Monitoring Data Quality for Very High Volume Transaction Systems*

Joseph Bugajski
Visa International

Robert L. Grossman†
Open Data Group

Eric Sumner
Open Data Group

Steve Vejcik
Open Data Group

July 7, 2006

This is a draft of the paper Joseph Bugajski, Robert L. Grossman, Eric Sumner and Steve Vejcik, Monitoring Data Quality for Very High Volume Transaction Systems, Proceedings of the 11th International Conference on Information Quality, 2006.

Abstract

In this paper we describe a new methodology for detecting data quality problems in high volume transaction streams called change detection using cubes of models or CDCM. We also describe how this system is deployed at Visa and two case studies that occurred during its first year of operation.

*This work was supported in part by the Visa International Data Interoperability Program and the U.S. Army Pantheon Project.

†Robert Grossman is the corresponding author. He is also a faculty member at the University of Illinois at Chicago.

1 Introduction

In this paper, we describe a system that we designed and developed to detect data quality problems in very high volume transaction streams. We also describe how this system is used at Visa, as well as some of experiences using it there.

Visa processes about 100 million authorization messages per day, and another 100 million clearing transactions, totaling greater than USD $4 trillion annual payment volume. There are over 20,000 member banks, over 24 million merchant locations, more than 1.3 billion Visa cards, and hundreds of national settlement networks. The high security, high volume processing network for all these activities is called VisaNet which can handle over 8,000 complex payments per second.

VisaNet connects a merchant's bank (known as the acquiring bank) with the cardholder's bank (known as the issuing bank). Often one or more third party payment processing services may also be involved as data moves from the merchant to the issuer.

Visa processes many point of sale (POS) payments in two steps. First, the merchant requests the card issuing banks approval to accept a card payment. This is called *payment authorization*. A merchant then sends approved and completed sales receipts to the acquirer to clear and settle the transaction with the cardholders' issuing banks. This is called a *clearing transaction*. Both forms of payment data are monitored for data quality problems and data interoperability problems.

We use the term *data interoperability* to refer to data quality problems arising when data is transformed inappropriately as it move from one data processing system to another. Data interoperability problems can be a source of difficulty for complex, distributed systems.

In this paper we introduce a methodology involving baselines for detecting data quality problems and data interoperability problems. We also describe how this methodology is used at Visa to discover discover potential data quality problems in either authorization or clearing data, and how Visa investigates these to determine the best course of action for mediating the problem.

When a data error or semantic inconsistency appears in an authorization messages that error may cause an Visa card issuing bank to inappropriately accept a defective payment or to decline a valid payment. Similarly, a data value or message semantic error in the clearing transaction may cause improper fees assessment, fraud or systemic abuse liability determination, or unnecessary costs for merchants, their acquiring banks, or the issuing banks.

If a data error carries from the transaction message into cardholders' statements, the cardholders may become confused about the source of the transaction, or be unable to reconcile their payments. This may prompt them to call their issuing bank for clarification, which will add costs for the bank and may lead to an inappropriate and expensive dispute. Data semantic errors also adversely effect risk analyses leading to higher rates of declines, exception items, increased merchant discount fees, and inappropriately assess transaction risk. In the worst case, invalid or fraudulent transactions pass risk detection engines leading to cardholder inconvenience, excess processing costs, and financial losses to Visa member banks.

Data quality issues may sometimes inconvenience cardholders; e.g., a legitimate payment is declined for reasons unknown to the cardholder, or an unfamiliar merchant listed in a monthly statement prompts a cardholder to call their bank for an explanation. Similarly, sometimes data quality patterns indicate possible misuse of Visa cards or the payment network; e.g., fraudulent transactions, rules irregularities. Every week, Visa measures billions of payment data records for quality patterns that might have inconvenienced cardholders or indicated card misuse. Baseline measurements provide early detection of these problems that are called "data interoperability problems."

Developing a data quality system given the sheer size and complexity of VisaNet is a major challenge. We feel that this paper makes the following two contributions. First, we have introduced a new methodology called change detection using cubes of models, or CDCM, for monitoring the data quality of very high volume, highly heterogeneous transactions systems. Second, we have operated this system for a year and demonstrated that it is practical and effectively monitor high volume transaction streams containing millions of entities.

A preliminary version of this work was presented in [2]. Although the preliminary work was based upon baselines, it did not describe the CDCM methodology. This paper also describes two cases that resulted during the operational deployment of this system during the past year.

Section 2 describes the data quality methodology we use. Section 3 describes the structure of the data quality program we developed. Section 4 describes two case studies of data quality problems identified by the program. Section 5 describes related work. Section 6 is the summary and conclusion.

2 Change Detection Using Cubes of Models

In this section, we give a quick review of the methodology we use, which is called change detection using cubes of models or CDCM.

We begin with a simple example. Assume that we are interested in monitoring the Point of Sale (POS) Condition Code for a single merchant to determine whether there is a statistically significant change in the distribution of values for this field after a system upgrade by one of the processing centers used by the merchant.

If we have sufficient data prior to the system change we can establish a baseline for the distribution of values for POS Condition Code and then compare the distribution of values after the system change and ask whether the difference in these distributions is statistically significant. We can also ask to detect such a difference as early as possible as we monitor the data after the system upgrade. See Table 2 for an example of a baseline distribution.

For cases like this, it is standard to use a change detection model [1]. A simple example is provided by the CUSUM model. To define this model, assume we have two Gaussian distributions with mean μ_i and variance σ_i^2, $i = 0, 1$.

$$f_i(x) = \frac{1}{\sqrt{2\pi\mu_i}} \exp \frac{-(x-\mu_i)^2}{2\sigma_i}$$

The log odds ratio is then given by

$$g(x) = \log \frac{f_1(x)}{f_0(x)}.$$

and can now define a CUSUM score Z_n as follows [1]:

$$Z_0 = 0.$$

$$Z_n = \max\{0, Z_{n-1} + g(x_n)\}.$$

An alert is issued where the Z_n exceeds a threshold.

In general the statistical distribution of the anomalous distribution is not known. In this case, if the change is reflected in the mean of the observations and the standard deviation is the same pre- and post-change, the generalized likelihood ratio (GLR) score can be used [1]:

$$G_k = \frac{1}{2\sigma^2} \max_{1 \le j \le k} \frac{1}{k-j+1} \left[\sum_{i=j}^{k} (x_i - \mu_0) \right]^2, \qquad k > 1$$

00	7.7451E-01
01	1.0511E-01
02	7.7413E-02
03	2.2383E-02
04	1.0412E-02
05	6.9560E-03
06	2.3694E-03
07	5.5585E-04
08	2.3791E-04
other	5.1907E-05
total	1.0000

Table 1: This table is an example of a baseline table for POS Condition Code for a region of granularity A for a processing entity B over a temporal period of length n days. By adjusting the granularity of the region, the entity B, and the temporal period different numbers of candidate alerts can be generated. The values are illustrative and are not the values associated with an actual baseline.

where μ_0 is the mean of the normal distribution and σ is the standard deviation of the both the normal and abnormal distributions, which are assumed to be Gaussian. Again, the detection procedure is to announce a change at the first up-crossing of a threshold by the GLR score.

Here is a simple example of how we could use a CUSUM or GLR score. Assume that a baseline distribution has been established for the POS Condition Code field as in Table 2 and also assume that each day we monitor the percentage or contribution associated with the POS Condition Code value 00, which represents normally about 77% of the transactions for the particular entity given whose distribution is shown in the table. Assume we monitor the contribution x_1, x_2, x_3, ... to the distribution associated with the payment field value 00 over the next few days, say day 1, 2, 3, ... We can then use the CUSUM or GLR score to alert us to an unusual situation.

The problem we address in this paper is how to scale this up to the millions of baselines that need to be computed to monitor data quality using this approach for VisaNet.

The basic idea is to introduce a data cube with several dimensions and for each dimension an ordered set of break points b_i that split that dimension into separate regions and the cube into separate cells. Given this collection

of cells, we build a separate baseline model for each cell in the data cube. For example, in our case, we could introduce the following dimensions:

1. time (e.g. daily, weekly, monthly, etc.)

2. geographic region (e.g. city, state, country, regional collection of countries, etc.)

3. business entity, (e.g. merchant, acquirer, etc.)

Since they are hundreds of payment field values, ten of thousands different acquirers, millions of merchants, and several different regions, this construction results in many millions of different baselines being estimated.

With so many different baselines, it is easy to produce so many alerts that they become unmanageable, since each one must be examined by a subject matter expert prior. Of course one way to reduce the number of alerts is to raise the threshold for an individual alert. Another way to reduce the number of alerts is to reduce the number of break points so that there are fewer cells in the data cube, and, hence, fewer baselines that need to be estimated.

On the other hand, if there are too few break points the baselines will be associated with so much data and so much heterogeneity that significant events will be missed and the alerts will no longer be meaningful.

Here is an overview of the basic methodology we used.

Change Detection Using Cubes of Models (CDCM).

1. The first step is to fix the dimensions of the data cube.

2. The second step is to determine the initial break points that are used to split the data cubes into cells. The break points in some dimensions were determined by business considerations, such as monitoring acquirers and merchants, while the break points for some dimensions, such as the temporal period for the baselines and the size of the region were determined by adjusting break points to balance the trade off between have enough alerts to be meaningful, but not so many that there were not manageable.

3. The third step is retrieve the required transactional data and compute the desired feature vectors for each cell in the data cube. With large amounts of transactional data and millions of cells, this may require some specialized software.

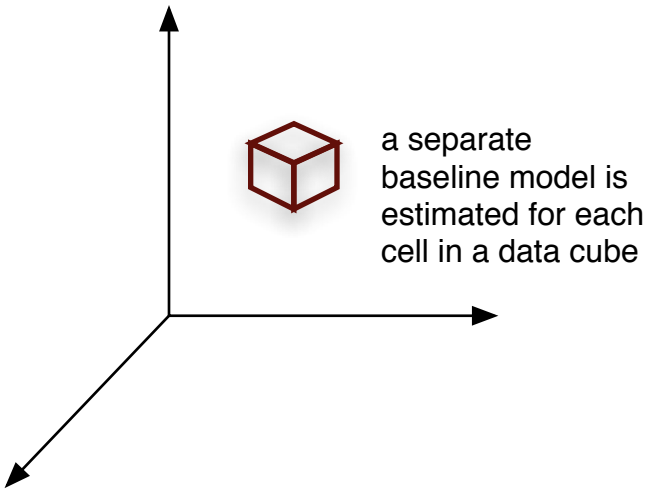

Figure 1: Separate baseline models are estimated and scored for each cell in a data cube. For the project described here, millions of separate baselines were monitored.

4. The fourth step is to compute the baselines by estimating the parameters required by either the CUSUM or GLR algorithm using the feature vectors associated with each cell in the data cube.

5. The fifth step is to use operational transactional data and the baselines to compute scores, and, using the scores, to compute candidate alerts.

6. The sixth step is to add or remote break points in selected dimensions in order to increase or decrease the number of candidate alerts as described above.

3 Program Structure

Program Mission and Governance. The Visa Data Interoperability Program was established in 2004 by the global council of the CIO's of Visa's operating units. The mission of the Program is to detect problems with Visa payment data that result in lost revenue or unnecessary costs for Visa member banks, or inconvenience to Visa cardholders or Visa card accepting

merchants. These problems can arise through errors of data architecture, inaccurate data values, illogical or inconsistent payment data, or errors by applications in how data is interpreted. The program is governed by a council of business executives and technical experts who set governance rules for data and message design.

Reference Model. The governance council adopted standards for analysis, checking validity of data field values, and measuring semantic consistency of transaction records. These rules are recorded in a Reference Model that is maintained by the Program and updated at least twice each year.

Scoring of Transactional Data Using the Monitor. The rules in the Reference Model are encoded in baseline models using the methodology described above. A system was designed and developed that uses these baseline models to monitor transactional payment data. For the purposes here, we call the system the Monitor. The Monitor receives daily samples of tens of millions of authorization messages and clearing transactions from a central ETL facility inside VisaNet. Statistically significant deviations from baselines that are associated with high business value generate what are called candidate alerts.

Investigation of Candidate Alerts. Candidate alerts are analyzed by program analysts and other subject matter experts to understand the issues that led to the candidate alert and to more carefully estimate the business value involved. If the program team believes that an issue is valid and sufficiently valuable that the cost of repair may be recovered through recaptured revenue or lower processing costs, and, furthermore, they believe that the issue is sufficiently clear that it may be explained accurately, they send a Program Alert to the customer relationship manager at the Visa operating region that is closest to the source of the problem. This may be a third party processor, an acquiring bank, or a VisaNet technical group.

Monitoring Program Alerts. The customer relationship manager works with the program analysts to explain the problem identified by the Program Alert to the bank or merchant and to work with them to estimate the cost required to fix the problem. The program team meanwhile reviews measurements to determine when and if the problem is resolved. If the data measurements indicate a resolution, then the business that effected the change is contacted once again to validate recovery of revenue or loss avoidance.

The issues described in the two case studies written in the next section all completed this process and were shown to have produced substantial

value for the merchant, processor, or bank indicated.

4 Two Case Studies

Sudden Changes in Sales Channel. Every Visa payment transaction bears some degree of risk until the transaction has been cleared, settled, and the cardholder makes their monthly payment. Some sales channels and merchants, and some payment conditions, carry inherently more risk than others.

A low risk transaction might be one where a Visa cardholder buys a modestly priced product from a merchant they know well. Moreover, the merchant knows the cardholder and that merchant routinely verifies the authenticity of every payment that they accept. For example, in-store sales by retailers to cardholders who regularly shop at their stores are known as *card present* transactions.

Conversely, a high risk transaction might be a large value purchase of a product or service on-line, where neither the merchant nor the cardholder may validate the authenticity of the other and the transaction completes as clear text on the Internet; these are called *card not present* transactions. Higher risk transactions are subject to more scrutiny by banks and by Visa than lower risk transactions. Also, higher risk transactions are more expensive to process than lower risk transactions.

A number of data fields in authorization messages carry information about the point of service (POS) environment, the merchant's line of business and location, evidence of cardholder's identity, transaction currency, amount, and bank routing instructions. Erroneous data in any of these fields may result in a payment transaction being classified as higher risk than it is, or as lower risk than it is. Data interoperability baselines are kept not only for individual fields in the payment record, but also for combinations of fields.

For example, a cardholder who has never traveled overseas, uses her Visa card on-line every month to pay the same merchants for her email service, cable television, and telephone bill, and the on-line payment services she uses process with a Visa approved security protocol, then those transactions would be considered relatively low risk. Conversely, should a her Visa card be used overseas to buy a $20,000 ring from merchant who indicates that the cardholder was not present for the transaction, then that transaction would be considered very high risk. If the data fields that describe these situations contain invalid or incomplete information, or the information presented by

combinations of valid data field values present an illogical payment description, then the true risk associated with these payments may be higher or lower than they appear to be.

For these reasons, Visa's data interoperability program uses baselines to find unusual, invalid, illogical, or ambiguous information in data fields, or in combinations of data fields, that may cause the risk associated with transactions to be characterized incorrectly. Each participant to the transaction adds information to the transaction messages that indicate the nature of the transaction and their handling of it.

Specifically, baseline scores are computed using the POS Condition Code field and the Merchant Channel and Category (MCC) field, which we now briefly describe.

The merchant channel and category field (MCC) is a four digit number assigned by the International Organization for Standards (ISO Technical Committee 68) to indicate the sales channel and category of goods of services sold by a merchant. For example the MCC field distinguishes between a tele-marketer and a retail merchant and between men's clothing and jewelry stores.

The POS entry mode field indicates the technique employed by the merchant to obtain the account number; e.g., magnetic stripe reader, chip card reader, key entry or telephone. The POS condition code field indicates operational characteristics of the card acceptance terminal; e.g., magnetic stripe and chip reading capability. Other fields define the transaction as having been accepted by mail order, telephone order or Internet and card authentication means employed at the POS.

In 2006, baseline measurements based upon these fields were used to detect a gradual change in merchant channel characteristics of a European acquiring bank. In early 2006, baseline measurements showed that the acquirer began processing more Internet and telephone order transactions then previously. Note that this meant that the acquirer was processing more cards not present transactions that carry higher risk than card present transactions, and these cost the bank more to accept. While not initially cause for alarm, several weeks of steady increase in card not present acceptance activity finally took the acquirer to the point of accepting a majority of their business as card not present transactions.

This activity raised a data interoperability alert that a program analyst then investigated. The analyst discovered the root cause of the problem by working directly with the acquiring bank. The problem was a software error introduced when the bank had updated a key system. When the bank changed their processing rules for another card company, they inadvertently

had changed processing rules for Visa transactions. Because the change caused their transactions to be reclassified to a higher risk category, the bank operations executives were delighted to know about the error. They corrected the software error, thereby saving millions of dollars of processing costs associated with higher risk, card not present payments.

Increase in Declines. Another example of the effectiveness of baseline measurements is related to third party processing services. Most airlines use trip scheduling services provided by one of the major airline reservation systems; e.g., Galileo, Amadeus, and Sabre. These services, also known as Global Distribution Systems (GDS) have Web site presence through Travelocity, Orbitz, individual airlines, and many other locations worldwide. Travel agents long have been subscribers to these reservations services, and the largest of these services connect directly to Visa processing systems via the Visa Global Airline program. Because Visa cardholders frequently use their cards to purchase airfare, any inconvenience to them is a concern to Visa.

Suppose a cardholder makes a travel reservation to fly from Dallas Texas to New York City using a secure on-line service like that provided by Travelocity. The reservation company behind Travelocity is Sabre, a spin-off of American Airlines. The Sabre system will send a request for payment authorization directly to Visa who route such requests to the card issuing bank. If the card issuing bank authorizes the payment, the airline reservation system issues the ticket and they send a sales draft to the airline's acquiring bank to collect payment for the airfare. In addition to cardholder and sales data, several other data fields in the payment authorization message provide the card issuing bank with information about the airline, the reservation system and the acquiring bank.

Baseline measurements of payment authorization messages from a global airline reservation system showed that under certain the circumstances the name of the airline and the country of operation where missing. This resulted in a sudden increase in authorization being declined for an airline reservation systems. Additional investigation indicated that when this occurred tens of millions of dollars of business were being lost to Visa competitors and to Visa banks that participate in these airline transactions. A Program alert was issued and the Program analysts worked with the acquiring bank and determined that losses were incurred when merchant name data turned up missing. By working with the manager of the Global Airline program, the problem was found to be a software update error that resulted in the loss of the airline name from the authorization messages but not the final sales

draft. The problem was corrected by the reservation service ending a serious inconvenience to Visa cardholders, the airline and increasing revenue to the banks.

5 Related Work

The use of statistical methods in data quality goes back at least to Deming [4]. Our approach is to estimate baselines and to measure statistically significant deviations from baselines. This is a standard approach in change detection [1]. In contrast, a common alternative approach for measuring data quality is engineering based (see for example, [7], [11] or [10]).

The work describe here is similar to other data and information quality methodologies [7], [9], [8], [6], [3], that contain components for defining, measuring, analyzing, and improving data and information quality issues.

On the other hand, the CDCM methodology described above is to the best of our knowledge the first data quality approach that tackles the complexity and heterogeneity of high volume transaction processing systems by using data cubes of models to break up the problem into smaller more manageable pieces.

6 Conclusion

In this paper, we have introduced a new methodology for detecting data quality problems in high volume transaction processing systems called change detection using cubes of models or CDCM. We have also described how this system was used at Visa and two case studies that occurred during its first year of operation.

The CDCM methodology compares operational transactional data to baseline models and issues candidate alerts when there are statistically significant deviations from the baselines that are also associated with high business value. Since all candidate alerts must be investigated by subject matter experts, the process runs into problems if too many candidate alerts are generated.

By using removing break points that divide the data cube into cells, fewer alerts are produced. On the other hand, if too many break points are removed so few alerts are produced that they are not meaningful.

The work described here was in part responsible for the baseline proposal pending before the Predictive Model Markup Language (PMML) Working

Group [5], which is the vendor led standards group for statistical and data mining models.

To summarize, the CDCM methodology has been deployed for two years at Visa and proven to be an effective mechanism for detecting potential data quality and data interoperability problems.

References

[1] M. Basseville and I. V. Nikiforov. Detection of Abrupt Changes: Theory and Application. Prentice Hall, 1993

[2] Joseph Bugajski, Robert Grossman, Eric Sumner, Tao Zhang, A Methodology for Establishing Information Quality Baselines for Complex, Distributed Systems, 10th International Conference on Information Quality (ICIQ), 2005.

[3] Tamraparni Dasu and Theodore Johnson, Exploratory Data Mining and Data Cleaning, Wiley, 2003.

[4] W. Edwards Deming, Elementary Principles of the Statistical Control of Quality: A Series of Lectures, JUSE, Tokyo, 1952.

[5] Data Mining Group, The Predictive Model Markup Language, Version 3.0, retrieved from www.dmg.org on July 5, 2006.

[6] DOD Guidelines on Data Quality Management (Summary), retrieved from tricare.osd.mil/rm/documents/fa/ DoDGuidelinesOnDataQualityManagement.pdf on March 20, 2004.

[7] Yang W. Lee, Diane M. Strong, Beverly K. Kahn, Richard Y. Wang, AIMQ: A Methodology for Information Quality Assessment, Information and Management, December 2002, Volume 40, Issue 2, pages 133–146.

[8] Ken Orr, Data Quality and Systems, Communications of the ACM, Volume 41, Number 2, 1998, pages 66–71.

[9] Leo L. Pipino, Yang W. Lee and Richard Y. Wang, Data Quality Assessment, Communications of the ACM, Volume 45, Number 4, 2002, pages 211–218.

[10] Thomas C. Redman, Data Quality: The Field Guide, Digital Press, Boston, 2001.

[11] D. M. Strong, Y.W. Lee and R.Y. Wang, Data Quality in Context, Communications of the ACM, Volume 40, Number 5, 1997, pages 1030-110.