Name: Raghu Reddy
Email: rkalvakole@gmail.com
Phone: (469) 751-2360
Current Role: Sr. Big Data/ Data Engineer
LinkedIn: https://www.linkedin.com/in/raghu-reddy-47823a213/

PROFILE SUMMARY:

- Over 8+ years of extensive hands - on Big Data Capacity with the help of Hadoop EcoSystem across internal and cloud- based platforms.
- Expertise in Cloud Computing and Hadoop architecture and its various components - Hadoop File System HDFS, MapReduce, Spark, Name node, Data Node, Job Tracker, Task Tracker, Secondary Name Node.
- Strong experience using HDFS, MapReduce, Hive, Spark, Sqoop, Oozie, and HBase.
- Deep knowledge of troubleshooting and tuning Spark applications and Hive scripts to achieve optimal performance.
- Experienced working with various Hadoop Distributions (Cloudera, Hortonworks, Map R, Amazon EMR) to fully implement and leverage new Hadoop features.
- Experience in developing Spark Applications using Spark RDD, Spark-SQL and Data frame APIs.
- Worked with real-time data processing and streaming techniques using Spark streaming and Kafka.
- Experience in moving data into and out of the HDFS and Relational Database Systems (RDBMS) using Apache Sqoop.
- Expertise in working with HIVE data warehouse infrastructure-creating tables, data distribution by implementing Partitioning and Bucketing, developing and tuning the HQL queries.
- Replaced existing MR jobs and Hive scripts with Spark SQL & Spark data transformations for efficient data processing.
- Deep knowledge of troubleshooting and tuning Spark applications and Hive scripts to achieve optimal performance.
- Experience developing Kafka producers and Kafka Consumers for streaming millions of events per second on streaming data.
- Database design, modeling, migration and development experience in using stored procedures, triggers, cursor, constraints and functions. Used My SQL, MS SQL Server, DB2, and Oracle
- Experience working with NoSQL database technologies, including MongoDB, Cassandra and HBase.
- Experience with Software development tools such as JIRA, Play, GIT.
- 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.
- Good understanding of the Data modeling (Dimensional & Relational) concepts like Star-Schema Modeling, Schema Modeling, Fact and Dimension tables.
- Experience in manipulating/analyzing large datasets and finding patterns and insights within structured and unstructured data.
- Experience working in different Google Cloud Platform Technologies like Big Query, Dataflow, Dataproc, Pubsub, Airflow.
- Design and Development of Ingestion Framework over Google Cloud and Hadoop cluster.

- Hands of experience in GCP, Big Query, GCS bucket, G - cloud function, cloud migration, cloud dataflow, Pub/suB cloud shell, GSUTIL, BQ command line utilities, Data Proc, Stack driver.
- Strong understanding of Java Virtual Machines and multi-threading processes.
- Experience in writing complex SQL queries, creating reports and dashboards.
- Proficient in using Unix based Command Line Interface.
- Strong experience with ETL and/or orchestration tools (e.g. Talend, Oozie, Airflow)
- Experience setting up AWS Data Platform - AWS CloudFormation, Development End Points, AWS Glue, EMR and
- Jupyter/ Sagemaker Notebooks, Redshift, S3, and EC2 instances
- Experienced in using Agile methodologies including extreme programming, SCRUM and Test-Driven Development (TDD)
- Used Informatica Power Center for (ETL) extraction, transformation and loading data from heterogeneous source systems into target databases.

TECHNICAL SKILLS:

- **Programming languages:** Python, PySpark, Shell Scripting, SQL, PL/SQL and UNIX Bash
- **Big Data**: Hadoop, Sqoop, Apache Spark, NiFi, Kafka, Snowflake, Cloudera, Horton Works, PySpark, Spark, Spark SQL
- **Data Modeling Tool**s: Erwin Data Modeler, ER Studio v17
- **Operating Systems**: UNIX, LINUX, Solaris, Mainframes
- **Data bases**: Oracle, SQL Server, My SQL, DB2, Sybase, Netezza, Hive, Impala
- **Cloud Technologies**: AWS, AZURE, GCP
- **IDE Tools**: Aginitiy for Hadoop, PyCharm, Toad, SQL Developer, SQL *Plus, Sublime Text, VI Editor
- **OLAP Tools**: Tableau, SSAS, Business Objects, and Crystal Reports 9
- **ETL/Data warehouse Tools:** Informatica 9.6/9.1, and Tableau.
- **Others**: AutoSys, Crontab, ArcGIS, Clarity, Informatica, Business Objects, IBM MQ, Splunk

PROFESSIONAL EXPERIENCE:

Client: Prudential Financial, Newark, NJ **Feb 2022 - present**
Role: Big Data/ Data Engineer
Roles & Responsibilities:

- Implemented Azure Data Factory (ADF) extensively for ingesting data from different source systems like relational and unstructured data to meet business functional requirements.
- Design and developed Batch processing and real-time processing solutions using ADF, Databricks clusters and stream Analytics.
- Created numerous pipelines in Azure using Azure Data Factory v2 to get the data from disparate source systems by using different Azure Activities like Move Transform, Copy, filter, for each, Databricks etc. Maintain and provide support for optimal pipelines, data flows and complex data transformations and manipulations using ADF and PySpark with Databricks.
- Automated jobs using different triggers like Events, Schedules and Tumbling in ADF.

- Created, provisioned different Databricks clusters, notebooks, jobs and autoscaling.
- Performed data flow transformation using the data flow activity.
- Used Polybase to load tables in Azure synapse.
- Implemented Azure, self-hosted integration runtime in ADF.
- Improved performance by optimizing computing time to process the streaming data by optimizing the cluster run time.
- Perform ongoing monitoring, automation, and refinement of data engineering solutions.
- Scheduled, automated business processes and workflows using Azure Logic Apps.
- Designed and developed a new solution to process the NRT data by using Azure stream analytics, Azure Event Hub and
- Service Bus Queue.
- Created Linked services to connect the external resources to ADF.
- Worked with complex SQL views, Stored Procedures, Triggers, and packages in large databases from various servers.
- Used Azure Devops & Jenkins pipelines to build and deploy different resources(Code and Infrastructure)in Azure.
- Ensure the developed solutions are formally documented and signed off by business.
- Worked with team members to resolve any technical issue, Troubleshooting, Project Risk & Issue Identification, and
- management.
- Worked on the cost estimation, billing, and implementation of services on the cloud.
- Experience managing Azure Data Lakes (ADLS) and Data Lake Analytics and an understanding of how to integrate with other Azure Services.
- Migration of on premise data (Oracle/ Teradata) to Azure Data Lake Store(ADLS) using Azure Data Factory(ADF V1/2).
- Work closely across teams (Support, Solution Architecture) and peers to establish and follow best practices while solving.
- customer problems
- Created infrastructure for on time extraction. transformation. and loading of data from a wide variety of data sources.

Environment: Azure Data Factory (ADF V2), Azure SQL Database, Azure functions Apps, Azure Data Lake, BLOB Storage, SQL server, Windows remote desktop, UNIX Shell Scripting, AZURE PowerShell, Data bricks, Python, ADLS Gen 2, Azure Cosmos DB, Azure Event Hub, Azure Machine Learning

Client: Lending Tree, Charlotte, NC **Jun 2021- Jan 2022**
Role: Senior Big Data Engineer
Roles & Responsibilities:

- Experience in data ingestion techniques for batch and stream processing using AWS Batch, AWS Kinesis, AWS Data Pipeline.
- Working on building centralized Data lake on AWS cloud utilizing primary services like s3,EMR, Glue, Athena.

- Build series of PYSpark applications using python and Hive scripts to produce various analytical data sets needed for data science and Marketing analytic's teams.
- Worked extensively on fine tuning spark applications and providing production support to various pipelines running in production.
- Build API based ingestions to bring in data to the data lake through the data pipelines built and provide access to different teams for the data.
- Worked on full spectrum of data engineering pipelines: data ingestion, data transformations and data analysis/consumption.
- Worked on analysis for different datasets and did complete EDA for the data based on the requirements from the Business Users.
- Worked on file based ingestion from various sources through the data pipelines built across the team to store the data to data lake.
- Written AWS Lambda code in Python for nested Json files, converting, comparing, sorting etc.
- Good understanding of the Data modeling (Dimensional & Relational) concepts like Star-Schema Modeling, Schema Modeling, Fact and Dimension tables.
- Experience in manipulating/analyzing large datasets and finding patterns and insights within structured and unstructured data.
- Experience in writing complex SQL queries, creating reports and dashboards.
- Proficient in using Unix based Command Line Interface.
- Worked on API Based ingestion to bring data to AWS S3 using AWS glue and AWS App flows data lake.
- Worked on exporting data to different applications used for marketing using AWS app flow and AWS glue jobs.
- Overall experience in working with large datasets and ingestion pipelines in AWS environment.

Environment: Apache Spark, Hadoop, PySpark, HDFS, Cloudera, AWS, Azure, Kafka, Snowflake, Docker, Jenkins, Ant, Maven, Kubernetes, Nifi, JSON, Teradata, DB2, SQL Server, MongoDB, Shell Scripting.

Client: AMEX, Phoenix, AZ **Sep 2020 - May 2021**

Role: Sr. Data Engineer / Big Data Engineer

Roles & Responsibilities:

- Meetings with business/user groups to understand the business process, gather requirements, analyze, design, develop and implement according to client requirements.
- Designing and Developing Azure Data Factory (ADF) extensively for ingesting data from different source systems like relational and Non relational to meet business functional requirements.
- Designed and Developed event driven architectures using blob triggers and DataFactory.
- Creating pipelines, data flows and complex data transformations and manipulations using ADF and PySpark with Databricks.
- Automated jobs using different triggers like Events, Schedules and Tumbling in ADF.
- Created, provisioned different Databricks clusters, notebooks, jobs and autoscaling.
- Ingested huge volume and variety of data from disparate source systems into Azure DataLake Gen2 using Azure Data Factory V2.
- Created several Databricks Spark jobs with Pyspark to perform several tables to table operations.
- Performed data flow transformation using the data flow activity.

- Implemented Azure, self-hosted integration runtime in ADF.
- Developed streaming pipelines using Apache Spark with Python.
- Created, provisioned multiple Databricks clusters needed for batch and continuous streaming data processing and installed the required libraries for the clusters.
- Improved performance by optimizing computing time to process the streaming data and saved cost to the company by optimizing the cluster run time.
- Perform ongoing monitoring, automation, and refinement of data engineering solutions.
- Designed and developed a new solution to process the NRT data by using Azure stream analytics, Azure Event Hub and Service Bus Queue.
- Created Linked service to land the data from SFTP location to Azure Data Lake.
- Extensively used SQL Server Import and Export Data tool.
- Working with complex SQL views, Stored Procedures, Triggers, and packages in large databases from various servers.
- Experience in working on both agile and waterfall methods in a fast pace manner.
- Generating alerts on the daily metrics of the events to the product people.
- Extensively used SQL Queries to verify and validate the Database Updates.
- Suggest fixes to complex issues by doing a thorough analysis of root cause and impact of the defect.
- Provided 24/7 On-call Production Support for various applications and provided resolution for night-time production job, attend conference calls with business operations, system managers for resolution of issues.
- Designs and implementing Scala programs using Spark Data frames and RDDs for transformations and actions on input data.
- Improved the Hive queries performance by implementing partitioning and clustering and Optimized file formats (ORC).

Environment: Azure Data Factory (ADF v2), Azure SQL Database, Azure functions Apps, Azure Data Lake, BLOB Storage, SQL server, Windows remote desktop, UNIX Shell Scripting, AZURE PowerShell, Data bricks, Python, ADLS Gen 2, Azure Cosmos DB, Azure Event Hub, Azure Machine Learning.

JP Morgan Chase, NYC, NY **December 2017 – Aug 2020**
Role: Big Data Engineer
Roles &
Responsibilities:

- Implemented Installation and configuration of multi-node cluster on Cloud using Amazon Web Services (AWS) on EC2.
- Handled AWS Management Tools as Cloud watch and Cloud Trail.
- Stored teh log files in AWS S3. Used versioning in S3 buckets where the highly sensitive information is stored.
- Integrated AWS DynamoDB using AWS lambda to store the values of items and backup teh DynamoDB streams
- Automated Regular AWS tasks like snapshots creation using Python scripts.
- Designed data warehouses on platforms such as AWS Redshift, Azure SQL Data Warehouse, and other high-performance platforms.
- Install and configure Apache Airflow for AWS S3 bucket and created dags to run the Airflow

- Prepared scripts to automate the ingestion process using Pyspark and Scala as needed through various sources such as API, AWS S3, Teradata and Redshift.
- Created multiple scripts to automate ETL/ ELT process using Pyspark from multiple sources
- Developed Pyspark scripts utilizing SQL and RDD in spark for data analysis and storing back into S3
- Developed Pyspark code to load from stg to hub implementing the business logic.
- Developed code in Spark SQL for implementing Business logic with python as programming language.
- Designed, Developed and Delivered teh jobs and transformations over the data to enrich the data and progressively elevate for consuming in the Pub layer of the data lake.
- Worked on Sequence files, Map side joins, bucketing, partitioning for hive performance enhancement and storage improvement.
- Wrote, compiled, and executed programs as necessary using Apache Spark in Scala to perform ETL jobs wif ingested data.
- Used Spark Streaming to divide streaming data into batches as an input to Spark engine for batch processing.
- Maintained Kubernetes patches and upgrades.
- Managed multiple Kubernetes clusters in a production environment.
- Wrote Spark applications for data validation, cleansing, transformation, and custom aggregation and used Spark engine, Spark SQL for data analysis and provided to the data scientists for further analysis
- Developed various UDFs in Map-Reduce and Python for Pig and Hive.
- Data Integrity checks have been handled using hive queries, Hadoop, and Spark.
- Worked on performing transformations & actions on RDDs and Spark Streaming data wif Scala.
- Implemented the Machine learning algorithms using Spark with Python.
- Implemented a Continuous Delivery pipeline with Docker and GitHub
- Worked with g-cloud function with Python to load Data in to Bigquery for on arrival csv files in GCS bucket
- Transformed batch data from several tables containing tens of thousands of records from SQL Server, MySQL, PostgreSQL, and csv file datasets into data frames using PySpark.
- Researched and downloaded jars for Spark-avro programming.
- Developed a PySpark program that writes dataframes to HDFS as avro files.
- Utilized Spark's parallel processing capabilities to ingest data.
- Created and executed HQL scripts that creates external tables in a raw layer database in Hive.
- Developed a Script that copies avro formatted data from HDFS to External tables in raw layer.
- Created PySpark code that uses Spark SQL to generate dataframes from avro formatted raw layer and writes them to data service layer internal tables as orc format.
- In charge of PySpark code, creating dataframes from tables in data service layer and writing them to a Hive data warehouse.
- Installed Airflow and created a database in PostgreSQL to store metadata from Airflow.
- Configured documents which allow Airflow to communicate to its PostgreSQL database.
- Developed Airflow DAGs in python by importing the Airflow libraries.
- Utilized Airflow to schedule automatically trigger and execute data ingestion pipeline.
- Process and load bound and unbound Data from Google pub/sub topic to Bigquery using cloud Dataflow with Python.

Environment: AWS, JMeter, Kafka, Ansible, Jenkins, Docker, Maven, Linux, Red Hat, GIT, Cloud Watch, Python, Shell Scripting, Golang, Web Sphere, Splunk, Tomcat, Soap UI, Kubernetes, Terraform, PowerShell.

Client: Apollo, Bengaluru, India **June 2015- July 2017**

Role: Hadoop Developer

Roles& Responsibilities:

- Worked on development of data ingestion pipelines using ETL tool, Talend & bash scripting with big data technologies including but not limited to Hive, Impala, Spark, Kafka, and Talend.
- Experience in developing scalable & secure data pipelines for large datasets.
- Gathered requirements for ingestion of new data sources including life cycle, data quality check, transformations, and metadata enrichment.
- Supported data quality management by implementing proper data quality checks in data pipelines.
- Delivered data engineer services like data exploration, ad-hoc ingestions, subject-matter-expertise to Data scientists in using big data technologies.
- Build machine learning models to showcase Big data capabilities using Pyspark and MLlib.
- Involved in converting Hive/SQL queries into Spark transformations using Spark data frames, Scala and Python.
- Experienced in developing Spark scripts for data analysis in both python and Scala.
- Wrote Scala scripts to make spark streaming work with Kafka as part of spark Kafka integration efforts.
- Built on premise data pipelines using Kafka and spark for real-time data analysis.
- Created reports in TABLEAU for visualization of the data sets created and tested Spark SQL connectors.
- Implemented Hive complex UDF's to execute business logic with Hive Queries.
- Developed a different kind of custom filter and handled predefined filters on HBase data using API.
- Implemented Spark using Scala and utilizing Data frames and Spark SQL API for faster processing of data.
- Handled importing data from different data sources into HDFS using Sqoop and performing transformations using Hive and then loading data into HDFS.
- Enhancing Data Ingestion Framework by creating more robust and secure data pipelines.
- Implemented data streaming capability using Kafka and Talend for multiple data sources.
- Worked with multiple storage formats (Avro, Parquet) and databases (Hive, Impala, Kudu).
- Working knowledge of cluster security components like Kerberos, Sentry, SSL/TLS etc.
- Involved in the development of agile, iterative, and proven data modeling patterns that provide flexibility.
- Knowledge on implementing the JILs to automate the jobs in the production cluster.
- Troubleshot user's analyses bugs (JIRA and IRIS Ticket).
- Worked with SCRUM team in delivering agreed user stories on time for every Sprint.
- Worked on analyzing and resolving production job failures in several scenarios.
- Implemented UNIX scripts to define the use case workflow and to process the data files and automate the jobs.
- Exporting of a result set from HIVE to MySQL using Sqoop export tool for further processing.
- Collecting and aggregating large amounts of log data and staging data in HDFS for further analysis.
- Experience in managing and reviewing Hadoop Log files.
- Used Sqoop to transfer data between relational databases and Hadoop.
- Worked on HDFS to store and access huge datasets within Hadoop.
- Good hands on experience with GitHub.

- Involved in review of functional and non-functional requirements.
- Installed and configured Hadoop MapReduce, HDFS, Developed multiple MapReduce jobs in java for data cleaning and preprocessing.
- Imported Legacy data from SQL Server and Teradata into Amazon S3.
- Migrating data from FS to Snowflake within the organization
- Created consumption views on top of metrics to reduce the running time for complex queries.
- Compare the data in a leaf level process from various databases when data transformation or data loading takes place. I need to analyze and look into the data quality when these types of loads are done (To look for any data loss, data corruption).
- As a part of Data Migration, wrote many SQL Scripts for Mismatch of data and worked on loading the history data from Teradata SQL to snowflake.
- Developed SQL scripts to Upload, Retrieve, Manipulate and handle sensitive data (National Provider Identifier Data I.e., Name, Address, SSN, Phone No) in Teradata, SQL Server Management Studio and Snowflake Databases for the Project.
- Worked on to retrieve the data from FS to S3 using spark commands.
- Built S3 buckets and managed policies for S3 buckets and used S3 bucket and Glacier for storage and backup on AWS.
- Using Nebula Metadata, registered Business and Technical Datasets for corresponding SQL scripts
- Experienced in working with the Spark ecosystem using Spark SQL and Scala queries on different formats like text file, CSV file.
- Developed spark code and spark-SQL/streaming for faster testing and processing of data.
- Closely involved in scheduling Daily, Monthly jobs with Precondition/Postcondition based on the requirement.
- Monitor the Daily, Weekly, Monthly jobs and provide support in case of failures/issues.
- Installed and configured Pig and wrote Pig Latin scripts.
- Wrote a MapReduce job using Pig Latin. Involved in ETL, Data Integration and Migration.
- Imported data using Sqoop to load data from Oracle to HDFS on a regular basis.
- Developing Scripts and Batch Job to schedule various Hadoop Program.
- Written Hive queries for data analysis to meet the business requirements.
- Creating Hive tables and working on them using Hive QL. Experienced in defining job flows.
- Importing and exporting data into HDFS from Oracle Database and vice versa using sqoop.
- Designed and implemented MapReduce-based large-scale parallel relation-learning system
- Setup and benchmarked Hadoop/HBase clusters for internal use
- Created Metric tables, End user views in Snowflake to feed data for Tableau refresh.
- Generated Custom SQL to verify the dependency for the daily, Weekly, Monthly jobs.

Environment: Hadoop, MapReduce, HDFS, Hive, Java, Hadoop distribution of Cloudera, Pig, HBase, Linux, XML, Java 6, Eclipse, Oracle 10g, PL/SQL, MongoDB, Toad