PRABHAT TRIPATHI

Big Data Engineer

Email: Pt.trip108@gmail.com

Number: (469)- 251 -2146

**PROFESSIONAL SUMMARY**

* Around 8 years of experience in big data application development, utilizing Hadoop ecosystem components such as Hadoop, Spark, Hive, Sqoop, Apache Airflow, and Databricks.
* Skilled in leveraging various cloud services of Microsoft Azure, Google Cloud Platform (GCP), and Amazon Web Services (AWS).
* Demonstrating comprehensive knowledge and understanding of distributed computing and parallel processing frameworks.
* Proficient in executing read and write operations on Hadoop Distributed File System (HDFS).
* Adept in handling Hadoop/Spark distributions, specifically Cloudera.
* Expert in designing and developing applications in Spark, utilizing Python and Scala.
* Experienced in working with Spark Data Frames for the parallel processing of datasets from diverse data sources.
* Accomplished in utilizing Airflow for scheduling, managing, and monitoring Spark Jobs on clusters.
* Proficient in creating Hive Tables and loading data from various file formats.
* Capable of managing vast data sets within data warehouses.
* Experienced in data cleansing utilizing Spark Map and Filter Functions.
* Competent in importing and exporting data to Hive and HDFS using Sqoop.
* Experienced in processing data with HiveQL for data analytics.
* Implemented techniques like Partitioning and Bucketing in HIVE and as well as in Bigquery.
* Adept in converting Hive/SQL queries into Spark transformations.
* Proficient in working with large data sets and implementing performance improvements.
* Experienced in handling file formats such as CSV, JSON, Parquet, and ORC.
* Capable of creating and driving ETL pipelines.
* Possessed strong knowledge of UNIX/LINUX commands and shell scripting.
* Exceptional communication and presentation skills; always eager to learn and adapt to emerging technologies.
* Highly motivated with the ability to work independently or collaboratively as part of a team, upholding the highest levels of professionalism.
  + Top of Form

**Work Experience**

**Senior data Engineer, Austin, TX April 05/2023 to Till Date**

**PayPal**

In this project, I was engaged in the analysis and transformation of diverse financial data, spanning from credit card transactions to funding mix Type U transactions. Responsibilities included tailoring scripts to meet evolving business standards and conducting thorough Data Quality Checks (DQC) to ensure accuracy. This was essential before migrating data from Teradata EDW to Google's Bigquery platform, ensuring a seamless data transition with maintained data integrity.

* Utilized credit card payment data and modified the script within the framework module, implementing necessary transformations according to job specifications. Conducted thorough testing on Voyager platform prior to transferring data to the production environment.
* Engaged with diverse datasets beyond payment metrics, including shipping and orders, while emphasizing a specific focus on payment metrics such as debit card and credit card transactions.
* Implemented an SQL script to effectively handle and manipulate multiple exchange rates, including RBS, Bloomberg, and yearly plan rates, ensuring accurate conversion of currency values to "USD".
* Collaborated within the module to address exceptional records by utilizing various scripts. Implemented a system to store these exception records for reprocessing in subsequent loads. Additionally, created a financial exception archive table to store new exception records once they were resolved. Ensured the insertion of processed data at the end of each day, enabling the capture of the same data in the following day's processing, specifically for even transactions.
* Managed and implemented a comprehensive Regression Testing Framework, proficiently minimizing potential risks by maintaining consistent functionality of pre-existing code during the integration of new code or subsequent modifications.
* Developed an automated data comparison tool using regression frameworks, effectively reconciling two database tables. This solution delivered detailed reports encompassing match count, total mismatch count, and overall record count, enhancing data accuracy and integrity. And shows regression match percentage based on formula our team has developed.
* Collaboratively authored a comprehensive Confluence page detailing the intricacies of the Regression Framework. Diligently documented the setup and execution procedures, supplemented by clear illustrations of query execution steps.
* Identified and addressed issues pertaining to database tables within Bigquery that were overly resource-intensive and exhibited long processing times. Leveraged comprehensive analysis of these tables to suggest and implement optimization strategies, effectively enhancing system efficiency and processing speed. My proactive contributions significantly improved the team’s performance and led to more efficient database management.
* Authored an efficient SQL script tasked with determining the total slot milliseconds and total bytes processed when using the COALESCE function across different job IDs. This was particularly useful when executing two distinct SQL scripts, each incorporating the COALESCE function in a unique manner, thus enabling optimized performance measurement and enhanced script execution.

**Environments:** SQL, Python, PayPal Notebook (Jupyter Notebook), Linux and bash scripting, GitHub, Citrix, Voyager, Bigquery, GCP.

**Big data Engineer Mar 2020 to Jan 2023**

**Lumen Technologies**

The goal of this project was to centralize a variety of data in inexpensive storage, enable fast processing, and can be leveraged by various mechanisms, for varied business needs. Accelerate business growth by helping build new products, insights and enable AI and Machine Learning capabilities.

**Responsibilities:**

* Implemented Databricks notebook that reads raw data in parquet format and writes the data into Azure Blob Storage container in delta format, which allows for upsert operations on delta tables created in Databricks, for the incremental data ingestion.
* Worked on script loading data from GCP Buckets to Big Query tables, querying the tables and loading the result.
* Worked on writing dataflow code in python to load streaming data from cloud pub/sub to big query.
* Worked on migrating data from on-premises to GCP.
* Used Dataproc cluster to run pyspark jobs for transformation and processing of data and write output data in Google cloud storage.
* Used Dataproc cluster on Google Cloud Platform (GCP) to run spark and Hadoop MR jobs.
* Load data from google cloud buckets to big query (data warehouse in Google) partitioned table.
* Implemented an Airflow DAG to automate the workflow for data ingestion and monitored the tasks.
* Used Airflow to automate the pipelines to run every day using Python.
* Automated and applied ETL to the raw data using spark-Hive SQL every 15 min using Apache Airflow/Composer and save the transformed data in Big query using cloud function.
* Used cloud function to append data from cloud storage to big query table and ran query on top of it.
* Worked on fixing the mapping of tables, used in BQ queries. Analyzed the tables columns and data resulting from the queries and made required changes in them to match the results in both.
* Created Cloud composer environment and ran DAG on it.
* Worked on a Databricks notebook to load data from Azure Databricks to Azure SQL database.
* Implemented a Databricks notebook code to monitor a notebook job run's and to load the status to a delta table.
* Developed Spark applications using Pyspark and Spark-SQL for data extraction, transformation, and aggregation from multiple file formats for analyzing & transforming the data to uncover insights into the customer usage patterns.
* Used AWS services like EC2 and S3 for small data sets processing and storage, Experienced in Maintaining the Hadoop cluster on AWS EMR.
* Experience in building scalable distributed data solutions using an EMR cluster environment with Amazon EMR.
* Implementing the pre-processing procedures along with deployment using the AWS services and creating virtual machines using EC2.
* Worked on importing data using Sqoop from RDBMS to HDFS and copied data from HDFS to Azure Blob storage by using Distcp (distributed copy) tool.
* Optimized SQL query using Pyspark which fetched the required results faster, in a few minutes compared to SQL.

**Environments:** Azure Blob Storage, Azure data lake, GCP, GCP Big Query, Databricks, HDFS, Spark, Spark SQL, Kafka, Python, SQL, Sqoop, Pyspark, Apache Airflow, Linux shell, Agile, spark data frame, Amazon EC2, Amazon EMR, S3

**Spark/Hadoop Engineer Apr 2018 to Feb 2020**

**Pacific Gas and Electric, San Francisco, CA**

The project was to move historical sensor logs of sold machinery from RDBMS to Hive and perform data analysis using spark SQL and store output to hive for use by the R&D team and setup Kafka for incremental loading for new data from sensors to be appended directly to the respective Hive tables.

**Responsibilities**

* Worked on large sets of structured and semi-structured historical data.
* Involved in working with Sqoop to import the data from RDBMS to Hive.
* Created partitioned and bucketed Hive tables to load the Data.
* Involved in cleansing and transforming the data.
* Used spark SQL to perform sort, join and filter the data.
* Copied the ORC files to amazon s3 buckets using Sqoop for further processing in amazon EMR.
* Wrote custom UDF’s in Spark SQL using Pyspark.
* Performed data Aggregation operations using Spark SQL queries.
* Copied output data back to Hive from Amazon S3 buckets using Sqoop after getting the output desired by the business.
* Helped client to access data from data warehouse which is hive on our case.
* Setup Kafka to subscribe to topics and load data directly to the Hive table.
* Compared the sensor data to a persisted table on a 24hr period to check if the machine is operating at optimal conditions and Used Kafka as a messaging system to notify the producer of that data and the maintenance department in case a maintenance is required.

**Environments:** HDP, AWS EMR, EC2, S3, HDFS, Apache Spark, Kafka, Amazon S3, EMR, Sqoop, Hive, Linux shell, Agile, Python.

**Spark/Hadoop Engineer Jan 2016 to Mar 2018**

**American family insurance, Madison, WI**

The goal of this project (cognitive search for life science) was to get the clinical test logs from different servers and data from different databases in real-time into HDFS and redesigning the imported data into useable format by cleansing and performing transformations on the data to provide an aggregated overview of the trials to be accessed by the clients. And store the data in hive for further processing by data scientists.

**Responsibilities**

* Used Sqoop to fetch data from RDBMS.
* Ingested and transformed the data using spark SQL.
* Created partitioned and bucketed Hive tables to load the Data.
* Exporting Spark SQL Data frames into hive tables stored as ORC Files.
* Ingested real-time logs from various Kafka producers.
* Used spark streaming to subscribe to desired topics(drug) for real time processing.
* Performed data Cleansing to meet business requirements.
* Responsible for performing sort, join, aggregations, filter, and other transformations on the data.
* Involved in Analysing data using HiveQL queries.
* Loading data from Linux Filesystems to HDFS and vice-versa using shell commands.

**Environments:** CDH, HDFS, Spark, Hadoop, Scala, Hive, Hue, Oozie, Sqoop, Kafka, Linux shell, Agile, python.

**Education:**

* Bachelor’s in mechanical engineering- The University of Texas at Arlington, 2016