**Akhil S**

**Email:** sakhil9099[@gmail.com](mailto:meghanayt07@gmail.com)  **PH: 470-296-4246**

**Linkedin - https://www.linkedin.com/in/akhil-s-231a94270/**

**Sr. Data Engineer**

**PROFESSIONAL SUMMARY**

* Around 10+ years of professional experience in IT industry including designing, developing, Analysis of big data applications in SPARK, Hadoop, Pig and HDFS environment and experience in Python.
* Highly experienced in importing and exporting data between HDFS and Relational Systems like MySQL and Teradata using Sqoop.
* GCP Certified Data Architect
* Knowledge on big-data database HBase and NoSQL databases MongoDB and Cassandra.
* Hands-on experience in scripting skills in Python, Linux, and UNIX Shell.
* Evaluated and recommended data engineering solutions, frameworks, and platforms to ensure that the COE leverages the most suitable technologies for the organization's data needs.
* Thorough understanding of various bigdata components such as HDFS, Job Tracker, Task Tracker, Name Node, Data Node YARN and MapReduce programming paradigm.
* Working with relative ease with different working strategies like Agile, Waterfall and Scrum methodologies.
* Expertise with the tools in Hadoop Ecosystem including Pig, Hive, HDFS, Map Reduce, Sqoop, Spark, Kafka, Yarn, Oozie, and Zookeeper, Hadoop architecture and its components.
* Experience with AWS services like S3, Athena, Redshift Spectrum, Redshift, EMR, Glue, Data pipeline, step functions, cloud watch, SNS and Cloud formation.
* Experienced with cloud: Hadoop-on-Azure, AWS/EMR, Cloudera Manager (also direct-Hadoop-EC2 (non EMR)
* Developed and maintained complex Oracle PL/SQL scripts, stored procedures, functions, and triggers to manage and manipulate data efficiently within the Oracle database.
* Experience in Agile Methodologies and extensively used Jira for Sprints and issue tracking.
* Experience in working with Map Reduce Programs, Pig Scripts and Hive commands to deliver the best results.
* Expertise in writing Apache Spark Streaming API on Big Data distribution in the active cluster environment.
* Developed Spark RDD and Spark Data Frame API for Distributed Data Processing.
* Good understanding and exposure to Python programming.
* Extensive knowledge on Amazon Web Services (AWS) EC2, S3, Elastic Map Reduce (EMR) and on Snowflake, Redshift, Identity and Access Management (IAM).
* Hands on experience on AWS cloud services (VPC, EC2, S3, RDS, Redshirt, Data Pipeline, EMR, Dynamo, Workspaces, RDS, SNS, SQS)
* Good experience of AWS Elastic Block Storage (EBS), different volume types and use of various types of EBS volumes based on requirement.
* Experience in building highly reliable, scalable Big data solutions on Hadoop distributions Cloudera, Horton works, AWS.
* Utilized the Power Platform, including Power BI, Power Apps, and Power Automate, to develop and maintain data solutions that enabled stakeholders to visualize, analyze, and interact with data for informed decision-making.
* Utilized PL/SQL for data extraction, transformation, and loading (ETL) processes, ensuring data integrity and consistency for analytics and reporting purposes.
* Worked and learned a great deal from Amazon Web Services (AWS) Cloud services like EC2, S3, EBS, RDS and VPC.
* Proficient in managing entire data science project life cycle and actively involved in all the phases of project life cycle including data acquisition, data cleaning, data engineering, features scaling, features engineering, statistical modeling, dimensionality reduction using Principal Component Analysis, testing and validation using ROC plot, K- fold cross validation and data visualization.
* Adept and deep understanding of Statistical modeling, Multivariate Analysis, model testing, problem analysis, model comparison and validation.
* Experience in using various packages in R and python like scikit-learn ggplot2, caret, dplyr, plyr, pandas, numpy, seaborn, SciPy, MatPlotLib, Beautiful Soup, Rpy2.

**TECHNICAL SKILLS:**

BigData Technologies: AWS EMR, S3, EC2-Fleet, Spark-2.2, 2.0 and 1.6, Hortonworks HDP, Hadoop, MapReduce, Pig, Hive, Apache Spark, SparkSQL, Informatica Power Center 9.6.1/8.x, Kafka, NoSQL, Elastic MapReduce(EMR), Hue, YARN, Nifi, Impala, Sqoop, Solr, Oozie

Databases: Cloudera Hadoop CDH 15.x, Hortonworks HDP, Oracle 10g/11g, Teradata, DB2, Microsoft SQL Server, MySQL, NoSQL, SQL databases.

Platforms (O/S): Red-Hat LINUX, Ubuntu, Windows NT/2000/XP.

Programming languages: Java, Scala, SQL, UNIX shell script, JDBC, Python, Perl.

Security Management: Hortonworks Ambari, Cloudera Manager, Apache Knox, XA Secure, Kerberos.

Web-technologies: DHTML, HTML, XHTML, XML, XSL (XSLT, XPATH), XSD, CSS, JavaScript, SOAP, RESTful, Agile, Design Patterns

Data warehousing: Informatica PowerCenter/Powermart/Dataquality/Bigdata, Pentaho, ETL Development, Amazon Redshift, IDQ, BigQuery Enterprise Warehouse, ALM Accelerator.

Database Tools: JDBC, HADOOP, Hive, No-SQL, SQL Navigator, SQL Developer, TOAD, SQL Plus, SAP Business Objects, SSIS, SSRS, Flask, Django, Fast API.

Data Modeling: Rational Rose, Erwin 7.3/7.1/4.1/4.0

Machine Learning Models - XGBoost, KNN, Linear Regression, Random Forest, Feedforward Nerual Network

Code Editors: Eclipse, Intellij, PyCharm

**PROFESSIONAL EXPERIENCE:**

**T-Mobile, Dallas, TX January 2021 to Present**

**Sr. Data Engineer**

**Responsibilities:**

* Design Architecture of data pipeline/ingestion as well as optimization of ETL workflows and developed syllabus/Curriculum data pipelines from Syllabus/Curriculum Web Services to HBASE and Hive tables.
* Provided training and support to team members on ALM Accelerator best practices, empowering them to effectively manage the end-to-end data engineering lifecycle.
* Performed data analysis, feature selection, feature extraction using Apache Spark.
* Leveraged ALM Accelerator to automate version control, code integration, and continuous integration/continuous deployment (CI/CD) for data engineering projects, reducing development cycle times and ensuring code consistency.
* Implemented performance tuning strategies within PL/SQL code, including query optimization, indexing, and efficient database schema design, enhancing overall database performance.
* Developed and enforced data engineering standards and guidelines within the COE, promoting consistency and quality in data pipelines, ETL processes, and data integration solutions.
* Designed and developed data models using DAX to support business intelligence and data visualization projects in Power BI, enabling efficient data analysis and reporting for stakeholders.
* Optimized existing DAX expressions and data models for performance improvements, resulting in faster report generation and improved user experience.
* Conducted data cleansing and transformation processes in conjunction with DAX measures to ensure data accuracy and consistency within Power BI datasets.
* Collaborated with stakeholders to define data governance, security, and compliance policies, and ensured their consistent application throughout data engineering processes within the COE.
* Machine Learning streaming libraries in Python and develop Python and bash scripts for automation and implemented Map Reduce jobs using Java API and Python using Spark
* Used Apache Kafka to aggregate web log data from multiple servers and make them available in Downstream systems for analysis and used Kafka Streams to Configure Spark streaming to get information and then store it in HDFS.
* Provided training and support to non-technical users in the organization to enable them to utilize the Power Platform tools effectively for data-related tasks.
* Involved in Migrating Objects from On prim to Snowflake and building up snow pipes, involved in doing data sharing, building up Table structures, schema and snowflake clone and involved in catapulting data from Teradata to snowflake to consume on Databricks.
* Built and maintained ETL pipelines using PL/SQL, integrating various data sources into the Oracle database, ensuring seamless data flow and processing.
* Used AWS Data Pipeline to schedule an Amazon EMR cluster to clean and process web server logs stored in Amazon S3 bucket and worked on setting up and configuring AWS's EMR Clusters and Used Amazon IAM to grant fine-grained access to AWS resources to users
* Used **SQL Server integration services (SSIS) and server reporting services (SSRS)** for the ETL Purposes.
* Used Machine Learning models - Random Forest, XGBoost, KNN models to predict the outcome.
* Automated data workflows and business processes using Power Automate, reducing manual data entry, enhancing data integrity, and increasing operational efficiency.
* Developed data pipeline for real time use cases using Kafka, Flume and Spark
* Streaming and extracting real time data using Kafka and spark streaming by Creating D-Streams and converting them into RDD, processing it and stored it into Cassandra.
* Demonstrated proficiency in GCP data services, including Google BigQuery, Google Cloud Storage, and Google Cloud Dataflow, for scalable data processing and storage.
* Designed, developed, and maintained data pipelines using GCP services, enabling the efficient ingestion, transformation, and loading of large datasets.
* Contributed to data modeling and schema design within Oracle databases using PL/SQL, ensuring scalability and performance in handling large volumes of data.
* Utilized Google BigQuery for ad-hoc queries and batch data processing, optimizing SQL queries for large-scale data analytics.
* Performed bulk load of JSON data from s3 bucket to snowflake and used Snowflake functions to perform semi structures data parsing entirely with SQL statements and design and development of full text search feature with multi-tenancy elastic search after collecting the real time data through Spark streaming.
* Used Hive QL to analyse the partitioned and bucketed data and compute various metrics for reporting and performed data transformations by writing MapReduce and Pig jobs as per business requirements
* Imported data from RDBMS systems like MySQL into HDFS using Sqoop and developed Sqoop jobs to perform incremental imports into Hive tables and Integrated
* MapReduce with HBase to import bulk amount of data into HBase using MapReduce programs.
* Involved in building up Datalake, loading and transforming the large sets of structured and semi structured data and created Data Pipelines as per the business requirements and scheduled it using Airflow and Oozie Coordinators.
* Worked on migrating MapReduce programs into Spark transformations using Spark and Scala, initially done using python (PySpark).
* Developed Pig scripts to help perform analytics on JSON and XML data and created Hive tables (external, internal) with static and dynamic partitions and performed bucketing on the tables to provide efficiency.
* Involved in file movements between HDFS and AWS S3 and extensively worked with S3 bucket, Redshift, Lambda, SNS, SQS, Kinesis, RDS and AWS Glue in AWS.
* Developed Spark jobs using Scala on top of Yarn/MRv2 for interactive and Batch Analysis and involved in querying data using SparkSQL on top of Spark engine for faster data sets processing and worked on implementing Spark Framework, a Java based Web Framework and developed Spark scripts by using Scala IDE as per the business requirement.
* Use of Docker and Kubernetes to manage micro services for development of continuous integration and continuous delivery.

**Environment**: Hadoop, Hive, HDFS, Pig, Sqoop, Python, SparkSQL, Machine Learning, MongoDB, AWS (S3, EC2, EMR, Athena, Glue, Kinesis, Lambda), Snowflake, Oozie, ETL, Databricks, Tableau, Spark, Spark-Streaming, Pyspark, KAFKA, Netezza, Cassandra, Cloudera Distribution, Java, Impala, Docker, Kubernetes, MySQL, Airflow, AWS, Agile-Scrum.

**Expedia Group May 2018 to December 2020**

**Sr. Data Engineer**

**Dallas, Texas**

**Responsibilities:**

* Design Architecture of data pipeline/ingestion as well as optimization of ETL workflows and developed syllabus/Curriculum data pipelines from Syllabus/Curriculum Web Services to HBASE and Hive tables.
* Performed data analysis, feature selection, feature extraction using Apache Spark.
* Used GCP Bigquery to perform analytical operations on the data, and GCP Firebase to leverage the API’s
* Leveraged SQL Server Integration Services to **building enterprise-level data integration and data transformations solutions**. Used Integration Services to solve complex business problems by copying or downloading files, loading data warehouses, cleansing and mining data, and managing SQL Server objects and data
* Machine Learning streaming libraries in Python and develop Python and bash scripts for automation and implemented Map Reduce jobs using Java API and Python using Spark
* Used Apache Kafka to aggregate web log data from multiple servers and make them available in Downstream systems for analysis and used Kafka Streams to Configure Spark streaming to get information and then store it in HDFS
* Involved in Migrating Objects from On prim to Snowflake and building up snow pipes, involved in doing data sharing, building up Table structures, schema and snowflake clone and involved in catapulting data from Teradata to snowflake to consume on Databricks.
* Ensured data security and compliance by implementing security measures within PL/SQL code, adhering to organizational and industry standards and best practices.
* Used AWS Data Pipeline to schedule an Amazon EMR cluster to clean and process web server logs stored in Amazon S3 bucket and worked on setting up and configuring AWS's EMR Clusters and Used Amazon IAM to grant fine-grained access to AWS resources to users
* Developed data pipeline for real time use cases using Kafka, Flume and Spark
* Streaming and extracting real time data using Kafka and spark streaming by Creating D-Streams and converting them into RDD, processing it and stored it into Cassandra.
* Performed bulk load of JSON data from s3 bucket to snowflake and used Snowflake functions to perform semi structures data parsing entirely with SQL statements and design and development of full text search feature with multi-tenancy elastic search after collecting the real time data through Spark streaming.
* Used Hive QL to analyse the partitioned and bucketed data and compute various metrics for reporting and performed data transformations by writing MapReduce and Pig jobs as per business requirements
* Imported data from RDBMS systems like MySQL into HDFS using Sqoop and developed Sqoop jobs to perform incremental imports into Hive tables and Integrated
* MapReduce with HBase to import bulk amount of data into HBase using MapReduce programs.
* Involved in building up Datalake, loading and transforming the large sets of structured and semi structured data and created Data Pipelines as per the business requirements and scheduled it using Airflow and Oozie Coordinators.
* Documented PL/SQL scripts, data processes, and workflows, contributing to the improvement of data handling procedures and adherence to industry standards.
* Worked on migrating MapReduce programs into Spark transformations using Spark and Scala, initially done using python (PySpark).
* Developed Pig scripts to help perform analytics on JSON and XML data and created Hive tables (external, internal) with static and dynamic partitions and performed bucketing on the tables to provide efficiency.
* Involved in file movements between HDFS and AWS S3 and extensively worked with S3 bucket, Redshift, Lambda, SNS, SQS, Kinesis, RDS and AWS Glue in AWS.
* Developed Spark jobs using Scala on top of Yarn/MRv2 for interactive and Batch Analysis and involved in querying data using SparkSQL on top of Spark engine for faster data sets processing and worked on implementing Spark Framework, a Java based Web Framework and developed Spark scripts by using Scala IDE as per the business requirement.
* Use of Docker and Kubernetes to manage micro services for development of continuous integration and continuous delivery.

**Environment**: Hadoop, Hive, HDFS, Pig, Sqoop, Python, SparkSQL, Machine Learning, MongoDB, AWS (S3, EC2, EMR, Athena, Glue, Kinesis, Lambda), Snowflake, Oozie, ETL, Databricks, Tableau, Spark, Spark-Streaming, Pyspark, KAFKA, Netezza, Cassandra, Cloudera Distribution, Java, Impala, Docker, Kubernetes, MySQL, Airflow, AWS, Agile-Scrum, GCP(Cloud run, Cloud SQL, BigQuery)

**First Republic Bank, New York, NY October 2016 to April 2018**

**Data Engineer**

**Responsibilities:**

* Prepared ETL design document which consists of the database structure, change data capture, Error
* Maintained end to end ownership for analyzed data, developed framework’s, Implementation building and communication of a range of customer analytics projects.
* USed GCP Cloud SQL, Cloud CDN, Run, Cloud functions to bring in insights to business.
* Orchestrated data workflows using tools like Apache Airflow on GCP or Cloud Composer, automating data processes and dependencies.
* Integrated GCP with various data sources, including on-premises databases, third-party APIs, and external data feeds, ensuring seamless data flow.
* Implemented data transformation and cleansing tasks using tools like Google Dataflow or Cloud Dataprep to prepare data for analysis.
* Developed and maintained data warehouses on GCP, using BigQuery and other tools to enable efficient storage and retrieval of structured data.
* Used Kafka producer to ingest the raw data into Kafka topics run the Spark Streaming app to process clickstream events.
* Used SSRS to deliver the right information to the right users. Users can consume the reports in a web browser on their computer or mobile device, or via email.SQL Server Reporting Services offers an updated suite of products:
* **Paginated reports** brought up to date, so you can create modern-looking reports, with updated tools and new features for creating them.
* **A modern web portal** you can view in any modern browser. In the new portal, you can organize and display paginated Reporting Services reports and KPIs. You can also store Excel workbooks on the portal.
* Experience on migrating SQL Database to Azure Datalake, Azure Data lake Analytics, Azure SQL Database, Databricks and Azure SQL Datastoreand controlling and granting database access and migrating on premise databases to Azure Data Lake Store using Azure Data Factroy.
* Architecture and Hands-on production implementation of the big data MapR Hadoop solution for Digital Media Marketing using Telecom Data, Shipment Data, Point of Sale (POS), exposure and advertising data related to Consumer Product Goods.
* Designed and configured Azure Cloud relational servers and databases to analyze current and future business requirements for real-time streaming applications using PySpark, Apache Flink, Kafka, and Hive.
* Migrated data from On-prem SQL server to Cloud databases, including Azure Synapse Analytics (DW) and Azure SQL DB.
* Set up separate application and reporting data tiers across servers using Geo replication functionality.
* Implemented Disaster Recovery and Failover servers in Cloud by replicating data across regions.
* Created pipeline jobs, scheduled triggers, and mapped data flows using Azure Data Factory(V2) and used Key Vaults to store credentials.
* Created tabular models on Azure analysis services for meeting business reporting requirements.
* Worked with Azure BLOB and DataLake storage and loaded data into Azure SQL Synapse analytics (DW).
* Designed SSIS Packages using Business Intelligence Development Studio (BIDS) to extract data from various data sources and load into SQL Server 2012 & 2016 database for further Data Analysis and Reporting by using multiple transformations.
* Utilized Databricks for distributed data processing and big data analytics tasks, such as machine learning, streaming analytics, and graph processing
* Created Elastic pool databases and scheduled Elastic jobs for executing TSQL procedures.
* Spark SQL is used as a part of Apache Spark big data framework for structured, Shipment, POS, Consumer, Household, Individual digital impressions, Household TV impressions data processing.

Environment: Map Reduce, HDFS, Hive, Python,Microsoft SQL Server, PySpark, Apache Flink, Kafka, Hive, Azure Cloud, Azure SQL DB, Azure, Synapse Analytics, Azure Data Factory (V2), Databricks, BLOB storage, Oracle, Informatica 9.6, SQL, MapR, Sqoop, Zookeeper, Data Pipeline, Jenkins, GIT, JIRA, Unix/Linux, Agile Methodology, Scrum.

**Data Engineer**

**Inline InfoTechn Inc. March 2015 to July 2016**

**Responsibilities:**

* Responsible for building scalable distributed data solution using Hadoop Cluster environment with Hortonworks distribution.
* Convert raw data with sequence data format, such as Avro and Parquet to reduce data processing time and increase data transferring efficiency through the network.
* Used SSIS ETL to deal with data transformation and data integration. **Loaded data into Data Warehouse, perform data mining, data cleansing, etc using SSIS**. It extracts data from Flat Files, XML files, SQL databases, etc. It provides GUIs for performing transformations, integrations, and building package
* Worked on building end to end data pipelines on Hadoop Data Platforms.
* Worked on Normalization and De-normalization techniques for optimum performance in relational and dimensional databases environments.
* Designed developed and tested Extract Transform Load (ETL) applications with different types of sources.
* Creating files and tuned the SQL queries in Hive Utilizing HUE. Implemented MapReduce jobs in Hive by querying the available data.
* Exploring with Spark to improve the performance and optimization of the existing algorithms in Hadoop using Spark context, Spark-SQL, Data Frame, pair RDD's.
* Experience with PySpark for using Spark libraries by using Python scripting for data analysis.
* Involved in converting HiveQL into Spark transformations using Spark RDD and through Scala programming.
* Created User Defined Functions (UDF), User Defined Aggregated (UDA) Functions in Pig and Hive.
* Worked on building custom ETL workflows using Spark/Hive to perform data cleaning and mapping.
* Implemented Kafka Custom encoders for custom input format to load data into Kafka portions.
* Support for the cluster, topics on the Kafka manager. Cloud formation scripting, security and resource automation.

**Environment**: Python, HDFS, MapReduce, Flume, Kafka, Zookeeper, Pig, Hive, HQL, HBase, Spark, Kafka, ETL, Web Services, Linux RedHat, Unix.

**Sr Data Engineer**

**Cummins, Delhi, India May 2012 to March 2015**

**Responsibilities:**

* Extensive knowledge/hands on experience in architecting or designing Data warehouse/Database, Modelling, building SQL objects such as tables, views, user defined/ table valued functions, stored procedures, triggers, and indexes.
* Created HBase tables from Hive and Wrote HiveQL statements to access HBase table's data.
* Developed complex Hive Scripts for processing the data and created dynamic partitions and bucketing in hive to improve the query performance.
* Developed MapReduce applications using Hadoop Map-Reduce programming framework for processing and used compression techniques to optimize MapReduce Jobs.
* Developed Pig UDF's to know the customer behavior and Pig Latin scripts for processing the data in Hadoop.
* Used Struts tag libraries and custom tag libraries extensively while coding JSP pages.
* Designed, Developed and Deployed data pipelines for moving data across various systems
* Resolve Spark and Yarn resource management issues in Spark including Shuffle issues, Out of Memory issues, heap space errors and schema compatibility.
* Monitor and troubleshoot performance of applications and take corrective actions in-case of failures and evaluate possible enhancements to meet SLAs.
* Import and export of data using Sqoop from or to HDFS and Relational DB Oracle and Netezza.
* Involved in converting Hive/SQL queries into Spark transformations using Spark Data Frames, Spark RDD.

**Environment**: Hadoop, Sqoop, MapReduce, SQL, Teradata, Snowflake, Hive, Pig, SQL,, Kafka, HBase, Apache, Informatica.

**Education**

* Bachelor of Technology (B.Tech) in Information Technology from Osmania University, Hyderabad, Telangana, India. (2009 – 2013)