LASYA RAO

Senior Big Data Engineer | +1 (281) 673-6089 | joginpallylasya@gmail.com

Professional Summary

- Around 9 years of solid experience as a data engineer, with practical knowledge of the Cycles of Analysis, Design, Development, and Testing in both Waterfall and Agile approaches.
- Proven proficiency with key Big Data Hadoop ecosystem components, including HDFS, YARN, MapReduce, Hive, Impala, Pig, Sqoop, HBase, Spark, Spark SQL, Kafka, Spark Streaming, Oozie, and Hue.
- Proficient knowledge of the Amazon Web Services (AWS) Cloud Platform, which includes EC2, EMR, S3, VPC, IAM, DynamoDB, Cloud Front, Cloud Watch, Route 53, Auto Scaling, and Security Groups.
- Knowledge of Microsoft Azure/Cloud Services, such as Azure SQL Server, Azure Databricks, Azure Data Lake, Azure Blob Storage, Azure Data Factory, and Azure Synapse Analytics.
- Solid Spark (spark streaming, spark SQL) working knowledge with Scala and Kafka. Used Scala to work on reading various data formats from HDFS.
- Experience with the Snowflake and Databricks environments.
- Experience with the Lake House Architecture driven by Databricks (Unity Catalog, DLT, DB-SQL, Databricks workflows) platform.
- To execute analytics on Hive data using the Cloudera platform, Spark Data Frames API was used, along with Spark Data Frame Operations to carry out the necessary data validations.
- Developed unique Kafka producers and consumers for various Kafka topics for publishing and subscribing.
- Knowledge of using Kafka and Kafka brokers to use Spark Structured Streaming to ingest data.
- Using Spark Data frames and Scala, I was involved in transforming Hive/SQL queries into Spark transformations.
- Worked on optimizing the performance of the snowflake and spark applications.
- Strong working knowledge of NoSQL databases including MongoDB, HBase, and DynamoDB.
- Worked with various formats of files like delimited text files, click stream log files, Apache log files, Avro files, JSON files, XML Files.
- While building the ETL pipelines, I spent a lot of time working with Snowflake, Star schema, and Business Vault modeling tools.
- Mastered in using different columnar file formats like RC, ORC and Parquet formats and has good understanding of various compression techniques used in Hadoop processing like G-zip, Zstd, Snappy, LZO etc.
- Knowledge of utilizing Sqoop to import and export data from f3 to relational database systems and vice versa, as well as to load data into partitioned Hive tables. Knowledge of ETL methods for data extraction, transformation and loading in corporate-wide ETL Solutions and Data Warehouse tools for reporting and data analysis.
- Proven ability to extract files from MongoDB using Sqoop, store them in HDFS, and process them.
- Vast expertise creating database scripts in SQL and PL/SQL and using a variety of databases.
- Possess extremely strong interpersonal skills, the capacity to work both independently and collaboratively, and the capacity to pick up new information quickly and effortlessly.

Skills

- **Big data Eco System:**

Hadoop, MapReduce, Spark, HDFS, Sqoop, YARN, Oozie, Hive, Impala, Apache Airflow, HBase.

- **Programming Languages:**

PL/SQL, SQL, Python, Scala, PySpark, Java

- **Database Management:**
- Snowflake, Teradata, Redshift, MySQL, SQL Server, Oracle.
- **NoSQL Databases:**

DynamoDB, Cassandra, HBase

- **Version Control:**

GIT, TFS, Azure DevOps

- **Workflow mgmt. tools:**

Apache Airflow, Oozie, Autosys

- **Visualization & ETL tools:**

Tableau, PowerBI, Informatica, Talend.

- **Methodologies:**

Agile, Scrum, Waterfall Model

- **Cloud Technologies:**

Azure & AWS

Work History

Senior Big Data Engineer **August 2021 - Current**

End Client – UBS -Weehawken, NJ

- Worked with AWS cloud services such as EC2, EMR, S3, RDS, auto scaling groups, IAM to build configuration to build analytical solutions.
- Hands on experience working with Lake House and Databricks clusters.
- Implemented merge logic on Delta tables within Data bricks.
- Worked on consuming data from Kafka topics and used spark structured streaming to load data into Delta Lake tables and Kafka topics.
- Worked with snowflake cloud data warehouse database for ingesting the data from s3 to snowflake.
- Implemented Row level security on snowflake objects (Row level security).
- Developed python-based Spark applications for performing data cleansing, event enrichment, data aggregation, de-normalization and data preparation needed for machine learning and reporting teams to consume.
- Worked extensively on performance tuning of spark applications.
- Streamed real time data by integrating Kafka with Spark for dynamic price surging using machine learning algorithm.
- Used python, the ETL pipeline was developed and programmed to collect data from Redshift data warehouse.
- Experience on processing large volumes of data from various sources
- Experience on Hadoop distribution including storing and managing large datasets. Worked on designing and implementing data storage solutions using technologies such as HDFS, HBase, and Kudu.
- Experience on Cloudera's Hadoop distribution which includes tools for transforming and analyzing large datasets.

- Experience on Cloudera machine learning tools, such as Cloudera Data Science Workbench and Cloudera Fast Forward Labs, used to design and implement machine learning solutions.
- Experience on Cloudera security features, including encryption, authentication, and authorization tools. Configured and managed these security features to ensure the security of data and systems.
- Developed python scripts for ingesting FTP servers' data as well as Sqoop jobs for ingesting to store it on data warehouse databases.
- Managed structured data that is designed to scale to a very large size across many commodity servers, with no single point of failure by using Cassandra.
- Working on Docker Hub, Docker Swarm, Docker Container network, creating Image files primarily for middleware installations & domain configurations. Evaluated Kubernetes for Docker Container Orchestration.
- Installed Docker Registry for local upload and download of Docker images and from Docker Hub and created Docker files to automate the process of capturing and using the images.
- Experience in integrating Jenkins with various tools like Maven (Build tool), Git (Repository), SonarQube (code verification), Nexus (Artifactory) and implementing CI/CD automation for creating Jenkins pipelines programmatically architecting Jenkins Clusters, and scheduled builds day and overnight to support development needs.
- Involved in Trouble Shooting, Performance tuning of reports and resolving issues within Tableau Server and Reports.
- Programmatically created CICD Pipelines in Jenkins using Groovy scripts, Jenkins file, integrating a variety of Enterprise tools and Testing Frameworks into Jenkins for fully automated pipelines to move code from Dev Workstations to all the way to Prod environment.
- Associated with composing Python scripts to computerize the way towards extricating weblogs utilizing Airflow DAGs.

Environment: Kafka, HBase, Docker, Kubernetes, AWS, EC2, S3, Lambda, Cloud Watch, Auto Scaling, EMR, Redshift, Jenkins, ETL, Spark, Hive, Athena, Sqoop, Pig, Oozie, Spark Streaming, Hue, Scala, Python, Databricks, GIT, Micro Services, Snowflake.

Big Data Engineer **September 2018- August 2021**

End Client – First Republic Bank, San Francisco, California

- Designed solutions to process high volume data stream ingestion, processing and low latency data provisioning using Hadoop Ecosystems Hive, Pig, Scoop and Kafka, Python, Spark, Scala, HBase and Druid.
- Strong understanding of AWS components such as EC2, S3, Lambda, Auto Scaling, Cloud Watch, Cloud Formation, Security groups and IAM.
- Used Hive SQL, Presto SQL, and Spark SQL for ETL jobs and using the right technology for the job to get done.
- Created Entity Relationship Diagrams (ERD), Functional diagrams, Data flow diagrams and enforced referential integrity constraints and created logical and physical models using Erwin.
- Created ad hoc queries and reports to support business decisions SQL Server Reporting Services (SSRS).
- Defined facts, dimensions and designed the data marts using the Ralph Kimball's Dimensional Data Mart modelling methodology using Erwin.

- Worked publishing interactive data visualizations dashboards, reports /workbooks on Tableau and SAS Visual Analytics.
- Experience in IBM DataStage system engineering on LINUX platform.
- Advanced knowledge on Confidential Redshift and MPP database concepts.
- Designing and building multi-terabyte, full end-to-end Data Warehouse infrastructure from the ground up on Confidential Redshift for large scale data handling Millions of records every day.
- Designed and implemented big data ingestion pipelines to ingest multi-TB data from various data source using Kafka, Spark streaming including data quality checks, transformation, and stored as efficient storage formats Performing data wrangling on multi-Terabyte datasets from various data sources for a variety of downstream purposes such as analytics using Spark.
- Optimizing and tuning the Redshift environment, enabling queries to perform up to 100x faster for Tableau and SAS Visual Analytics
- Develop Data warehousing systems by using Informatica tools.
- Worked hands on with ETL process using Informatica to load data into oracle database.
- Handled importing of data from various data sources, performed transformations using Hive, MapReduce, and loaded data into HDFS.
- Migrated on premise database structure to Confidential Redshift data warehouse.
- Was responsible for ETL and data validation using SQL Server Integration Services.
- Exception handling in python to add logs to the application.
- Analyzed the system for new enhancements/functionalities and perform Impact analysis of the application for implementing ETL changes.
- Developed SSRS reports, SSIS packages to Extract, Transform and Load data from various source systems.
- Compiled data from various sources to perform complex analysis for actionable results.
- Built performant, scalable ETL processes to load, cleanse and validate data.
- Analyse the existing application programs and tune SQL queries using execution plan, query analyser, SQL Profiler, and database engine tuning advisor to enhance performance.
- Implementing and Managing ETL solutions and automating operational processes.
- Wrote various data normalization jobs for new data ingested into Redshift.
- Collaborate with team members and stakeholders in design and development of data environment.
- Preparing associated documentation for specifications, requirements, and testing
- Optimized the TensorFlow Model for efficiency.
- Participated in the full software development lifecycle with requirements, solution design, development, QA implementation, and product support using Scrum and other Agile methodologies.

Environment: Oracle, Kafka, Python, Terraform, Redshift, Informatica, AWS, EC2, S3, SQL Server, Erwin, RDS, NOSQL, Snowflake Schema, MySQL, Dynamo DB, PostgreSQL, Tableau, Git Hub.

Big Data Engineer **April 2017 - September 2018**

Nike – Beaverton, Oregon

- Experience in Azure Data factory which includes design, development, and implementation.

- Experienced on Migrating SQL database to Azure data lake, Azure SQL Database, Databricks and Azure Data Factory.
- Data ingestion to one or more azure data services like azure data factory, data bricks
- Data extraction from various sources, Transformation and Loading into the target SQL Server Database. Implemented Copy activity, Custom Azure Data Factory Pipeline Activities for On-cloud ETL processing.
- Worked 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.
- Automize the Power BI reports, dashboards and Azure Data Factory (ADF) pipelines when source data updated.
- Created Pipelines and Load the data in Azure SQL Datawarehouse through Data Lake and ADF activities. 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.
- Experienced in the progress of real time streaming analytics data pipeline. Confidence in building connections between event hub and streaming analytics.
- Interacted with multiple teams who are responsible for Azure Platform to fix the Azure Platform Bugs.
- Providing 24/7 support for on-call on Azure configuration and Performance issues.
- Good hands-on Azure Data Factory, worked on creating dependencies of activities in Azure Data Factory.
- Transforming data in Azure Data Factory with the ADF transformations and creating pipelines with GUI.
- Developed and maintained end-to-end operations of ETL data pipelines and worked with large data sets in ADF.
- Developed data pipeline using Sqoop to ingest customer behavioral data and purchase histories into HDFS for analysis.
- Delivered de normalized data for Power BI consumers for modeling and visualization from the produced layer in Data Lake
- Tested Apache TEZ, an extensible framework for building high performance batch and interactive data processing applications, on Pig and Hive jobs.
- Exposed transformed data in Azure Spark Databricks platform to parquet formats for efficient data storage.
- Enhanced and optimized product Spark code to aggregate, group and run data mining tasks using the Spark framework.
- Worked on product positioning and messaging that differentiate Hortonworks in the open-source space.
- Involved in importing the real time data to Hadoop using Kafka and implemented the Oozie job for daily imports.
- Involved in complete big data flow of the application starting from data ingestion from upstream to HDFS, processing and analyzing the data in HDFS.
- Created Partitioned and Bucketed Hive tables in Parquet File Formats with Snappy compression and then loaded data into Parquet hive tables from Avro hive tables.

Environment: Scala, Azure, HDFS, Yarn, MapReduce, Hive, Sqoop, Flume, Oozie, Kafka, Impala, Spark SQL, Spark Streaming, Eclipse, Oracle, Teradata, PL/SQL UNIX Shell Scripting.

Data Engineer **March 2014- December 2016**

End Client – IDBI Bank - India

- Involved in all phases of the Big Data Implementation including requirement analysis, design, development, building, testing, and deployment of Hadoop cluster in fully distributed mode.
- Worked with ETL data flow using Informatica power center.
- Performed load and retrieve unstructured data (CLOB, BLOB etc.)
- Imported Legacy data from SQL Server and Teradata into Amazon S3 data lake.
- Developed Hive jobs to transfer 8 years of bulk data from DB2, MS SQL Server to HDFS layer.
- Implemented Data Integrity and Data Quality checks in Hadoop using Hive and Linux scripts.
- Automated the DDL creation process in hive by mapping the DB2 data types.
- Experience in Hadoop framework, HDFS, MapReduce processing implementation.
- Tuning Hadoop performance with high availability and involved in recovery of Hadoop clusters.
- Responsible for coding Java Batch, Restful Service, Map Reduce program, Hive queries, testing, debugging, Peer code review, troubleshooting and maintaining status report.
- Designed Business classes and used Design Patterns like Data Access Object, MVC etc.
- Used AVRO, Parquet file formats for serialization of data.
- Developed several test cases using MR Unit for testing Map Reduce Applications
- Experience in using HBase as a backend database for application development.
- Support/Troubleshoot hive programs running on the cluster and involved in fixing issues arising out of duration testing.

Environment: Informatica, Hadoop, Hive, Pig, HBase, Avro, Parquet, Oracle 9i, SQL*Plus, PL/SQL, MS Access, UNIX Shell Scripting.