**Deekshitha**

**Data Engineer**

**(737) 800-0869 ||** [**dekshethar@gmail.com**](mailto:dekshethar@gmail.com) **|| TX**

**PROFESSIONAL SUMMERY:-**

* **9 years** of demonstrated experience in IT industry with expert-level skills in **Big Data Hadoop** Ecosystem, Apache Spark, **PySpark, Scala, Python, Kafka, Data Warehousing, Data Pipeline, Business Intelligence, Snowflake, Data Analytics.**
* Proficient in Azure – Data Lakes, Amazon Web Services (AWS) – EC2, S3, EMR, ETL, Informatica, **Google cloud Platform** (GCP) Glue and Presto and Data bricks.
* Expert level knowledge **on Hadoop Distributed File System** (HDFS) architecture and YARN.
* Experienced on Hive partitioning, bucketing, optimization code through set parameters and perform different types of joins on Hive tables and implement Hive SerDe like **Avro, JSON**.
* Extensively utilized Hive for Processing and Analyzing logs, joining large tables, Batch Jobs, HiveQL for Ad-hoc interactive queries to summarize and analyze large data sets.
* Experienced in importing and exporting data from different databases like **SQL Server, Oracle, Teradata and Netezza.**
* Extensive experience in leveraging data serialization formats like AVRO, Protocol Buffer and Columnar formats like **RCFile, ORC** and **PARQUET** file formats.
* Experienced in automating Oozie workflows and Job Controllers for job automation – Shell, Hive, Sqoop and email notifications.
* Experience in **Zookeeper configuration** as to provide cluster coordination services.
* Extensive experience in **creating RDD’s** and **Data Sets** in Spark from local file system and HDFS.
* Hands on experience in writing different **RDD** (Resilient Distributed Datasets) transformations and actions using Scala.
* Experience in analyzing data using R, SQL, Microsoft Excel, Have, **Pyspark, Spark SQL** for data Mining, **Data cleansing** and **machine learning.**
* Experience in Extraction, Transformation and Loading (ETL) of data from multiple sources like Flat files and Databases.
* Build data pipelines in airflow in **GCP** for ETL related jobs using different airflow operators.
* Created **Data Frames** and performed analysis using **Spark SQL** and used RDD and DF APIs to access variety of data sources using **Scala, PySpark, Pandas** and **Python**.
* Excellent knowledge on **Spark core architecture**.
* Created ERM transient and long running clusters and in AWS for data processing (ETL) and log analysis.
* Deployed various **Hadoop** applications in EMR - **Hadoop, Hive, Spark, HBase, Hue, HCatalog, Glue**, Oozie and Presto etc. based on the needs.
* Experience in integrating Hive with AWS S3 to read and write data from and to S3 and created partitions in Hive .
* Extensively worded on **ETL/Data pipelines** to transform data and load from **AWS** S3 to Snowflake or viceversa.
* Extensively utilized **EMRFS** (Elastic Map Reduce File System) for reading and writing SerDe data between **HDFS** and **EMRFS.**
* Experience in using **Presto on EMR** to query different types of data sources **RDBMS** and **NoSQL** databases.
* Experience in creating ad-hoc reporting, development of data visualizations using enterprise reporting tools like **Tableau**, **Power BI,** **Business Objects** and **Alteryx**.

**TECHNICAL SKILLS-**

# **Big Data Tools :** Hadoop, Hive, Apache Spark, PySpark, HBase, Kafka, Pig, Map Reduce, Zookeeper and Flume.

# **Cloud Technologies :** Azure (Data Bricks, Azure Data Factory, Azure Data Lake, Azure pipelines, Azure Functions, Blob storage)AWS (EC2, S3 Bucket, Amazon Redshift, Lambda, IAM, Kinesis), Snowflake, GCP(Big Query, Cloud SQL, Cloud Storage, Cloud SDK, Cloud APIs, and other tools like Dataflow, Dataproc, Data prep, Data Studio).

# **ETL Tools :** SSIS,DBT, Informatica.

# **Relational** **Databases :** MS SQL Server, MySQL, Oracle, PostgreSQL, Netezza.

* **No SQL Databases** :Cassandra, MongoDB, HBase

# **Programming Language:** Python, R, Scala, JSON, HTML.

* **Scripting**  : Python, Shell scripting

# **IDE’s :** PyCharm, Jupiter Notebook.

* **Build Tools :** Apache Maven and SBT, Jenkins, Bitbucket
* **Version Control :**GIT, SVN
* **CI/CD** : Jenkins, Azure .

# **Machine Learning :** Linear Regression, Logistic Regression, Decision Tree, SVM, KNN, K mean.

# **Packages :** NumPy, Pandas, Matplotlib, Scikit-learn, Seaborn, PySpark.

# **Reporting Tools :** Tableau, Power BI, SSRS.

# **Operating Systems :** Windows, Linux, MacOS.

**WORK EXPERIENCE**

**Fannie Mae, Plano, Texas||Azure Data Engineer Aug 2023 – Till Date**

Project Descriptions: In this Project, Client Application provides the customer details experience through real time connecting with customers in various ways. The objective is to serve the customer more efficiently by conducting user research and usability studies to understand how customers are interacting, utilizing with the application, client services and building a highly scalable, high-availability, and high performance.

Responsibilities:

* Develop standardized Azure Data Factory pipelines for ingesting diverse data sources into Azure Data Lake Storage (ADLS).
* Establish metadata framework within Azure Data Factory for improved data management and organization.
* Parameterize linked services and datasets in Azure Data Factory to enhance pipeline flexibility and reusability.
* Utilize Databricks notebooks with PySpark to register and transform raw data into structured formats.
* Write SparkSQL transformations in Databricks notebooks to facilitate data movement across different layers in ADLS and Databases.
* Implement automation workflows using Azure Logic Apps and Automation Runbooks for efficient task management.
* Contribute to SQL Server database development and optimization.
* Perform performance tuning and query optimizations to enhance database efficiency.
* Involved in Snowflake data warehouse migration, managing External Stages, Tables, Stored Procedures, and Views.
* Utilized SnowSQL &amp; Snowpark for Data transformations within Snowflake Data warehouse.
* Employed Snowpipe for real time data ingestions from Data Lake into Snowflake Data warehouse.
* Provide hands-on training and mentorship to onboard new team members, fostering knowledge sharing.
* Collaborate closely with users to identify and resolve issues, ensuring enhanced user experience.
* Experienced in writing real-time processing and core jobs using Spark Streaming with Kafka as a data pipeline system.
* Migration of on-premises data (SQL Server / MongoDB) to Azure Data Lake Store (ADLS) using Azure Data Factory (ADF V1/V2).

**Environment:** Azure Data factory, Azure Databricks, Azure SQL, Synapse, Data Lake, Snowflake Data Warehouse, Kafka, MS SQL Server, SSRS, SQL Server Integration Services (SSIS), Microsoft Visual Studio, SQL Server Management Studio, Jenkins, PL/SQL, T-SQL, Spark eco-system, Pyspark, Big Data, Agile methodology, Agile safe, Kanban, Scrum.

**Change Healthcare, Lombard, IL || Data Engineer Dec 2022 – July 2023**

Project Descriptions: Project majorly focuses on expanding and optimizing data and data pipeline architecture, as well as building and maintaining data workflow, designing optimal ETL data pipeline and infrastructure required for optimal extraction, transformation, and loading of data from a wide variety of data sources. As a Data Engineer, involved in maintaining the huge data and designing developing predictive data models for business users according to the requirement.

Responsibilities:

* Handled importing of data from various data sources, performed data control checks using **PySpark** and loaded data into HDFS.
* Developed python scripts, UDFs using both Data frames/SQL/Data sets and RDD/Kafka in Spark 1.6 for data Aggregation, queries and writing data back into OLTP system through Sqoop.
* Experienced in handling large datasets using Partitions, PySpark in Memory capabilities, Broadcasts in PySpark, effective & efficient Joins, Transformations and other during ingestion process itself.
* Involved in writing live Real-time Processing and core jobs using Spark Streaming with Kafka as a data pipe-line system.
* Implemented robust error handling and retry mechanisms within Lambda functions, ensuring fault tolerance and reliability of serverless applications.
* Developed and maintained IAM policies and security configurations to enforce compliance with industry standards (such as PCI DSS, HIPAA, or GDPR) and internal security policies, ensuring data confidentiality and integrity.
* Used Spark API over Cloudera Hadoop YARN to perform analytics on data in HDFS.
* Streaming pipeline that uses PySpark to read data from Kafka, transform it and write it to HDFS.
* Designed Data Marts by following Star Schema and Snowflake Schema Methodology, using Data Modeling tools.
* Worked on Snowflake database on queries and writing Stored Procedures for normalization.
* Worked with Snowflake’s stored procedures, used procedures with corresponding DDL statements, used JavaScript API to easily wrap and execute numerous SQL queries.
* Involved in performing unit testing and integration testing.
* I have substantial work experience in software projects development life cycle, utilizing the core principles of Agile methodologies.
* I have experience in closely collaborating with offshore development and production support teams.
* I have experience collaborating closely with offshore development and production support teams, with responsibilities that include gathering daily status updates and conveying them to senior leadership.

**Environment:** AWS Glue, Snowflake, HDFS, Hive, Kafka, Spark 1.8, Linux, Python 2, SQL Server Database, Jira, Service Now, Confluence, Agile methodologies (SCRUM Framework), AWS (EC2, S3, EMR, Lambda, Step Function).

**RYAN, Hyderabad, INDIA || GCP Data Engineer Oct 2018 – July 2021**

Project Description: The project main center is optimizing data management, processing, and analysis to improve the efficiency and accuracy of tax related operations. To enhance the company’s services, data infrastructure and capabilities to streamline tax related processes, ensure compliance with regulatory requirements, and provide better insights for clients.

Responsibilities:

* Developed multiple data pipelines using cloud services, worked on Map reduce for data distribution to reduce the data load.
* Hands on experience in Google Cloud as Cloud Storage, Dataflow, Cloud Composer, Bigquery, Cloud Functions, Cloud Pub/sub and Dataproc.
* Experience in IBM Console for monitoring of streaming jobs.
* Experience in testing the data through streaming jobs for Events and Outages.
* Writing Python scripts to load the data from Bigquery to Bigquery using Dataflow and Composer.
* Experience in data validation and analysis for Prod defects.
* Running Cron Jobs using the Omega Data to GCP and checking the logs in Omega.
* Experienced in GCP Jobs and Egress jobs testing once the migration is done.
* Experience in writing and creating Hive tables in Omega and data validation.
* Working experience with Support team and took the responsibility for the issues in production.
* Experience in solving priority issues and involving in SOC calls while there is any production issues.
* Handling all priority incidents created by the end users and providing the solution on time via Service Now.
* Experience in creating Teradata scripts, pyspark scripts to load the data from Teradata tables to Hadoop.
* Responsible for L2 support for environment and application related issues.
* Handling Change Requests and Service Requests.
* Troubleshooting production issues under client defined SLA’s.
* Responsible for Google Production Support for environments and application related issues.
* Service Now applications implemented: Incidents, Change Requests, Service Requests, Configuration, Dashboards.
* Involving with business teams and different teams to solve the critical issues.

**Environment:** GCP, Cloud SQL, Big Query, Cloud DataProc, GCS, Cloud SQL, Cloud Composer, Hadoop, Hive, Map Reduce,Teradata, SAS, Teradata, Spark, Python, SQL Server, Service Now, Confluence, IBM Console.

**Ceequence Technologies Hyderabad, India || Jr. Data Engineer Aug 2014 – Sep 2018**

Project Description: The primary goal of the project is to focus on the collection, integration, and analysis of data from different sources. This focus will result in greater insights and more effective support for decision making. creating Corporate Data Warehouse and migrating data from the OLTP systems to the Corporate Data Warehouse. SSIS was used as an ETL tool for extracting the Data from various sources running on Oracle, DB2 and MS SQL Server databases and Generate reports on Power BI to cover weekly, monthly, quarterly and annual historic information.

Responsibilities:

* Created action filters, parameters, and calculated sets for preparing dashboards and worksheets using Power BI.
* Developed Snowflake views to load and unload data from and to an AWS S3 bucket, as well as transferring the code to production.
* Developed visualizations and dashboards using Power BI.
* Performing ETL testing activities like running the Jobs, Extracting the data using necessary queries from database transform, and upload into the Data warehouse servers.
* Created dashboards for analyzing POS data using Power BI.
* Converting Hive/SQL queries into Spark transformations using Spark data frames, Scala Python.
* Running Spark SQL operations on JSON, converting the data into a tabular structure with data frames, and storing and writing the data to Hive and HDFS.
* Developing shell scripts for data ingestion and validation with different parameters, as well as writing custom shell scripts to invoke spark Employment.
* Tuned performance of Informatica mappings and sessions for improving the process and making it efficient after eliminating bottlenecks.
* Worked on complex SQL Queries, Cassandra procedures and convert them to ETL tasks.
* Worked with PowerShell and UNIX scripts for file transfer, emailing and other file related tasks.
* Created a risk-based machine learning model (logistic regress, random forest, SVM, etc.) to predict which customers are more likely to be delinquent based on historical performance data and rank order them.

**Environment:** MS-SQL Server, SQL Server Integration Services (SSIS), Import and Export Data wizard, TFS, SQL Server Reporting Services, Power BI, SQL Server Analysis Services (SSAS),SQL Profiler, Python 3.0, SSIS, Spark.