Subahani Shaik

Email: subahani.de6@gmail.com

Phone No: +91 7838876077

Objective Summary

Over 4+ years of IT experience in implementing end-to-end Datawarehouse and Data engineering solutions. expertise in the field of data engineering using AWS cloud services. End-to-end data ingestion, modeling, processing, transformation, and analysis solutions with a focus on Spark
SQL, Python, AWS, Unix Shell scripting, Tableau, Oracle, GitHub, SQL, PySpark, AWS Glue, AWS Lambda, and Athena. Experience with SQL database migration, RDS, Redshift, and S3, Athena, Glue.

PROFESSIONAL SUMMARY

- Having 4 yrs of experience as Data Engineer. Experience in Analysis, Design and Implementation of Business Applications.
- Experience in Agile methodology in current project.
- Experience in writing SQL queries using analytical functions, aggregate, with clause, joins, hints etc.
- Good Exposure to **Pyspark ,sqoop,hive , hdfs , aws services , unix shell scripting and python .**
- Experience in using spark API's on cloud era for importing data from Relation database **to HDFS, AWS S3 Bucket in different file format and then loading it to different data Warehouse(Redshift).**
- Have experience on **AWS (S3, Redshift,Glue,Lambda, CI/CD, Athena).**
- Experience implementing batch and real-time data pipelines using **AWS Services, S3, Lambda, DynamoDB, Redshift, Glue.**
- Strong SQL skills to query data for validation, reporting and dashboarding.
- Developed **Glue job and Lambda functions to migrate the data across different Data layers .**
- Developed Unix Shell scripts to onboard the historical data from on - premise to S3.
- Developed Redshift procedures similar to on premise ETL logic.
- Strong experience **in CI (Continuous Integration)/ CD (Continuous Delivery) software development pipeline stages like Commit, Build, Automated Tests, and Deploy**.
- working knowledge on SQL DB/DW and developing pipelines, scheduling pipelines, storage repo and deploying from development environment to test and production.
- Implemented SCD type1, 2, email notifications using logic app, rollback using batch id and error handling.
- Creating Notebooks by using pycharm with Python and PYSPARK scripting language • Worked on EMR clusters to maintain data quality with required format.
- Creating Logic apps to integrate applications, data, services, systems across various enterprises or organization.
- Experience developing Spark applications **using Pyspark and Spark-SQL for data extraction, transformation, and aggregation from multiple file formats for analyzing & transforming the data to uncover insights into customer usage patterns**.
- Experienced in using spark scripts to do transformations, event joins, filters, and pre-aggregations before storing the data in HDFS.
- Experience working with Map Reduce programs using Apache Hadoop for Big Data.
- **Experience installing, configuring, supporting, and monitoring Hadoop clusters using Apache, Cloudera distributions, and AWS.**
- Good Knowledge of **Amazon Web Service (AWS) concepts like EMR and EC2 web services successfully loaded files to HDFS from Oracle, SQL Server, Teradata, and Netezza using Sqoop.**
- Excellent knowledge of Big Data infrastructure, distributed file systems -HDFS, parallel processing -MapReduce framework.
- Good understanding and exposure to Python programming.

TECHNICAL SKILLS

- Cloud : AWS(ATHENA, EC2 ,S3, EMR, RDS, REDSHIFT)
- Big Data : PySpark, Apache Kafka
- Code Repository Tools : GitLab
- Database : SQL Server (2008 R2 to 2019), MySQL, PostgreSQL
- Operating System : Windows, Linux,
- Other tools : Pycharm, MS Office, Putty

PROFESSIONAL EXPERIENCE

Sr. Data Engineer **Jan 2022 – July 2022**

Ford Motor

Responsibilities:

- Implemented solutions utilizing Advanced AWS Components: EMR, EC2, etc. integrated with Big Data/Hadoop Distribution Frameworks: Hadoop YARN, MapReduce, Spark, Hive, etc.
- Used AWS Athena extensively to ingest structured data from S3 into multiple systems, including RedShift, and to generate reports.
- Created on-demand tables on S3 files using Lambda Functions and AWS Glue using Python and PySpark.
- Performed end-to-end Architecture and implementation assessment of various AWS services like Amazon EMR, Redshift, S3, Athena, Lambda ,Step Functions Glue, and Kinesis.
- Install and configure Apache Airflow for S3 bucket and Snowflake data warehouse and created dags to run the Airflow.
- Experienced in performance tuning of Spark Applications for setting right Batch Interval time, correct level of Parallelism and memory tuning .
- Developed PySpark and Spark SQL code to process the data in Apache Spark on Amazon Glue to perform the necessary transformations based on the STMs developed.
- Creating Notebooks by using Pycharm with Python and PYSPARK scripting language
- Worked on EMR clusters to maintain data quality with required format. • Creating Logic apps to integrate applications, data, services, systems across various enterprises or organizations.
- Worked on logic app create a business process graphically using the workflow engine and visual designer and connect them through connecters
- Creating logic app to send email notifications to different users when an event happens in various applications, services, and systems, etc.
- Developed shell scripts to validate the QC checks across the data set and scheduled using autosys tool.
- Worked on enhancement of procedures and packages in Oracle PLSQL.
- Using Spark-SQL to Load JSON data and create Schema RDD and loaded into Hive Tables and handled structured data using Spark-SQL.

Data Engineer **Oct 2019 to Dec 2021**

Sat Software Infrastructure Pvt. Ltd.

Responsibilities:

- Expertise in creating and architecting various data pipelines, including full-cycle ETL and ELT processes for data ingestion and Data transformation in AWS and Spark.

- Skill in building pyspark code for iterative Spark algorithms to optimize performance
- Involved in job management using Fair scheduler, developed job processing scripts using Oozie workflow, and was responsible for developing scalable distributed data solutions using Hadoop.
- Worked with ETL engineers to make sure that data is thoroughly cleaned and the data warehouse is current for Pig reporting purposes
- Selected and created data was put in CSV files and then organized and stored in AWS Redshift using AWS EC2.
- To speed up testing and data processing, Spark code was created utilizing Scala and Spark-SQL/Streaming.
- Participated in the implementation of SQOOP, which facilitates the transfer of data from multiple RDBMS sources to Hadoop systems and vice versa.
- Developed a task scheduling program to run across numerous servers in an EC2 environment.
- Skilled at minimizing query response time by integrating Hive QL with Impala.
- Data from AWS S3 was imported into Spark RDDs, and the RDDs underwent various transformations and operations.
- Worked with NoSQL databases like HBase in making HBase tables to load expansive arrangements of semi-structured data
- Performed data preparation, trained, tested, and feature engineering for additional predictive analytics using Python Pandas

Environment Oracle SQL, Spark, AWS S3, Amazon Redshift, AWS EMR, AWS RDS, DynamoDB, Lambda, Hive, HDFS, Sqoop

Data Engineer **May 2018 to Sep 2019**

Sat Software Infrastructure Pvt. Ltd.

Responsibilities:

- Designed and developed PySpark applications for processing and analyzing large-scale datasets from various sources, resulting in improvement in data processing efficiency.
- Implemented ETL pipelines using PySpark to transform and load data into a data lake, enabling seamless integration with downstream applications.
- Collaborated with data scientists and analysts to understand business requirements and translate them into efficient Spark-based solutions.
- Optimized Spark jobs by tuning configurations, employing caching strategies, and leveraging Spark UI for performance analysis, leading to reduction in processing time.
- Integrated PySpark applications with cloud-based storage systems (e.g., S3, GCS) and databases (e.g., Redshift) to enable data storage and retrieval.
- Conducted thorough testing and debugging of Spark applications, identifying and resolving issues to ensure data accuracy and reliability.
- Mentored and trained junior PySpark developers, fostering a collaborative and knowledge-sharing environment within the team.
- **Environment:** Microsoft SQL Server 2012/2014, Oracle 12c, SQL Server Data Tools 2016 (SSDT), Maven, Git, Bit Bucket, Windows Server XP, MSOffice 2013, Excel 2013,MS Office365.