**RAHUL RAJ**

**Email:** [**rahulsamba7@gmail.com**](mailto:shalinikataram01@gmail.com) **PH: 9133461701**

**Sr. Data Engineer**

**PROFESSIONAL SUMMARY**

* Over 6 years of IT experience in Design, Development, Maintenance and Support of **Big Data Applications**
* Over 3+ yearsof experience in Data Engineering, Data Pipeline Design, Development and Implementation as a Sr. Data Engineer/Data Developer and Data Modeler.
* Strong experience in Software Development Life Cycle (SDLC) including Requirements Analysis, Design Specification and Testing as per Cycle in both Waterfall and Agile methodologies.
* Strong experience in writing scripts using Python API, PySpark API and Spark API for analyzing the data.
* Python Libraries PySpark, Pytest, Pymongo, PyExcel, Psycopg, NumPy and Pandas.
* Hands On experience on Spark Core, Spark SQL, Spark Streaming and creating the Data Frames handle in SPARK with Scala.
* Experience in NoSQL databases and worked on table row key design and to load and retrieve data for real time data processing and performance improvements based on data access patterns.
* Extensive experience in Hadoop Architecture and various components such as HDFS, Job Tracker, Task Tracker, Name Node, Data Node, and Map Reduce concepts.
* Experience in building large scale highly available Web Applications. Working knowledge of web services and other integration patterns.
* Developed Simple to complex Mapreduce and Streaming jobs using Java and Scala language.
* Developed Hive scripts for end user / analyst requirements to perform ad hoc analysis with Hive to handle less important bulk ETL jobs.
* Hands-on use of Spark and Scala API to compare the performance of Spark with Hive and SQL, and Spark SQL to manipulate Data Frames in Scala.
* 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.
* Hands on Spark MLlib utilities such as classification, regression, clustering, collaborative filtering, dimensionality reduction.
* Hands-on experience in developing and deploying enterprise-based applications using major Hadoop ecosystem components like MapReduce, YARN, Hive, HBase, Flume, Sqoop, Spark MLlib, Spark GraphX, Spark SQL, Kafka.
* Experience in application of various data sources like Oracle SE2, SQL Server, Flat Files and Unstructured files into a data warehouse.
* Able to use Sqoop to migrate data between RDBMS, NoSQL databases and HDFS.
* 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 using Apache Flume, Kafka, PowerBI and Microsoft SSIS.
* Skilled in System Analysis, E-R/Dimensional Data Modeling, Database Design and implementing RDBMS specific features.
* Knowledge of working with Proof of Concepts (PoC's) and gap analysis and gathered necessary data for analysis from different sources, prepared data for data exploration using data munging and Teradata.
* Well experience in Normalization and Denormalization techniques for optimum performance in relational and dimensional database environments.
* Experience in developing customized UDF's in Python to extend Hive and Pig Latin functionality.
* Experience on Migrating SQL database to Azure data Lake, Azure data lake Analytics, Azure SQL Database, Data Bricks and Azure SQL Data warehouse and controlling and granting database access and Migrating On premise databases to Azure Data lake store using Azure Data factory.
* Experience in Developing Spark applications using Spark - SQL in Databricks for data extraction, transformation and aggregation from multiple file formats for analyzing & transforming the data to uncover insights into the customer usage pattern.
* Experience in Data Analysis, Data Profiling, Data Integration, Migration, Data governance and Metadata Management, Master Data Management and Configuration Management.
* Expertise in designing complex Mappings and have expertise in performance tuning and slowly changing Dimension Tables and Fact tables
* Extensively worked with Teradata utilities Fast export, and Multi Load to export and load data to/from different source systems including flat files.
* Experienced in building Automation Regressing Scripts for validation of ETL process between multiple databases like Oracle, SQL Server, Hive, and MongoDB using Python.
* Proficiency in SQL across several dialects (we commonly write MySQL, PostgreSQL, Redshift, SQL Server, and Oracle)
* Exposure with Snowflake Data Cloud in building efficient data pipelines using SnowPipe and extensively used snowSQL for performing queries and creating database objects.
* Experience in designing star schema, Snowflake schema for Data Warehouse, ODS architecture.
* Skilled in System Analysis, E-R/Dimensional Data Modeling, Database Design and implementing RDBMS specific features.
* Ability to work effectively in cross-functional team environments, excellent communication, and interpersonal skills.

**PROFESSIONAL EXPERIENCE**

**Sr. Data Engineer**

**Salesforce, Bellevue, Washington February 2020 to Present**

**Responsibilities:**

* Performed data analysis and developed analytic solutions. Data investigation to discover correlations / trends and the ability to explain them.
* Worked with Data Engineers, Data Architects, to define back-end requirements for data products (aggregations, materialized views, tables - visualization)
* Developed frameworks and processes to analyze unstructured information. Assisted in Azure Power BI architecture design
* Experience in Data Analysis using Python - Data Cleaning, Data Transformation, Data Visualization by using the following libraries - Pandas, NumPy, Matplotlib and Seaborn.
* Designing and Developing Oracle PL/SQL and Shell Scripts, Data Import/Export, Data Conversions and Data Cleansing.
* Responsible for importing data from PostgreSQL to HDFS, HIVE using SQOOP tool.
* Experienced in migrating HiveQL into Impala to minimize query response time.
* Implemented Avro and Parquet data formats for Apache Hive computations to handle custom business requirements.
* Designed and implemented Sqoop for the incremental job to read data from DB2 and load to Hive tables and connected to Tableau for generating interactive reports using Hive server2.
* Used Sqoop to channel data from different sources of HDFS and RDBMS
* Developing Spark applications using Scala and Spark-SQL for data extraction, transformation, and aggregation from multiple file formats. Using Kafka and integrating with Spark Streaming. Developed data analysis tools using SQL and Python code.
* Developed various Mappings with the collection of all Sources, Targets, and Transformations using Informatica Designer
* Developed Mappings using Transformations like Expression, Filter, Joiner and Lookups for better data messaging and to migrate clean and consistent data
* Designed and implemented Sqoop for the incremental job to read data from DB2 and load to Hive tables and connected to Tableau for generating interactive reports using Hive server2.
* Used Sqoop to channel data from different sources of HDFS and RDBMS.
* Developed Spark applications using Pyspark and Spark-SQL for data extraction, transformation and aggregation from multiple file formats.
* Used Spark Streaming to receive real time data from the Kafka and store the stream data to HDFS using Python and NoSQL databases such as HBase and Cassandra
* Developed Spark Applications by using Scala and Implemented Apache Spark data processing project to handle data from various RDBMS and Streaming sources.
* Data ingestion to one or more Azure services (Azure Data Lake, Azure Storage, Azure SQL, Azure DW) and processing the data in Azure Databricks
* Worked with data migration from Data Lake to Hive, Python, Pyspark, Azure Data Factory (ADF) since it's a great SaaS solution to compose and orchestrate Azure.
* Used Snowflake functions to perform semi structured data parsing entirely with SQL statements.
* Migrated on premise database structure to Confidential Redshift data warehouse.
* Experience in Developing Spark applications using Spark - SQL in Databricks for data extraction, transformation and aggregation from multiple file formats for analyzing & transforming the data to uncover insights into the customer usage patterns.
* Worked on Dimensional and Relational Data Modeling using Star and Snowflake Schemas, OLTP/OLAP system, Conceptual, Logical and Physical data modeling using Erwin.
* Experience in developing Spark applications using Spark-SQL in Databricks for data extraction, transformation, and aggregation from multiple file formats for Analyzing& transforming the data to uncover insights into the customer usage patterns.
* Extract Transform and Load data from source Systems to Azure Data Storage services using a combination of Azure Data factory, T-SQL, Spark SQL, and U-SQL Azure Data Lake Analytics.
* Worked on Docker containers snapshots, attaching to a running container, removing images, managing Directory structures and managing containers.
* Experienced in day - to-day DBA activities including schema management, user management (creating users, synonyms, privileges, roles, quotas, tables, indexes, sequence), space management (table space, rollback segment), monitoring (alert log, memory, disk I/O, CPU, database connectivity), scheduling jobs, UNIX Shell Scripting.
* Developed complex Talend ETL jobs to migrate the data from flat files to database. Developed Talend ESB services and deployed them on ESB servers on different instances.
* Developed stored procedures/views in Snowflake for loading Dimensions and Facts.
* Developed merge scripts to UPSERT data into Snowflake from an ETL source.
* Extensively used SnowPipe for design data pipelines to migrate data from the source systems to the target systems ensuring a proper data flow within the system.
* Used SnowSQL for writing queries and interacting with the data and creating database objects and admin related tasks in Snowflake Data Cloud.

**Environment**: Hadoop, HDFS, Hive, Sqoop, Data Lake, Sqoop, Oozie, Kafka, Spark, Scala, Microsoft Azure, Azure Data factory, U-SQL Azure Data Lake Analytics, Azure SQL, Azure DW, Databricks, GitHub, Docker, Talend Big Data Integration, Snowflake, SnowPipe, SnowSQL, Oracle, Sql Server, MySQL, No SQL, MongoDB, Hbase, Cassandra, Unix, Shell Scripting, Python - Libraries like Pandas, Numpy, Matplotlib,MLib, SciPy, Seaborn.

**Data Engineer**

**Advantasure, Southfield, MI January 2017 to November 2019**

**Responsibilities:**

* Performed data extraction, transformation, loading, and integration in data warehouse, operational data stores and master data management
* Developed Spark Applications by using Scala and Implemented Apache Spark data processing project to handle data from various RDBMS and Streaming sources.
* Data ingestion to one or more Azure services (Azure Data Lake, Azure Storage, Azure SQL, Azure DW) and processing the data in Azure Databricks
* Implemented Apache Airflow for authoring, scheduling and monitoring Data Pipelines
* Developed Mappings using Transformations like Expression, Filter, Joiner and Lookups for better data messaging and to migrate clean and consistent data
* Worked on setting up high availability for major production clusters and designed automatic failover control using zookeeper and quorum journal nodes.
* Designed several DAGs (Directed Acyclic Graph) for automating ETL pipelines
* Provide troubleshooting and best practices methodology for development teams.
* This includes process automation and new application onboarding.
* Produce unit tests for Spark transformations and helper methods. Design data processing pipelines.
* Multiple batch jobs were written for processing hourly and daily data received through multiple sources like Adobe, No-SQL databases.
* Leveraged cloud and Azure computing technologies for automated machine learning and analytics pipelines.
* Designed and implemented configurable data delivery pipeline for scheduled updates to customer facing data stores built with Python
* Involved in building the ETL source to Target mapping to load data into Data warehouse.
* Developed a batch data ingestion pipeline using Sqoop and Hive to ingest, transform and analyze Supply Chain data.
* Use Object-Oriented Programming concepts to build Hive UDFs in Python that could be reused across the pipeline.
* Extensively working on Hive queries to load data from various sources like Teradata, DB2, Oracle, Mainframe etc.
* Extensively working on developing YAML files which is passed to the internal framework scripts having all the required JARs to move data.
* Migrating the transformed data to Azure data lake to be consumed by consumers depending on the business need.
* Creating tables in Hive in ORC format using correct compression ensuring no data loss.
* Worked on performance tuning of the queries, cost-based optimizations to improve the performance and latency of the queries.
* Writing shell scripts to automate the migration process and validate the data after loading.
* Created data validation application in Scala for validating the data of source and target.
* Working on performance tuning of the code with help of Hive dynamic partitioning, set parameters, mapper, and reducer tuning.
* Providing on call support to all the production jobs every alternate week.
* Coordinate with Quality Assurance teams to ensure the developed code is accurately transformed into working product.
* Participate in daily meetings and stand ups to provide updates and project progress to the team.
* Analyzed the system for new enhancements/functionalities and perform Impact analysis of the application for implementing ETL changes
* Implemented a Continuous Delivery pipeline with Docker, and GitHub, Azure
* Built performant, scalable ETL processes to load, cleanse and validate data
* Collaborate with team members and stakeholders in design and development of data environment
* Preparing associated documentation for specifications, requirements, and testing

**Environment**: Azure Data factory, U-SQL Azure Data Lake Analytics, Azure SQL, Azure DW, Databricks, GitHub, Docker, Talend Big Data Integration, Snowflake, Oracle, Sql Server, MySQL, No SQL, MongoDB, Hbase, Cassandra, Python- PySpark, Pytest, Pymongo, PyExcel, Psycopg, Matplotlib, NumPy and Pandas.

**Data Analyst /Data Engineer**

**Highmark Health Care, Pittsburg, PA January 2016 to December 2017**

**Responsibilities:**

* Gather business requirements, definition and design of the data sourcing, and worked with the data warehouse architect on the development of logical data models.
* Created sophisticated visualizations, calculated columns and custom expressions and developed Map Chart, Cross table, Bar chart, Tree map and complex reports which involves Property Controls, Custom Expressions.
* Investigated market sizing, competitive analysis and positioning for product feasibility. Worked on Business forecasting, segmentation analysis and Data mining.
* Extensively used Agile methodology as the Organization Standard to implement the data Models. Used Micro service architecture-based services interacting through a combination of REST and Apache Kafka message brokers.
* Created several types of data visualizations using Python - Matplotlib and Tableau. Extracted data using SQL Queries to create reports.
* Performed reverse engineering using Erwin to redefine entities, attributes, and relationships existing database.
* Analyze functional and non-functional business requirements and translate into technical data requirements and create or update existing logical and physical data models. Developed a data pipeline using Kafka to store data into HDFS.
* Performed Regression testing for Golden Test Cases from State (end to end test cases) and automated the process using python scripts.
* Developed Spark jobs using Scala for faster real-time analytics and used Spark SQL for querying
* Generated graphs and reports using the ggplot package in RStudio for analytical models. Developed and implemented R and Shiny application which showcases machine learning for business forecasting.
* Developed predictive models using Decision Tree, Random Forest, and Naïve Bayes.
* Used pandas, NumPy, seaborne, SciPy, Matplotlib, Scikit-learn, NLTK in Python for developing various machine learning algorithms. Expertise in R, Matlab, python and respective libraries.
* Research on Reinforcement Learning and control (TensorFlow, Torch), and machine learning model (Scikit-learn).
* Performed K-means clustering, Regression and Decision Trees in R. Worked on data cleaning and reshaping, generated segmented subsets using NumPy and Pandas in Python.
* Implemented various statistical techniques to manipulate the data like missing data imputation, principal component analysis and sampling.
* Responsible for design and development of Python programs/scripts to prepare, transform and harmonize data sets in preparation for modeling.
* Worked with Market Mix Modeling to strategize the advertisement investments to better balance the ROI on advertisements.
* Implemented clustering techniques like DBSCAN, K-means, K-means++ and Hierarchical clustering for customer profiling to design insurance plans according to their behavior pattern.
* Used Grid Search to evaluate the best hyper-parameters for my model and K-fold cross validation technique to train my model for best results.
* Used Python 3.X (NumPy, SciPy, pandas, scikit-learn, seaborn) and Spark 2.0 (PySpark, MLlib) to develop a variety of models and algorithms for analytic purposes.
* Performed Data Cleaning, features scaling, features engineering using pandas and NumPy packages in python and build models using deep learning frameworks
* Implemented Univariate, Bivariate, and Multivariate Analysis on the cleaned data for getting actionable insights on the 500-product sales data by using visualization techniques in Matplotlib, Seaborn, Bokeh, and created reports in Power BI.
* Processed the image data through the Hadoop distributed system by using Map and Reduce then stored into HDFS.
* Created Session Beans and controller Servlets for handling HTTP requests from Talend
* Performed Data Visualization and Designed Dashboards with Tableau and generated complex reports including chars, summaries, and graphs to interpret the findings to the team and stakeholders.
* Wrote documentation for each report including purpose, data source, column mapping, transformation, and user group.
* Utilized Waterfall methodology for team and project management.
* Used Git for version control with the Data Engineer team and Data Scientists colleagues.

**Environment**: Spark, YARN, HIVE, Pig, Scala, Mahout, NiFi, TDD, Python, Hadoop, Azure, DynamoDB, Kibana, NOSQL, Sqoop, MYSQL

**PROFESSIONAL SKILLS**

Data Modeling Tools: Erwin Data Modeler, ER Studio v17

Programming Languages: SQL, PL/SQL, UNIX

Methodologies: RAD, JAD, System Development Life Cycle (SDLC), Agile

Cloud Platform: AWS, Azure, Google Cloud.

Databases: Oracle 12c/11g, Teradata R15/R14.

OLAP Tools: Tableau, SSAS, Business Objects, and Crystal Reports 9

ETL/Data warehouse Tools: Informatica 9.6/9.1, and Tableau.

Operating System: Windows, Unix, Sun Solaris, Mac OS

Big Data Tools: Hadoop Ecosystem: Map Reduce

**EDUCATION DETAILS**

Jawaharlal Nehru Technological University

Bachelor of Technology - Computer Science