**DATA ENGINEER**

**Name:** Akhil Reddy

**Email:** [akhilko1214@gmail.com](mailto:akhilko1214@gmail.com)

**Phone:** **(940)-843-8773 LinkedIn:** [**https://www.linkedin.com/in/akhilkreddy/**](https://www.linkedin.com/in/akhilkreddy/)

**Data Engineer, Hadoop & Python, Spark Developer**

Data Analysis, Data Profiling, Data Integration, Migration, Data governance, metadata management, and Master Data Management and Configuration Management

**PROFESSIONAL PROFILE:**

* Having 9+ years of IT industry experience as a Data Engineer with solid understanding of Data Modelling, Data Validation, Evaluating Data Sources, and strong understanding of **Data Warehouse / Data Mart Design, Data** **Architects**.,**ETL, BI, OLAP** and Client/Server applications on **AWS** and **Azure**.
* **Solid Knowledge of AWS services like AWS EMR, Redshift, S3, EC2, and concepts, configuring the servers for auto-scaling and elastic load balancing**.
* Expertise in using **AWS S3** to stage data and to support data transfer and data archival. Experience in using **AWS Redshift** for large scale data migrations using **AWS DMS** and implementing CDC (change data capture).
* Expert in writing **Python** scripts, Complex **SQL** queries and optimizing the queries in **Oracle, SQL** Server, and Teradata. Excellent **Software Development Life Cycle** (SDLC) with good working knowledge of testing methodologies, disciplines, tasks, resources, and scheduling.
* Excellent knowledge in **Data Analysis**, **Data Validation**, **Data Cleansing**, **Data Verification** and identifying data mismatch.
* Performed data analysis and data profiling using complex **SQL** on various sources systems including **Oracle** and **Teradata**.

Proficient in all phases of SDLC and currently working in conditions utilizing Agile (SCRUM), Waterfall, Product mangers ,synapase and Test- Driven Development (TDD) advancement approaches.

* Extract, Transform and load data from sources systems to **Azure Data Storage** services using a combination of **Azure data factory, T-SQL, Spark SQL.**
* Experience in **Big data Hadoop, Spark, Python, SQL, Power BI, Tableau** andother **Data Engineering tools** and **Azure/AWS platform cloud services.**
* Experience in **Data Warehousing**, building **ETL pipelines, dashboarding and Big Data Technologies**.
* Data ingestion to one or more **Azure services (Azure Data Lake, Azure storage, Azure SQL)** and processing data in Azure Data bricks and Azure Data Factory (ADFv2).
* **Hands on experience on Azure Cloud Technologies like Azure SQL Database, Azure Data Factory, Azure Databricks, Azure Data Lake store, Azure Data Lake Analytics, Azure Synapse Analytics, Azure storage accounts (Gen1 & Gen2), Azure Key Vault, IoT Hub.**
* Expertise in **Data Migration, Data Profiling, Data Ingestion, Data Cleansing**, **Transformation, Data Import,** and **Data Export** using multiple **ETL** **tools** such as **Informatica** **Power Centre.**
* Working knowledge of **Spark RDD, Data frame API, Data set API, Data Source API, Spark SQL** and **Spark Streaming**.
* Experience in exporting as well as importing the data using **Sqoop** between **HDFS** to Relational Database systems and vice-versa and load into **Hive tables**, which are partitioned.
* Worked on **HQL** for required **data extraction and join operations** as required and having good experience in optimizing **Hive Queries.**
* Experience in **Partitions, bucketing** concepts in **Hive**, and designing both **Managed** and **External tables** in **Hive** to optimize performance.
* Experience in working with business intelligence and data warehouse software, including **SSAS/SSRS/SSIS, Business Objects, Amazon Redshift, Azure Data Warehouse,** and **Teradata.**
* Profound experience in creating real-time data streaming solutions using **Apache Spark /Spark Streaming, Kafka, and Flume.**
* Responsible for identifying and implementing new SQL Server features designed to improve query performance; this included utilizing new T-SQL language enhancements, performing in-memory optimization, and performance tuning with Extended Events.
* Used **Apache** **NiFi** to automate the data movement between different Hadoop Systems.
* Good experience in handling messaging services using **Apache Kafka**.
* Knowledge in Data mining and Data warehousing using **ETL** Tools and Proficient in Building reports and dashboards in **Tableau** (BI Tool).
* Worked parallelly in both **GCP** and **Azure** clouds coherently.
* Have Extensive Experience in IT data analytics projects, Hands on experience in migrating on premise **ETL**s to **Google Cloud Platform (GCP)** using cloud native tools such as **BIG query**, **Cloud Data Proc**, **Google Cloud Storage**, **Composer**.
* Excellent knowledge of job workflow scheduling and locking tools/services like **Oozie** and **Zookeeper.**
* Good understanding and knowledge of **NoSQL** databases like **HBase** and **Cassandra.**
* Good understanding of **Amazon Web Services (AWS)** like **EC2** for computing and **S3** as a storage mechanism and **EMR**, **Step functions, Lambda, Redshift, DynamoDB**.
* Hands-on working experience on Microsoft **Azure** services like **HDInsight** Clusters, **BLOB**, **ADLS**, **Data Factory** and **Logic Apps.**

**TECHNICAL SKILLS:**

|  |  |
| --- | --- |
| **Hadoop Components / Big Data** | PySpark, Airflow, HDFS, MapReduce, Hive, HCatalog, HBase, Sqoop, Impala, Zookeeper, Kafka, Yarn. |
| **Programming Languages** | Python, Scala, SQL, PySpark, PowerShell, T-SQL. |
| **Cloud Platform** | AWS (Lambda, S3, EC2, EMR, RDS), Microsoft Azure (Azure Databricks, Azure Data Factory, Azure Data Explorer, Azure HDInsight, ADLS), GCP |
| **Reporting and ETL Tools** | AWS GLUE, Tableau, Power BI. |
| **Databases** | Oracle, SQL Server, , MS Access, NoSQL Database (HBase, MongoDB, DynamoDB) |
| **Big Data Technologies** | Hadoop, HDFS, Hive, Oozie, Sqoop, Spark, Machine Learning, Pandas, NumPy, Seaborn, Impala, Zookeeper, Airflow, Informatica, Snowflake, Data Bricks, Kafka, Cloudera |
| **Data Analysis Libraries:** | Pandas, NumPy, SciPy, Scikit-learn, Matplotlib |
| **Containerization** | Docker, Kubernetes |
| **CI/CD Tools** | Jenkins, GitLab, Bamboo. |
| **Software Methodologies** | Agile, Scrum, Waterfall |
| **Development Tools** | Eclipse, PyCharm, IntelliJ, SSMS, Microsoft Office Suite (Word, Excel, PowerPoint) |
| **Version Control Tools** | GitHub and Azure DevOps |

**EDUCATION:**

Bachelor of Technology in Computer Science and Engineering from **JNTUH – 2014**

**PROFESSIONAL EXPERIENCE:**

**Amex, Phoenix, AZ Nov 2021 – Present**

**Senior Data Engineer**

* Worked with business/user groups for gathering the requirements and working on the creation and development of pipelines.
* To build the **batch data pipelines** we used **airflow** and to build the **streaming data pipelines** we have used the **Kafka**.
* Developed Spark Streaming script which consumes topics from distributed messaging source **Kafka** and periodically pushes a batch of data to spark for real-time processing
* Worked on creating **Azure Data Factory** and managing policies for Data Factory and Utilized Blob storage for storage and backup on Azure.
* Worked on developing the process and ingested the data in Azure cloud from web service and load it to **Azure SQL DB**.
* Worked with **Spark** applications in **Python** for developing the distributed environment to load high volume files using **PySpark** with different schema into PySpark Data frames and process them to reload into **Azure SQL DB** tables.
* Designed and developed the pipelines using **Databricks** and automated the pipelines for the **ETL** processes and further maintenance of the workloads in the process.
* Worked on creating ETL packages using **SSIS** to extract data from various data sources like Access database, Excel Spreadsheet, and flat files, and maintain the data using **SQL Server**.
* Worked with **ETL** operations in **Azure Databricks** by connecting to different relational databases using **Kafka** and used **Informatica** for creating, executing, and monitoring sessions and workflows.
* Worked on automating data ingestion into the Lakehouse and transformed the data, used **Apache Spark** for leveraging the data, and stored the data in **Delta Lake**.
* Ensured data quality and integrity of the data using **Azure SQL Database** and automated **ETL** deployment and operationalization.
* Used Data bricks, Scala, and Spark for creating the data workflows and capturing the data from Delta tables in **Delta Lakes**.
* Performed Streaming of pipelines using **Azure Event Hubs** and Stream Analytics to analyze the data from the data-driven workflows.
* Worked with **Delta Lakes** for consistent unification of Streaming, processed the data, and worked on ACID transactions using **Apache Spark**.
* Worked with **Azure Blob Storage** and developed the framework for the implementation of the huge volume of data and the system files.
* Responsible for identifying and implementing new SQL Server features designed to improve query performance; this included utilizing new T-SQL language enhancements, performing in-memory optimization, and performance tuning with Extended Events.
* Served as the resident SME on performance tuning stored procedures, functions, T-SQL scripts, indexes and SSIS packages.
* Implemented various tasks and transformations for data cleansing and performance tuning of packages by using SSIS.
* Deployed applications in clustered SOA 12c environment and performance tuning of SOA middleware.
* Implemented of distributed stream processing platform with low latency and seamless integration, with data and analytics services inside and outside **Azure** to build your complete big data pipeline.
* Worked with **PowerShell** scripting for maintaining and configuring the data. Automated and validated the data using **Apache Airflow**.
* Worked on optimization of **Hive** queries using best practices and right parameters and using **Hadoop, YARN**, **Python, and PySpark**.
* Used Sqoop to extract the data from **Teradata** into **HDFS** and export the patterns analyzed back to Teradata.
* Worked on **Kafka** to bring the data from data sources and keep it in **HDFS** systems for filtering.
* Used Accumulators and Broadcast variables to tune the Spark applications and to monitor the created analytics and jobs.
* Tracked **Hadoop** cluster job performance and capacity planning and tuning Hadoop performance for high availability and Hadoop cluster recovery.
* Worked on a **Data Migration** project to migrate data from different sources (**Teradata, Hadoop, DB2**) to **Google Cloud Platform (GCP)** using UDP framework and transforming the data using **Spark Scala scripts**.
* Worked on creating data ingestion processes to maintain Global Data Lake on the **GCP cloud and Big Query**.
* Worked with **Tableau** for generating reports and created Tableau dashboards, pie charts, and heat maps according to the business requirements.
* Worked with all phases of **Software Development Life Cycle** and used **Agile methodology** for development.

**Environment:** Python, SQL, Cassandra DB, Azure Data Lake Storage Gen 1, Azure Data Factory, Azure SQL DB, Spark, Databricks, SSIS, SQL Server, Kafka, Informatica, Apache Spark, Delta Lake, Azure Event Hubs, Stream Analytics, Azure Blob Storage, PowerShell, Apache Airflow, Hadoop, YARN, PySpark, Hive, Teradata, Sqoop, HDFS, Spark, Agile.

**Starbucks, Seattle, WA. Jan 2021 – Nov 2021**

**Senior Data Engineer**

**Responsibilities:**

* Worked in complete **Software Development Life Cycle (SDLC)** process by analyzing business requirements and understanding the functional workflow of information from source systems to destination systems.
* Utilizing analytical, statistical, and programming skills to collect, analyze and interpret large data sets to develop data-driven and technical solutions to difficult business problems using tools such as **SQL, and Python**.
* Worked on designing **AWS EC2** instance architecture to meet high availability application architecture and security parameters.
* Created **AWS S3 buckets** and managed policies for **S3 buckets** and **Utilized S3 buckets** and Glacier for storage and backup. Worked on **Hadoop cluster** and data querying tools to store and retrieve data from the stored databases.
* Worked on designing and developing the **SSIS Packages** to import and export data from **MS Excel, SQL Server**, and Flat files.
* Worked on Data Integration for extracting, transforming, and loading processes for the designed packages.
* Designed and deployed automated **ETL workflows** using **AWS lambda**, organized and cleansed the data in **S3 buckets** using **AWS Glue** and processed the data using **Amazon Redshift**.
* Worked within the **ETL architecture** enhancements to increase the performance using query optimizer.
* Implemented the data that is extracted using **Spark, Hive**, and large data sets using **HDFS**.
* Worked on Streaming data transfer, data from different data sources into **HDFS, No SQL databases**.
* Created **ETL Mapping** with Talend Integration Suite to pull data from Source, apply transformations, and load data into the target database.
* Performance tuning to optimize SQL queries using query analyzer.
* Worked on scripting with **Python** in **Spark** for transforming the data from various files like Text files, CSV and JSON.
* Loaded the data from different relational databases like **MySQL** and **Teradata** using **Sqoop jobs**.
* Worked on processing the data and testing using Spark SQL and on real-time processing by Spark Streaming and Kafka using Python.
* Scripted using **Python** and **PowerShell** for setting up baselines, branching, merging, and automation processes across the process using **GIT**.
* Worked with the implementation of the **ETL architecture** for enhancing the data and optimized workflows by building **DAG**s in **Apache Airflow** to schedule the **ETL** jobs and additional components in **Apache Airflow** like Pool, Executors, and multi-node functionality.
* Used various Transformations in **SSIS Dataflow**, Control Flow using for loop Containers and Fuzzy.
* Worked on creating **SSIS** packages for Data Conversion using data conversion transformation and producing the advanced extensible reports using **SQL Server Reporting Services**.
* Deployed application to **GCP** using Spinnaker (rpm based) launched multi-node **Kubernetes** cluster in **Google Kubernetes Engine (GKE)** and migrated the dockized application from **AWS to GCP**.

**Environment:** Python, SQL, AWS EC2, AWS S3 buckets, Hadoop, PySpark, AWS lambda, AWS Glue, Amazon Redshift, Spark Streaming,GCP, Apache Kafka, SSIS, Informatica, ETL, Hive, HDFS, NoSQL, Talend, MySQL, Teradata, Sqoop, PowerShell, GIT, Apache Airflow.

**Blue Cross Blue Shield, Louisville, KY Nov 2018 – Dec 2020**

**Data Engineer**

**Responsibilities:**

* Develop a data set process for data mining and data modeling and recommend ways to improve data quality, efficiency, and reliability.
* Extract Transform and Load data from Sources Systems to Azure Data Storage services using a combination of Azure Data Factory, T-SQL, **Spark** SQL, and U-SQL Azure Data Lake Analytics Data Ingestion to one or more Azure Services - (**Azure Data Lake, Azure Storage, Azure SQL, Azure DW**) and processing the data in In **Azure Databricks**.
* Created Pipelines in ADF using Linked Services/Datasets/Pipeline/ to Extract, Transform, and load data from different sources like Azure SQL, Blob storage, Azure SQL Data warehouse, write-back tool and backwards.
* Developed **Spark** applications using **PySpark** and **Spark**-SQL for data extraction, transformation, and aggregation from multiple file formats for analyzing & transforming the data to uncover insights into the customer usage patterns.
* Responsible for writing **Hive Queries** to analyze the data in Hive warehouse using **Hive Query Language (HQL).** Involved in developing **Hive DDLs** to create, drop and alter tables.
* Extracted the data and updated it into **HDFS** using **Sqoop Import** from various sources like Oracle, Teradata, SQL server, etc.
* Created **Hive staging tables** and **external tables** and joined the tables as required.
* Implemented **Dynamic Partitioning**, **Static Partitioning**, and **Bucketing**.
* Installed and configured Hadoop Map Reduce, Hive, HDFS, Pig, Sqoop, Flume, and Oozie on the Hadoop cluster.
* Worked on Microsoft Azure services like **HDInsight Clusters, BLOB, ADLS, Data Factory** and Logic Apps and did POC on Azure Data Bricks.
* Implemented Sqoop jobs for data ingestion from Oracle to Hive.
* Worked with various formats of files like delimited text files, clickstream log files, Apache log files, **Avro files, JSON files, and XML** Files. Mastered in using different columnar file formats like RC, ORC, and Parquet formats.
* Developed custom the Unix/BASH **SHELL** scripts for pre and post validations of the **master** and **slave** nodes, before and after the configuration of the name node and data nodes
* Developed job workflows in **Oozie** for automating the tasks of loading the data into HDFS.
* Implemented compact and efficient file storage of big data by using various file formats like **Avro, Parquet, and JSON** and using compression methods like **GZip, and Snappy** on top of the files.
* Exploring with **Spark,** improving performance and optimization of the existing algorithms in Hadoop using **Spark Context, Spark-SQL, Data Frame,** and **Pair RDDs**.
* Worked on **Spark** using **Python** as well as **Scala** and **Spark SQL** for faster testing and processing of data.
* Worked on various data modeling concepts like star schema, and **snowflake** schema in the project.
* Extensively used **Stash, Bit-Bucket, and GITHUB** for code control purposes.
* Migrated **Map reduce**s jobs to **Spark jobs** for achieving a better performance.

**Environment:** Hadoop 2.7, HDFS, Microsoft Azure services like HDInsight, BLOB, ADLS, Logic Apps, etc., Hive 2.2, Sqoop 1.4.6, snowflake, Apache Spark 2.3, Spark-SQL, ETL, Maven, Oozie, Java 8, Python3, Unix.

**Tesco, Bengaluru, India Nov 2015 – Sep 2018**

**Data Engineer**

**Responsibilities:**

* Prepared specification documents (**BRD / FRD**), based on business rules given by Business.
* Prepared Business Process and Data Process Models using **MS Visio**.
* Extensive experience in Stored Procedures, **Triggers, Functions, Cursors, Views**, Materialize Views, and Analytical Functions.
* Performed ETL processes to data ready for creating business analysis visuals which help the leadership team to make the right business decisions.
* Worked on visuals as Employee Info, project details, skills, leaves, calendar, time sheet
* POC: On premise to Cloud Azure data migration.
* Created and shared the dashboards within the organization for updating and editing the report as per the business requirement.
* Developed system standards, architecture, scenarios, detailed screen specifications and documented logical and physical data mode
* Implemented several **DAX** functions for various fact calculations for efficient data visualization in Power BI
* Utilized Power BI gateway to keep dashboards and reports up to-date with on premise data
* Created Power BI dashboard and generated reports using the SQL server tables as source. Lock objects on the reports.
* Performed Incremental load with several Dataflow tasks Control Flow Tasks using SSIS.
* Deployed, Tested and Scheduled SSIS packages.
* Prepared the Layouts by placing the fields to the appropriate place according to the requirement of the final Report.
* Performed data extraction, transformation, and loading (**ETL**) between systems using SQL tools such as **SSIS.**

**Environment**: Power BI, SQL Server 2012, AZURE, Google Big Query, Tableau, Excel, MS SQL Server 2014/2012, Server Integration Services (SSIS), SQL Server Reporting Services (SSRS), SQL Server Analysis Services (SSAS), SQL Server Management Studio (SSMS) and Oracle 11g.

**Amazon, India May 2014- Nov 2015**

**Data Engineer**

**Responsibilities:**

* Working with open-source Apache Distribution then Hadoop admins have to manually set up all the configurations- **Core-Site, HDFS-Site, YARN-Site, and Map Red-Site**. However, when working with popular Hadoop distributions like Hortonworks, Cloudera, or MapR the configuration files are set up on startup and the Hadoop admin need not configure them manually.
* Used Sqoop to import data from Relational Databases like MySQL, and Oracle.
* Involved in importing structured and unstructured data into HDFS.
* Responsible for fetching real-time data using Kafka and processing using Spark and Scala.
* Worked on Kafka to import real-time weblogs and ingested the data to Spark Streaming.
* Developed business logic using Kafka Direct Stream in Spark Streaming and implemented business transformations.
* Worked on Building and implementing a real-time streaming ETL pipeline using Kafka Streams API.
* Worked on Hive to implement Web Interfacing and stored the data in Hive tables.
* Migrated Map Reduce programs into Spark transformations using Spark and Scala.
* Experienced with Spark Context, Spark-SQL, and Spark YARN.
* Responsible for data extraction and data ingestion from different data sources into Hadoop Data Lake by creating ETL pipelines using Sqoop, Hive and Spark.
* Responsible for importing data to HDFS using Sqoop from different RDBMS servers and exporting data using Sqoop to the RDBMS servers after aggregations for other ETL operations.
* Experience in designing and developing applications in PySpark using python to compare the performance of Spark with Hive.
* Developed Spark modules, write complex queries, functions, views to generate reports for business users.
* Created automated python scripts to convert the data from different sources and to generate the ETL pipelines.
* Developed end to end data processing pipelines that begin with receiving data using distributed messaging systems Kafka through persistence of data into Hive and enrich the data using PySpark.
* Worked on Sequence files, RC files, Map side joins, bucketing, partitioning for Hive performance enhancement and storage improvement.
* Worked on connecting Cassandra database to the Amazon EMR File System for storing the database in S3.
* Implemented usage of Amazon EMR for processing Big Data across a Hadoop Cluster of virtual servers on **Amazon Elastic Compute Cloud (EC2)** and Amazon Simple Storage Service (S3).
* Deployed the project on **Amazon EMR with S3** connectivity for setting backup storage.

**Environment:** Hadoop, Map Reduce, Hive, Spark, Oracle, GitHub, Tableau, UNIX, Cloudera, Kafka, Sqoop, Scala, NIFI, HBase, Amazon EC2, S3.