6/26/2020

LIU,HONGYANG 17201091/1

uNIVERSITY OF MALAYA

WQD7007 BIG DATA MANAGEMENT

PART 1 REPORT

Table of contents

Application of Big Data Techniques in the Financial Field: A framework work for Taking Credit card Fraud Detection as an Example

**Abstract**

People are likely to use credit card for transactions when they bug stuff in supermarket, shopping mall or online shopping. While the credit card has the risk that they may be stolen or used by fraudsters. The fraudulent behavior has caused the customers to have a huge money loss annually. Fraudsters often improve their technology to attack the system or deceive other people to illegally stolen credentials from the card holders and then they could loot them off their money. Meanwhile, the transactions using credit card also happed every day and the number of transactions has arrived at around 150 million every 24 hours in the world. Technically, it’s very difficult to find the fraud behavior manually and the traditional databases are also hard to support the fraud detection in such huge number. Hence, in this report, we decide to propose a big data pipeline to handle the technical problem, such as collecting, storage, accessing or analyzing using provided big data resources.

Keyword: Credit Card Fraud, Big Data Storage, Big Data Pipeline

1. **Introduction**

1.1 Background

Big data has been a buzz word that it prevalent in financial realm. The big data is similar to other term words, it has not a single definition and it has evolved since Roger Moungalas firstly coined the term in 2005. At that time, the big data refers to the data sets are impossible to be processed using traditional techniques. While the big data are not literally referring to the large volume of data sets that contain information, it also includes different dimensions of data sets. According to the study conducted by some organizations, the data has taken increased exponentially in recent years and most of data generated in the last 2 years. The transactions of credit card also have this tendency, most of the credit card transactions generated in recent years. However, the banks face a challenge that it is difficult to handle the huge voluminous data sets for fraud detection using traditional computer techniques. One reason is that with the business growing, the data sets of transactions also increased dramatically. On the other handle, the new E-pay methods has become booming and it often connected with credit card to pay and this way brings more complex data structure to the banks.

The financial industry has utilized big data computer techniques to handle the increasing data to understand user’s behavior and apply data analytics methods to detect credit card fraud detection behavior. The new techniques could improve the performance of fraud transactions detection and prevent the data money loss to some extent. Besides, with the advancements in big data techniques, such as Hadoop, Spark, Hbase. The large amounts of data could be easily stored and analyzed by the banks to extract useful values and gain insights.

* 1. Problem Statement

Due to the exponentially increasing credit card transactions and the challenging fraud behaviors, the financial sectors have to utilize bid data techniques to store, access and manage the datasets. Credit card transactions data should be fetch and analyzed by system in a short time so that the system could detect the suspicious behavior and interrupt the transaction to avoid money loss.

Nowadays, the traditional techniques has two main problems:

* The traditional techniques have little capability to store large voluminous data sets
* The traditional techniques have could not support online data accessing and analyzing
  1. Objective

Fortunately, big techniques are increasingly used to solve big data issues including offline analyzing problem and online analyzing problems. They could help handle the large amounts of data using an unprecedented speed. In this report, we would review the big data techniques and solve the big data issue related with credit card fraud detection based on the three aspects:

* To Identify a most suitable big data techniques to store the credit card transactions
* To Find out a suitable tool to access and analyze the data sets
* To Propose a big data pipeline framework to help solve credit card detection problem

**2.Data Collection**

**Credit Card Transactions data**

We find the data sets “Credit Card Fraud Detection” in Kaggle which is a data science community. The data sets was generated in September 2013 and has 284807 transaction records occurred. While most of them are genuine data sets, accounting for almost 99%.

**A close up of a logo

Description automatically generated**

**Figure 1: Credit Card Transactions**

According to the statistics [2] , Financial sectors has enormous customer data, while these sectors also face the risk that some fraudsters may apply credit card with fraudulent intention or pretend to other users stealing money. On the other hand,

organizations in Financial sector has exploited the credit card transactions to detect fraud behavior and prevent the financial loss caused by fraudster. They would manage the credit card transactions data and apply machine learning algorithms to do data mining, from which they could find the suspicious accounts or transactions related to fraudsters.[1] In addition, financial sector need the transaction data to do data analyzing and protect the customers’ rights.

**7V’s** **characteristics**

[3] Developed a structure to describe the 5’v characteristics volume, velocity, variety, veracity and value of big data. However, in this report, we would describe the 7’v characteristics of credit card transactions. In table 1, we would talk about the characteristics of credit card transactions.

|  |  |
| --- | --- |
| **7 V’s** | **Characteristics** |
| Volume | The amounts of transactions data sets is quite large, if we collected the data monthly or annually to analyze, it would extend PB. |
| Velocity | The speed of generation of data is quite fast, the transactions data would be generated in very short time. |
| Variety | The data is heterogeneous and comes from the online transaction like payments, deposit/ withdrawals and customer demographics. |
| Variability | The variability refers to how long the data would valid and stored. The data sets in fraud detection are constantly changing, as the fraud behavior often happed limited times. |
| Veracity | The transaction of data sometimes has the poor data quality. |
| Value | From the transaction data sets, the financial industry could gain sights from the data and analyze the datasets to find fraudsters. |
| Visualization | The transactions data sets could be visualized through data visualization methods. |

**Table1 7V’s Characteristics for credit card transactions[1]**

1. **Data Storage**

In reality, the data storage techniques could be split into two categories(Table2):

* SQL-based DBMS
* NoSQL(Not Only SQL)

The Relational DBMS like Mysql, Oracle has been widely used by many enterprises to store transactions data sets. This type of databases has relational tables to store the data and support transactions like ACID properties. Hence, the SQL-based DBMS often used for OLTP scenarios.

The NoSQL databases like MongoDB or Hbase often do not support ACID manipulations, but they are good at retrieve documents or unstructured data.

|  |  |  |
| --- | --- | --- |
| **Name** | **Category** | **Data Structure** |
| **Relational database** | **DBMS** | **B+ tree, Hash Table** |
| **Hbase** | **NoSQL** | **Table,Map** |
| **MongoDB** | **NoSQL** | **Document-based** |
| **XML database** | **NoSQL** | **XML** |

**Table 2 Evaluations of data storage techniques**

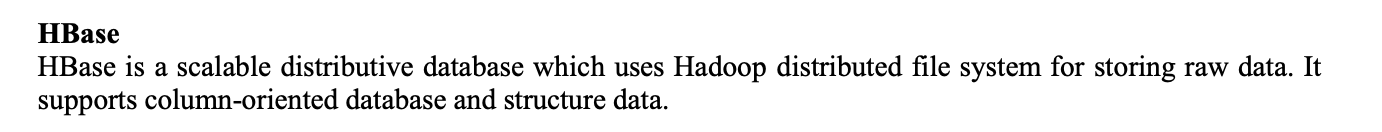
Some credit card transactions are structured like customer’s profiles like age, gender, phone and this data are suitable to store in Relational database. While with the variety increasing, the unstructured may also appear and the data become heterogeneous. The Relational database are difficult to store the unstructured datasets. Besides, with the volume of datasets growing, the traditional database are also challenging for handling the large amounts of data. The multi-dimensions of big data make the traditional databases complicated.

* **Change it for data detection.**

As this transactions data sets are collected from the end devices or clients,

the credit cards transaction datasets are mainly composed of structured data sets like numeric type or string type. Hence, they main be stored in DBMS firstly for supporting ACID properties.

While in the data analyzing part, we need integrate the transactions data sets and customer profiles information. The datasets have been large and the dimension has also increased, we need NoSQL database to store the credit card transactions datasets. The table2 shows that Mongodb is more likely to support documents data structure and xml database is more suitable for XML data sets. While Hbase is an excellent choice for semi-structured datasets and it is column based database that support billions of rows with millions. Consequently, we decide to use Hbase for accomplishing the storage task.



1. **Process to store the data resource**

**Sqoop:** Sqoop is tool used for imported the data sets from RDBMS to Hadoop(Hive, Hbase, HDFS).

It has several advantages:

* Parallel import
* Support most of RDBMS databases
* Directly load data into Hbase

Base on the advantages above, we decide to use this tool import credit card transactions records from bank RDBMS databases into. The process that stores the data resource could be splited into two steps[]:

Create HBase table:

Create ‘Transactions’,’Account’,’Customer’

Execute Sqoop import:

Sqoop Import

**HBase Table Structure:**

A picture containing player, game

Description automatically generated

1. **access data**

* Spark-HBase Connector

We have use Spark-HBase Connector(SHC) to connect the HBase database with Apache Spark. SHC was provided by Hortonworks and it could help fetch the datasets directly from the HBase instead of the process loading the datasets into memories first and then access the data into Spark.

* Spark Mlib

We also need use Spark MLib to access the data imported from HBase and apply machine learning Algorithms on the transactions data sets to do classification and detect fraud datasets.

We decide to apply supervised learning algorithms on the datasets:

Table

|  |  |
| --- | --- |
| **Algorithm** | **Description** |
| Decision tree classifier | Decision Tree are easily to cope with categorial features and do not need feature scaling methods. |
| Random forest classifier | Random Forests is an algorithm that randomly pick up features as the root nodes. It combines multiple decision tree and use ensembles methods to avoid overfitting. |

After the Machine learning part, we could gain a model used for detecting the fraud transactions in credit card. We can also improve the model with the data increased.

1. **Big data pipeline**

**A picture containing device

Description automatically generated**

Figure 6

Conclusion

Batch processing for credit card transactions. We would load the huge amounts of tanscation data using batch processing framework. The datasets was first stored in Relational databases in banking system, we need collect the data sets from the databases. The recommended tool is Sqoop, as it supports batch loading of large size of data. And then , we use SHC(spark-connector) to load the datasets into Spark for data preprocessing. The SHC has advantages that loading data sets directly into Spark and reduce the intermediate procedures. Finaly, we tuilize the machine learning algorithms to traing the datasets and build a fraud detection model.

Howerver, There is a latency that the time data appears in the database and the time for data analyzing. Hence, the we in further work. We could improve the big data framework to make it support real-time data preprocessing.

Reference:

[1]Scope of Big Data Analytics in Industrial Domain

[2] <https://www.finextra.com/blogposting/17847/big-data-in-the-financial-services-industry---from-data-to-insights>

[3] **A SURVEY ON BIG DATA ANALYSIS – AN OVERVIEW**

[4] Big data: A survey

[5] <https://www.dummies.com/programming/big-data/hadoop/importing-data-into-hbase-with-sqoop/>

[6] <https://medium.com/@contactsunny/connect-apache-spark-to-your-hbase-database-spark-hbase-connector-f61f591b75df>

[7] <http://spark.apache.org/docs/latest/ml-classification-regression.html#decision-trees>