**Akhila K**

*Cloud Data Engineer*

✉️**:** [kakhila809@gmail.com](mailto:kakhila809@gmail.com)

**☏ :** (972) 584-7655

**Career Summary**

* 9+ years of experience in IT industry specializing in Databricks lake house, migration of Hadoop and SQL databases to various Azure services, such as Azure Data Lake, Azure Data Lake Analytics, Azure SQL Database, Data Bricks, and Azure SQL Data Warehouse. I excel in efficiently overseeing data access management and facilitating the transfer of on-premises databases to Azure Data Lake store through the use of Azure Data Factory.
* Specialized in developing Spark applications using Spark-SQL within the Databricks platform. I have a track record of successfully extracting, transforming, and aggregating data from multiple file formats. This enables comprehensive analysis and transformation, ultimately leading to the valuable insights into customer usage patterns.
* Creating contemporary data warehousing solutions through the utilization of Databricks in conjunction with Aws or Azure technologies.
* Having strong grasp of Spark Architecture, which includes its essential components like Spark SQL, Data Frames, Spark Streaming, Driver Node, Worker Node, Stages, Executors, and Tasks.
* Having a deep understanding of Hadoop cluster architecture, which includes distributed file systems, parallel processing, scalability, fault tolerance, and high availability.
* In Databricks, acquired comprehensive expertise in the effective use of a diverse range of Hadoop ecosystem tools. These tools encompass HDFS, MapReduce, Yarn, Spark, Kafka, Hive, Impala, HBase, Sqoop, Pig, Oozie, Zookeeper.
* Engaging in technical conversations with client architects and collaborating with team members,
* Proficiency in employing Databricks DELTA lake for analytical data lake applications.
* Expertise in Databricks across several areas, including optimizing Hadoop cluster performance, debugging, monitoring, data transformations, mapping, cleaning, and troubleshooting.
* Expertise in Python programming for Databricks, making effective use of features like Lists, Dictionaries, Tuples, the Pandas framework, and the boto3 package to access and read files from Cloud.
* Hands-on expertise in crafting data pipelines for MLOPS (Machine Learning Operations).
* Beyond my technical skills, I possess exceptional communication abilities and a robust work ethic. I am proactive, a valuable team player, and consistently uphold a positive attitude.
* Skilled in Requirements Analysis, Agile Methodologies, Data Modeling, ETL, Data Integration, Data Analysis, Data Cleaning in web application developments and reporting.
* Having domain knowledge in Finance, Logistics, and Health insurance.
* Having proficiency in visualization tools like Power BI and Excel, encompassing the use of formulas, Pivot Tables, Charts, and DAX Commands.
* Having hands-on experience with Agile software development methodologies, including Scrum and Sprint, as well as traditional models like Waterfall and TDD for Databricks.
* Experience in SSIS, SAS, Tableau Prep or other ETL tools to automate data flows for reporting
* Experience in User Requirement Gathering and Technical Specification Documentation.
* Strong command of UNIX, Linux, and Windows operating systems within the Databricks environment.

**Technical skills:**

|  |  |
| --- | --- |
| **Cloud Technologies and Services** | Azure Data Factory, Microsoft Azure-Databricks, Logic Apps, Functional App, Data Lake, Blob Storage, Azure DevOps, Azure Database Migration Service, SQL Database, SQL Data Warehouse, Google Cloud Platform and AWS |
| **Big Data Ecosystems** | Databricks Lakehouse, Apache Spark, HDFS, YARN, Map-reduce, Sqoop, Hive, Oozie, Pig, Spark, Zookeeper, Cloudera Manager, Kafka, Spark Streaming |
| **Hadoop Distributions** | Apache Hadoop , Cloudera CDP, Hortonworks HDP |
| **Scripting language** | Python, PySpark, SparkSQL, SQL, Scala, R, shell scripting, HiveQL. |
| **NoSQL Database** | MongoDB, HBase, Cassandra |
| **Database** | MySQL, Oracle, Teradata, MSSQL SERVER, PostgreSQL, DB2 |
| **Version Control** | GIT, GitHub |
| **BI Tools** | Tableau, PowerBI |

**Experience:**

**Role: Data Engineer | June 2022 – Present**

**Client: Akamai Technologies, San Jose.**

**Responsibilities:**

* Implemented end-to-end data pipelines using Azure Data Factory to extract, transform, and load (ETL) data from diverse sources into Snowflake.
* Designed and implemented data processing workflows using Azure Databricks, leveraging Spark for large-scale data transformations.
* Built scalable and optimized Snowflake schemas, tables, and views to support complex analytics queries and reporting requirements.
* Developed data ingestion pipelines using Azure Event Hubs and Azure Functions to enable real-time data streaming into Snowflake.
* Strategically employed Azure Blob Storage for streamlined data file storage and retrieval, implementing compression and encryption methodologies to optimize storage costs and bolster data security.

• Seamlessly merged Azure Data Factory with Azure Logic Apps, orchestrating intricate data workflows and initiating targeted actions based on specific triggers.

* Executed seamless migration efforts from Microsoft SQL Server to Azure SQL Database, ensuring a seamless transition.

•. Created reusable SSIS Packages for extracting data from diverse formats, including Multi formatted Flat files, Excel, XML files, efficiently loading them into UL Database and DB2 Billing Systems.

* Engineered, deployed, and closely monitored SSIS Packages, employing SSIS Designer to craft packages for exporting heterogeneous data from sources like OLE DB (Oracle) and Excel to SQL Server.
* Implemented robust data governance and quality assurance protocols through Azure Data Factory and Snowflake, safeguarding data accuracy and uniformity.
* Pioneered data replication and synchronization strategies between Snowflake and other platforms via Azure Data Factory and Change Data Capture techniques.
* Developed and deployed Azure Functions to facilitate data preprocessing, enrichment, and validation tasks within data pipelines.
* Developed data processing workflows using Azure Data Factory to ingest, transform, and load data while integrating NumPy and pandas for data manipulation tasks.
* Utilized pandas to efficiently handle missing data, reducing data preparation time and enhancing the accuracy of analytical processes on Azure.
* Implemented data transformations with NumPy and pandas in Python scripts for Azure Functions, enabling event-driven data processing and automation.
* Employed NumPy for advanced mathematical computations and statistical analysis on Azure Databricks, enhancing insights generation and decision-making processes.
* Constructed advanced analytics and machine learning workflows using Azure Machine Learning and Snowflake, paving the way for predictive analytics and data-centric insights.
* Engineered and enacted data archiving and retention protocols by leveraging Azure Blob Storage and Snowflake's Time Travel functionality.
* Created tailored monitoring and alerting mechanisms using Azure Monitor and Snowflake Query Performance Monitoring (QPM), proactively identifying and resolving performance bottlenecks.
* Integrated Snowflake seamlessly with Power BI and Azure Analysis Services, fostering interactive dashboards and reports that empower self-service analytics for business users.
* Fine-tuned data pipelines and Spark jobs within Azure Databricks to enhance performance, encompassing Spark configuration optimization, strategic caching, and the application of data partitioning techniques.
* Devised comprehensive data cataloging and lineage solutions, harnessing tools such as Azure Purview and Apache Atlas to establish a comprehensive comprehension of data assets and their interrelations.
* Engaged in collaborative endeavors with diverse teams, including data scientists, data analysts, and business stakeholders, to grasp data requisites and deliver robust, scalable data solutions.

.

**Environment**: Azure Databricks, Data Factory, Logic Apps, Azure Database Migration Service, Snowflake, Functional App, MS SQL, Oracle SQL, Python, Scala, PySpark, Shell scripting, GIT, JIRA, Jenkins, Kafka, ADF Pipeline, Power BI.

**Role: Databricks Data Engineer | August 2020 – May 2022**

**Client: Thermo Fisher Scientific, Carlsbad, CA**

**Responsibilities:**

* Contributed to building our cloud infrastructure in AWS. Automated Cloud deployments using Chef, Python and AWS Cloud Formation templates.
* Optimizing of existing algorithms in Hadoop using Spark Context, Spark-SQL, Data Frames and Pair RDD's.
* Applied transformation Using Spark on the dataset.
* Created HBase tables to store data depending on column families.
* Extensively written Hive queries for data analysis to meet the business requirement.
* Involved in adding and decommissioning the data nodes.
* Responsible for analyzing using Spark SQL queries result with Hive queries.
* Implemented Spark using Scala and Spark SQL for faster testing and processing of data.
* Implemented Spark using Scala and utilizing Data frames and Spark SQL API for faster processing of data.
* Using the Maven for the deployments and processed structured, semi structured such as XML and unstructured data as well.
* Involved in file movements between HDFS and AWS S3 and extensively worked with S3 bucket in AWS.
* Created data partitions on large data sets in S3 and DDL on partitioned data.
* Implemented rapid-provisioning and life-cycle management for using Amazon EC2 and custom Bash scripts.
* Experience in developing/consuming Web Services (REST, SOAP, JSON) and APIs (Service-oriented architectures).
* Maintained the repository management tools like Artifactory and Nexus to store the WAR, JAR files which are deployed by using Chef in Jenkins tool.
* Built Chef based CI/CD solutions to improve developer productivity and rapid deployments.

**Environment:** PL/SQL, DOM, SOAP, JUnit, Hibernate 3.0, JDBC, MS SQL Server, Aws, Amazon S3, AWS Lambda, Amazon RDS, Amazon EC2.

**Role: Big Data Developer | July 2018 – July 2020**

**Client: Vanguard, Peoria, IL.**

**Responsibilities:**

* Executed meticulous data aggregation and analysis on expansive datasets, leveraging the prowess of Apache Spark, Scala, and Hive, thereby elevating the depth of insights provided to the business.
* Put into action extensive big data ecosystems such as Hadoop, Spark, and Cloudera to proficiently load and transform substantial volumes of structured, semi-structured, and unstructured data.
* Engineered a seamless integration between HBase and Hive within the Analytics Zone, masterfully crafting and optimizing HBase tables to expedite and enhance data querying efficiency.
* Harnessed the potency of Hive queries and Spark SQL to effectively dissect and process data, aligning with specific business requisites and emulating MapReduce functionalities.
* Championed the automation of deployments via YAML scripts, resulting in expedited and more efficient build and release procedures.
* Conducted a seamless migration of data from RDBMS (Oracle) to Hadoop using Sqoop, augmenting data management and processing capabilities for enhanced operational efficiency.
* Leveraged JIRA as an instrumental tool to streamline issue management and project workflow, effectively enhancing project organization and overall efficiency.
* Collaborated actively with fellow team members to diagnose and resolve JVM-related issues, leading to noticeable improvements in system performance and stability.
* Employed Git as a robust version control tool to expertly oversee the code repository, ensuring meticulous code management and precise tracking of alterations.

**Environment**: Sqoop, MYSQL, HDFS, Apache Spark, Scala, Hive Hadoop, Cloudera, HBASE, Kafka, MapReduce, Zookeeper, Oozie, Data Pipelines, RDBMS, AWS, EC2, Python, PySpark, shell script, Ambari, JIRA.

**Role: Big data Developer | May 2016 – June 2018**

**Client: Caterpillar, Peoria, IL.**

**Responsibilities:**

* Prepared an ETL framework using Sqoop, Pig, and Hive to bring in data from various sources and make it available for consumption.
* Processed HDFS data and created external tables using Hive, along with developing scripts for table ingestion and repair for reuse across the project.
* Developed ETL jobs using Spark and Scala to migrate data from Oracle to new MySQL tables.
* Utilized Spark (RDDs, Data Frames, Spark SQL) and Spark-Cassandra Connector APIs for various tasks, including data migration and business report generation.
* Developed a Spark Streaming application for real-time sales analytics.
* Analyzed source data, efficiently handled data type modifications, and used Excel sheets, flat files, and CSV files to generate Power BI ad-hoc reports.
* Analyzed SQL scripts and designed solutions using PySpark.
* Extracted data from various data sources into HDFS using Sqoop.
* Handled data import from various sources, performed transformations using Hive and MapReduce, and loaded data into HDFS.
* Extracted data from MySQL into HDFS using Sqoop.
* Implemented automation for deployments using YAML scripts for streamlined builds and releases.
* Worked with Apache Hive, Apache Pig, HBase, Apache Spark, Zookeeper, Flume, Kafka, and Sqoop.
* Implemented data classification algorithms using MapReduce design patterns.
* Extensively worked on creating combiners, partitioning, and distributed cache to enhance the performance of MapReduce jobs.
* Utilized Git and GitHub repositories to maintain the source code and enable version control.

**Environment**: Hadoop, Hive, spark, PySpark, Sqoop, Spark SQL, Shell Script, Cassandra, YAML, ETL, Talend.

**Role: Data Analyst | February 2014 to November 2016**

**Client: Yana Software Private Limited Hyderabad, India**

**Responsibilities:**

* Worked with Business Analyst and helped represent the business domain details and prepared low-level analysis documentation.
* Created Hive tables and created Sqoop jobs to import the data from Oracle/SQL Server to HDFS
* Developed Oozie workflows and scheduled them in Control-M as daily jobs to load incremental updates from the RDBMS source systems.
* Wrote different pig scripts to clean up the ingested data and created partitions for the daily data.
* Prepared pig scripts and Spark SQL to handle all the transformations specified in the S2TM's and to support SCD2 and SCD1 scenarios.
* Wrote different UDF's to convert the date format and to create hash value using MD5 Algorithm in Java.
* Implemented Partitioning and bucketing in Hive based on the requirement.
* Involved in converting Hive SQL queries into Spark transformations using Spark SQL and Scala.
* Experienced in implementing Spark RDD transformations, actions to implement business analysis and worked
* Create Sqoop import jobs to import source tables from Microsoft SQL Server.
* Create Sqoop export jobs to export target tables to Teradata and to make the target tables available to the reporting layer.
* Worked with BI and QA team to test the application and fixed the defects immediately.
* Leveraged open-source monitoring toolkit Prometheus to capture pod metrics and built sample dashboards in Splunk.
* Involved in Unit and Integration level testing and prepared supporting documents for deployment.

**Environment:** HDP 2.4, Hadoop 2.6, Hive 0.14, Pig 0.14, HBase, Spark1.6, Scala, Kafka, Oozie, SQL Server, Jenkins, Nexus, Shell, Java, CSS, HTML,Eclipse.

**Education:**

* **Bachelor’s Degree (Vidya Jyothi Institute of Technology (2010 – 2014))**

**Major: Computer Science**