**Alekhya V**

alekhya.v1108@gmail.com https://www.linkedin.com/in/alekhya-v-91b269233/ **+1 857-574-9745**

- **Around 9 years** of professional software development experience and expertise in **Cloud Engineering, Hadoop Ecosystem, Big Data, Data Warehousing.**
- Experience in huge scope application improvement utilizing Big Data biological system - **Hadoop** (HDFS, MapReduce, Yarn), **Spark**, **Hive**, **Impala**, **HBase**, **Airflow**, **AWS**, **Azure**.
- Experienced in working with databases **Oracle, MySQL,** and **SQL server.**
- Implemented **Spark Scripts** using **Scala, Spark SQL** to access hive tables into spark for faster processing of data
- Hands on experience with working on **Spark** using both **Scala** and **Python**
- Performed various actions and transformations on **spark RDD's** and **Data Frames**.
- Experience in handling large datasets using Partitions, **Spark** in memory capabilities, Broadcasts in **Spark** with Scala, Effective Joins, Transformations and other during ingestion process itself.
- Experience in building and architecting multiple Data pipelines, end to end **ETL and ELT process** for Data ingestion and transformation in Spark and coordinating tasks among the team.
- Hands on Experience in Writing **Python** Scripts for Data Extract and Data Transfer from various data sources.
- Experience in implementing Data warehouse solutions in **AWS Redshift** worked on various projects to migrate data from on database to **AWS Redshift, RDS and S3.**
- Practical Experience with **AWS** services like **Amazon EC2, S3, EMR, Amazon VPC, Amazon Elastic Load Balancing, IAM, Auto Scaling, CloudFront, CloudWatch, SNS,SQS,** and to stimulate resources
- Experience in loading data to **Azure Data Lake, Azure SQL Database**, **Azure SQL Data Warehouse to** control and grant database access and in building data pipelines using **Azure Data Factory**
- Adequate experience with **Azure** services like, **Stream Analytics**, **Active Directory**, **Blob Storage**, **Storage Explorer**.
- Compute engine, cloud load balancing, cloud storage, **cloud SQL**, stack driver monitoring, and cloud deployment manager.
- Strong Hadoop and platform support experience with all tools and services in major **Hadoop distributions** - **Cloudera, Amazon EMR, Azure, and Hortonworks**.
- Proficient in handling and ingesting terabytes of **Streaming data** (Spark streaming, Strom), **Batch Data**, **Automation** and **Scheduling** (Airflow).
- Profound knowledge in **MapReduce, Apache Crunch, Hive,** and **Splunk** for Hadoop jobs.
- Expertise in leveraging **Spark** Components such as **Spark SQL**, **Data Frames, Datasets, Spark-ML,** and **Spark Streaming** to create production-ready Spark applications.
- Strong working experience with **SQL** and **NoSQL** databases data modeling, tuning, disaster recovery, backup and creating data pipelines.
- Experienced in scripting with **Python** (**PySpark**), **Scala** and **Spark-SQL** for development, aggregation from various file formats such as **XML, JSON, CSV**, **Avro, Parquet, ORC**.
- Great experience in data analysis using **HiveQL,** tables, **Latin** queries, custom MapReduce programs and achieved improved performance.
- Experience in to develop search engines on unstructured data within **NoSQL** databases in **HDFS**.
- Extensive knowledge in all phases of **Data Acquisition**, **Data Warehousing** (gathering requirements, design, development, implementation, testing, and documentation), **Data Modeling**(analysis using Star Schema and **Snowflake** for FACT and Dimensions Tables), **Data Processing** and **Data Transformations** (Mapping, Cleansing, Monitoring, Debugging, Performance Tuning and Troubleshooting Hadoop clusters).

**Technical Skills:**

| | |
|---|---|
| **Big Data Ecosystem** | HDFS, Yarn, MapReduce, Spark, Hive, Airflow, StreamSets, HBase |
| **Hadoop Distributions** | Apache Hadoop 2.x/1.x, Cloudera CDP, Hortonworks HDP, **Amazon AWS** - EMR, EC2, EBS, S3, Athena, Glue, Elasticsearch, SQS, DynamoDB, Redshift, ECS, Kinesis, **Microsoft Azure** - Databricks, Data Lake, Blob Storage, Azure Data Factory, SQL Database, SQL Data Warehouse, Cosmos DB, Azure Active Directory) |
| **Scripting Languages** | Python, Scala, HiveQL. |
| **Cloud Environment** | Amazon Web Services (AWS), Microsoft Azure |
| **NoSQL Database** | HBase, DynamoDB |
| **Database** | MySQL, Oracle, Teradata, MS SQL SERVER, PostgreSQL, DB2 |
| **ETL/BI** | Snowflake, Redshift |
| **Version Control** | Git, Bitbucket |

**PROFESSIONAL EXPERIENCE:**

**Client: United Health Group (Optum), MN** **June 2022-Present**
**Role: Senior Data Engineer**

**Responsibilities:**

- Built a scalable technical architecture and data processing layer on Azure Cloud to solve business problems for a leading health insurance company.
- Performed ETL activities using Azure Data Factory and Databricks.
- Set up a development environment on IntelliJ to develop code in Scala.
- Structured code in different files as per organization standards and following best practices.
- Developed a validation framework in Spark Scala to ensure data quality as well as check for accuracy and completeness of data.
- Built test functions to locally test the developed functions.
- Performed integration tests to validate end to end ETL pipeline.
- Use Azure Data Factory to run Databricks jar and schedule ETL jobs.
- Use Azure DevOps Release pipelines to deploy ADF using arm templates from one environment to another.
- Utilized Spark Scala to distribute data processing on large streaming datasets to improve ingestion and processing speed of the data.
- Used stored procedure, lookup, execute pipeline, data flow, copy data, azure function features in ADF.
- Hands on experience implementing Spark jobs performance tuning.
- Performed monitoring and management of clusters by using Azure HDInsight
- Writing complex SQL queries using joins, group by, nested queries.
- Involved in solving complex problems, so be sure to highlight your problem-solving skills and any specific examples of how you have applied these skills in your work.
- Experience with designing and implementing data pipelines using Azure data services.
- Experience with data modeling and data management in Azure.
- Experience with working with large datasets and designing scalable solutions.
- Troubleshooting and debugging data-related issues.
- Developing and maintaining data integration solutions using Azure Data Factory. Implemented Kubernetes for container orchestration and deployment of applications.

***Environment:*** *Hadoop, MySQL, HBASE, Azure, HDFS, Apache Spark, Python, Scala, Oracle 11g, PL/SQL, UNIX, Tableau*

**Client: Bio-Rad- Hercules, CA** **June 2021-May 2022**
**Role: Senior Data Engineer**

**Responsibilities:**

- Created and enforced policies to achieve **HIPAA** compliance.
- Monitor System health and logs and respond accordingly to any warning or failure conditions.
- Involved in maintaining various Unix Shell scripts.
- Installed and configured EC2 instances on Amazon Web Services (**AWS**) for establishing clusters on cloud.
- Created **S3** buckets in the **AWS** environment to store files.
- Configured **S3** buckets with various life cycle policies to archive the infrequently accessed data to storage classes based on requirement.
- Created data pipeline for different events of ingestion, aggregation and load consumer response data in **AWS** S3 bucket into Hive external tables
- Migrated 160 tables from **Oracle to HDFS and HDFS to Redshift.**
- Involved in file movements between HDFS and **AWS S3**.
- Implemented **Python** code for retrieving the Social Media data.
- Developed a data pipeline using **Kinesis** to store data into HDFS.
- Involved in importing the real-time data to Hadoop using **Kinesis** and implemented job for daily imports.
- Automated all the jobs starting from pulling the Data from different Data Sources like MySQL to pushing the result set Data to Hadoop Distributed File System using **Sqoop.**
- Import the data from different sources like **HDFS/HBase into Spark RDD.**
- Involved in running Hadoop streaming jobs to process terabytes of text data.
- Worked on creating custom ETL scripts using **Python** for business related data.
- Converted all jobs to run in **EMR** by configuring the cluster according to the data size.
- Performed the ETL operations using Elastic Map Reduce and Redshift
- Used **Spark Streaming** to divide streaming data into batches as an input to Spark engine for batch processing
- Developed counters on **HBase** data to count total records on different tables.
- **ETL** Data Cleansing, Integration & Transformation using scripts for managing data from disparate sources.
- Used **HBase** to store majority of data which needs to be divided based on region.
- Developed Spark code using **Scala and Spark-SQL/Streaming** for faster testing and processing of data.
- Worked on Scala code base related to Apache Spark performing the Actions, Transformations on RDDs, DataFrames & Datasets using SparkSQL and Spark Streaming Contexts.
- created Spark jobs for processing data from **S3** data lake to Redshift
- Used Coalesce and repartition on data frames while optimizing the Spark jobs.
- Used **Spark API** over Cloudera Hadoop YARN to perform analytics on data in Hive.
- Performed performance tuning for Spark Steaming e.g. setting right Batch Interval time, correct level of Parallelism, selection of correct Serialization & memory tuning.
- Analyze business requirements and data sources from Excel, Oracle, and SQL Server for design, development, testing, and production rollover of reporting and analysis projects within **Tableau**.
- Implemented Fair schedulers on the Job tracker to share the resources of the Cluster for the MapReduce jobs given by the users.
- Created EBS volumes for storing application files for use with **EC2** instances whenever they are mounted to them.
- Worked on Creating **Kinesis** topics, partitions, writing custom partitioned classes.
- Worked on Creating **Kinesis Adaptors** for decoupling the application dependency.
- Exported the analyzed data to the relational databases using Sqoop for visualization and to generate reports.
- Worked in Agile development environment in sprint cycles of two weeks by dividing and organizing tasks.

***Environment**: Hadoop, CDH 5, MapReduce, Hive QL, MySQL, HBase, AWS, HDFS, HIVE, Apache Spark, Python, Scala, Cloudera, Hue Editor, Oracle 11g, PL/SQL, UNIX, Tableau.*

**Client: WFS Corp, Miami-Florida** **June 2020-May 2021**
**Role: Data Engineer**

**Responsibilities:**

- Participated in requirements sessions to gather requirements along with business analysts and product owners.
- Designing and Developing **Azure Data Factory (ADF)** extensively for ingesting data from different source systems like relational and non-relational to meet business functional requirements.
- Designed and Developed event driven architectures using **blob trigger**s and **Data Factory.**
- Creating pipelines, data flows and complex data transformations and manipulations using **ADF and PySpark with Databricks.**
- Automated jobs using different triggers like **Events, Schedules and Tumbling in ADF.**
- Created, provisioned different **Databricks clusters, notebooks, jobs and autoscaling.**
- Ingested huge volume and variety of data from disparate source systems into **Azure DataLake Gen2 using Azure Data Factory V2.**
- Created several **Databricks Spark jobs with Pyspark** to perform several tables to table operations.
- Performed **data flow** transformation using the data flow activity.
- Implemented **Azure,** self-hosted integration runtime in ADF.
- Implement **Tableau server** user access control for various Dashboard requirements
- Implement complex business rules in **python using Panda libraries, numpy, Scikit** learn
- Optimize **ELT** workloads against Hadoop file system implementing **HIVE SQL for transformation**
- Containerize data wrangling jobs in **Docker containers utilizing Git and Azure**
- Developed streaming pipelines using **Apache Spark** with Python.
- Created, provisioned multiple **Databricks clusters** needed for batch and continuous streaming data processing and installed the required libraries for the clusters.
- Improved performance by optimizing computing time to process the streaming data and saved cost to the company by **optimizing the cluster run time.**
- Perform ongoing monitoring, **automation, and refinement of data engineering solutions**.
- Designed and developed a new solution to process the **NRT data by using Azure stream analytics, Azure Event Hub and Service Bus Queue.**
- Created Linked service to land the data from **SFTP location to Azure Data Lake.**
- Extensively used **SQL Server Import and Export Data tool.**

***Environment:** Azure Data Factory (ADF), Data Factory, PySpark, Azure DataLake Gen2 using Azure Data Factory V2, PySpark, Azure, Tableau, Python, Hive SQL, Docker, Git, Apache Spark, NRT data, Azure Stream , Service Bus Queue, SFTP Azure Data Lake, SQL server Import.*

**Client: Mind Tree** **Oct 2017 - Dec 2019**
**Role: Big Data Engineer**

**Responsibilities:**

- Worked with **Hortonworks** distribution. Installed, configured, and maintained a **Hadoop** cluster based on the business and the team requirements.
- Experience with big data components like **HDFS**, **MapReduce**, **YARN**, **Hive**, **HBase, Sqoop and Ambari.**
- Involved in end to end implementation of **ETL** pipelines using **Python** and **SQL** for high volume analytics, also reviewed use cases before on boarding to **HDFS**. Capturing data and importing it to HDFS using and for semi-structured data and Sqoop for existing relational databases.
- Used **Cloudera Hue** and **Zeppelin notebooks** to interact with HDFS clusters. Used **Cloudera Manager, Search** and **Navigator** to configure and monitor resource utilization across the cluster.
- Enhanced scripts of existing modules written in **Python**. Migrated **ETL** jobs to scripts to apply transformations, joins, aggregations and to load data into HDFS.
- Developed an **ETL** pipeline to extract archived logs from disparate sources and stored in **AWS S3** for further processing using **PySpark**. Used **Cron** schedulers for weekly automation.
- Designing and developing jobs to get the files from transaction systems into data lake raw zone.
- Responsible for loading, managing and reviewing terabytes of log files using **Ambari** and **Hadoop streaming jobs**.
- Involved in writing rack topology scripts and **map** reduce programs to parse raw data.
- Used **Sqoop** to migrate data between traditional **RDBMS** and **HDFS**. Ingested data into HDFS from **Teradata**, **Oracle,** and **MySQL**. Identified required tables and views and exported them into **Hive**. Performed ad-hoc queries using **Hive joins, partitioning, bucketing** techniques for faster data access.
- Working with the new **Business Data Warehouse (BDW)** improved query/report performance, reduced the time needed to develop reports and established self-service reporting models in Cognos for business users.
- Used to automate the data flow between disparate systems. Designed dataflow models and complicated target tables to obtain relevant metrics from various sources.
- Developed **Bash** scripts to get log files from FTP server and executed **Hive** jobs to parse them.
- Implemented various **Hive** queries for analytics and called them from a client engine to run on different nodes. Worked on writing APIs to load the processed data to **HBase** tables.
- Created **External tables**, optimized **Hive** queries and improved the cluster performance by 30%.
- Performed data analysis using **HiveQL**, and custom **MapReduce** programs Python.
- Enhanced scripts of existing modules written in **Python**. Migrated **ETL** jobs to scripts to apply transformations, joins, aggregations and to load data into HDFS.
- Responsible for collecting, scrubbing, and extracting data from var generate reports, dashboards, and analytical solutions. Helped in debugging the **Tableau** dashboards.
- Troubleshooted defects by identifying root cause and fixing them during the QA phase. Used **SVN** for version control.

.

---

**EDUCATION:**

- Bachelor of Technology in Information Technology, GRIET.