Diligent Planning, Right Strategy in Big Data Projects – Key for Success

By **Chuck Rehberg,** CTO, Trigent Software Inc

Background

A company has trading software where they provide the ability to buy and sell a wide variety of products. The company provides a browser-based application used by more than one million clients world-wide per day. Each client has their own secure account. Unfortunately for the company, they have found that it is possible for a group of [seemingly] diverse clients to work together to manipulate trading activity in a variety of ways. It seems such "coordinated attacks" are often carried out over a relatively short time period and at different times during the day. It also seems the attackers mostly use stolen client accounts.

The company has the ability to monitor the website activity on a session-by-session basis at a fine granularity (i.e. mouse and key strokes). They record the apparent IP address where the website is being accessed and they track the sequence of Buy and Sell transactions for each account by session.

The company would like to know which accounts have [likely] been hacked and which accounts are [likely] being used to manipulate trading activity. Ideally, the company would like to know this in real time to allow them to mitigate any damage.

The company created a project with the goals to:

1. Find a way to identify accounts being used to manipulate trading activity.
2. Find a way to identify accounts that have been hacked.

The approach

The company's techies immediately jumped on Big Data technologies for this task. The task seemed perfect. They were already collecting every keystroke and button push. Fortunately, they discovered (albeit slowly) that they needed to:

1. Carefully align their work with the company's business goals,
2. Better understand the available data, and
3. Know how this data can contribute to the goals

The project team assumed the usage profile of a hacked account would generally look different than the typical usage profile (i.e. usual sequence of mouse clicks and key strokes) of the account's "rightful" owner. They assumed that each account owner tends to establish a set of somewhat predictable patterns of usage leading to a given transaction.

After wider discussion, it became clear that a mere change in account usage pattern would not be strong enough evidence to flag an account for the relatively expensive task of validation. In addition, establishing and maintaining usage profiles given the number of accounts and the high frequency of transactions may be too much of a stretch for their first Big Data project. So at the very least, they would need to find a way to reduce the number of account sessions to check.

Account transaction data was collected and maintained by a different group. This group monitored the state of current transactions and ensured their fulfillment. They also performed a variety of trend analysis and supported market research. Any time a coordinated attack was discovered or suspected, this group was charged with examining the historical data to determine how it happened and who might have done it. Over time, they began noticing activity patterns characteristic of attacks.

By working with expert fraud investigators, an initial set of profiles for coordinated attacks were identified. The patterns were specified in terms of activities usually directed at one or more products, often sequential and often within limited time

intervals. The products being targeted were usually linked in some way, providing an additional source of information. The location and distribution of the attacking sessions were also a factor. Occasionally, other random activities and products would be included within any given session, perhaps to avoid initial detection or to use the same hacked account later.

changes are flagged. For each such potentially "under attack product", the sessions involved are examined to determine if they match any of the existing coordinated attack profiles. [Note: longer term sustained attacks usually originate from continuously changing accounts and require a modification to this approach.]

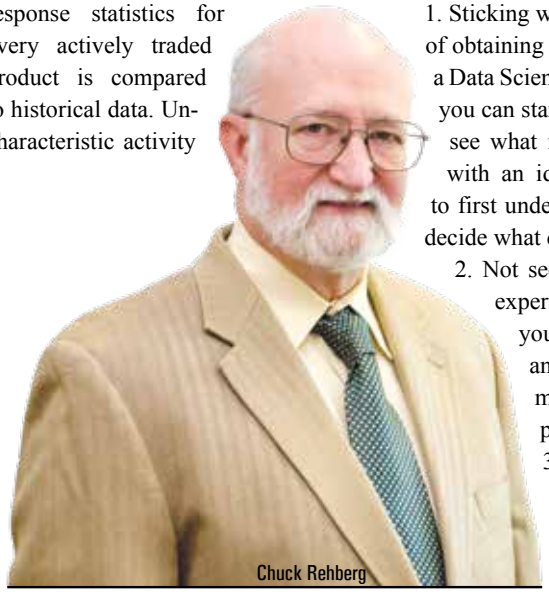
The rapid adoption of new technologies and business models creates a need for more innovative risk management solutions

The Plan

The "profiles of a coordinated attempt to manipulate trading activity" were expressed as a set of pattern recognition rules. A rules-based system was chosen that can identify sequential time constrained patterns, handle high-noise data, and work at speed.

Streaming transaction session data is collected in a NoSQL database. As soon as a session ended, the data is transferred to a Hadoop cluster. This introduces a slight real-time delay since a client's account session can span multiple minutes or longer.

Current transaction/response statistics for every actively traded product is compared to historical data. Uncharacteristic activity changes are flagged. For each such potentially "under attack product", the sessions involved are examined to determine if they match any of the existing coordinated attack profiles. [Note: longer term sustained attacks usually originate from continuously changing accounts and require a modification to this approach.]


Chuck Rehberg

Finally, the usage profile of each session identified as fitting a coordinated attack profile is compared to historical usage patterns to determine the probability that the account was hacked. When a "hacked account usage pattern" coincides with transactions that fit the profile of a coordinated attack, the account is flagged for verification and a remediation is initiated.

Epilogue

In my experience, projects like this often do not end well. There are so many ways they can go wrong. Some of the top ones I've seen are:

1. Sticking with the data you have instead of obtaining the data you need. If you are a Data Scientist on a voyage of discovery, you can start with the data you have and see what it shows. If you are starting with an identified problem, it is best to first understand the problem and then decide what data you need.
2. Not seeking help from the domain experts. You may find some of what you are doing is unnecessary and some of what is needed may not fit the technology you planned to use.
3. Not understanding the nature of the problem. Are you looking for specific needles in a "needle stack", discovering trends over time, or finding data correlations? Or some of each? This has critical implications on skill sets and technology choices.
4. Not being receptive to new ideas or new requirements. Sometimes after a Big Data project is underway, an alternative [algorithmic] approach is discovered to work better.

Recommendations:

1. Create an information architecture. To the extent possible (within your constraints), understand how each unit of data relates to other units and how the information will be used
2. Understand your available sources of data. Know how and when the data is available and can be accessed.
3. Most importantly, "understand the goal"

Here are some Critical Questions:

1. Do you have the data you need to meet the goal? If your answer is yes, can you test this assumption up front?
2. What do you gain from using massively scalable technology? Are you anticipating a future need? Can you make that case?
3. What is the time value of the answers you seek? (i.e. Does the value of the answer diminish over time? Is the answer intended to be actionable now? Is later ok? How much later?)
4. How accurate does the answer need to be?
5. What kinds of answers are the users/consumers looking for?
6. Does the planned target system support the user/consumer'slevel of understanding?

Big Data technology increases the possibility of

(1) Finding individual data lost in a sea of noise

(2) Discovering trends over time and using them to predict or even change the future

(3) Discovering/ verifying correlations between people, places, genes, environment, products, the environment and much more.

However to be successful, the use of Big Data tools need to live within an established information architecture. CR