Akhila Sai

akhilasai0829@gmail.com | 216-245-7422

Summary

Over 9 years of extensive professional experience in information technology, I possess a broad skill set spanning Big Data, Hadoop, Spark, Hive, Impala, Sqoop, Flume, Kafka, SQL tuning, ETL development, report development, SAS, database development, and data modeling, with a strong foundation in Oracle database architecture. My expertise extends to Big Data analytics, utilizing Hadoop ecosystem tools like MapReduce, HDFS, Yarn/MRv2, Pig, Hive, HBase, Spark, Kafka, Flume, Sqoop, Oozie, Avro, Solr, Zookeeper, GCP, and Spring Boot. I have hands-on experience with test-driven development (TDD), behavior-driven development (BDD), and acceptance test-driven development (ATDD). Additionally, I excel in migrating databases to Snowflake and leveraging Google Cloud Platform (GCP) services, including BigQuery, Dataflow, Pub/Sub, and Cloud Storage, for designing and implementing scalable data processing solutions. My proficiency extends to managing databases, Azure Data Platform services, and building multiple Data Lakes. I have substantial expertise in Text Analytics, data visualizations using R and Python, and dashboards using tools like Tableau and Power BI. I specialize in building and optimizing data pipelines on GCP with Apache Beam, Cloud Composer, and Cloud Functions. I am skilled in multiple databases like MongoDB, Cassandra, MySQL, Oracle, and MS SQL Server, working with various file formats and container orchestration using ECS, ALB, and Lambda. I have created Snowflake Schemas, developed Automation Regressing Scripts, and utilized Kubernetes and Docker for CI/CD processes. My proficiency extends to GCP's data engineering stack, data architecture design on GCP with services like Cloud Dataproc, Cloud Composer, and Data Catalog, and using analytical applications like R, SPSS, Rattle, and Python for trend analysis and relationship identification. I am well-versed in Hadoop ecosystem components, Jenkins, Docker, Kubernetes, and have experience in enterprise application development and deployment.

Technical Skills:

Big Data/Hadoop Technologies:	Map Reduce, Spark, Spark SQL, Azure, Spark Streaming, Kafka, PySpark,Pig, Hive, HBase, Flume, Yarn, Oozie, Zookeeper, Hue, Ambari Server
Languages:	Java, Scala, Python (NumPy, SciPy, Pandas, Genism, Keras), Shell Scripting
NO SQL Databases:	Cassandra, HBase, MongoDB, Maria DB
Web Design Tools:	HTML, CSS, JavaScript, JSP, jQuery, XML
Development Tools:	Microsoft SQL Studio, IntelliJ, Azure Data bricks, Eclipse, NetBeans.
Development Methodologies:	Agile/Scrum, UML, Design Patterns, Waterfall
Build Tools:	Jenkins, Toad, SQL Loader, PostgreSQL, Talend, Maven, ANT, RTC, RSA, Control-M, Oozie, Hue, SOAP UI
Reporting Tools:	MS Office (Word/Excel/Power Point/ Visio/Outlook), Crystal Reports XI, SSRS, Cognos.
Databases:	Microsoft SQL Server, MySQL, Oracle 11g, 12c, DB2, Teradata, Netezza
Operating Systems:	All versions of Windows, UNIX, LINUX
Cloud:	MS Azure, GCP

Professional Experience

Client: AgFirstColumbia, SC **April 2022 to Present**
Role: Sr. Data Engineer

Responsibilities:

- Imported data from diverse sources like HDFS/HBase into Spark RDD and established a data pipeline using Kafka and Storm to store data in HDFS.
- Proficient in developing a log producer in Scala that actively monitors application logs, performs incremental log transformation, and dispatches them to a Kafka and Zookeeper-based log collection platform.
- Expertise in configuring and administrating the Hadoop Cluster, working with major Hadoop Distributions such as Apache Hadoop and Cloudera.

- Designed and implemented scalable data processing pipelines on GCP, employing tools like Cloud Dataflow and Apache Beam to efficiently transform and analyze vast volumes of data.
- Converted date-related data into formats compatible with applications by creating Apache Pig UDFs.
- Effectively configured Zookeeper to coordinate and support Kafka, Spark, Spark Streaming, HBase, and HDFS.
- Professionally handled Tableau Servers for ETL, Teradata, and other EDW data integrations and developments.
- Successfully integrated Kafka with Spark Streaming for real-time data processing.
- Architected and optimized data storage solutions on GCP, utilizing services like BigQuery and Cloud Storage to ensure data reliability, availability, and performance for downstream analytics and reporting.
- Extensive experience utilizing HDFS, Map Reduce, Hive, Spark, Sqoop, Oozie, and HBase.
- Implemented real-time data streaming solutions using GCP's Pub/Sub and Dataflow, enabling the processing and analysis of streaming data for immediate insights and decision-making.
- Devised an automated system using Shell scripts for sqooping the job.
- Pioneered the implementation of a log producer in Scala that actively monitors application logs, performs incremental log transformation, and dispatches them to a Kafka and Zookeeper-based log collection platform.
- Implemented data governance and security measures on GCP to ensure compliance with data protection regulations and industry best practices.
- Provided consulting on Snowflake Data Platform Solution Architecture, Design, Development, and deployment, with a focus on promoting a data-driven culture across enterprises.
- Worked proficiently with Oracle Databases, Redshift, and Snowflakes.
- Created Matillion jobs for Redshift to process data from multiple source systems into S3, processing all S3 CSV files into Redshift and unloading the resulting data set into S3 for further consumption in SageMaker/ML predictive modeling.
- Diligently documented requirements, including available code, to be implemented using Spark, Hive, HDFS, HBase, and Elastic Search.
- Transformed and aggregated data for analysis by implementing workflow management of Sqoop, Hive, and Pig scripts.
- Conducted performance tuning and optimization of GCP data solutions to enhance data processing speed, reduce costs, and improve overall system performance.
- Involved in the design and deployment of Hadoop clusters and various Big Data analytic tools, including Pig, Hive, HBase, Oozie, Zookeeper, Sqoop, Flume, Spark, Impala, and Cassandra with Hortonworks Distribution.
- Developed Kafka consumer API in Scala for data consumption from Kafka topics.
- Created Map Reduce programs to parse raw data, populate staging tables, and store refined data in partitioned tables in the EDW.
- Developed a Python script to interface with REST APIs and extract data to GCP.
- Implemented monitoring and alerting solutions on GCP to proactively detect and address data pipeline issues, ensuring high availability and reliability of data processing workflows.
- Developed Scala scripts using both DataFrames/SQL and RDD/MapReduce in Spark for data aggregation, queries, and writing data back into OLTP systems through Sqoop.
- Utilized Spark streaming to receive real-time data from Kafka and store the streaming data to HDFS using Scala, as well as NoSQL databases such as HBase and Cassandra.
- Integrated Oozie with Hue and scheduled workflows for multiple Hive, Pig, and Spark Jobs.
- Created and managed Splunk DB connect identities, connections, inputs, outputs, lookups, and access controls.
- Possessed experience with Splunk Enterprise Security app (ES), conducting data investigations and analysis to address security vulnerabilities, incidents, and penetration techniques.
- Generated dashboards, reports, and alerts for real-time monitoring in Splunk, Tableau, and JasperSoft.

Environment: Hadoop, Oracle, Scala, Spark-Sql, PySpark, Python, Kafka, Sas, Sql, Mdm, Oozie, Ssis, T-Sql, Etl, Hdfs, Cosmos, Pig, Sqoop, Ms Access, Splunk, GCP, Splunk, HBase, Oozie, Kafka, Snowflake, Airflow, GCP

Client: ascena retail group Patskala, Ohio **January 2020 to March 2022**
Role: Data Engineer

Responsibilities:
- Performed data ingestion from Sqoop and Flume, extracting data from Oracle databases.
- Worked on the implementation of various stages of data flow within the Hadoop ecosystem, including ingestion, processing, and consumption.

- Developed PIG UDFs to convert date and timestamp formats from unstructured files into the required date formats and processed them accordingly.
- Imported and exported data into HDFS and Hive using Sqoop and Kafka.
- Demonstrated proficiency in developing MapReduce programs using Apache Hadoop for processing Big Data.
- Validated Sqoop jobs and Shell scripts, performing data validation to ensure correct data loading without discrepancies. Conducted migration and testing of both static and transactional data from one core system to another.
- Scripted the creation, truncation, dropping, and alteration of HBase tables to store data after the execution of MapReduce jobs for subsequent analytics.
- Established self-service reporting in Azure Data Lake Store Gen2 using an ELT approach.
- Implemented monitoring and Azure log analytics to alert the support team regarding the usage and statistics of daily runs.
- Took responsibility for extensive data ingestion using Sqoop and HDFS commands, accumulating 'partitioned' data in various storage formats such as text, JSON, Parquet, and more. Also involved in loading data from the LINUX file system to HDFS.
- Developed a data warehouse model in Snowflake for over 100 datasets using WhereScape.
- Possessed experience with Agile Methodologies, including Scrum stories and sprints, in a Python-based environment, along with expertise in data analytics and data wrangling.
- Conducted performance tuning for Phoenix/HBase, Hive queries, and Spark.
- Installed Kafka to gather data from various sources and store it for consumption.
- Utilized the custom File System plugin, enabling Hadoop MapReduce programs, HBase, Pig, and Hive to work unmodified and access files directly.
- Wrote PySpark and Spark SQL transformations in Azure Databricks to perform complex transformations for implementing business rules.
- Designed and developed Hive, HBase data structures, and Oozie workflows.
- Enhanced Hive and Pig core functionality by developing custom UDFs, UDTFs, and UDAFs.
- Built and maintained the environment on Azure IAAS and PAAS.
- Implemented continuous integration/continuous development best practices using Azure DevOps, ensuring code versioning.
- Scheduled various Snowflake jobs using NiFi.
- Handled data import from various sources, performed transformations using Hive and MapReduce, and loaded data into HDFS.
- Architected and implemented medium to large-scale BI solutions on Azure using Azure Data Platform services, including Azure Data Lake, Data Factory, Data Lake Analytics, Stream Analytics, Azure SQL DW, HDInsight/Databricks, and NoSQL DB.
- Made extensive use of Azure Portal, Azure PowerShell, Storage Accounts, Certificates, and Azure Data Management.
- Extracted, transformed, and loaded data from source systems to Azure Data Storage services using a combination of Azure Data Factory, T-SQL, Spark SQL, and U-SQL Azure Data Lake Analytics. Ingested data into one or more Azure Services such as Azure Data Lake, Azure Storage, Azure SQL, and Azure DW, and processed the data in Azure Databricks.
- Developed ETL jobs using Spark-Scala to migrate data from Oracle to new Hive tables.
- Experienced in working with various scripting technologies, including Python and Unix shell scripts.
- Developed both simple and complex MapReduce jobs using Hive and Pig.
- Created Spark code using Scala and Spark SQL/Streaming for faster testing and data processing.
- Developed Java-Spring-based middleware components to fetch data from HBase using the Phoenix SQL layer for various web UI use cases.

Environment: Sqoop, Hive, Azure, Json, XML, Kafka, Snowflake, Python, Map Reduce, Oracle, Agile Scrum, Map Reduce, Pig, Spark, Scala, Hive, Azure, Azure Data Bricks, DAX, Azure Data Lake, Kafka, Python

Client: CITRIX Fort lauderdale, FL **September 2017 to December 2019**
Role: Data Engineer

Responsibilities:
- Contributed to an Agile environment and utilized Rally tool to manage user stories and tasks.

- Implemented Apache Sentry to enforce access restrictions on Hive tables at a group level.
- Designed and configured Kafka topics within a new Kafka cluster across all environments.
- Developed multiple Tableau dashboards to address various business requirements.
- Implemented partitioning, dynamic partitions, and buckets in Hive to optimize data access.
- Established a composite server for data virtualization needs, creating multiple views with restricted data access using a REST API.
- Devised and executed data architecture on GCP, encompassing data modeling, schema design, and data integration strategies to facilitate efficient data processing and analysis.
- Exported analyzed data to relational databases using Sqoop for visualization and report generation for the BI team with Tableau.
- Installed a Kerberos-secured Kafka cluster with no encryption in both Dev and Prod environments, also configuring Kafka ACLs.
- Developed machine learning models to showcase Big Data capabilities using PySpark and MLlib.
- Migrated MapReduce jobs to Spark for improved performance.
- Leveraged GCP services such as Dataflow, Dataproc, and BigQuery to construct data pipelines for batch and real-time data processing, ensuring timely and accurate data availability for business insights.
- Transformed Hive/SQL queries into Spark transformations using Spark RDDs, Python, and Scala.
- Built Apache Spark applications for data processing from various streaming sources.
- Demonstrated a strong understanding of Tealeaf architecture and components, proficiently working with Spark Core, SparkSQL, Scala, and Cascading. Streamlined data streaming to Spark using Kafka.
- Handled Spark, Spark Streaming, Spark MLlib, Snowflake, Scala, and managed Data Frames in Spark with Scala.
- Constructed data pipelines using Spark, Hive, Pig, Python, Impala, and HBase for ingesting customer data.
- Converted Hive/SQL queries into Spark transformations using Spark RDDs, Python, and Scala.
- Implemented data transformation and cleansing processes using GCP tools like Dataflow or Cloud Dataprep to ensure data quality and consistency throughout the data pipeline.
- Queried and analyzed data from Cassandra using CQL for quick searching, sorting, and grouping.
- Performed joins on various tables in Cassandra using Spark and Scala and conducted analytics on top of them.
- Applied advanced Spark procedures like text analytics and processing with in-memory processing.
- Implemented Apache Drill on Hadoop to join data from SQL and NoSQL databases and store it in Hadoop.
- Ingested data from various sources into Hadoop and Cassandra using Kafka.
- Conducted performance monitoring and optimization of GCP data solutions, identifying bottlenecks and implementing optimizations to enhance data processing speed and resource utilization.
- Collaborated with data scientists and analysts to design and deploy machine learning pipelines on GCP using services such as AI Platform and AutoML, enabling the deployment and scaling of machine learning models.
- Utilized SQL Server Reporting Services (SSRS) to create and format various report types, including Cross-Tab, Conditional, Drill-down, Top N, Summary, Form, OLAP, Sub reports, ad-hoc reports, parameterized reports, interactive reports, and custom reports.
- Designed and developed Oracle PL/SQL scripts, Shell Scripts, data import/export processes, data conversions, and data cleansing.
- Designed and implemented data architecture on GCP, covering data modeling, schema design, and data integration strategies to facilitate efficient data processing and analysis.
- Developed Spark applications using Spark SQL in Databricks for data extraction, transformation, and aggregation from multiple file formats to analyze and transform data, uncovering insights into customer usage patterns.

Environment: Map Reduce, HDFS, Hive, Pig, Impala, Kafka, Cassandra, Spark, Scala, Solr, Java, SQL, Tableau, PIG, Zookeeper, Sqoop, Kafka, Teradata, GCP

Client: Limerock, India **April 2015 to June 2017**
Role: Data Engineer

Responsibilities:
- Created a mapping document to establish column mappings between source and target systems.
- Developed a Python utility to validate HDFS tables against source tables.

- Architected and implemented data solutions on Azure, harnessing services such as Azure Data Factory, Azure Databricks, and Azure SQL Database to construct scalable and efficient data pipelines for both batch and real-time data processing.
- Wrote Python code to fetch and manipulate data.
- Devised an ETL process using Informatica to load data from flat files and Excel files into the target Oracle Data Warehouse database.
- Automated all data extraction jobs from the FTP server to populate Hive tables using Oozie workflows.
- Implemented data transformation and cleansing processes using Azure Data Factory, Azure Databricks, or Azure Functions to maintain data quality and consistency across the data pipeline.
- Developed Python wrapper scripts to extract specific date ranges from Sqoop using custom properties required for the workflow.
- Engaged in filtering data stored in S3 buckets using Elasticsearch and subsequently loaded the data into Hive external tables.
- Crafted and implemented user-defined functions (UDFs) to enhance functionality in both PIG and HIVE.
- Performed regular data imports and exports between MySQL and HDFS using Sqoop.
- Implemented data storage and management solutions on Azure, making use of services such as Azure Blob Storage, Azure Data Lake Storage, and Azure SQL Database to efficiently store and organize both structured and unstructured data.
- Developed a shell script to generate staging and landing tables with identical schemas as the source and generate properties utilized by Oozie Jobs.
- Worked extensively with NoSQL databases like HBase to create HBase tables for ingesting large sets of semi-structured data from various sources.
- Designed and optimized data models and schemas within Azure SQL Database or Azure Synapse Analytics (formerly SQL Data Warehouse) to support efficient data querying and analysis.
- Created Oozie workflows to execute Sqoop and Hive actions.
- Implemented data transformation and cleansing processes using Azure Data Factory, Azure Databricks, or Azure Functions to maintain data quality and consistency across the data pipeline.
- Developed and maintained data ingestion processes, encompassing data extraction, transformation, and loading (ETL), leveraging Azure services such as Azure Data Factory and Azure Logic Apps to ensure reliable and timely data integration from diverse sources.
- Generated various graphs for business decision-making using Python's matplotlib library.

Environment: Python, HDFS, Spark, Hive, Sqoop, Oozie, ETL, Pig, Oracle 10g, My SQL, No SQL, Hbase, Windows, MS Azure

Client: Edvensoft Solutions India Pvt. Ltd, India **July 2013 to March 2015**
Role: Hadoop Developer

Responsibilities:
- Installed the Oozie workflow engine to execute multiple Hive and Pig Jobs.
- Developed Map/Reduce Jobs ranging from simple to complex using Hive and Pig.
- Collaborated closely with stakeholders to identify and comprehend data requirements, translating them into technical solutions on GCP to facilitate effective data-driven decision-making.
- Implemented Avro and Parquet data formats for Apache Hive computations to address custom business requirements.
- Conceived, implemented, and deployed a series of custom parallel algorithms for various customer-defined metrics and unsupervised learning models within the customer's existing Hadoop/Cassandra cluster.
- Set up and configured Hive, Pig, Sqoop, Flume, and Oozie on the Hadoop cluster.
- Cooperated with DevOps teams to deploy and manage data engineering solutions on GCP, utilizing infrastructure-as-code techniques and making use of services like Cloud Deployment Manager or Terraform.
- Made extensive use of SSIS transformations, including Lookup, Derived Column, Data Conversion, Aggregate, Conditional Split, SQL Task, Script Task, and Send Mail Task, among others.
- Established data security measures and access controls on GCP, including encryption, key management, and IAM policies, ensuring compliance with data privacy and security standards.
- Conducted data cleansing, enrichment, mapping, and automated data validation processes to ensure efficient reporting of meaningful and accurate data.

- Developed and maintained documentation and best practices for data engineering processes on GCP, fostering knowledge sharing and ensuring consistency across the team.
- Implemented Apache Pig scripts for loading data from and storing data into Hive.

Environment: Hive, Hadoop, Cassandra, Pig, Sqoop, Ooze, Hive, and Scala. Python, MS Office, GCP