**Dilnaz R**

**Sr. Cloud Data Engineer/ GCP Data Engineer**

**+1-469-848-0845 |** dilnazw29@gmail.com**| Plano TX**

**PROFESSIONAL SUMMARY:**

* Around 10 years of IT experience in Data Analysis and Engineering with strong expertise in the planning and implementation of large-scale Big Data applications using technologies such as Apache Spark, Hadoop, Hive, Sqoop, Python & Scala.
* Collaborated with AWS and GCP cloud services, including Cloud Storage, Dataproc, Data Flow, Big Query, EMR, S3, Glacier, and EC2 with EMR Clusters.
* Proficiency in GCP services like Dataproc, Google Cloud Storage (GCS), Cloud Functions, and Big Query.
* Implemented and managed data warehousing solutions on GCP, including Big Query and Google Cloud SQL, for scalable and high-performance analytics.
* Designing and implementing ETL pipelines using Apache Airflow to automate data workflows.
* Developed and executed large-scale data processing workflows using Dataproc with tools like Hadoop, Spark, and Hive.
* Experience in creating the clusters on AWS EMR using Cloud Formation Script and integrating with the Jenkins job to trigger the Spin up and Spin down jobs.
* Extensive experience with Informatica (ETL Tool) for Data Extraction, Transformation and Loading.
* Extensive experience in SQL scripting and performing SQL operations such as Joins and Managing External tables.
* Experience in building batch and streaming ingestion flows in Apache Ni-Fi and Databricks.
* Extensive knowledge in Data storage and Analysis in Azure data storage environments like Azure SQL, Blob storage, Azure SQL, Azure Cosmos DB, and Polybase.
* Hands on experien ce in GCP, Big Query, GCS, cloud functions, Cloud dataflow, Pub/Sub, cloud shell, GSUTIL, bq command- line utilities, Data Proc.
* Hands-on experience in Big Data components like MapReduce, HDFS, HBase, Hive, Sqoop, Pig, Zookeeper, Oozie, Apache Spark, Impala.
* Worked related to downloading Big Query data into pandas or Spark data frames for advanced ETL capabilities.
* Migrate data into RV Data Pipeline using DataBricks, Spark SQL and Scala.
* Strong experience in Data Analysis, Data Profiling, Data Cleansing & Quality, Data Migration, Data Integration, Data Ingestion and Data Transformation.
* Used Databricks for encrypting data using server-side encryption.
* Data Ingestion to Azure Data Services such as Data Lake, Storage, SQL, DWand processing the data in In Azure Databricks.
* Knowledge in developing high-quality Azure Data Lakes and Data Warehouse and designing the team to scale.

**TECHNICAL SKILLS:**

|  |  |
| --- | --- |
| **Big Data Frameworks** | HDFS, MapReduce, Apache Spark, Apache Hive, YARN, HBase, Apache HBase, Spark Streaming, Spark SQL, Sqoop, Zookeeper, Kafka, Informatica PowerCenter |
| **Languages & Scripting** | Scala, Python, Shell Script, SQL, SAS. |
| **Hadoop Distribution** | Cloudera , EMR, Virtual Machines, Data Proc |
| **Cloud Technologies** | AWS EC2, S3, GCP, GCP Big Query, GCP Data Proc, EMR, Lambda, Kinesis, ELB, RDS, Glue, SNS, SQS, EBS, CloudFormation, Microsoft Azure Synapse, ADF, Data Lake Storage, Azure SQL, Pipelines |
| **Development Tools** | IntelliJ, Microsoft SQL Studio, Eclipse, SBT, Scala Unit |
| **Database** | MySQL, MS-SQL server, Postgres, Hive, HBase, MongoDB |
| **Operating Systems** | Unix, Linux, Windows, Mac OS |
| **Version Control** | GIT, Bitbucket |
| **BI Tools** | Tableau, Power BI. |

**PROFESSIONAL EXPERIENCE:**

**Client: Voyager Sopris Learning, Dallas, TX Jan 2022 – Present**

**Role: Sr. GCP Data Engineer**

**Responsibilities:**

* Proficiency in GCP services like Data proc, Google Cloud Storage (GCS), Cloud Functions, and Big Query.
* Developed structured streaming code to ingest data from Kafka to on-prem Hadoop Clusters (Cloudera)
* Designed and implemented data pipelines in GCP using Airflow, employing different airflow operators.
* Utilized GCP's Cloud Shell SDK to configure services like Data Proc, Storage, and Big Query.
* Created PySpark Code to run on the Airflow services and write the data to the GCP Buckets as Parquet files
* Created Data Frames and performed analysis using Spark SQL and used RDD and DF APIs to access variety of data sources using Scala, PySpark, Pandas and Python.
* Developed syncs between the GCP Buckets and Big Query tables.
* Create multiple external and internal tables on the GCP Big query.
* Write SQL Queries against the tables on Big Query and efficiently use the time travel to capture incremental data updates and find anomalies in the data using quality checks
* Worked in both GCP and AWS cloud environments parallelly.
* Efficiently created and used objects using Scala OOPS concepts and processed data using RDD and data frames.
* Storing data files in Google Cloud S3 Buckets daily basis, Using DataProc, Big Query to develop and maintain GCP cloud base solutions.
* Played a key role in migrating the data infrastructure to Google Cloud Platform (GCP), utilizing Big Query and Dataflow for data processing and analysis.
* Designed and implemented an Apache Airflow-based pipeline to process real-time data streams, resulting in up-to-the-minute insights for business stakeholders.
* Contributed to migrating on-premises Hadoop systems to GCP.
* Created the Scala Spark Streaming code as framework and utilized the jars and deployed multiple application to capture and load the streaming real-time data to Hadoop Storages.
* Developed Python and PySpark code for ETL and used them in AWS services like AWS Lambda, Glue, Ec2, EMR, etc.
* Performed information purging and applied changes utilizing Databricks and Spark information analysis.
* Develop the data consistency scripts by integrating the SQL files and connectors to the warehouses such as kudu impala and hive from Active Batch.
* Manage performance tuning methodology in optimizing SQL, ETL mappings, HIVE- managed / ORC tables.
* Worked extensively on writing various SQL queries for workflow specific data analysis.
* Extensively used PySpark, Python for building and modifying existing pipelines.
* Extensively utilized Databricks notebooks for interactive analysis utilizing Spark APIs.

**Environment:** Python, Spark, Scala, Hadoop, Cloudera, GCP, ETL, Hive, Shell Scripting, Databricks, Python, SQL, Python Modules.

**Client: Northern Trust, Chicago IL Jun 2019 – Dec 2021**

**Role: AWS Data Engineer**

**Responsibilities:**

* Developed the concept of In-State Management for Processing and Handling constant data updates by Storing Data in Spark in memory in a streaming application.
* Build a program with Python sdk with Apache beam framework and execute it in Cloud Dataflow to stream pub sub messages into big query tables.
* Expertise in AWS data migration between different database platforms like Local SQL Server to Amazon RDS, EMR HIVE, and experience in managing and reviewing Hadoop log files in AWS S3.
* Worked extensively with importing metadata into Hive using Python and migrated existing tables and applications to work on AWS cloud (S3).
* Installed application on AWS EC2, configured the storage on AWS S3 buckets, and worked closely with AWS EC2 infrastructure teams to troubleshoot complex issues.
* AWS EMR to process big data across Hadoop clusters of virtual servers on Amazon Simple Storage Service (S3).
* Built and supported several AWS, multi-server environment's using Amazon EC2, EMR, EBS, and Redshift and deployed the Big Data Hadoop application on the AWS cloud.
* Collaborated wif Data Analysts to understand the data and their end requirements and transform the data using PySpark, SparkSQL and created data frames
* Developed Schema using Scala Struct Types to parse and process ingestion data using data frames in Spark.
* Used Case classes to build data frames in Scala and write to different Kudu tables.
* Provide Application support for Big Data applications on Active Batch and Batch Ingestion pipelines on AWS.
* Provide Application Support for PySpark, Python Applications running on AWS, and Python Servers
* Develop PySpark code and apply Spark SQL Transformations on the data and load the data to S3 Buckets
* Develop EMR Spin up or Spin Down process using the AWS CloudFormation Script and integrated with Jenkins for the Spin up/Down.
* Created, provisioned different Databricks clusters needed for batch and continuous streaming data processing and installed the required libraries for the clusters.
* Creating ETL mappings and enhancing existing mappings to facilitate the data load in system.
* Developed hands-on experience on AWS services like S3, EMR, Kinesis, Glue, and CloudWatch.

**Environment:** AWS, Python, Spark, Spark Streaming, Databricks, Spark SQL, Snowflake, Redshift, S3, Glue, GCP, Cloud Watch, Shell Scripting, AWS EMR, SQL.

**Client: Norfolk Southern, Raleigh, NC Jan 2018 – May 2019**

**Role: Big Data Developer/** **Hadoop Developer**

**Responsibilities:**

* Data Ingestion from relational databases into HDFS using Sqoop import/export and also created Sqoop Job, Evaluate, and incremental jobs.
* Created Partitions, Bucketing and Indexing concepts for optimization as part of hive data modelling.
* Created several Databricks Spark jobs with Pyspark to perform several tables to table operations.
* Responsible for installation and configuration of Hive, Pig, Sqoop, Flume and Oozie on the Hadoop Cluster.
* Involved in developing Hive DDLs to create, alter and drop Hive tables.
* Involved in converting Hive/SQL queries into Spark transformations using Spark RDDs and experience in using Spark-Shell and Spark Streaming.
* Strong knowledge in NOSQL column-oriented databases like HBase and their integration with Hadoop cluster using connectors.
* Developed PySpark scripts for ETL using DataProc.
* Designed workflow by scheduling Hive processes for Log file data which is streamed into HDFS using Flume.
* Experienced in performing in memory batch processing using Spark Streaming (Spark and Spark-SQL and Spark-shell).
* Analyzed large and critical data sets using Cloudera, HDFS, HBase, Hive, UDF, Sqoop, YARN, PySpark and Apache Spark
* Tested Apache, an extensible framework for building high performance batch and interactive data processing applications, on Pig and Hive jobs.
* Pre-processed large sets of structured and semi-structured data, with different formats like Text Files, Avro, Parquet, ORC, Sequence Files, and JSON Record.
* Responsible for continuous monitoring and managing the Hadoop Cluster using Cloudera Manager.

**Environment:** Cloudera ,Linux, HDFS, MapReduce, Shell Scripting, Java, Talend, Hive, Pig, Spark, Databricks, Impala, Sqoop, Flume, Eclipse, Apache.

**Client: Tech Vedika, Hyderabad, India Jan 2014 – Nov 2017**

**Role: Big Data Developer**

**Responsibilities:**

* Performed the creation, update, and manipulate using SQL.
* Handled Slowly Changing Dimensions techniques in the ETL developments.
* Create Shell Scripts to trigger the SQL Codes on the Hadoop servers.
* Creating functional and technical ETL mapping specification document for data mappings.
* Wrote complex SQL, Procedures, Functions, and Packages to validate data and testing processes.
* Developed advanced correlated and un-correlated sub-queries in T- SQL to develop complex reports.
* Develop and deploy the outcome using spark and Scala code in Hadoop cluster running on GCP.
* Develop/support ETL transformation mapping and enhancing existing mappings to facilitate data load into DWH.
* SQL Scripts to perform Data Quality checks on the HDFS files after SAP data services workflows are executed.
* Support Data migration from SQL Warehouse to Hadoop warehouse, acquire expertise on the Hadoop storage systems.
* Developed weekly, monthly reports related to the marketing and financial departments using Teradata SQL.
* Improved the performance of the T- SQL queries and Stored procedures by using SQL profiler, Execution plan, SQL performance monitor, and Index tuning advisor.
* Developed ETL Specifications and Mappings using Informatica PowerCenter tool for Data loading.

**Environment:** Java, MySQL, Pig Scripts, SQL Scripts, ETL, Shell Scripts, HBase, Sqoop, Zookeeper, Oozie, Oracle, Shell Scripting.