HARSHITA C

harshitach48@outlook.com | LinkedIn |+1 (972)-763-5673

Overall, 7+ years of experience in Data Engineering, expertise in designing, implementing, and developing the Data Pipeline, Data Migration, Data Modelling, Data transformation, Data Munging, Data Enrichment **using the cloud development platforms and native big data tools.** Leveraged the cutting-edge technologies like Databricks, Apache Spark and Airflow to drive efficiency and performance. Proficient in building data architecture that facilitates real-time insights using different cloud platforms like AWS, Azure and GCP and focused on automation and analytics, guaranteeing smooth dataflow and optimal decision making

PROFESSIONAL SUMMARY

- Utilized **AWS services** like **EC2** and **S3** for efficient data storage and processing. Experienced in managing Hadoop clusters on **AWS EMR** in optimizing resource allocation for big data applications
- Strong experience in writing scripts using **Python API** and **PySpark** for analyzing data
- Practical knowledge of Apache Spark, including **Spark Core, Spark SQL, and Spark Streaming** for building real-time data pipelines and managing **Spark DataFrames** for efficient structured data processing
- Experienced in **NoSQL** databases, including table row key design, and to load and retrieve data for real-time data processing and performance improvements based on access patterns
- Proficient in Azure and GCP for optimizing big data processes using **BigQuery, Dataproc, Cloud Functions, GCS, Azure Databricks, Azure Data Factory**. Capable at refining data pipelines for seamless data transmission between GCP and Azure using Azure Data Factory
- Expertise in **Power BI** reporting with **Azure Analysis Services** for interactive dashboards. Skills in optimizing performance between pre-aggregated Azure datasets and direct queries in GCP query
- Expertise **in AWS services including S3, EC2, SNS, SQS, RDS, Neptune, EMR, Kinesis, Lambda, Step Functions, Cloud Formation, Event Bridge, Glue, Redshift, Athena, DynamoDB** to build high performance data solutions. Familiar with **CloudWatch** and **IAM** for monitoring the system
- Worked on Hadoop architecture, including **HDFS, MapReduce, JobTracker, NameNode, DataNode** and resource management for high performance in distributed data processing
- Well-versed in automating complex data operations by using **Apache Airflow** and **Oozie** for workflow management and assuring dependable ETL pipelines and monitoring
- Strong background in **CI/CD and automation**, utilizing tools like **Maven, SBT, Git, SVN, and Jenkins** to streamline development workflows, version control, and continuous integration
- Experienced in in Microsoft Business Intelligence tools, developing **SSIS** for integration, **SSAS** for analysis and **SSRS** for reporting, building Key Performance Indicators and **OLAP** cubes
- Proficient in **SQL** across multiple platforms, including **MySQL, PostgreSQL, Redshift, SQL Server, and Oracle**, with a deep understanding of **query optimization, indexing strategies, and database performance tuning** to handle complex analytical workloads
- Experience in orchestrating the **Airflow** workflow to migrate the data from legacy systems to the target **snowflake** table
- Expert in **data warehousing methodologies**, including **Kimball and Inmon approaches**, designing **enterprise data warehouses (EDWs)** from scratch to support **advanced business intelligence and analytics capabilities**
- Hands-on experience with **schema design**, including **Star and Snowflake schemas**, optimizing **relational data storage** for **high-performance analytics** in enterprise data warehousing
- Experience on Migrating **SQL database** to **Azure data Lake, Azure data lake Analytics, Azure SQL Database, Data Bricks and Azure SQL Data warehouse** and controlling and granting database access and Migrating On premise databases to **Azure Data Lake** store using **Azure Data factory**
- Experience in configuring & development skills with **Azure Data Lake**, **Azure Data Factory**, **Azure SQL Data Warehouse**, **Azure Blob**, **Azure Storage** Explorer, and other Azure services such as **Databricks**, **Stream Analytics**, **Synapse Analytics**, **SQL Database, Azure HD insights**, **Azure SQL Datawarehouse**, **Cosmos DB**
- In order to optimize inference pipelines for high-performance AI workloads and guarantee scalability and real-time processing in cloud environments, vLLM models were deployed using vLLM-deploy on **Azure AKS**
- Constructed **AI agents** for real-time decision making by integrating structured and unstructured data using Azure Databricks, Spark Streaming and Cosmos DB
- Strong experience in Software Development Life Cycle (**SDLC**) which covers requirement analysis till deployment to guarantee robust and scalable, with **Agile** methodology for seamless project management

TECHNICAL SKILLS

Hadoop/Big Data Technologies	HDFS, Apache NIFI, Map Reduce, Sqoop, Flume, Pig, Hive, Oozie, Impala, Zookeeper, Ambari, Storm, Spark, and Kafka
Cloud Services	Azure: Azure Data Lake, Azure Data Factory, Azure SQL Data Warehouse, Azure Blob, Azure Storage Explorer, Azure Databricks, Azure Stream Analytics, Azure Synapse Analytics, Azure SQL Database, Azure HDInsight, Cosmos DB, Azure Analysis Services, Azure Kubernetes Service (AKS), Azure Active Directory (Azure AD) AWS: S3, EC2, SNS, SQS, RDS, Neptune, EMR, Kinesis, Lambda, Step Functions, Cloud Formation, Event Bridge, Glue, Redshift, Athena, DynamoDB, CloudWatch, IAM, CloudTrail, SNS GCP: BigQuery, Dataproc, Cloud Functions, Google Cloud Storage (GCS)
No SQL Database	Cassandra, MongoDB, DynamoDB
Hadoop Distribution	Horton Works, Cloudera
Build and Deployment Tools	Maven, Sbt, Git, SVN, Jenkins
Programming and Scripting	SQL,Shell Scripting, Python HiveQL, PySpark
Databases	Oracle, MY SQL, MS SQL Server, Vertica, Teradata
Analytics Tools	Tableau, Microsoft SSIS, SSAS and SSRS
Operating Systems	Linux, Unix, Windows 8, Windows 7, Windows Server 2008/2003
BI Tools	Power BI, Tableau, SSAS, SSIS, SSRS

PROFESSIONAL EXPERIENCE

Client: Verizon Communications Inc. **April 2024 – Present**
Role: Senior Data Engineer/ AI Engineer

Project Overview: Verizon, a global leader in telecommunications, is committed to delivering scalable, data-driven solutions by leveraging Azure cloud services to enhance data integration, transformation, and real-time analytics. This project utilized Azure Databricks, Azure Data Factory, HDInsight, Cosmos DB and Azure SQL Data Warehouse for smooth data transformation and integration. AI-driven insights and smooth workflows for business operations were made possible using Apache Airflow, Spark and Power BI.

Environment: *Pyspark, Spark, Spark SQL, MySQL, Cassandra, Snowflake, MongoDB, Flume, VSTS, AZURE services (Azure HDInsight, Data Bricks (ADBX), Data Lake (ADLS), Cosmos DB, DevOps, Azure AD, Blob Storage, Data Factory), Git, Scala, Hadoop 2.x (HDFS, MapReduce, Yarn), Airflow, Hive, Sqoop, HBase, PowerBI, MySQL, PostgreSQL, Spark, Cassandra, Scala shell, PySpark, Sqoop, Kafka, Oracle, hive, Zookeeper*

Responsibilities:

- Developed the **Spark Scala scripts** and **UDFs** to read from **Azure Blob Storage** to perform **transformations** on large datasets using **Azure Databricks**
- Created **scalable data intake pipelines** on **Azure HDInsight Spark clusters**, utilizing **Spark SQL** and **Cosmos DB (SQL API and Mongo API)** to store and process **structured and unstructured conversational data**, **intelligent AI agents** for **real-time analytics**
- Configured **Spark Streaming** to receive **real-time data** from **Apache Flume** and store the stream data using **Scala** to **Azure Table**, utilized **Spark Streaming API** to stream data from various sources and **optimized existing Scala code** for improved performance
- To preprocess and modify **massive datasets** stored in **Azure Blob Storage**, we created **AI-driven chatbot pipelines** with **Azure Databricks**, guaranteeing **effective answer generation** for **conversational AI models**
- By creating **ETL processes** in **Azure Data Factory**, moving data from **on-premises databases (MySQL, Cassandra)** to **Azure Blob Storage**, and using **PySpark transformations** to train and optimize **LLM-based chatbots** installed on **Azure SQL Data Warehouse**, **LLaMA-based NLP models** were integrated

- In order to **automate model serving** for **AI agents** and enable **low-latency LLM inference** for **chatbot applications**, **vLLM-deploy** was implemented on **Azure AKS**, leveraging **linked services** in **Azure Data Factory**
- Developed the **Spark DataFrames** from various **datasets** and applied **business transformations** and **data cleansing** operations in **Azure Databricks**
- Developed the **Python scripts** to build the **ETL pipeline** and **Directed Acyclic Graph (DAG) workflows** in **Airflow** and **Apache NiFi**
- Designed **custom-built input adapters** using **Spark** and **Hive** to **ingest and analyze data** in **Airflow**, then ingested the **enriched data** to **Snowflake**
- Utilized **Azure Data Factory** and **Airflow** to convert **enriched data** from different sources into **Snowflake**
- Migrated the existing **Oozie workflow** to **Apache Airflow** for **daily incremental loads**, getting data from **RDBMS**
- Implemented **Dimensional Data Modelling** to deliver **Multi-Dimensional STAR schemas** and developed **Snowflake Schemas** by **normalizing dimension tables** as appropriate
- Designed & implemented **Azure Subscriptions**, **data factories**, **Virtual Machines**, **SQL Azure Instances**, **SQL Azure DW instances**, and **HD Insight clusters**, and installed **DMGs** on **VMs** to connect to **on-premise servers**
- Managed **resources and scheduling** across the cluster using **Azure Kubernetes Service (AKS)**, creating, configuring, and managing a **cluster of Virtual Machines** to handle **online and batch workloads** for **analytics** and **machine learning applications**
- Used **Azure DevOps** and **VSTS (Visual Studio Team Services)** for **CI/CD**, **Active Directory** for **authentication**, and **Apache Ranger** for **authorization**
- Used **Scala** for **concurrency support** and developed **map-reduce jobs** using **Scala** to **compile program code** into **bytecode for the JVM** for **data processing**
- Proficient in utilizing **data** for **interactive Power BI dashboards** and **reporting** based on **business requirements**

Client: GAF **February 2023 – March 2024**
Role: Senior Data Engineer

Project Overview: GAF, a leading roofing and waterproofing manufacturer, built a scalable AWS data lake to combine data from SQL server, Hive and PostgreSQL. Leveraging Snowflake, MongoDB and Spark, the project optimized data transformation and also automated ETL pipelines with Apache Airflow, AWS Glue and Lambda. Fraud detection was made possible by real-time streaming using Kafka, enhanced business intelligence and decision-making with Tableau dashboards and query optimization.

Environment: *Amazon EC2, Amazon S3, Amazon ECS, Amazon Lambda, Amazon RDS, Amazon Elastic Load Balancing, Elastic Search, Amazon SQS, AWS Identity and access management, AWS Cloud Watch, Amazon EBS and Amazon CloudFormation, Scala,Spark,Mongo DB, Snowflake,Airflow,Pyspark, SparkSQl.Kafka, Tableau.*

Responsibilities:

- Worked on **Big Data Integration**, Analytics based on **Spark, Hive, PostgreSQL, Snowflake, MongoDB** and ingested the data into data lake from different sources and performed various transformations like sort, join, aggregations, filter to process various AWSs
- Implemented Data warehouse solutions in **AWS Redshift**, worked on various projects to migrate data from one database to **AWS Redshift**, **RDS, ELB, EMR, Dynamo DB and S3**
- Automated data flow between the software systems using **Apache Airflow** and migrated data from legacy Teradata to **MongoDB** built **ETLs** to load the data into **MongoDB**
- Built **ETLs** to load the data from **Presto, PostgreSQL, Hive, SQL** Server to **Snowflake** using **Apache Airflow,** Python and **Spark** and configured **Apache Airflow** with Python and Unix to submit the **Spark** batch jobs in EMR Cluster
- Developed the robust data pipelines for pulling the data from **SQL Server, Hive**. Landed the data in **AWS S3** and loaded into **Snowflake** after transforming and developed data processing triggers for **Amazon S3** using **AWS Lambda** with Python and **AWS Glue**
- Created ETL jobs using **Spark** to perform data migrations, data loads into **HDFS, Hive** from different source systems, provisioned multiple **Databricks clusters** needed for batch and continuous streaming data processing by installing the required libraries for the clusters

- Implemented **Spark** jobs for data preprocessing, validation, normalization, transmission and configured multiple **Spark** jobs to obtain efficient run time
- Implemented real-time data streaming with Kafka for distributed processing, custom topic polling, and anomaly detection. Built dashboards to track catalog updates, pricing, API calls, and fraud transactions, supporting data science teams in real-time model performance analysis
- Converted data from AWS S3 into **Snowflake** and used dimensional modelling techniques to organize it
- Developed **Snowflake** views to load and unload data from and to an **AWS S3 bucket**, as well as transferring the code to production
- Snowflake queries have been optimized for the improvement in reporting efficiency and made them quicker data retrieval for BI applications such as **Tableau** and **Power BI**
- Created interactive dashboards using Tableau to give leadership teams insights of their data and important KPIs
- Developed Executive Dashboards by collecting the requirement from department directors and stakeholders and mapped the data columns from source to target and performed analysis by querying to customize the data structure to align **tableau** visualization and special business requirements
- Designed and developed data quality framework utilizing **apache** beam to collect data quality metrics on, **Teradata, Mongo DB** supporting individual tables and entire databases/key space.
- Responsible for installing and configuring **Apache Hadoop** clusters and various tools **(Hbase, Redshift, Spark, Kafka, Sqoop, Hive, Kinesis**) on the **AWS cloud** and created various types of data visualizations using Python and **Tableau.**
- Developed and deployed Lambda functions in **AWS** using pre-built **AWS Lambda Libraries**, as well as Lambda functions in Scala using custom libraries and installed and configured **Apache Airflow** for **S3 bucket** and **Snowflake** data warehouse and created dags to run the **Airflow.**
- Worked on building Kubernetes template driven application deployment on **AWS** using **Spark** and **Big Data technologies** such as **PySpark, SparkSQL.**
- Deployed performance testing scripts programmed in **Scala**, **Kubernetes cluster,** to aggregate data from **Mongo DB cluster**, **Cloud Spanner, Beam SDK** to support digital platforms.

Client: Pena4 Tech Solutions India Private Limited **December 2020 – December 2022**
Role: Data Platform Engineer

Project Overview: The project built an extensible AWS data integration and analysis system using services like S3, EMR, Redshift, Athena and Glue for smooth ETL process and querying. AWS Glue and Airflow automated processes while Snowflake and Databricks improved big data processing. Real-time reporting was made achievable by using Tableau and Power BI, monitored them using CloudWatch, CloudTrail and SNS.

Environment*: Spark, Spark SQL, Python, Pyspark, Databricks, AWS services (EMR, Redshift, EC2, S3, Glue, Cloud watch, cloud trail, SNS, DynamoDB), Snowflake, Snow pipe, Shell scripting, MySQL, PostgreSQL, Enterprise DB, Jenkins, IntelliJ, Oracle, Git, Tableau.*

Responsibilities:

- Well versed in working with **Spark RDD, Data Frame API, Data set API, Data Source API, Spark SQL and Spark Streaming Spark Context, Spark-SQL, Data Frame, Pair RDD, and Spark YARN.**
- Worked on building **a** centralized **Data Lake** on AWS Cloud utilizing primary services like **S3**, **EMR**, **Redshift** and **Athena**
- Developed Spark Applications by using **Python** for data processing Projects to handle data from various **RDBMS** and Streaming sources
- Developed the **Pyspark** scripts to optimize the run time and the efficiency of the existing algorithms of Hadoop
- Experienced in data manipulation using python for loading and extraction as well as with python libraries such as **NumPy**, **SciPy** and **Pandas** for data analysis and numerical computations
- Configured & monitored the **Apache Airflow dags** to migrate the data from S3 bucket to **Snowflake data** warehouse
- Developed **Lambda** functions to create ad-hoc tables to add schema and structure the data in S3 and performed data validation, filtering, sorting and transformation for every data change in the **Dynamo DB** table and load the transformed data into **PostgreSQL** database
- Designed and developed **ETL** processes in **AWS Glue and** Python to migrate campaign data from external sources like S3, ORC/Parquet/Text files into **AWS Redshift**

- Improved **Snowflake** query performance to accommodate analytical workloads that require speed
- Developed Spark applications using **SPARK-SQL** in **Databricks** for data extraction, transformation, and aggregation from multiple file formats for analyzing & transforming the data to uncover insights
- Configured **Snow Pipe** to pull the data from **S3 buckets** and stored incoming data in the **Snowflakes** staging area and worked with micro batching to ingest millions of files on **Snowflake** cloud when files arrive to staging area
- Maintained **Data Marts** in Data Warehouse consisting of **Star Schema Schemas** (Facts/Dimensions) and **Snowflake Schemas** (Extended Star) utilizing Type II **Slowly Changing Dimensions** (SCD) for History Retention
- Used **Cloud watch**, **Cloud Trail** and **SNS** to monitor resources such **as EC2, CPU memory, Amazon DB services, Dynamo DB tables, Elastic Block Store (EBS)** volumes to set alarms for notification or automated actions and to monitor logs for a better understanding and operation of the system
- Developed the Python script using **Boto3** library to download file from AWS S3 bucket and utilized **Python script** in SSIS package for ETL processing
- Developed **Hive** Tables, with CREATE TABLE, LOAD DATA, ADD Partition, Hive Command Line and **HiveQL** with Data Storage on HDFS and S3
- Designed columnar families in Dynamo DB **architecture, replication strategy** and Ingested data from **RDBMS and AWS S3**
- Used **Spark QL** to analyze the **partitioned** and **bucketed data** and executed **Spark SQL queries on Parquet tables**
- Created data pipeline for different events such as ingestion, aggregation and load consumer response data in AWS S3 bucket to serve as feed for **Tableau dashboards**
- Created reports with complex calculations, designed dashboards for analyzing **POS** data and developed visualizations and drafted on Ad-hoc reporting **Tableau**
- Extensively worked with automation tools like **Jenkins** for continuous integration and continuous delivery (CI/CD) and to implement the End-to-End Automation

Client: Schemax Expert Techno Craft Pvt. Ltd. **July 2018 – November 2020**
Role: ETL Developer

Project Overview: With a focus on corporate applications, business intelligence and data integration, Schemax Expert Techno Craft Pvt. Ltd. offers robust solutions that facilitate productivity in operations and data-driven decision making. With the objective to improve data analysis and report, the project included creating interactive Tableau dashboards, obtaining data from various sources, and constructing reliable ETL pipelines with Informatica.

Environment: *Oracle, MySQL, SQLite, NO SQL, RDBMS, SQL Server, Strategy, SSIS, SSRS, SSAS, Qlikview, Tableau, Zeeplin Seaborn, Bokeh, ggplot, iplots, Shiny*

Responsibilities:

- Using **Informatica**, performed **ETL** on data from various heterogeneous data sources and destinations based on the business requirements.
- Created Tables, Stored Procedures, extracted data using **T-SQL** for business users to extract, load data and performed SQL queries.
- Developed conversion scripts using **SQL, PL/SQL**, stored procedures, functions, and packages to migrate data from SQL server database to Oracle database. Standardized QA standards and practices across teams where possible.
- Extracted data from different sources like Oracle, flat files, XML, DB2 and SQL Server loaded into **DWH**.
- Responsible for Designing, Development, and testing of the database and Developed Stored Procedures, Views, Triggers and developed Python-based **API** to track revenue.
- Created reusable utilities, programs in python to perform repetitive tasks such as sending emails, comparing data, and performed Backend **SQL** Data Testing on Oracle, **Teradata**, Sybase and DB2 database using **SQL/PLSQL** queries.
- Designed and developed all the staging tables needed to transform and store data from OLTP environment prior to export to **data warehouse**. Also created dimension, fact tables for **DWH**.
- Optimized current pivot tables' reports using **Tableau** and proposed an expanded set of views in the form of interactive dashboards using line graphs, bar charts, heat maps, tree maps, trend analysis, Pareto charts and bubble charts to enhance data analysis.

- Created dashboard style of layouts using various components of **QlikView** like List boxes, Multi boxes, slider, current selections box, buttons, charts, text objects, bookmarks
- Worked with various file formats (delimited text files, click stream log files, web server log files, **JSON files, XML Files**) and performed effective data cleaning without any data loss.
- Engaging with development teams, **QA**, Implementation, and others for providing deployment services from initial development through production deployments.
- Documented all build and release process related items. Level one support for all the build and deploy issues encounter during the build process.
- Monitoring metrics for usage thresholds and prepared weekly, monthly **OBIEE** test validation reports for management reviews and performed weekly **scrum** meetings.

Client: IMerit Technology Services Pvt. Ltd. **March 2017 – May 2018**
Role: Data Engineer

Project Overview: AI-driven solutions and data annotation are the areas of expertise for IMerit Technology Services Pvt. Ltd. In order to improve business insights, I developed machine learning models, automated data aggregation, and streamlined pipelines as a data engineer.

Environment: *SQL, MySQL, MS Office, Lucid chart, Jupyter, R 3.1, Python, SSRS, SSIS, SSAS, HBase, HDFS, Hive, Pig, Microsoft Office.*

Responsibilities:

- Used **MS Excel, MS Access,** and SQL to write and run various queries.
- Worked extensively on creating tables, views, and **SQL** queries in **MySql.**
- Worked with internal architects and assisted in the development of current and target state data architectures.
- Expertise in all areas of business operations to identify systems needs and requirements.
- Perform troubleshooting, fixed, and deployed many Python bug fixes of the two main applications that were the main source of data for both customers and the internal customer service team.
- Wrote Python scripts to parse **JSON** documents and load the data in the database.
- Worked on unstructured and structured data from multiple sources and automated the data analysis and aggregation using **Python scripts**.
- Extensively performed large data read/writes to and from CSV and Excel files using **pandas.**
- Performed Exploratory Data Analysis, trying to find trends and clusters.
- Built models using techniques like **Regression,** Tree-based ensemble methods, Time Series forecasting**, LSTM, LDA KNN**, Clustering.
- Analyzing various logs that are been generating and predicting/forecasting the next occurrence of an event with various **Python libraries.**
- Generating various capacity planning reports (graphical) using Python packages like **Numpy, matplotlib.**

EDUCATION

Master of Science in Computer Science
University of Texas at Arlington
Relevant Coursework: Artificial Intelligence, Big Data, Machine Learning, Data Mining, Web Data Management

CERTIFICATIONS

Salesforce Certified AI Associate
Salesforce Certified JavaScript Developer 1