Name : Gaurav Raj Thapa

Email: tgauravraj18@gmail.com
Ph#: 501-518-1004

Professional Summary:

- ✓ Over **5+ years** of experience in Data Engineering, Data Pipeline Design, Development and Implementation as a **Data Engineer/Data Developer and Data Modeler.**
- ✓ Experience in **SDLC (Software Development Life Cycle)** and was involved in all phases in projects.
- ✓ Experience with **Spark Architecture** including **Spark Core, Spark SQL, Spark Streaming** and **Spark MLlib.**
- ✓ Experience in **Spark** using **Scala** for loading data from local file systems, HDFS, Amazon S3, Relational and NoSQL databases using Spark SQL, import data into RDD, DataFrames and ingesting the data from a range of sources using **Spark Streaming.**
- ✓ Experience in **Data Warehouse/Data mart, ODS, OLTP and OLAP** implementations teamed with project scope, Analysis, requirements gathering, data modelling, Effort Estimation, **ETL Design, development,** System testing, Implementation **and production support.**
- ✓ Experienced in using **Tableau Online** and **Tableau Server.**
- ✓ Hands on experience in working with **Tableau Desktop, Tableau Server** and **Tableau Reader** in various versions.
- ✓ Experience in building and publishing **POWER BI** reports utilizing complex calculated fields, table calculations, filters, parameters.
- ✓ Experience in streaming applications using **Apache Kafka.**
- ✓ Experience fine-tuning Spark applications utilizing various concepts like Broadcasting, increasing shuffle parallelism, caching/persisting **Data** Frames, sizing executors appropriately to utilize the available resources in the cluster effectively etc.,
- ✓ Experience in **Data Analysis, Data Validation, Data Cleansing, Data Verification** and identifying data mismatch, data quality, metadata management, and master data management.
- ✓ Experience structural modifications using **Map-Reduce** and analyze data using visualization/reporting tools **(Tableau).**
- ✓ Experienced in using **Pig scripts** to do transformations, event joins filters and pre-aggregations before storing the data into **HDFS.**
- ✓ Hands on experience working **Amazon Web Services (AWS)** using **Elastic Map Reduce (EMR), Redshift, and EC2** for **data processing.**
- ✓ Experience in Dimensional Modeling using **Star and Snowflake schema** methodologies of Data Warehouse and Integration projects.
- ✓ Experience with the **Apache Airflow** engine that can easily schedule and run my complex data pipelines which will make each task to get executed in a correct order.
- ✓ Experience in integration of various data sources with multiple Relational Databases like **SQL Server, Teradata, and Oracle.**
- ✓ Good communication skills, work ethics and the ability to work in a team efficiently with good leadership skills.

Technical Skills:

Databases	Snowflake, AWS RDS, Teradata, Oracle, MySQL, Microsoft SQL, Postgre SQL.
NoSQL Databases	MongoDB, Hadoop HBase and Apache Cassandra.
Programming Languages	Python, SQL, Scala, MATLAB.
Cloud Technologies	AWS, Docker
Data Formats	CSV, JSON
Querying Languages	SQL, NO SQL, PostgreSQL, MySQL, Microsoft SQL
Integration Tools	Jenkins
Scalable Data Tools	Hadoop, Hive, Apache Spark, Pig, Map Reduce, Sqoop.
Operating Systems	Red Hat Linux, Unix, Windows, macOS.
Reporting & Visualization	Tableau, Matplotlib.

Professional Experience:

Client: Comcast, Philadelphia, PA **Dec 2022 – Till Date**

Role: Data Engineer

Responsibilities:

- ✓ Worked with the business users to gather, define business requirements and analyze the possible technical solutions.
- ✓ Developed **Spark** applications using **Pyspark** and **Spark-SQL** for data extraction, transformation and aggregation from multiple file formats for analyzing and transforming.
- ✓ Developed **Spark code** using **Scala** and **Spark-SQL** for faster testing and data processing.
- ✓ Developed predictive analytics using **Apache Spark Scala APIs.**
- ✓ Developed of **Python APIs** to dump the array structures in the Processor at the failure point for debugging.
- ✓ Designed and Developed **Pig Latin** scripts and **Pig** command line transformations for data joins and custom processing of **Map reduce** outputs.
- ✓ Developed **ETL's** for **Data Extraction, Data Mapping** and data Conversion using **SQL, PL/SQL** and various **ETL scripts.**
- ✓ Involved in **ETL** processes, Data warehousing methodologies and concepts including **star schemas, snowflake schemas, dimensional modeling and reporting tools, Operations Data Store concepts, Data Mart** and **OLAP** technologies.
- ✓ Implemented Visualized BI Reports with **Tableau.**
- ✓ Administered user, user groups, and scheduled instances for reports in **Tableau.** Monitoring of **Tableau Servers** for its high availability to users.
- ✓ Deployed web embedded **power BI** dashboards refreshed using gateways by using workspace and data source.
- ✓ Involved in all phases of **data** mining, **data** collection, **data** cleaning, developing models, validation and visualization.
- ✓ Involved in Manipulating, cleansing & processing **data** using Excel, Access and SQL and responsible for loading, extracting and validation of client **data**.
- ✓ Involved in **Data Architecture, Data profiling, Data analysis, data mapping** and Data architecture artifacts design.
- ✓ Wrote **AWS Lambda** functions in **python for AWS's Lambda** which invokes python scripts to perform various transformations and analytics on large data sets in EMR clusters.
- ✓ Involved in **modeling (Star Schema methodologies)** in building and designing the logical data model into Dimensional Models.
- ✓ Install and Configure **Apache Airflow for S3 bucket** and snowflake data warehouse and created DAGs to run the **Airflow.**
- ✓ Instantiated, created, and maintained **CI/CD** (continuous integration & deployment) pipelines and apply automation to environments and applications
- ✓ Involved in developing **DAGS** using **Airflow** orchestration tool and monitored the weekly processes.
- ✓ Created **Stored Procedures** to transform the Data and worked extensively in **PL/SQL** for various needs of the transformations while loading the data.
- ✓ Worked in **Agile** methodology, Attended SCRUM meetings, and standup meetings.
- ✓ Involved in Daily and Weekly Status meetings.

Environment: Spark, Scala, Pyspark, Python, PIG, AWS, Docker, Restful, HDFS, Tableau, Snowflake, Apache Airflow, Power BI, ETL, Agile and SQL.

Client: Delta Airlines, Atlanta, GA **Dec 2021 – Nov 2022**

Role: Data Engineer

Responsibilities:

- ✓ Involved in Requirement gathering phase to gather the requirements from the business users to continuously accommodate changing user requirements.
- ✓ Developed **Spark** scripts by using **Python** in **PySpark** shell command in development.
- ✓ Worked on migrating **Map Reduce** programs into **Spark transformations** using **Spark and Scala.**
- ✓ Implemented **Spark** Scripts using **Scala, Spark SQL** to access hive tables into **spark** for faster processing of data.

- ✓ Responsible for Writing the Data Quality checks, based on the existing source code, using **Python and PySpark dataframe work** in **Databricks** platform (which Improved process time).
- ✓ Responsible for extracting the data and loading the data using the **Python.**
- ✓ Design and develop efficient **PySpark** programs using cloud-based data platforms (EMR) to extract/transform/load data in between various data warehouse applications.
- ✓ Developed **ETL procedures** to transform the data in the intermediate tables according to the business rules and functionality requirements.
- ✓ **ETL** Restarting capability for a date or date range or from point of failure or from beginning
- ✓ Created report schedules on **Tableau server.**
- ✓ Validated the **Tableau insight** center reports to make sure all the data is populated as per requirements.
- ✓ Created automated **python** scripts to convert the data from different sources and to generate the **ETL** pipelines.
- ✓ Worked on Azure Data factory to integrate data of both on-prem and cloud (Azure SQL DB) and applied transformations to load back to Azure synapse. Managed configured and scheduled resources across the cluster using Azure Kubernetes service

- ✓ Involved on creating multiple kind of Report in **Power BI** and present it using Story Points.
- ✓ Leveraged SQL scripting for **data** modeling, enabling streamlined **data** querying and reporting capabilities, which contributed to improved insights into customer **data**.
- ✓ Collaborated with end users to resolve **data** and performance-related issues during the on boarding of new users.
- ✓ Developed **Airflow pipelines** to efficiently load **data** from multiple sources into **Redshift** and monitored job schedules.
- ✓ Designed the data marts in dimensional data modeling using **star** and **snowflake schemas.**
- ✓ Performed analysis on the unused user navigation data by loading into **HDFS** and writing **MapReduce** jobs.
- ✓ Design and Develop ETL Processes in **AWS** Glue to migrate Campaign data from external sources like S3, ORC/Parquet/Text Files into **AWS Redshift.**
- ✓ Involved in developing and writing **Pig** scripts to store unstructured data into **HDFS**
- ✓ Created and modified several database objects such as Tables, Views, Indexes, Constraints, Stored procedures, Packages, Functions and Triggers using **SQL and PL/SQL.**
- ✓ Involved in **Agile** methodologies, daily scrum meetings, spring planning.
- ✓ Actively participated and provided feedback in a constructive and insightful manner during weekly Iterative review meetings to track the progress for each iterative cycle and figure out the issues.

Environment: Spark, Scala, Pyspark, Python, PIG, AWS, Docker, Restful, HDFS, Tableau, Snowflake, Apache Airflow, Power BI, ETL, Agile and SQL.

Client: Western Union, Milwaukee, WI **May 2020 – Jul 2021**
Role: Data Engineer
Responsibilities:

- ✓ Involved in Requirement gathering phase to gather the requirements from the business users to continuously accommodate changing user requirements.
- ✓ Developed **spark** applications for performing large scale transformations and denormalization of relational datasets.
- ✓ Involved in developing **Spark code** using **Scala and Spark-SQL** for faster testing and processing of data and exploring of optimizing it using **SparkContext, Spark-SQL, PairRDD's.**
- ✓ Used **Spark** for interactive queries, processing of streaming data and integration with popular NoSQL database for huge volume of data.
- ✓ Developed Spark code using **scala** and Spark-SQL/Streaming for faster testing and processing of data.
- ✓ Designed data and **ETL** pipeline using **Python and Scala** with **Spark.**
- ✓ Developed **Map Reduce** programs for applying business rules on the data.
- ✓ Built real time data pipelines by developing **Kafka** producers and **spark** streaming applications for consuming.
- ✓ Used **Apache Kafka** functionalities like distribution, partition, replicated commit log service for messaging.
- ✓ Migrated an existing on-premises application to **Amazon Web Services (AWS)** and used its services like **EC2 and S3** for small data sets processing and storage, experienced in maintaining the **Hadoop cluster** on **AWS EMR.**

- ✓ Utilized **Spark, Scala, Python** for querying, preparing from big data sources.
- ✓ Wrote pre-processing queries in **python** for internal **spark** jobs.
- ✓ Worked on **PySpark APIs** for data transformations.
- ✓ Developed solutions to pre-process large sets of structured, semi-structured data, with different file formats like **Text, Sequence, XML, and JSON.**
- ✓ Created **Hive** tables and involved in data loading and writing **Hive UDFs**. Developed Hive UDFs for rating aggregation
- ✓ Created **HBase** tables to load large sets of structured data.
- ✓ Designing and creating **SQL Server** tables, views, stored procedures, and functions.
- ✓ Involved in weekly walkthroughs and inspection meetings, to verify the status of the testing efforts and the project as a whole.

Environment: Spark, Scala, Hive, JSON, AWS, MapReduce, Hadoop, Python, XML, NoSQL, HBase, and Windows.

Company: Verisk, Nepal **May 2018 – Apr 2020**
Role: Data Engineer
Responsibilities:

- ✓ Performed Data analysis, Data Profiling and Requirement Analysis.
- ✓ Developed **Spark code** using **Scala and Spark-SQL/Streaming** for faster testing and processing of data.
- ✓ Performed **Spark jobs** with the **Spark core, SparkSQL** libraries for processing the data.
- ✓ Worked on migrating **MapReduce** programs into Spark transformations using **Scala.**
- ✓ Integrated data quality plans as a part of **ETL** processes.
- ✓ Building data pipelines and complex **ETL** to process external client data using **Python, Spark.**
- ✓ Performed Data cleaning and Preparation on **XML** files.
- ✓ Optimized **MapReduce** Jobs to use **HDFS** efficiently by using various compression mechanisms
- ✓ Developed various **Python** scripts to find vulnerabilities with **SQL Queries** by doing SQL injection, permission checks and analysis.
- ✓ Wrote reports using **Tableau Desktop** to extract data for analysis using filters based on the business use case.
- ✓ Performed **Tableau** type conversion functions when connected to relational data sources.
- ✓ Worked with building data warehouse structures, and creating facts, dimensions, aggregate tables, by dimensional modeling, **Star** and **Snowflake schemas**.
- ✓ Use **SQL** queries and other tools to perform data analysis and profiling.
- ✓ Analyzed the **SQL scripts** and designed the solution to implement using **PySpark.**
- ✓ Used **Agile** methodology named SCRUM for all the work performed.
- ✓ Involved in weekly walkthroughs and inspection meetings, to verify the status of the testing efforts and the project as a whole.

Environment: Spark, Scala, Hive, JSON, AWS, MapReduce, Hadoop, Python, XML, NoSQL, HBase, and Windows.
