Niharika Bharatula

Data Engineer

niharika.bharatula09@gmail.com | 214-809-7711 | TX | LinkedIn

SUMMARY

- **Over 5+** years of Industry experience as a Data Engineer with Good understanding of Business Requirements Gathering, Data warehousing, Data Modelling, Evaluating Data Sources, Translating Requirements into Specifications.
- Experience in development methodologies like **Agile** and **Waterfall**. Also, with Agile frameworks such as **Scrum**, Backlog grooming & Hands on with user stories. End to End experience in designing and deploying data visualizations using **Tableau**.
- Strong experience with **Python (2.x,3.x)** to develop analytic models and solutions.
- Good knowledge of Machine Learning, Mathematical Modelling and Operations Research.
- Strong knowledge on Cloud Technologies & **BI Tools** Such as **Terraform, AWS, AWS EMR, AWS Data pipeline, AWS Redshift, AWS (EC2, Lambda, VPC), MS Azure, Hadoop, Apache spark & Snowflake, Informatica, Teradata SSIS &SSRS**.
- Experience in creating visualizations, interactive dashboards, reports and data stories using **Tableau & Power BI**.
- Knowledge of relational databases – **MySQL, Oracle, Teradata, Microsoft SQL Server**.
- Familiar in applying Design & Implementation of Data Extraction, Transformation & Loading using **SQL/Python** analyse **MySQL** & Migration of the same. Comfortable with **Python** and Relational databases.
- Passionate about technology, interacting with **product owners** and **technical stakeholders**, delivering exceptional results with good teamwork skills.
- Expertise in **MS Office** suite especially **MS Excel, MS SharePoint, MS Word, MS PowerPoint, MS Visio & MS Project** and also change management tool **Google Suite.**

SKILLS

Methodologies: SDLC, Agile, Waterfall

Language: SQL, Python, Scala, PowerShell

Relational Database: MySQL, Oracle, Teradata, MS SQL Server, PostgreSQL, Snowflake

NoSQL Database: PL/SQL, T-SQL, Cassandra, Mongo DB.

Visualization Tools: Tableau, Power BI, Teradata SSIS, SSRS

Data Technologies: Hadoop, MapReduce, Spark, HDFS, Sqoop, Hive, Zookeeper, Apache, Apache Airflow, Cloudera, HBase.

Cloud Technologies: Amazon AWS (EMR, EC2, EBS, RDS, S3, Athena, AWS Glue, Elasticsearch, AWS Lambda, DynamoDB, Amazon Redshift, Amazon ECS, Quick Sight & Amazon Kinesis) & Microsoft Azure (Databricks, Azure Data Lake, Azure Blob Storage, Azure Data Factory, Azure SQL Database, Azure SQL Data Warehouse, Azure Cosmos DB & Azure Active Directory).

Hadoop Technologies: Apache Hadoop 2.x/1.x, Cloudera CDP and Hortonworks HDP**.**

Version Control Tools: Git, GitHub

Other Skills: Microsoft (Excel, Word, PowerPoint), Critical Thinking, Communication Skills, Presentation Skills, Problem- Solving.

Operating System: Windows, Linux

EXPERIENCE

BNY Mellon, TX | Big Data Engineer — Jan 2023 - Current

- Utilize Agile methodologies to monitor steer and develop project objectives.
- Strongly Functions with **Python, Scala, PowerShell & SQL** to develop data models and solutions.
- Created views in **Tableau** Desktop that were published to internal team for review and further data analysis and

customization using filters and actions. Executed in using various packages of libraries such as **Pandas, Py Torch** & **Py Spark**.

- Involved in building the **ETL** architecture and Source to Target mapping to load data into Data warehouse.
- Implemented data engineering and **ETL** solutions leveraging **CI/CD** software including **Jenkins, maven & GitHub**.
- Worked with **google data** & other **google cloud APIs** for monitoring, query & billing related analysis for **Big Query** usage.
- Designed various Jenkins jobs to continuously integrate the processes and executed **CI/CD** pipeline using Jenkins.
- Extracted Bullet graphs to determine profit generation by using measures and dimensions data from **MySQL & MS Excel**.
- Developed and support **Scalable ETL** components to aggregate and move data from a variety of structured and unstructured data sources to the **Data warehouse/Snowflake.**
- Lead the development of an automated **DEVOPS CI/CD** solution using GitHub. Jenkins & python to improve unit testing program modules to guarantee logic works correctly and to ensure application performs optimally with regards to business requirement.
- Used Cloud technologies as per client requirements as **Terraform & Google Cloud Platform (GCP**).
- Plans and coordinates the administration with relational Database such as **MySQL, Oracle, Teradata, Microsoft SQL Server & PostgreSQL**, to ensure accuracy, appropriate and effective use of data, including database definition, structure, documentation, long-range requirements and operational guidelines.
- Worked with Microsoft products such as **SQL Server Reporting Services (SSRS**) and **Power BI** to develop dashboards and visualizations for business users and upper management.
- Serve as a subject matter expert in transforming large amounts of data and creating business intelligence reports, using state-of-the-art big data technologies such as **Hive, Spark, Scoop and NIFI** for data ingestion, as well as Python/Bash scripting and Apache Airflow for scheduling jobs in **Google's cloud-based environments** (GCP).
- Successfully migrated an **Oracle SQL ETL to run on Google Cloud Platform (GCP**) using cloud Data proc and Big Query, as well as Cloud Pub/Sub for triggering airflow jobs.

Capital One, VA | Data Engineer Jan 2022 – Dec 2022

- Utilized **Apache Spark** with **Python** to develop and execute **Big Data Analytics and Machine learning** applications, executed machine Learning use cases under **Spark ML and Matplotlib**.
- Filtering and cleaning data using **Scala code and SQL Queries**.
- Implemented Installation and configuration of multi-node cluster on **Cloud** using **Amazon Web Services (AWS)** on **EC2**.
- Developed Automation Regressing Scripts for validation of ETL process between multiple databases like **AWS Redshift, Oracle, MongoDB, T-SQL,** and **SQL Server** using **Python**.
- Worked with **Hadoop** infrastructure to storage data in **HDFS storage** and use **Spark / HIVE SQL** to migrate underlying **SQL** codebase in **AWS**.
- Collected data using **Spark** Streaming from **AWS S3** bucket in near-real-time and performs necessary Transformations.
- Developed **Spark/Scala, Python** for regular expression (regex) project in the **Hadoop/Hive** environment with Linux/Windows for big data resources. Export tables from **Teradata** to HDFS using **Sqoop** and build tables in **Hive**.
- Data sources are extracted, transformed & loaded to generate CSV data file with **Python** programming & **SQL queries**.
- Have worked on **Spark, Scala/Java/Python, Python, REST, JSON, NoSQL databases, relational databases, Jenkins/Maven, Cloud Infrastructure** to name a few. Analyzing **SQL** scripts and designed the solution to implement using **Py Spark**.
- Use **Spark SQL** to load **JSON** data and create Schema **RDD** and loaded it into **Hive Tables** and handled structured data using **Spark SQL**. Have experience using **Gitlab/GitHub.**
- Design, build and operationalize large scale enterprise data solutions and applications using one or more of **AWS data** and analytics services in combination with 3rd parties **- Spark, EMR & Lambda**.

BCBS, TX | Data Engineer May 2021 – Dec 2021

- Worked extensively on **Azure Data Profiling, Data clustering, Data Mapping, Data Visualizations and Data Quality.**
- Created **Tableau** Dashboards with interactive views, trends, and drill downs along with user level security.

- Maintained the smooth frequency with **Azure** packages as **Azure Data Lake, Azure Data Factory, Azure Databricks, Azure Synapse Analytics, Azure Data Migration & Apache Spark**.
- Used **Pandas, Py Torch & Py Spark** in **Python** libraries for developing various machine learning algorithms.
- Wrote the **MySQL** with **NoSQL DB** queries on data staging tables and **data warehouse** tables to validate the data results which includes **PL/SQL, T-SQL, Cassandra & Mongo DB.**
- Performed **Data Clustering** and **Data visualization** using **Py Torch** and **Scikit** learn and used it to perform feature selection.
- Scheduled data refresh on **Tableau** Server for weekly and monthly increments based on business change to ensure that the views and dashboards were displaying the changed data accurately.
- Generated and ran basic **SQL** queries and data collection spreadsheets to support assigned business areas.
- Functioning with dirty data from various sources like **MySQL, Oracle, Teradata, Microsoft SQL Server, PostgreSQL** & performing **data cleansing** using ETL procedure. Using rest **API** with **Python** to ingest Data from and some other site to **BIGQUERY**.
- Used **Agile** with **Scrum** methodology throughout the project, Involved in weekly and daily basis release management.
- Created a data warehouse model for over 100 datasets in **Snowflake** using **Where cape** and built reports in Looker based on **Snowflake** connections. Interpreted requirements and modelled attributes from various source systems stored in **Oracle, Teradata** and **CSV files**, then loaded them into the **Teradata Warehouse**.
- Conducted data extraction and Exploratory data analysis, seeking to understand the reasons behind the data.
- Possessed hands-on experience in implementing LDA, Naive Bayes, and expertise in **Random Forests, Decision Trees, Linear** and **Logistic Regression, SVM, Clustering, neural networks** and Principal Component Analysis.

Dell Technologies, India | Data Engineer **Jan 2017 – Dec 2019**

- Utilized **Apache Spark** with **Python** to develop and execute **Big Data Analytics and Machine learning** applications, executed machine Learning use cases under **Spark ML and Matplotlib**.
- Aggregation on the fly to build the common learner data model and persists the data in **HDFS**.
- Developing **Spark** programs with **Python**, and applied principles of functional programming to process the complex structured data sets.
- Maintained the smooth frequency with **Azure packages** as **Azure Data Lake, Azure Data Factory, Azure Databricks, Azure Synapse Analytics, Azure Data Migration & Apache Spark**.
- Generate **metadata**, create Talend ETL jobs, mappings to load **data warehouse & data lake**.
- Designed and Developed Real Time Stream Processing Application using **Spark, Kafka, Scala** and **Hive** to perform Streaming ETL and apply Machine Learning.
- Reduced access time by refactoring information models, query streamlining and actualized **Redis** store to help **Snowflake**.
- Converting **Hive/SQL** queries into **Spark** transformations using **Spark RDDs** and **Py spark.**
- Loaded and transformed large sets of structured, semi structured and unstructured data using **Hadoop/Big Data concepts.**
- Developed reusable objects like **PL/SQL** program units and libraries, database procedures and functions, database triggers to be used by the team and satisfying the business rules.
- Worked on **Continuous Integration CI/Continuous Delivery (CD)** pipeline for **Azure Cloud** Services using **CHEF**.
- Developed automated processes in **Azure cloud** for ingesting data daily from web services and loading it into **Azure SQL DB**, and analyzed data using **Azure Data Lake**, **Blob** and **Databricks.**
- Used Cloud technologies as per client requirements as **Terraform, MS Azure HDInsight, Hadoop & Apache spark.**
- Designed **Kafka** producer client using Confluent **Kafka** and produced events into Kafka topic.
- Participated in **Azure beta programs**, offered beta programs for new **Azure** features by Microsoft.

EDUCATION

Masters in Data Science **2021**
University of North Texas, TX

Bachelors in Computer Science **2018**
Guru Nanak Institution of Technical Campus (GNITC), Hyderabad India