CLOUDPHYSICS

Bringing the Power of Big Data Analytics to IT Operations

Big Data analytics is the process of examining large amounts of data of a variety of types to uncover hidden patterns, unknown correlations and other useful information. Such information can provide competitive advantages over rival organizations and result in business benefits, such as more effective products and operations, better marketing and increased revenue.

The use of Big Data analytics has become commonplace across almost every discipline: sales, marketing, ecommerce, social media, mobile apps and more. In every case:

- Massive amounts of structured and unstructured data and metadata are captured and analyzed to discover patterns and correlations
- Algorithms are developed to predict trends and behaviors
- Managers translate that information into better decisions and improved performance.

We're most familiar with consumer-based applications of Big Data, with Amazon.com being the canonical example of using customer behavior as captured through transactions and use of the website to develop new products and services like AmazonPrime. Online ad targeting pioneered by the likes of DoubleClick, is another classic example in which billions of user-related events, like clickstreams, are analyzed to create profiles that drive the selection and display of targeted ads.

But there is one domain that has yet to fully embrace Big Data: IT. Big Data is still relatively foreign to IT operations. Big irony? Yes, especially if we consider the volume of data generated by technology. For example, a medium-sized data center with 50 hosts and 500 virtual machines (VMs), throws off more data by the minute than all the Twitterers tweeting in the world.

Running IT Operations is a Big Data problem.

Does IT really need Big Data? After all, there is no shortage of management tools out there to help us manage and optimize our environments. Unfortunately, those tools were never designed to handle the complexity and constant change inherent in modern infrastructure.

Visibility has gone down

Virtualization and software defined everything have allowed us to achieve unprecedented consolidation of applications onto hardware and given us freedom from that hardware by providing containers to wrap apps and services. But with that abstraction, we've also lost visibility into the behavior and performance of systems. It's harder to tell what particular aspect of what specific system at what specific layer of abstraction—or what combination of them—is having an effect on the application we're trying to understand or fix right now.

"Smart leaders across industries will see using Big Data for what it is: a management revolution."

Harvard Business Review, 2012

Complexity has gone up

The web of connections and dependencies we have to deal with now is too big and complex for any admin, engineer or architect keep in their head. There are multiple layers of systems that have to be understood: from configuration information to performance counters to event logs and more; from physical hosts to virtual machines to virtual disks to physical storage and more. All of which are emitting multiple streams of data that have to be collected, understood and analyzed to get a handle on what's going on at any point in time. It's not enough to understand each stream of data independently. We have to correlate all that data, mash it up and and figure out how thing interrelate.

Silos amplify the pain

We have virtualization management tools, physical server management tools, storage management tools, network management tools, log management tools, service management tools, ad infinitum. Some of these tools work at the virtual layers, some work at physical layers and a few work across layers, but only for one kind of IT silo. So the data coming from all our systems gets split across all these tools which usually don't interact and don't have a way of dealing with cross-systems data. In order to work together, we have to bring our tools and data with us to the table, translate everything into something that makes sense, then put that data together and try to draw out correlations using spreadsheets and whiteboards. We have to hope we get it right.

This adds up to fundamental uncertainty.
Less visibility + more complexity + silos = MORE PAIN.

We don't know what affects what and precisely how in any given circumstance. Landmines look like fruit baskets. We can react only to those things we can see and know are problems. We can't proactively seek the earlier indications of weaknesses and minor failures that will lead to big future failures because we don't know the full web of causation and interdependencies. Most of our best practices are more like best guesses.

Where are the tools that fix this?

The tools that give us deep visibility only do so for one silo, just one of many moving parts. The tools that give us visibility into more than one moving part are shallow and present the data in silos. The tools that collect data from everything do no correlation or analysis across systems. Every tool limits what data you can mine, what kinds of questions you can ask and what kinds of analysis you can do. They impose models on us instead of letting us decide what is and isn't important.

Solving simple problems takes hours. Solving slightly complicated problems takes days, manual effort, and a few tools. Solving really complex problems takes weeks or months, teams of people, and dozens of tools along with spreadsheet acrobatics to correlate the information coming in from everything and everyone.

Most of our best practices are more like best guesses.

Using Big Data to solve IT problems.

In IT, our Big Data is machine data, the metadata coming out of the physical and virtual systems that make up our infrastructures. This includes every last bit of operational data–things like events, tasks, configuration information and performance counters–coming out of servers, storage, networking, hypervisors, operating systems, middleware and applications. And it's BIG.

By collecting that machine metadata and mashing it up it's possible to build a living picture of what's happening. When we do that, we can:

- **Create visibility.** Look deep into what's going on at all layers of the infrastructure, how resources are being used, where and over time. Put eyes and ears where we don't have any to get the information we can't or that is hard for us to get. Put the metadata in one place and in context. Let us ask questions of the metadata, thus asking questions of our systems.
- **Make complexity manageable.** Correlate and build context automatically. Find the needles in the haystack, or the right needle in a stack of a needles. Tell us what we don't know, translate what we see now into what will happen later.
- **Cross silos of data and knowledge.** Pull together metadata from different systems, different clusters, different data centers, different clouds, different organizations. Understand not only what's going on, but how best to respond to it, backed by ever more metadata about cause and effect across ever more systems.

Without this type of Big Data analytics, if we're trying to solve a performance problem, we will look at: monitoring tools, logs, CPU, paging, network I/O contention in the VM and in the hardware, storage IO/contention in the VM and in the hardware, cache hits/misses/evictions, and on and on. This might involve one person or a dozen people. It could take hours of staring at graphs or days of looking at the problem, trying to isolate the data, collecting that data, dumping everything into spreadsheets and doing manual analysis to hunt down the needle in the haystack that's mucking up the works.

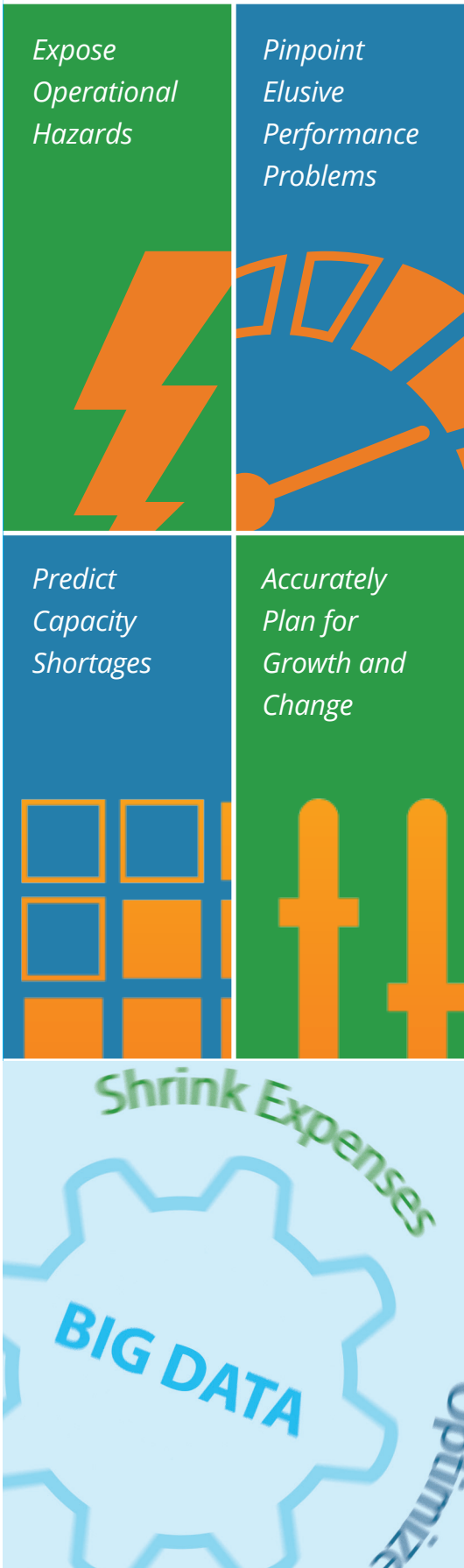
With Big Data analytics, we just look at whatever is experiencing the problem and automatically rule out what isn't related while also ruling in every possible culprit--because it is multi-layer, cross-silo and context-aware. All of the work we had to do before can be compressed into a few queries which can be made with a few clicks. We can reduces hours, days and weeks to minutes.

CloudPhysics brings Big Data analytics to virtualized infrastructure.

The CloudPhysics team, comprised of virtualization domain experts and data scientists, delivers a powerful analytics platform that lies at the heart of our solution to:

- Collect cross-silo and multi-layered metadata, putting it in one place to let you ask questions across all systems.
- Correlate the data automatically, building context as you go, pointing out things of interest, faults and causes, and suggesting solutions.
- Take patterns, analysis, and knowledge from thousands of data centers to help each individual organization better manage its own.

WHITE PAPER:
Bringing the Power of Big Data Analytics to IT Operations



CloudPhysics continuously collects systems metadata from VMware virtual infrastructures using a lightweight, agentless virtual appliance that accesses a read-only API to grab configuration, events and performance data from vCenter. It does this at a level of granularity unknown in the industry, particularly for performance data. The data isn't rolled up or summarized after an hour, a day, a week or ever. So you're always working with the best information possible--even when it's about last week.

Then it normalizes all that metadata into a single, flexible data model that shows how everything is connected and interrelated. This is key to keeping context and enables drilling down into what's going on from any starting point, while automatically ruling out what doesn't matter and ruling in what does as you go. Metadata from any system, virtual or physical, is automatically correlated to all the related metadata from related systems.

Finally, CloudPhysics makes the analytics available to customers for day to day use, helping them proactively address hidden hazards, unpredictable performance, and operational waste. As a result, IT teams can avoid service disruptions, improve overall datacenter health, and reduce capital and operating expense.

About CloudPhysics

CloudPhysics provides data-driven insights for smarter IT, giving IT teams more power than ever before to understand, troubleshoot and optimize their virtualized datacenters and drive better operational decision making. The company, based in Mountain View, Calif., serves thousands of users worldwide. For more information, visit: http://www.cloudphysics.com