Name: **Gayathri Mattaparthi**
Contact: 269-753-9922
Mail ID: ggayi284@gmail.com
LinkedIn: http://linkedin.com/in/gayathri-gayi

Professional Summary:

- Around 6 years of IT experience as a Data Engineer with expertise in Big Data technologies like Spark, MapReduce, Hive, Kafka, and HDFS. Proficient in programming languages such as Python, Scala, and Java for developing distributed data solutions, analytical applications, and ETL pipelines leveraging the big data ecosystem components and AWS/Azure cloud services.
- Experience in working with Apache Hadoop open-source distribution, including technologies like HDFS, MapReduce, Python, Pig, Hive, HBase, SQOOP, Oozie, Zookeeper, Spark, Spark-Streaming, Storm, Kafka, Cassandra, Impala, Snappy, Greenplum, and MongoDB.
- Skilled in managing all stages of the Workday project lifecycle, including requirement gathering, planning, design, documentation, coding, testing, and deployment.
- Proficient in utilizing Cloudera and Hortonworks distributions for data engineering projects.
- Hands-on experience with Azure Cloud services such as Azure Data Factory, Azure Data Lake Storage, Azure Synapse Analytics, Azure Analytical services, Azure Cosmos, NOSQL DB, Azure HDInsight, and Data Bricks.
- Familiarity with Amazon Web Services (AWS) Cloud Platform, including services like EC2, S3, VPC, ELB, IAM, DynamoDB, CloudFront, CloudWatch, Route 53, Elastic Beanstalk, Auto Scaling, Redshift, CloudFormation, Kinesis, SQS, SNS, SES, Lambda, and EMR.
- Good understanding of Informatica tools like Snap logic, Informatica, and Ab Initio, and supporting ETL teams with mapping documents and transformations.
- Excellent understanding of Hadoop architecture and its components such as HDFS, Job Tracker, Task Tracker, Name Node, Data Node, and MapReduce programming paradigm.
- Proficient in designing star schema, Snowflake schema for Data Warehouse, and ODS architecture.
- Experience in Data Migration from Teradata to AWS Snowflake Environment using Python and BI tools like Alteryx.
- Extensive knowledge of developing Spark Streaming jobs using Scala, PySpark, and Spark-Shell.
- Expertise in ETL using Spark-SQL, Spark Core, and real-time data processing using Spark Streaming.
- Experience in converting SQL Server stored procedures in old legacy systems into ETL jobs using Data stage 11.5.
- Strong experience working with various tools and technologies like Data stage, SAP Data Services, Teradata, SAP, SQL Server, and UNIX.
- Proficient in working with various file formats like Avro, Parquet, Orc, JSON, and CSV.
- Experience in developing customized UDFs in Python to extend Hive and Pig Latin functionality.
- Extensively worked with Teradata utilities like Fast Export and Multiload to export and load data to/from different source systems.

□

- Proficient in ETL processes to Extract, Transform, and Load source data into respective target tables to build Data Marts.
- Experienced in building Automation Regression Scripts for validation of ETL processes between multiple databases like Oracle, SQL Server, Hive, and MongoDB using Python.

Developed Spark Applications using Scala, Java, and implemented Apache Spark data processing projects to handle data from various RDBMS and Streaming sources. Utilized different Spark Modules like Spark Core, Spark RDDs, Spark Data Frames, and Spark SQL.

Technical Skills:

Languages	Shell scripting, SQL, PL/SQL, Python, R, PySpark, Pig, Hive QL, Scala.
Python Packages	NumPy, Pandas, Matplotlib.
Big Data Ecosystem	HDFS, MapReduce, Hive, Pig, Sqoop, Flume, Oozie, Zookeeper, Kafka, Apache Spark, Spark Streaming, HBase, Flume, Impala.
Databases	Oracle 10g/11g/12c, SQL Server, MySQL, Teradata, PostgreSQL, MS Access, Snowflake, NoSQL Database (HBase, MongoDB).
Cloud Technologies	Amazon Web Services (AWS), Microsoft Azure
Version Control	GIT, GIT HUB.
IDE & Tools, Design	Eclipse, Visual Studio, Net Beans, Junit, CI/CD, SQL Developer, MySQL, SQL Developer, Workbench, Tableau
Operating Systems	Windows 98, 2000, XP, Windows 7,10, Mac OS, Unix, Linux
Data Engineer/Big Data Tools/Cloud/ETL/Visualization/Other Tools	Data bricks, Hadoop Distributed File System (HDFS), Hive, Pig, Sqoop, MapReduce, Flume, YARN, Oozie, Zookeeper, etc. AWS, Azure Data bricks, Azure Data Explorer, Azure, Linux, Bash Shell, Unix, etc., Tableau, Power BI.

Education Details:

- Master’s in data science, Western Michigan University – Dec 2023.
- Bachelors in computer science & engineering, at JNTU AP in the year 2019.

Certifications:

- Microsoft Certified: Azure Data Engineer Associate.
- AWS Certified: Solution Architect Associate.

Professional Experience:

Client: Facebook, Columbus, OH **Feb 2023 – Dec 2023**

Role: Data Engineer (Azure)

Responsibilities:

- Designed and developed ETL pipelines in Azure Databricks to ingest and process real-time streaming data from Kafka, loaded with IoT devices.

□

- Developed complex transformation logic using Python and PySpark to handle nested JSON files.
- Implemented incremental data loading solutions using Auto Loader in Azure Databricks.
- Designed Medallion architecture in Databricks Lakehouse implementation, including efficient data models for silver and gold layers, and various performance optimization techniques for delta tables such as Zordering, Optimize, Auto-optimize, etc.
- Implemented cost optimization solutions using all-purpose and job clusters.
- Expert in troubleshooting and debugging Spark applications using Spark UI.

Implemented PoC solutions using advanced concepts such as Delta Live Tables, Unity Catalogue, etc., in Databricks.

- Installed Hadoop, Cassandra, MapReduce, and HDFS, and developed multiple MapReduce jobs in Pig and Hive for data.
- Installed and configured Hadoop ecosystem components like HBase, Flume, Pig, and Sqoop.
- Created parameterized ETL pipelines in Azure Synapse for migrating data from on-premises source systems like Oracle, SQL, and File Host to Delta Lake tables in ADLS Gen2.
- Automated pipelines using a metadata-driven framework to load multiple sources of tables and files in a single master pipeline using Azure Data Factory.
- Developed complex ETL pipelines to handle Slowly Changing Dimension (SCD) Type 1 and Type 2 using Dataflow in Azure Data Factory.
- Created event-based and schedule triggers to execute pipelines in Synapse automatically based on given conditions.
- Converted Hive/SQL queries into Spark transformations using Spark RDDs and Scala.
- Created secrets in key vaults and parameterized linked service connections to protect connection details of source systems.

□

- Proficient in performance optimization across ADF and Synapse Pipelines, such as DIU, Degree of parallelism in Copy activity, batch count in for each activity, etc.
- Designed scalable and efficient data warehouse ETL Pipelines in Snowflake, ensuring optimal performance and ease of maintenance.
- Designed and developed Power BI visuals to present data compellingly and understandably, enhancing datadriven decision-making.
- Strong knowledge of Data modeling. Have implemented Denormalized and normalized data models across various data solutions.
- Strong exposure to advanced SQL concepts like window functions, CTE, complex joins, stored procedures, etc.
- Developed and implemented CI/CD pipelines using Azure DevOps and GitHub to automate deployments.
- Environment: Azure Data bricks, Azure Data Factory, Azure Data Lake Storage Gen2, Azure SQL, Azure Synapse Analytics, Teradata, Informatica, PySpark, SQL, Python, Oracle PLSQL, PowerShell Scripts, Azure DevOps, Microsoft SQL Server.

Client: MasterCard, St Louis, MO **Aug 2022 – Jan 2023**
Role: Data Engineer (Azure)
<u>Responsibilities:</u>

- Created pipelines using Azure Synapse Studio to pull the M365 Data and transformed the data using Azure □ Notebook and loaded it to Cosmos DB.
- Performed transformations using Pyspark and loaded the data into Cosmos DB to make it available to Analysts.

□

- Defined Roles and policies based on the business requirements.
- Assisted senior resources in the assessment, analysis, and implementation of Microsoft Graph Data Connect
- Acted as a technical lead for the project team and coordinated with different stakeholders including the platform.
- Engineering, operations, data science &amp; Microsoft technical support to help implement the solution.
- Analyzed the Workday data with Microsoft Viva Insights and created visualizations using Power BI and helped.
- Employees to improve their focus time, wellbeing, and Productivity.

Developed ETL process for the Workday data using Python and loaded the data into HIVE and SQL for the analysts.

- Created Hive and SQL views for the Workday and Beeline data.
- Generated alerts with the list of the impacted Objects on file load failure.
- Created bash script to automate the ETL process for Workday.
- Created scripts to generate DDL scripts for hive and SQL.
- Generated a Data Lineage to read the workday logs and get insights of the data ingestion to display the lineage graphically.

Client: Accenture, India **Oct 2019 – Dec 2021**
Role: Data Engineer (AWS)
Responsibilities:

- Extensively used AWS Athena to import structured data from S3 into other systems or to generate reports.
- Building use cases in Snowflake by bringing various sources using Attunity.
- Worked with Spark to improve the speed and optimization of Hadoop's current algorithms.
- Migrated an existing on-premises application to AWS, utilizing services like EC2 and S3 for data processing and storage. • Developed Scala scripts and UDFs in Spark for data aggregation, queries, and writing data back into RDBMS through Sqoop.
- Utilized Spark-Streaming APIs for real-time changes to the common learner data model from Kinesis.
- Performed end-to-end architecture and implementation evaluations of different AWS services such as Amazon EMR, Redshift, S3, Athena, Glue, and Kinesis.
- Created Apache Presto and Apache Drill configurations on an AWS EMR cluster to integrate different databases such as MySQL and Hive.
- Implemented data integration using Attuity to bring data from different sources to Snowflake.
- Developed end-to-end data lake solutions in Hadoop and Snowflake.
- Developed Spark jobs using Scala for faster data processing in a test environment.
- Developed multiple POCs using Scala and deployed them on the Yarn cluster to compare performance with Hive and SQL.
- Involved in data extraction, transformation, and loading (ETL) using Informatica PowerCenter.
- Designed and maintained Informatica PowerCenter mappings for extraction, transformation, and loading between Oracle and Teradata.
- Consulted on Snowflake Data Platform Solution Architecture, Design, Development, and Deployment.
- Developed and implemented ETL pipelines on S3 parquet files in a data lake using AWS Glue.
- Implemented SQL Alchemy for complete access to SQL.
- Dealt with Python OpenStack APIs and used Python scripts for database updates and file manipulation.
- Developed Spark scripts to perform ETL using Glue jobs, extracting data from S3 using a crawler, and creating a data catalog to store metadata.
- Designed and developed modules in Python deployed in AWS Glue using Spark library and Python.

□

- Worked extensively with the Teradata database, creating pipelines to load data into the EDW.
- Created Python notebooks on Azure Databricks for processing datasets and loading them into Azure SQL Databases.
- Worked on ETL migration services by creating and deploying AWS Lambda functions for serverless data pipelines.
- Environment: Python, Java, Data bricks, PySpark, Kafka, AWS S3, Delta Lake, Snowflake, Cloudera CDH, Hive, Impala, Kubernetes, Flume, Apache Nifi, Shell scripting, SQL, Sqoop, Oozie, Oracle, SQL Server, HBase, Power BI, Agile Methodology.

Client: Value Labs, India **Jan 2018 – Sep 2019**
Role: Data Engineer Trainee
Responsibilities:

- Learn and understand the fundamentals of data engineering and big data technologies.
- Assist in the implementation of ETL processes to extract, transform, and load data.
- Gain hands-on experience with tools such as Apache NiFi or AWS Glue for data integration.
- Collaborate with team members to design and develop basic data pipelines.
- Acquire knowledge in managing and processing data using Hadoop Distributed File System (HDFS).
- Work on data analysis tasks using tools like Hive, Pig, and Sqoop within the Hadoop ecosystem.
- Develop basic MapReduce programs for data processing and analysis.
- Support the creation and maintenance of data warehouses and databases.
- Participate in troubleshooting data-related issues and contribute to solutions.
- Engage in learning sessions to enhance skills in data visualization using tools like Tableau or Power BI.

Note: References will be provided on request