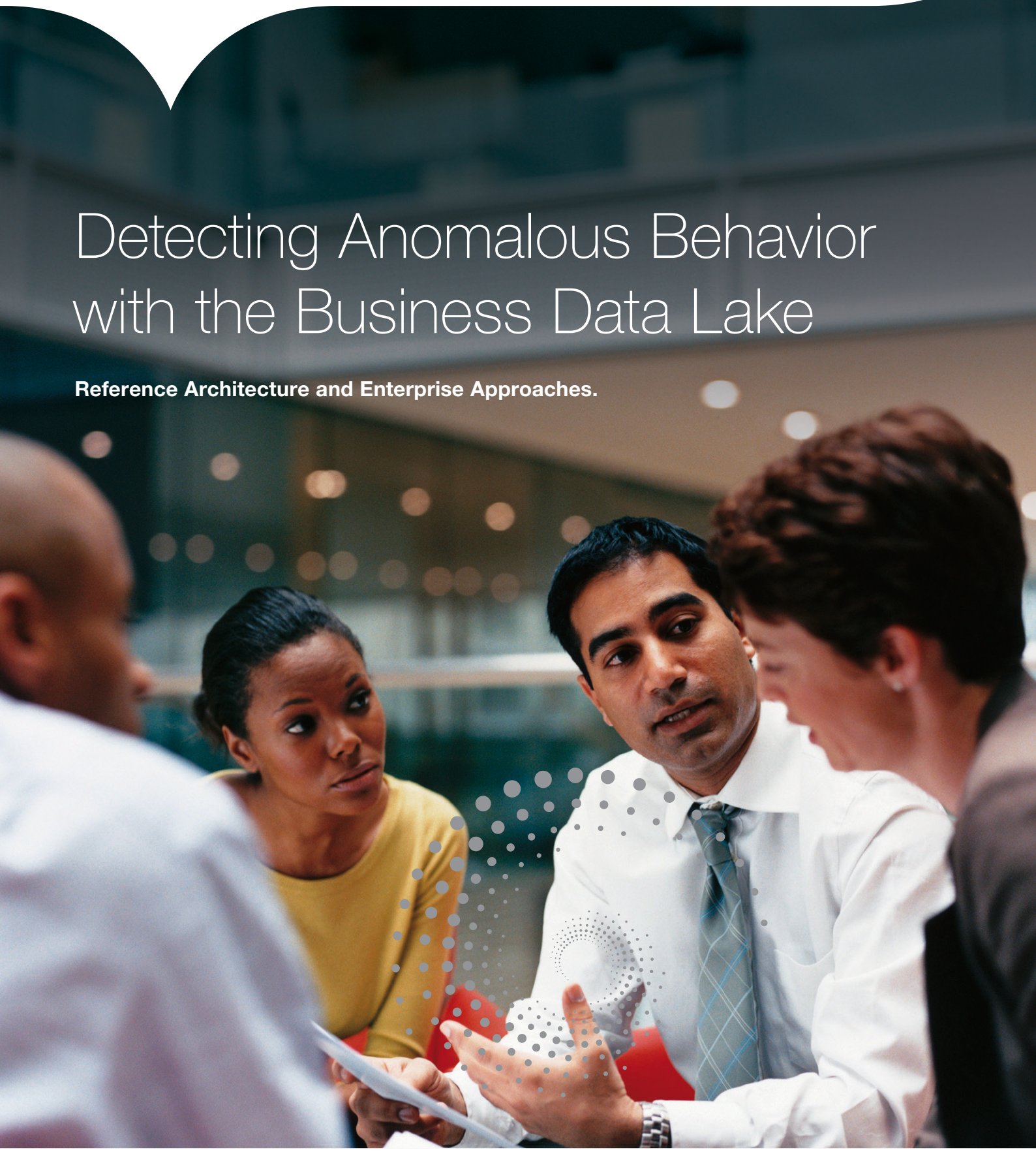
Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

Pivotal™

Detecting Anomalous Behavior with the Business Data Lake

Reference Architecture and Enterprise Approaches.



Reference Architecture and Enterprise Approaches

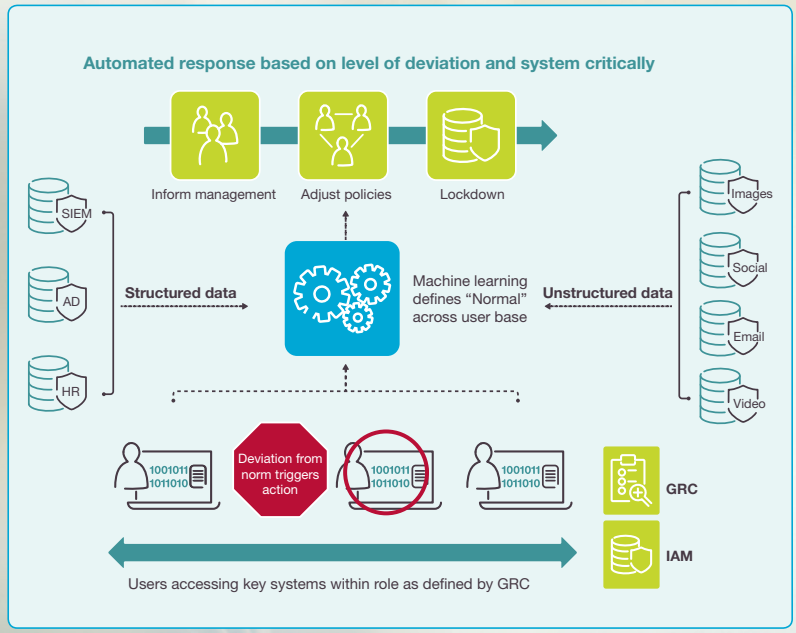
Anomalous behavior detection gives businesses an approach to identify “in-role users”, whether human or machine, who behave abnormally but within agreed security and compliance frameworks. In this paper, we will show you the common reference architecture for the ingestion, storage, analysis and action layers of an enterprise capability that provides not only multiple use case fulfilment, but also an extensible open capability for the enterprise to leverage.

It enables the platform to create the response and act based on the insights discovered.

This is because it:

- Provides more agility and a more capable response to security threats
- Provides the ability to respond to threats that were previously impossible to deal with because of time, cost or complexity issues
- Builds disposable security insights to respond to one-time business events (e.g. restructuring).

Fig 1. Builds capability that detects anomalous behavior and provides insight for investigation



Business Data Lake as Underlying Reference Architecture

THE UNDERLYING PRINCIPLE FOR BUILDING AN ANOMALOUS BEHAVIOR DETECTION CAPABILITY LIES IN THE BUSINESS DATA LAKE

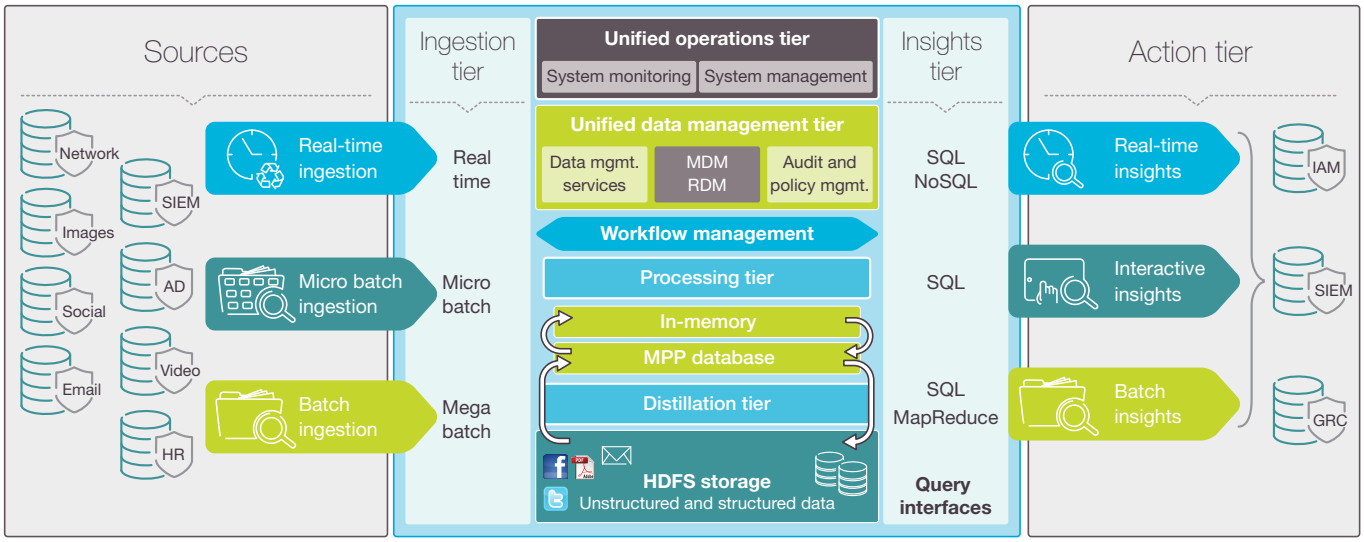
The underlying principle for building an anomalous behavior detection capability lies in the Business Data Lake. It provides the ability for the enterprise to:

- **Ingest** – Capture data from a wide variety of sources, traditional (e.g. security logs) and new (HR data, social media).
- **Store** – Store everything in one environment, with its history.
- **Analyze** – Use advanced algorithms to discover new insights.
- **Surface** – Share insights with business domain experts.
- **Act** – Take appropriate, process-driven actions based on business exposure and risk.

By taking the Business Data Lake approach, the organization can:

- Initiate an enterprise class capability with low barriers of entry.
- Ingest structured and unstructured data at scale and at sustainable cost points.
- Move analytics workloads between real time and batch to match business need.

Fig 2. Ingest structured and unstructured data to the Business Data Lake; create insight and take action



Anomalous Behavior Detection as an Abstract Application Layer

In building a capability to detect anomalous behavior, the solution acts as an application layer on top of the Business Data Lake. The application is geared to create three main types of analytics:

- **Capture time analytics** – to identify interesting characteristics of data right at the time of capture. This includes basic characteristics, interesting characteristics, and indicators of compromise. The key is to use tooling that creates metadata out of these interesting characteristics. The metadata can be used for further analytics or to facilitate investigations.
- **Streaming analytics** – to analyze metadata in real time to spot concurrent sessions or actions happening over a short time window that might indicate a threat. Streaming analytics can be based on combinations of events or deviations from a "baseline" normal count of a piece of metadata. Streaming analytics appliances need not be deployed right at the point of collection, but can be deployed in parallel throughout the environment for enhanced scalability.
- **Batch analytics** – to build patterns that represent the anomalous behavior to identify "low and slow" type attacks, and patterns that occur over extended periods of time. Batch analytics is performed using various analytic techniques including:
  - Behavioral analysis
  - Cluster analysis
  - Anomaly detection
  - Machine learning

Batch analytics, with these advanced methods, facilitates the wider set of use cases that we call "detecting anomalous behavior". This is often behavior by users operating within the security policy but acting abnormally in role.

Ingestion of Data Types and Key Considerations

In deploying Abnormal Behavior Detection, the main variances in deployment arise in four areas:

1. **Source ingestion**
2. **Implementation of the algorithms for behavioral modeling**
3. **Insight to action**
4. **Interconnects to wider enterprise systems for automated action**

Source Ingestion

Source ingestion typology is driven by the use case in question, but can be broadly spread into these areas, the storage demands for which vary depending on the needs for both near real-time and long-term analysis.

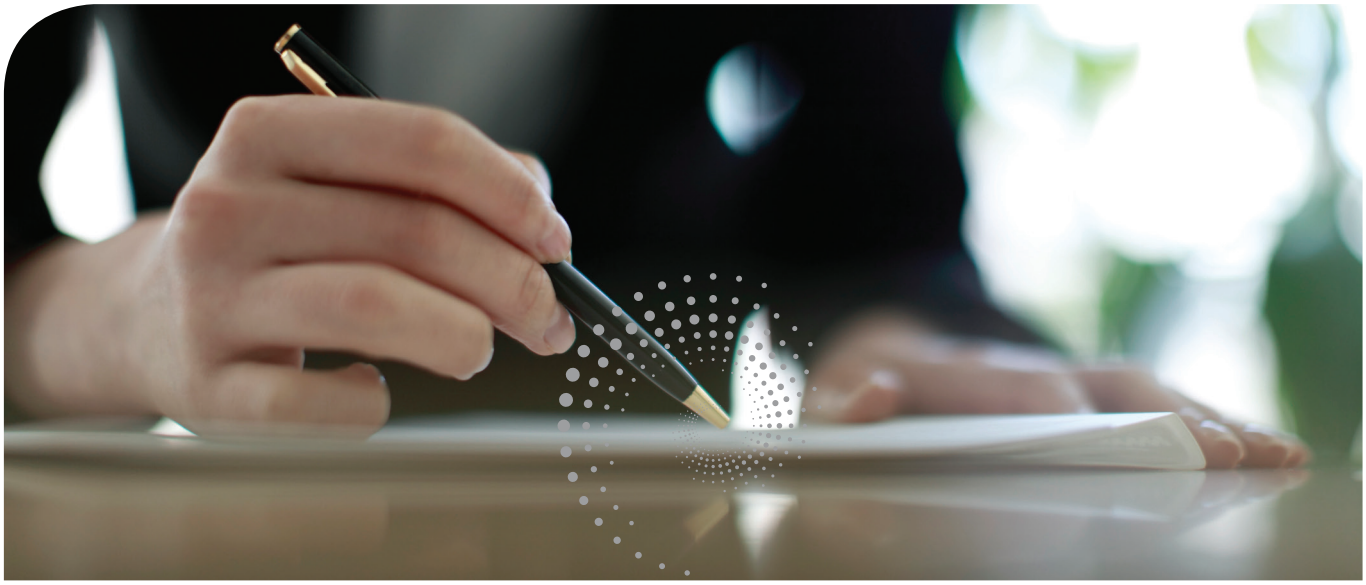
- **Network activity** – performing full packet capture to carry out session reconstruction and analysis of packet data.
- **Device and application log data** – collecting log and event data from devices and applications that support business and IT activity.
- **Security product logs** – logs and alerts from existing security vendors' products.

Contextual data

- **Asset data** – this includes the collection of technical configuration data, as well as the business context such as what business processes the system supports, or the criticality of the system.
- **Vulnerability data** – data that can add additional context to an investigation (e.g. when the system was last scanned and what vulnerabilities were present) or help prioritize response attacks on vulnerable systems.
- **Identity data** – additional contextual information about the user, their location, their job function, and the privileges they have.

Wider enterprise data

- **Structured data sets** – to add context to a user's permitted role and operating constraints – governance, risk and compliance plus identity and access management. More broadly, other valuable data sets include employee workgroup information, access rights, privileges and other HR-centric information; employee badging records; document location and security information, etc.
- **Unstructured data sets** – these can add rich context to behavioral analysis including, but not limited to, instant messaging, email, service tickets, administrators' executed commands, etc.



Algorithms for Behavioral Modeling

The central concept of detecting anomalous behavior is the use of data science to create models and algorithms that can determine the "normal" behavior for users, devices or applications and then detect variation from that behavior so that action can be taken. The algorithms and the supporting data sets needed vary significantly. Special attention should be paid to:

- Data sovereignty for the data sets being used.
- Data privacy and variance by geography.
- Legality and employee rights in active monitoring and analysis.
- Likely threat window and optimal time window for response and the compute/storage needed (cost benefit analysis) – batch vs. real-time needs.

Insight to Take Action

Generating insight is important – surfacing that insight for action, either automated or manual, is where the enterprise can both automate and create the right level of response. Consider:

- Creating a response matrix based on threat risk vs operational needs.
- Overheads to security investigation teams when tuning the algorithms for potential fraud.
- Taking interim automated action based on risk and potential impact of detected behavior. This may be anything from an automated email to the employee's line manager flagging up exceptional behavior through to automated removal of access rights.

Integrating Action to Enterprise Systems

The Business Data Lake approach provides you with action tiers – real time, interactive and batch – that can now be connected to wider enterprise systems to enhance the outcomes from the discovery of anomalous behavior. Connectors can be built to:

- **Security information and event management (SIEM)**
  - Bidirectional information share with SIEM for security incident management.
- **Governance, Risk Management, and Compliance (GRC) platforms**
  - Inbound GRC rules sharing with the analytics platform.
  - Outbound updates to GRC to inform on incidents and provide wider organizational impact analysis.
  - Visibility of changes to risk based on breaches.
- **Identity and Access Management**
  - Automated changes to permissions for employees: For example, reduction in permissions to selected applications, suspension of remote or mobile access.
- **Automated quarantine of resources**
  - Ability via management tooling (and via SIEM) to quarantine suspect servers, devices or virtual machines during investigations.
- **Physical security**
  - Enable links to site and pass card based security to control physical entry/exit.

Proving Out Use Cases for the Enterprise

The real success factor is obtaining proof of the data science model.

The architecture we have scales from a small proof of concept to a full line-of-business requirement and through to a pan-enterprise service. You can prove out use cases from “laptop scale” to “enterprise scale” with a common architecture and low barriers to entry.

Many proof points can be easily tested by the enterprise – based both on a hypothesis about the current business (unproven fraud for example) and previous fraud (where bad actors were identified).

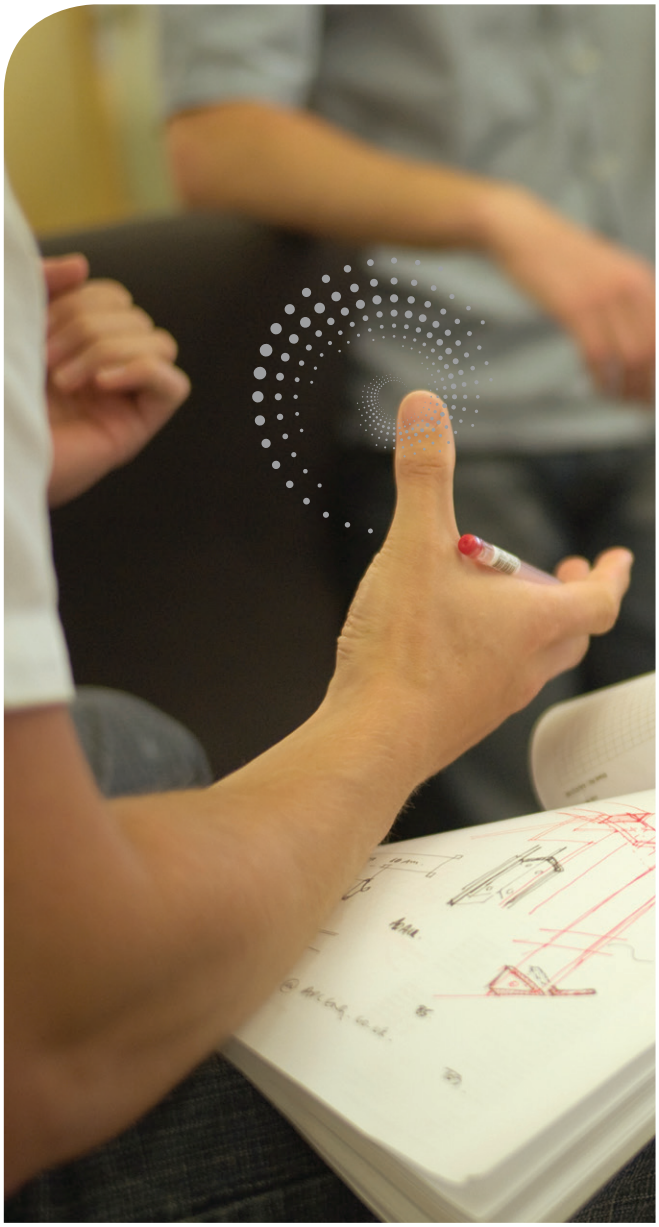
Proofs of concept do not, in our experience, require extensive investment, and clients can often make key discoveries through cost-effective environments that can deliver the insight on a batch or multi-hour basis – real time is not required. The real success factor is obtaining proof of the data science model.

We find that the following can help accelerate the testing of the model within proofs of concept:

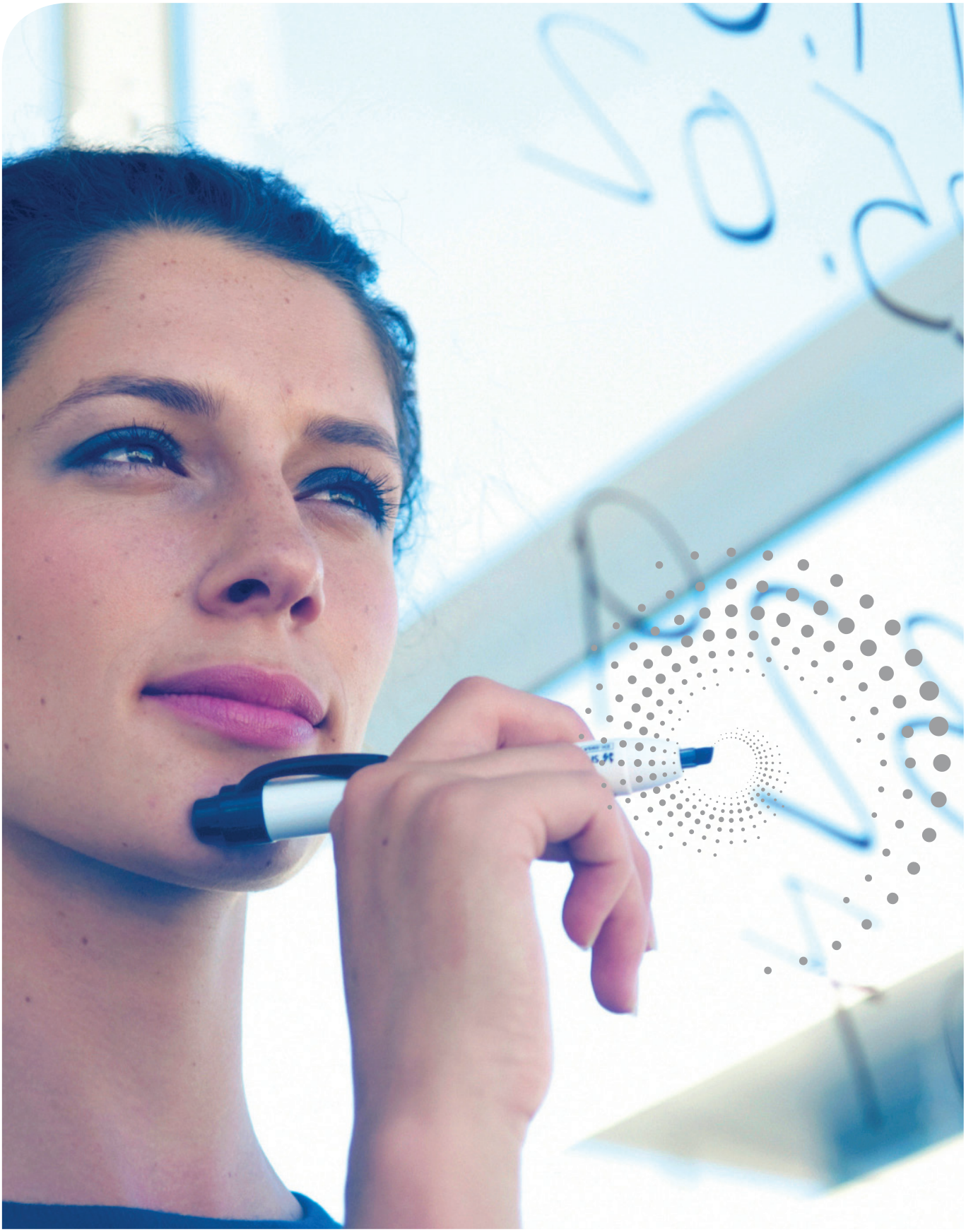
- Tightly defined use cases, ideally from live or proven cases.
- Re-proving cases where the anomalous behavior can be found “blind”, without knowledge of who the bad actors are amidst anonymous data.
- Use of cloud or temporary infrastructures to trial the solution with a low cost of entry.

Delivering the Architecture

In scaling out a wider enterprise service to detect and react to threats within multiple lines of business or multiple analytics in parallel, enterprises must ingest data from a variety of sources. The Business Data Lake architecture addresses these through the use of leading big data technologies, including Hadoop and in-memory and parallel computing platforms. Supporting elements of the approach include:



- Ingestion and use of non-traditional data sets to augment the detection of abnormal behavior – HR performance database, video, instant messaging, email, social networks, depending on the availability.
- Where network data needs to be collated, collection and capture-time analytics get deployed close to where the activity occurs. This allows the system to scale across locations more effectively. It also minimizes the impact on WAN connections, since the system can be configured to transfer only metadata, not raw data, across these connections.
- An architecture that scales to support the business as use case and demand grows and evolves, with scalable infrastructure and predictable cost metrics:
  - Careful understanding of total cost of ownership (TCO) impacts of cloud-based infrastructure vs dedicated on-premise/off-premise infrastructure.
  - Selection of analytics platforms that give you the most flexibility as the business adapts its demands from batch to real-time outcomes – and vice versa without cumbersome license models or penalties for moving between usage types.
- Enabling by creating a common service center and capability to:
  - Provide scalability to wider business users and use cases.
  - Maximize re-use of algorithms across multiple divisions and avoid silos of use case implementation.



Analytics Roadmap and Wider Business Value

Building a point capability for anomalous behavior detection is not about solving one use case or a narrow set of security issues. By taking a Business Data Lake approach, you can build wider enterprise capabilities:

1. Enhancing the anomalous behavior detection to address wider use cases; you can run these from the same architecture, and just need to add new processing engines and operational interfaces.
2. By building from a foundation of the Business Data Lake – complemented with wider technologies, layered applications, and services – you are able to provide a platform for future defense capability:
    a. Implementing advanced machine learning to improve rules base and detection thresholds. With machine learning, the system can also discover which previously unnoticed parameters are useful for detecting an attack.
    b. Anticipate attacks – The prioritization system sets thresholds for priority risks and triggers immediate investigation by incident response teams. The system provides alerts if it detects behavior that could indicate an attack is being planned, such as multiple failed authentication attempts in quick succession.
    c. Enhance through federated sharing of threats – You can develop built-in mechanisms to automatically exchange threat intelligence data with other organizations (for example, concerted external lateral movement attacks from a common threat targeting the industry sector). There are refined masking capabilities and high levels of trust in the information-sharing network.
    d. Automated quarantine of resources – The system automatically moves the attacked resource to a different physical location. At the new location, the system instantiates controls around the resource specific to the type of attack underway.
3. Deliver wider enterprise value. By taking a common architectural approach that leverages the Business Data Lake, it is possible to create wide-scale capability for the enterprise, including not just anomalous behavior detection but also much wider business value that can create valuable analytics insights. This is because data to answer wider business opportunities and challenges can be made available for shared use. The organization can gain greater value from the data collected to develop applications and analytics for adjacent use cases such as:
    a. Customer buying patterns, augmenting social media, web, mobile and contact center information to drive increased revenue
    b. IT and business capacity planning
    c. IT downtime impact analysis
    d. Shadow IT detection
    e. Employee attrition prediction

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

Pivotal™

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**
and **www.pivotal.io/big-data/businessdatalake**
Or
contact us at **bim@capgemini.com**

MCOS_GI_AP_20140905