Rakesh Kumar Reddy Poreddy

rakesh22rreddy@gmail.com +1 (216) 9033355 LinkedIn: LinkedIn/In/Rakesh

PROFESSIONAL SUMMARY

- 9+ years' experience in Software/Application Development using Python, Scala, C, SQL, and in-depth understanding of Distributed Systems Architecture and Parallel Processing Frameworks.
- Deep knowledge and strong deployment experience in Hadoop and bigdata ecosystems - HDFS, MapReduce, Spark, Pig, Sqoop, Hive, Oozie, Kafka, zookeeper, and HBase.
- Knowledge on current trends in data technologies, data services, data virtualization, data integration, Master Data Management.
- Used various Hadoop distributions (Cloudera, Hortonworks, Amazon EMR, Microsoft Azure HDInsight) to fully implement and leverage new Hadoop features.
- Constructing and manipulating large datasets of structured, semi-structured, and unstructured data and supporting systems application architecture using tools like SAS, SQL, Python, R, Minitab, Power BI, and more to extract multi-factor interactions and drive change.
- In-Depth understanding of Snowflake Multi-cluster Size and Credit Usage Played a key role in Migrating Teradata objects into the Snowflake environment.
- Responsible for converting all ETL logic into SQL queries, create INFA mapping to load into Netezza and Snowflake database.
- Experience in moving data into and out of the HDFS and Relational Database Systems (RDBMS) using Apache Sqoop.
- Hands on experience in Oracle ADF 11g Architecture, JDeveloper, Oracle ADF Development (ADF, ADF Faces, ADF TaskFlows and Business Components) with Web Services.
- Designed and Developed ETL Processes in AWS Glue to migrate Campaign data from external sources like S3, ORC/Parquet/Text Files into AWS Redshift.
- Experience object-oriented programming (OOP) concepts using Python, C++, and PHP.
- Experience with Snowflake Multi-Cluster Warehouses.
- Experience with Snowflake Virtual Warehouses.
- Involved in Migrating Objects from Teradata to Snowflake.
- Expertise in working with Hive data warehouse infrastructure creating tables, data distribution by implementing Partitioning and Bucketing, developing, and tuning the HQL queries.
- Strong experience in tuning Spark applications and Hive scripts to achieve optimal performance.
- Wrote MapReduce jobs using Pig Latin, Optimized the existing Hive and Pig Scripts.
- Developed Spark Applications using Spark RDD, Spark-SQL, and Data Frame APIs.
- Proficient in using QlikView and Tableau.
- Experience with healthcare interoperability tools and protocols, including HL7, FHIR, CDA and EDI.
- Worked on FHIR and NCPD methodologies based on business requirements for information transfer as per the regulations.
- Experienced in designing standards for using normalized data structures, de-normalized structures, and dimensional structures. Defines common design patterns for modeling various types of relationships.
- Experienced in Batch processes, Import, Export, Backup, Database Monitoring tools, and Application support and experienced in big data analysis and developing data models using Hive, PIG, and Map-reduce, SQL with strong data architecting skills designing data-centric solutions.
- Experienced in Teradata SQL queries, Teradata Indexes, Utilities such as Mload, Tpump, Fast load, and Fast Export.
- Excellent in performing data transfer activities between SAS and various databases and data file formats like XLS, CSV, DBF, MDB, etc.
- Experienced in Data Scrubbing/Cleansing, Data Quality, Data Mapping, Data Profiling, Data Validation in ETL.
- Excellent Knowledge of Ralph Kimball and Bill Inmon's approaches to Data Warehousing.
- Excellent experience and knowledge in developing Informatica Mappings, Mapplets, Sessions, Workflows, and Worklets for data loads from various sources such as Oracle, Flat Files, DB2, SQL Server, etc.
- Experienced in writing UNIX shell scripting and hands-on experience with scheduling shell scripts using Control-M.

TECHNICAL SKILLS

- Hadoop/Big Data Technologies: HDFS, Apache NIFI, Map Reduce, Sqoop, Flume, Pig, Hive, Oozie, Impala, Zookeeper, Ambari, Storm,Spark, and Kafka.
- No SQL Database: HBase, Cassandra, MongoDB.
- Hadoop Distribution: Horton Works, Cloudera, MapReduce.
- Databases: Oracle SQL server, MY SQL, MS SQL Server, Vertica, Teradata, Snowflake.
- Programming and Scripting: Java, SQL, JavaScript, Shell Scripting, Python, Pig Latin, HiveQL.
- Java Technologies: J2EE, Java Mail API, JDBC.
- Analytics Tools: Tableau, Power BI, Microsoft SSIS, SSAS and SSRS.
- Web Dev. Technologies: HTML, XML, JSON, CSS, JQUERY, JavaScript.
- Operating Systems: Linux, Unix, Windows 8, Windows 7, Windows Server 2008/2003.
- AWS Services: Amazon EC2, Amazon S3, Amazon Simple DB, Amazon MQ, Amazon ECS, Amazon Lambdas, Amazon Sagemaker, Amazon RDS, Amazon Elastic Load Balancing, Elastic Search, Amazon SQS, AWS Identity and access management, AWS Cloud Watch, Amazon EBS and Amazon CloudFormation.
- Azure Services: Azure Data Lake, Data factory, Azure Databricks, Azure SQL database, Azure SQL Datawarehouse, Azure Functions, Azure Synapse, Azure HD Insights, Azure Blob Storage, Azure Event Hub, Azure streaming Analytics.
- GCP Services: BigQuery, Cloud DataProc, GCS.
- CI/CD tools: Git, Jira, Jenkins.
- Network protocols: TCP/IP, UDP, HTTP, DNS, DHCP.

PROFESSIONAL WORK EXPERIENCE

German Town Technologies, Portsmouth, VA – Data Integrations Engineer **Apr 2023 – Present**

- Designed, built, and launched efficient and reliable data pipelines in Databricks to move and transform large-scale research data, ensuring high performance and scalability.
- Led the refactoring and optimization of existing data flows from multiple source systems, enhancing performance and reducing processing time for daily batch and real-time streaming data.
- Architected and implemented real-time data processing solutions using Azure Databricks and Spark Streaming, enabling the ingestion, processing, and analysis of high-velocity data streams from various sources.
- Integrated Azure Databricks with Apache Kafka and Azure Event Hubs to seamlessly ingest streaming data, providing real-time analytics and insights for critical business operations.
- Developed custom connectors and integrations to streamline the ingestion and processing of streaming data from IoT devices and real-time data feeds into Azure Databricks.
- Utilized Azure Data Factory pipelines and Scala scripts with Spark to optimize data processing for large-scale, sensitive drug manufacturing data, improving efficiency and accuracy.
- Implemented Azure Purview for comprehensive data governance, facilitating automated data discovery, classification, and lineage tracking across heterogeneous data sources. Enhanced data security and compliance with GDPR, HIPAA, and CCPA standards by definingand enforcing data access policies.
- Developed and maintained robust data models, encompassing conceptual, logical, and physical structures, to support healthcare analytics, ensuring data consistency and accuracy through effective normalization, indexing, and partitioning practices.
- Orchestrated the creation and management of intricate database schemas, optimizing data retrieval and reporting efficiency for clinical and operational insights within the healthcare domain.
- Implemented Infrastructure as Code (IaaC) using Terraform to design and deploy Azure resources, ensuring consistent, scalable, and reproducible environments. Managed infrastructure configurations as code, minimizing manual errors and enabling version-controlled infrastructure changes, thus enhancing efficiency and reliability of the system.
- Led the implementation of Event-Driven Architecture (EDA) principles to enhance responsiveness and scalability of data integrations forcritical business processes. Implemented Azure Event Hub and Azure Service Bus to establish dependable communication channels between microservices and data systems, facilitating seamless real-time data processing and analytics.

- Developed and optimized Azure Functions in .NET C# to execute business logic for event processing and data workflow orchestration within the Azure ecosystem. Ensured high performance and scalability to efficiently manage large volumes of data streams, meeting the client's rigorous performance standards.
- Designed and implemented resilient data integration pipelines capable of ingesting, transforming, and delivering real-time data from various sources like IoT devices and enterprise systems to Azure data stores. Implemented comprehensive data validation, enrichment,and normalization processes to uphold data quality and integrity throughout the integration lifecycle, ensuring reliable and accurate insights for stakeholders.
- Collaborated with cross-functional teams to define streaming data requirements, ensuring seamless integration with downstream systems and data warehouses for unified data processing.
- Implemented end-to-end monitoring and logging solutions leveraging Azure Monitor and Application Insights to proactively detect performance bottlenecks, errors, and security threats. Conducted thorough performance tuning and optimization exercises to boost the efficiency and cost-effectiveness of Azure resources, resulting in substantial cost savings for the client.
- Played a pivotal role in cross-functional collaboration, liaising with data engineers, software developers, and business stakeholders to comprehend requirements, pinpoint technical hurdles, and deliver inventive solutions in line with business goals. Additionally, served asa subject matter expert (SME), offering mentorship to junior team members, thus cultivating a culture of perpetual learning and knowledge exchange within the organization.
- Maintained regulatory compliance with HIPAA and GDPR standards through the implementation of rigorous security protocols, including data encryption, access controls, and robust security measures within the Azure environment. Conducted routine security assessments and audits to proactively identify and address potential vulnerabilities, safeguarding the confidentiality, integrity, and availability of sensitive research data.
- Used ETL to implement the Slowly Changing Transformation, to maintain Historical data in Legacy Databricks – Hive Metastore Data warehouse.
- Implemented Copy activity and custom Azure Data Factory Pipeline Activities to support custom use-cases specific to each system beingintegrated.
- Built analytics tools that utilize data pipelines to provide actionable insight into research, operational efficiency and other key business performance metrics.
- Worked with internal and external stakeholders to assist with data related technical issues and support data infrastructure needs.
- Implemented end-to-end streaming data workflows, from data ingestion to transformation and visualization, providing stakeholders with real-time dashboards and actionable insights.

QBE, NY - Data Engineer
Jan 2021 -Mar 2023

- Developed and maintained robust processes for data transformation, structuring, metadata management, dependency tracking, andworkload management using Databricks.
- Conducted complex data analysis by analyzing, organizing, interpreting, and assembling raw data, while defining new data collection andanalysis methodologies to enhance data quality and reliability.
- Provided production and on-call support for multiple projects, ensuring system reliability and performance.
- Created Tables & built views on top of them to cater the business needs. Used Databricks Internal connectors with Power BI to createdashboards and support daily reporting.
- Utilized Azure pipelines and python scripts to perform data migration from different sources including Teradata, SQL Server, andincremental file ingestion into Snowflake.
- Create pipelines in ADF using linked services to extract, transform and load data from multiple sources like Azure SQL, Blob storage andAzure SQL Data warehouse.
- Utilized Control-M to schedule, orchestrate, and monitor complex job workflows, ensuring timely and accurate execution of dataprocessing tasks.
- Established and maintained comprehensive data governance frameworks, including data quality standards, metadata management, andaccess controls.
- Spearheaded the migration of complex data sets from Teradata to Snowflake, ensuring data integrity and compliance with minimaldowntime.

- Used ETL to implement the Slowly Changing Transformation, to maintain Historical data in the Data warehouse.
- Designed DDL statements, complex SQL queries, views, and indexes to create and test tables.
- Optimized Snowflake performance for large-scale data analytics, resulting in a 40% improvement in query response times.

PNC Bank, Pittsburgh, PA - Sr. Data Engineer **Oct 2018 - Dec 2020**

- Developed Spark Applications by using Python and Implemented Apache Spark data processing project to handle data from various RDBMS and Streaming sources. Responsible for performing sort, join, aggregations, filter, and other transformations on the datasets using Spark Extract Real-time feed using Kafka and Spark Streaming and convert it to RDD and process data in the form of Data Frame and save the data as Parquet format in HDFS.
- Collaborated with Analysts and other departments to report and review the documents for tiebreaker analysis and metadata documents. Responsible for data services and data movement infrastructures. Experience with ETL concepts, building ETL solutions and Data modelling.
- Architected several DAGs (Directed Acyclic Graph) for automating ETL pipelines.
- Worked extensively in creating a Scala code to move staging process to standardization meeting business requirements.
- Involved in processing and wrangling of huge data volumes of log files data at scale. involved in Data wrangling, cleaning, and parsing the log data sets to extract structured attributes with meaningful information from each log message.
- Developed Python scripts, UDFs using both Data frames/SQL/Data sets and RDD/MapReduce in Spark for Data Aggregation, queries and writing data back into OLTP system through Sqoop.
- Worked with Amazon Web Services (AWS) using EC2 for hosting and Elastic map reduce (EMR) for data processing with S3 as a storage mechanism.
- Experienced in writing live Real-time Processing and core jobs using Spark Streaming with Kafka as a data pipe-line system.
- Worked on AWS Data pipeline to configure data loads from S3 to into Redshift.
- Using AWS Redshift, I Extracted, transformed, and loaded data from various heterogeneous data sources and destinations.
- Worked with X12 EDI standards for healthcare data, HEDIS, and HIPAA.
- Validated the reports and files according to HIPAA X12 standards.
- Validation of HL7 messages for results to be sent from organizational database to hospitals.
- Used Data Transformation Studio to configure and deploy data transformation service required to load HL7 messages from message queue to relational tables.
- Wrote scripts and indexing strategy for a migration to Confidential Redshift from SQL Server and MySQL databases.
- Used AWS glue catalog with crawler to get the data from S3 and perform SQL query operations using Crawlers and scheduled the job and crawler using workflow feature.
- Worked on SnowSQL and Snowpipe and converted Talend Joblets to support the snowflake functionality.
- Created Snowpipe for continuous data load and used COPY to bulk load the data into Snowflake.
- Created data sharing between two snowflake accounts and created internal and external stage to transform data during load stage.
- Worked on Apache NiFi to support scalable directed graphs for data routing, transformation and automating the movement of data between disparate systems.
- Implemented Lambda to configure Dynamo DB Autoscaling feature and implemented Data Access Layer to access AWS DynamoDB data.
- Experience in performance tuning a Cassandra cluster to optimize writes and reads.
- Worked on cloud deployments using Maven, Docker, and Jenkins.

Exl Service.com (I) Pvt.Ltd, India - Data Engineer **Apr 2017 - Jan 2018**

- Experience in Designing and Developing Spark applications using PySpark 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.
- Successfully transferred dev environment data to QA testing and scheduled airflow testing jobs.
- Involved in designing the Dim processes and assisted in creating domain tables.
- Created airflow jobs using python code and scheduled development jobs.
- Involved in Extract, Transform and Load (ETL) 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. Involved in Data Ingestion to one or more Azure Services (Azure Data Lake, Azure Storage, Azure SQL, Azure DW) and processing the data in In

Azure Databricks.

- Architected & implemented medium to large scale BI solutions on Azure using Azure Data Platform services (Azure Data Lake, Data Factory, Data Lake Analytics, Stream Analytics, Azure SQL DW, HDInsight/Databricks, NoSQL DB).
- Developed Spark applications in Python (PySpark) on distributed environment to load huge number of CSV files with different schema.
- Improved performance by optimizing computing time to process the streaming data and saved cost to company by optimizing the clusterrun time.
- Worked with Text, Avro, and Parquet file formatted and snappy as a default compression.
- Design & implement migration strategies for traditional systems on Azure (Lift and shift/Azure Migrate, other third-party tools) workedon Azure suite: Azure SQL Database, Azure Data Lake (ADLS), Azure Data Factory (ADF) V2, Azure SQL Data Warehouse, Azure Service Bus, Azure key Vault, Azure Analysis Service (AAS), Azure Blob Storage, Azure Search, Azure App Service, Azure data Platform Services.
- Design and implement end-to-end data solutions (storage, integration, processing, visualization) in Azure.
- Experience managing Azure Data Lakes (ADLS) and Data Lake Analytics and an understanding of how to integrate with other AzureServices. Worked on USQL and how it can be used for data transformation as part of a cloud data integration strategy.
- Established database standards for operations, upgrades, migrations and onboarding new applications and/or customers.
- Used Cosmos DB for storing catalog data and for event sourcing in order processing pipelines.
- Designed and developed user defined functions, stored procedures, triggers for Cosmos DB.
- Implemented CI/CD pipelines using Azure DevOps platform and built deployment pipelines to manage the release process. Used Azure DevOps as a source code repository for Azure Databricks notebooks to automatically deploy a notebook from dev to prod Azure Databricks Workspace.
- Used Azure Repo as a SVM tool.
- Created various reports using Tableau and QlikView based on requirements with the BI team.

High Radius Technologies, India- Data Engineer **Jul 2015 - Mar 2017**

- Design and development of IT solutions using Big Data tools.
- Created Pipelines in ADF using Linked Services/Datasets/Pipeline/ to extract, transform, and load data from different sources like AzureSQL, Blob storage, Azure SQL Data warehouse, write-back tool and backwards.
- Participated in detailed technical design, development, implementation, and support of Data applications. Develop, construct, test,automate and maintain Data Pipelines.
- Extensively worked with Spark-SQL context to create data frames and datasets to pre-process the model data.
- Responsible for estimating the cluster size, monitoring, and troubleshooting of the Hadoop cluster.
- Used Zeppelin, Jupyter notebooks and Spark-Shell to develop, test and analyze Spark jobs before Scheduling Customized Spark jobs.
- Undertake data analysis and collaborated with down-stream, analytics team to shape the data according to their requirement.
- Experienced in performance tuning of Spark Applications for setting right Batch Interval time, correct level of Parallelism and memory tuning.
- To meet specific business requirements wrote UDF's in Scala and Stored procedures.
- Replaced the existing MapReduce programs and Hive Queries into Spark application using Scala.
- Responsible for documenting the process and cleanup of unwanted data.
- Responsible for Ingestion of Data from Blob to Kusto and maintaining the PPE and PROD pipelines.
- Expertise in creating HDInsight cluster and Storage Account with End-to-End environment for running the jobs.
- Developed Json Scripts for deploying the Pipeline in Azure Data Factory (ADF) that process the data using the Cosmos Activity.
- Hands-on experience on developing PowerShell Scripts for automation purpose.
- Created Build and Release for multiple projects (modules) in production environment using Visual Studio Team Services (VSTS).
- Experience in using Scala Test Fun Suite Framework for developing Unit Tests cases and Integration testing.
- Involved in running the Cosmos Scripts in Visual Studio 2017/2015 for checking the diagnostics.
- Worked in Agile development environment in sprint cycles of two weeks by dividing and organizing tasks.