Srijana Raut

+1469-848-7938 || LinkedIn || srijanaraut567@gmail.com

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 all stages of **SDLC (Agile, Waterfall),** writing Technical Design document, Development, Testing and Implementation of **Enterprise level Data mart and Data warehouses.**
- Utilized **SSIS tool** to extract, transform and load data from various source like data warehouse and data lakes. Also, integrated with various source like **SQL Server, Oracle, Excel, third Party APIs.**
- Experience in developing **Spark** streaming jobs by developing **RDD's (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 the **big data** as per the requirement.
- Experience in using **Python and SQL** for **Data Engineering** and **Data Modeling.**
- Experience in **Extraction, Transformation and Loading (ETL)** data from various sources into Data Warehouses, as well as data processing like collecting, aggregating, and moving data from various sources.
- Experience creating Visual report, 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 stores them in **HDFS.**
- Hands on experience on **Star Schema** Modeling, **Snow-Flake** Modeling, **FACT** and **Dimensions Tables**, **Physical and Logical Data Modeling** using **Erwin**.
- Experience in Importing and exporting data into **HDFS and Hive** using **Sqoop.**
- Hands on experience working **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 instance.
- Experienced with **Integration Services (SSIS), Reporting Service (SSRS)** and **Analysis Services (SSAS).**
- Strong skills in analytical, presentation, communication, problem solving with the ability to work independently as well as in a team and had the ability to follow the best practices and principles defined for the team.

EDUCATION

University of South Dakota, Vermillion, SD, USA **Aug 2022 – Dec 2023**

Master of Computer Science

GPA – 4.0/4.0

Course Work – Machine Learning, Artificial Intelligence, Computer Vision, Quantum Computing, Pattern Recognition, Database, Distributed System etc.

TECHNICAL SKILLS

Programming Language	**Python, Scala, SQL, Java (Basics)**
SQL Databases	Oracle, MySQL, PostgreSQL, SQL Lite, PL/SQL,
NoSQL Databases	MongoDB, Hadoop, HBase, Redis, Big Table, Apache Cassandra, Dynamo DB(AWS)
Modern Database	Teradata, Snowflake, Redshift, Databricks
Cloud Technologies	GCP, AWS, Azure, DBT, SSIS
Scalable Data Tools	Hadoop, Hive, Apache Spark, Apache Flink, Apache Storm, Apache Kinesis, Pig, Map Reduce, Sqoop
Hadoop Core Services	HDFS, Map Reduce, Spark, YARN, Hive, Pig, Scala, Kafka, Sqoop, Flume, Impala, Oozie, Zookeeper
Hadoop Distribution	Horton Works, Cloudera
Containerization, Orchestration and Version Control Tool	Docker, Kubernetes, Temporal, Luigi GitHub
ML, DL, NLP Algorithms	Linear Regression, Logistic Regression, SVM, Decision Trees, Random Forest, CNN, RNN, Yolo, LLM (Llama2, Mistral)

ML/DL Library	TensorFlow, PyTorch, Scikit learn, Lang Chain
Visualization& Reporting Tools	Tableau, Looker, Power BI, Matplotlib, Seaborn, Pandas, NumPy

EXPERIENCE

American Airlines, Dallas TX **Jan 2023 – Present**
Data Engineer

- Involved in Analysis, Design, System architectural design, Process interfaces 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 S3storage** and do required transformations in **Scala** using **Spark framework.**
- Developed **Scala scripts, UDF's** 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 pipeline using **Kafka, Spark and Hive** to ingest, transform and analyzing data.
- Developed **Pig Scripts, Pig UDFs and Hive Scripts, Hive UDFs** to analyses **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 upload 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.
- Developed **ETL** framework using Spark and Hive (including daily runs, error handling, and logging) to useful data.
- Involved in creating technical design documents, source to target mapping documents and test case documents to reflect **ETL** process.
- Implemented **Apache Airflow** for authoring, scheduling, and monitoring Data Pipelines Designed several DAGs (Directed Acyclic Graph) for automating ETL pipelines.
- Created **airflow DAGs** to sync files from box, analyze data quality, and alert for missing files.
- Used **Apache Kafka** to aggregate web log data from multiple servers and make them available in downstream systems for analysis.
- Utilized **AWS** services with focus on **big data** architect /analytics / enterprise Data warehouse and business intelligence solutions to ensure optimal architecture, scalability, flexibility, availability, 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 into **HDFS** and writing **MapReduce** jobs.
- Creating **Hive tables, loading** and analyzing data using hive scripts. Implemented Partitioning, Dynamic Partitions, Buckets in **Hive.**
- Extracted the data from **Teradata** into **HDFS** using the **Sqoop.**
- Worked on different file formats like **Text**, **Sequence files, Avro, Parquet, JSON, XML files and Flat files** using **Map Reduce Programs.**
- Involved in creating, 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 & configured various sprints to execute.
- 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, Python, PySpark, MapReduce, Apache Kafka, ETL, Tableau, Airflow, Pig, Hive, HDFS, AWS, Sqoop, XML, JSON, MongoDB, SQL, Agile and Windows.

UFG Insurance, Cedar Rapids, Iowa **Nov 2020 – July 2022**
Data Engineer

- 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, Spark SQL** using **Scala.**
- Using **Apache Kafka** for Streaming purpose.
- 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 Map, Pie Charts and Bar Charts and Density Chart.
- Involved in building the **ETL architecture** and Source to Target mapping to load data into Data warehouse.
- **Extract, transform, and load (ETL)** data from multiple federated data sources (**JSON**, **relational database**, etc.) with Data Frames in **Spark**.
- Built models using **Python** and **PySpark** to predict probability of attendance for various campaigns and events.
- Designed and implemented end-to-end cloud solutions on **Microsoft Azure**, leveraging services such as **Azure Data Lake Storage**, **Azure SQL Database**, **and Azure Databricks**.
- Specialized in performance tuning and optimizing **Azure** and **Databricks** solutions, conducting in-depth analyses and implementing enhancements to improve query performance, minimize data processing latencies, and enhance overall system efficiency.
- Engineered and executed ETL pipelines, data transformations, and analytics workflows within Databricks notebooks.
- 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 include data from **csv, databases,** and **log data** from servers.
- Involved in loading and transforming large Datasets from relational databases into **HDFS** and vice-versa using **Sqoop** imports and export.
- Used **SSIS tools** for data transformation like data cleansing, merging, handling robust error to capture log and processing errors. Managed and monitored scheduled jobs, resolving issues and maintaining continuous data flow using SSIS.
- 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, meetings and solving any technical issues.

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

Verisk Pvt Ltd, Lalitpur Nepal **Aug 2018 – Oct 2020**
Data Engineer

- Orchestrated a data pipeline integrating a data warehouse with third-party applications, utilizing Python, REST APIs, Docker, AWS ECS, and Airflow, enhancing operational efficiency.
- Engineered an ETL Pipeline for efficient extraction of emails from a corporate domain, followed by transformation and ingestion into PostgreSQL database in JSON format, optimizing data management processes.
- Trained and documented initial deployment and Supported product stabilization/debugging at the deployment stage. Worked on SQL for backend data transactions and validations.
- Used Python to write Data into JSON files for testing Django Websites, Created scripts for data modelling and data import and export.
- Created, and maintained CI/CD continuous integration & deployment pipelines and apply automation to environments and applications.
- Created views for reporting purpose which involves complex SQL queries with sub-queries, inline views, multi table joins, with clause and outer joins as per the functional needs in the Business Requirements Document (BRD).

Environment: Python, SQL, PostgreSQL, CI/CD Pipelines, Data Modelling, Data Pipeline, ETL Process, NoSQL Database.