Name: Hrishi Reddy
Email: hrishi.rl3223@gmail.com
Ph: +1 (505) 333-8765

PROFESSIONAL SUMMARY

- IT professional with 9+ years overall experience, specializing in **Big Data ecosystem, Data Acquisition, Ingestion, Modeling, Storage Analysis, Integration, Data Processing**, and **Database Management**.
- Experience in application development, implementation, deployment, and maintenance using **Hadoop** and **Spark**-based technologies like **Cloudera, Hortonworks**, **Amazon EMR**, and **Azure HDInsight**.
- A **Data Science enthusiast** with strong Problem solving, Debugging, and Analytical capabilities, who actively engage in understanding and delivering to business requirements.
- Ample work experience in **Big-Data ecosystem** - Hadoop (HDFS, MapReduce, Yarn), Spark, Kafka, Hive, Impala, HBase, Sqoop, Pig, Airflow, Oozie, Zookeeper, Ambari, Flume.
- Good knowledge of **Hadoop** cluster architecture and its key concepts - Distributed file systems, Parallel processing, High availability, fault tolerance, and Scalability.
- Complete knowledge of **Hadoop architecture** and daemons **of Hadoop clusters,** which include Name node, Data node, Resource manager, Node Manager, and Job history server**.**
- Expertise in developing Spark applications for interactive analysis, batch processing and stream processing, using programming languages like PySpark, Scala.
- Advanced knowledge in Hadoop based Data Warehouse (**HIVE**) and database connectivity (**SQOOP**).
- Ample experience using **Sqoop** to ingest data from **RDBMS** - Oracle, MS SQL Server, Teradata, PostgreSQL, and MySQL.
- Experience in working with various streaming **ingest services** with Batch and Real-time processing using **Spark Streaming, Kafka, Confluent, Storm, Flume, and Sqoop.**
- Proficient in using **Spark API** for streaming real-time data, staging, cleaning, applying transformations, and preparing data for machine learning needs.
- Experience in developing end-to-end ETL pipelines using **Snowflake**, **Alteryx,** and **Apache NiFi** for both relational and non-relational databases (**SQL** and **NoSQL**).
- Strong working experience on **NoSQL** and their integration with the Hadoop cluster- **HBase**, **Cassandra**, **MongoDB, DynamoDB**, and **CosmosDB**.
- Experience with **AWS** cloud services to develop cloud-based pipelines and Spark applications using **EMR**, **LAMBDA** and **Redshift**.
- Extensive knowledge in working with **Amazon EC2** to provide a solution for computing, query processing, and storage across a wide range of applications.
- Expertise in using **AWS S3** to stage data and to support data transfer and data archival. Experience in using **AWSRedshit** for large scale data migrations using **AWS DMS** and implementing **CDC** (change data capture).
- Strong experience in developing **LAMBDA** functions using **Python** to automate data ingestion and tasks.
- Working knowledge of**Azure** cloud components (HDInsight, Databricks, DataLake, Blob Storage, Data Factory, Storage Explorer, SQL DB, SQL DWH, CosmosDB).
- Experienced in building data pipelines using **Azure Data Factory**, **Azure Databricks,** and loading data to **Azure Data Lake, Azure SQL Database**, **Azure SQL Data Warehouse,**and controlling database access.
- Extensive experience with **Azure** services like **HDInsight**, **Stream Analytics**, **Active Directory**, **Blob Storage**, **Cosmos DB**, and **Storage Explorer**.
- Good knowledge in understanding the security requirements and implementation using **Azure Active Directory, Sentry, Ranger**, and **Kerberos for** authentication and authorizing resources.
- Experience in all phases of **Data Warehouse** development like requirements gathering, design, development, implementation, testing, and documentation.

- Extensive knowledge of Dimensional **Data Modeling** with Star Schema and Snowflake for FACT and Dimensions Tables using Analysis Services.
- Good experience in the development of **Bash scripting**, **T-SQL,** and **PL/SQL Scripts**.
- Sound knowledge in developing highly scalable and resilient **RestfulAPI**s, **ETL** solutions, and third-party platform integrations as part of the Enterprise Site platform.
- Experience in designing interactive dashboards, and reports, performing ad-hoc analysis and visualizations using **Tableau, Power BI, Arcadia,** and **Matplotlib**.
- Experience in implementing pipelines using **ELK(Elasticsearch, logstash, kibana)** and developing stream processes using **Apache Kafka**.
- Sound knowledge and experience in programming languages like **Python**, **Scala**.
- Experience in using various IDEs like **Eclipse**, **IntelliJ**, and repositories **SVN** and **Git** version control systems.
- A team player with strong communication, interpersonal, problem-solving, and debugging skills. Ability to quickly adapt to new environments and technologies.
- Successfully working in a fast-paced environment, both independently and in a collaborative way. Expertise in complex troubleshooting, root-cause analysis, and solution development.

TECHNICAL SKILLS

Big Data Ecosystem	HDFS, Yarn, MapReduce, Spark, Kafka, Kafka Connect, Hive, Airflow, StreamSets, Sqoop, HBase, Flume, Pig, Ambari, Oozie, ZooKeeper, Nifi, Sentry
Hadoop Distributions	Apache Hadoop 2.x/1.x, Cloudera CDP, Hortonworks HDP
Cloud Environment	Amazon Web Services (AWS), Microsoft Azure,GCP
Databases	MySQL, Oracle, Teradata, MS SQL SERVER, PostgreSQL, DB2
NoSQL Database	DynamoDB, Hbase
AWS	EC2, EMR, S3, Redshift, EMR, Lambda, Kinesis Glue, Data Pipeline
Microsoft Azure	Databricks, Data Lake, Blob Storage, Azure Data Factory, SQL Database, SQL Data Warehouse, Cosmos DB, Azure Active Directory
Operating systems	Linux, Unix, Windows 10, Windows 8, Windows 7, Windows Server 2008/2003, Mac OS
Software's/Tools	Microsoft Excel, Statgraphics, Eclipse, ShellScripting, ArcGIS, Linux, Jupyter Notebook, PyCharm, Vi / Vim, Sublime Text, Visual Studio, Postman
Reporting Tools/ETL Tools	Informatica, Talend, SSIS, SSRS, SSAS, ER Studio, Tableau, Power BI, Arcadia, Data stage, Pentaho
Programming Languages	Python (Pandas, Scipy, NumPy, Scikit-Learn, StatsModels, Matplotlib, Plotly, Seaborn, Keras, TensorFlow, PyTorch), PySpark, T-SQL/SQL, PL/SQL,HiveQL, Scala, UNIX Shell Scripting
Version Control	Git, SVN, Bitbucket
Development Tools	Eclipse, NetBeans, IntelliJ, Hue, Microsoft Office Suite (Word, Excel, PowerPoint, Access)

PROFESSIONAL EXPERIENCE

Client: Parker Hannifin Corporation, Oxnard, CA **Dec 2021 - Present**
Role: Sr Data Engineer

Responsibilities:

- Worked on Apache Spark data processing project to process data from RDBMS and several data streaming sources and developed Spark applications using Python on AWS EMR.
- Designed and deployed multi-tier applications leveraging AWS services like (EC2, Route 53, S3, RDS, DynamoDB) focusing on high availability, fault tolerance, and auto-scaling in AWS Cloud Formation.
- Configured and launched AWS EC2 instances to execute Spark jobs on AWS Elastic Map Reduce (EMR).
- Performed data transformations using Spark Data Frames, SparkSQL, Spark File formats, Spark RDDs.
- Effectively worked in Informatica version based environment and used deployment groups to migrate the objects.
- Loaded data from various sources like RDBMS (MySQL, Teradata) using Sqoop jobs.
- Handled JSON datasets by writing custom Python functions to parse through JSON data using Spark.
- Developed a preprocessing job using Spark Data Frames to flatten JSON documents to flat files.
- Improved performance of cluster by optimizing existing algorithms using Spark.
- Developed ETL programs using Informatica to implement the business requirements.
- Aggregated logs data from various servers and made them available in downstream systems for analytics by using Apache Kafka. Improved Kafka performance and implemented security.
- Developed batch and streaming processing apps using Spark APIs for functional pipeline requirements.
- Automated data storage from streaming sources to AWS data lakes like S3, Redshift and RDS by configuring AWS Kinesis (Data Firehose).
- Performed analytics using real time integration capabilities of AWS Kinesis (Data Streams) on streamed data
- Create new mapping designs using various tools in Informatica Designer like Source Analyzer, Warehouse Designer, Mapplet Designer and Mapping Designer
- Effectively used Informatica parameter files for defining mapping variables, workflow variables, FTP connections and relational connections
- Performed reporting analytics on data from AWS stack by connecting it to BI tools (Tableau, Power Bi).
- Imported data from AWS S3 into SparkRDD performed transformations and actions on RDD's.
- Worked with database administrating team on SQL optimization for databases like Oracle, MySQL, MS SQL.
- Assisted in configuring and implemented MongoDB cluster nodes on AWS EC2 instances.
- Identified executor failures, data skewness, and runtime issues by monitoring Spark apps through Spark UI.
- Ensured database performance in production by stress testing AWS EC2 of DynamoDBenvironments.
- Automated deployments and routine tasks using UNIX Shell Scripting.
- Collaborated with the Data Science team building machine learning models on Spark EMR cluster to deliver the data needs under business requirements.
- Worked in an agile environment to implement projects and enhancements with weekly SCRUMs.

Environment: Hadoop 2.x, Spark v2.0.2, Hive, Sqoop, Informatica, Kafka, Spark streaming, ETL, Scala, Python (Pandas, Numpy), PySpark, GIT (version control), MySQL, MS SQL, MongoDB, AWS (EC2, S3, EMR, RDS, Lambda, Kinesis, Redshift, Cloud Formation)

Client: ADP, Atlanta, GA **Jan 2021 – Dec 2021**
Role: Data Engineer

Responsibilities:

- Designed and deployed data pipelines using Azure cloud platform (HDInsight, DataLake, DataBricks, Blob Storage, Data Factory, Synapse, SQL, SQL DB, DWH, and Data Storage Explorer).

- Developed custom-built ETL solution, batch processing, and real-time data ingestion pipeline to move data in and out of the Hadoop cluster using PySpark and Shell Scripting.
- Integrated on-premises data (MySQL, Hbase) with cloud (Blob Storage, Azure SQL DB) and applied transformations to load back to Azure Synapse using Azure Data Factory.
- Migrated an entire oracle database to BigQuery and also build Data pipelines in airflow in GCP for ETL related jobs using different airflow operators.
- Built and published Docker container images using Azure Container Registry and deployed them into Azure Kubernetes Service (AKS).
- Imported metadata into Hive and migrated existing tables and applications to work on Hive and Azure.
- Created complex data transformations and manipulations using ADF and Scala.
- Hands-on experience on-premises ETL's to google cloud platform (GCP) using cloud native tools such as Big Query, Cloud Data Proc, and Google Cloud Storage.
- ConfiguredAzure Data Factory (ADF) to ingest data from different sources like relational and non-relational databases to meet business functional requirements.
- Used Informatica Power Center for (ETL) extraction, transformation and loading data from heterogeneous source systems into target database
- Optimized workflows by building DAGs in Apache Airflow to schedule the ETL jobs and implemented additional components in Apache Airflow like Pool, Executors, and multi-node functionality.
- Improved performance of Airflow by exploring and implementing the most suitable configurations.
- Configured Spark streaming to receive real-time data from Apache Flume and store the stream data using Scala to Azure Table and DataLakeis used to store and do all types of processing and analytics. Created data frames using Spark Dataframes.
- Designed cloud architecture and implementation plans for hosting complex app workloads on MS Azure.
- Performed operations on the transformation layer using Apache Spark RDD, Data frame APIs, and Spark SQL and applied various aggregations provided by Spark framework.
- Provided real-time insights and reports by mining data using Spark Scala functions. Optimized existing Scala code and improved the cluster performance.
- Processed huge datasets by leveraging Spark Context, SparkSQL, and Spark Streaming.
- Enhanced reliability of Spark cluster by continuous monitoring using Log Analytics and Ambari WEB UI.
- Improved the query performance by transitioning log storage from Cassandra to Azure SQL Datawarehouse.
- Implemented custom-built input adapters using Spark, Hive, and Sqoop to ingest data for analytics from various sources (Snowflake, MS SQL, MongoDB) into HDFS. Imported data from web servers and Teradata using Sqoop, Flume, and Spark Streaming API.
- Improved efficiency of large datasets processing using Scala for concurrency support and parallel processing.
- Developed map-reduce jobs using Scala for compiling program code into bytecode for the JVM for data processing.Ensured faster data processing by developing Spark jobs using Scala in a test environment and used Spark SQL for querying.
- Improved processing time and efficiency by using Spark applications like batch interval time, level of parallelism, memory tuning.Monitored workflows for daily incremental loads from RDBMSs (MongoDB, MS SQL, MySQL).
- Implemented indexing to data ingestion using Flume sink to write directly to indexers deployed on a cluster.
- Delivered data for analytics and Business intelligence needs by managing workloads in GCP.
- Improved security by using Azure DevOps and VSTS (Visual Studio Team Services) for CI/CD, Active Directory, and Apache Ranger for authentication. Managed resources and scheduling across the cluster using Azure Kubernetes Service**.**

Environment: Hadoop, Spark, Hive, Sqoop, HBase, Flume, Ambari, Scala, MS SQL, MySQL, Snowflake, MongoDB,Git, Data Storage Explorer, Python, Azure (Data Storage Explorer, ADF, AKS, Blob Storage)

Client: Gainwell technologies, Columbus. OH **May 2018 - Dec 2020**
Role: Data Engineer

Responsibilities:

- Implemented and maintained the monitoring and alerting of production and corporate servers/storage using AWS Cloud watch.
- Managed servers on the Amazon Web Services (AWS) platform instances using Puppet, Chef Configuration management.
- Developed PIG scripts to transform the raw data into intelligent data as specified by business users.
- Worked in AWS environment for development and deployment of Custom Hadoop Applications.
- Worked closely with the data modelers to model the new incoming data sets.
- Involved in start to end process of Hadoop jobs that used various technologies such as Sqoop, PIG, Hive, Map Reduce, Spark and Shell scripts (for scheduling of few jobs.
- Expertise in designing and deployment of Hadoop cluster and different Big Data analytic tools including Pig, Hive, Oozie, Zookeeper, SQOOP, flume, Spark, Impala, Cassandra with Horton work Distribution.
- Involved in creating Hive tables, Pig tables, and loading data and writing hive queries and pig scripts
- Assisted in upgrading, configuration and maintenance of various Hadoop infrastructures like Pig, Hive, and HBase.
- Exploring with the Spark improving the performance and optimization of the existing algorithms in Hadoop using Spark Context, Spark-SQL, Data Frame, Pair RDD's, Spark YARN.
- Developed Spark code using Scala and Spark-SQL/Streaming for faster testing and processing of data. Configured deployed and maintained multi-node Dev and Test Kafka Clusters.
- Performed transformations, cleaning and filtering on imported data using Hive, Map Reduce, and loaded final data into HDFS.
- Worked on tuning Hive and Pig to improve performance and solve performance related issues in Hive and Pig scripts with good understanding of Joins, Group and aggregation and how it does Map Reduce jobs
- Exploring with the Spark improving the performance and optimization of the existing algorithms in Hadoop using Spark Context, Spark-SQL, Data Frame, Pair RDD's, Spark YARN.
- Developed Spark code using Scala and Spark-SQL/Streaming for faster testing and processing of data.
- Import the data from different sources like HDFS/HBase into Spark RDD.
- Developed a data pipeline using Kafka and Storm to store data into HDFS.
- Performed real time analysis on the incoming data.
- Used Spark Streaming to divide streaming data into batches as an input to Spark engine for batch processing.
- Implemented Spark using Scala and SparkSQL for faster testing and processing of data.

Environment: Apache Hadoop, HDFS, MapReduce, Sqoop, Flume, Pig, Hive, HBASE, Oozie, Scala, Spark, Linux.

Client: Citizens Financial Group, Columbus, OH **Jun 2013 – Jul 2017**
Role: Data Engineer

Responsibilities:

- Participated in the analysis, design, and development phase of the Software Development Lifecycle (SDLC).
- Developed test-driven web applications using Java J2EE, Struts 2.0 framework, Spring MVC, Hibernate framework, JavaScript, and SQL Server database with deployments on IBM WebSphere.
- Designed and developed NSEP, which is an online web application where students can register, find, search, and apply for the jobs available. Utilized Java J2EE, JavaScript, SQL, HTML, CSS,and XML on Eclipse.

- Designed & developed a web Portal using Struts Framework, J2EE. Developed newsletter as part of process improvement tasks using HTML and CSS to report the weekly activities.
- Developed front-end, User Interface using HTML, CSS, JSP, Struts, Angular, and NodeJS, and session validation using SpringAOP.
- Extensively used Java multi-threading to implement batch Jobs with JDK 1.5 features and deployed it on the JBoss server.
- Ensured High availability and load balancing by configuring and Implementing clustering of Oracle on WebLogic Server 10.3.
- Improved productivity by developing an automated system health check tool using UNIX shell scripts.

Environment: Java/J2EE, Spring, Oracle, Linux, JDBC, Git,HTML, CSS, Angular, NodeJS, Postman, Servlets, Struts, JSP, WebLogic, PL/SQL, Eclipse.