A black background with a black square

Description automatically generated with medium confidence**Sandeep R**

[**Sandeeprdy1291@gmail.com**](Sandeeprdy1291@gmail.com)

[**+1**](mailto:shravani.reddydondla@gmail.com|+1) **(475)-266-6606**

**LinkedIn:** [**http://www.linkedin.com/in/rdy1291**](http://www.linkedin.com/in/rdy1291)

**Senior Data Engineering**

**Professional Summary:**

* Over **9+** years of professional IT experience in Software Development, specializing as a **Big Data Engineer**, **Cloud Engineer (AWS/Azure), and Hadoop Developer**, with a strong understanding of the **Hadoop framework**.
* Excellent working experience in **Scrum / Agile framework, Iterative and Waterfall project execution methodologies.**
* Experience in architecting, designing, installation, configuration, and management of **Apache Hadoop** Clusters, MapR, Hortonworks & Cloudera Hadoop Distribution.
* I have expertise in **Hadoop Architecture** and various Components Such **as HDFS, YARN**, High availability, job Tracker, Task Tracker, **Name Node, Data Node** and **MapReduce Programming Paradigm.**
* Good experience in **Python Object-Oriented Programming**
* Proficient in **Agile** methodologies, including extreme programming, **SCRUM**, and **Test-Driven Development (TDD)**.
* Experience in developing Map Reduce Programs using **Apache Hadoop** for analyzing the **big data** as per the requirement.
* Experience working with various **Hadoop** distributions (**Cloudera, Hortonworks, Map R, and Amazon EMR)** to fully implement and leverage new **Hadoop** features.
* Skilled in developing custom UDFs for Pig and **Hive**, incorporating **Python/Java** methods and functionality into Pig Latin and **HQL** (HiveQL), and utilizing **UDFs** from Piggybank UDF Repository.
* Expertise in migrating **SQL** databases to **Azure Data Lake, Azure Data Lake Analytics, Azure SQL Database, Data Bricks, and Azure SQL Data Warehouse**. Proficient in controlling and granting database access and migrating on-premises databases to **Azure Data Lake Store using Azure Data Factory**.
* Deep knowledge of troubleshooting and tuning Spark applications and **Hive scripts** to achieve optimal performance.
* Strong experience with ETL and/or orchestration tools such as **Talend, Oozie, and Airflow**.
* Proficient in setting up **AWS Data Platform components**, including **AWS CloudFormation**, **Development End Points, AWS Glue, EMR, Jupiter/Sagemaker Notebooks, Redshift, Dynamo DB, S3, and EC2 instances**.
* Experience in data migration from **RDBMS** to **Snowflake cloud data warehouse**.
* Skilled in developing **Spark** applications using **Spark RDD**, **Spark SQL, and Dataframe APIs**.
* Worked with real-time data processing and streaming techniques **using Spark Streaming and Kafka**.
* Proficient in designing, creating, revising, and managing reports generated from operational and analytical systems **using SSRS, Tableau, Power BI, and Crystal Reports**.
* Hands-on experience in setting up workflows using **Apache Airflow and Oozie workflow engine for managing and scheduling Hadoop jobs.**
* Good understanding of partitions and bucketing concepts in Hive, designing managed and external tables to optimize performance.
* Strong understanding of **Spark** Architecture with **Databricks**, Structured Streaming, setting up **AWS and Microsoft Azure with Databricks, Databricks Workspace for Business Analytics, and managing clusters in Databricks.**
* Expertise in working with **HIVE data warehouse infrastructure**, including **creating tables, implementing partitioning, and bucketing for data distribution,** and **developing and tuning HQL queries**.
* Replaced existing MR jobs and Hive scripts with **Spark SQL** and Spark data transformations for efficient data processing.
* Proficient in **database design, modeling, migration, and development**, using **stored procedures**, **triggers**, **cursors**, **constraints**, and **functions**. Experienced with **MySQL, MS SQL Server, DB2, and Oracle**.
* Designed and developed spark pipelines to ingest real time event-based data from **Kafka** and other message queue systems and processed huge data with **spark batch processing** into **data warehouse hive.**
* Experience working with **NoSQL database technologies**, including **MongoDB**, **Cassandra**, and **HBase**.
* Proficient in software development tools such as **JIRA, Play, and Git.**
* Experienced in developing **Kafka** producers and consumers for streaming millions of events per second on streaming data.
* Solid understanding of **data modeling concepts**, including **dimensional and relational models such as star-schema modelling, a schema modelling, fact, and dimension tables.**
* Skilled in writing complex **SQL queries**, **creating reports**, and **developing dashboards**.
* Capable of **organizing, coordinating,** and managing multiple tasks simultaneously.

**Technical Skills:**

|  |  |
| --- | --- |
| **Big Data Ecosystem** | HDFS, Map Reduce, Azure Data Lake, Data Factory, Azure Databricks, Sqoop, Flume, Pig, Hive, Oozie, Impala, Zookeeper, Spark, snowflake, Spark Streaming and Kafka |
| **Programming Languages** | Python, R, Scala, SAS, SQL, HiveQL, PL/SQL, UNIX shell Scripting, Pig Latin |
| **Cloud Technologies** | AWS EC2, VPC, EBS, SNS, RDS, EBS, S3, Autoscaling, Lambda, Redshift, Cloud Watch, Azure Cloud, Azure Data Factory (ADF v2), Azure functions Apps, Azure Data Lake, BLOB Storage, Azure Cosmos DB, Data bricks. |
| **Scripting Languages** | Shell, Python, Perl, Unix, JavaScript. |
| **Operating Systems** | UNIX, LINUX, Ubuntu, Windows Vista/7/8/10 |
| **Databases** | Snowflake, MySQL, Teradata, Oracle, MS SQL SERVER, PostgreSQL, DB2. |
| **IDE Dev. Tools** | Eclipse, SOAP UI, Ant, Maven, PyCharm, and Jenkins. |
| **ETL/ Reporting Tools** | Informatica, SSIS, SSRS, SSAS, Tableau, Power BI, QlikView, Arcadia. |
| **No SQL Databases** | HBase, Cassandra, MongoDB, DynamoDB and Cosmos DB |
| **Version Controllers** | Git, SVN, Bitbucket |
| **Web Dev. Technologies** | HTML, XML, JSON, CSS, jQuery, JavaScript |
| **DevOps Tools** | Jenkins, Docker, Maven |
| **Web Services** | SOAP, JMS, Apache Tomcat, WebLogic, JBOSS, Apache HTTP Server. |
| **Methodologies** | RAD, JAD, UML, System Development Life Cycle (SDLC), Jira, Confluence, Agile, Waterfall Model. |

**Project Experience:**

**Client: Equinix Inc – Palo, California Jan 2022 to Present**

**Role: Senior Data Engineer**

**Responsibilities:**

* Implemented a real-time data pipeline for streaming data using **Kafka and Spark Streaming**.
* Designed and maintained optimal data pipeline architecture in **Microsoft Azure cloud** using **Data Factory** and **Azure Databricks**.
* **Developed architecture** for a data services ecosystem that encompassed Relational, **NoSQL**, and **Big Data technologies.**
* Wrote, executed, modified, documents the daily X12N provider, Rx, dental, institutions, and professional claim files from health insurance companies to extract, transform, and load data intoboth **MongoDB** and Oracle database using various ingestion data quality techniques andstrategies.
* The systems are used by DevOps, big data analytic and Analysis, business intelligence, artificial intelligence, data scientist teams.
* Conversion from **MongoDB** back to **Oracle** for better DSS type query processing.
* Utilized **SQL Server Integration Services (SSIS)** for data extraction, transformation, and loading from multiple sources into the target system.
* Migrated data from **Oracle** and SAS to **Hive** and **Azure Data Lake**.
* Developed complex **SQL** scripts and automated resulting scripts and workflows using **Apache Airflow and shell scripting for daily production execution.**
* Creating Documentation on **confluence and updating Jira stories on timely basis.**
* Used Talend for loading data from various sources into the **Hadoop ecosystem**.
* Created **DAX queries** to generate computed columns in **Power BI**.
* Loaded data into **Snowflake DB** in the cloud from various sources.
* Developed production-level Machine Learning classification models and ensemble classification models from scratch using **Python and PySpark** for predicting **binary values** within specific time limits.
* Provided day-to-day support for **GIT** across different projects, including designing and maintaining **GIT** repositories and access control strategies.
* Delivered end-to-end **PySpark ETL pipelines** on **Azure Databricks for data transformation, orchestrated through Azure Data Factory (ADF) and scheduled using Azure Automation Accounts and Tidal Scheduler.**
* **Utilized Jenkins** and pipelines for building and deploying Microservices to **Docker registry and Kubernetes.**
* Led the migration of **data** from on-premises **SQL** Server to Azure Synapse Analytics (DW) and **Azure SQL DB**, ensuring a seamless transition to cloud-based database solutions.
* Responsible for data ingestion using **Sqoop** and **HDFS commands,** collecting partitioned data in various storage formats such as text, **JSON,** and Parquet.
* Ensured successful **ETL/ELT** processes and data loading into **Snowflake DB.**
* Wrote **UNIX shell scripts** for job automation and scheduled cron jobs using Crontab.
* Utilized Kubernetes and Docker for the runtime environment in the **CI/CD system for building, testing, and deploying.**
* Developed **BDD (Behaviour Driven Development)** and **TDD (Test Driven Development)** features, scenarios, and step definitions using **Cucumber, Gherkin, and Ruby**.
* Created pipelines in **Azure Data Factory (ADF)** to extract, transform, and load data from various sources such as **Azure SQL**, **Blob storage**, **Azure SQL Data Warehouse**, and write-back tools.
* Implemented Kafka producer and consumer applications on a **Kafka cluster with the assistance of ZooKeeper**.
* Leveraged Spring **Kafka API** calls to process messages smoothly on the **Kafka cluster**.
* Recreated existing **SQL Server objects** in **Snowflake.**
* Developed **PySpark** and **Spark SQL** transformations in **Azure Databricks** to manage complex business rule implementations.
* Used Power **BI** and Power Pivot for data analysis **prototypes** and employed Power View and Power Map for report visualization.
* Utilized Apache Spark DataFrames, **Spark SQL**, and **Spark MLLib** extensively, developing and designing POCs using **Scala**, **Spark SQL**, and **MLlib libraries.**

**Environment:** Spark-Streaming, Hive, Scala, Hadoop, Azure, Data Bricks, Data Lake, Data Factory, Data Storage, Azure SQL, Kafka, Airflow, Oozie, Spark, Sqoop, Confluence, Docker, Spark SQL, TDD, pig, ETL/ELT, Power BI, Talend, Impala, Oozie, Hbase, Nifi, Zookeeper, Snowflake, Unix/Linux Shell Scripting, Python, PyCharm, CI/CD, Jenkins, Docker, Kubernetes, Microservices, Linux, Shell Scripting, Git

**Client: Finra – Rockville, Maryland Jul 2020 – Dec 21**

**Role: Data Engineer/Hadoop Developer**

**Responsibilities:**

* Implemented data ingestion from various sources into **Hadoop** and **Cassandra** using **Kafka.**
* Supported **MapReduce programs** running on the **cluster** and loaded data from the **UNIX** file system to **HDFS.**
* Developed **Spark** applications using **Spark SQL** in **Databricks** to extract, transform, and aggregate data from multiple file formats, enabling analysis and uncovering insights into customer usage patterns.
* Created, **debugged**, scheduled, and monitored jobs using **Airflow** and **Oozie.**
* Implemented partitioning, dynamic partitions, and buckets in **Hive** for efficient data access.
* Worked on **Spark Streaming** with Kafka for real-time data processing and submission of jobs.
* Designed **AWS** CloudFormation templates to provision **VPC,** **subnets**, and **NAT**, ensuring successful deployment of web applications and database templates.
* Loaded data from various sources, including databases and files, into **Hive** using Talend tool.
* Designed and developed **ETL** processes in **AWS** Glue to migrate campaign data from external sources (**S3**, ORC/Parquet/Text Files) into **AWS** Redshift.
* Utilized Talend big data components, such as **Hadoop** and **S3** Buckets, along with **AWS** services for **Redshift integration**.
* Exported analyzed data to relational databases using **Sqoop** for visualization and report generation with Tableau for the BI team.
* Installed and configured **Apache Airflow** for **S3** bucket and **Snowflake data warehouse**, creating **DAGs** to automate workflows.
* Extracted, aggregated, and consolidated Adobe data within **AWS Glue** using **PySpark.**
* Developed **Spark** jobs to consume data from **Kafka** topics, perform data validations, and push data into HBase and Oracle databases.
* Built visualizations and reports in Tableau using **Snowflake** data.
* Selected and generated data into CSV files, stored them in **AWS** S3 using AWS EC2, and structured and stored them in **AWS** Redshift.
* Configured **ZooKeeper**, **Kafka**, and **Logstash** clusters for data ingestion and optimized Elasticsearch performance.
* Estimated cluster size, monitored, and troubleshooted **Spark Databricks** clusters.
* Utilized **NiFi** data pipeline for processing large datasets and configured lookups for data validation and integrity.
* Unit evaluated data between **Redshift** and **Snowflake**.
* Imported data from various sources into **HDFS using Sqoop**, performed transformations using **Hive** and MapReduce, and loaded the data into **HDFS**.
* Generated reports on predictive analytics using **Python and Tableau**, visualizing model performance and prediction results.
* Used **Jenkins** pipelines to facilitate building microservices, deploying them to **Docker** registry, and managing them with **Kubernetes**.
* Validated Looker reports with Redshift database.
* Developed exception handling code and pushed it into an exception **Kafka** topic.
* Designed **ETL** pipelines to retrieve datasets from **MySQL** and **MongoDB** into AWS S3 bucket, managing access permissions for bucket and objects.
* Architected and designed serverless application **CI/CD** using **AWS Serverless** (**Lambda**) application model.

**Environment:** Hadoop, Kafka, Spark, Spark Databricks, Sqoop, ETL/ELT, Talend, Airflow, Oozie, AWS Tableau, Spark SQL, Spark-Streaming, ETL/ELT, Hive, Scala, pig, Impala, Oozie, Nifi, Hbase, Zookeeper, CI/CD, Jenkins, Kubernetes, Docker, Micro services, Python, Snowflake, Unix

**Client: Elevance Health – Austin, Texas Jan 2019 – Jun 2020**

**Role: Big Data Engineer**

**Responsibilities:**

* Installed, Configured and Maintained Apache Hadoop clusters for application development and Hadoop tools like Hive, Pig, Hbase and HDFS.
* Successfully migrated databases to the **SQL Azure** cloud platform and performed performance tuning to optimize their performance.
* Installed and configured **Sqoop** for seamless import and export of data between relational databases and **Hive.**
* Demonstrated expertise in **Python and Scala**, developing user-defined functions (UDFs) for Hive and Pig using **Python.**
* Created **HiveQL** queries for efficient data retrieval from **HBase** tables and imported work order data into **Hive** tables.
* Experience with **Hadoop** ecosystem components including **Hadoop**, **MapReduce**, **HDFS**, **HBase,** **Hive,** **Sqoop**, **Pig**, **Zookeeper**, and **Flume**.
* Ingested data into Azure services such as **Azure Data Lake**, **Azure Storage**, **Azure SQL, and Azure DW**, and processed it using **Azure Databricks.**
* Implemented **Apache Airflow** for authoring, scheduling, and monitoring data pipelines, designing multiple DAGs to automate ETL pipelines.
* Possess strong experience in ETL concepts, building **ETL** solutions, and data modeling.
* Installed and configured **Hadoop** MapReduce and HDFS, developed **Java** and **Scala** MapReduce jobs for data cleaning and preprocessing.
* Utilized Azure Portal, Azure PowerShell, Storage Accounts, Certificates, and **Azure Data Management** extensively.
* Leveraged **OOZIE** Operational Services for batch processing and dynamically scheduling workflows.
* Architected ETL transformation layers and developed **Spark** jobs for data processing.
* Used **HBase** for storing **Kafka** topics, partition numbers, and offsets values, and employed Phoenix jar for HBase table connection.
* Provided L1 support for **Kafka**-related Jira requests.
* Developed Scala scripts and **UDFs** using **DataFrames, SQL, DataSets,** and **RDD in Spark** for data aggregation, querying, and writing **data** back into **OLTP** systems through **Sqoop.**
* Utilized HBase for storing a massive portion of data requiring regional division.
* Designed **Oozie** workflows for job scheduling and batch processing.
* Wrote Pig scripts to generate **MapReduce** jobs and performed ETL procedures on **HDFS** data.
* Created self-service reporting in **Azure Data Lake** Store **Gen2** using an **ELT approach**.
* Proficient in writing **Sqoop scripts** for importing and exporting data between **RDBMS** and **HDFS.**
* Implemented **Python codebase** for branch management over **Kafka features.**
* Designed and implemented a configurable data delivery pipeline for scheduled updates to customer-facing data stores using **Python.**
* Migrated MapReduce programs to **Spark transformations** using **Spark** and **Scala.**
* Built performant and **scalable ETL** processes for loading, cleansing, and validating data.

**Environment:** Hadoop, Apache Hadoop 2.0.0, HDFS, Hive, Python,Azure, Data Bricks, Data Lake, Data Storage, Data Factory, ETL/ELT, Airflow, Kafka, Map reduce, Scala, spark, Hbase, pig, **zookeeper**, **Sqoop, Flume, Oozie**

**Client: Markk Softech PVT LTD – Hyderabad, India Dec 2012 – Dec 2016**

**Role: Data Engineer/Hadoop Developer**

**Responsibilities:**

* Implementing and Managing **ETL** solutions and automating operational processes.
* Defined and deployed monitoring, metrics, and logging systems on **AWS.**
* Installed and configured **Hadoop MapReduce**, **HDFS**, developed multiple **MapReduce** jobs in **Java** and **Nifi** for data cleaning and pre-processing.
* Worked on **Big data** on **AWS cloud services** i.e., **EC2, S3, EMR** and **DynamoDB**
* Migrated on premise database structure to Redshift **data warehouse.**
* Developed parameter and dimension-based reports, drill-down reports, matrix reports, charts, and Tabular reports using **Tableau Desktop.**
* Was responsible for **ETL** and **data validation using SQL Server Integration Services.**
* Measured Efficiency of **Hadoop/Hive** environment ensuring **SLA** is met.
* Worked on **AWS Data Pipeline** to configure data loads from S3 to into Redshift
* Optimized the Tensor Flow Model for efficiency and **Spark SQ**L for **ETL** jobs and using the right technology for the job to get done.
* Strong understanding of **AWS** components such as **EC2 and S3**
* Created Tableau scorecards, dashboards, Heat maps using show me functionality.
* Used **JSON** schema to define table and column mapping from S3 data to Redshift.
* Developed**Kafka** consumer **API** in **Scala** for consuming data from **Kafka** topics.
* Created ad hoc queries and reports to support business decisions **SQL Server Reporting Services (SSRS).**
* Optimizing and tuning the Redshift environment, enabling queries to perform up to one hundred times faster for **Tableau** and **SAS** Visual Analytics.
* Developed**PIG UDF'S** for manipulating the data according to Business Requirements and worked on developing custom **PIG** Loaders.
* Created group and users in **tableau server**.
* Experienced in working with various kinds of data sources such as **Teradata** and **Oracle**. Successfully loaded files to **HDFS** from **Teradata,** and load loaded from **HDFS** to **hive** and **impala.**
* Worked on **Oozie**workflow engine for job scheduling.
* Experience in setting up the whole app stack, **setup,** and **debug** log stash to send **Apache** logs to **AWS Elastic** search.
* Developed **SSRS** reports, **SSIS** packages to Extract, Transform and Load data from various source systems.

**Environment:** AWS, EC2, S3, Lambda, Redshift, Nifi, SQL Server, Erwin, Kafka, Spark, Pig, Hive, HDFS, Oracle, Informatica, RDS, MySQL, Docker, PostgreSQL, Tableau, Git Hub, Hadoop**,** Hive, HDFS, Spark, Oozie, Map Reduce, Scala, Python, Pyspark, AWS, Oracle 10g, OLTP, Tableau.