**Abhishek J**

**rishitha.23g@gmail.com**

**+1 469 277 7354**

# **PROFESSIONAL SUMMARY**:

* 5+ Years of Professional experience in Analysis, Design, Development, and Implementation as a Data Engineer.
* Excellent understanding knowledge of Hadoop Architecture and various daemons such as Job Tracker, Task Tracker, Name Node, Data Node, Resource Manager.
* Knowledge of designing and implementing complete end-to-end Hadoop infrastructure using MapReduce, Hive, PIG, Sqoop, Oozie, Spark, HBase and Zookeeper and Airflow.
* Experienced in HDFS data storage and running Map Reduce jobs.
* Proficient in data migration from various databases to HDFS using SQOOP.
* Extensively worked on Spark using Python and Scala on cluster for computational (analytics), installed it on top of Hadoop. Performed advanced analytical application by making use of Spark with Hive and SQL, Oracle.
* Involved in converting Hive/SQL queries into Spark transformations using Spark RDD and PySpark concepts.
* Worked with distributions such as Cloudera, Hortonworks on premises clusters and the cloud clusters.
* Dealt with Improving the performance and optimization of the existing algorithms in Hadoop using Spark Context, Spark-SQL, Data Frame, Pair RDD's, YARN.
* Expertise in writing custom Kafka consumer code and modifying existing producer code in Python to push data to Spark-streaming jobs.
* Hands on experience in integration, configuration and maintaining of Hadoop clusters in different environments and also production.
* Ample knowledge on Apache Kafka, Apache Storm to build data platforms, pipelines, and storage systems; and search technologies such as Elastic search.
* Hands on experience in Automation of Sqoop incremental imports by using Sqoop and automating jobs using Oozie.
* Good at implementing Kafka custom encoders for custom input format to load data in partitions.
* Designed and developed the core data pipeline, that involves Java and Scala coding and using Kafka and Spark frameworks.
* Expertise in using Kafka for log aggregation solution with low latency processing and distributed data consumption and widely used Enterprise Integration Patterns (EIPs).
* Extensively used Stash, Bit-Bucket and GITHUB for the code control purpose.
* Proficient in handling and writing SQL queries and Extensive experience in development of Bash scripting, T-SQL, and PL/SQL scripts.
* Extensive experience in creating the Workflows, Worklets, Mappings, Mapplets, Reusable transformations and scheduling the Workflows and sessions using Informatica PowerCenter
* Extensive experience using Microsoft software products including Microsoft Office Suite (Word, Excel, Access, Outlook, PowerPoint, and Publisher) for Windows 7/NT/XP and Vista.
* Strong conceptual, analytical, and design skills and excellent communication skills with leadership qualities.
* Excellent teamwork spirit and capable of learning new technologies and concepts.
* Good knowledge and hands on experience on python modules such as Numpy, Pandas, Matplotlib, Scikit-learn, Pyspark modules.

# **TECHNICAL SKILLS**:

|  |  |
| --- | --- |
| **Programming Languages** | Python, Scala (Functional), Java, SQL, PySpark. |
| **Hadoop ecosystem** | HDFS, Yarn, Apache Airflow, SQOOP, HBase, Hive, Cassandra, Spark, Kafka, Zookeeper with Hortonworks, Cloudera and MapR distributions.  |
| **Cloud services** | Snowflake, AWS (S3, Redshift, Glue, Sage Maker, EMR, EC2 Instance, Lamda, Dynamo DB), Azure (Data bricks, Data Lake, Data factory Logic Apps, Azure cosmos DB, ADLS, Blob), SQL data Warehouse. |
| **Web Technologies** | JDBC, HTML, Java Script, React JS, and CSS |
| **Databases** | Oracle, MS-SQL Server, MySQL, HBase, Cassandra, Mongo DB |
| **Version Control Tools** | GitHub, Bitbucket. |
| **Operating System** | Windows, Linux, Unix, Macintosh HD. |

**EDUCATION:**

* **Southern Arkansas University, Magnolia, AR – 2022**

**Masters in computer and information Science**

* **Malla Reddy Institute of Technology and Science, Hyderabad, INDIA - 2016
Bachelor Degree in Electronic And communication Engineering**

**Professional Experience**:

**Client: ALBERTSONS, CA JULY 2022 – PRESENT**

**Role: Sr. Data Engineer**

**Project Description:**

Implemented Agile methodology for data warehouse development using Kanbanize, leveraging Azure Databricks for organizing and visualizing data in notebooks and dashboards. Performed ETL processes from various source systems to Azure Data Storage services, utilizing Azure Data Factory, T-SQL, Spark SQL, and U-SQL for data ingestion and processing in Azure Databricks. Managed and optimized Spark Databricks clusters, performed data aggregation and validation on Azure HDInsight using Python scripts, and monitored the Hadoop cluster. Developed and modified database objects and performed extensive data manipulation using SQL and PL/SQL, integrating data from multiple sources with SAS/SQL and SAS/Macros. Created data warehouse models in Snowflake, designed ETL processes using SSIS, and developed dashboards with Tableau. Supported data governance, quality, and metadata management, optimizing data solutions with Spark, Scala, Python, and MapReduce for business analytics.

**Responsibilities**:

* Used **Agile Methodology** of **Data Warehouse** development using **Kanbanize**.
* Working Experience on **Azure Databricks cloud** to organize the data into notebooks and making it easy to visualize data using dashboards.
* **Performed ETL on data from different 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. 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.**
* Worked on managing the Spark Databricks by proper troubleshooting, estimation, and monitoring of the clusters.
* Performed Data Aggregation, Validation and on **Azure HDInsight** using spark scripts written in **Python.**
* Performed monitoring and management of the Hadoop cluster by using **Azure HDInsight.**
* Involved in **extraction, transformation** and loading of data directly from different source systems **(flat files/Excel/Oracle/SQL)** using **SAS/SQL**, **SAS**/**macros**.
* Generated **PL/SQL scripts** for **data manipulation, validation,** and materialized views for remote instances.
* Created partitioned tables in **Hive**, also designed a **data warehouse** using **Hive external tables** and also created **hive queries** for analysis.
* Created and modified several database objects such as **Tables, Views, Indexes, Constraints, Stored procedures, Packages, Functions** and **Triggers** using **SQL** and **PL/SQL**.
* Created large datasets by combining individual datasets using various inner and outer joins in **SAS/SQL** and dataset sorting and merging techniques using **SAS/Base.**
* Extensively worked on **AS Visual Analytics** & **SAS** Web Report Studio for data presentation and reporting.
* Extensively used **SAS/Macros** to parameterize the reports so that the user could choose the summary and sub-setting variables to be used from the web application.
* Developed **data warehouse models** in snowflake for various datasets using **WhereScape**.
* Responsible for translating business and data requirements into logical data models in support Enterprise data models, **ODS, OLAP, OLTP** and Operational data structures.
* Created **SSIS** packages to migrate data from heterogeneous sources such as **MS Excel, Flat files** and **CVS files.**
* Provided thought leadership for architecture and the design of **Big Data** Analytics solutions for customers, actively drive **Proof of Concept (POC)** and **Proof of Technology (POT)** evaluations and to implement a **Big Data solution.**
* Worked on **Shell** **scripts** for running **SAS** programs in batch mode on **UNIX**.
* Wrote **Python scripts** to parse **XML** documents and load the data in database.
* Used **Python** to extract weekly information from **XML files**.
* Integrated **Nifi** with Snowflake to optimize the client session running.
* Provided technical support during delivery of **MDM** **(Master Data Management)** components.
* Developed Spark scripts by using **Scala** **shell** commands as per the requirement.
* Worked with **Data Governance, Data Quality** and **Metadata** Management team to understand project.
* Used **Spark SQL** to process huge amounts of structured data to aid in better analysis for our business teams.
* Implemented Optimized join base by joining different data sets to get top claims based on state using **Map** **Reduce**.
* Developed dashboards in **Tableau** **Desktop** and published them on to **Tableau** **Server** which allowed end users.
* Worked on **QA** the data and adding **Data** **sources**, snapshot, caching to the report.
* Involved in troubleshooting at database levels, error handling and performance tuning of queries and procedures.

**Environment**: HDFS, Snowflake, Microsoft Azure services like HDInsight, BLOB, ADLS, Logic Apps, Hive, Klafka, Azure data bricks, Azure data factory, Synapse, Data Lake, Apache Spark, Spark-SQL, ETL, Logic Apps, Maven, Oozie, Scala, Python, Unix shell scripting.

**Client: Edward Jones, MO June 2021 – June 2022**

**Role: Data Engineer**

**Project Description:**

Developed and deployed machine learning models using AWS SageMaker, working with performance metrics such as f-1 score, precision, and AUC. Orchestrated ETL jobs using AWS Step Functions, Lambda, and Glue, integrating Java and Node.js APIs for Lambda to manage AWS services. Implemented ETL processes to transition data from DynamoDB to Redshift using Python-based AWS Lambda functions and developed workflows in Oozie for automating data preprocessing tasks in HDFS. Managed CI/CD processes with Git, Jenkins, Docker, and Kubernetes, and integrated Kafka with Spark Streaming for real-time data processing. Utilized AWS EMR for MapReduce jobs and Snowflake for transforming and querying ORC data, and leveraged Tableau and Alteryx for data visualization and analytics.

**Responsibilities:**

* Used the AWS Sage Maker to quickly build, train and deploy the Machine Learning (ML) models.
* Worked with performance metrics such as f-1 score, precision, recall, log-loss, accuracy, and AUC etc.
* Worked with Machine learning algorithms like Regressions (linear, logistic), SVMs and Decision trees.
* Orchestrated multiple ETL jobs using AWS step functions and lambda, also used AWS Glue for loading and preparing data Analytics for customers.
* Involved in writing Java and Node.js API for Amazon Lambda to manage some of the AWS services.
* Worked on AWS Lambda to run servers without managing them and to trigger run code by S3 and SNS.
* Developed data transition programs from DynamoDB to AWS Redshift (ETL Process) using AWS Lambda by creating functions in Python for certain events based on use cases.
* Implemented the AWS cloud computing platform by using RDS, Python, Dynamo DB, S3 and Redshift.
* Moved data from the traditional databases like MS SQL Server, MySQL and Oracle into the Hadoop by using HDFS.
* Worked on developing workflow in Oozie to automate the tasks of loading data into HDFS and preprocessing with Pig.
* Involved in the creation of Hive tables, loading and analyzing the data by using hive queries.
* Have worked on installation and configuration of EC2 instances on AWS (Amazon Web Services) for the establishment of clusters on the cloud.
* Worked on CI/CD solution, using Git, Jenkins, Docker and Kubernetes to setup and configure big data architecture on AWS cloud platform.
* Exposed to all aspects of software development life cycle (SDLC) like Analysis, Planning, Developing, Testing, implementing and post-production analysis of the projects. worked through Waterfall, Scrum/Agile Methodologies.
* Worked on Informatica on cloud in AWS for POC to build source to target mapping.
* Develop **Informatica ETL** code to load the data to DWH tables.
* Testing of **Informatica ETL** jobs developed
* Used Pig as ETL tool to do event joins, transformations, filter and some pre-aggregations.
* Wrote Spark-Streaming applications to consume the data from Kafka topics and write the processed streams to HBase.
* Working on integrating Kafka Publisher in spark job to capture errors from Spark Application and push into database.
* Used **Amazon EMR** for map reduction jobs and test locally using Jenkins
* Involved in processing log files generated from various sources to HDFS for further processing through Elastic Search, Kafka, Flume & Talend and processing the files using Piggybank.
* Loading and Transforming ORC data into S3 using Snowflake and performed queries using SnowSQL.
* Writing custom Kafka consumer code and modified existing producer code in Python to push data to Spark-streaming jobs.
* Configuring a MongoDB cluster with High Availability, Load balancing and performing CRUD operations.
* Loading data from flat files into the target database using ETL processes by applying business logic for inserting and updating records when loaded.
* Performed day-to-day Database Maintenance tasks including Database Monitoring, Backups, Space, and Resource Utilization.
* Used Tableau Data Visualization tool for reports, integrated tableau with Alteryx for Data & Analytics.
* Building Nifi data pipelines in docker container environment in development phase.
* Creating Docker images to support models in different formats and developed using different languages.
* Used Maven in building and deploying code in Yarn cluster.
* Used JIRA as the Scrum Tool for Scrum Task board and work on user stories.
* Worked on creating Oozie workflow and the coordinator jobs to remove the jobs in time for availability of data.
* Experience in building scripts by using Maven and performing continuous integrations systems like Jenkins.

**Environment**: AWS S3, Redshift, AWS glue, AWS SageMaker, AWS EMR, Nifi, Docker, Kafka, PySpark, Kubernetes, Python, Hive, Active Batch, GitHub, SQL, Cloud Informatica, Tableau, S3, Redshift Linux, Apache Spark, Scala, Spark-SQL, HBase.

**Tapestry, New York, USA December 2020 – May 2021**

**Role: Data Engineer**

**Project Description:**

I led ETL efforts to integrate and transform data from various sources using tools like Sqoop, Kafka, and Oozie, and developed data pipelines with Spark, Hive, and NiFi for data ingestion and processing. I worked extensively with AWS services, including setting up CI/CD pipelines using Jenkins, Docker, and Kubernetes, and developed Lambda functions for event-driven processing. Using Spark Streaming and Kafka, I processed real-time data for analytics, and utilized AWS Glue for ETL tasks into Redshift. I also handled data migration and transformation tasks in Hadoop and Hive, created visualization dashboards in Tableau, and participated in Agile methodologies to ensure project alignment with business goals.

**Responsibilities:**

* Collected and interpreted raw data for data processing.
* Load and transform data visa using scripting languages and tools (Ex: Python, Linux shell, Sqoop)
* Led ETL efforts to integrate transform and map data from multiple sources using python.
* Imported data sets with data ingestion tools like **Sqoop, Kafka, Flume and Oozie**
* Extensively worked with moving data across cloud architectures including **redshift, hive, S3 buckets**
* Provided enhanced operational support on the process of Data Migration, Data Profiling, Data Cleansing, Transformation, Integration, Data Import, and Data Export.
* Developed data pipeline using Flume, Sqoop, Hive and Spark to ingest **subscriber data, provider data and claims** into HDFS for analysis.
* Collected data using Spark Streaming from AWS S3 bucket in near-real-time and performs necessary Transformations and Aggregations on the fly to build the common learner data model and persistence the data in **HDFS**.
* Involved in working with Hive for data retrieval process and writing Hive queries to load and process data in Hadoop file system.
* Worked on NiFi data Pipeline to process large set of data and configured Lookup's for Data Validation and Integrity.
* Develop a script using Jenkins with the integration of the GIT repository for the build, testing, code review and the deployment of the build Jar file, shell-scripts and OOZIE workflows to the destination HDFS paths.
* Working knowledge of **Spark RDD, Data Frame API, Data set API, Data Source API, Spark SQL and Spark Streaming.**
* Load **D-Stream data** into **Spark RDD** and do in memory data Computation to generate output response.
* Experience in working with **Maps, Density Maps, Tree Maps, Heat Maps, Pareto charts, Bubble chart and Bullet Chart, Pie charts, BarCharts,** and **Line charts.**
* I worked on sort & filters of **tableau** like **Basic Sorting, basic filters, quick filters, context filters, condition filters, top filters** and **filter operations.**
* I worked with various formats of files like **delimited text files, click stream log files, Apache log files**, **Avro files, JSON files, XML Files.** Mastered in using different columnar file formats like **RC, ORC and Parquet formats.**
* Worked on CI/CD solution, using **Git, Jenkins, Docker and Kubernetes** to setup and configure big data architecture on AWS cloud platform.
* Developed Spark code by using Scala and Spark-SQL for faster processing and performing data transformations on Member Enrollment, Eligibility, Claims data.
* Administered and Engineered Jenkins for managing weekly Build, Test and Deploy chain, GIT with Dev/Test/Prod Branching Model for weekly releases.
* Extensively used Stash Git-Bucket for Code Control and Worked on AWS Components such as Airflow, Elastic Map Reduce (EMR), Athena and Snowflake.
* Developed scripts for loading application call logs to S3 and used AWS Glue ETL to load into Redshift for data analytics team.
* Used **AWS EMR** clusters for creating **Hadoop** and **spark** clusters. These clusters are used for submitting and executing **Scala** and **python** applications in production.
* Responsible for developing data pipeline with Amazon AWS to extract the data from weblogs and store in **HDFS**.
* Creating and configuring AWS Elastic Load Balancers and Auto Scaling Groups to distribute the traffic and to have a cost-efficient, fault-tolerant and highly available environment.
* Creating AWS Lambda functions to run python scripts, Used AWS Lambda with java to perform event-driven processing.
* Implemented branch incidents competitive process jobs using AWS stack. It processes our records
using Spark, S3, Lambda, Python, Redshift based on various parameters.
* Worked with DevOps team to Culturize NIFI Pipeline on EC2 nodes integrated with Spark, Kafka, Postgres running on other instances using SSL handshakes.
* Worked on **Apache Nifi** to automate the data movement between **RDBMS** and **HDFS**.
* Created shell scripts to handle various jobs like **Map Reduce**, **Hive**, **Pig**, **Spark** etc, based on the requirement.
* Used **Hive**, **Impala** and **Sqoop** **utilities** and **Oozie** **workflows** for **data** **extraction** and **data** **loading**.
* Experienced in writing live Real-time Processing and core jobs using **Spark Streaming** with **Kafka** as a Data pipe-line system.
* Used **Hive QL** to analyze the **partitioned** and **bucketed data, Executed Hive queries on Parquet tables**.
* Stored in **Hive** to perform data analysis to meet the business specification logic.
* **Extracted, Transformed, and Loaded** (**ETL**) and Data Cleansing of data from various sources like XML files, Flat files, and Databases and Involved in UAT, Batch testing, test plans.
* Written several **Map reduce Jobs** using **Pyspark, Numpy** and used **Jenkins** for Continuous integration.
* Used Spark for interactive queries, processing of streaming data and integration with NoSQL database for huge volume of data.
* Involved in scrum meetings, provided daily and weekly status updates.
* Working closely with the product leads to planning quarterly OKRs, define and implement new KPIs.
* Good exposure to automated CI/CD pipelines for code deployment using **Jenkins, Docker, Kubernetes**
* Good experience with the following Version Control / Management tools: Bitbucket, Git, SVN, Github.
* Experienced in Strong scripting skills in Python and UNIX shell

**Environment**: Spark, Scala, AWS Glue, AWS EMR, Redshift, Microservices, GitHub, PySpark, DataPro, Tableau, SQOOP, Python, Linux, NIFI, JSON, HDFS, Kafka, PowerBI, MySQL.

**Client: Axis Bank, Hyderabad August 2018 - September2020**

**Role: Data Engineer**

**Project Description:**

I have extensive experience building distributed, high-performance systems using Spark and Scala, including developing applications for loading and streaming data into NoSQL databases like MongoDB and HDFS. On Microsoft Azure, I worked with HDInsight Clusters, Azure Data Lake, Data Factory, and Databricks for data extraction, transformation, and analysis. I managed Azure infrastructure and implemented BI solutions using Azure services, ensuring efficient data processing and visualization. Additionally, I developed machine learning algorithms with Spark, optimized SQL queries, and created SSIS packages for ETL processes, ensuring data integrity and performance in large-scale data environments.

**Responsibilities:**

* Experience in building distributed high-performance systems using Spark and Scala.
* Experience developing Scala applications for loading/streaming data into NoSQL databases (MongoDB) and HDFS.
* Worked on Microsoft Azure services like HDInsight Clusters, BLOB, ADLS, Data Factory and Logic Apps and done POC on Azure Data Bricks.
* Managed Azure Infrastructure Azure Web Roles, Worker Roles, VM Role, Azure SQL, Azure Storage, Azure AD Licenses, Virtual Machine Backup and Recover from a Recovery Services Vault using Azure PowerShell and Azure Portal.
* Architect & implement medium to large scale BI solutions on Azure using Azure Data Platform services (Azure Data Lake, Data Factory, Data Lake Analytics, Stream Analytics, Azure SQL DW, HDInsight/Databricks, NoSQL DB)
* Working Experience on Azure Databricks cloud to organize the data into notebooks and making it easy to visualize data using dashboards.
* Worked in Developing **Spark** applications using **Spark - SQL** in **Databricks** for data extraction, transformation, and aggregation from multiple file formats for analyzing & transforming the data to uncover insights into the customer usage patterns.
* 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.**
* Responsible for estimating the cluster size, monitoring, and troubleshooting of the Spark Databricks cluster.
* Worked on various compression and file formats like Avro, Parquet, and Text formats.
* Responsible for writing Hive Queries for analyzing data in Hive warehouse using HQL.
* Perform T-SQL tuning and optimizing queries for and SSIS packages.
* Designed Distributed algorithms for identifying trends in data and processing them effectively.
* Creating an SSIS package to import data from SQL tables to different sheets in Excel.
* Used Spark and Scala for developing machine learning algorithms that analyze clickstream data.
* Used Spark SQL for data pre-processing, cleaning, and joining very large data sets.
* Performed data validation with Redshift and constructed pipelines designed over 100TB per day.
* Co-developed the SQL server database system to maximize performance benefits for clients.
* Assisted senior-level data scientists in the design of ETL processes, including SSIS packages.
* Database migrations from traditional data warehouses to spark clusters.
* Ensure the data warehouse was populated only with quality entries by performing regular cleaning and integrity checks.
* Created new cubes and modified existing cubes using SSAS 2008 to make data available
* For the decision makers
* Designed user defined hierarchies in SSAS including parent-child hierarchy.
* Used Oracle relational tables and used them in process design.
* Developed SQL queries to perform data extraction from existing sources to check format accuracy.
* Developed automated tools and dashboards to capture and display dynamic data.
* Installed a Linux operated Cisco server and performed regular updates and backup and used MS excel functions for data validation.
* Coordinated data security issues and instructed other departments about secure data transmission and encryption.

**Environment**: Hadoop, HDFS, Microsoft Azure services like HDInsight, BLOB, ADLS, Logic Apps, Hive, Sqoop, Apache Spark, Spark-SQL, ETL, Maven, Oozie, Scala, Python3, Unix shell scripting, T-SQL, MongoDB, HDFS, Spark SQL, **MS SQL Server Analysis Services (SSAS). Relational Databases, Redshift, SSIS, SQL, Linux, Data Validation, MS Excel.**

**Client: Cybage Software Private Limited Hyd, India June 2017 - August 2018**

**Role: Data Analyst and Quality Analyst**

**Project Description:**

I collaborated closely with business leaders, IT teams, and vendors to ensure clear communication and resolution of data-related issues. I performed thorough data analysis using tools like Microsoft Excel to evaluate data quality, debugged issues between source systems and the Enterprise Data Warehouse (EDW), and implemented data reconciliation and integrity checks. Additionally, I supported the business by identifying reporting needs, developing specifications, and creating test cases and scripts for data testing and validation in the EDW environment.

**Responsibilities:**

* Collaborated with business leaders, IT groups and vendors to ensure issues communicated clearly and accurately.
* Perform data analysis to evaluate the data quality and resolve the data related issues
* Responsible for improving data quality and for designing or presenting conclusions gained from analyzing data using Microsoft Excel as statistical tool.
* Performed Data reconciliation of source, Referential Integrity, Null and Default value checks, and Business Rules applied.
* Debugged data issues between source systems and EDW.
* Profiled data in the sources prior to defining the data cleaning rules.
* Verified the data loads as well as the reports.
* Involved in identifying reporting needs and helping the business in creating specifications and worked closely with report development team to achieve the reporting goal.
* Assisted the project manager in setting up the direction for business analysis team.
* Developed custom routines for generating test cases.
* Created the Test Specifications, Test Scripts and Test Categories for the testing of data in the EDW.
* Attended defective triage meetings with the end users and developers.

**Environment**: ETL Tools, data reconciliation, data cleansing, and data integration, Informatica PowerCenter, Talend Data Integration, IBM DataStage, SQL, data querying, data manipulation, and data validation tasks, Reporting Tools, visualizations, Tableau, Power BI, SAP BusinessObjects, Test Management Tools, test cases, test scripts, Examples include HP ALM (Application Lifecycle Management), JIRA, TestRail, Collaboration Tools, Microsoft Teams, Slack, Zoom.