Rehitha Kilaru
Sr Data Engineer | kilaru.rehitha27@gmail.com | +1 (720) 800-6667

SUMMARY

- Around 8+ years of experience as an IT professional specialized in Big Data Analytics using Hadoop Ecosystem.
- Hands - on experience in installing, configuring, and using Hadoop ecosystem components like HDFS, MapReduce, Spark, Yarn, Kafka, Sqoop, Flume, Pig, Hive, HBase, Airflow, Storm, Oozie, Zookeeper, Impala.
- Experience with AWS Services like EC2, S3, EMR, RDS, VPC, Elastic Load Balancing, IAM, Auto Scaling, RedShift, DynamoDB, Security Groups, Kinesis, Glue, EKS, CloudWatch, Lambda, SNS, SES, SQS.
- Experience in building data pipelines using Azure Data Factory, AzureDatabricks, and loading data to Azure Data Lake, Azure SQL Database, Azure SQL Data Warehouse to control and grant database access.
- Worked with Azure services like HDInsight, Event Hubs, Stream Analytics, ActiveDirectory, Blob Storage, Cosmos DB.
- Expertise in building PySpark and Spark-Scala applications for interactive analysis, batch and stream processing.
- Experience in configuring Spark Streaming to receive real-time data from Kinesis and store stream data to HDFS.
- Worked with different Hadoop Distributions - Cloudera, Amazon EMR, Azure HDInsight, and Hortonworks.
- Extensively used Spark Data Frame API over Cloudera platform to perform analytics on Hive data and used Spark Data Frame Operations to perform requiredValidations in the data.
- Involved in loading structured and semi-structured data into spark clusters using SparkSQL and Data Frames API.
- Capable of handling and ingesting terabytes of streaming data (Kinesis, Spark streaming, Storm), batch data, automation, and scheduling (Oozie, Airflow).
- Good Experience in implementing and orchestrating data pipelines using Oozie and Airflow.
- Proven expertise of Spark Components such as Spark Core, Spark SQL, MLlib, GraphX, Data Frames, Datasets, Spark-ML, and Spark Streaming to create production-ready Spark applications.
- Experience in working with NoSQL databases like HBase and Cassandra.
- Proficiency in RDBMS like MySQL, PostgreSQL, Redshift, SQL Server, and Oracle.
- Experience in Data Analysis, Data Profiling, Data Integration, Data Migration, Data Governance, and Metadata Management, Master Data Management, and Configuration Management.
- Knowledge in modeling, tuning, disaster recovery, backup, and data pipeline creation.
- Developed scripts in Python (PySpark), Scala, and SparkSQL in Databricks for development, aggregation from various file formats such as XML, JSON, CSV, Avro, Parquet, and ORC.
- Used Amazon S3 to handle data transfer over SSL, and the data is immediately encrypted as it is uploaded.
- Ingested data into Snowflake Cloud Data Warehouse using Snowpipe.
- Extensive experience in working with micro batching to ingest millions of files on Snowflake cloud when files arrive at the staging area.
- Having good knowledge in writing MapReduce jobs through Pig, Hive, and Sqoop.
- Worked in developing Hive scripts for extraction, transformation, loading of data into Data warehouse.
- Expertise in writing end-to-end Data processing Jobs to analyze data using MapReduce, Spark, and Hive.
- Knowledge of tools like Tableau, Power BI, and Microsoft Excel for data analysis and generating data reports.
- Extensive experience in all phases of the Software Development Life Cycle (SDLC) from analysis, design, development, testing, implementation, and maintenance with timely delivery against deadlines.

- Extensively used Terraform in AWS Virtual Private Cloud to automatically set up and modify settings by interfacing with the control layer.
- Experience in designing interactive dashboards, reports, performing ad-hoc analysis and visualizations using Tableau, Power Bl, and Matplotlib.
- Sound experience in building production ETL pipelines between several source systems and Enterprise Data Warehouse by using Informatica PowerCenter, SSIS, SSAS, and SSRS.
- Used Azure AD, Sentry, and Ranger for maintaining security.
- Experience in the implementation of Continuous Integration(Cl), Continuous Delivery, and Continuous Deployment (CD) on various applications using Jenkins, Docker, and Kubernetes.
- Hands-on Experience in deploying Kubernetes Cluster on the cloud with master/minion architecture.
- Experienced in providing support on AWS Cloud infrastructure automation with multiple tools including Gradle, Chef, Docker and monitoring tools such as Splunk and CloudWatch.
- Worked in both Agile and Waterfall environments.
- Used Git, Bitbucket, and SVN version control systems.

TECHNICAL SKILLS

- **Big Data ecosystem:** HDFS, MapReduce, Spark, Yarn, Kafka, Hive, Airflow, Sqoop, HBase, Flume, Pig, Ambari, Oozie, Zookeeper, NiFi, Cassandra, Scala, Impala, Storm, Splunk, Tez, Flink, Stream Sets, Sentry, Ranger, Kibana.
- **Hadoop Distributions:** Apache Hadoop 2.x/1.x, Cloudera CDP, Hortonworks HDP
- **Cloud Platforms(AWS/Azure):** Amazon AWS - EMR, EC2, EBS, RDS, S3, Atana, Glue, Elasticsearch, Lambda, SQS, DynamoDB, Redshift, Kinesis Microsoft Azure - Databricks, Data Lake, Blob Storage, Azure Data Factory, SQL Database, SQL Data Warehouse, Cosmos DB, Active Directory GCP - BIG QUERY, DATAPROC,CLOUDSTROGE
- **Scripting Languages:** Python, Java, Scala, R, Shell Scripting, HiveQL, Pig Latin
- **NoSQL Database:** Cassandra, Redis, MongoDB, Neo4j
- **Database:** MySQL, Oracle, Teradata, MS SQL SERVER, PostgreSQL, DB2
- **ETL/Bl Tools:** Tableau, Power BI, Snowflake, Informatica, Talend, SSIS, SSRS, SSAS, ER Studio
- **Operating Systems:** Linux (Ubuntu, Centos, RedHat), Unix, Macintosh, Windows (XP/7/8/10/11)
- **Methodologies:** Agile/Scrum, Waterfall.
- **Version Control:** Git, SVN, Bitbucket
- **Others:** Docker, Kubernetes, Jenkins, Chef, Ansible, Jira, Machine learning, NLP, Spring Boot, Jupyter Notebook, Terraform.

PROFESSIONAL EXPERIENCE

Capital One, Plano, Texas **July 2024 to Present**
AWS Data Engineer
Responsibilities:

- Worked on creating pipelines and analytics using big data technologies such as Hadoop, Spark.
- Imported data from SQL Server to AWS Redshift and utilized Spark to execute transformations and actions to get the required outcome.
- Developed cloud-based serverless pipelines using AWS Lambda to export data from Hive to Redshift.
- Experience in the development of ETL data pipelines using Python, PySpark, Redshift, Amazon EMR, S3.
- Created Hive UDFs for custom analytical functions used to generate business reports.
- Installed/Configured/Maintained Apache Hadoop clusters for application development and Hadoop tools like Hive, Pig, HBase, Flume, Oozie, and Sqoop.
- Developed a data pipeline using Delta Lake that led to a client revenue increase
- Imported data from S3 Glacier to Hive using Spark on EMR Clusters.

- Used Crontab and shell scripting to import clickstream data from AWS S3 to Hive.
- Created a pipeline to ingest real-time(stream) data from Kinesis and store it in HDFS using Spark Streaming.
- Extensively used Lambda functions for pre-processing data ingested into S3 buckets.
- Created spark apps to stream data from Kinesis to HDFS, which integrates with Apache Hive to make data available for HQL querying instantly.
- Implemented real-time data streaming pipeline using AWS Kinesis, Lambda, andDynamo DB and deployed AWS Lambda code from Amazon S3 buckets.
- Created DDL and DML scripts in Hive and Redshift for generating tables and views, loading data.
- Developed ETL applications and used AWS Glue to run them.
- Worked on Big Data-Hadoop infrastructure for batch processing
- Programmed in Hive, Spark SQL, T-SQL and Python to streamline the incoming data and build the data pipelines to get the useful insights, and orchestrated pipelines.
- Used Amazon Elastic Kubernetes Service (Amazon EKS) to run and scale Kubernetes applications in the cloud or on-premises.
- Managing Databricks Notebooks, Delta Lake wif Python, Delta Lake.
- Design and develop custom, scalable, reusable, and resilient applications to integrate various components, increase consistency, automate tasks, alerts, assist in monitoring/diagnosing and processing of data in the Data Lake using Hadoop components in Hadoop cluster using SPARK
- Strong experience on Hadoop distributions like Cloudera and Hortonworks Platforms.
- Used Spark API over Hortonworks Hadoop YARN to perform analytics on data in Hive.
- Implemented Partitioning and Bucketing in HIVE for increasing performance benefit and helping in logically organizing data.
- Optimized the Glue workloads for different types of data loads by choosing the right compression, cluster type, instance type storage type to analyze data with low cost and high scalability.
- Used Data Frame API in Python for converting the distributed collection of data organized into named columns, developing predictive analytics using Apache SparkAPIs.
- Supporting continuous storage in AWS using Elastic Block Storage, S3 and created Volumes and configured Snapshots for
  EC2 instances.
- Worked on ETL migration services by developing and deploying AWS Lambda functions for generating serverless data pipeline which can be written to GlueCatalog and can be queried from Atana.
- Experience with Amazon EC2, S3, IAM users, groups, roles, VPC, Subnet, security groups, Network ACLs, Redshift, EMR, Atana, Glue, CloudWatch.
- Designed and Develop ETL Processes inAWSGlue, write Python Glue scripts to parse the Flat Files, CSV to required format either Cv or Parquet files and move the data from S3DataSources to Data warehouse.
- Responsible for building scalable distributed data solutions using the EMR cluster environment wif Amazon EMR.
- Created a serverless ETL on AWS Lambda to process new files in the S3 bucket and catalog them right away.
- Developed a Python module to access Jira and create issues for all database owners, notifying them every seven days if the issue isn't resolved.
- Used AWS EMR to transform and move large amounts of data into and out of otherAWSdatastores and databases, such as Amazon Simple Storage Service (Amazon S3) and Amazon DynamoDB
- Involved in database design and data modeling for OLTP and OLAP databases, utilizing Entity-Relationship modeling and dimension modeling.
- Monitor, maintain and support the new instances of Data Lake on AWS.
- AWS SQS was used to transfer the processed data to the next working teams for processing.
- Integrated data from MySQL to the Client Portal using AWS Lambda services.
- HandledAWSManagement Tools as Cloud watch and Cloud Trail.
- Administered and configured the ELK Stack (Elasticsearch, Log Stash, Kibana) on AWS, performed log analysis.

- Manage code repositories using Git to ensure the integrity of the codebase is maintained at all times.
- Experience developing Airflow workflows for scheduling and orchestrating the ETL process.
- Ensured system architecture met business requirements, constantly worked with different teams to ensure every aspect of architecture is beneficial to the company.
- Connected Redshift to Tableau for creating a dynamic dashboard for the analytics team.

Environment: Hadoop, Spark, Spark SQL, Hive, MySQL, Apache Spark Streaming, Kafka, AWS RDS, AWS Lambda, AWSEC2, AWS S3, AWS Redshift, EMR, Glue, Atana, Kinesis, AWS Data Pipeline, Jira, Shell Scripting, Crontab

Verizon, Tampa, Florida **December 2021 to June 2024**
Azure Data Engineer
Responsibilities:

- Designed and developed Hadoop-based Big Data analytic solutions and engaged clients in technical discussions.
- Worked on multiple Azure platforms like Azure Data Factory, Azure Data Lake, Azure SQL Database, Azure SQL Data Warehouse, Azure Analysis Services, HDInsight.
- Worked on the creation and implementation of custom Hadoop applications in the Azure environment.
- Created ADF Pipelines to load data from an on-prem to Azure SQL Server database and Azure Data Lake storage.
- Developed complicated Hive queries to extract data from various sources (Data Lake) and to store it in HDFS.
- Used Azure Data Lake Analytics, HDInsight/Databricks to generate Ad Hoc analysis.
- Developed custom ETL solutions, batch processing, and real-time data ingestion pipeline to move data in and out of Hadoop using PySpark and shell scripting.
- Implemented large Lambda architectures using Azure Data platform capabilities like Azure Data Lake, Azure Data Factory, Azure Data Catalog, HD Insight, Azure SQL Server, Azure ML and Power BI.
- Data Ingestion to at least one Azure Services - (Azure Data Lake, Azure Storage, Azure SQL, Azure DW) and processing the data in In Azure Databricks.
- Worked on all aspects of data mining, data collection, data cleaning, model development, data validation, and data visualization.
- Experienced in managingAzureDataLake Storage (ADLS), Databricks Delta Lake and an understanding of how to integrate with otherAzureServices.
- Worked on building data pipelines using Azure Data Factory, Azure Databricks, loading data to Azure Data Lake.
- Handled bringing in enterprise data from different data sources into HDFS usingSqoop and performing transformations using Hive, Map Reduce, and then loading data into HBase tables.
- Responsible for estimating the cluster size, monitoring, and troubleshooting of the Hadoop cluster.
- Used Zeppelin, Jupyter notebooks, and Spark-Shell to develop, test, and analyzeSpark jobs before Scheduling Customized Spark jobs.
- Worked with Azure BLOB and Data lake storage and loading data into Azure SQL Synapse analytics(DW).
- Performing hive tuning techniques like partitioning, bucketing, and memory optimization.
- Developed Spark applications using PySpark and Spark-SQL for data extraction, transformation, aggregation from various file formats for analyzing & transforming the data to uncover insights into customer usage patterns.
- Analyze, design, and build Modern data solutions using Azure PaaS service to support visualization of data.
- Using Data bricks utilities called widgets to pass parameters on run time fromADF to Data bricks. data in InAzure Databricks.
- Design Data Lake storage solution for Data science Project using Azure Data factory Pipelines.

- Integrateddatastorage options wif Spark, notably wifAzureDataLake Storage and Blob storage.
- Hands-on experience on creating Spark cluster in both HDInsight's and Azure Databricks environment.
- Created an Oozie workflow to automate the process of loading data into HDFS and Hive.
- Created tables using NoSQL databases like Base to load massive volumes of semi-structured data from sources.
- Created, provisioned numerous Databricks clusters needed for batch and continuous streaming data processing and installed the required libraries for the clusters.
- Worked on creating tabular models on Azure analysis services for meeting business reporting requirements.
- Developed SSIS modules to move data from a variety of sources like MS Excel, Flat files, and CSV files.
- Designed, developed, and deployed Business Intelligence solutions using SSIS,SSRS, and SSAS.
- Implemented a variety of MapReduce tasks in Scala for data cleansing and data analysis in Impala.
- Fetched live stream data using Spark Streaming and Kinesis.
- Imported and exported the data using Sqoop from HDFS to Relational Database systems and vice-versa and loaded into Hive tables, which are partitioned.

Environment: Azure Data Factory, Azure Databricks, Azure Data Lake, Blob Storage, HDFS, MapReduce, Spark, SQL, Hive, HBase, HDInsight, Kafka, Oozie, NiFi, Jenkins, OLAP, OLTP, Scala, SSIS, Agile.

IBM, India **January 2019 to November 2020**
Big Data Engineer
Responsibilities:

- Involved in the requirement collecting phase to collect needs from business users to accommodate changing user requirements constantly.
- Created a Data Quality Framework for Spark that does schema validation and data profiling (PySpark).
- Developed Spark code for quicker data testing and processing utilizing Scala and Spark-SQL/Streaming.
- Used Python and Scala with Spark to design data and ETL pipeline.
- Developed very complicated Python and Scala scripts that are sustainable, easy to use, and meet application requirements, data processing, and analytics through the usage of built-in libraries.
- Contributed to the design of Spark SQL queries, Data frames, data import from data sources, transformations, read/write operations, and saving the results to an output directory in HDFS/AWS S3
- Created Pig Latin scripts to import data from web server output files and to store it in HDFS.
- Created Tableau tools to help internal and external teams see and extract information from big data platforms.
- Responsible for conducting Hive queries and running Pig scripts on raw data to analyze and clean it.
- Created Hive tables, imported data, and wrote Hive queries.
- Worked on ETL (Extract, Transform, Load) processing, which includes data source, data transformation, mapping, conversion, and loading.
- Used multiple compression algorithms to optimize MapReduce jobs to make the most of HDFS.
- Worked with AWS Elastic Cloud Compute (EC2) infrastructure for computational operation, while Simple Storage Service (S3) was used as a storage method.
- Experience in building and architecting multiple Data pipelines, end to end ETL and ELT process for Data ingestion and transformation in GCP.
- Configured AWS CLI and performed necessary actions on the AWS services using scripting.

- Able to execute and monitor Hadoop and Spark tasks on AWS using EMR, S3, and Cloud Watch services.
- Configured the monitoring and alerting of production and corporate servers/storage usingCloud Watch.
- Developed Docker containers by merging them with workflow to make them lighter.
- Developed several MapReduce programs to extract, transform, and aggregate data from a variety of file formats including XML, JSON, CSV, and other compressed file formats.
- Migrated existing data from Teradata and SQL Server to Hadoop and performed ETL operations on it.
- Good knowledge in querying data from Cassandra for searching, grouping and sorting.
- Optimizing existing algorithms in Hadoop using Spark Context, Spark-SQL, DataFrames, and Pair RDD's.
- Good knowledge in using Apache NiFi to automate the data movement between different Hadoop systems.
- Created AWS Lambda, EC2 instances provisioning on AWS environment and implemented security groups, administered Amazon VPC's
- Developed ETL Pipelines in and out of data warehouse, develop major regulatory and financial reports using advanced SQL queries.
- Experience in using theAWSservices Atana, Redshift and Glue ETL jobs.
- Experience in using Terraform to create Infrastructure as Code onAWS.
- Scripting experience in PySpark, which involves cleansing and transformation of data.
- Used the AWS Kinesis to gather and load data onto HDFS, using Sqoop to load data from relational databases.
- Developed job processing scripts using Oozie workflow to automate data loading into HDFS.
- Developed SQL queries for both dimensional and relational data warehouses and performed data analysis.
- Good experience wif use-case development, wif Software methodologies likeAgile.

Environment: HDFS, Spark, Spark SQL, PySpark, Scala, Python, AWS S3, EC2, CLI, EMR, Cloud Watch, Docker, Data Frames, Pair RDD's, NiFi, SQL, Pig Latin, Hive, Tableau, MapReduce.

Volvo Group India Private Limited, Bangalore, India **August 2016 to December 2018**
Data Engineer
Responsibilities:

- Performed advanced ETL development activities using Informatica, PL/SQL, and Oracle.
- Responsible for replicating mappings, mapplets, sessions, and workflows to load the data from source to target database from Informatica Power Center.
- Hands on experience in GCP, Big Query, GCS bucket, * - cloud function, cloud data flow, Pub/sub cloud shell, GSUTIL, BQ command line utilities, Data Proc, Stack driver
- Provided data analysis to support new project definition/design as well as to troubleshoot issues.
- Created, updated, and maintained project documents including business requirements, functional and non-functional requirements, functional and technical design, data mapping, etc.
- Analyzed and evaluated required data sources.
- Data was extracted, transformed, and loaded into the data warehouse successfully.
- Hands on experience in migrating on premise ELs to Google Cloud Platform (GCP) using cloud native tools such as BIG query, Cloud Data Proc, Google Cloud Storage, Composer
- Used cloud shell SDK in GCP to configure the services Data Proc, Storage, BigQueryer specified functions, stored procedures, indexes, and triggers were developed.
- Participated in the creation and implementation of development, testing, and deployment standards.
- Effectively increased the performance of long-running views and stored procedures.
- Created views, complex queries and mapped them through ETL job implementations.
- Involved in ETL process development to test and production environments

- Data was extracted from a variety of sources, including flat files and Oracle, and uploaded into target systems.
- Executed Workflows and Sessions using Workflow Monitor.
- Created test plans, test cases, test scenarios, test strategies, defect management to ensure quality assurance to test all the business requirements.

Environment: Informatica power center, SQL, Oracle SQL Developer, UNIX, Putty, JIRA, Microsoft Office, MySQL.GCP