Rakesh Kanparthi

Data Engineer-Big Data,Azure

[Rakeshhrao4@gmail.com](mailto:Rakeshhrao4@gmail.com)

7042866995

Professional Summary:

* Having 5+ years of data engineering with elite experience in Big Data Technologies, ETL, HadoopEcosystem, MySQL,AWS, Map Reduce,python,pyspark,kafka, RDBMS, business reporting,Azure and Aws
* Hands on experience on Hadoop ecosystem components like HDFS, Hive, YARN, Sqoop, Hbase, impala, oozie, zookeeper, Python, kafka, pyspark.
* In depth understanding of Hadoop Architecture including YARN and various components such as HDFS, Resource Manager, Node Manager, Name Node, Data Node.
* In - depth understanding of Spark Architecture including spark core, spark sql, data frames, spark streaming, spark Miliband Spark Real time Streaming.
* Hands on experience in Analysis, Design, Coding and Testing phases of Software Development Life Cycle (SDLC).
* Proficient with Apache Spark ecosystem such as Spark, Spark Streaming using Scala and Python.
* Hands on experience with AWS (Amazon Web Services), Elastic Map Reduce (EMR), Storage S3, EC2 instances, lambda, kinesis.
* Experience with ETL tools such as Airbyte.
* Worked comprehensively with Apache Sqoop and developed Sqoop scripts to interface data from a MySQL database into the Hadoop Distributed File System (HDFS)
* Hands on experience in various Bigdata application phases like data ingestion, data analytics and data visualization.
* Experience in usage of Hadoop distribution like Cloudera, Hortonworks distribution & Amazon AWS.
* Experience in transferring data from RDBMS to HDFS and HIVE table using pyspark.
* Migrating the coding from Hive to Apache Spark and Scala using Spark SQL, RDD.
* Understanding of flume architecture to load the data from multiple sources directly into HDFS.
* Good knowledge on impala, sparksql, storm, avro, kafka, hue and aws and knowledge on ide tools such as eclipse, netbeans, and maven.
* Created new ETL packages to validate, extract, transform and load data into data warehouse databases.
* Strong knowledge of version control systems like SVN and GITHUB.
* Worked with Tableau to connect with Hive Data Warehouse to represent the data in dashboards
* Extract Transform and Load data from Sources Systems to Azure Data Storage services using a combination of Azure Data Factory, T-SQL, Spark SQL and U-SQL Azure Data Lake Analytics.
* 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.
* Experience in implementing data solutions in Azure including Azure SQL, Azure Synapse, Cosmos DB, Databricks, ADLS, Blob Storage, ADF and Azure Stream Analytics.
* Experience on building data pipelines using Apache Spark (Azure Databricks), Scala.
* Worked on Microsoft azure services like HDInsight Clusters, BLOB, ADLS, Data Factory and Logic Apps and also done POC on Azure Databricks
* Involved in developing code to write canonical model JSON records from numerous input sources to Kafka Queues.
* Created data pipeline using Kafka, HBase, Spark and Hive to ingest, transform and analyze customer behavior.
* Implemented POCs using Kafka, Strom, HBase for processing streaming data.
* Implemented Kafka Custom encoders for custom input format to load data into Kafka Partitions.
* Implemented Kafka High level consumers to get data from Kafka partitions and move into HDFS.
* Developed Kafka Producers and Consumers for data streaming.
* Expertise in integrating Kafka with Spark streaming for high-speed data processing.
* Extensive experience in using Microsoft Business Intelligence studio products like SSIS, SSAS, SSRS for implementation of ETL methodology in data extraction, transformation and loading.
* Good understanding of Spark -Streaming using kafka. Experience in writing kafka consumers.
* Basic Knowledge on nifi, kylin and zeppelin with Apache Spark.
* Experience in NoSQL Column-Oriented Databases like Hbase, Cassandra and its Integration with Hadoop cluster.
* Involved in Cluster coordination services through Zookeeper.
* Good level of experience in Core Java, J2EE technologies as JDBC, Servlets, and JSP.
* Worked extensively with sqoop importing data from MYSQL and storing in HDFS.
* Experienced in extract transform and load (ETL) processing large datasets of different forms including structured, semi-structured and unstructured data.
* Experience in designing and developing applications in spark using Scala to compare the performance of spark with Hive and sql/oracle.
* Strong experience handling in huge volumes of data, SQL and ETL,batch processing.
* Good Business knowledge on Finance and banking Sector domain.
* Experience in NoSQL database HBase, MongoDB and Cassandra
* Ability to plan, manage, motivate, and work efficiently as an independent or collaboratively in a team.
* Self-motivated, enthusiastic, and always keen to learn new methodologies and techniques

Technical Skills:

|  |  |
| --- | --- |
| Databases | Oracle, Mysql, Hive, SQL Server |
| BigData Technologies | HDFS, Hive, pySpark, NiFi, Map Reduce, Pig, YARN, Sqoop, Oozie, Zookeeper, and Flume. |
| Programming Languages | Python, Java, SQL, PL/SQL, Scala, json, xml. |
| Cloud Computing | AWS EC2, S3, RDS, DynamoDB, SNS, SQS, kinesis, step functions, lambda functions and **Azure SQL Database, Azure Data Lake, Azure Data Factory (ADF).** |
| Techniques | Datamining, Clustering, Data Visualization, Data Analytics |
| NoSQL Databases | Hbase, Cassandra |
| Container Platforms | Docker, Kubernetes, CD/CI, Jenkins |

EDUCATION:

Bachelor of Technology in Computers Science, May, 2014

Jawaharlal Nehru Technology University, India

Masters in Computer's software Engineering, May, 2017

Oklahoma Christian University, Oklahoma

Professional Experience:

Role: Big Data engineer,

Wellsfargo

Charlotte, NC August 2021 – Present

* Analysing and troubleshooting the JIRA-tickets, follow up with the stake holders and giving resolutions on priority basis.
* Involved in the technical designing of the project based on the requirements.
* Apache Spark SQL is used for comparing the data from different sources.
* Code is implemented using Scala and Spark to get the best performance and make the code more readable.
* Developed the validation code for Hive data quality rules by using Spark/Scala.
* Extensively worked on the Hive database to create the tables and develop the data validation scripts and automated the process.
* Good experience of AWS Elastic Block Storage (EBS), different volume types and use of various types of EBS volumes based on requirement.
* Worked and learned a great deal from AmazonWebServices (AWS) Cloud services like EC2, S3, EBS, RDS and VPC.
* Implemented AWS provides a variety of computing and networking services to meet the needs of applications

Big Data engineer

Bank of America

Charlotte, NC February 2021 – August 2021

Responsibilities:

* Hands on experience in handling Hive tables using Spark SQL.
* Efficient in writing MapReduce Programs and using Apache Hadoop API for analyzing the structured and unstructured data.
* Expert in working with Hive data warehouse tool-creating tables, data distribution by implementing partitioning and bucketing, writing and optimizing the HiveQL queries.
* Debugging Pig and Hive scripts and optimizing MapReduce job and debugging Map reduce job.
* Hands-on experience in managing and reviewing Hadoop logs.
* Good knowledge about YARN configuration.
* Imported and exported data from different Relational Database Systems like Mysql and Oracle into HDFS and Hive and vice-versa, using Sqoop
* Worked on Mapping the source table with the target table.
* Used Data Lineage tool to find the usage of the columns of source tables in the downstream so, that no downstream get impacted with this migration.
* Extensively worked on Data validation between Hive source tables and target tables using automation Python Scripts.
* Worked on converting the Data Ingestion pipelines from the Hive to Presto to improve the performance of the Pipeline Created by the Data Analyst for the Business.
* Performed job functions using Spark API's in Scala for real time analysis and for fast querying purposes.
* Hands on experience in AWS Cloud in various AWS services such as Redshift cluster, Route 53 domain configuration.
* On demand secure EMR launcher with custom spark submit steps using S3 Event, SNS, KMS and Lambda function.
* Extensive knowledge of working on NiFi.
* Migrated an existing on-premises application to AWS.
* Used AWS services like EC2 and S3 for small data sets.
* Experienced in writing complex SQL Queries, Stored Procedures, Triggers, Views, Cursors, Joins, Constraints, DDL, DML and User Defined Functions to implement the business logic.
* Good experience in developing Hive DDLs to create, alter and drop Hive TABLES.
* Designed and published visually rich and intuitive Tableau dashboards and Crystal Reports for executive decision making.
* Monitored System health and logs and responded accordingly to any warning or failure conditions.

Cerner,

Kansas City, MO

Data Engineer – Spark, Scala,Azure Jan2020 - December2020

Responsibilities:

* Involved in analyzing the system and business.
* Involved in importing data from Oracle to HDFS using OOZIE & PYSPARK.
* Written Hive complex Structure fields and return complex data type.
* Developed the different ingestion pipeline components using Spark and Python.
* Writing spark sql scripts to load and process data into Hadoop File System.
* Design and created common templates for ETL processing jobs.
* Worked on analyzing Hadoop stack & different Big Data analytics tools like Pig, Hive, HBase database & Sqoop.
* Excel Spreadsheets, Text file, MS Access and CSV files to data warehouse.
* Used MDX to interact with the back end OLAP Services
* Created new ETL packages to validate, extract, transform and load data into data warehouse databases, data mart databases using SSIS packages.
* Designed and Developed ETL pipelines that will read data from various sources like CRM, ADP (HR and Pay role) and JDE(Operational) and loaded into SQL Datawarehouse using Stored Procedures, Triggers and Other Database Objects.
* Experienced working on larger datasets.
* Worked with configuring checkpoints, package logging, error logging and event handling to redirect error rows and fix the errors in SSIS.
* Primarily involved in Data Migration using SQL, SQL Azure, Azure storage, and Azure Data Factory, SSIS, PowerShell.
* Load data into Amazon Redshift and use AWS Cloud Watch to collect and monitor AWS RDS instances within Confidential.
* Coordinated with the BI team to visualize the transformed data into a dashboard using Tableau
* Designed and developed ETL/ELT processes to handle data migration from multiple business units and sources including Oracle, Postgres, Informix, MSSQL, Access and others.
* Developed and executed a migration strategy to move Data Warehouse from an Oracle platform to AWS Redshift.
* Used BI Tools such as ThoughtSpot and SAP Tools to create and maintain Reports
* SAP Data Services Integrator ETL developer with strong ability to write procedures to ETL data into a Data Warehouse from a variety of data sources including flat files and database links (Postgres, MySQL, Oracle).
* Design and Develop ETL Processes in ADF to migrate Campaign data from external sources like Azure blob Storage, Azure Data Lake and Using ORC/Parquet/Text Files into Azure Datawarehouse (Azure synapse).
* Design, Develop, test, and deploy ELT/ETL pipelines using Azure Data Factory.
* Implemented CDC type2 to handle incremental data updates in HDFS.
* Created Spark applications for data preprocessing for greater performance.
* Involved in migrating oozie workflow code to spark flows for greater performance.
* Developed Spark code and Spark-SQL for faster testing and processing of data.
* Involved in converting Hive or SQL queries into Spark transformations using Python and Scala.
* Design and Develop ETL Processes in ADF to migrate Campaign data from external sources like Azure blob Storage, Azure Data Lake and Using ORC/Parquet/Text Files into Azure Datawarehouse (Azure synapse).
* Design, Develop, test, and deploy ELT/ETL pipelines using Azure Data Factory.
* Great Understanding of AWS compute services such as EC2, Elastic MapReduce (EMR), EBS and accessing Instance metadata.
* Worked extensively on PYSPARK to build Big Data flow.

Bank of America,

Charlotte, NC August 18 – December19

Data Engineer – Spark, Scala, Hive

Responsibilities:

* Worked with internal and external data sources on improving data accuracy / coverage and generate recommendation on the process flow to accomplish the goal.
* Worked on json formats and converted all json files to hive tables using spark/java.
* Involved in spark performance tuning.
* Implemented Partitioning, Dynamic Partitions and Buckets in HIVE for increasing performance benefit and helping in organizing data in a logical fashion.
* Written Hive queries for data analysis to meet the business requirements. Designed and developed User Defined Function (UDF) for Hive.
* Developed spark applications in python(PySpark) on distributed environment to load huge number of CSV files with different schema in to Hive ORC tables.
* Worked on NoSQL databases like HBase and imported data from Oracle and processed data using Hadoop Tools and exported to Cassandra NoSQL database.
* Involved in creating Hive tables (Managed tables and External tables), loading, and analyzing data using hive queries.
* Good knowledge about the configuration management tools like SVN/CVS/GitHub.
* Exposure on Spark Architecture and how RDD’s work internally by involving and processing the data from Local files, HDFS and RDBMS sources by creating RDD and optimizing for performance.
* Working on Airflow which is a job trigger where the code is written in hive query language and Scala. This helps to read, back fill, and write the data for particular time Frame from hive tables to HDFS locations.
* Involved in converting Hive/SQL queries into Spark transformations using Spark RDD, Scala, and Python.
* Configured Spark streaming to receive real time data from the Apache KAFKA and store the stream data to the HDFS using Scala.
* Developed framework for converting existing PowerCenter mappings and to PySpark(Python and Spark) Jobs.
* Involved in converting MapReduce programs into Spark transformations using Spark RDD in Scala
* Worked towards creating real time data streaming solutions using Apache Spark / Spark Streaming, Kafka.
* Good Knowledge about scalable, secure cloud architecture based on Amazon Web Services (leveraging AWS cloud services: EC2, Cloud Formation, VPC, S3, etc.
* Collected the logs data from web servers and integrated in to HDFS using Flume.
* Analyzed the data by performing Hive queries (HiveQL) and running Pig Scripts (Pig Latin).
* Expert in creating the SFTP Connection to the internal and external source to get data in secured manner without any breakage.
* Handle the production Incidents assigned to our workgroup promptly and fix the bugs or route it to the respective teams and optimized the SLA’s.

TCS,

India , May 2014 – May 2015

Software Developer,ETL and Hadoop

Responsibilities:

* Responsible for building scalable distributed data solutions using Hadoop
* Worked comprehensively with Apache Sqoop and developed Sqoop scripts to interface data from a MySQL database into the Hadoop Distributed File System (HDFS)
* Strong core Scala, including experience with collections, type variance, implicit parameters and conversion.
* Functional programming experience in Scala, including higher-order functions, partial functions, partial application, nested functions.
* Responsible for data extraction and data ingestion from different data sources into Hadoop Data Lake by creating ETL pipelines using Pig, and Hive
* Professional experience coding in Scala in a production environment.
* Worked on debugging and performance tuning of Hive & Pig Jobs.
* Developed technical specifications of the ETL process flow.
* Experienced in creating data pipeline integrating kafka with spark streaming application used Scala for writing applications.
* Used python sub-process module to perform UNIX shell commands.
* Used sparksql for reading data from external sources and processes then the data using Scala computation framework.
* Involved in unit testing and documentation of the ETL process.
* Increased ETL performance by 55% by implementing best ETL practices after analyzing inherent operational pattern.
* Acted as point of contact between product users, onsite team, dev team and SMEs resulting in successful communication.
* Improved project performance by 70% by seamlessly auditing and moving ETL jobs from DEV to QA environment.
* Written shell scripts for extracting the data to perform analytics on top of it.
* Integrated Oozie with the rest of the Hadoop stack supporting several types of Hadoop jobs (such as Map-Reduce, Pig, Hive, and Sqoop) as well as system specific jobs (such as Java programs and shell scripts)
* Used pig and hive Upon the Hcatalog tables to analyze the data and create schema for the Hbase tables in Hive.
* Responsible for managing and scheduling jobs on Hadoop Cluster.