SREENIVAS REDDY

Email: sreenivasreddybigdata@gmail.com **Contact:** +1 (571) 699-7988

PROFESSIONAL SUMMARY:

- 7 Years of experience in software development life cycle, development, and system application architecture.
- Fluid understanding of multiple programming languages, including C#, Java, Scala, and Python.
- Good Knowledge of Hadoop Architecture and various daemons such as Job Tracker, Task Tracker, Name Node, Data Node, and Resource Manager.
- Experienced in HDFS data storage and running Map Reduce jobs.
- Good exposure to Hadoop ecosystems.
- Expertise with the Hadoop ecosystem tools like Spark, Hive, HBase, Pig, Storm, Kafka, SQOOP, Flume, Zookeeper, Oozie, and Ni-Fi.
- Proficient in Azure Data Factory, Azure Databricks, Azure SQL Database, and Azure Synapse Analytics
- Proficient in data migration from various databases to HDFS using SQOOP and vice-versa using Flume.
- Expertise in managing Airflow DAGs, Monitoring the tasks, runs, Failures, and logs.
- Hands-on Experience in using AWS glue with Apache Airflow for scheduling workflow right from data ingestion followed by processing data in different formats and finally giving gold-level data to the end-users.
- Experienced with handling AWS services, data storage tools such as Amazon S3, Data integration tools such as AWS Glue, and Data warehousing tools such as AWS Redshift.
- Hands-on working skills with different file formats like Parquet, ORC, SEQ, AVRO, JSON, RC, CSV, and compression techniques like Snappy, GZip, and LZO.
- Experienced in designing both time-driven and data-driven automated workflows using Oozie.
- Extending Hive and Pig core functionality by using custom UDF, UDTF, and UDAF in Spark Scala.
- Hands-on experience in writing Spark jobs and Spark streaming API using Java, Scala, and Python.
- Experienced in handling Spark SQL, Streaming, and complex analytics using Spark over Cloudera Hadoop YARN.
- Good understanding of the architecture of Cassandra as well as its integration with Spark Scala.
- Proficient experience with DevOps essential tools like Clear Case, GIT, Ant, Maven, Jenkins, Chef, Puppet, Ansible, and Docker.
- Well-versed with Ansible Playbooks, modules, and roles. Wrote playbooks with Python SSH as the wrapper to manage AWS nodes.
- Skilled in Cassandra maintenance and performing tuning on both database and server.
- Good at implementing Kafka custom encoders for custom input format to load data in partitions.
- Designed and developed the core ETL pipeline, that involves Java and Scala coding and using Kafka and Spark frameworks.
- Expertise in using Kafka for log aggregation solution with low latency processing and distributed data consumption and widely used Enterprise Integration Patterns (EIPs).
- Experienced in working with Hadoop Storage and Analytics framework over AWS cloud using tools like SSH, Putty, and Mind-Term.
- Hands-on experience in configuring EC2 instances in VPC network and managing security through IAM and monitoring server's health through Cloud watch.

- Experienced in the ETL process, Data Modeling and Mapping, Data Integration, Business Intelligence, and Data Analysis, Data Validation and Data Cleaning, Data Verification, and Identifying data mismatch.
- Solid understanding of OLAP concepts and challenges, especially with large data sets and mapping, analysis, and documentation of OLAP reports.
- Excellent Java development skills using Springs, J2SE, Struts, Servlets, Junit, JSP, JDBC, Hibernate, and JPA for object mapping with the database.
- Good working experience in using Spring modules like Spring Core Container Module, Spring Application Context Module, Spring MVC Framework Module, and Spring ORM Module in web applications.
- Expertise in creating Dashboards and Alerts using Splunk Enterprise, Tableau, and Monitoring using DAGs.
- Skilled in monitoring servers using Nagios, Datadog, Cloud watch, and using ELK stack Elastic search Logstash.
- Highly motivated self-starter with good communication and interpersonal skills.
- Techno-functional responsibilities include interfacing with users, identifying functional and technical gaps, designing custom solutions, leading developers, and producing documentation and production support.
- Good team player, dependable resource, and ability to learn new tools and software quickly as required.
- Good Domain Knowledge of Automobile, Banking, Retail, Healthcare, and Insurance.

TECHNICAL SKILLS:

PROGRAMMING LANGUAGES	C#, Java, Scala, Python, and Shell Scripting
BIG DATA ECOSYSTEM	Spark, Hive, HBase, SQOOP, Oozie, Storm, Flume, Pig, Kafka, NIFI, Zookeeper, MapReduce
CLOUD	AWS EMR, EC2, S3, RDS, Azure Databricks, Azure Data Factory, GCP
DBMS	SQL Server, MySQL, PL/SQL, Oracle, Cassandra, Vertica, Versant
WEB TECHNOLOGIES	HTML, JavaScript, XML, JQuery, Ajax, CSS
WEB SERVICES	Web Logic, Web Sphere
IDEs	Eclipse, IntelliJ, Visual Studio, WinSCP
DevOps	GitHub, Jenkins, Ansible, Chef, Docker, Nagios, Puppet
OPERATING SYSTEMS	Windows, Unix, Linux, Solaris, CentOS
FRAMEWORKS	MVC, Struts, Maven, Junit, Log4J, ANT, Tableau, Splunk, Aqua-data Studio
J2EE TECHNOLOGIES	Spring, Servlets, J2SE, JSP, JDBC

WORK EXPERIENCE:

Signant Health, Blue bell, PA **Aug'22 – Sep'23**

Sr Data Engineer

Responsibilities:

- Architected and implemented end-to-end ETL pipelines utilizing AWS Glue, Snowflake, PySpark, and SQL to extract data from various databases such as Oracle and SQL Server into Snowflake's scalable cloud-native data model.
- Designed performant Snowflake schemas for structured and semi-structured datasets using advanced features like clustering keys and materialized views to optimize query performance.
- Developed complex ETL transformations in PySpark within the AWS Glue framework for processing large-scale datasets efficiently on AWS EMR clusters.
- Coordinated data acquisition, transformation, analysis, storage, and database management using Azure Databricks and Azure SQL Database, overseeing a huge data volume within Azure Data Factory.
- Executed data integration, ETL operations, and data modeling in Azure Synapse Analytics, ensuring smooth data migration across development (DEV), system integration testing (SIT), user acceptance testing (UAT), and production (PROD) environments.
- Performed thorough analysis and management of mapping data flows within Data Factory, enhancing real-time operations in Data Lake Storage and optimizing high-speed processing, resulting in increase in efficiency.
- Collaborated on team-based initiatives to develop Databricks notebooks in accordance with Azure DevOps standards.
- Incorporated serverless computing with AWS Lambda functions to trigger the execution of Glue jobs based on event-driven triggers or time-based schedules.
- Terraformed infrastructure resources required for the ETL process including configuring Amazon S3 buckets as staging areas, defining VPCs for secure connectivity between sources, databases, and Snowflake warehouse instances etc., enabling Jenkins CI/CD pipeline automation.
- Implemented continuous integration/continuous deployment (CI/CD) processes using Jenkins in conjunction with Terraform scripts:
- Set up version control systems such as Git/GitHub repositories to manage code changes seamlessly across development stages.
- Leveraged declarative configuration management through Terraform's Infrastructure-as-a-Code approach to provision resources like EC2 instances, RDS database instances ensuring consistency throughout different environments (development/staging/production).
- Orchestrated automated deployments by integrating Jenkins job workflows with source code repositories along with triggering appropriate tests via build agents/pipelines within a controlled environment.
- Implemented parameterized builds/jobs allowing users/passwords/secrets injection securely while deploying infrastructure stacks related specifically to ETL tasks accessing sensitive data sources/repos/schemas.
- Ensured robust monitoring & logging mechanisms during ETL processes leveraging multiple toolsets.
- Utilised CloudWatch Logs/Metrics/Dashboards along-with third-party tools such as Splunk/DataDog/Prometheus+Grafana/etc. for centralized real-time visibility into health, latency, and resource utilization of various components involved in ETL workflows.
- Configured alerts & notifications for triggering proactive responses to failures/errors/slowdowns based on predefined thresholds via SNS/SQS/Slack/Email.
- Implemented data quality checks using SQL queries within Glue jobs to validate the integrity and accuracy of loaded datasets into Snowflake.

- Optimized query performance by utilizing Snowflake's automatic clustering capabilities, query caching mechanisms, and workload management techniques like warehouses sizing/tuning (concurrency/scaling/upgrading).
- Employed best practices for security measures such as encryption at rest with AWS KMS, transit encryption with SSL/TLS between AWS services/components while transferring data during staging/loading/extraction processes.

Environment: AWS Glue, Amazon S3, Amazon Redshift, RDS (Relational Database Service), EMR (Elastic MapReduce), Oracle SQL, T-SQL (SQL Server), ANSI SQL, Snowflake Data Warehouse, Terraform, Jenkins CI/CD pipelines, Spark, Scala, Python, Oracle, Jenkins, AWS.

THALES, Melbourne, FL **May'21 – Jul'22**

Senior Data Engineer

Responsibilities:

- Led migration from Oracle to Redshift using AWS Athena and AWS S3.
- Designed and implemented a real-time ETL pipeline to process semi-structured data by integrating raw records from data sources using Pyspark and Kafka.
- Administered, monitored, and resolved incidents for ETL and data warehouse including failed loads and performance. Implement long-term fixes to address underlying root causes
- Did Performance tuning & profiling of dataflow.
- Created ETL scripts for Ad-hoc requests to retrieve data from analytical sites.
- Using Apache Airflow managed, structured, and organized ETL pipelines using Directed Acyclic Graphs (DAGs).
- Used Airflow to program workflows including the creation, scheduling, and monitoring of workflows.
- Used Apache Spark as a data processing framework to perform processing tasks on very large data sets.
- Used Spark to import customer information data from Oracle database into HDFS for data processing along with minor cleansing.
- Developed MapReduce jobs to calculate the total usage of data by commercial routers in different locations using Horton work distribution.
- Involved in information gathering for new enhancements in Spark, Production support for field issues and label installs for Hive scripts and MapReduce jobs.
- Managed GIT repositories for branching, merging, and tagging.
- Used Maven to build RPMs from source code in Scala checked out from GIT repository, with Jenkins being the Continuous Integration Server and Artifactory as repository manager.
- Responsible for Setting up UNIX/Linux environments for various applications using shell scripting.
- Set up scalability for application servers using a command-line interface for Setting up and administering DNS system in AWS using Route53.
- Managed users and groups using Amazon Identity and Access Management (IAM).
- In-depth understanding of the principles and best practices of Software Configuration
- Designed and developed server-side applications on the Linux platform in a fast-paced environment.

- Translated customer business requirements into technical design documents, established specific solutions and led the efforts including programming in Spark Scala and testing that culminate in client acceptance of the results.
- Expertise in Object-Oriented Design (OOD) and end-to-end software development experience working on Scala coding and implementing mathematical models in Spark Analytics.
- Software translating the business requirements into Use Cases and Diagrams conducting reviews of Codes and test cases analyzing change requests enhancements managing release plans for business apps
- Developed oozie workflows and scheduling jobs through Hue.
- Used AWS Lambda to perform data validation, filtering, sorting, and other transformations for every data change in an HBase table and load the transformed data to RDS.
- Loading data from different servers to the S3 bucket and setting appropriate bucket permissions.
- Configured routing to send JMS files to interact with the application for real-time data using Kafka.
- Managed Zookeeper for cluster coordination and Kafka Offset monitoring.
- Optimized legacy queries to extract customer information from Oracle.
- Reviewed HDFS usage and system design for future scalability and fault tolerance.
- Managed and reviewed Hadoop log files using Spark to identify issues when a job fails.

Environment: HDFS, Spark, Scala, Python, Shell Scripting, Hive, HBase, Oracle, MapReduce, Logstash, Jenkins, Versant, Java, Kafka, Horton works, GIT, ClearCase, Zookeeper, Ansible, AWS.

VERTEX, Austin, TX **Jun'19 – Mar'21**

Big Data Developer

Responsibilities:

- Installed and configured Hive, HDFS, and the NIFI implemented HDP Hadoop cluster. Assisted with performance tuning and monitoring.
- Involved in loading and transforming large sets of structured data from router location to EDW using a NIFI ETL pipeline flow.
- Developed PySpark code and Spark-SQL for faster testing and processing of data.
- Worked on Data serialization formats for converting complex objects into sequence bits by using Parquet, ORC, AVRO, JSON, and CSV formats.
- Created Hive tables to load large data sets of structured data coming from WADL after the transformation of raw data.
- Created reports for the BI team using SQOOP to export data into HDFS and Hive.
- Developed custom NIFI processors for parsing the data from XML to JSON format and filtering broken files.
- Created Hive queries to spot trends by comparing fresh data with EDW reference tables and historical metrics.
- Used PySpark to convert panda's data frame to Spark Data frame.
- Used KafkaUtils module in PySpark to create an input stream that directly pulls messages from Kafka broker.
- Worked on partitioning Hive tables and running scripts parallel to reduce the run time of the scripts.
- Extensively worked on creating an End-to-End ETL pipeline orchestration using NIFI.
- Implemented business logic by writing UDFs in Spark Scala and configuring CRON Jobs.

- Provided design recommendations and resolved technical problems.
- Assisted with data capacity planning and node forecasting.
- Involved in performance tuning and troubleshooting the Hadoop cluster.
- Developed H-catalog Streaming code to stream the JSON data into Hive (EDW) continuously.
- Administrated Hive, and Kafka installing updates, patches, and upgrades.
- Supported code/design analysis, strategy development, and project planning.
- Managed and reviewed Hadoop log files.
- Evaluated suitability of Hadoop and its ecosystem to project and implemented various proof of concept applications to eventually adopt them to benefit from the Hadoop initiative

Environment: Spark, Scala, Hive, Maven, Microservices, GitHub, Splunk, PySpark, Tableau, Tidal, SQOOP, Java 1.8, Linux, Aqua-data studio, NIFI, Google cloud, J2EE, HDFS, Kafka, MySQL

INFOMAX SOLUTIONS, San Diego, CA **Nov'17 – May'19**

Data Engineer

Responsibilities:

- Developed ETL jobs across multiple platforms using Spark Scala, Hadoop, and Vertica.
- Configured Hive Meta-store with Vertica database and vice versa using SQOOP.
- Designed Data flow to pull the data from Rest API using Apache NIFI with SSL context configuration enabled.
- Created a POC for the demonstration of retrieving the JSON data by calling Rest service and converting it into CSV by creating data flow and loading it into Vertica by calling a Unix script in NIFI.
- Developed custom processors in Java using maven to add functionality in Kafka for additional tasks.
- Wrote complex SQL queries, and PL/SQL stored procedures and convert them to ETL tasks in Spark.
- Created and maintained documents related to business processes, mapping design, data profiles, and tools.
- Extracted weblogs by using Spark Streaming job, which is written in JavaScript and continuously tracked using Oozie.
- Wrote PySpark jobs with RDDs, Pair RDDs, Transformations and actions, and data frames for data transformations from relational sets.
- Automated the development environment using Vagrant and Shell provisioning.
- Used Aqua-data studio to collaborate with the Vertica database for performance tuning and visual analysis.
- Created Dashboards and sets of data using Tableau for business decision purposes and estimating the sales on location bases.
- Developed UDFs to extract and trim the raw data using Spark Scala.
- Responsible for developing Spark coding for extracting data using JSON Reader function.
- Connected Tableau server to publish dashboards to a central position for portal integration.
- Created visual trends and calculations in Tableau on customers and product data as per client requirements.
- Designed Splunk Dashboards for monitoring the pipeline jobs in production.
- Created Alerts using Splunk for failed and late-running jobs.

Environment: Spark, Scala, Hadoop, Vertica, GitHub, Splunk, PySpark, Pig, Tableau, Tidal, Map Reduce, VSQL, SQOOP, Python Scripting, Shell, Linux, Aqua-data studio, NIFI, Vagrant, Oozie.

THRINAINA INFORMATICS Ltd., Hyderabad, India **Feb'16 – Nov'17**

Java Developer

Responsibilities:

- Worked on both WebLogic Portal 9.2 for portal development and WebLogic 8.1 for Data services programming.
- Developed the presentation layer using JSP, HTML, CSS, and client validations using JavaScript.
- Used GWT to send Ajax requests to the server and update data in UI dynamically.
- Developed Hibernate 3.0 in Data Access Layer to access and update information in the database.
- Used JDBC, SQL, and PL/SQL programming for storing, retrieving, and manipulating the data.
- Involved in designing and developing the eCommerce site using JSP, Servlets, EJBs, JavaScript, and JDBC.
- Used Eclipse 6.0 as IDE for application development and configured Struts framework to implement MVC design patterns.
- Validated all forms using Struts validation framework and implemented Tiles framework in the presentation layer.
- Designed and developed GUI using JSP, HTML, and CSS. Worked with JMS for messaging interface.
- Used XML for ORM mapping relations with the java classes and the database.
- Used Subversion as the version control system. Extensively used Log4j for logging the log files.

Environment: Java/J2EE, Oracle, SQL, PL/SQL, JSP, EJB, Struts, Hibernate, WebLogic, HTML, CSS, Servlets, UML, Junit, Log4j, Eclipse

QUALIFICATION:

- Rowan University **GPA: 3.94**

Master of Science in **Data Science**

- VNR VJIET **GPA: 3.5**

Bachelor of Technology in **Electronics and Communication Engineering**