Detecting Anomalous Behavior with the Business Data Lake

Protecting the Enterprise with Next-Generation Analytics.
Data Science and Analytics helps you to move beyond the threats you can see to anticipate the threats that you can’t

Our Anomalous Behavior Detection Solution addresses security issues that conventional methods can’t. It can help to detect and prevent theft of data or Intellectual Property (IP), for instance at the behest of nation states or organized crime, or by a disenchanted employee. The solution can quickly identify when a user is behaving in a way that is abnormal for them and take appropriate action to limit what they can do, or flag up the situation for managerial attention. It can also predict when anomalous behavior is likely to occur, flagging events of interest for further investigation for potential security breach.

The solution builds on Capgemini’s Business Data Lake (BDL), a joint development between Capgemini and Pivotal. The BDL has the benefit of being able to store vast amounts of information, both structured and unstructured, affordably, and present analytics outcomes to the business. Information is nothing without analytics however and we apply data science techniques to the information in the BDL to achieve the Anomalous Behavior Detection Solution. These techniques leverage machine learning to build up a mathematical model of how your business operates and provide insight into when individuals or systems are acting outside of standard operating procedures and may be presenting a risk to your business.

To ensure that your business is getting continual value from the solution we are able to first identify the challenge, secondly prove we can solve it through a PoC and finally industrialize and even support that so it becomes a standard part of your businesses anti-fraud and security measures.
The most elusive security issues could pose the biggest threat

Over the past 20 years, enterprises have been fairly successful in combating “the threats you can see”. Measures have included antivirus, firewalls, access controls, and physical security of the data center, along with Governance, Risk and Compliance (GRC) methodologies. Some organizations have also introduced security operations centers that capture and analyze network traffic to discover threats.

What most organizations have yet to do is find a way to tackle internal threats from employees and others who operate within their approved authority, but start to abuse delegated business privileges and rights. Staff may have been subjected to psychological manipulation, blackmail, bribery or be driven by simple greed to sell sensitive corporate information to nation states, competitors or organized crime. Examples such as Nortel Networks show the reality of this threat and its impact on the valuable IP of a company.

An example of the challenge is a system administrator who starts passing downloaded data to unauthorized individuals or organizations. Even advanced security solutions can’t detect this, because the individual is accessing only the systems that they’re allowed to access.

Organizations need a way to pre-empt attacks

Most businesses are already well aware that they need to identify anomalous behavior, behavior that steps outside the norm, before the damage is done, but until recently, this hasn’t been practical.

The challenge has been that companies lack the analytic capability to identify anomalous behavior fast enough for action and reliably enough to be taken. This has been because historically the technologies required to undertake what is a hugely complex mathematical task have been prohibitively expensive.

The second challenge around the information required for that analytics is not its existence but its availability. The information is fragmented across multiple systems used by different organizational functions and is often thrown away due the cost of storage and an inability to drive insight from it.

To address the challenge of anomalous behavior therefore two pieces must be addressed, firstly the mathematical analytics and data science must be available at a reasonable cost and secondly there must be a way which cost effectively consolidates the information required.
Capgemini’s Anomalous Behavior Detection Solution

Capgemini now offers a solution that addresses these two challenges and can spot anomalous and suspicious behavior by applying advanced, machine learning algorithms to a broad and deep big data set. By applying innovative data science techniques to huge volumes of data, you can detect atypical patterns of behavior fast, and often take action before a threat becomes critical. The solution:

- Learns what is normal, and the difference between what is “approved” (actions that someone should be taking in line with their role), what is “allowed” (actions that someone has the ability to take within their role, although not explicitly approved is allowed) and what is “suspicious” behavior that indicates a clearly different pattern of behavior.
- The Data Science model identifies anomalies and provides an initial score which allows you to categorize, prioritizing the risk for further investigation, taking action immediately or teaching the mathematical model that this behavior is either approved or allowed.
- Creates alerts so you can react before an anomaly becomes a problem.

The mathematical models allow the response to be tailored to specific behavior based on its risk to the business. Suppose an analyst building quantitative analysis models for a bank accesses a system that they have never looked at before. Because there could be a valid reason and model has learnt that the role regularly requires new information sets, it might be enough to notify a manager. At the other extreme, if a system administrator suddenly downloaded four terabytes of data, and the HR system simultaneously flagged up that they had just received a poor appraisal score, the system could switch off their access pending investigation.

The solution can also detect events that are anomalous in IT, rather than human, terms. If someone who travels a lot logs in from Japan that might be fine, but if they had logged in from the UK only four hours before, then obviously the user may not be who they seem, and urgent action will be required to protect data and IP.

The advantage of using machine learning models is that they can be used to provide you with predictive analytics which alert you before any anomalous behavior occurs. The mathematical models become able to identify the trends that lead to fraud or a breach. This enables you to put in place new processes which can then react as a threat occurs to prevent it at source.

How it works

Data science, machine learning and the MADLib library are at the heart of this next generation analytics solution. The machine learning algorithms in MADLib represent the industrialization of traditionally high-end analytics into an open source library and are used to analyze the historical behavior of users and systems to understand what constitutes normal behavior; these algorithms are then weighted based on the risk of different threats to your business to create a baseline. By combining transaction, network, HR and even external data a view is created of both what represents the standard operating procedures for your business and what represents a business threat. The models can then be integrated into HR or security systems to enable you to react to threats as, or before, they occur.

The benefit of this approach is not limited to uncovering anomalous human behavior. These same algorithms can also locate and prioritize suspicious machine-to-machine activity. For instance if network traffic normally flowed between two systems and suddenly was being routed to a third or even external system. This suspicious activity could be benign, for instance a new server deployment; however, it could also indicate activity of malicious or ill intent.

All this depends on the ability to analyze huge amounts of data – both at speed and in batch modes. The Business Data Lake (BDL) makes that possible.
The role of the BDL
The BDL solves the problem of how to ingest and store all of the diverse information sets a business requires without everything fitted within a single enterprise data warehouse (EDW), something impractical when looking at diverse analytical challenges.

EDWs and security systems are good for storing structured enterprise data, but have their limitations when it comes to identifying behavioral changes. For example, organizations usually have lengthy processes for storing new types of data in the EDW – and rightly so, but it makes it hard to store additional data needed for behavioral analysis, especially unstructured data like emails, video or instant messaging.

The other challenge of the EDW is that its schema needs to be optimized for normal enterprise use and the mathematical models of anomaly analytics often require a different approach to perform effectively.

In addition, given today’s “networked business” model, a lot of the data needed for monitoring isn’t within the control of your organization but is managed by third parties – for example, social media, emails and partner data – so it doesn’t readily fit into the EDW.

A final challenge is that volumes of data are becoming huge – security log data, for instance, can be measured in petabytes. Social interaction data – which some organizations need is not only huge but also unstructured, whereas the EDW is best suited to comparatively small-scale, a few hundred terabytes, of structured data.

The BDL is a fundamentally different approach that complements EDWs and overcomes their limitations. It makes it possible to store huge volumes of data, including unstructured data, affordably. There’s no need for branded software or hardware – commodity platforms can be used.

The BDL technology extracts and distills the views of the data that you need – whether these are disposable or shared across the business. It comes with analytic tools provided at a price point that wasn’t previously possible.

The BDL is an important enabler of anomalous behavior analysis because it:

- Can handle extremely large data volumes, including historical data, fast enough to distil the view needed to judge whether a particular pattern is anomalous.
- Can bring together disparate data types: both IT data such as network logs, and enterprise data such as HR records or social media.
- Is flexible enough to support learning about behavior, including complex interactions (since anomalous behavior can involve more than one individual).
- Is affordable, putting a sophisticated approach within reach of most organizations.
- Does not require extensive up-front work to define a single enterprise schema, meaning it compresses the time to value.

Fig 2. Ingest structured and unstructured data to the Business Data Lake; create insight and take action
Benefits of the Anomalous Behavior Detection Solution

Protection from potentially catastrophic loss. This approach could save millions by preventing fraud – even more by protecting vital IP. It can counter IP leakage, theft of sensitive material, and abuse of company systems, and complement network-level detection by providing early warning of social engineering attacks. You can not only minimize exposure time and loss, but potentially predict leakage areas before an attack happens so that you can take appropriate action.

You don’t have to know what you’re looking for – for one client, we not only identified a person known to be creating false records but also a number of others apparently using the same loophole.

A fast, affordable approach. Because the Business Data Lake takes an augmentation approach, leveraging your existing EDW and related investments, we can make maximum use of any relevant technology or data that you already have, increasing your return on existing investments. You can initiate a proof of concept (POC) to test your hypotheses with minimal investment and delay. A full solution can be up and running in a few months, thanks to the pre-written algorithms and other accelerators for your industry that Capgemini and Pivotal can supply. We make maximum use of any relevant technology or data that you already have, increasing your return on existing investments.

A light touch. You no longer have to create extensive inflexible GRC rules for every system and task. Instead you focus on risks, and take the level of action that’s appropriate when an anomaly is found – which could be as simple as emailing a manager asking them to check why an employee is accessing particular data.

A general-purpose analytic resource. The BDL can be used for many types of analytics, extending the analytic capabilities of your business while complementing existing investments in security and EDW technologies. Governance can be applied where it matters, in the same way as to any other data store.

All this means that implementing anomalous behavior analytics isn’t simply another silo of information. It becomes a core part of your company’s information fabric, not only reducing risk but also enabling new insights and reports on the information it leverages.

Fig 3. Build a platform for future defense capability
Why work with Capgemini?

Capgemini have been a pioneer of this new approach of integrated information solutions. We have worked closely with the Pivotal data science team on the mathematical approaches required and our own internal data science team is collaborating with Pivotal on continually improving the approach. As a technology partner, we have deep expertise in Pivotal’s technologies and data science skills backed up by our own dedicated Pivotal infrastructures in Europe and North America. We can provide anything from a rapid proof of concept (POC) to comprehensive integration, roll out and scalability services. If you’re interested in carrying out a POC, we can provide a working application in a few weeks. We take an anonymized sample of your data and show what insights you could gain by applying these techniques.

Once we have worked with you to identify the potential benefit for your organization, we can then scale the POC to provide a solution that meets wider line of business or full enterprise analysis, depending on your needs. We can provide a strategic roadmap to move from small-scale POC to a complete enterprise managed service, with analytics delivered in both real time and from long-term analysis.
About Capgemini

With almost 140,000 people in over 40 countries, Capgemini is one of the world’s foremost providers of consulting, technology and outsourcing services. The Group reported 2013 global revenues of EUR 10.1 billion.

Together with its clients, Capgemini creates and delivers business and technology solutions that fit their needs and drive the results they want. A deeply multicultural organization, Capgemini has developed its own way of working, the Collaborative Business Experience™, and draws on Rightshore®, its worldwide delivery model.

Find out more at [www.capgemini.com/bdl](http://www.capgemini.com/bdl) and [www.pivotal.io/big-data/businessdatalake](http://www.pivotal.io/big-data/businessdatalake)

Or contact us at [bim@capgemini.com](mailto:bim@capgemini.com)