Diagram

Description automatically generated**Shahreen Tanzila**

**Saint louis, MO**

**314-514-5359**

[**tasha102295@gmail.com**](mailto:tasha102295@gmail.com)

**Senior AWS Data Engineer**

**Professional Summary:**

* Big Data Engineer with over 8+ years of overall experience as software developer in design, development, deploying and large scale supporting large scale distributed systems.
* 8+ years of experience working on large scale, full lifecycle data implementation projects.
* 2+ years professional development experience with some of the AWS/Azure/GCP stack.
* Hands on experience on complete Software Development Life Cycle **SDLC** for the projects using methodologies like Agile and hybrid methods.
* 2+ years of experience in a client facing role
* Expert in using **Python** for **Data Engineering** and **Modeling**.
* Data importing and exporting by using **Sqoop** from **HDFS** to Relational Database Systems and vice-versa.
* Hands-on experience with **Amazon EC2, Amazon S3, Amazon RDS, VPC, IAM, Amazon Elastic Load** **Balancing, Auto Scaling, Cloud Front, CloudWatch, SNS, SES, SQS** and other services of the AWS family.
* Excellent knowledge of **Microsoft Office** with an emphasis on **Excel**.
* Proficient in data mining tools like **R, SAS, Python, SQL Alchemy, Excel, BigData Hadoop** eco-systems.
* Experience in developing customized **UDF's** in **Python** to extend **Hive** and **PigLatin** functionality.
* Experience in extracting source data from Sequential files, **XML** files, **CSV** files, transforming and loading it into the target **Data warehouse**.
* Transforming and retrieving the data by using **Pig, Hive, SSIS** and **Map Reduce**.
* Excellent Knowledge of **RalphKimball** and **BillInmon's** approaches to **Data Warehousing**.
* Extensive experience in Text Analytics, developing different Statistical **Machine Learning**, **Data Mining solutions** to various business problems and generating data visualizations using **R, Python**.
* Good knowledge in **OLAP, OLTP**, **Business Intelligence** and **Data Warehousing** concepts with emphasis on **ETL** and Business Reporting needs.
* Hands on experience for Data Integrity process and **Data Modelling** concepts.
* Experience in Data mining with querying and mining large datasets to discover transition patterns and examine financial reports.
* Responsible for designing and building a **Data Lake** using **Hadoop** and its ecosystem components.
* Automated the cloud deployments using **puppet**, chef, python and **AWS Cloud Formation** Templates.
* Migrated an existing on-premises application to **AWS**. Used **AWS** services like EC2 and S3 for small data sets processing and storage and deployed using **puppet.**
* Strong background in various **Data modeling** tools using **ERWIN**, **ER Studio** and **PowerDesigner**.
* Expertise in **DataReporting**, **Ad-hocReporting**, Graphs, Scales, PivotTables and **OLAP** reporting.
* Hands on experience on data modeling with **Star schema** and **Snowflake schema.**
* Perform structural modifications using **Map - Reduce, HIVE** and analyze data using visualization/ reporting tools.
* Good experience in using Data Modelling techniques to find the results based on SQL and PL/SQL queries.
* Experience working with different databases, such as **Oracle, SQL Server, MySQL** and writing stored procedures, functions, joins, and triggers for different Data Models.
* Experience in **Impala, Solr, MongoDB, HBase** and **Spark, Kubernetes.**
* An excellent experience in generating ad-hoc repots using **Crystal Report**
* Experience in Performance Tuning and query optimization techniques in transactional and **Data Warehouse** Environments
* Experience in develop data set processes for data modelling, and Data mining. Recommend ways to improve data reliability, efficiency and quality.

**Technical Skills:**

|  |  |
| --- | --- |
| **Languages** | Scala, Python, SQL, Java, PL/SQL, Linux shell scripts |
| **Hadoop Ecosystem** | HDFS, Spark/PySpark, YARN, Hive, Hbase, Impala, Zookeeper, Sqoop, Oozie, Drill, Flume, Spark/PySpark, Solr and Avro, |
| **AWS Cloud Stack** | S3, EC2, IAM, Lambda, Cloud Watch, Cloud Formation, Auto Scaling, Security Groups, Redshift, EMR, Dynamo DB, AWS glue, |
| **Enterprise Technologies** | J2EE, Event Hub, Kinesis, JDBC, Kafka |
| **Operating Systems** | Windows, Linux, UNIX, Mac-OS. |
| **IDEs** | Eclipse, IntelliJ |
| **Relational Databases** | Oracle, SQL, DB2, MySQL, Teradata |
| **NoSQL databases** | Hbase, Cassandra, Mongo DB |
| **Build & Management Tools** | ANT, MAVEN, SVN. |
| **Query Languages** | SQL, PL/SQL. |
| **Methodologies** | SDLC, OOAD, Agile. |
| **Continuous Integration Tools** | Jenkins, Docker, Kubernetes |

**Project Experience:**

**Client:AIG, PA Feb 2021 – till date**

**Role: Senior AWS Data Engineer**

**Responsibilities:**

* Collaborated with Business Analysts, SMEsacross departments to gather business requirements, and identify workable items for further development.
* Utilized **Agile and Scrum** methodology for team and project management.
* Loaded data into S3 Buckets using AWS Glue and Pyspark involved in filtering data stored in S3 buckets using Elasticsearch and loaded data into Hive external tables.
* Partnered with **ETL** developers to ensure that data is well cleaned and the data warehouse is up-to-date for reporting purpose by **Pig**.
* Worked on Big data on **AWS cloud services** i.e. **EC2, S3, EMR, Lambda, Auto Scaling, Cloud Watch, Cloud Formation, Security Groups** and **DynamoDB**
* Selected and generated data into csv files and stored them into **AWS S3** by using **AWS EC2** and then structured and stored in **AWS Redshift**.
* Performed the migration of **Hive** and **Map Reduce** Jobs from on - premise **MapR** to **AWS** cloud using EMR and **Qubole**
* Developed **Spark** code using **Scala** and **Spark**-**SQL/Streaming** for faster testing and processing of data.
* Involved in **SQOOP** implementation which helps in loading data from various RDBMS sources to **Hadoop** systems and vice versa.
* Automated the data processing with **Oozie** to automate data loading into the **Hadoop** Distributed File System.
* Hands on experience in installing configuring and using **Hadoop** ecosystem components like **Hadoop** **Map Reduce** **HDFS** **HBase** **Hive** **Sqoop** **Pig** **Zookeeper** and **Flume**
* Experience in using **Kafka** and **Kafka** brokers to initiate **spark** context and processing livestreaming.
* Used **PySpark** and **Pandas** to calculate the moving average and RSI score of the stocks and generated them into data warehouse.
* Experience other **Hadoop** ecosystem tools in jobs such as **ZooKeeper**, **Oozie**, **Implala**
* Experience in working with **Map reduce** programs, **Pig** scripts and **Hive** commands to deliver the best results
* Utilized **Spark**, **Scala**, **Hadoop**, **HBase**, **Kafka**, **Spark** **Streaming**, **MLLib**, **Python**, a broad variety of machine learning methods including classifications, regressions, dimensionally reduction etc.
* Expertise in writing the **Scala** code using higher order functions for the iterative algorithms in **spark** for performance consideration
* Configured **Zookeeper**, worked on **Hadoop** High Availability wif **Zookeeper** failover controller, add support for **scalable**, fault-tolerant data solution.
* Migrated an existing on-premises application to **AWS**. Used **AWS** services like EC2 and S3 for small data sets processing and storage and deployed using **puppet.**
* Developed **Pig** Scripts, **Pig** UDFs and **Hive** Scripts, **Hive** UDFs to load data files.
* Developed **Map Reduce** programs to parse the raw data, populate staging tables and store teh refined data in partitioned tables in the EDW
* Used **Spark** Streaming to receive real time data from the **Kafka** and store the stream data to HDFS using **Python** and **NoSQL** databases such as **HBase** and **Cassandra**
* Involved in integration of **Hadoop** cluster with **spark** engine to perform **BATCH** and **GRAPHX** operations.
* Responsible for importing data from **Postgres** to **HDFS, HIVE using SQOOP tool**.
* Experienced in migrating **Hive QL** into Impala to minimize query response time.
* Implemented Avro and parquet data formats for **apache Hive** computations to handle custom business requirements.
* Implemented the **Spark** **Scala** code for Data Validation in **Hive**
* Imported data from **AWS** S3 into **Spark** RDD, Performed transformations and actions on RDD's.
* Utilized **SQOOP**, **Kafka**, Flume and **Hadoop** File system APIs for implementing data ingestion pipelines
* Developed various Mappings with the collection of all Sources, Targets, and Transformations using **Informatica** Designer
* Configure **Zookeeper** to coordinate and support **Kafka**, **Spark**, **Spark** Streaming, **HBase** and HDFS
* Experience d in developing **Map Reduce** programs using Apache **Hadoop** for working with Big Data
* Performed data preprocessing and feature engineering for further predictive analytics using **Python Pandas**.
* Used the **AWS**-CLI to suspend an **AWS** Lambda function. Used **AWS** CLI to automate backups of ephemeral data-stores to S3 buckets, EBS.
* Used **Hbase**/**Pheonix** to support front end applications that retrieve data using row keys
* Developed custom **Kafka** producer and consumer for different publishing and subscribing to **Kafka** topics.
* Generated report on predictive analytics using **Python** and **Tableau** including visualizing model performance and prediction results.
* Expertise in using **Docke**r to run and deploy the applications in multiple containers like **Docker Swarm** and **Docker Wave.**
* Developed complex **Talend ETL jobs** to migrate the data from **flat files** to database. Pulled files from **mainframe into Talend** execution server using multiple **ftp** components.
* Used **Sqoop** to channel data from different sources of **HDFS** and **RDBMS**
* Architect and design serverless application CI/CD by using **AWS Serverless (Lambda**) application model.
* Developed **stored procedures/views in Snowflake** and use in **Talend** for loading Dimensions and Facts.
* Developed merge scripts to **UPSERT** data into **Snowflake** from an ETL source
* Used **Git** for version control with colleagues.

Environment: Hdfs, Hive, Spark, Kafka, Linux, Python, Numpy, Pandas, Tableau, GitHub, AWS EMR, EC2, S3, Redshift, Lambda, Auto Scaling, Cloud Watch, Cloud Formation, Pig, Map Reduce, Cassandra, Snowflake, Unix, Shell Scripting, Git.

**Client: TIkTOK, CA May 2019 – Feb 2021**

**Role: AWS Data Engineer**

**Responsibilities:**

* Used Agile methodology in developing the application, which included iterative application development, weekly Sprints, stand up meetings and customer reporting backlogs.
* Extensive experiences working with AWS cloud platform(EC2,S3, Lambda and Glue)
* Monitoring resources and Applications using **AWS Cloud Watch,** including creating alarms to monitor metrics such as **EBS, EC2, ELB, RDS, S3, SNS** and configured notifications for the alarms generated based on events defined.
* Writing code and creating **hive** jobs to parse the logs and structure them in tabular format to facilitate effective querying on the log data.
* Created and managed cloud VMs with **AWS EC2** Command line clients and AWS management console.
* Migrated on premise database structure to Confidential Redshift data warehouse. Worked on AWS Data Pipeline to configure data loads from **S3** into **Redshift**
* Created functions and assigned roles in **AWS Lambda** to run python scripts, and **AWS Lambda** using **java** to perform event driven processing. Created **Lambda** jobs and configured Roles using **AWS CLI.**
* Write documentation of program development, subsequent revisions and coded instructions in the project related **GitHub** repository.
* Deployment support including change management and preparation of deployment instructions.
* Prepare release notes, validation document for user stories to be deployed to production as part of release.
* Design the incremental, historical extract logic to load the data from flat files into Massive Event Logging Database (MELD) from various servers.
* Writing technical design document based on the data mapping functional details of the tables.
* Extracting batch and Real time data from **DB2, Oracle, Sql server, Teradata, Netezza** to Hadoop (HDFS) using **Teradata TPT, Sqoop, Apache Kafka, Apache Storm.**
* Assist with the analysis of data used for the **tableau** reports and creation of dashboards.
* Design and implement large scale distributed solutions in **AWS.**
* Optimized **Map Reduce** Jobs to use **HDFS** efficiently by using various compression mechanisms.
* Development of efficient **pig** and **hive** scripts with joins on datasets using various techniques.
* Automating the ETL tasks and data work flows for the data pipeline of the ingest process through **UC4** scheduling tool.
* Solved performance issues in **Hive** and **Pig** scripts with understanding of Joins, Group and aggregation and how does it translate to **MapReduce** jobs.
* Used ORC and Parquet file formats in **Hive**.
* Handled importing of data from various data sources, performed transformations using Hive, MapReduce, Spark and loaded data into **HDFS.**
* Developed UDFs in **Java** as and when necessary to use in **PIG** and **HIVE queries.**
* Automated the cloud deployments using **puppet**, chef, python and **AWS Cloud Formation** Templates.
* Design and build ETL workflows, leading the efforts of programming data extraction from various sources into Hadoop file system, implement end to end ETL workflows using **Teradata, SQL, TPT, SQOOP** and load to **HIVE** data stores.
* Experience in change implementation, monitoring and troubleshooting of **AWS Snowflake** databases and cluster related issues
* Analyze and develop programs by considering the extract logic and the data load type using Hadoop ingest processes using relevant tools such as **Sqoop, Spark, Scala, Kafka, Unix shell scripts** and others.
* Developing **Apache Spark** jobs for data cleansing and pre-processing.
* Writing spark programs to improve the performance and optimization of the existing algorithms in Hadoop using **spark context, spark-sql, data frame, pair RDD's, spark yarn.**
* Using **Scala** language to write programs for faster testing and processing of data.

**Environment**: RHEL, HDFS, Map-Reduce, Hive, AWS, EC2, S3, Lambda, Redshift, Pig, Sqoop, Oozie, Teradata, Oracle SQL, UC4, Kafka, GitHub, Hortonworks data platform distribution, Spark, Scala.

**Client: Verizon, San Diego, CA Feb 2017 – May 2019**

**Role: Big Data Engineer**

**Responsibilities:**

* Developed SSRS reports, SSIS packages to Extract, Transform and Load data from various source systems
* Implementing and Managing ETL solutions and automating operational processes.
* Was responsible for **ETL** and data validation using **SQL Server Integration Services.**
* Measured Efficiency of Hadoop/Hive environment ensuring SLA is met
* Optimized the **Tensor Flow** Model for efficiency
* Created Entity Relationship Diagrams (ERD), Functional diagrams, Data flow diagrams and enforced referential integrity constraints and created logical and physical models using **Erwin.**
* Defined and deployed monitoring, metrics, and logging systems on **AWS.**
* and **Spark SQL** for ETL jobs and using the right technology for the job to get done.
* Compiled data from various sources to perform complex analysis for actionable results
* Strong understanding of **AWS** components such as **EC2** and **S3**
* Managed security groups on **AWS,** focusing on high-availability, fault-tolerance, and auto scaling using Terraform templates. Along with Continuous Integration and Continuous li with **AWS Lambda** and AWS code pipeline.
* Defined facts, dimensions and designed the data marts using the Ralph Kimball's Dimensional Data Mart modeling methodology using Erwin.
* Worked on Big data on **AWS** cloud services i.e**. EC2, S3, EMR** and **DynamoDB**
* Created ad hoc queries and reports to support business decisions **SQL Server Reporting Services (SSRS).**
* Analyze the existing application programs and tune **SQL queries** using execution plan, query analyzer, SQL Profiler and database engine tuning advisor to enhance performance.
* Optimizing and tuning the **Redshift** environment, enabling queries to perform up to 100x faster for **Tableau** and SAS Visual Analytics.
* Migrated on premise database structure to Confidential Redshift data warehouse

**Environment:** AWS, EC2, S3, SQL Server, Erwin, Oracle, Redshift, Informatica, RDS, NOSQL, MySQL, Dynamo DB, Docker, PostgreSQL, Tableau, Git Hub

**Client: GE Commercial Finance Division, Hyderabad, India Jan 2015 – Nov 2016**

**Role: Data Engineer**

**Responsibilities:**

* Worked on development of data ingestion pipelines using ETL tool, Talend & bash scripting with big data technologies including but not limited to **Hive**, Impala, **Spark**, **Kafka**, and Talend.
* Experience in developing **scala**ble & secure data pipelines for large datasets.
* Gathered requirements for ingestion of new data sources including life cycle, data quality check, transformations, and metadata enrichment.
* Developed **Map reduce**/**Spark** **Python** modules for machine learning & predictive analytics in **Hadoop** on **AWS**.
* Developed data pipeline using **Flume, Sqoop, Pig** and **Java Map Reduce** to ingest customer behavioral data into HDFS for analysis.
* Supported data quality management by implementing proper data quality checks in data pipelines.
* Build machine learning models to showcase Big data capabilities using Py**spark** and MLlib.
* Enhancing Data Ingestion Framework by creating more robust and secure data pipelines.
* Implemented data streaming capability using **Kafka** and Talend for multiple data sources.
* Involved in **SQOOP** implementation which helps in loading data from various RDBMS sources to **Hadoop** systems and vice versa.
* Worked with multiple storage formats (Avro, Parquet) and databases (**Hive**, Impala, Kudu).
* S3 – Data Lake Management. Responsible for maintaining and handling data inbound and outbound requests through big data platform.
* Used **Sqoop** to transfer data between relational databases and **Hadoop**.
* Knowledge on implementing the JILs to automate the jobs in production cluster.
* Worked with SCRUM team in delivering agreed user stories on time for every Sprint.
* Worked on analyzing and resolving the production job failures in several scenarios.
* Implemented UNIX scripts to define the use case workflow and to process the data files and automate the jobs.

**Environment:** Spark, AWS, Redshift, Python, HDFS, Hive, Pig, Sqoop, Scala, Kafka, Shell scripting, Linux, Jenkins, Eclipse, Git, Oozie, Talend, Agile Methodology.

**Client: JK Infotech, Hyderabad, India Oct 2013 – Jan 2015**

**Role: Hadoop Developer**

**Responsibilities:**

* Developed **PIG scripts** for the analysis of semi structured data.
* Worked on **Oozie workflow** engine for job scheduling.
* Worked with **Avro Data** Serialization system to work with JSON data formats.
* **Hive Context**, with transformations and actions (**map, flat Map, filter, reduce, reduce** by Key).
* Migrated **ETL jobs** to **Pig scripts** to do Transformations, even joins and some pre-aggregations before storing the data onto **HDFS**.
* Written **Hive** jobs to parse the logs and structure them in tabular format to facilitate effective querying on the log data.
* Worked on different file formats like Sequence files, XML files and Map files using **MapReduce** Programs.
* Developed **PIG UDF'S**for manipulating the data according to Business Requirements and also worked on developing custom **PIG Loaders.**
* Worked on various performance optimizations like using distributed cache for small datasets, partition and bucketing in **Hive**, doing map side joins etc.
* Performed pre-processing on a dataset prior to training, including standardization, normalization.
* Worked with heterogeneous source to Extracted data from **Oracle database, XML** and flat files and loaded to a relational **Oracle warehouse**.
* Evaluated model accuracy by dividing data into training and test datasets and computing metrics using evaluators.
* Performed tuning of **SQL queries** and Stored Procedure for speedy extraction of data to resolve and troubleshoot issues in **OLTP** environment.

**Environment:** Hive, HDFS, Oozie, Map Reduce, Oracle 10g, SQL, OLTP, Windows, MS Office

**Educational Qualifications:**

**Bachelor’s in computer science, GPA (2013) 3.5/4**

Jawaharlal Nehru Technological University (2013 May)