R.E.A.L.
BIG DATA
FOR IT OPERATIONS

A Solution Buying Guide

rocana

CONTENTS

THERE'S A LOT OF NOISE IN THE INDUSTRY ABOUT BIG DATA AND BIG DATA ANALYTICS

Vendor claims are all over the map. Not every provider that calls itself a Big Data expert actually is.

A lot of the companies vying for your business are really offering the same old technology with a shiny new label. Legacy technologies weren't purpose-built to handle the challenges of Big Data in complex environments.

Environments that include cloud computing, virtual machines, mobile devices, and other advanced technologies require a solution especially designed to handle *real* Big Data.

TWO KEY QUESTIONS

Wading through vendor claims is confusing. Make the first cut by asking if a solution is:

- An investment that will deliver a real competitive advantage?
- More than just legacy technology that is being rebranded?

MAKING SENSE OF THE OPTIONS

There are lots of moving parts to monitoring modern IT infrastructures. Legacy IT solutions focus on issues within specific parts of an application, or only give visibility to limited amounts of system information.

You have to go beyond that.

Your solution needs to cross operational boundaries to give you answers that the 'search first' approach of legacy log management tools simply can't deliver. Determining whether a solution will be able to cross those boundaries can be as hard as managing your IT operations.

If only there were a clear set of criteria you could use to find the champions among the contenders.

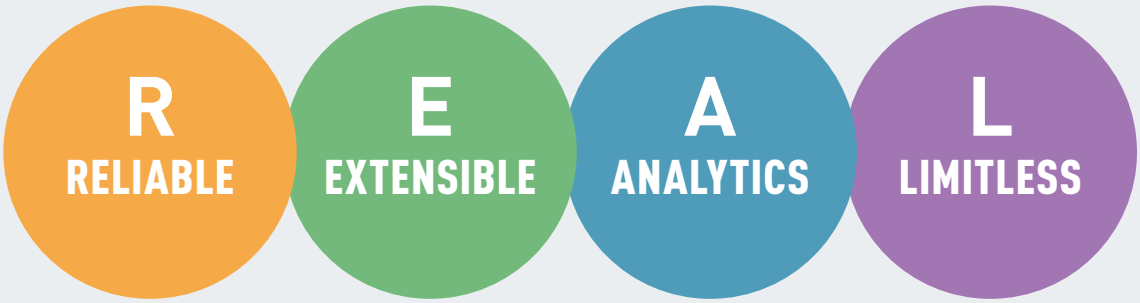
Oh, wait. There is...

THE R.E.A.L. BIG DATA TEST

One system no longer equates to one server, and applications are no longer monolithic. We've outgrown those ideas with the advent of cloud and virtualization, precipitating the massive influx of machine data gained from mobile devices, micro-services, and the Internet of Things.

The complexity of modern architectures requires a new way of thinking in order to get out of reactive mode and take control of your infrastructure. The end game is knowing what you want to accomplish, and actually being able to collaborate and solve problems.

Getting there starts with a simple acronym: **R.E.A.L.**



UNDERSTAND THESE FOUR CAPABILITIES TO MAKE A BUYING DECISION THAT MAKES SENSE

UNDERSTANDING R.E.A.L.

R
RELIABLE

The entire Big Data framework, from data-in-transit to data-at-rest, must be reliable and robust to prevent data loss. You need to be certain you are collecting all the data that's sent into your Big Data solution before you can trust any of it.

E
EXTENSIBLE

Your solution must be built on an open architecture and commodity hardware so that you always have access to your data using industry-standard tools. You must be able to integrate reference data and easily share data in a controlled manner.

A
ANALYTICS

Whether you need real-time streaming analytics now or in the future, your solution needs to provide statistical modeling, data visualizations, and "data replays" for model testing—all as data arrives and is implemented using industry standards.

L
LIMITLESS

Your long-term data needs are hard to predict, but you know they will grow quickly. That means your solution should be able to collect data at both your current rate of generation and by at least two orders of magnitude (100x) greater in order to be ready for future needs.

THE ELEMENTS OF RELIABILITY

Reliability starts a with fundamentally sound architecture:

- Guaranteed delivery of events as soon as they are seen by the collection system's agents or APIs
- Record and store metrics in tandem with logs, such as CPU usage, memory, disk and network activity at any interval and granularity required (e.g. every five seconds or less)
- Retain data online and accessible for as long as necessary, limited only by cost of storage

Three Levels of *Real Time*

Reliability means **low latency data delivery.**

- High Frequency Trading Real Time – sub-millisecond
- Human Real Time – sub-second
- Near Real Time – within 15 seconds

Data that arrives at some "undetermined time in the future" is about as useful as data that never arrives. Anything outside of a <15 second window is simply not real time by any reasonable definition of the term.

A SIMPLE GUIDELINE FOR LATENCY

Your solution should deliver logs and metrics (analyzed, searchable and viewable) in between Human Real Time and Near Real Time, depending on complexity. **No query should ever take hours.**

Support for Your Data Collection Strategy

The pricing and TCO for your solution should not inhibit the collection of data; look for pricing models that are independent of data volumes such as those based on number of users. This frees you to make decisions on how long to keep data and at what level of granularity, based on your business needs, not on your budget.

A good vendor should be heavily invested in R&D focused on reducing storage requirements. That effort will return advantages of increasing efficiencies in storage and power density for their customers.

A SIMPLE GUIDELINE FOR DATA COLLECTION

Users should be encouraged to collect all data and keep it for as long as they need. How long is that? The answer depends on the user's purpose. Think about compliance reporting and analytical model testing, and remember to leave room for unforeseen business needs.

THE ELEMENTS OF EXTENSIBILITY

Open architecture means *choice*.

Technically speaking, event data from IT systems is a mess. It is typically in the form of poorly documented unstructured or semi-structured text. But the data is super-rich, with lots of hidden gems that can be used to improve IT operations and security.

Practically speaking, the richness of the data means you will start out intending to accomplish one thing, but will quickly find many more unanticipated uses.

WHO REALLY OWNS YOUR DATA?

With the never ending variety of operations and uses for operational event data, it's important that you have complete control of your data, which means:

- Open data structures so that you can modify and enrich as needed with no arbitrary limits
- Open streaming protocols so you can integrate best-of-breed products while taking advantage of the breadth and depth of community knowledge to implement solutions
- Open data formats so you completely control the data and how you use it

Don't Confuse APIs with Openness

If you've ever been surprised by missing functions in an API, you'll know why open formats are better. And make sure you not only own your data, but have access to it even if you are out of compliance with license terms or maintenance fees.

It can be a nasty surprise when your apps stop functioning because a vendor has decided to hold your data hostage—something that can't happen with true open systems.

Flexibility is Key

Data collection methods must be flexible enough to support the incredible variety of sources in a global enterprise. While syslog is extremely important, it is just one of many sources. **Beware of systems that can only onboard data from sources whose formats are known in advance or are not user-extensible.**

A SIMPLE GUIDELINE FOR FLEXIBILITY

Trying to be all things to all people is a recipe for disaster. Look for a solution that is built on known, open standards, and lets you manage, access, and control your own data.

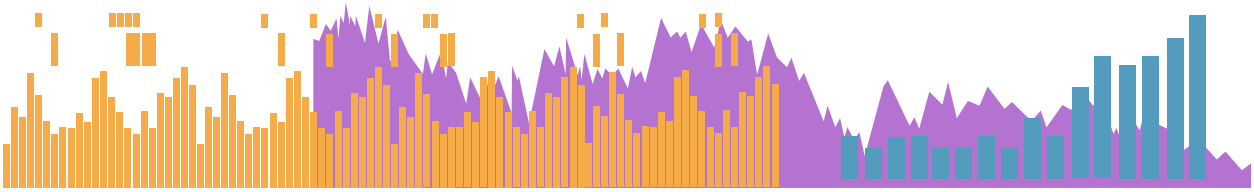
What to look for:

- An open format that allows the system to capture new event types without requiring configuration within the platform before receiving data
- The ability to retain both raw and extracted data.
- The ability to transform data and/or reprocess it later, without affecting the fidelity of the raw data, is critical for compliance and security.

Data is for Everyone. Extend it throughout the Enterprise.

IT operations and security are not the only uses for operational data. Users access this data for a variety of purposes, including improving billing systems, implementing capacity planning, performing churn analysis, and much more. You have lots of teams that can use your data to deliver valuable business insights.

- Data scientists use R or SAS to investigate churn optimization or fraud
- Web developers can analyze patterns to improve user experience
- Business analysts can use familiar languages and tools based on SQL



How can all of these assorted groups make use of your data? Maintain open file formats and schema specifications so the data can be processed by all the different systems that are used in your lines of business.

CAUTION!

Minimize license costs and reduce TCO by avoiding tools that were never designed for R.E.A.L. Big Data, but have been jerry-rigged to address Big Data needs.

Publish/Subscribe

Look for a mechanism to create streams of data for sharing via a publish/subscribe model. The use of custom real-time data streams should have no impact on users' ability to concurrently search and analyze data. Publish/subscribe capability can be used for actions like:

- Shipping data to spreadsheets and BI dashboarding tools
- Forwarding data to purpose-built analytic tools that are not R.E.A.L. Big Data systems but serve legacy purposes and require a subset of operational event data

THE ELEMENTS OF ANALYTICS

In the IT and security spaces, there are lots of tools that let users search their data. **But search, by definition, is not analysis.**

Search is a powerful tool when you know what you are looking for—but it's not very helpful for analysis. Search also places a heavy burden on IT administrators, requiring users to have a general idea of all systems and infrastructure components in order to be able to find critical events and begin analysis.

But sys admins and security operations can **no longer maintain tribal knowledge** of all the applications, services, systems, and associated interdependencies. The volume and complexity are simply too great. As a result, they can **no longer simply search** for problem areas.

WITHOUT ANALYSIS, IT'S PARALYSIS

Look for a tool that learns from your data, automatically applies analytics, and provides visualizations to help you spot problems and rapidly find answers. Instead of spending fruitless hours trying to search mountains of event data, analytic solutions can comb through massive haystacks to identify possible issues and provide the context that enables even novice Site Reliability Engineers to become power operators.

Legacy systems that have tried to build analytic applications on top of search-based paradigms have largely failed. The batch-oriented nature of search is the polar opposite of streaming analytics and compromises time-series analysis. Search-oriented systems also struggle with true "sliding windows" and out-of-order event sequences.

ANOMALY DETECTION ENGINES

The ability to perform real-time analysis against many high-volume data feeds is a critical requirement for both IT operations and security use cases. A solution should include an anomaly detection engine that can:

- Use historical data to construct a quantitative representation of the data distribution exhibited by each metric being monitored
- Compare new data points against these representations and assign a score
- Determine whether the new data point is an anomaly, based on a threshold derived from recent observations of the data
- Continually adjust the anomaly detection model and thresholds to adapt to changing conditions

One of the key advantages of this approach is that the thresholds are not static, but instead evolve with the data.

THE ELEMENTS OF LIMITLESSNESS

How big is "Big?" Bigger than you think.

When people talk about operational data, such as logs, metrics, network diagnostics, and application instrumentation, they tend to view "Big" through a historical lens—they look at what they're currently collecting and place themselves into one of three buckets:

SUB-100GB/DAY UP TO 1TB/DAY MORE THAN 1TB/DAY

But what you're collecting now is very different from what you should be or will be collecting...

And you're not alone; most companies are limited by both legacy system scalability limits and onerous per-byte or per-CPU pricing schemes. They're left struggling to gain visibility and control of their infrastructure because their aging systems aren't a match for the environments they're dealing with today.

Collecting truly "Big" data volumes is no longer a luxury. It's a necessity.

In order to triage IT and security issues that may overlap dozens or hundreds of systems—including virtualized applications and network, storage, and compute tiers—**businesses can no longer afford to cherry-pick data into siloed monitoring applications.**

Everything operational must be stored in a single repository that makes it easier to correlate events from different systems, and delivers true analytics, not just search.

A SIMPLE GUIDELINE FOR "BIG"

"Big" is two orders of magnitude greater than you currently need:

- If you're at 1TB/day today, your solution needs to be proven at 100TB/day
- Your solution must be able to keep petabytes of data online
- Your data needs to be available for analysis in seconds, no matter how old it is

Purpose-Built Solutions

Purpose-built solutions are architecturally different from the ground up. To extract value from machine data, a solution should be architected specifically to handle:

- Real-time streaming
- Long-term full fidelity data retention
- High-speed search
- Machine learning and analytics

No Architectural Limits

When you consider the compute power required to index, ingest, and apply dozens of anomaly detection techniques in Human Real Time on an infrastructure that contains tens of thousands of machines, millions of VMs and containers, arrays of critical network devices, and billions of loglines and metrics per hour, you realize why this is not a problem that can be addressed on a single rack-mounted appliance with a procedure as limiting as search, or on a platform built for a bygone era.

Your solution needs to:

- Build on multiple execution engines
- Combine a wide range of data sets
- Develop algorithmic techniques that are tightly coupled with out-of-the-box visualizations

THE PETABYTE AND BEYOND

Your IT Operations solution should not constrain the IT staff's ability to execute because of scalability limits or onerous pricing models. The scale and complexity of modern infrastructures require real-time analytics and visualizations.

In turn, these require a platform designed to capture, store, and analyze the totality of your operational data. That totality may have been measured in gigabytes a few years ago, but today tens of TBs per day are normal for many organizations, and that number is growing quickly.

Soon, a petabyte of operational event data under management will be commonplace.

That's R.E.A.L. Big Data. Be sure to choose a solution that's up to the challenge.

R.E.A.L. BIG DATA FOR IT OPS: A CHECKLIST

- Reliable event data collection and lossless storage.
- Rapid availability of data: time from event to ability to query data is measured in human real-time
- Maintain and query all data, even data retained for months or years
- Retention of both unaltered raw and transformed data
- Data access, direct and/or through APIs, is unfettered; no license restrictions or "shut-off valves"
- Out-of-the-box support for analytics like anomaly detection and purpose-built visualizations
- Support for both streaming and query-based integration with 3rd party tools
- Able to scale to 100-times your current data volumes
- Pricing independent of data collection and retention policies

READY TO GET R.E.A.L.?

Rocana: The R.E.A.L. Architecture for IT Ops

Rocana has taken a Big Data approach to the problems of unifying monitoring for IT and security operations, providing enterprises with the ability to maintain control of their modern, global-scale infrastructure.

The scale and complexity of modern infrastructures require real-time analytics and visualizations that span the entire spectrum of IT components.

The number of components and interplay between them is too great and too dynamic for IT administrators to effectively monitor. IT teams need tools that will monitor the flow of data and rapidly identify operational issues, then help them get to the root cause of the problem, quickly.

In turn, this requires a platform that has been designed to analyze all of your operational data. What may have been measured in gigabytes a few years ago is already reaching 10's of TBs per day for many organizations and growing quickly. Soon, petabytes of operational event data under management will be commonplace.

That's R.E.A.L. Big Data.

To learn more about applying R.E.A.L. to your Big Data initiatives, visit www.rocana.com.

About the Author

Don Brown, Co-Founder and COO of Rocana

Don Brown innately understands what it takes to make customers successful. As Director of Architectural Services at Cloudera, Don worked as an advisor for many Fortune 100 companies, assisting in both strategic and tactical aspects of their Big Data deployments. He led the global post-sales team during Cloudera's expansion from early Hadoop adopters to mainstream mission critical deployments.

Before assuming the leadership position at Cloudera, Don worked as a Principal Solution Architect, working on dozens of the earliest and largest Hadoop implementations in the world. Don started his career in the data center as the Principal Infrastructure and Security Architect at CompuCredit after working as an intern at IBM doing encryption research to Lead Software Architect at a mobile development company. Don studied Political Science and Statistics at James Madison University. @tensigma

Rocana, Inc. | 548 Market St #22538, San Francisco, CA 94104
+1 (877) ROCANA1 | info@rocana.com | www.rocana.com