Name: Prav. Dhakal
Email: dhaprav3@gmail.com
Ph#: 872-215-5473

Professional Summary:

- Over **6+ years** of experience in **Data Engineering, Data Pipeline Design**, Development, and Implementation as a **Data Engineer/Data Developer and Data Modeler.**
- Experience in all stages of **SDLC (Agile, Waterfall),** writing Technical Design documents, Development, Testing, and Implementation of **enterprise-level Data mart and Data warehouses.**
- Experience in developing **Spark** streaming jobs by developing **RDDs (Resilient Distributed Datasets)** using **Scala, PySpark, and Spark-Shell.**
- Hands-on use of **Spark and Scala APIs** to compare the performance of **Spark** with **Hive and SQL, and Spark SQL** to manipulate Data Frames in **Scala.**
- Strong experience in writing scripts using **Python API, PySpsark API, and Spark API** for analyzing the data.
- Expertise in **Python and Scala, user-defined functions (UDF)** for **Hive** and **Pig** using **Python.**
- Experience in developing **Map Reduce** Programs using **Apache Hadoop** for analyzing big data as per the requirement.
- Experience in using **Python and SQL** for **Data Engineering** and **Data Modeling.**
- Extensive experience with **Informatica (ETL Tool)** for Data Extraction, Transformation and Loading.
- Extensive experience in building Data Warehouses/Data Marts using **ETL** tools **Informatica Power Center (9.0/8.x/7.x).**
- Experience creating Visual reports, Graphical analysis, and Dashboard reports using **Tableau, Informatica** of historical data saved in **HDFS,** and data analysis using Splunk enterprise edition.
- Experience in writing **Map-Reduce** Jobs in **Python** for processing large sets of structured, semi-structured, and unstructured data sets and storing them in **HDFS.**
- Hands-on experience designing and building data models and data pipelines on **Data Warehouse** focus and **Data Lakes.**
- Hands-on experience in **Star Schema** Modeling, **Snow-Flake** Modeling, **FACT** and **Dimensions Tables**, **and Physical and Logical Data Modeling** using **Erwin**.
- Hands-on experience working with **Amazon Web Services (AWS)** using **Elastic Map Reduce (EMR), Redshift, and EC2** for data processing. Used **Amazon Web Services** Elastic Compute Cloud (AWS EC2) to launch cloud instances.
- Experience in Importing and exporting data into **HDFS and Hive** using **Sqoop.**
- Experienced with **Integration Services (SSIS), Reporting Services (SSRS),** and **Analysis Services (SSAS).**
- Good Hands-on Experience with **NoSQL** databases like **MongoDB, Cassandra, and HBase.**
- Experience in working with databases, such as **Oracle**, **SQL Server, and My SQL.**
- Strong skills in analytical, presentation, communication, and problem solving with the ability to work independently as well as in a team and the ability to follow the best practices and principles defined for the team.

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: Seven Eleven, Irving TX **Sep 2022 – Till Date**
Role: Data Engineer
Responsibilities:

- Involved in Analysis, Design, System architectural design, Process interface design, and design documentation.

- Developed **Spark code** using **Scala** and **Spark-SQL/Streaming** for faster testing and processing of data.
- Developed **Spark** jobs to clean data obtained from various feeds to make it suitable for ingestion into **Hive tables** for analysis.
- Developed the batch scripts to fetch the data from **AWS S3 storage** and do required transformations in **Scala** using **Spark framework.**
- Developed **Scala scripts, and UDFs** using both **Data frames/SQL and RDD/MapReduce** in **Spark** for Data Aggregation, queries, and writing data back into **RDBMS** through **Sqoop.**
- Responsible for designing and building new data models and schemas using **Python and SQL.**
- Built **Spark jobs** using **PySpark** to perform **ETL** for data in **S3 Data Lake.**
- Involved in developing data pipelines using **Kafka, Spark, and Hive** to ingest, transform, and analyze data.
- Developed **Pig Scripts, Pig UDFs, and Hive Scripts, Hive UDFs** to analyze **HDFS data.**
- Involved in **ETL** process consisting of data transformation, data sourcing, mapping, conversion, and loading.
- Performing **ETL** testing activities like running the Jobs, Extracting the data using necessary queries from database transform, and uploading into the **Data warehouse** servers.
- Developed connections for **Tableau Application** to core and peripheral data sources like **Flat files, Microsoft Excel, Tableau Server, Amazon Redshift Database, Microsoft SQL Server, etc.** to Analyze complicated data.
- Used **Apache Kafka** to aggregate web log data from multiple servers and make them available in downstream systems for analysis.
- Utilized **AWS** services with a focus on big data architect /analytics/enterprise Data warehouse and business intelligence solutions to ensure optimal architecture, scalability, flexibility, availability, and performance, and to provide meaningful and valuable information for better decision-making.
- Prepared scripts to automate the ingestion process using **Python** and **Scala** as needed through various sources such as **API, AWS S3, Teradata, and Snowflake.**
- Performed analysis on the unused user navigation data by loading it into **HDFS** and writing **MapReduce** jobs.
- Creating **Hive tables, loading** and analyzing data using Hive scripts. Implemented Partitioning, Dynamic Partitions, and Buckets in **Hive.**
- Implemented **Apache Airflow** for authoring, scheduling, and monitoring Data Pipelines Designed several DAGs (Directed Acyclic Graphs) for automating ETL pipelines.
- Created airflow DAGs to sync files from the box, analyze data quality, and alert for missing files.
- Worked on different file formats like **Text**, **Sequence files, Avro, Parquet, JSON, XML files and Flat files** using **Map Reduce Programs.**
- Worked on designing, building, deploying, and maintaining **Mongo DB.**
- Involved in creating, and modifying **SQL queries,** prepared statements, and stored procedures used by the application.
- Implemented the project under **Agile** Project Management Environment and followed SCRUM iterative incremental model & and configured various sprints to execute.
- Actively participated and provided feedback constructively and insightfully during weekly Iterative review meetings to track the progress for each iterative cycle and figure out the issues.

Environment: Spark, Scala, Python, PySpark, MapReduce, Apache Kafka, ETL, Tableau, Pig, Hive, HDFS, AWS, Sqoop, XML, JSON, MongoDB, SQL, Agile and Windows.

Client: Southwest Airlines, Dallas TX **May 2020 – Aug 2022**
Role: Data Engineer
Responsibilities:

- Gathered, analyzed, and translated business requirements to technical requirements, communicated with other departments to collect client business requirements and access available data.
- Developed various **spark applications** using **Scala** to perform various enrichments of user behavioral data (click stream data) merged with user profile data.
- Involved in developing production-ready **spark** application using **Spark RDD APIs, Data frames, Spark-SQL** and **Spark-Streaming API's.**
- Involved in implementing advanced procedures like text analytics and processing using **Apache Spark** written in **Scala.**
- Involved in converting **Hive/SQL queries** into **Spark transformations** using **Spark RDDs, and Spark SQL** using **Scala.**

- Using **Apache Kafka** for Streaming purposes.
- Design and implement secure data pipelines into a **Snowflake data warehouse** from **on-premises** and **cloud data sources.**
- Developed Simple to complex **MapReduce** Jobs using **Hive and Pig**. Developed **Shell and Python scripts** to automate and provide Control flow to **Pig scripts.**
- Involved in **Extraction, Transformation, and Loading (ETL)** of data from multiple sources like **Flat files, XML files, and Databases.**
- Developed **Tableau** data visualization using Cross tabs, Heat maps, Box and Whisker charts, Scatter Plots, Geographic maps, Pie Charts Bar Charts, and Density charts.
- Built models using **Python** and **Pyspark** to predict the probability of attendance for various campaigns and events.
- Built an ETL framework for Data Migration from on-premises data sources such as Hadoop to AWS using Apache Airflow and Apache Spark (PySpark).
- Created Airflow Scheduling scripts in Python.
- Working with **AWS stack S3, EC2, Snowball, EMR, Athena, Glue, Redshift, DynamoDB, RDS, Aurora, IAM, Firehose, and Lambda.**
- Worked on **Kafka** messaging platform for real-time transactions and streaming of data from APIs and databases to Reporting tools for analysis.
- Involved in creating **Data Lake** by extracting customer's data from various data sources to **HDFS** which includes data from **CSV, databases,** and **log data** from servers.
- Worked on custom **Pig Loaders and Storage classes** to work with a variety of data formats such as **JSON, Compressed CSV, etc.**
- Developed **NoSQL** database by using CRUD, Indexing, Replication, and Sharing in **MongoDB.**
- Designing and creating **SQL Server tables, views, stored procedures, and functions**.
- Used **Agile (SCRUM)** methodologies for Software Development.
- Actively participating in the code reviews, and meetings and solving any technical issues.

Environment: Spark**,** Scala, Python, PySpark, ETL, Tableau, Pig, Map Reduce, AWS, Kafka, Hive, Apache Kafka, HDFS, Pig, JSON, Sqoop, NoSQL, MongoDB, SQL, Agile and Windows.

Client: Merck Group, West Point, PA **Nov 2019 – Apr 2020**
Role: Data Engineer
Responsibilities:

- Worked with the business users to gather, define business requirements, and analyze the possible technical solutions.
- Developed **Spark** applications by using **Scala and Python** and implemented **Apache Spark** for data processing from various streaming sources.
- Developed **Spark** scripts by using **Scala** and **Python** shell commands as per the requirement.
- Developed **Spark**-Streaming applications to consume the data from **Kafka topics** and to insert the processed streams into **HBase.**
- Involved in writing **Spark** applications using **Scala** to perform various data cleansing, validation, transformation, and summarization activities according to the requirement.
- Built reusable **Hive UDF libraries** for business requirements which enabled users to use these **UDFs** in **Hive Querying.**
- Developed **Map Reduce** programs for applying business rules to the data.
- Design and develop **Tableau visualizations** which include preparing Dashboards using calculations, parameters, calculated fields, groups, sets, and hierarchies.
- Designed and developed end-to-end **ETL process** from various source systems to Staging area, from staging to **Data Marts and data load.**
- Involved in writing **Pyspark** User Defined Functions (UDF's) for various use cases and applied business logic wherever necessary in the **ETL** process.
- **Data gathering, data cleaning, and data wrangling** performed using **Python.**
- Using **Amazon Web Services (AWS)** for storage and processing of data in the cloud.
- Performed incremental loads as well as full loads to transfer data from **OLTP** to the **Data Warehouse** of **snowflake** schema using different data flow and control flow tasks and provided maintenance for existing jobs.

- Performing **Sqoop** jobs to land data on **HDFS** and running validations.
- Creating **Hive** tables, loading and analyzing data using Hive scripts. Implemented Partitioning, Dynamic Partitions, and Buckets in **HIVE.**
- Wrote complex **SQL scripts and PL/SQL** packages, to extract data from various source tables of data warehouse.
- Involved in **Agile** methodologies, daily scrum meetings, and spring planning.
- Actively participating in the code reviews, and meetings and solving any technical issues.

Environment: Spark, Scala, Python, PySpark, ETL, Tableau, Map Reduce, Hive, AWS, Snowflake, Datawarehouse, Sqoop, HDFS, SQL, Agile and Windows.

Client: Western Union, Milwaukee, WI **Feb 2018 - Oct 2019**
Role: Data Engineer
Responsibilities:

- Worked with the business users to gather, define business requirements, and analyze the possible technical solutions.
- Developed **Spark** scripts by using **Python** shell commands as per the requirement.
- Developed **spark code** and **spark-SQL/streaming** for faster testing and processing of data.
- Developed **Spark/Scala, Python** for regular expression (RegEx) project in **Hadoop/Hive** environment for big data resources.
- Developed **PIG** scripts for the analysis of semi-structured data.
- Used **Pig** as **an ETL** tool to do transformations, event joins, filters, and some pre-aggregations before storing the data onto **HDFS.**
- Built key business metrics, Visualizations, dashboards, and reports with **Tableau.**
- Involved in building the **ETL** architecture and Source to Target mapping to load data into the Data warehouse.
- Developed **Map Reduce** jobs for **data cleaning and manipulation.**
- Involved in writing **Pyspark** User Defined Functions (UDF's) for various use cases and applied business logic wherever necessary in the **ETL** process.
- **Data gathering, data cleaning, and data wrangling** performed using **Python.**
- Written Programs in **Spark** using **Scala** for Data quality check.
- Created **Hive** tables as per requirement as internal or external tables, intended for efficiency.
- Worked on **Snowflake** environment to remove redundancy and load real-time data from various data sources into **HDFS using Kafka.**
- Used **AWS S3** to store large amounts of data in identical/similar repositories.
- Wrote complex **SQL scripts and PL/SQL** packages, to extract data from various source tables of data warehouse.
- Actively participating in the code reviews, and meetings and solving any technical issues.

Environment: Spark, Scala, Hadoop, Python, Pyspark, AWS, MapReduce, Pig, ETL, HDFS, Hive, HBase, SQL, Agile and Windows.

References: These will be provided upon request.