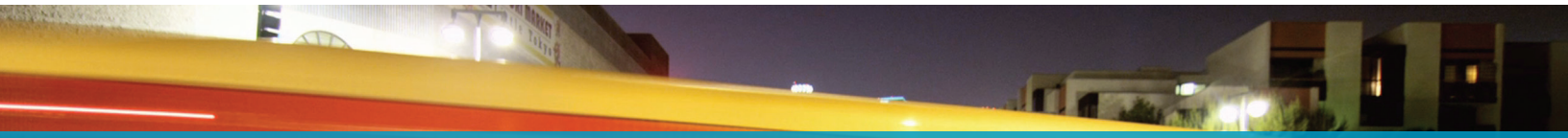
proofpoint



# CIO Series:
# How Big Data Can Help Enterprises Build Better Security Defenses

Proofpoint CIO Series – 1

# Contents

# Introduction

New database technologies are enabling enterprises to collect, store, and analyze large volumes of data in significantly less time. Collectively, these technologies have earned the nickname 'Big Data' and have found application and usage in a variety of diverse consumer, internet, medical, research, and enterprise applications. In this paper, we will explore some of the basic aspects of Big Data and discuss how CIOs and CISOs can apply these technologies to address the next generation of security challenges enterprises face today.

# Big Data – The Basics:

Big Data technologies radically transform three aspects of data management: Volume, Velocity, and Variety.[1] Each aspect provides a unique benefit, leveraged individually or in concert with another, to solve specific problems effectively and in a manner not replicated by any other available technology. We will explore each of these in brief below.

## Volume

About 2.5 exabytes, in other words 2.5 billion gigabytes, of data are created worldwide each day. As impressive as that number is, it's expected to double in 40 months.[2]

Organizations find themselves holding a growing pool of data, but need to leverage Big Data techniques to process and unlock knowledge held within large volumes of data. As one of the largest retailers in the world, Walmart's stores and online properties generate about 1 million transactions every hour, generating data that is stored in databases that have grown to 2.5 petabytes.[3] To maintain operational excellence and competitive pricing, Walmart must leverage this data to predict demand, understand customer preferences, and determine inventory levels. Walmart employs Big Data technologies that are capable of providing the 'Volume' benefits to solve their data analysis needs and gain real–time insight into operations.

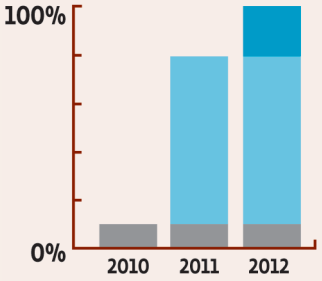
Scientific research applications also take advantage of the 'volume' aspect of Big Data technology. It took the Human Genome Project billions of dollars and over a decade of analysis to decode the first genome. Big Data along with other adjunct technologies have accelerated this analysis process a million–fold due to their capability of working with larger quantities of data. Decades have been reduced to days, and costs for decoding a genome have dropped to a few thousand dollars.[4]

Solutions such as the MongoDB data store make use of more flexible data models than traditional approaches and are capable of managing vast amounts of data, while also providing the accuracy that business–critical services require.

## Velocity

'Velocity' refers to the speed at which enterprises collect data and query it for answers. One of the most remarkable aspects of the Big Data revolution is that even though data volumes are growing exponentially, query response times are getting faster. Even in data stores that span multiple terabytes, it is sometimes possible to get answers with split–second responsiveness.

**EVERYDAY BUSINESS AND CONSUMER LIFE CREATES 2.5 BILLION GIGABYTES OF DATA PER DAY**




**90% of the data in the world today has been created in the last two years alone.**

Source IBM

[1] http://blogs.gartner.com/doug–laney/files/2012/01/ad949–3D–Data–Management–Controlling–Data–Volume–Velocity–and–Variety.pdf
[2] McAfee and Brynjolfsson, "Big Data: The Management Revolution," Harvard Business Review, October 2012.
[3] "Data, data everywhere," The Economist, http://www.economist.com/node/15557443
[4] http://www.forbes.com/sites/sap/2012/04/16/how–cloud–and–big–data–are–impacting–the–human–genome–touching–7–billion–lives/

The velocity of analysis is critical, because amassing high volumes of data would simply be a burden if enterprises could not analyze that data in time to effectively act on it. For Walmart to be able to analyze its rapidly growing 2.5 petabytes of data quickly enough to make strategic business decisions at stores around the world is impressive. High volume, high velocity business intelligence would be impossible if Walmart were to rely on traditional RDBMS and disk technology. Simply reading a terabyte of data from a disk takes about 2.5 hours.[5] To read 2,500 terabytes, even with RAID optimizations, would take many days—and by then the backload of shopping hour data to be analyzed would have grown exponentially. But with Big Data techniques, Walmart can analyze all these transactions in near real time and quickly make decisions about inventory levels, pricing, and other factors that directly affect its bottom line.

High velocity analysis of vast data stores will continue to benefit other industries, as well. PASSUR Aerospace uses Big Data to collect and analyze weather conditions, flight patterns of airline jets, and past take-off and arrival patterns for airports all over the U.S. By taking data 'snapshots' of air traffic and weather conditions every 4.6 seconds and delivering intelligence updates to subscribers in near real time, PASSUR is able to substantially improve the accuracy of timetables used by ground crews responsible for incoming flights. PASSUR estimates that this added efficiency is saving major airports millions of dollars each year.[6]

Big Data solutions such as Apache Cassandra and Apache Hadoop enable enterprises to load and query data exponentially faster than traditional data technologies. Cassandra was originally developed by Facebook to provide millions of users with near real-time access to their mailboxes. Hadoop was originally developed as a part of a search engine project for Yahoo!. Users of social networks and search engines expect rapid performance, and thanks to Big Data they are able to get it.

## Variety

Not only is there more data, there are also more types of data. Storage of structured data, such as customer records or financial transactions that in smaller quantities might be recorded in an RDBMS, is growing exponentially. Enterprises are also collecting and analyzing unstructured data, including new data types, for example GPS data from mobile devices and image data from digital cameras. Many enterprises are also collecting high volumes of sensor data—everything from weather monitoring systems for logistics to motion controls in factory automation systems. Given this trend, it is estimated by IDC that by 2020 most enterprises will find themselves managing 75 times more sensor data than they are today.[7]

More traditional businesses like retailers also find themselves juggling disparate data types. The most successful ecommerce platforms combine traditional structured data such as product SKUs, product descriptions, and prices with more diverse types of data such as photographs, demo videos, customer comments, ratings, and instructional guides formatted as PDF files. Retrieving these different types of data fast enough to satisfy impatient Internet shoppers would be difficult with traditional database queries, but it is possible with Big Data.

**2.9 million emails** sent every second.

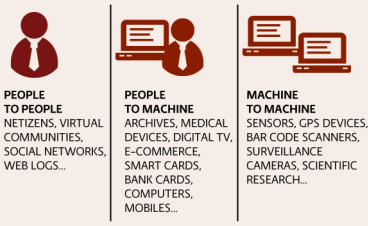
**30 billion pieces of content** were added to Facebook this past month by 600 million plus users.

**Zynga processes 1 petabyte of content** for players every day: a volume of data that is unmatched in the social game industry.

**More than 2 billion videos** were watched on YouTube... yesterday.

The average teenager sends **4,762 text messages** per month.

**32 billion searches** were performed last month... on Twitter.




Source: Gartner, Wipro

[5] Tom White, Hadoop: The Definitive Guide, O'Reilly Media/Yahoo Press
[6] McAfee and Brynjolfsson, "Big Data: The Management Revolution," Harvard Business Review, October 2012
[7] http://www.zdnet.com/data-volume-to-hit-1-8zb-in-2011-2062301103/

To accommodate a broader range of data types, including unstructured data, many enterprises are complementing traditional SQL–based data storage architectures with more flexible data storage systems. So–called NOSQL (Not Only SQL) technologies such as Apache CouchDB and MongoDB forego traditional database schemas and data access methodologies in favor of looser, 'eventually consistent' database architectures that offer greater speed and flexibility.

## The Age of Targeted Attacks, Social Engineering, and Data Theft

While enterprises are starting to adopt such Big Data technologies and feel bullish about their ability to analyze more data than ever before, most IT and InfoSec teams are feeling much less confident about having the necessary tools, skill sets, and processes to protect that data from hackers, competitors, criminal syndicates, and prying state actors.

Inbound security attacks have become more sophisticated. A decade ago, inbound security attacks were like acts of vandalism—attacks were brash, high–volume, and much more easily detectable through signature–based approaches. Attacks have evolved to be much more criminal in nature, and are now designed to be narrowly focused and stealthy. The motivation is to steal data—as much data as possible for as long as possible. Instead of overloading servers to cause denial of service, new attacks leverage techniques that aim to sneak in malicious software onto corporate computer systems and send back valuable data that attackers can monetize. These targeted attacks quietly exfiltrate data over hours, days, months, or even years.[8] Eighty–five percent of data breaches in 2011 were discovered only after many weeks had passed.[9] Because these attacks are so long–lived, they are also commonly referred to as 'Advanced Persistent Threats' or 'APTs'.

How effective have the attacks been? Over two–thirds of large enterprises report having been subject to attacks, but the actual number of victims is probably much higher. Security expert and former counterterrorism czar Richard Clarke believes that every major company in the United States has been penetrated by targeted attacks.[10] Data stolen includes customer records, source code, advanced design algorithms and other intellectual property, credentials to financial accounts and other protected systems, HR records, and more.

**29%** OF BREACHES USED SOCIAL ENGINEERING

**79%** OF THOSE BREACHES PHISHING WAS USED AS THE ATTACK VECTOR:

**95%** OF ESPIONAGE ATTACKS INVOLVE PHISHING

Source: Verizon Threat Report, 2013

Targeted attacks typically take the form of blended threats combining a highly targeted phishing email message (sometimes called 'spear–phishing'), and a spoofed or compromised website that has been set up to deliver malware to a user's machine. The phishing messages are carefully crafted. They often include personal information gleaned from social networks and public websites such as Facebook, LinkedIn or personal blogs, or they use terms likely to be of interest to their recipients. For example, a spoofed email message to a company vice president with many job openings in her department might pretend to link to an article about hiring practices, or even a fake candidate trying to reach out to the hiring manager with a link to their online resume and industry citation.

The goal of each message is to be believable enough as to lure the recipient into trusting the message, and eventually clicking on an embedded link in the email message which would

[8] http://krebsonsecurity.com/2011/10/who–else–was–hit–by–the–rsa–attackers/
[9] Verizon's 2012 Data Breach Investigations Report
[10] http://www.smithsonianmag.com/history–archaeology/Richard–Clarke–on–Who–Was–Behind–the–Stuxnet–Attack.html

lead the user to a malicious website that compromises the user's machine or tricks them into giving up their credentials. If the attackers are crafty and want to avoid immediate detection by the users and security monitoring solutions, the destination website might be safe most of the time and 'pulsed' with malware only intermittently. Scanning the website for malware when the phishing message arrives might not raise any alarms, because the website might not be dangerous at the moment when the email arrives. But eventually one or more of the message recipients will return to the site at some later time and encounter its malware, and the attacker's trap will be sprung.

Those traps are working. In 2012, Verizon estimates that the industry saw 621 reported breaches, in which 29% of them used some form of social engineering, with 79% of those leveraging phishing as the vector for attack.[12] Buoyed by these successes, attackers are getting bolder and attacking more frequently. These attacks are targeting enterprises across industries, as well as government agencies at the federal, state, and local levels.[13]

**4X** MORE **PHISHING ATTACKS IN 2013** COMPARED TO THE 2012 REPORT

Source: Verizon Threat Report, 2013

Traditional IT security solutions, such as firewalls, web and email gateways, and other network perimeter defenses, do a poor job of defending against the more sophisticated targeted attacks. Why? Targeted attacks don't look like the traditional mass, malware-laden attacks these defense systems were designed to detect using known reputation analysis and signature matching. The new attacks arrive in the form of very few spoofed messages from an innocuous sender, or even worse from legitimate email accounts that have been compromised, rather than in torrents of spam messages spawned by botnets. This alone defeats perimeter security that relies on 'reputation' as the IP reputation inherited by the attackers during the spoof or use of compromised accounts is typically good with no history of reported bad behavior. Because the messages do not contain malware themselves, they can slip past antivirus engines and filters in firewalls.

Attackers take advantage of the behavioral patterns of email recipients, timing email delivery to maximize the chances of users clicking on links, preferably when users are checking email from home or another remote location off the corporate network (and hence deprived of enterprise security solutions like firewalls and gateways that may be in place). For example, sending phishing email on a late Friday evening or early Monday morning is more likely to catch an employee at home, and unfortunately has proven to be an effective tactic for avoiding network perimeter malware detection systems.

**12 HRS**

AMOUNT OF TIME THAT ELAPSES BETWEEN SENDING A **PHISHING CAMPAIGN & RECEIVING 1/2 THE TOTAL TARGET CLICKS**

Source: Verizon Threat Report, 2013

Are enterprises defenseless against sophisticated and targeted attacks like these? Fortunately, Big Data technologies we discussed earlier in this paper could not have come at a better time to help enterprises combat these attacks.

## Big Data to the Rescue

The advancements in Big Data is timely not just for scientific research and business intelligence, but also for enterprise data security. Big Data analytics applied towards volumes of enterprise email behavior data offer IT departments the most effective defense against targeted attacks.

[11] http://threatpost.com/en_us/blogs/rsa-phishing-attacks-net-687m-date-2012-082412
[12] http://www.verizonenterprise.com/DBIR/2013/
[13] According to the U.S. Computer Emergency Readiness Team, 51.2 percent of reported attacks on federal, state and local government agencies in 2011 involved phishing. http://gcn.com/articles/2012/09/26/20-most-common-words-phishing-attacks.aspx

Cutting–edge IT defense systems leverage dynamic behavioral models that are created and maintained using Big Data techniques. These behavioral models are built around observing normal mail flow characteristics for every employee mailbox, and then used to analyze inbound email in real–time against what is considered 'normal' mail flow. This approach can detect the subtle characteristics of targeted phishing ('spear–phishing') attacks, increasing their suspiciousness and calling them out as 'anomalies' in mail flow that are worth monitoring. Examples of these characteristics include:

- **Sender's domain.** Has anyone in the company ever sent email to this domain before? Has the company ever received email from this domain before? How recently was the domain registered? Is it registered with a registrar known to be used by hackers or spammers? Has the particular recipient ever sent or received email from the sender?
- **Message contents.** Are the contents of the message unusual? Is there a suspicious attachment? Does the message include a link to a site that was recently created? Does the site contain potentially malicious behavior–ware when the user clicks on a link to visit it?
- **Message delivery.** Was the message sent at an unusual time (e.g., at 3 a.m. even though its contents discuss business issues?) Is there anything else anomalous about the delivery of the message or its header contents?

Of course, analyzing millions of messages across a company's user base in real time is beyond the capabilities of traditional IT defense systems and RDBMS products. But analysis on this grand scale is possible with Big Data technologies that provide the benefits discussed earlier associated with Volume, Velocity and Variety of data.

It is now possible to scan terabytes of inbound email for various specific attributes and relationships to detect anomalies that suggest the presence of a targeted attack. Suspicious messages can then be quarantined or deleted before a user clicks on a dangerous link and activates the attack payload. Through continuous, comprehensive scrutiny, Big Data analytics give enterprises their best opportunity to defend against targeted attacks.

Enterprises should maintain their perimeter defenses and signature–based solutions which serve the necessary function of stopping traditional attacks like promotional spam, high–volume phishing, and common malicious attachments. To protect against targeted attacks and longlining attacks however, CIOs and CISOs should consider complementing the enterprise's traditional malware defenses with new solutions based on Big Data technologies.

## About Proofpoint

Proofpoint, Inc. (NASDAQ: PFPT) is a leading security–as–a–service provider that focuses on cloud–based solutions for threat protection, compliance, archiving & governance and secure communications. Organizations around the world depend on Proofpoint's expertise, patented technologies and on–demand delivery system to protect against phishing, malware and spam, safeguard privacy, encrypt sensitive information, and archive and govern messages and critical enterprise information.

# About Proofpoint Targeted Attack Protection

Proofpoint Targeted Attack Protection is a cloud-based security-as-a-service offering from Proofpoint that leverages Big Data infrastructure and applies advanced analytics techniques to detect, respond to, and manage new forms of attack including highly targeted and socially-engineered phishing attacks.

## Targeted Attack Protection includes:

- the Proofpoint Anomalytics Service, which examines hundreds of variables in real time—including email message characteristics, but also the email traffic history of the message recipient—to identify anomalies at a per-user level that could indicate a potential threat.
- the Proofpoint URL Clicktime Defense Service which rewrites URLs in suspicious email as instructed by the Anomalytics Service, and examines URLs both at delivery and at click-time to ensure that the Web destinations are not malicious. The URL Clicktime Defense Service helps protect employees and their devices whether within or beyond the enterprise network perimeter, and provides the ability to protect users from successfully delivered email in which the destination Web site for URLs may turn malicious at some future point after delivery.
- the Proofpoint Malware Analysis Service which uses a combination of IP analysis, page security analysis, and malware analysis sandboxing—performed entirely in the cloud—to ensure that each time a user clicks a link, the resulting URL payload is inspected, regardless of user location, and before malware has the opportunity to take effect. The Malware Analysis Service uses its multi-pronged approach to be effective against advanced threats and malware that are designed to bypass traditional signature-based defenses. It is effective against web-based malware attacks that are delivered using documents, Java, Javascript and Flash exploits, cross-site scripting, DOM attacks, malicious images, and compressed archives like ZIP and RAR, etc.
- the Proofpoint Threat Insight Service provides a graphical, web-based threat analysis dashboard using Big Data techniques to manage, analyze and present the relevant data. This that enables IT staff to answer crucial questions in real-time, like who received these targeted emails, when they received it, and who has clicked on the malicious links. This enables incident response teams to take efficient action, and spend their time on the events that matter.

By applying Big Data analytics and advanced machine-learning techniques, Proofpoint Targeted Attack Protection provides 24/7 defense against targeted attacks.

**To learn more about Longlining Attacks, download** Longline Phishing: A New Class of Advanced Phishing Attacks **White Paper.**

**Download** CIO Series: Why Sandboxing is a Necessary but Insufficient Defense against Targeted Attacks.

**For information about Proofpoint Targeted Attack Protection**, please visit www.proofpoint.com/tap or call +1 (877) 634-7660.

proofpoint™

Proofpoint, Inc.
892 Ross Drive, Sunnyvale, CA 94089
Tel: +1 408 517 4710
www.proofpoint.com