Bharath Gangula

605-480-5104

Bharathreddygangula1@gmail.com

PROFESSIONAL SUMMARY:

- Data Engineer with 5 years of experience.
- Data experience includes ingestion, pipeline development, pipeline automation, integrations, ETL, and migrations.
- Over 4 years of experience with Airflow, Spark, Scala, Python, and PySpark.
- Orchestrated data pipelines using Apache Airflow to interact with services like Azure Databricks, Azure Data Factory, Azure Data Lake, and Azure Synapse Analytics.
- Developed Spark applications using PySpark and Spark-SQL for data transformation, and aggregation from multiple file formats for analyzing & transforming the data to uncover insights into the customer usage patterns.
- Worked on Big Data integration and Analytics based on Spark, Hive, PostgreSQL, Snowflake, and MongoDB.
- Other experience includes Hadoop, Azure, AWS, Databricks, Azure SQL, and Testing
- Performed ETL operations on the business data and developed a spark pipeline that efficiently executes ETL activities.
- Experience in Setting up the build and deployment automation for Terraform scripts using Jenkins
- Knowledgeable with Reporting tools such as Power BI, Data Studio and Tableau.

Linkedin:

- **https://www.linkedin.com/in/bharathsimha-reddy-013763264/**

TECHNICAL SKILLS:

AWS: Amazon EC2, Amazon S3, Amazon Simple DB, Amazon MQ, Amazon ECS, Amazon Lambda, Amazon RDS, Amazon Elastic Load Balancing, Amazon SQS, IAM, AWS Cloud Watch, Amazon EBS, AWS Glue
Azure Cloud Services (PaaS & IaaS): Azure Blob Storage, Azure Monitoring, Azure Search, Data Factory, Azure SQL, Azure Analysis Services, Azure Synapse Analytics (DW), Azure Data Lake, Azure Active Directory
Hadoop/Big Data Technologies: Hadoop, Map Reduce, Pig, Sqoop, Hive, Oozie, Spark, Zookeeper, Flume
ETL Tools: Oracle Data Integrator (ODI), Informatica, Azure Data Factory
Hadoop Distribution: Horton Works, Cloudera
Programming & Scripting: Python, Scala, SQL, Shell Scripting
Databases: Oracle, SQL Server, MySQL, HBase, Mongo DB, Redshift, Snowflake

PROFESSIONAL EXPERIENCE:

Madison Cyber Labs - Madison,SD **08/2021 – Present**
Data Engineer/Data Scientist

- The project involved building data pipelines in data integration by extracting large sets of data from numerous internal and external data sources. Also, involved in hosting the data in a Data warehouse using Azure Data Factory (ADF), PySpark and transforming data into MS Azure Data Lake.
- Gathered requirements for Analysis, Design, Development, testing, and implementation of business rules.
- Migrated data from on-prem SQL Database to Azure Synapse Analytics using Azure Data Factory.
- Ingested huge volume and variety of data from disparate source systems into Azure Data Lake using Azure Data Factory.
- Performed ETL operations for Data cleansing, Filtering, Standardizing, Mapping, and Transforming of Extracted data from multiple sources such as Azure Data Lake, and on-prem SQL DB.
- Applied Python scripting to enhance ETL processes, resulting in improved data accuracy and reduced processing times.
- Utilized Python in conjunction with Linux environments to orchestrate complex data workflows
- Applied various ADF dataflow transformations such as Data Conversion, Conditional Split, Derived Column, Lookup, join, Union, Aggregate, pivot, and filter and performed data flow transformation using the data flow activity.
- Experience in Linux system administration, including server setup, maintenance, and troubleshooting.

- Developed Spark applications using PySpark and Spark-SQL for data transformation, and aggregation from multiple file formats for analyzing & transforming the data to uncover insights into the customer usage patterns.
- Worked on predicting the cluster size, monitoring, and troubleshooting the Spark data bricks cluster.
- Worked on data ingestion, transformation, and loading processes using Snowflake's SQL capabilities and integration with other data tools
- Worked on leveraging Snowflake's features like clustering, materialized views, and query optimization techniques to enhance data processing efficiency.
- Worked on automating and validating the created data-driven workflows extracted from the ADF using Apache Airflow.
- Orchestrated data pipelines using Apache Airflow to interact with services like Azure Databricks, Azure Data Factory, Azure Data Lake, and Azure Synapse Analytics.
- Used ADF as an orchestration tool for integrating data from upstream to downstream systems.
- Worked on maintaining and tracking the changes using version control tools like SVN and GIT.
- Designed and implemented CI/CD pipelines using Jenkins and GitLab CI/CD, enabling continuous integration, testing, and deployment.
- Expertise in package management, system monitoring, and security configurations..
- Worked on Ansible for infrastructure automation and configuration management.
- Written playbooks, inventory management, application deployment, and application management across several servers. strong familiarity with Ansible recommended practices, and effective use of Ansible solutions
- Worked with building data warehouse structures, and creating facts, dimensions, and aggregate tables, by dimensional modeling, and Star and Snowflake schemas.
- Worked on defining the CI/CD process and support test automation framework in the cloud as part of the build engineering team.

Environment: Azure Synapse Analytics, SQL Database, Azure Data Lake Storage (ADLS), and Azure Data Factory, SQL Database, Azure Synapse Analytics, Azure Data Factory, Teradata, HDFS, Sqoop, Azure Data Lake, Azure Data Factory, ETL, SQL DB, Oracle, SQL Server, Teradata, Azure Data Share, PySpark with Databricks, Apache Airflow

Innodatatics - Hyderabad, India **02/2019 – 07/2021**
Data Engineer/Data Analyst

- The aim of the project is to build the enterprise-scale data platform used by data scientists/analysts to accelerate the impact of data-driven insights. The project provides essential tools and frameworks to support the entire analytics lifecycle, from efficient data ingestion and transformation workflows. The technology stack used in this project are Spark, AWS Cloud, Hadoop technologies and Python.
- Used Hadoop ecosystem with AWS EMR to build a scalable distributed data system.
- Worked with Python, Hive to create simple to complicated jobs.
- Implemented the Apache Spark data processing module to handle data from multiple RDBMS and Streaming sources, then used Python to compile Apache Spark applications.
- Developed applications using Scala and functional programming concepts. I am well-versed in Scala syntax, collections, and functional programming concepts such as higher-order functions, currying, and monads
- Used Sqoop, as an ETL component, was used to extract data from MySQL and load it into HDFS.
- Performed ETL operations on the business data and developed a spark pipeline that efficiently executes ETL activities.
- Wrote Hive scripts to analyze customer behavior data.
- Worked with EMR and environment setup on Amazon AWS EC2 instances for pipelines in AWS.
- Provisioned the highly available EC2 Instances using Terraform and cloud formation and wrote new plugins to support new functionality in Terraform.
- Created DAGs in Airflow to automate the process using Python schedule jobs.
- Worked extensively with AWS S3 buckets and was involved in file transfers between HDFS and AWS S3.
- Loaded data into Amazon Redshift and utilized AWS Cloud Watch to capture and monitor AWS RDS instances within the environment.
- Developed and implemented a migration strategy from an Oracle platform to AWS Redshift for the Data Warehouse.
- Worked on Big Data integration and Analytics based on Spark, Hive, PostgreSQL, Snowflake, and MongoDB.
- Created diff types of Redshift DB tables such as physical/Temp/Deep copy.
- Worked on Vacuum types and Analyze in Redshift.
- Developed the PySpark code for EMR and AWS Glue tasks.
- Imported data from several sources, transformed with Spark, then loaded into Hive.
- Worked with the Spark Core, Spark Streaming, and Spark SQL modules.

- Utilized Cloud watch logs to move application logs to S3 and created alarms based on exceptions.
- Loaded data from AWS S3 to Redshift using Glue.
- Worked on Production support for the EMR cluster, mainly troubleshooting memory and spark job application issues.
- Developed AWS Lambda function to monitor the EMR cluster status updates and the jobs.
- Expertise in establishing an optimum data integration platform that can handle growing data volumes.
- Used Sqoop to export the analyzed data to relational databases for visualization and report generation by our BI team.
- While researching Spark's modules, worked with Data Frames, RDD, and Spark Context.
- Worked with PySpark for using Spark's Library's by scripting in Python to analyze data.

Environment: AWS EMR, S3, EC2, Redshift, Glue, Lambda, Hadoop, Hive, Sqoop, Zookeeper, Spark, Kinesis, PySpark, Spark Core, Teradata, Python, SQL, Splunk, Snowflake, MongoDB, Oracle, MySQL, Tableau, Jira

360Digitmg – Hyderabad,India **06/2018 – 01/2019**
Data Engineer/Data Analyst

- The goal of the project is to support and implement high performance and data-centric solutions using comprehensive big data capabilities for the company's data platform environment. The project is implemented using Bigdata technologies and Azure Cloud and AWS Cloud.
- Implemented the Snow SQL and Snow pipe for continuous data load from csv files to snowflake.
- Redesigned the views from SQL server to snowflake by replacing the snowflake functions with SQL server functions in snowflake.
- Designed the data marts in data modeling using the star schema and snowflake schema.
- Worked on cloud technologies with AWS and Azure which will be used for stages to load data into Snowflake.
- Worked on loading the source files from local to AWS S3 and then loaded to snowflake from S3 as external storage.
- Developed complex views in snowflake based on the requirements given by client and worked on exploring multiple tables/views to create the view.
- Implemented the cloning, data retention periods, and the fail-safe mechanism in Snowflake to recover the data.
- Loaded the data from Azure Blob Storage as an External storage to snowflake which includes continuous loading of data.
- Primarily involved in Data Migration using SQL, SQL Azure, Azure Storage, Azure Data Factory, SSIS, PowerShell.
- Created and selected the virtual warehouses in Snowflake based on the query load performance.
- Used Spark-Streaming APIs to perform necessary transformations and actions on the data got from Kafka.
- Worked on analyzing Hadoop clusters and different big data analytic tools including Pig, Hive.
- Migrated MapReduce jobs into Spark jobs and used Spark SQL and Data frames API to load structured data into Spark clusters.
- Configured the snow pipe in snowflake environment using AWS by configuring SQS events in order to trigger Snow pipe.
- Performed loading into snowflake by bulk loading using COPY.
- Worked with snowflake cloud data warehouse and AWS S3 bucket for integrating data from multiple source system which include loading JSON Formatted data into snowflake table.
- Automated the SQL Scripts in Snowflake using the Tasks.
- Wrote UDF's and Stored Procedures to load the data from one phase to another phase in Snowflake.

Environment: Azure (Blob Storage, Data Lake, Databricks, SQL Azure, Azure Data Factory), AWS (EC2, S3, SQS), Snowflake, Data Warehousing, Hadoop, Pig, Hive, MapReduce, Spark, Spark SQL, Power BI, Jira, SQL Server, GitLab, SQL, SSIS, PowerShell

EDUCATION

Dakota State University – Madison, SD. Dec 2022
Master of Science in Analytics (M.S.)

SREC University – India. May 2018
Bachelor of Technology in Mechanical Engineering (B.Tech.)