SWETHA REDDY

Email: swethareddyy2017@gmail.com **Contact Number:** +1 334-575-2603 **Location**: Montgomery Alabama.

SUMMARY OF EXPERIENCE

- Data Engineer with over 4+ years of experience in leveraging statistical analysis, machine learning, and data visualization techniques to extract insights and drive business decisions.
- Proven track record of effectively communicating complex insights derived from PowerBI and Tableau visualizations to non-technical stakeholders.
- Working knowledge of Hadoop, Pig, Hive, HDFS, MapReduce, Sqoop, Storm, Spark, Airflow, Snowflake, Teradata, Flume, Kafka, Yarn, Oozie, and Zookeeper ecosystems.
- In-depth knowledge of Map Reduce and Hadoop Infrastructure; extensive exposure to Data technologies and the Hadoop ecosystem.
- Practical knowledge of Databricks Workspace User Interface, Managing Databricks Notebooks, and Unified Data Analytics with Databricks, as well as Delta Lake with Spark SQL and Python.
- Give the development team using Pyspark as an ETL platform direction. ensures the definition and fulfilment of quality standards. For quicker data processing, optimize the Pyspark jobs for the Kubernetes cluster.
- Data Bricks Workspace for Business Analytics, Cluster Management in Data Bricks, Managing the Machine Learning Lifecycle, and Setting Up AWS and Microsoft Azure with Data Bricks
- Proficient in utilizing MapReduce, Spark, and Hive to write end-to-end data processing jobs for data analysis.
- Proficiency with Info Park MLLib and exposure to the Apache Spark ecosystem, including Spark-Core, SQL, Data Frames, and RDD. Thorough knowledge of structured streaming and Databricks in Spark Architecture.
- Continuous learner in both PowerBI and Tableau, staying updated with the latest features and best practices to optimize data visualization and analysis workflows.
- Has experience working with Python libraries like NumPy, SciPy, and Pandas for data analysis and numerical calculations, as well as loading and extracting data using Python.
- Strong background and comprehension in the development and implementation of extensive data and analytics solutions using Snowflake Data Warehouse.

TECHNICAL SKILLS:

- **Fundamentals:** Machine Learning Algorithms, Exploratory Data Analysis, A/B Testing, Time Series Analysis
- **Programming and Markup Languages:** Python, Java, R, C++, Scala, Unix shell script, Cobol, SQL, and, PL/SQL, JavaScript, TypeScript, JSON, MATLAB, PowerBI and Tableau
- **Frameworks:** Pandas, NumPy, Scikit-learn, TensorFlow, KERAS, Flask, React, Flask, Apache Spark, Apache Flink, Kafka
- **Hadoop:** HDFS, Hive, Pig, Sqoop, Yarn, Spark, SQL, Kafka, Horton work and Cloudera Hadoop
- **Technologies and Tools:** Git, Docker, Amazon Web Service (AWS), Tableau, Google Cloud Platform, Kubernetes, Power BI, Apache Airflow, Talend, Amazon Redshift, Google Big Query
- **Hadoop/Spark Ecosystem:** Hadoop, MapReduce, Pig, Hive/impala, YARN, Kafka, Flume, Oozie, Zookeeper, Spark, Airflow, MongoDB, Cassandra, HBase, and Storm.
- **Hadoop Distribution:** Cloudera distribution and Horton works.
- **Programming Languages:** Scala, Hibernate, JDBC, JSON, HTML, CSS, SQL, R, Shell Scripting
- **Script Languages:** JavaScript, jQuery, Python.
- **Databases:** Oracle, SQL Server, MySQL, Cassandra, Teradata, PostgreSQL, MS Access, Snowflake, NoSQL, Database (HBase, MongoDB).
- **Operating Systems:** Linux, Windows, Ubuntu, Unix
- **Web/Application server:** Apache Tomcat, WebLogic, WebSphere Tools Eclipse, NetBeans
- **Data Visualization Tools:** Tableau, Power BI, SAS, Excel, ETL
- **OLAP/Reporting:** SQL Server Analysis Services and Reporting Services.
- **Cloud Technologies:** MS Azure, Amazon Web Services (AWS).
- **Machine Learning Models:** Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbor (KNN), Principal Component Analysis, Linear Regression, Naïve Bayes.

PROFESSIONAL EXPERIENCE:

Dell Technologies **Jan 2023 to Present**
Data Engineer

- The individual has extensive experience in requirements gathering, business analysis, design and development, testing, and implementation of business rules.
- Expertise in developing Spark applications using Spark-SQL in Databricks for data extraction, transformation, and aggregation from multiple file formats.
- Proficient in leveraging PowerBI to create interactive dashboards and reports, enabling data-driven decision-making across the organization.
- Experienced in designing visually compelling visualizations using Tableau, transforming complex data into actionable insights for stakeholders.

- Experience in extracting, transforming, and loading data from sources systems to Azure Data Storage services using a combination of Azure Data Factory, T-SQL, Spark SQL, and U-SQL Azure Data Lake Analytics.
- The individual has also developed ETL integration patterns using Python on Spark and created a framework for converting existing PowerCenter mappings to Pyspark jobs.
- Translated business requirements into maintainable software components and understand their impact on both technical and business aspects.
- The individual has designed and developed ETL pipelines in Azure cloud, orchestrated all data pipelines using Azure Data Factory, and created custom alerts platforms for monitoring.
- Also created Databricks Job workflows that extract data from SQL servers and upload files to SFTP using Python and Python. The individual has been involved in the full lifecycle of projects, including requirement gathering, system designing, application development, enhancement, deployment, maintenance, and support.
- Developed data extraction, transformation, and loading jobs from various sources into Teradata using BTEQ, Fast Load, Fast Export, Multi Load, and stored procedures.
- Worked on Informatica Advanced concepts and implemented Push down Optimization technology and pipeline partitioning. Skilled in data extraction, transformation, and loading (ETL) processes within PowerBI, ensuring accurate and reliable data analysis.
- Demonstrated ability to develop advanced calculations and custom visualizations in Tableau, enhancing data exploration and analysis capabilities.
- Performed bulk data load from multiple data sources to Teradata RDBMS using BTEQ, Multi Load, and Fast Load. They have also designed, created, and tuned physical database objects to support normalized and dimensional models.
- Responsible for performance monitoring, resource and priority management, space management, user management, index management, access control, and execute disaster recovery procedures.
- Python and Shell scripts to automate Teradata ELT and Admin activities, performed application-level DBA activities, and developed UNIX scripts for the loading process.

Environment: Spark-Streaming, Hive, Scala, Hadoop, Kafka, Spark, Sqoop, Docker, Spark SQL, TDD, pig, NoSQL, Impala, Oozie, HBase, PowerBI and Tableau Data Lake, Zookeeper, Azure, Unix/Linux Shell Scripting, Python, PyCharm, Informatica, Informatica PowerCenter, Linux, Shell Scripting.

Trigent **Feb 2018 to Jun 2021**
Data Engineer

- Involved extensively in setting up and installing the Cloudera Hadoop distribution.
- Using in-memory computing capabilities such as Apache Spark built in Scala, advanced procedures such as text analytics and processing were implemented. Possess real-time analysis experience using HDP 2.2's Kafka-Storm platform.
- Created Spark apps to handle large-scale relational dataset denormalization and transformations.
- Proficient in connecting PowerBI and Tableau to various data sources, including SQL databases, Excel spreadsheets, and cloud platforms. Experienced in creating dynamic drill-down reports and interactive filters in PowerBI, facilitating user-driven data exploration.
- Using Sqoop, loaded data from relational database management systems and Flume, dynamically produced files into the cluster. Using Sqoop to export data into HDFS and Hive, created reports for the BI team.
- By putting the unused user navigation data into HDFS and creating MapReduce tasks, analysis was carried out on the data. The analysis gave the lucent team and the new APM front-end developers some useful information.
- Implemented the Oozie task for daily imports and engaged in real-time data importation to Hadoop via Kafka.
- Used Apache Storm to integrate with Apache Kafka and handle data from several servers in real-time for the enterprise.
- Oversaw and examined Hadoop log files. Developed data transformation APIs using Pyspark.
- Strong understanding of data modelling concepts and techniques, utilizing both PowerBI and Tableau to build robust data models for analytics.
- Assisted in leveraging Scala and Spark RDDs to translate Hive/SQL queries into Spark transformations.
- Importing data from SQL databases into Hadoop to carry out the imported data's consolidations and validations.
- Writing new UDFs for data analysis allows you to extend the fundamental capabilities of Hive and Pig.
- RHEL version 5.6 replaced the existing Linux version.
- Familiarity with PowerBI integration capabilities with Microsoft Office Suite, enabling seamless collaboration and sharing of insights.
- Expertise in hardening, Linux Server, and Compiling, Building, and installing Apache Server from sources with minimum modules. Worked on JSON, Parquet, and Hadoop File formats.
- Worked on different Java technologies like Hibernate, Spring, JSP, and Servlets and developed code for both the server side and client side for our web application. Used Git hub for continuous integration services.

Environment: Agile Scrum, MapReduce, Hive, Pig, Sqoop, PowerBI and Tableau, Spark, Scala, Oozie, Flume, Java, HBase, Kafka, Python, Storm, JSON, Parquet, GIT, JSON SerDe, Cloudera.

EDUCATION:

Masters in Computer Science Aug 2021 to Dec 2022.

VISA:

- STEM OPT.