Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 2, No. 1, February : 2023

REAL TIME CREDIT CARD FRAUD DETECTIONUSING SPARK 2.2

Nakshatra Naik, Omkar Goilkar, Samarth Korgaonkar, Pranav Bhatt Department of Electronics and Telecommunication, Atharva College Of Engineering, Mumbai-095, Maharashtra

Abstract

Abstract The use of computer technology to aid modern fraud transaction problems has revolutionised the banking industry. As there is a tremendous increase in the fraud transac- tions, the study highlights the new technologies help in transforming enterprise into a digital world. The code and server setup is completed using Intellij framework. Project demonstration is done using Intellij Development Environment. Dashboard and real time fraud transaction is demonstrated using Apache Spark Standalone Cluster using Airflow automation.

Keywords: *Apache Spark; Cassandra; Kafka; Intellij; Springboot; Node.js; Airflow Automation.*

Introduction

A credit card allows you to make purchases and pay for them later. In that sense, it's like a short-term loan. When you use a credit card to make a purchase, you're essentially using the credit card company's money. The biggest advantage of a credit card is its easy access to credit. Credit cards function on a deferred payment basis, which means you get to use your card now and pay for your purchases later. The money used does not go out of your account, thus not denting your bank balance every time you swipe. The use of credit cards has replaced cash transactions asthe preferred method of payment.

A large concern on fraud; Credit card fraud is the most common type of identity theft. Credit card fraud is the most common type of identity theft because of the increasing popularity of credit cards. Every year, millions of people fall victim to credit card fraud in INDIA alone due to the estimated 1.5 billion credit cards in the country. Lost or stolen credit cards are sometimes used by criminals to commit fraud. Some people make illegal transactions without ever touching their credit cards. An unauthorised transaction that is carried out with the credit cards, a fraud can take place by physically stealing the card or by stealing the card information via phishing or credit card skimming and using it to make payments, or withdraw cash.

This era of human history where computing has moved from mainframe's PC's to big data and Machine Learning. These technologies contribute significantly in every industry and provideseamless solutions to detect fraud transactions.

1. Related Work

The organisation of the paper is as follows. Section 1 explains the need of automation in banking industry, section 3, 4 and 5 defines the significant concept of Apache Spark framework,Kafka and Apache Cassandra database. Section 6 gives information about the implementation of the study with 5 subsections of Dataset, Experimental Setup, Fraud Detection Architecture, Air Flow Automation and Clean up which comprises information about the main architecture of the study. Section 7 talks about the results and real time fraud alerts displayed on the dashboard.

2. Apache Spark

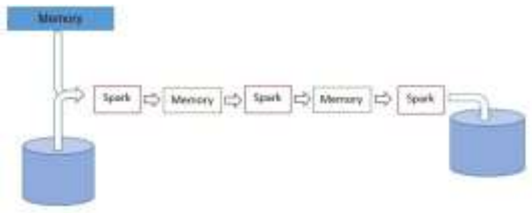

Figure 1. Apache Spark

Spark is an open-source analytics engine for processing large data sets, a distributed processing framework. Apache Spark is an essential component of the Hadoop ecosystem. The framework acts as a single platform for different kinds of data processing techniques such as SQL processing, Machine learning and graph processing. Using a single framework such as Apache Spark, all of these data processing tasks can be accomplished. Data is represented as RD, Data frame or data- set. Spark provides various libraries for data processing. For SQL and ETL processing Spark SQL is used. Real time processing is done by using Spark Streaming library. Machine learning uses Spark ML library. For graph processing Spark GraphX library is used. The framework acts as an in-memory data processing engine, so besides reading the data from the file system, it can also read data from the memory and retain the data in memory.

3. Kafka

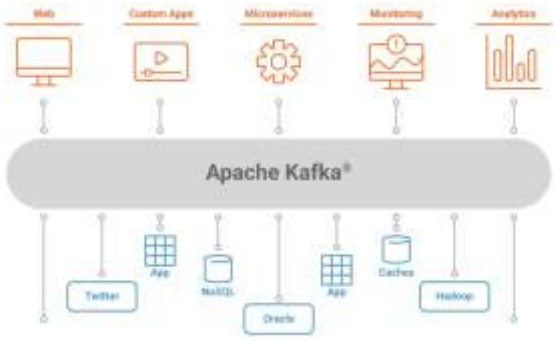

Figure 2. kafka

Apache Kafka is primarily used to build real-time streaming data pipelines and applications that adapt to the data streams. It combines messaging, storage, and stream processing to all low storage and analysis of both historical and real-time data. This is a distributed Persistent Circular Message Queue that is not implemented on a single system but instead on a cluster of systems. For durability, messages are stored on the disk, while for faster access, they are stored in memory. Old messages are deleted automatically every week by default. Through its function as a message broker, Kafka facilitates and processes communication between two applications. In the project, the Credit Card App will send messages to Kafka and applications like Samza. Spark Streaming consumes transaction details from Kafka and will process the details in real time. Web application will directly consume the data from Kafka and will display it on the web UI.

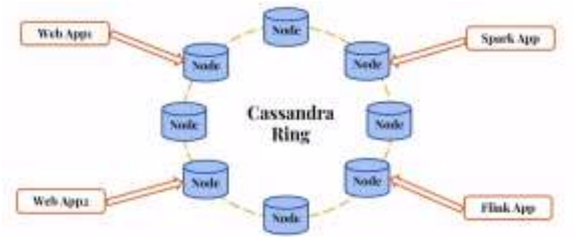

Figure 3. Apache Cassandra

4. Apache Cassandra

The Cassandra database management system is an open source, distributed, NO-SQL database designed to support large amounts of data. Used as a storage layer. Cassandra follows a ring architecture. It is not implemented as a master slave model hence there is no single point of failure. Every node is equal, if any node goes down there is another node to help clients to access the data. Data is directly returned to the memory. Cassandra is linearly scalable and consists of flexible storage. Data must be denormalized through features like 'collections' before storing the data because it does not support joins between column families.

5. Implementation

This section contains an overview of the software installation, code setup and project demon-stration to actively perform the study.

5.1. Data-set

Existing data is present in the resource directory. In this directory there are customer.csv and transactions.csv files. Customer.csv file contains all the details regarding the customer. Transac- tion.csv contains all the transaction details such as credit card number, merchant name, transaction ID and transaction timestamp, etc. These two data sets are imported into the Cassandra database.

5.2. Experimental Setup

The study uses virtual box, Ubuntu and Intellij IDE to run the Spark Jobs. Firstly, connect to Cassandra server through Cassandra console client and create tables and databases in Cassandra. Next we need to start the dashboard server. Whenever there is a fraud transaction it will be displayed here. Zookeeper server is bundled with Kafka, go to kafka and start the zookeeper server. Lastly, start the kafka server and kafka topic, our topic is Credit Card Transaction.

5.3. Fraud Detection Architecture

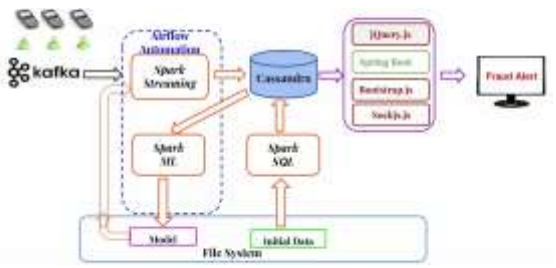

Figure 4. Fraud Detection Architecture

Figure shows that the existing file systems are imported from file system to cassandra , to import the files we'll run spark jobs. The Spark SQL reads the data and will save the data into the Cassandra database. Now, run Spark detection training- Spark ML job which reads the data from cassandra and will train the model accordingly. The model will be saved on the file system. After that will start streaming jobs, it will load the model that was created by previous Spark ML jobs, it will also start consuming the credit card transactions from kafka and will predict whether the Transactions are fraud or not. These predictions are saved on the Cassandra database. After Streaming jobs the Kafka Producer will start. Kafka producer takes input arguments. It will randomly generate the credit card transactions and will publish it to kafka. The spark streaming jobs that are waiting for these messages will consume these messages and will predict if the transactions are fraud or not and will save the predictions to cassandra. From Cassandra database fraud transactions are alerted on the dashboard. The spring boot is used to display fraud transactions on the dashboard.

5.4. Apache Airflow Automation

Air flow automation is required to run the spark jobs on a regular basis. To update the model with the new transaction details we need to run the spark jobs on a regular basis. Spark Jobs reads the old and new transaction data and will create a new updated model. In the project Apache Airflow is used for automation.

5.5. Cleanup.

The cleanup is required so the project can be implemented from scratch. In a cleanup process we need to stop the servers. Always stop the Kafka server first and then the zookeeper server.

6. Result and Conclusion

The results obtained are data or facts obtained from research. In this study, we initiated the discussion by describing the current situation of the credit card business with respect to fraud issues. Although new technology is available and widely supported by banks and merchants globally to lessen or perhaps eradicate the repercussions of credit card fraud, some researchers are starting to challenge its design and implementation. This paper suggests building a model based on the spending behaviour of the card holders and using it to detect anomalous transactions. Benefits of implementing such a detection system will lessen the phone and SMS costs shouldered by the banks; instead of sending SMS transaction notifications to all customers, messages will be sent to those customers with detected anomalous transactions.

References

[1] A NoSQL Database Approach for Modeling Heterogeneous and Semi-Structured Information, 2018 9th International Conference on Information, Intelligence, Systems and Applications (IISA).

[2] Credit Card Fraud Detection Using Machine Learning, 2021 International Conference on Advances in Computing, Communication, and Control (ICAC3)

[3] Machine learning and its applications: A review, 2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC)

[4] Online Transaction Fraud Detection System, 2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)

[5] Credit Card Fraud Detection Using Machine Learning, 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)

.