**Susmitha Reddy**

**susmitha.usk10@gmail.com**

832-390-4838

**Professional Summary:**

* Around **8 Years** of experience in designing data-intensive applications using **Big Data Analytical**, Cloud **Data engineering** (Aws,), Data Visualization, **Data Warehouse**, Reporting, and Data Quality solutions.
* Implemented large Lambda architectures using Azure Data platform capabilities like Azure Data Lake, **Azure Data Factory**, **HDInsight, Azure SQL Server, Azure ML,** and Power BI.
* Practical understanding of **Data modeling** (Dimensional & Relational) concepts like Star - Schema Modelling, **Snowflake Schema Modelling**, Fact and **Dimension tables**.
* Strong understanding of the architecture and capabilities of **Big Query** and **Snowflake**, enabling efficient storage, processing, and analysis of large-scale datasets.
* Hands-on experience working in various cloud (**GCP, AWS, and Azure**) environments.
* Expertise in transforming business requirements into analytical models, designing algorithms, building models, developing **Data Mining**, Data Acquisition, Data Preparation, **Data Manipulation**, Feature Engineering, **Machine Learning Algorithms**, Validation and Visualization, and reporting solutions that scale across massive volumes of structured and unstructured Data.
* IT experience in Big Data technologies, **Spark**, and database development. Good experience in Amazon Web Service (AWS) concepts like **EMR** and **EC2** services which provide fast and efficient processing **of Teradata Big Data Analytics**.
* Strong experience working with **HDFS, MapReduce, Flume, Kafka, Oozie, Pig, and HBase**.
* Experience with creating **ETL** framework with **Python**.
* Experience with Requests, **NumPy, Matplotlib, SciPy, PySpark, and Pandas** Python libraries during the development lifecycle and experience in developing APIs for the application using **Python.**
* Good Knowledge of Big Data Tools such as **Apache Kafka, Apache Spark, Airflow, Hive, Sqoop**, and **Delta Lake** with different file formats like Avro, CSV, and Parquet.
* Expert in designing Parallel jobs using various stages like Join, Merge, Lookup, remove duplicates, Filter, Dataset, look up file set, Complex flat file, Modify, Aggregator & XML.
* Extensive experience working with ETL tools **Airflow /SSIS** in **OLAP** and **OLTP** environments.
* Experience in building data ingestion, **ETL,** and data processing pipelines using On-Prem and Cloud services.
* Coordinating with Business Users, functional Design team, and testing team during the different phases of project development and resolving the issues.
* Hands-on experience in handling database issues and connections with **SQL** and **No SQL** databases such as **MongoDB, and SQL server**.
* Good knowledge of Teradata and Snowflake databases.

**Technical Skills:**

* **Big Data Tools**: Hadoop Ecosystem: Map Reduce, Spark 2.3, Airflow 1.10.8, Nifi 2, HBase 1.2, Hive 2.3, Pig 0.17 Sqoop 1.4, Kafka 1.0.1, Oozie 4.3, Hadoop 3.0
* **Visualization Tools:** Tableau, Looker, Power BI, Qlik Sense, Microsoft Excel, Python (Seaborn, Matplotlib), SSIS, SSRS
* **Database:** MYSQL, SQL Server, PostgreSQL, Azure ML Studio, Oracle
* **Python Libraries:** Pandas, NumPy, TensorFlow, Keras, Scikit-learn, PyTorch
* **Cloud Technologies:** AWS, GCP, Snowflake, Microsoft Azure
* **Data Modelling:** Logistic Regressions, Decision Trees, Multivariate Regression, Neural Networks, Random Forests
* **Methodologies:** Agile/Scrum, Waterfall
* **Version Tools:** GIT, SVN
* **Tools:** Microsoft Office, DAX, PowerPoint, SharePoint, Hypothesis Testing
* **Certification:**- Snowflake SnowPro Core Certified

**Professional Experience:**

**Amdocs Inc, Champaign, IL Jul 2022 – Till Date.**

**Senior Data Engineer**

**Responsibilities:**

* Designed, developed, and tested dimensional data models using Star and Snowflake schemas by the **Kimball methodology**.
* Managed AWS tools, including CloudWatch and CloudTrail, to monitor and track cloud resources and activities.
* Implemented user versioning in S3 buckets for highly sensitive data storage.
* Integrated **AWS DynamoDB** with **AWS Lambda** to manage data storage and backup through DynamoDB streams.
* Installed and configured **Apache Airflow** for AWS S3 buckets and created Directed Acyclic Graphs (DAGs) for workflow automation.
* Developed scripts for automating data ingestion processes using **PySpark** and **Scala**, connecting to various sources, including APIs, AWS S3, Teradata, and Redshift.
* Migrated on-prem Informatica **ETL** processes to **AWS cloud** and **Snowflake**.
* Employed Spark streaming to divide streaming data into batches and process them in the Spark engine.
* Wrote **Spark** applications for data validation, cleansing, transformation, custom aggregation, and analysis, making data available to data scientists.
* Proficient in building **Snowpipe** and experienced in Data Sharing within the **Snowflake Database**, Schema, and Table structures.
* Migrated on-prem **Informatica ETL** processes to **AWS cloud** and **Snowflake**.
* Expertise in migrating existing databases from on-premises to AWS Redshift using various AWS services.
* Created, modified, and executed DDL in **AWS Redshift** and **Snowflake** tables for data loading.
* Monitored resources and applications using AWS CloudWatch, including setting up alarms to track metrics for **EBS, EC2, ELB, RDS, S3**, and **SNS**, with customized notifications.
* Scheduled jobs using Oozie for efficient workflow management.
* Managed the import and export of data from **Snowflake**, **Oracle**, and **DB2** into **HDFS** and **HIVE** using Sqoop for analysis, visualization, and reporting.
* Used filters, quick filters, sets, parameters, and calculated fields on **Tableau** and **Power BI** reports.

**Environment**: Python, Power BI, AWS Glue, Athena, SSRS, AWS S3, AWS Redshift, AWS EMR, AWS RDS, DynamoDB, SQL, Tableau, Distributed Computing, Snowflake, Spark, Kafka, MongoDB, Hadoop, Pyspark, Oozie, HBase, Hive, Pig, Docker.

**Saras America, Hyderabad, India Jan 2020 – Nov 2021**

**Data Engineer**

**Responsibilities:**

* Created ETL mapping document and ETL design templates for the development team.
* Created external tables on top of S3 data which can be queried using AWS services.
* Developed ETL pipelines in and out of the data warehouse using a combination of **Python** and Snowflakes Snows Writing **SQL** queries against **Snowflake**.
* Utilized **Azure Machine Learning** to develop predictive models and apply advanced analytics techniques, empowering data-driven decision-making and optimizing business processes.
* Conducted ETL testing, including extracting data from databases, transforming data, and uploading it to data warehouse servers using **SQL** and **NoSQL**.
* Proficient in designing and developing multidimensional and tabular data models using SQL Server Analysis Services (**SSAS).**
* Worked on developing **Pig scripts** to create the relationship between the data present in the **Hadoop cluster.**
* Performed advanced procedures like text analytics and processing using the in-memory computing capabilities of **Spark** using **Python** and **Scala.**
* Worked on implementing **data lake** and was responsible for data management in **Data Lake.**
* Developed **Ruby Script** to map the data to the production environment.
* Experience in analyzing data using **Hive, HBase** and custom Map Reduce program.
* Developed **Hive UDFs** and **Pig UDFs** using **Python script.**
* Experienced in working with IBM Data Science tool and responsible for injecting the processed data into **IBM Data Science tool.**
* Created HBase tables to load large sets of structured, semi-structured, and unstructured data coming from **UNIX, NoSQL,** and a variety of portfolios.
* Worked on **Oozie Workflow** Engine in running workflow jobs with actions that run **Hadoop Map/Reduce and Pig jobs.**
* Responsible for configuring the cluster in **the IBM cloud** and maintaining the number of nodes as per requirement.
* Created parameterized reports, drill-down reports, and sub-reports in **SSRS** to cater to diverse reporting requirements.
* Implemented data visualization techniques for creating Dashboards and reports using **Python** and **Tableau** to present complex data insights to stakeholders, reducing time spent on data analysis.
* Collaborated with cross-functional teams to develop and implement **machine learning algorithms** for maintenance and collect and report key metrics.
* Strong understanding of data mining algorithms and the ability to implement data mining models within **SSAS** for predictive analytics and pattern recognition.
* Hands-on experience with **Snowflake utilities, SnowSQL, SnowPipe,** and **Big Data** model techniques using **Python.**

**Environment**: Amazon Web Services (AWS), Hadoop, Hive, Presto, PySpark, Databricks, Airflow, Apache Tez, MySQL, PostgreSQL, SQL Server, Python, Scala, Spark, Spark-SQL, Tableau, Data Lake, SnowSQL.

**Mouritech, Hyderabad, India Nov 2017 – Dec 2019**

**Data Engineer**

**Responsibilities:**

* Experience in architecting, designing, installing, configuring, and managing **Apache Hadoop Clusters, MapR,** and **Hortonworks** & **Cloudera Hadoop** Distribution.
* Developed **ETL data pipelines** using **Spark, Spark streaming,** and **Scala.**
* Building Real-Time **Data pipelines** that code modularization, with dynamic configurations to handle multiple data transfer requests without making the production job changes.
* Loaded data from **RDBMS** to **Hadoop** using **Sqoop**
* Worked collaboratively to manage build-outs of large data clusters and real-time streaming with **Spark.**
* Responsible for loading **Data pipelines** from web servers using **Sqoop, Kafka,** and **Spark Streaming API**
* Monitoring the hive metastore and the cluster nodes with the help of Hue.
* Developed **Spark** code using **Scala** and **Spark-SQL/Streaming** for faster processing of data.
* Created **AWS EC2** instances and used **JIT** servers.
* Developed various UDFs in Map-Reduce and **Python** for **Pig and Hive.**
* Data Integrity checks have been handled using hive queries, **Hadoop**, and **Spark.**
* Worked on performing transformations & actions on **RDDs** and Spark Streaming data with **Scala.**
* Responsible for handling Streaming data from web server console logs.
* Installed Oozie workflow engine to run multiple **Hive** and **Pig Jobs.**
* Developed **PIG Latin** scripts for the analysis of semi-structured data.
* Used Hive and created **Hive tables** and was involved in data loading and writing Hive UDFs.
* Used **Sqoop** to import data into HDFS and Hive from other data systems.
* Installed and configured **Apache Hadoop** to test the maintenance of log files in the **Hadoop cluster.**
* Installed and configured **Hive, Pig, Sqoop, Flume,** and **Oozie** on the **Hadoop cluster.**
* Worked on developing **ETL processes (Data Stage Open Studio)** to load data from multiple data sources to **HDFS** using **FLUME** and **SQOOP** and performed structural modifications using Map Reduce, and **HIVE.**
* Involved in **NoSQL** database design, integration, and implementation
* Loaded data into **NoSQL database HBase.**
* Developed **Kafka** producer and consumers, **HBase** clients, **Spark, and Hadoop** MapReduce jobs with components on **HDFS, and Hive.**
* Very good understanding of Partitions, and bucketing concepts in Hive and designed both Managed and External tables in Hive to optimize performance.
* Developed workflows for complete end-to-end **ETL process** starting with getting data into HDFS,
* validating and applying business logic, storing clean data in hive external tables, and exporting data from
* **hive to RDBMS** sources for reporting and escalating data quality issues.
* Handled importing of data from various data sources performed transformations using **Spark** and loaded data into hive.
* Involved in performance tuning of **Hive (ORC table)** for design, storage, and query perspectives.
* Developing and deploying using Horton Works HDP 2.3.0 in production and HDP 2.6.0 in the development environment.
* Worked extensively with **Sqoop** for importing and exporting the data from HDFS to Relational Database systems

**Environment**: Hive, Sqoop, ETL Process, Spark, NoSQL, HBase, Kafka, HDFS, Apache, Pig, FLUME, Hadoop cluster, Sqoop, Oozie, Scala, AWS EC2, JIT, Data Pipelines, RDBMS, Python

**Sonata Software Limited, Hyderabad, India Mar 2015 – Oct 2017**

**Big Data Engineer**

**Responsibilities:**

* Suggestions on converting to **Hadoop** using MapReduce, **Hive, Sqoop, Flume,** and **Pig Latin**.
* Collecting and aggregating large amounts of log data using **Apache Flume** and staging data in HDFS for further analysis
* Extracted data from various locations and loaded them into the **Oracle** table using **SQL\*LOADER.**
* Responsible for cluster maintenance, adding and removing cluster nodes, cluster monitoring, and troubleshooting, managing and reviewing data backups, and reviewing **Hadoop log files.**
* Developed the **Pig Latin** code for loading, filtering, and storing the data.
* Create, develop, modify, and maintain Database objects, **PL/SQL** packages, functions, stored procedures, triggers, views, and materialized views to extract data from different sources.
* Handled Import of data from various data sources, performed transformations using **Hive**, and MapReduce, and loaded data into HDFS.
* Extracted the data from **Teradata** into HDFS using **Sqoop**.
* Analyzed the data by performing Hive queries and running **Pig** scripts to know user behavior.
* Installed **Oozie workflow** engine to run multiple **Hive.**
* Developed **Hive** queries to process the data and generate the data cubes for visualization.
* Handled importing of data from various data sources, performed transformations using **Hive**, and MapReduce, loaded data into HDFS, and Extracted the data from **Oracle database** into HDFS using **Sqoop.**
* Importing and exporting data into HDFS and **Hive** using **Sqoop.**
* Analyzed the data by performing **Hive queries** and running **Pig scripts** to know user behavior
* Continuous monitoring and managing of the **Hadoop cluster** through Cloudera Manager.
* Installed **Oozie** workflow engine to run multiple **Hive and Pig** jobs.
* Developed **Hive queries** to process the data and generate the data cubes for visualization.
* Worked on loading the data from **MySQL** to **HBase** where necessary using **Sqoop**
* Responsible for building scalable distributed data solutions using **Hadoop.**
* Worked hands-on with the **ETL process.**
* Responsible for cluster maintenance, adding and removing cluster nodes, cluster monitoring, and

troubleshooting, managing and reviewing data backups, and managing and reviewing **Hadoop log files**

**Environment**: Hadoop, Sqoop, ETL, MySQL, Hive, Teradata, Data Pipelines, RDBMS, Python

Oozie, Pig, Oracle, HBase, SQL, Apache, FLUME, HDFS