**Deekshitha S**

**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, Matillion, **Google cloud Platform (GCP)** Glue and Presto and **Data bricks, Snowflake.**
* 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.**
* Expertise in **SSAS, SSIS, SSMS, SSRS** for Data migration, Analysis, Validation.
* 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, Data Factory, Data Lake, pipelines, Blob storage, AWS (EMR, S3 Bucket, Amazon Redshift, Lambda, IAM, Kinesis), Snowflake (Snowflake Data warehouse, Snowpipe), Matillion, GCP(Big Query, Cloud SQL, Cloud Storage, Cloud SDK, Cloud APIs, and other tools like Dataflow, Dataproc, Data prep, Data Studio).

# **Containerization / Orchestration**: Dockers,Kubernetes, Apache Airflow.

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

# **Databases :** MS SQL Server, My SQL, Oracle, Teradata, PostgreSQL.

* **No SQL Databases** :Cassandra, MongoDB, HBase
* **Querry Languages** : SQL, T-SQL, Pl/SQL, SSAS.

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

* **Scripting**  : Python, Power Shell scripting

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

* **Build Tools :** Apache Maven and SBT, Jenkins, Bitbucket
* **Version Control :**GIT, SVN
* **CI/CD** : Jenkins, Azure .
* **CRM tool** : Salesforce.
* **Software Tools** :CLS, LOS, LMS

# **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|| 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:

* Leveraged AWS Glue for Data integration, data cataloging, schema discovery, and automated ETL job generation, reducing manual effort and improving efficiency.
* Handled importing of data from various data sources, performed data control checks using PySpark and loaded data into HDFS.
* Designed and implemented data integration solutions using APIs and data connector for seamless data flow between CLS and external systems, enhancing overall data interoperability and operational efficiency.
* Processed the data using AWS Glue, AWS Lambda, and Amazon Kinesis for real-time and batch data.
* Utilized AWS Lambda for data processing tasks like Data Validation, Enrichment, real-time event processing.
* Used AWS Managed Streaming for Apache Kafka (MSK) to handle high-throughput data ingestion.Involved in Real-time data processing, core jobs using Spark Streaming with Kafka as a data pipe-line system.
* leveraged Scala libraries and frameworks to interact with databases (e.g., JDBC), message queues (e.g., Kafka), cloud storage (e.g., AWS S3, Google Cloud Storage), and other data sources or APIs.
* Designed Data Marts by following Star Schema, Snowflake Schema Methodology, using Data Modeling tools.
* wrote Scala code to connect to data sources, extract data, and load it into Spark's Resilient Distributed Dataset (RDD) or Data Frame for further processing.
* Implemented data warehousing solutions on Amazon Redshift, optimizing query performance and reducing data retrieval times.
* Designed and implemented ETL processes to consolidate data from various CLS into a centralized data warehouse.
* Involved in performing unit, System, Regression, integration testing.
* Integrated LOS with other financial systems by engineered robust data integration solutions using APIs and data connector for data exchange.
* Monitored and managed Kafka infrastructure using AWS CloudWatch, AWS CloudTrail, and custom monitoring solutions to ensure system health and performance.
* Worked 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.

**Environment:** AWS (Redshift, DynamoDB, S3, EMR, Lambda, Kinesis, IoT)**,** Dockers,Kubernetes, AWS Glue, CLS, LOS, LMS, Kafka, Spark, Python, MY SQL, SQL, T-SQL, Jira, Service Now, Scala, Agile methodologies.

**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:

* Develop standardized Azure Data Factory pipelines for ingesting diverse data sources into (ADLS).
* Established metadata framework within ADF for improved data management.
* Applied Scala functions for the data filter, Mining, aggregate, and join datasets, applying business logic and data transformation rules as needed.
* Parameterize linked services, datasets in ADF to enhance pipeline flexibility, 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.
* Integrated T-SQL Scripts with SSIS to automate the data processing workflow, Streamline data pipeline orchestration.
* Implement automation workflows using Azure Logic Apps for efficient task management.
* Contribute to SQL Server database development and optimization.
* Developed robust ETL (Extract. Transform, Load) processes using SSIS to efficiently move data from various source systems into the data warehouse.
* Perform performance tuning and query optimizations to enhance database efficiency.
* Involved in Snowflake data migration, managing External Stages, Tables, Stored Procedures, and Views.
* Processed the HDFS framework within the Spark eco system for seamless data integration.
* Experienced in writing real-time processing and core jobs using Spark Streaming with Kafka as a data pipeline system.
* Interacted with cluster management systems like Apache Hadoop YARN or Kubernetes to deploy and manage Spark applications using Scala.
* Facilitated with Git Lab / Git Hub for Collaboration with team members, CI/CD, Issue tracking, Documentation.
* Migration of on-premises data (SQL Server / MongoDB) to ADLS using ADF.

**Environment:** Azure (Data factory, Databricks, SQL data warehouse, Synapse, Data Lake), Map Reduce, HDFS, MS SQL Server, SSIS, Microsoft Visual Studio, SQL Server Management Studio, Dockers,Kubernetes, PL/SQL, T-SQL, GIT, Pyspark, Scala, Big Data, Agile methodology.

**RYAN, Hyderabad, INDIA || GCP Data Engineer Sep 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.
* Worked on Cloud Storage, Dataflow, Cloud Composer, Bigquery, Cloud Pub/sub and Dataproc.
* 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.
* Migrated on-premises data infrastructure to GCP, reducing operational costs by [percentage] and improving system reliability by implementing fault-tolerant architecture.
* Running Cron Jobs using the Omega Data to GCP and checking the logs in Omega.
* Designed and implemented scalable data pipelines on Google Cloud Platform (GCP) using tools such as Dataflow, Dataproc, and Apache Beam, achieving [specific metrics or goals].
* Experience in writing and creating Hive tables in Omega and data validation.
* Developed and optimized BigQuery data models and queries to support complex analytical and reporting requirements, resulting in [specific improvements in query performance or data accessibility].
* 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 & providing the solution on time via Service Now.
* Experience in creating Teradata scripts, to load the data from Teradata tables to Hadoop.
* Responsible for L2 support for environment and application related issues.
* Implemented CI/CD pipelines for automated deployment of data pipelines and infrastructure as code using tools like Cloud Composer (Airflow) and Cloud Build, reducing deployment time by [percentage].
* Handled Change Requests and Service Requests.
* Designed and implemented data security best practices on GCP, including encryption at rest and in transit, IAM policies, and auditing, achieving compliance with [specific industry standards or regulations].
* Collaborated with cross-functional teams to gather and analyze business requirements, translating them into technical specifications and scalable data solutions on GCP.
* Conducted performance tuning and optimization of GCP services to meet SLAs, including troubleshooting and resolution of data processing bottlenecks and issues.

**Environment:** GCP, BigQuery, Dataflow, Dataproc, Pub/Sub, BigTable, Firestore, Dataflow, Apache Beam, Cloud Composer (Airflow), Hadoop, Hive, Map reduce, Teradata, Airflow, Jenkins, Python, SQL, T-SQL, SQL Server, Service Now.

**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:

* Designed, Deployed, and Managed data solutions using Azure cloud services such as ADF, Azure SQL Database, Azure Databricks, Azure Synapse Analytics (formerly SQL Data Warehouse), and Azure Cosmos DB.
* Built and Maintained data pipelines and ETL processes using Azure Data Factory to orchestrate data movement and transformation across various sources and destinations.
* worked with both SQL and NoSQL databases such as SQL Database, Cosmos DB, and Table Storage, ensuring optimal database design, performance tuning, and data management.
* Performed data warehousing solutions using Azure Synapse Analytics (formerly SQL Data Warehouse), including schema design, data partitioning, indexing strategies, and optimizing query performance.
* Fabricated data modeling techniques for OLAP (Online Analytical Processing) and OLTP (Online Transaction Processing) environments, ensuring efficient data retrieval and analysis capabilities for reporting and business intelligence purposes.
* Implemented data encryption, access controls, and auditing to ensure data protection, security and regulatory compliance.
* Wrote PL/SQL, SQL Queries for Data Analysis, Data Validation, Data Transformation.
* Performed ETL process using Informatica for on premises data to Oracle Data base for Data process.
* Monitored Azure data solutions for performance metrics, troubleshooting issues, and implementing optimization strategies to improve data processing efficiency and reduce costs.
* Created action filters, parameters, and calculated sets for preparing dashboards and worksheets using Power BI.
* Developed visualizations and dashboards using Power BI.
* Documented skills and experience collaborating with cross-functional teams, stakeholders, and business users to gather requirements, define data architecture, and deliver scalable data solutions on Azure.

**Environment:** Azure, Oracle, SQL, PL/SQL, Informatica, SSIS, Import and Export Data wizard, TFS, Power BI.