**Name: Thej**

**BigData Engineer**

**Contact no:** **+1 214-217-5944| Email ID:teja@masitlabs.com**

**Over 10+ years** experience as software professional with rich and varied experience in analysis, **development**, design and **implementation** various **data** **engineering** projects. Posses strong analytical thinking, self-motivation, and an energetic approach with collaborative team skills and leadership skills.

### PROESSIONAL SUMMARY

* Strong experience as **Big Data/Hadoop developer** and **java Developer** with strong emphasis in **Hadoop Ecosystem** using **Spark, Mapreduce , Hive, HBase,Snowflake, Sqoop,Kafka and Hbase**.
* Strong experience in programming languages like **Java**, **Scala** and **Python**.
* Strong understanding of **Distributed** systems design, **HDFS** architecture, internal working details of **MapReduce** and **Spark** processing frameworks.
* Extensively worked on **Pyspark** and **Scala** on Cluster computations
* Migrated Map-Reduce jobs to Spark applications built on **Scala**
* Solid experience developing **Spark** Applications for performing highly scalable data transformations using **RDD**, **Dataframe**, **Spark**-**SQL** and **Spark** **Streaming**.
* Can work parallelly in both **GCP** and Azure Clouds coherently.
* Strong experience troubleshooting **Spark** failures and fine-tuning long running **Spark** applications.
* Strong experience working with various configurations of **Spark** like **broadcast** thresholds, increasing shuffle **partitions**, **caching**, **repartitioning** etc., to improve the performance of the jobs using **Pyspark**.
* Worked on Spark **Streaming** and **Structured** **Spark** **streaming** including **Kafka** for real time data processing.
* Experience in using the cloud services like **Amazon** **EMR**, **S3,** **EC2**, **Red** **shift** and **Athena**.
* In depth knowledge on import/export of data from Databases using **Sqoop**.
* Consultation wif clients on best practices for **data migration**, providing support, and updating implementation team and clients on project status.
*  Very keen in knowing newer techno stack that Google Cloud platform (**GCP**) adds.
* Well versed in writing complex **hive** queries using analytical functions.
* Knowledge in writing custom **UDF’s** in **Hive** to support custom business requirements.
* Experienced in working with structured data using **HiveQL**, **Join** operations, writing custom **UDFs** and optimizing **Hive** queries.
* Solid experience in using the various file formats like **CSV**, **TSV**, **Parquet**, **ORC**, **JSON** and **AVRO**.
* Experienced working with various Hadoop Distributions (**Cloudera**, **Hortonworks**, Amazon **EMR**) to fully implement and leverage various Hadoop services.
* Strong Experience in working with Databases like **Oracle**, and **MySQL**, **Teradata**, Netezza and proficiency in writing complex SQL queries.
* Experience with Snowflake Virtual Warehouses
* In-depth Knowledge of Data Sharing in Snowflake.Database, Schema and Table structures.
* Experience in Migrating Teradata objects into Snowflake Environment.
* Experienced working with JIRA for project management, **GIT** for source code management, **JENKINS** for continuous integration and Crucible for code reviews.
* Proficient in Core **Java** concepts like Multi-threading, Collections and Exception Handling concepts.
* Experience in version control tools like SVN, **GitHub** and CVS.
* Experience of developing applications with Model View Architecture (MVC2) using Spring Framework and J2EE Design Patterns.
* Strong team player with good communication, analytical, presentation and inter-personal skills.
* Expertise in developing both Front End & Back End applications using Java, Servlets, JSP, Web Services, Struts, Hibernate, JDBC, JavaScript, JSON, HTML, and XML.
* Experience with Oozie Workflow Engine in running workflow jobs with actions that run Hadoop MapReduce and Pig jobs
* Expertise in Data Warehousing, Data Migration, **Data Integration** and Reporting using SQL Service Integration Services, SQL Server Analysis Services, SQL Server Reporting Services and Tableau.
* Strong Experience in working with Databases like Oracle, and MySQL, Teradata, Netezza and proficiency in writing complex SQL queries.
* Good understanding of NoSQL databases and hands on work experience in writing applications on NoSQL databases like **HBase**, Cassandra and MongoDB.
* Experience with Talend to develop processes for extracting, cleansing, transforming , integrating, and loading data into data mart database.
* Experienced using the Talend with Big data using Hive, Sqoop, Hdfs components.
* Front-End Developer specializing in React.js ,building responsive, scalable web applications with modern JavaScript, reusable components, and API integrations.
* Strong focus on performance optimization, UI/UX design, and accessibility.
* Experience in Full SDLC cycle which involves architecture, analysis, design, development, testing, Implementation, deployment, Enhancements, and production support using Agile and Waterfall Methodologies.
* Exposure building Java applications and create micro services using spring boot.
* Proficient in Core Java concepts like Multi-threading, Collections and Exception Handling concepts.
* Experience in version control tools like SVN, GitHub and CVS.
* Strong team player with good communication, analytical, presentation and inter-personal skills.

### Technical Proficiencies

**Big Data Technologies:** Spark, MapReduce, Hive, HBase, Impala, Hue, Sqoop, Kafka, Oozie, Flume.

**Programming Languages:** Java,Scala and Python

**Cloud Infrastructure:** AWS S3, Redshift, Athena, Glue, EMR, Step Functions, **Cloud Storage (GCS), Cloud Composer, Big query, Dataproc, Spanner,Azure Data Lake, Data Factory, Data Lake Analytics, HDInsight/Databricks.**

**Databases:** Oracle, Teradata,Snowflake MySQL, SQL Server.

**Version Control:** CVS, SVN and Github.

**Build and Other Tools:** ANT, Maven, Jenkins, Jira, Confluence

### Education

* Bachelors OF Technology in JAWAHARLAL NEHRU TECHNOLOGICAL UNIVERSITY ANANTAPUR.(2009-2013)

### Work Experience

**Lowe’s, Charlotte, Nc 02/2023 – present**

**Big Data Engineer**

**Description :** Building a customer view table in gold layer which has all the information about a customer in 360 degree view considering privacy regulations, transactions related, suppressions and optouts. This view is customer ID level so that we can join multiple buckets on ID and pull information from different sources. This consolidated tables has every information if at all we need few other fields like clickstream browsing activity we need to join on ID, aggregate and predict the activity. This layer is primarily used by data science teams to build models for business teams and build scorecards, dashboards for strategies. This whole project is migrated from Hadoop to GCP(Lift and Shift), but found gaps and rewrite code to capture missing transactions, customer ID changing logic and PII information policy tags, masking the data activities performed.

**Responsibilities:**

* Design and Develop an End-to-End Large-scale streaming data pipeline on Google cloud platform to provide data for building a **Customer Intelligence Platform (Customer 360)** which organize, collect, standardize customer data to maximize customer lifetime value and organization market share.
* Working on Developing a **Pyspark-streaming** compliance application integrating with Kafka and processing the compliance requirements across the data sources in our system on Google Cloud Platform.
* Involved in developing mapping spreadsheets for source to target data mapping with physical naming standards, datatypes, volumetric, domain definitions and corporate meta-data definitions.
* Implementing **Data Warehousing techniques** to decrease the processing time like read and write into BigQuery tables.
* Working on developing pyspark alert fail mechanism to ensure data merges and demerges within the range.
* Migrated high volume **clickstream data pipelines** to Google cloud platform ensuring real-time ingestion and processing.
* Built ETL workflows to transform raw clickstream data into analytics-ready formats, leveraging Big query for storage and analytics.
* Deployed Machine learning models into production integrating with Apache Airflow to provide insights on Data for Stakeholders.
* **Encrypted and decrypted** the data to ensure security and authorization to protect sensitive Personal Data on Google Cloud Platform using Python scripting
* Developing, testing and Monitoring existing Data pipelines on Apache Airflow for scheduling and establishing connection plugins with **Spark and ML models** for Computational Complexity using Python scripts on Google Cloud Platform.
* Built Airflow Dags on **Google Cloud Platform** to schedule Data pipelines for Moving data from Google Cloud Platform to Hadoop.
* Design and Develop an on-prem End-to-End Large-scale data pipeline to provide data for building a Platform (Segment Builder) which serves the business for all the metrics required to implement strategies.
* The pipeline pulls the data from all the required hive sources accordingly make transformations and load the final data into Elastic search.
* Building large scale pyspark applications across the Machine learning platform Kubeflow to provide real time to Data science teams.
* Working on building real time data streaming pipelines with Cassandra to provide pull the data for real time machine learning models.
* Working on tuning spark applications to reduce processing times to meet SLA which involves huge datasets on on-prem.
* Built CI/CD pipelines using Jenkins for Code base and Application Deployment into Production.
* Working with different Analytical teams in Agile Development Environment to meet deadlines.

**Environment**: Spark, Kafka,, Hive, Java, **data migration**, Scala,PySpark,Airflow,Jenkins, **GCP**, Bitbucket, GCS(Google cloud Storage, Dataproc, Big Query.

**Client: UnitedHealth group 02/2021-02/2023**

**Role: Senior Data Engineer**

**Responsibilities**:

* Understanding and implementing based on Business Requirement which is provided by Business Analyst.
* Develop, design data models, data structures and ETL jobs for data acquisition and manipulation purposes.
* Develop deep understanding of the data sources, implement data standards, maintain data quality, and master data management.
* Build Data Build Tool (DBT) Models to implement the code migration to Snowflake
* Used Snowflake’s Snowpipe, Storage Integration object to automate the ingestion of data from the AWS S3 bucket. Further transformations applied on the tables using Stored Procedures.
* Build and Maintain End to End Data pipelines from Azure to Snowflake Data Lake
* Design batch functions in Snowflake for processing the data from Azure to Snowflake.
* Expert in developing JSON Scripts for deploying the Pipeline in Azure Data Factory (ADF) that process the data. Expert in using Databricks with Azure Data Factory (ADF) to compute large volumes of data.
* Performed ETL operations in Azure Databricks by connecting to different relational database source systems using jdbc connectors.
* Developed Python scripts to do file validations in Databricks and automated the process using ADF.
* Developed an automated process in Azure cloud which can ingest data daily from web service and load into Azure SQL DB.
* Successfully employed Snowflake's data sharing, Snow pipe, cloning, and time travel functionalities.
* Developed Streaming pipelines using Azure Event Hubs and Stream Analytics to analyze data for dealer efficiency and open table counts for data coming in from IOT enabled poker and other pit tables. O Analyzed data where it lives by Mounting Azure Data Lake and Blob to Databricks.
* Used Logic App to take decisional actions based on the workflow.
* Developed custom alerts using Azure Data Factory. SQLDB and Logic App.
* Developed Databricks ETL pipelines using notebooks. Spark Data frames. SPARK SQL and python scripting.
* Developed complex SQL queries using stored procedures, common table expressions (CTES), temporary table to support Power BI reports.
* Worked with enterprise Data Modeling team on creation of Logical models.
* Implemented complex business logic through T-SQL stored procedures. Functions. Views and advance query concepts.
* Developed and implemented complex data pipelines using Synapse to orchestrate data workflows, ensuring reliability and scalability.
* Designed and optimized data models within Synapse for efficient storage and querying, leveraging familiarity with both relational and big data concepts.
* Development level experience in Microsoft Azure providing data movement and scheduling functionality to cloud-based technologies such as Azure Blob Storage and Azure SQL Database.
* Independently manage development of ETL processes - development to delivery.

Environment: Azure (ADF, Databricks) python, Synapse, PySpark, Jenkins, git, Databricks, Data lake, delta lake

**BCBS, Kansas city, Mo 01/2020 – 01/2021**

**Big Data Developer**

**Responsibilities:**

* Responsible for ingesting large volumes of **user behavioral** data and **customer profile** data to Analytics Data store.
* Developed custom multi-threaded **Java** based ingestion jobs as well as **Sqoop** jobs for ingesting from FTP servers and data warehouses.
* Developed **Scala** based **Spark** applications for performing data cleansing, event enrichment, data aggregation, de-normalization and data preparation needed for machine learning and reporting teams to consume.
* Experienced in working with spark eco system using **Spark SQL** and **Scala** queries on different formats like Text file, CSV file.
* Developed **Python** scripts to automate data sampling process. Ensured the data integrity by checking for completeness, duplication, accuracy, and consistency.
* Developed **Python** Scripts to filter and Aggragate data.
* Data sources are extracted, transformed and loaded to generate CSV data files with **Python** programming and SQL queries.
* Worked on troubleshooting spark application to make them more error tolerant.
* Worked on fine-tuning **spark** applications to improve the overall processing time for the pipelines.
* Involved in migrating existing traditional ETL jobs to PySpark and Hive Jobs on new cloud data lake.
* Wrote **Kafka** producers to stream the data from external rest apis to Kafka topics.
* Wrote **Spark-Streaming** applications to consume the data from Kafka topics and write the processed streams to **HBase**.
* Experienced in handling large datasets using **Spark** in Memory capabilities, using broadcasts variables in Spark, effective & efficient joins, transformations and other capabilities.
* Worked extensively with Sqoop for importing data from **Oracle**.
* Experience working for **EMR** cluster in AWS cloud and working with S3.
* Involved in creating **Hive** tables, loading and analysing data using hive scripts.
* Implemented **Partitioning**, **Dynamic** **Partitions**, **Bucketing** in **HIVE**.
* Good experience with continuous Integration of application using **Jenkins**.
* Used Reporting tools like **Tableau** to connect with Impala for generating daily reports of data.
* Collaborated with the infrastructure, network, database, application and BA teams to ensure data quality and availability.
* Designed, documented operational problems by following standards and procedures using JIRA.

**Environment**: AWS EMR, Spark, Hive, S3, Athena, Sqoop, Kafka, HBase,Scala,Python,PySpark,Linux,Unix

**Big Data Developer 07/2018 – 12/2019**

**Pinterest, SFO, CA**

**Responsibilities:**

* Involved in requirement analysis, design, coding and implementation phases of the project.
* Loaded the data from Teradata to HDFS using Teradata Hadoop connectors.
* Converted existing MapReduce jobs into Spark transformations and actions using Spark RDDs, Data frames and Spark SQL Apis.
* Written new spark jobs in Scala to analyse the data of the customers and sales history.
* Used Kafka to get data from many streaming sources into HDFS.
* Involved in collecting and aggregating large amounts of log data using Apache Flume and staging data in HDFS for further analysis.
* Designed and developed responsive, mobile-first web interfaces using React.js, HTML5, and CSS3.
* Built modular, reusable UI components with React functional components and hooks (useState, useEffect, useRef).
* Implemented state management using Context API and Redux to handle complex application states.
* Integrated RESTful APIs and handled asynchronous data fetching with Axios and fetch.
* Developed dynamic client-side routing using React Router for multi-page navigation.
* Applied component-level styling using Tailwind CSS, Styled-Components, and Material-UI.
* Ensured cross-browser compatibility and optimized page load performance using lazy loading and code splitting (React.lazy, Suspense).
* Collaborated closely with UX/UI designers to implement pixel-perfect designs while maintaining accessibility (WCAG 2.1) standards.
* Created a Python based web application using Python scripting for data processing, MySQL for the database, and HTML/CSS/ and HighCharts for data visualization.
* Good experience in Hive partitioning, Bucketing and performed different types of joins on Hive tables.
* Created Hive external tables to perform ETL on data that is generated on daily basics.
* Written HBase bulk load jobs to load processed data to Hbase tables by converting to HFiles.
* Performed validation on the data ingested to filter and cleanse the data in Hive.
* Created Sqoop jobs to handle incremental loads from RDBMS into HDFS and applied Spark transformations.
* Loaded the data into hive tables from spark and used ORC columnar format.
* Developed oozie workflows to automate and productionize the data pipelines.
* Developed Sqoop import Scripts for importing reference data from Teradata.
* Environment: HDP, MapReduce, Spark, Yarn, Hive, Tez, HBase, Oozie, Sqoop, Flume, Teradata.

**Client: Enstar Group, Columbia,SC 02/2017 to 07/2018**

**Role: BigData / Hadoop Developer**

**Responsibilities:**

* Involved in writing Spark applications using Scala to perform various data cleansing, validation, transformation and summarization activities according to the requirement.
* Load the data into Spark RDD and Perform in-memory data computation to generate the output as per the requirements.
* Developed data pipelines using Spark, Hive and Sqoop to ingest, transform and analyze operational data.
* Developed Spark jobs, Hive jobs to summarize and transform data.
* Worked on performance tuning of Spark application to improve performance.
* Performance tuning the Spark jobs by changing the configuration properties and using broadcast variables.
* Real time streaming the data using Spark with Kafka. Responsible for handling Streaming data from web server console logs.
* Worked on different file formats like Text, Sequence files, Avro, Parquet, JSON, XML files and Flat files using Map Reduce Programs.
* Developed daily process to do incremental import of data from DB2 and Teradata into Hive tables using Sqoop.
* Wrote Pig Scripts to generate Map Reduce jobs and performed ETL procedures on the data in HDFS.
* Analyzed the SQL scripts and designed the solution to implement using Scala.
* Solved performance issues in Hive and Pig scripts with understanding of Joins, Group and Aggregation and how does it translate to MR jobs.
* Work with cross functional consulting teams within the data science and analytics team to design, develop and execute solutions to derive business insights and solve client’s operational and strategic problems.
* Exported the analyzed data to the relational databases using Sqoop for visualization and to generate reports for the BI team.
* Extensively used Hive/HQL or Hive queries to query data in Hive Tables and loaded data into HBase tables.
* Extensively worked with Partitions, Dynamic Partitioning, bucketing tables in Hive, designed both Managed and External tables, also worked on optimization of Hive queries.
* Involved in collecting and aggregating large amounts of log data using Flume and staging data in HDFS for further analysis.
* Assisted analytics team by writing Pig and Hive scripts to perform further detailed analysis of the data.
* Designing Oozie workflows for job scheduling and batch processing.

**Environment**: Java, Scala, Apache Spark, MySQL, CDH, IntelliJ IDEA, Hive, HDFS, YARN, Map Reduce, Sqoop, PIG, Flume, UNIX Shell Scripting, Python, Apache Kafka.

**Client : Sopra Steria, India 01/2015 to 12/2016**

**Role: Hadoop Developer**

**Responsibilities:**

* Worked on migrating MapReduce programs into Spark transformations using Spark and Python.
* Developed Spark jobs using python along with Yarn/MRv2 for interactive and Batch Analysis.
* Queried data using Spark SQL with Spark engine for faster data set processing.
* Extensively used Elastic Load Balancing mechanism with Auto Scaling feature to scale the capacity of EC2 instances across multiple availability zones in a region to distribute incoming high traffic for the application with zero downtime.
* Created Partitioned Hive tables and worked on them using HiveQL.
* Involved in creating Hive tables, loading with data, and writing hive queries which will run internally in map reduce pattern.
* Used Data Frames and Datasets APIs for performing analysis on Hive tables.
* Monitored Hadoop cluster using Cloudera Manager, interacting with Cloudera support and log the issues in Cloudera portal and fixed them as per the recommendations.
* Responsible for Cloudera Hadoop Upgrades and Patches and Installation of Ecosystem Products through Cloudera manager along with Cloudera Manager Upgrade.
* Used Sqoop for large data transfers from RDBMS to HDFS/HBase/Hive and vice-versa.
* Worked on continuous integration tools like Jenkins and automated jar files at the end of the day.
* Developed Unix shell scripts to load many files into HDFS from Linux File System.
* Monitored workload, job performance and capacity planning using Cloudera Manager.
* Used Impala connectivity from the User Interface (UI) and query the results using Impala SQL.
* Used Zookeeper to coordinate the servers in clusters and to maintain the data consistency.
* Continuously monitored and managed the Hadoop cluster using Cloudera manager and Web UI.
* Used OOZIE Operational Services for batch processing and scheduling workflows dynamically.
* Managed and scheduled several jobs to run over a certain period on Hadoop cluster using Oozie.
* Supported the setting up of QA environment and implemented scripts with Pig, Hive and Sqoop.
* Followed Agile Methodology for entire project and supported testing teams.
* Worked with customers and product manager to prioritize and validate requirements.
* Completed plans for long term goals using Microsoft Project.
* Coordinated the work efforts of 8-person team for various projects. Helped team complete tasks successfully and on-time and resolved obstacles encountered by team members.
* Coordinated and participated in weekly estimation meetings to provide high-level estimates (Story Points) for backlog items.

**Environment:** Hadoop, HDFS, Hive, MapReduce, Impala, Sqoop, SQL Talend, Python, PySpark, Yarn, Pig, Oozie, Linux-Ubuntu, AWS, Tableau, Maven, Jenkins, Cloudera, JUnit, agile methodology.