6/26/2020

WQD7007 BIG DATA MANAGEMENT

PART 1 REPORT

LIU,HONGYANG 17201091/1
UNIVERSITY OF MALAYA

Credit Card Fraud Detection Model on Financial Transaction Data

LIU,HONGYANG
University of Malaya, Kuala Lumpur, Malaysia

Abstract

People are likely to use the credit card for paying the bills when they buy stuff in supermarkets, shopping malls, or online. However, the credit card has the security risk that they may be stolen or used by fraudsters. The fraudulent behavior has caused customers to have a huge money loss annually. Fraudsters often improve their technology to attack the system or deceive customers and illegally stolen their credentials and then loot them off money. Meanwhile, the transactions that people use credit cards have arrived at about 150 million every 24 hours in the world(Jaidhan et al., 2019). Technically, it's very difficult for banks to find the fraud behavior manually and the traditional databases are also hard to support fraud detection in such huge datasets. Hence, in this report, we decide to propose a big data pipeline to handle these technical problems, such as data collecting, storage, accessing, and analyzing using provided big data resources.

Keywords: Credit Card Fraud, Big Data Storage, Big Data Pipeline, Transaction data

1 Introduction

1.1 Background

Big data has been a buzz word that is prevalent in the financial realm. The big data is similar to other term words, it has not a single definition since it was coined. People often taken large volume of data as big data.

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. Big data refers to the data sets that are impossible to be processed using traditional techniques. According to the study conducted by some organizations, the data has taken increased exponentially in the 21 century, and most of them generated in the last 2 years.

The transactions of credit cards also have this tendency that most of the credit card transactions generated in recent years. In addition, the banks face a challenge that it is hard to handle the large 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 hand, the new E-pay payments have become booming and it often connected with the Internet, from which the new E-pay brings higher risks to credit card fraud.

The financial industry has utilized big data computer techniques to process the increasing data sets and apply data analytics methods to understand users' behavior for detecting fraud behavior. The new technique could also improve the performance of the model in fraud transaction detection and prevent the data money loss to some extent. Besides, with the advancements in big data tools, such as Hadoop, Spark, HBase, the large amounts of data sets could be easily stored and analyzed by the analyzers for extracting useful values and gain insights.

1.2 Problem Statement

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

Nowadays, traditional techniques have two main problems:

- The traditional tools have little capability to store large size of data sets.
- The traditional tools could not support online data accessing and analyzing

1.3 Objective

Fortunately, big data tools are widely used to solve big data problems related to offline and real-time analysis problems. These tools can help process large amounts of data at an unprecedented speed. In this report, we will review big data tools and resolve big data issues related to credit card fraud detection.

- To identify the most suitable big data techniques to store the credit card transactions
- To find out a suitable tool to build credit card detection model
- To propose a big data pipeline framework to help solve credit card offline detection problem

2.Data Collection

2.1 Credit Card Transactions data

We find the data sets “Credit Card Fraud Detection” in Kaggle that is a data science community. The data sets were generated in September 2013 and have 284807 transaction records and genuine data sets account for almost 99%(ULB, M. L. 2018).

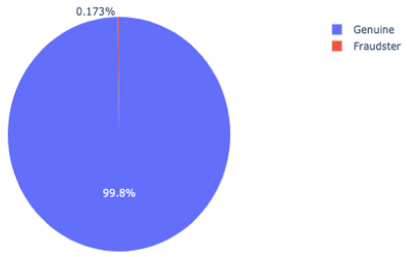

Figure 2.1: Credit Card Transactions

According to the statistics (Jaidhan et al., 2019), financial sectors have enormous customer data, while these sectors also face the risk that some fraudsters may apply for credit cards with fraudulent intention or pretend to other users stealing money. On the other hand, organizations in the financial sector has exploited the credit card transactions to detect fraud behavior and prevent financial loss caused by fraudsters. They would manage the credit card transaction data and apply machine learning algorithms to build models, from which they could find the suspicious accounts or transactions related to fraudsters(Shah & Shah, 2019). Besides, the financial sector needs transaction data to do data analyzing and protect the customers’ rights.

2.2 7V's characteristics

Gupta & Nimbre (2019) developed a structure to describe the 5'v characteristics volume, velocity, variety, veracity and In table 2.2, we would talk about the characteristics of credit card transactions. value of big data. However, in this report, we would describe the 7v characteristics for credit card transactions.

Table 2.2 7V's Characteristics for credit card transactions(Shah & Shah, 2019)

7 V's	Description	Characteristics
Volume	Data Size	The amounts of transactions data sets is quite large, if we collected the data monthly or annually to analyze, it would extend PB.
Velocity	Date generation speed	The speed of generation of data is quite fast, the transactions data would be generated in very short time.
Variety	Data formats	The data is heterogeneous and comes from the online transaction like payments, deposit/ withdrawals and customer demographics.
Variability	Data change time	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	Data trustworthiness	The transaction of data sometimes has the poor data quality.
Value	Data value	From the transaction data sets, the financial industry could gain sights from the data and analyze the datasets to find fraudsters.
Visualization	Data representation	The transactions data sets could be visualized through data visualization methods.

3 Data Storage

In reality, data storage tools could be classified for two categories (Table 3.1)

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

Relational DBMS, such as Mysql, Oracle, and SQL Server has been widely used by many enterprises to store transaction data sets. This type of database has relational tables to store data and supports transactions (such as ACID properties). Consequently, the SQL-based DBMS often used for OLTP scenarios.

However, the NoSQL databases like MongoDB or Hbase often do not support ACID manipulations, but they are good at retrieve documents or unstructured data (Caldarola & Rinaldi, 2015).

Table 3.1 Evaluations of databases

Name	Category	Data Structure
Relational database	DBMS	Tree, Hash Table
HBase	NoSQL	Table, Map
MongoDB	NoSQL	Document-based
XML database	NoSQL	XML

Data Collection Part – DBMS

As credit card transaction datasets are collected from the end devices or clients, the datasets are mainly composed of structured data sets like numeric type or string type. Hence, they mainly would be stored in the DBMS database firstly for supporting ACID properties.

Data Storage Part - HBase

The Relational database is difficult to store the unstructured datasets. Besides, with the volume of datasets growing, the traditional database is also challenging for handling large amounts of data. The multi-dimensions of big data make the traditional databases complicated.

While in the data analysis part, we need to integrate the transaction data sets and customer profiles information. The datasets become large and the dimensions of datasets would also increase. Hence, we need NoSQL database to store the credit card transaction datasets.

The table3.1 shows that MongoDB is more likely to support document data structure and the XML database is more suitable for XML data sets. While HBase is a scalable database and supports the column-oriented database and structure data. It is an excellent choice for semi-structured datasets and supports billions of rows with millions of columns. Consequently, we decide to use HBase for accomplishing the storage task.

4 Process to store the data resource

Sqoop: Sqoop is a 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 Sqoop to import credit card transaction records from bank RDBMS databases into HBase.

1. Create HBase table:

```
Create 'Transactions','Account','Customer'
```

2. Execute Sqoop import:

```
Sqoop Import
```

HBase:

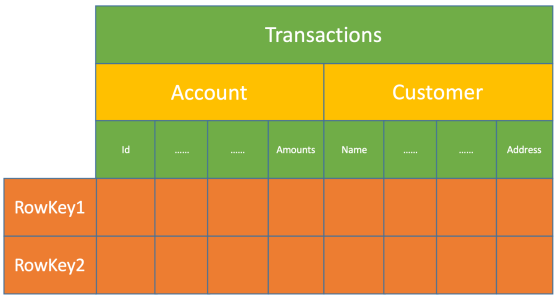

Figure 4.1 HBase Structure

Figure 4.1 shows that the credit card transactions has been imported into HBase datasets.

The transactions table has two column families: Account and Customer. The column names are determined by the import data from RDBMS(e.g MySQL). The row keys represent the transaction records. HBase could support millions of transactions.

5 Access data & Build model

Access data (Spark-HBase Connector)

Spark-HBase Connector(SHC) is a bridge connecting the HBase database with Apache Spark. SHC was provided by Hortonworks. It could help fetch the data records directly from HBase instead of the two steps that loading the datasets into memories first and then loading the data into Spark(Srinidhi, S. 2020).

Build Model (Spark MLlib**)**

We would use Spark MLlib to access the data imported from HBase and apply machine learning Algorithms on the transaction data sets to do classification and detect fraud datasets. Table 5.1 shows the supervised learning algorithms supported by Spark MLlib.

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

Table 5.1 Description of Machine Learning

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.

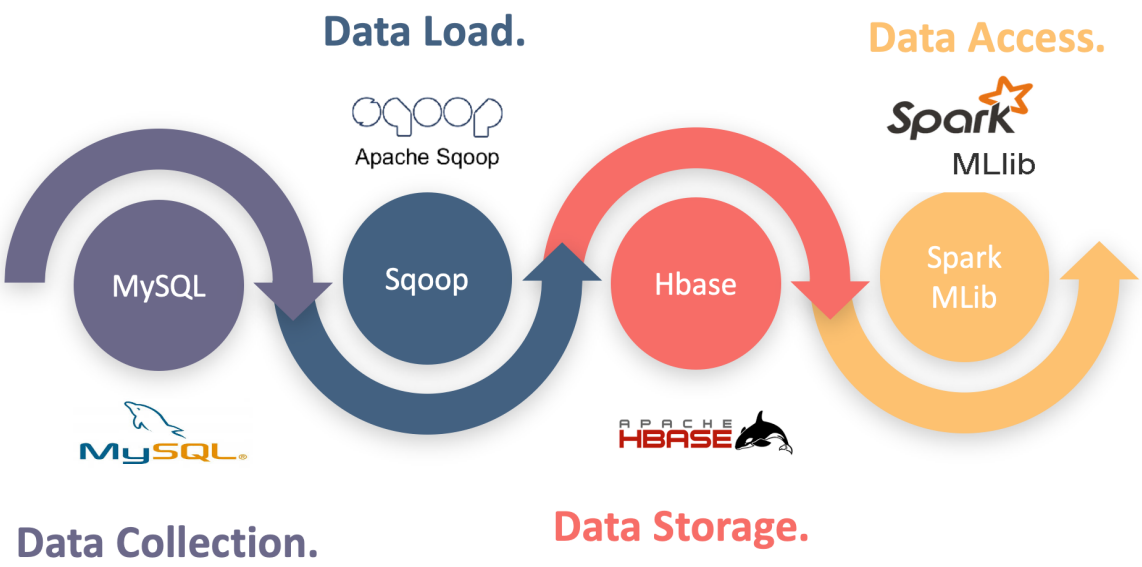

Figure 6.1 Big Data Pipeline

6 Big Data Pipeline

Figure 6.1 illustrates the big data pipeline process. Firstly, the financial transaction data were stored in MySQL database in the banking system. And then, the data sets were loaded from MySQL to HBase by Sqoop. The integration and upload process happened daily or every two days.

The SparkMLlib processes the data for supervised learning. Finally, the output would be a training model for credit card detection.

7 Conclusion

This report shows the batch processing for credit card transactions. The framework aims at solving the credit card offline detection problem. The datasets were stored in relational databases in the banking system. We need to 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 reduces the intermediate procedures. Finally, we utilize machine learning algorithms to train the datasets and build a fraud detection model.

Hence, in further work. We could improve the big data framework to make it support real-time data preprocessing.

References

Caldarola, E. G., & Rinaldi, A. M. (2015). Big Data: A Survey - The New Paradigms, Methodologies and Tools: *Proceedings of 4th International Conference on Data Management Technologies and Applications*, 362–370. https://doi.org/10.5220/0005580103620370

Classification and regression. (n.d.). Retrieved June 28, 2020, from http://spark.apache.org/docs/latest/ml-classification-regression.html#decision-trees

Gupta, H., & Nimbre, S. (2019). *A SURVEY ON BIG DATA ANALYSIS – AN OVERVIEW*. 7.

Jaidhan, B. J., Madhuri, B. D., & Pushpa, K. (2019). *Application of Big Data Analytics and Pattern Recognition Aggregated With Random Forest for Detecting Fraudulent Credit Card Transactions (CCFD-BPRRF)*. *7*(6), 6.

Shah, K. R., & Shah, D. S. (2019). *Scope of Big Data Analytics in Industrial Domain*. *06*(10), 4.

Srinidhi, S. (2020, January 09). Connect Apache Spark to your HBase database (Spark-HBase Connector). Retrieved June 28, 2020, from https://medium.com/@contactsunny/connect-apache-spark-to-your-hbase-database-spark-hbase-connector-f61f591b75df

ULB, M. L. (2018, March 23). Credit Card Fraud Detection. Retrieved June 28, 2020, from https://www.kaggle.com/mlg-ulb/creditcardfraud