**Giridhar M**

**PH: 704-837-8374**

**Professional Summary:**

* Around 10 years of IT experience in software design, development, implementation, and support of business applications.
* Experience in Big data Hadoop, **Hadoop Ecosystem** components like MapReduce, Sqoop, Flume, Kafka, Pig, Hive, Spark, Storm, HBase, Airflow, Oozie, and Zookeeper.
* Strong experience in migrating other databases to Snowflake.
* Worked extensively on installing and configuring Hadoop ecosystem architecture work.
* components Hive, **SQOOP, HBase**, Zookeeper, and Flume
* Good Knowledge in writing Spark Applications in Python (**Pyspark**).
* Working with the data extraction, transformation, and load using **Hive**, Sqoop, and HBase.
* Experienced in working with Spark ecosystem using **SCALA** and **HIVE Queries** on different data formats like Text file and parquet.
* Effectively managed and prioritized work, collaborated with cross-functional teams, and delivered high-quality solutions on time by implementing Agile methodologies such as Scrum and Kanban.
* Extensively worked on **Spark** using Scala on the cluster for computational (analytics), installed it on top of Hadoop performed advanced analytical applications by making use of Spark with Hive and SQL/Oracle.
* Hands-on experience in installing, configuring and using Apache Hadoop ecosystem components like Hadoop Distributed File System (**HDFS**), **MapReduce**, PIG, HIVE, HBASE, Apache Crunch, ZOOKEEPER, SCOOP, Hue, Scala, **AWS Glue, Lambda functions, Step functions, CloudWatch, SNS, DynamoDB, and SQS.**
* Good Knowledge of Amazon Web Service (**AWS**) concepts like EMR and EC2 web services which provide fast and efficient processing of Teradata Big Data Analytics.
* Expertise in containerization technologies, including Docker and Kubernetes, and containerized data application deployment and management on GCP Kubernetes Engine (GKE) to ensure scalability and dependability.
* Led data integration and migration initiatives from on-premises or other cloud platforms to Google Cloud Platform, ensuring minimal business disruption and a seamless transition.
* Familiarity with Calculated Measures and assigned Aggregation levels based on Dimension Hierarchies to optimize querying and analysis capabilities within the data warehouse environment.
* Solid Experience and understanding of Implementing large scale Data warehousing Programs and E2E Data Integration Solutions on Snowflake Cloud, AWS Redshift, Informatica Intelligent Cloud Services (IICS - CDI) & Informatica PowerCenter integrated with multiple Relational databases (MySQL, Teradata, Oracle, Sybase, SQL server, DB2)
* Utilizing Stackdriver and other tools, I implemented monitoring and logging solutions on GCP in order to detect anomalies, optimize resource utilization, and monitor performance in a cost-effective manner.
* Collaborated with clients and stakeholders to ascertain their perspectives on GCP data engineering solutions, solicit their input, and offer technical direction and assistance.
* Experience in building Snowpipe.
* Strong experience in using SQL for advanced analytical functions and writing complex PL/SQL packages, stored procedures, functions, cursors, triggers, views, and materialized views.
* Expertise in Big Data like Hadoop (Azure, Hortonworks, Cloudera) distributed systems, MongoDB, and **NoSQL.**
* Experience in Migrating SQL database to **Azure Data Lake, Azure Data Lake Analytics**, **Azure SQL Database, Data Bricks** and **Azure SQL Data warehouse, Azure HDInsight** and controlling and granting database accessandMigrating On-premise databases to **Azure Data Lake store** using Azure Data factory.
* Hands-on Experience developing **GCP** Big Query projects with Airflow as a scheduler.
* Possesses excellent knowledge of databases, data warehouses, and business intelligence concepts. Expertise in designing and developing ETL framework end to end from the source system to target OLAP systems using ETL tools such as Talend, Data Stage, and SSIS.
* Experience in both Logical and Physical data modeling for enterprise data warehouse and applications, ensuring alignment with business requirements and performance optimization.
* Expertise in architecting and designing Enterprise Data Warehouse solutions, ensuring scalability and efficiency for large-scale applications.
* Experience in designing and developing various components of Data Warehouse Architecture, including Enterprise Data Warehouse (EDW), Operational Data Store (ODS), Operational Data Marts, and Data Marts.
* Had Experience using **GCP** **Data Proc** cluster to with Hadoop tools like Pyspark and Hive.
* Experience in Developing **Spark** applications using **Spark - SQL** in **Databricks** for data extraction,
* Well-versed with Design and Architecture principles to implement Big Data Systems.
* Acumen on Data Migration from Relational Database to Hadoop Platform using **SQOOP**.
* Experienced in migrating ETL transformations using Pig Latin Scripts, transformations, and **join** operations.
* Good understanding of MPP databases such as HP Vertica and Impala.
* Proficient in developing ETL processes using Dimensional Modeling for both ROLAP and MOLAP environments, along with designing ETL technical architecture and managing BI load dependencies.
* Hands-on experience in configuring and working with Flume to load the data from multiple sources directly into HDFS
* Expertise in relational databases like Oracle, My SQL, and SQL Server.
* Strong analytical and problem-solving skills, highly motivated, good team player with very Good communication & interpersonal skills

**Technical Skills:**

|  |  |
| --- | --- |
| **Languages** | R, SQL, Python, Shell scripting, Java, Scala, C++. |
| **Databases** | Oracle 11g, SQL Server, MS Access, MySQL, MongoDB,  Cassandra, PL/SQL, T-SQL, |
| **Big Data**  **Ecosystems** | Hadoop, MapReduce, HDFS, HBase, Hive, Pig, Impala, kafka, Spark  MLLib. PySpark, Sqoop, AVRO. |
| **BI and**  **Visualization** | Tableau, SSAS, SSRS, Informatica |
| **Version Controls** | GIT, SVN, GitLab, Bitbucket |
| **Data Engineer/Big Data Tools / Cloud / Visualization / Other Tools** | Databricks, Hadoop Distributed File System (HDFS), ERWIN, Data Dimensional Modeling, Hive, Pig, Sqoop, MapReduce, Flume, YARN, Hortonworks, Cloudera, Mahout, MLlib, Oozie, Zookeeper, etc. AWS, AWS Lambda, AWS Glue, Data stage 9x/7.5/7.1, Redshift, Athena, Azure Databricks, Azure Data Explorer, **Azure HDInsight**, GCP, BigQuery, PubSub, Salesforce, Google Shell, Linux, PuTTY, Bash Shell, Unix, etc., Tableau, Power BI, SAS, Matplotlib, Seaborn, Bokeh. |

**Education:** Bachelor’s Degree in ECE from JNTUH, 2012.

**PROFESSIONAL EXPERIENCE**

**The Exchange Jan 2021 to Till Date**

**Sr Data Engineer**

**Responsibilities:**

* Expertise in designing and deployment of **Hadoop clusters** and different Big Data analytic tools including **Pyspark, Hive**, **Snowflake, and Airflow as schedulers**.
* Implemented advanced procedures like text analytics and processing using in-memory computing capabilities like Apache Spark written in Python.
* Leveraged Azure Synapse's unified analytics platform to design and execute end-to-end data analytics solutions for big data analytics, data integration, and data warehousing.
* Deployed Azure Synapse workspaces and provisioned resources such as SQL pools, Apache Spark pools, and data integration runtimes to support diverse data processing workloads.
* Day to-day responsibility includes developing ETL Pipelines in and out of data warehouse, develop major regulatory and financial reports using advanced SQL queries in snowflake.
* Developed stored procedures to extract the data from different sources and load it into **data warehouse.**
* Implemented **Spark** using Python and Spark SQL for faster testing and processing of data.
* Involved in converting **Hive/SQL** queries into Spark transformations using Spark RDDs, and Python.
* Worked with Spark to create structured data from the pool of unstructured data received.
* Documented the requirements including the available code which should be implemented using Spark, Hive, and HDFS.
* Monitoring and maintaining the best possible performance, availability, and scalability for GCP database services, including Cloud SQL, Bigtable, Firestore, and Spanner.
* Implement One time Data Migration of Multistate level data from SQL server to Snowflake by using Python and SnowSQL.
* Created and executed end-to-end machine learning processes by automating model creation, training, and deployment through the use of Amazon SageMaker Pipelines.
* Used SageMaker Pipelines to define dependencies between various processing steps and guarantee consistent execution while orchestrating sophisticated machine learning pipelines.
* Worked on ETL process of data loading from different sources and data validation process from staging area to Actavis data warehouse.
* Worked with ETL team involved in loading data to staging area to data warehouse. Provided all business rules for the database for loading data.
* Stage the API or Kafka Data(in JSON file format) into Snowflake DB by Flattening the same for different functional services.
* Using SageMaker Processing or AWS Lambda, new pipeline components were developed to integrate custom data preprocessing, feature engineering, or model evaluation logic, extending the capability of SageMaker Pipelines.
* Collaborating to build and deploy GCP solutions that are in line with corporate goals with cross-functional teams made up of developers, architects, and operations.
* Chosen and produced information into CSV records and put away them into **AWS S3** by utilizing AWS EC2 and afterward organized and put away in AWS Redshift.
* Developing and writing SQLs and stored procedures in Teradata. Loading data into Snowflake and writing Snow SQL scripts.
* Developed **Glue** Jobs to read data in CSV format in the raw layer and write data to parquet format in the publish layer.
* Scheduled **Glue** jobs along with AWS Lambda functions using step functions.
* Developed a **Glue** job that delete, update, and incremental loads from source to target.
* Extract Real-time feed using **Kafka** and **Spark** Streaming and convert it to RDD and process data in the form of Data Frame and save the data as Parquet format in HDFS.
* Experienced in transferring Streaming data, data from different data sources into **HDFS, No SQL** databases
* Created **ETL Mapping** with Talend Integration Suite to pull data from Source, apply transformations, and load data into the target database.
* Creating Databricks notebooks using SQL, Python, and automated notebooks using jobs.
* Creating Spark clusters and configuring high-concurrency clusters using Azure Databricks to speed up the preparation of high-quality data.
* Used **PySpark and Pandas** to calculate the moving average and RSI score of the stocks and generated them in the data warehouse.
* Worked on design and development of Informatica mappings, workflows to load data into staging area, data warehouse and data marts in **SQLServer** and **Oracle.**
* Developed a PySpark program that writes Dataframes to HDFS as avro files.
* Encoded and decoded JSON objects using PySpark to create and modify the Dataframes in Apache Spark
* Ingested data from RDBMS and performed data transformations, and then exported the transformed data to Cassandra as per the business requirement.
* Developed multiple Kafka Producers and Consumers from scratch as per the software requirement specifications.
* Created on-demand tables on S3 files using **Lambda Functions and AWS Glue using Python and PySpark.**
* Optimized existing algorithms in Hadoop using **Spark Context, Spark-SQL, Data Frames, and Pair RDDs**.
* Developed pipeline for POC to compare performance/efficiency while running pipeline using the **AWS EMR Spark cluster.**
* Extensive experience in building ETL jobs using Jupyter notebooks with Apache Spark.
* Running analytics on power plant data using **Pyspark API**with **Jupyter**notebooks in on-premise cluster for certain transforming needs.
* Implemented Spark using Scala and utilizing Data frames and Spark SQL API for faster processing of data.
* Real-time streaming of the data using Spark with Kafka.
* Designed and developed data loading strategies, and transformation for business to analyze the datasets.
* Experienced in writing **Spark** Applications in **Scala** and **Python** (**Pyspark**).
* implemented design patterns in Python for the application.
* Performing **ETL** testing activities like running the Jobs, Extracting the data using necessary queries from database transform, and upload into the **Data warehouse** servers.
* Develop quality code adhering to Python coding Standards and best practices.
* Used Spark and **Spark-SQL** to read the parquet data and create the tables in hive using Python.
* Implemented **Spark** using Scala and Spark SQL for faster testing and processing of data.
* Implemented Spark using Scala and utilizing **Data frames**and Spark SQL API for faster processing of data.
* Collected data using spark streaming in near-real-time and performed necessary transformations and aggregations to build the data model persists the data in HDFS
* Processed Multiple Data sources input to same Reducer using Generic Writable and Multi-Input format.
* Involved in performance tuning of spark jobs using Cache and using complete advantage of the cluster environment.

**Environment:** AWS, Hadoop, Hive, Java, Kafka, Pyspark, Yarn, DynamoDB, shell Scripting, Python, Spark and Scala, Maven, MySQL, Airflow

**State of SC Jul 2017 to Dec 2020**

**Sr Data Engineer**

**Responsibilities:**

* Experience in Big Data Analytics and design in the **Hadoop ecosystem** using MapReduce Programming, Spark, Hive, Pig, Sqoop, HBase, Oozie, Impala, Kafka
* Performing hive-tuning techniques like partitioning, bucketing, and memory optimization.
* Worked on different file formats like parquet, orc, JSON, and text files.
* Coordinated and automated data transformation and transfer activities across hybrid and multi-cloud systems using Azure Synapse Data Integration.
* Worked on SnowSQL and Snowpipe
* Converted Talend Joblets to support the snowflake functionality.
* Created Snowpipe for continuous data load.
* Monitor and tune ETL processes for performance improvements; identify, research, and resolve data warehouse load issues.
* Created and carried out Spark notebooks and jobs in Azure Synapse Studio to glean patterns and insights from intricate datasets.
* Designed and implemented end-to-end data pipelines using Terraform and AWS services including S3, Redshift, and Glue.
* Proficient in managing GCP cloud infrastructure, including the creation and administration of IAM (Identity and Access Management) policies, virtual machines, storage containers, and networks.
* Enforced data privacy and regulatory compliance by implementing stringent data governance and security measures on GCP, including encryption, access controls, and compliance policies.
* BigQuery was utilized to develop and optimize data analytics solutions on GCP, facilitating the analysis of large datasets for business intelligence and decision-making in a timely and cost-effective manner.
* By incorporating machine learning models and algorithms into data pipelines on Google Cloud Platform (GCP) via services such as TensorFlow and AI Platform, sophisticated predictive capabilities and analytics were enabled.
* Utilizing AWS CodePipeline and AWS CodeBuild, SageMaker Pipelines were integrated with CI/CD pipelines to automate the deployment of machine learning models into production settings.
* SageMaker Model Integrated In order to track model quality and drift in real-time, monitor into machine learning pipelines. When performance deviations are found, alarms are sent out and workflows are retrained.
* Automating the provisioning and maintenance of GCP resources through the use of technologies such as Terraform or Deployment Manager to implement Infrastructure as Code (IaC).
* Re-platforming, re-architecting, or lift-and-shift tactics to optimize workloads for cloud environments.
* Employing GitLab CI/CD, Cloud Build, Jenkins, and other tools to create and manage CI/CD pipelines on GCP.
* Worked on migrating MapReduce programs into Spark transformations using **Spark** and **Scala**, initially done using Python (PySpark).
* Data collection and transformation mappings and design of the data warehouse data model.
* Used **spark SQL** to load data and created schema RDD on top of that which loads into **hive** tables and handled structured using spark SQL.
* Automated infrastructure provisioning and configuration using Terraform, reducing deployment time and improving overall system reliability.
* Worked on analyzing **Hadoop cluster** using different big data analytic tools including Flume, Pig, Hive, HBase, Oozie, Zookeeper, Sqoop, Spark, and Kafka.
* Define virtual warehouse sizing for Snowflake for different type of workloads.
* Clear and thorough documentation for team members and stakeholders was ensured by employing Markdown or AWS CloudFormation templates to define pipeline setups, dependencies, and execution routines.
* Developed projects with **GCP** **BigQuery** and **Airflow** as schedulers.
* Created a Hadoop instance in the **DataProc** and developed ETL pipelines using Spark Scala and Hive.
* Loading data in Kafka every 15 minutes on an incremental basis to **BigQuery** raw using Google **DataProc**, GCS bucket, Hive, Spark, Scala, Python, Gsutil, and shell script.
* Monitor and tune ETL processes for performance improvements; identify, research, and resolve data warehouse load issues.
* Using rest API with Python to ingest data to from the Dashboard system to **Bigquery**.
* As a Big Data Developer implemented solutions for ingesting data from various sources and processing the Data-at-Rest utilizing Big Data technologies such as Hadoop, MapReduce Frameworks, MongoDB, Hive, Oozie, Flume, Sqoop Talend, etc.
* Explored with **Spark** improving the performance and optimization of the existing algorithms in Hadoop using **Spark Context, Spark -SQL, Data Frame, Pair RDDs, Spark, YARN, and pyspark.**
* Used Oozie workflow engine to manage interdependent Hadoop jobs and to automate several types of Hadoop jobs such as Java map-reduce Hive, **Pig**, and **Sqoop**.
* The Databricks platform follows best practices for securing network access to cloud applications.
* Hands-on experience with **git bash commands** like git pull to pull the code from the source and develop it as per the requirements, git add to add files, git commit after the code build and git push to the pre-prod environment for the code review and later used screwdriver. Yaml which builds the code, generates artifacts which released into production
* Performed data validation which does the record-wise counts between the source and destination.
* Involved in the data support team as the role of bug fixes, schedule changes, memory tuning, schema changes loading the historic data.
* Worked on implementation of some checkpoints like hive count check, Sqoop records check, done file create check, done file check, and touch file lookup.
* Worked on both **Agile** methodologies

**ENVIRONMENT**: Hadoop, Map Reduce, HDFS, Hive, Cassandra, Sqoop, Oozie, SQL, Kafka, Spark, Scala, Java, GitHub, Teradata Big Data Integration, Impala, GCP, Data proc, Airflow, Big query.

**M&T Bank, New York, NY May 2014 to Jun 2017**

**Data Engineer**

**Responsibilities:**

* Worked extensively on Hadoop Components such as **HDFS, Job Tracker, Task Tracker, Name Node, Data Node, YARN, Spark and Map Reduce programming.**
* Converting the existing relational database model to the **Hadoop ecosystem**.
* Worked with Linux systems and RDBMS database regularly to ingest data using Sqoop.
* Strong experience in working with **ELASTIC MAPREDUCE** and setting up environments on Amazon AWS EC2 instances.
* Performing ETL testing activities like running the Jobs, Extracting the data using necessary queries from database transform, and upload into the Data warehouse servers.
* Documented architecture designs, data models, and deployment configurations within Azure Synapse workspaces, ensuring clear and comprehensive documentation for stakeholders and team members.
* Ability to spin up different **AWS instances** including EC2-classic and EC2-VPC using cloud formation templates.
* Collected data using **Spark Streaming** from AWS S3 bucket in near-real-time and performs necessary Transformations and Aggregations to build the data model and persist the data in **HDFS.**
* Managed and reviewed **Hadoop and HBase log files**.
* Worked extensively with importing metadata into Hive and migrated existing tables and applications to work on Hive.
* Designed and implemented **HIVE queries** and functions for evaluation, filtering, loading, and storing of data.
* Analyze table data and implement compression techniques like Teradata Multivalued compression.
* Involved in the ETL process from design, development, testing, and migration to production environments.
* Involved in writing the **ETL test scripts** and guided the testing team in executing the test scripts.
* Involved in performance tuning of the ETL process by addressing various performance issues at the extraction and transformation stages.
* Guide the development team working on **PySpark** as an ETL platform.
* Writing Hadoop MapReduce jobs to run on Amazon EMR clusters and creating workflows for running jobs.
* Generating analytics reporting on probe data by writing EMR (elastic map reduce) jobs to run on Amazon VPC cluster and using Amazon data pipelines for automation.
* Have a good understanding of **Teradata MPP** architecture such as Partitioning, Primary Indexes,
* Good knowledge of Teradata Unity, Teradata Data Mover, OS PDE Kernel internals, Backup and Recovery
* Created HBase tables to store variable data formats of data coming from different portfolios.
* Optimize the **Pyspark jobs** to run on Kubernetes Cluster for faster data processing.
* Created Partitions and buckets based on State to further process using Bucket-based Hive joins.
* Involved in transforming data from Mainframe tables to **HDFS**, and **HBase tables using Sqoop.**
* Creating Hive tables and working on them using **HiveQL**.
* Creating and truncating HBase tables in hue and taking backup of submitter ID.
* Developed data pipeline using Kafka to store data in HDFS.
* Used Spark API over Hadoop YARN as an execution engine for data analytics using Hive.
* Continuous monitoring and managing of the **Hadoop cluster through Cloudera Manager**.
* Involved in the review of **functional and non-functional** requirements.
* Developed ETL Process using **HIVE** and **HBASE**.
* Prepared the Technical Specification document for the **ETL job development.**
* Responsible for managing data coming from different sources.
* Loaded the CDRs from relational DB using **Sqoop** and other sources to the Hadoop cluster by using **Flume**.
* Installed and configured **Apache Hadoop, Hive, and Pig** environment.

**Environment**: Hadoop, HDFS, Pig, Hive, Java, Flume, Sqoop, Oozie, Python, Shell Scripting, SQL Talend, Spark, HBase, Elastic search, Linux- Ubuntu, Kafka.

**Cuspyd, Hyderabad, India (Intern ) Jun 2012 to Jul 2013**

**Hadoop Developer**

**Responsibilities:**

* Worked extensively on Hadoop Components such as **HDFS, Job Tracker, Task Tracker, Name Node, Data Node, YARN, Spark and Map Reduce programming.**
* Converting the existing relational database model to the **Hadoop ecosystem**.
* Worked on development of data ingestion pipelines using ETL tool, Talend &amp; bash scripting with big
* data technologies including but not limited to Hive, Impala, Spark, Kafka, and Talend.
* Experience in developing scalable &amp; secure data pipelines for large datasets.
* Gathered requirements for ingestion of new data sources including life cycle, data quality check,
* transformations, and metadata enrichment.
* Collecting and aggregating large amounts of log data and staging data in HDFS for further analysis.
* Monitor the Daily, Weekly, and Monthly jobs and provide support in case of failures/issues.
* Delivered data engineer services like data exploration, ad-hoc ingestions, and subject-matter-expertise to Data scientists in using big data technologies.
* Build machine learning models to showcase big data capabilities using PySpark and MLlib.
* Knowledge of implementing the JILs to automate the jobs in the production cluster.
* Troubleshooted user&#39; s analyses bugs (JIRA and IRIS Ticket).
* Worked with the SCRUM team in delivering agreed user stories on time for every Sprint.
* Worked on analyzing and resolving production job failures in several scenarios.
* Implemented UNIX scripts to define the use case workflow to process the data files and automate the jobs.
* Utilizing Terraform to automate infrastructure provisioning and setup for data processing pipelines, consistency and scalability across cloud environments and platforms are guaranteed.
* Used Terraform to manage infrastructure as code simplifies the deployment and upkeep of data engineering infrastructure, allowing for version control, cooperation, and repeatability while cutting down on operational overhead.

**Environment:** Spark, Redshift, Python, Java, HDFS, Hive, Pig, Scala, Kafka, Shell scripting, Linux, Jenkins,

Eclipse, Git, Oozie, Talend.