Girija Polamreddy

girijapolamreddy03@gmail.com | +1 (716) 520-8611 | LinkedIn | GitHub | Portfolio

---

Summary:

- Seasoned **Big Data Engineer** with over 5 years of expertise, adept in **analysis, design, development, implementation, maintenance, and support across the entire Big Data ecosystem.** Proficient in utilizing **Hadoop Development** and related technologies for comprehensive data solutions.

- Proven track record of leveraging **AWS** cloud services such as **EC2, VPC, S3, Glue, EMR, RedShift, CloudWatch**, and **Lambda** functions to architect scalable and resilient Big Data systems.

- Skilled in constructing **data pipelines** and performing operations on **Microsoft Azure**, ensuring seamless data processing and analytics across cloud platforms.

- Masterful in **Shell and Bash Scripting**, harnessing scripting languages to automate and optimize data workflows and system operations.

- Experienced in setting up **clusters in EMR** and implementing efficient data storage strategies in **S3**, ensuring high-performance and cost-effective data management.

- Agile practitioner, actively involved in the **Software Development Life Cycle**, delivering iterative and value-driven solutions.

- Specialized in **real-time streaming** solutions using Kafka, empowering organizations to harness the power of continuous data processing. Proficient in **Sqoop f**or seamless data transfers between **RDBMS and HDFS/HBase/Hive.**

- Proven expertise in developing **high-throughput ETL pipelines** and constructing robust **data lakes**, facilitating efficient data processing and analytics at scale.

- Skilled in **ETL Processes, Data mining**, and implementing Web reporting features for **Data warehouses** through Business Object, enabling insightful reporting and decision-making.

- Proficient in handling **large-scale data processing** using technologies such as **Spark, Scala, Python, Apache Kafka, Pig/Hive, and Impala**, both in **batch and streaming modes.**

- Extensive hands-on experience with **Spark components** like **Spark SQL** and **Spark Streaming**, enabling advanced data processing and **real-time analytics**.

- Deep involvement in implementing **Hive and Pig scripts** for efficient data transformations, migrating data from **OLAP/OLTP** systems to **HDFS**, and developing **Change Data Capture (CDC) logic.**

- Experienced in creating, scheduling, and monitoring workflows using **Apache Airflow** with **Python**, automating data pipelines and orchestrating complex data workflows.

- Expert in working with the **Hadoop ecosystem**, including **HDFS, MapReduce, Spark, Kafka, HBase, Scala, Pig, Impala, Sqoop, Oozie, Flume, and Zookeeper**. Proficient in **utilizing Spark SQL, Spark Streaming, and AWS services such as EMR, S3, Airflow, Glue, and Redshift.**

- In-depth understanding of **Hadoop architecture,** key components, and programming models like **MapReduce** with proficiency in data modeling, cleansing, profiling**,** and analysis techniques.

- Collaborative team player with a strong ability to contribute to both development and maintenance phases of projects, ensuring the successful delivery and long-term stability of data solutions**.**

Technical Skills:

Big Data Ecosystem	HDFS, Yarn, MapReduce, Spark, Kafka, Kafka Connect, Hive, Airflow, Impala, Sqoop, HBase, Flume, Oozie, Zookeeper
Hadoop Distributions	Cloudera, Hortonworks, Apache.
Cloud Environments	AWS EMR, EC2, S3, AWS Redshift, Airflow, Microsoft AZURE, Google Cloud platform.
Operating Systems	Linux, Windows, Mac OS
Programming Languages	Python, SQL, Scala, Java, C, C++
Databases	Oracle, SQL Server, MySQL, HBase, MongoDB, RedShift, DynamoDB, PostgreSQL
Deployment	AWS Cloud (S3, RDS, EC2, ECS), GCP, Docker, Kubernetes
Tools	Informatica, Git, Eclipse, IntelliJ Ide, Visual Studio Code, Jupyter Notebook, Tableau, Power BI, Postman, JIRA, JMeter, Mocha, Maven, Gradle
Repositories	GitHub, SVN.
Scripting Languages	bash/Shell scripting, Linux/Unix
Methodology	Agile, Waterfall

Professional Experience

Moore Archive | University at Buffalo, New York **2022 Apr – 2022 Dec**

Programming Assistant(Student Assistant)

- Utilized **XML, HTML, CSS,** and **JavaScript** to enhance the website's user interface, including the Notebook Viewer for easy access to Marianne Moore's works.

- Employed **XSLT, XQuery,** and **XPath** for data manipulation and organization of Marianne Moore's works.
- Used **Joomla CMS (Hubzero 2.0)** for website management and maintenance.
- Applied **Python** for various backend tasks and automation.
- Implemented **OCR** tools for digitization of Marianne Moore's works.
- Managed data using **JSON** and ensured version control with Git (**GitHub**).
- Utilized **TimelineJS** for creating interactive timelines on the website.
- Worked with **EVT** and **CWRC-GitWriter** for text encoding and editing.

Environment: HTML, CSS, XML (TEI), XSLT XQuery, XPath, JSON, Javascript, Joomla CMS (Hubzero 2.0), Python, OCR (various), JSON, Git (GitHub), TimelineJS EVT, CWRC-GitWriter, TEI P5-compliant RelaxNG.

Pike solutions – Hyderabad, India **2020 Jan – 2022 Jan**

Sr. Data Engineer

Responsibilities:

- Implemented efficient export of event weblogs to **HDFS** by creating a HDFS sink and configured **Spark Streaming** with **Scala** to store real-time data from **Apache Kafka** in HDFS.
- Designed and developed **POCs** in Spark using Scala to compare performance with **Hive** and **SQL/Oracle,** showcasing expertise in data processing and analysis.
- Demonstrated extensive knowledge of **Hadoop architecture,** including **YARN, HDFS**, and **MapReduce** concepts.
- Created multi-node Hadoop and Spark clusters in **AWS instances**, generating and storing terabytes of data in AWS HDFS.
- Developed Spark code from scratch using Scala and conducted initial testing using both Hive and SQL contexts.
- Designed and implemented a system using **Kafka and Spark** to collect and process data from multiple portals.
- Proficient in working with **Hive, AWS Athena, and Redshift**, creating external tables with partitions and processing data to HDFS using **Sqoop.**
- Managed and supported enterprise Data Warehouse operations, including advanced predictive application development using Cloudera and Hortonworks HDP.
- Created a **Virtual Data Lake** using AWS Redshift, S3, Spectrum, and Athena services for querying large amounts of data stored on S3.
- Implemented an on-demand, secure **EMR launcher** with custom Spark submit steps using **S3 Event, SNS, KMS,** and Lambda functions.

- Uploaded and processed terabytes of data from structured and unstructured sources into HDFS using Sqoop, implementing business logic with Spark/Scala for a Rating Engine.

- Developed **ETL data pipelines** using Hadoop big data tools, including HDFS, Hive, Presto, Apache Nifi, Sqoop, Spark, Elastic Search, and Kafka.

- Proficient in data ingestion using **Flume, Pig**, and Sqoop for customer data histories in HDFS.

- Extensive experience in loading and transforming structured, semi-structured, and unstructured data in various formats such as **text, zip, XML, and JSON**.

- Skilled in using **Airflow** for job scheduling and automation, leveraging Hive query language and Scala for data operations.

- Extracted real-time feeds with **Spark Streaming**, processed data into Cassandra, and integrated distributed messaging queues using **Apache Kafka and Zookeeper**.

- Provided specifications for Hadoop cluster size, resource allocation, and distribution by writing JSON specification files.

Environment: Hadoop, Spark, HDFS, Hive, Pig, HBase, AWS, EMR, Big Data, Oozie, Sqoop, Kafka, Apache Nifi, Flume, Zookeeper, MapReduce, Cassandra, Scala, Linux, NoSQL, MySQL, SQL Server.

Pike solutions – Hyderabad, India **2018 Jan – 2019 Dec**

Data Engineer

Responsibilities:

- Assessed the compatibility of Hadoop and its ecosystem with the project, implementing and validating them through various proof-of-concept (POC) applications for potential integration into the **Big Data Hadoop Initiative.**

- Successfully installed and configured **Hadoop MapReduce** and **HDFS**, developing multiple Java and Scala MapReduce jobs for **data cleansing** and **preprocessing.**

- Built and maintained **AWS data pipelines**, leveraging resources like **AWS API Gateway**, **Snowflake, DynamoDB, AWS Lambda functions**, and AWS S3. Designed workflows to receive responses from AWS Lambda, retrieve data from Snowflake, and convert responses into **JSON** format.

- Developed **Spark jobs** using **RDDs, Pair RDDs**, **Transformations, Actions, and data frames to transform relational datasets and enable advanced data processing.**

- Implemented **data quality strategies** as an integral part of **ETL processes**, ensuring **data accuracy and integrity** throughout the pipeline.

- Created complex **SQL queries** and **PL/SQL stored procedures**, translating them into ETL tasks to extract, transform, and load data.

- Maintained comprehensive documentation associated with **business processes, mapping design, data profiles,** and tools, ensuring **knowledge sharing and project continuity.**

- Wrote **MapReduce code** to process and parse data from various sources, storing the parsed data in HBase and Hive using **HBase-Hive integration**.

- Leveraged **AWS S3** for **data transformations** according to business requirements, optimizing data processing and storage.

- Managed **data movement** between **Oracle and HDFS** using **Sqoop,** fulfilling the needs of business users and facilitating seamless data integration.

- Utilized **Oozie Scheduler** to automate pipeline workflows and orchestrate **map-reduce jobs**, ensuring timely data extraction and processing.

- Actively participated in **knowledge sharing sessions** with team members, fostering collaboration and continuous learning.

- Planned, deployed, monitored, and maintained **Amazon AWS cloud infrastructure**, including multiple EC2 nodes and VMs, to meet environmental requirements.

- Managed and reviewed Hadoop log files, troubleshooting issues and ensuring system stability.

- Created **Hive queries and UDFs** for **data analysis** and **transformation** in HDFS, enabling efficient data processing and analytics.

- Developed Hive scripts to implement control table logic and designed and implemented partitioning and buckets in Hive for optimized query performance.

- Developed complex **Hive queries and Spark scripts** to meet advanced data processing needs.

- Created **test scripts** to support **test-driven development** and **continuous integration**, ensuring code quality and reliability.

- Managed the loading and transformation of large **structured, semi-structured, and unstructured datasets,** handling diverse data formats efficiently.

- Worked with Spark using **Scala and Spark SQL** for faster data testing and processing, harnessing the power of distributed computing.

- Utilized the **Hue browser** for interacting with Hadoop components, simplifying and enhancing the user experience.

- Transformed **Hive/SQL** queries into Spark transformations using Spark RDDs, Python, and Scala, leveraging Spark's advanced capabilities for data processing.

- Operated on the **Hortonworks distribution** of Hadoop and utilized **Ambari** for **cluster health monitoring,** ensuring optimal performance and reliability of the Hadoop ecosystem.

Environment: Hadoop, Hive, Scala, GitHub, Spark, Tableau, Sqoop, HDP, Python, Shell Scripting, AWS, Linux.