Spark programs to load, parse, refined and store sensor data into Hadoop and process analyzed and aggregated data for visualizations.
* Utilized HDFS and PySpark to manipulate and transform data, employing MapReduce for parallel processing.
* Implemented Hive and Pig for data processing tasks, ensuring optimal performance and scalability.
* Utilized Python for scripting and automation in data engineering workflows.
* Implemented AWS CloudFormation for infrastructure automation and AWS Lake Formation for data lake management.
* Developed PL/SQL scripts as part of the data quality assurance process, to execute data purification, enrichment, validation, and profiling. Built-in functions and procedural logic were utilized to detect and resolve data anomalies and inconsistencies.
* Designed and optimized Amazon Redshift for high-performance data warehousing.
* Implemented Kinesis for real-time data streaming, ensuring timely and accurate data processing.
* Managed the development and implementation of a real-time data streaming and analytics system built on Apache Kafka, Apache Flink, and cloud-based managed services like Amazon Kinesis.
* Orchestrated containerized applications using Docker and managed container orchestration with Kubernetes.
* Implemented and optimized DynamoDB for efficient storage and retrieval of NoSQL data.
* Utilized AWS Glue for ETL processing, ensuring data integrity and consistency.
* Implemented dbt models to streamline real-time data workflows and utilized dbt's modular SQL-based approach to create and manage data transformations.
* Managed codebase efficiently using Bitbucket for version control.
* Implemented Jenkins for automated CI/CD pipelines, ensuring rapid and reliable software delivery.
* Proficient in handling data in JSON and XML formats.
* Collaborated in an Agile environment using Jira, actively participating in Scrum ceremonies.
* Implemented CloudWatch for real-time monitoring and logging of AWS resources.
* Developed visualizations using Amazon QuickSight for meaningful insights and reporting.

**Environment:** AWS, RESTful APIs, Hadoop, Spark, HDFS, PySpark, MapReduce, Hive, Pig, Python, CloudFormation, Redshift, Kinesis, Docker, Kubernetes, DynamoDB, AWS Glue, Bitbucket, Jenkins, JSON, XML, Jira, Agile, Scrum, Cloud Watch, QuickSight.

**Horizon BCBS, Newark, NJ | Data Engineer | March 2020 - April 2022**

**Roles & Responsibilities:**

* Implemented end-to-end data pipelines using Azure Data Factory, ensuring efficient and timely ETL processes.
* Utilized Azure Databricks to process large datasets, optimizing Spark jobs for enhanced performance.
* Integrated Snowflake data warehouse for scalable and flexible cloud-based analytics, enabling advanced data analytics and insights generation.
* Created scalable data pipelines that use Apache Kafka, Azure Functions, or Azure Event Hubs for event-driven data processing and ingestion to ingest streaming data into Cosmos DB.
* Skilled in utilizing Snowflake's capabilities to translate complex technical and functional requirements into detailed designs, fostering effective communication and alignment between technical and business stakeholders.
* Orchestrated infrastructure deployment with ARM templates, ensuring seamless resource provisioning on Azure.
* Developed Python scripts for data transformation and manipulation, enhancing ETL capabilities.
* Proficiently utilized Informatica for data extraction, transformation, and loading processes.
* Managed large-scale data storage and processing using Hadoop, ensuring scalability and reliability.
* Designing the data transformation mappings and data quality verification programs using Informatica and PL/SQL.
* Developed Spark applications in Scala, optimizing data processing workflows.
* Implemented and optimized Hive and Pig scripts for efficient data querying and processing on Hadoop.
* Executed data transfers between Hadoop and relational databases using Sqoop.
* Maintained version control of the codebase using Git, ensuring collaboration and code integrity.
* Implemented CI/CD pipelines with Azure DevOps, automating data testing and deployment processes.
* Designed and maintained MySQL databases, ensuring data integrity and high availability.
* Managed NoSQL databases like Cassandra, optimizing data storage and retrieval.
* Implemented data serialization formats like Avro and Parquet for efficient storage and processing.
* Developed and maintained shell scripts for automating routine data engineering tasks.
* Created interactive and insightful data visualizations using Power BI.
* Defined and deployed infrastructure as code using ARM templates for Azure services.
* Implemented monitoring and alerting solutions with Prometheus and managed project workflows using JIRA.

**Environment:** Azure, ARM templates, Python, ETL, Informatica, Hadoop, Scala, Sqoop, PySpark, Git, Azure DevOps, MySQL, Cassandra, Avro, Parquet, Snowflake, Shell scripting, Power BI, Prometheus, JIRA.

**Olio Global AdTech, Mumbai, India | Data Engineer | Nov 2017 - Jan 2020**

**Roles & Responsibilities:**

* Designed and implemented robust ETL processes using SSIS for efficient data extraction, transformation, and loading.
* Managed and optimized data storage and processing using the Hadoop ecosystem, ensuring high-performance data workflows.
* Developed Spark applications to process large datasets, improving data processing speed and efficiency.
* Utilized Hive and Pig for data processing and querying within the Hadoop ecosystem, ensuring seamless data analysis.
* Integrated REST APIs to extract and push data, leveraging Python for scripting and automation in data pipelines.
* Implemented and maintained databases in MySQL and MongoDB, ensuring data integrity and efficient retrieval.
* Collaborated on code repositories using Git and GitHub, facilitating version control and team collaboration.
* Implemented build automation using Maven and ANT for streamlined project development.
* Worked in an Agile Scrum environment, participating in sprint planning, reviews, and retrospectives.
* Implemented monitoring and visualization solutions using Grafana for real-time data insights.
* Set up and maintained Jenkins pipelines for continuous integration and continuous delivery (CI/CD) of data applications.
* Automated infrastructure provisioning and management for dbt using Terraform for scalable and reproducible environments.
* Developed Tableau dashboards for data visualization and reporting, providing actionable insights.
* Managed project tasks and workflows using Jira, ensuring efficient project tracking and collaboration.

**Environment:** SSIS, Hadoop, Spark, Hive, Python, MySQL, MongoDB, GitHub, Maven, ANT, GCP, Agile, Scrum, Grafana, Jenkins, Terraform, Tableau, Jira.

**NewVision Software, Hyderabad, India | Business Intelligence (BI) Developer| Feb 2015 - Oct 2017**

**Roles & Responsibilities:**

* Constructed detailed data dictionaries outlining data structures, relationships, and definitions, enhancing communication and comprehension among cross-functional teams.
* Applied Python, Pandas, and NumPy for efficient data manipulation and analysis tasks.
* Executed SQL queries, and PL/SQL to extract, transform, and load (ETL) data, ensuring seamless data processing.
* Implemented advanced techniques such as partitioning, indexing, and parallel processing in PL/SQL scripts to streamline data extraction from source systems, transformation, and loading into the data warehouse.
* Utilized Hadoop and Spark for distributed data processing, enhancing scalability and performance.
* Employed Hive and Oracle databases to store and retrieve data efficiently.
* Demonstrated proficiency in Linux and Git for version control and collaborative development.
* Collaborated with MarTech teams, utilizing data insights for personalized customer engagement and interfacing with MarTech tools for campaign optimization.
* Applied Excel for data analysis and visualization tasks, providing actionable insights.
* Implemented ETL processes for handling XML data, ensuring data integrity and consistency.
* Tracked and managed project tasks using JIRA, ensuring timelines and goals were met.

**Environment:** Python, Pandas, NumPy, SQL, PL/SQL, Hadoop, Spark, Hive, Oracle, Linux, Git, Excel, ETL, XML, JIRA.