**Dinesh Donikena**

**AWS Data Engineer**

**Phone**: 469-573-2358

**Email**: dineshdd2748@gmail.com

**Professional Summary**

* Around 11 Years of experience in the field of Datawarehouse and AWS Data Engineering and Results-oriented and highly skilled professional with expertise in AWS, Snowflake, and Big Data technologies.
* Experienced in utilizing AWS Glue for ETL workflows, enabling efficient data extraction, transformation.
* Designed and implemented end-to-end data integration and ETL using AWS Glue, orchestrating data workflows to efficiently process and transform large-scale datasets.
* Proficient in managing user access and permissions to AWS services using IAM.
* Developed Python AWS serverless lambda with concurrent and multi-threading to make the process faster and asynchronously executing the callable.
* Proficient with Python web frameworks such as Flask or Django for building RESTful APIs and web services to expose data endpoints for consumption by other applications.
* Implemented data pipelines using AWS services including EC2, S3, Athena, Redshift, Glue, Lambda functions, Step functions, EMR, CloudWatch, DynamoDB, RDS, and Kinesis.
* Expertise in AWS S3 for scalable and cost-effective data storage and retrieval.
* Experience in overseeing data lifecycle policies within Amazon S3, effectively reducing storage expenses by automating the archival and removal of data according to utilization patterns and compliance guidelines.
* Implemented performance tuning in Redshift using Distribution Keys, Sort Keys, and Partitioning.
* Skilled in designing roles, views, and implementing performance tuning techniques to enhance Snowflake system performance.
* Experienced in partitioning strategies and multi-cluster warehouses in Snowflake to ensure optimal query performance and scalability.
* Proficient in using SnowSQL for complex data manipulation tasks and developing efficient data pipelines.
* Proficient in utilizing virtual warehouses, caching, and Snow pipe for real-time data ingestion and processing in Snowflake.
* Strong knowledge of Snowflake's time travel feature for auditing and analysing historical data.
* Extensive experience in leveraging window functions, Snowflake arrays, regular expressions, and JSON parsing for advanced data analysis and manipulation.
* Highly proficient in Snowflake scripting to automate ETL processes, data transformations, and data pipelines.
* Strong knowledge of AWS CloudWatch for monitoring and managing AWS resources, setting up alarms, and collecting metrics.
* Proficient in integrating AWS SNS and SQS for real-time event processing and messaging.
* Ensure compliance with regulatory standards such as HIPAA by maintaining confidentiality of provider data.
* Adept in utilizing Databricks Delta Lake and Unity Catalog for seamless data management and integration.
* Employed cost optimization strategies on AWS Databricks by leveraging spot instances and auto-scaling, resulting in significant cost savings.
* Proactively performed frequent optimizations on Databricks clusters and data pipelines to maintain optimal performance levels, ensuring smooth and reliable data processing.
* Proficient in developing and optimizing Spark and Spark Streaming applications for real-time data processing.
* Proficient in using Hadoop ecosystem tools such as Apache Hive, Apache Spark, and Apache Hadoop MapReduce for processing and analysing data stored in different file formats.
* Import the data from different sources like HDFS/HBase into Spark RDD and perform computations using PySpark to generate the output response.
* Proficient in working with various Hadoop file formats including ORC, Avro, Parquet, and CSV, with a deep understanding of their characteristics and optimal use cases.
* Experienced in scheduling and workflow management using IBM Tivoli, Control-M, Oozie, and Airflow for efficient job orchestration.
* Proficient in version control systems like Git, GitLab, and VSS for code repository management and collaboration.

|  |  |
| --- | --- |
| **Technical Skills** | |
| **AWS Services:** | S3, Redshift, EMR, SNS, SQS, Athena, Glue, CloudWatch, Kinesis, Route 53, IAM, EC2, Lambda, Quicksight, CloudFormation, API Gateway, DynamoDB. |
| **Big Data Technologies:** | HDFS, SQOOP, PySpark, Hive, MapReduce, Spark Streaming, HBASE. |
| **Hadoop Distribution:** | Cloudera, Horton Works. |
| **Languages:** | SQL, PL/SQL, T-SQL, Python, HiveQL, Scala. |
| **Operating System:** | Windows (XP/7/8/10/11), UNIX, LINUX, UBUNTU |
| **Database:** | Teradata, Oracle, SQL, PostgreSQL |
| **SQL Server Tools** | SSIS, SSMS, SSAS |
| **Scheduling:** | AWS Step Function, IBM Tivoli, Control-m, Oozie, Airflow. |
| **Version Control:** | GIT, GitHub, VSS. |
| **Methodology:** | Agile, Scrum. |
| **IDE & Build Tools:** | Eclipse, Visual Studio |
| **Data Visualization Tools** | Quicksight, PowerBI, Tableau, Seaborn, Plotly, Matplotlib |
| **Machine Learning** | Supervised, Semi-Supervised, Unsupervised Learning, A/B experimentation, Regression, Classification. |

**Work Experience:**

**Role: Sr AWS Data Engineer**   **Apr 2022 - Present**

**Client: Albertsons,** **North Phoenix, Arizona.**

**Responsibilities:**

* Developed ETL Workflows using AWS Glue to extract, transform and load data from various sources into S3.
* Leveraged AWS Glue's serverless architecture and auto-scaling capabilities to build cost-effective and scalable data pipelines, reducing infrastructure overhead and improving operational efficiency.
* Configured AWS Glue Data Catalog to catalog and organize metadata from various data sources, providing a centralized metadata repository for discovering, understanding, and querying datasets.
* Transformed data in S3 using AWS Glue and optimized it used techniques like partitioning, formatting.
* Developed and optimized complex SQL queries in Amazon Athena to analyze large-scale datasets stored in S3.
* Created and maintained data catalog in AWS Glue for seamless integration with Amazon Athena, ensuring data consistency and accessibility.
* Designed and implemented data streaming solutions using AWS Kinesis for real-time data processing.
* Implemented data governance and metadata management practices to ensure data integrity, lineage, and compliance across ORC, Avro, Parquet, and CSV datasets.
* Orchestrated data replication tasks using AWS Glue, AWS Step Functions, or AWS Lambda to automate data movement and transformation processes.
* Implemented event-driven architectures with AWS Lambda, enabling real-time processing and triggering of functions in response to various AWS services.
* Implemented logging, monitoring, and error handling mechanisms for APIs, enabling proactive issue detection and resolution.
* Implemented automation scripts in Python to provision and manage AWS resources using AWS SDK, enhancing operational efficiency.
* Proficient in utilizing Pandas library for data manipulation, cleaning tasks within Python environments.
* Experienced in working with JSON, XML, and other data serialization formats in Python to parse and transform structured and semi-structured data.
* Leveraged proprietary internal tools and SDKs to streamline data integration processes, ensuring seamless flow across diverse data sources.
* Developed and maintained data pipelines using PySpark, ensuring efficient data ingestion, cleaning, and transformation, leading to streamlined data workflows and enhanced data quality.
* Conducted performance tuning and optimization of PySpark jobs, fine-tuning parameters, and configurations to maximize resource utilization and minimize processing time.
* Designed and implemented replication workflows to synchronize changes in real-time or batch mode to AWS S3.
* Proficient in SQL programming, including writing stored procedures, functions, triggers, and packages to automate business logic and data processing tasks.
* Proficient in writing complex SQL queries to retrieve, update, and manipulate data from relational databases.
* Developed Snowflake scripting solutions for analysing and querying large volumes of data for insights for inventory optimization.
* Developed complex SnowSQL queries to extract, transform, and load data from various sources into Snowflake.
* Implemented partitioning techniques in Snowflake to improve query performance and data retrieval.
* Designed and implemented Snowflake stages to efficiently load data from various sources into Snowflake tables.
* Created and managed different types of tables in Snowflake, such as transient, temporary, and persistent tables.
* Configured and managed multi-cluster warehouses in Snowflake to handle high-concurrency workloads effectively.
* Defined roles and access privileges in Snowflake to ensure proper data security and governance.
* Implemented Snowflake caching mechanisms to improve query performance and reduce data transfer costs.
* Utilized snow pipe for real-time data ingestion into Snowflake, ensuring continuous data availability.
* Utilized AWS Quicksight to build interactive dashboard and visualizations on data for optimizing inventory.
* Developed and deployed an inventory optimization dashboard using Amazon Quicksight to provide real-time insights into inventory levels, turnover rates, and stockouts.
* Utilized AWS CloudWatch for monitoring and managing resources, setting up alarms, and collecting metrics.
* Designed and configured workflows for data processing and ETL pipelines.
* Utilized AWS Step functions to build complex workflow that triggers by events in response to data ingestion.
* Designed and implemented AWS CloudFormation templates to provision and manage infrastructure as code.
* Developed complex CloudFormation templates using YAML to define AWS resources such as S3 buckets, IAM policies, and Lambda functions.
* Defined user access using AWS IAM to ensure compliance with data governance and security policies.
* Utilized AWS Secrets Manager to mask credential and rotating keys after specific interval for enhanced security.
* Utilized Git, and GitLab for version control.

**Environment**: AWS Glue, AWS Kinesis DataStreams, API Gateway, AWS S3, Snowflake, AWS CloudWatch, AWS CloudFormation, AWS Quicksight, AWS Step functions, AWS IAM, AWS Secrets Manager, Git, GitLab.

**Role: AWS Data Engineer Sep 2020 - Mar 2022**

**Client: EXL Services, Maricopa, Arizona.**

**Responsibilities:**

* Developed and optimized ETL workflows using AWS Glue to extract, transform, and load data from diverse sources into Redshift for efficient data processing.
* Implemented data quality checks and data validation rules within AWS Glue jobs to ensure data accuracy, completeness, and consistency across the data lifecycle.
* Integrated AWS Glue with AWS services such as Amazon S3, RDS, Redshift, Athena to ingests, process, and analyze data across heterogenous data sources and formats.
* Designed and implemented data streaming solutions using AWS Kinesis, enabling real-time data processing and analysis.
* Configured and fine-tuned Redshift to achieve high-performance data processing and streamlined querying.
* Worked on AWS Data pipeline to configure data loads from S3 to Redshift.
* Developed and optimized Spark applications using PySpark to extract, transform, and load data on EMR clusters.
* Monitored and troubleshoot EMR cluster health, performance, and job execution using AWS CloudWatch logs and metrics.
* Proficient in designing, implementing, and optimizing Amazon Redshift data warehouses to support large-scale analytics and data processing workloads efficiently.
* Competent in managing Redshift clusters, including provisioning, scaling, monitoring and troubleshooting to ensure optimal performance.
* Experienced in schema design, table distribution, sort key selection and compression strategies in Amazon Redshift to maximize query performance and minimize storage costs.
* Leveraged Amazon Athena’s query capability to perform joins and aggregations across multiple data sources including Amazon S3, RDS and Redshift.
* Integrated Lambda functions with other AWS services such as S3, SQS, and SNS to orchestrate data workflows and build scalable serverless applications.
* Automated query execution and result delivery using AWS Lambda functions triggered by Amazon Athena query completion events, enhancing operational efficiency.
* Skilled in database schema design and normalized principles, ensuring efficient storage and retrieval of data.
* Skilled in developing data pipelines and ETL processes using python frameworks like Apache Airflow to automate data ingestions and transformation.
* Designed and deployed automated ETL workflows using AWS Lambda and S3 once data arrives in the raw buckets.
* Proficient in using python libraries like SQLAlchemy and pandas to interact with relational databases such as MySQL and PostgreSQL for data manipulation, querying, and analysis.
* Proficient in writing SQL scripts and stored procedures to automate repetitive tasks and implement business logic within the database.
* Proficient in using SQL to generate ad-hoc reports, dashboards, and data visualizations to support decision-making processes and business intelligence initiatives.
* Proficient in leveraging PySpark for large-scale data processing tasks, including data manipulation, transformation, and analysis, resulting in optimized performance and improved data insights.
* Integrated AWS SNS and SQS to enable real-time event processing and efficient messaging.
* Implemented robust IAM policies and roles to ensure secure user access and permissions for AWS resources.
* Proficient in integration Terraform into CI/CD pipelines for automate infrastructure provisioning and management.
* Experienced in state management with Terraform, ensuring accurate tracking of infrastructure changes.
* Experience working with automated snapshot backup, EC2 instances, cloud watch alarms and SNS notifications.
* Created calculated fields, parameters, and sets in Tableau to perform advanced data manipulation and support dynamic filtering and aggregation.
* Developed interactive dashboards and visualizations using Tableau Desktop to analyze and present complex datasets for business stakeholders.
* Implemented a data pipeline using Kafka, Spark and Hive for ingestion, transformation, and analysis of data.
* Import the data from different sources like HDFS/HBase into spark RDD and perform computations using PySpark to generate the output response.
* Handled importing of data from various data sources, performed transformation using Hive, and MapReduce, and loaded data into HDFS.
* Proficient in utilizing version control systems such as Git, GitLab, and VSS for efficient code repository management and collaborative development processes.
* Ensure compliance with data governance and security policies.

**Environment**: AWS Glue, Redshift, AWS SNS, AWS SQS, AWS Data Pipeline, IAM, AWS Kinesis, Hive, MapReduce, HDFS, Kafka, Spark, Terraform, EC2, Airflow, Git, GitLab, VSS.

**Role: Data Engineer Feb 2019 - Aug 2020**

**Client: Kaiser Permanente, Pleasanton, California**

**Responsibilities:**

* Imported data from MySQL to HDFS on a regular basis using Sqoop for efficient data loading.
* Performed aggregations on large volumes of data using Apache Spark and Scala and stored the results in the Hive data warehouse for further analysis.
* Implemented data transformation tasks for Facets API using big data tools such as Apache Spark and Apache Hadoop.
* Leveraged Kafka for real-time data ingestion and streaming, ensuring timely updates to Facets API.
* Worked extensively with Data Lakes and big data ecosystems, including Hadoop, Spark, Hortonworks, and Cloudera.
* Loaded and transformed structured, semi-structured, and unstructured data sets efficiently.
* Capable of integrating Python scripts with SQL databases, NoSQL databases, and data warehouses for data extraction, loading, and transformation tasks.
* Utilized PySpark and Spark SQL for faster testing and processing of data in Spark.
* Utilized Kafka and Spark Streaming to process streaming data for specific use cases.
* Developed data pipelines using Flume and Sqoop to ingest customer behavioural data into HDFS for analysis.
* Utilized various big data analytic tools, such as Hive and MapReduce, to analyze Hadoop clusters.
* Wrote Hive queries and used Hive QL to simulate MapReduce functionalities for data analysis and processing.
* Migrated data from RDBMS (Oracle) to Hadoop using Sqoop for efficient data processing.
* Implemented CI/CD pipelines for building and deploying projects in the Hadoop environment.
* Utilized JIRA for issue and project workflow management.
* Utilized PySpark and Spark SQL for faster testing and processing of data in Spark.
* Used Spark Streaming to process streaming data in batches for efficient batch processing.
* Leveraged Zookeeper to coordinate, synchronize, and serialize servers within clusters.
* Utilized the Oozie workflow engine for job scheduling in Hadoop.
* Utilized PySpark in SparkSQL for data analysis and processing.
* Used Git as a version control tool to maintain the code repository.

**Environment**: Sqoop, MYSQL, HDFS, Apache Spark Scala, Hive Hadoop, Cloudera, Kafka, MapReduce, Zookeeper, Oozie, Data Pipelines, RDBMS, Python, PySpark, Ambari, JIRA.

**Role: Big Data Developer May 2017 - Jan 2019**

**Client: CNA Chicago - Chicago, Illinois**

**Responsibilities:**

* Developed ETL jobs using Spark -Scala to migrate data from Oracle to new MySQL tables.
* Rigorously used Spark -Scala (RRDs, Data frames, Spark SQL) and Spark - Cassandra -Connector APIs for various tasks (Data migration, Business report generation, etc.)
* Developed Spark Streaming application for real-time sales analytics.
* Proficient in Python programming language, with a focus on data analysis, manipulation, and visualization using libraries such as NumPy, pandas, and Matplotlib.
* Prepared an ETL framework with the help of Sqoop, Pig, and Hive to be able to frequently bring in data from the source and make it available for consumption.
* Analysed the source data and handled it efficiently by modifying the data types. Used Excel sheets, flat files, and CSV files to generate PowerBI ad-hoc reports.
* Analysed the SQL scripts and designed the solution to implement using PySpark.
* Extracted the data from other data sources into HDFS using Sqoop.
* Extracted the data from MySQL into HDFS using Sqoop.
* Implemented automation for deployments by using YAML scripts for massive builds and releases.
* Apache Hive, Apache Pig, HBase, Apache Spark, Zookeeper, Flume, Kafka and Sqoop.
* Extensively worked on creating combiners, Partitioning, and distributed cache to improve the performance of MapReduce jobs.
* Worked on GIT to maintain source code in Git and GitHub repositories.

**Environment**: Hadoop, Hive, spark, PySpark, Sqoop, Spark SQL, Cassandra, YAML, ETL.

**Role: Hadoop Developer Feb 2013 - May 2017**

**Client: Deloitte, Orlando, Florida**

**Responsibilities:**

* Developed data validation processes, load processes, and test cases using PL/SQL, MySQL Stored Procedures, Functions, and Triggers in Oracle 9i.
* Implemented data partitions to enable parallel processing of data.
* Designed and maintained a fully integrated and multi-level help system for the modules.
* Leveraged SQL Loader for bulk loads from external data files into database tables.
* Created XML documents for generating dynamic SQL statements for different Compensation groups.
* Utilized Informatica extensively to extract data from databases and load it into the data warehouse repository.
* Performed SQL Tuning to optimize SQL queries for improved performance.
* Managed and optimized automated Oracle jobs using CRON Tab.
* Enhanced and fine-tuned PL/SQL code to improve the data load process.
* Executed logical and physical data model implementations and conducted reverse engineering using Erwin.

**Environment**: Oracle 9i, TOAD, SQL, SQL Loader, PL/SQL, Windows NT, UNIX, Forms 6i, Reports 6i, Informatica, Erwin, XML.