Predictive Performance

An exciting look at the future measuring and improving the performance of digital experiences.

WHITE PAPER

Limelight NETWORKS

Predictive Performance

In this whitepaper you will learn:

- How current methods to measure digital experience performance don't capture the whole picture
- How to develop a more comprehensive measurement model
- The future of digital experience performance is all about acting before it happens

> **"Last-mile connectivity is often overlooked when companies consider web acceleration services."**

A Story about Digital Experience Performance

Paul, bleary-eyed from another late-night-into-early-morning-performance session, stared at his laptop as he tried, unsuccessfully, to will the numbers to go down. Only it was no good. The website was still too slow.

He knew he was spending more time each day trying to figure out how to make everything go faster. It wasn't necessarily his job but it had been assumed he'd figure out how to improve the site's performance. The meeting was still fresh in his mind even though it was a few weeks past—his boss and Sam, the VP of Marketing. Sam had been pretty wound-up. He kept claiming that the website performance was causing people to drop off the site which, in turn, was undermining conversions and generating revenue. Paul's boss had only nodded and said, "we'll fix it."

We meaning me, Paul thought as he made another configuration change on the Apache server and pushed it live. I wonder if that will help at all.

But what bothered Paul the most was everything under the surface. Over the past few months the marketing folks had inundated the IT department with requests for new functionality on the website—Facebook integration, video, live chat, and more.

Integration with systems that we have no control over, Paul thought. He kicked off a synthetic browser test and watched the results. No speed improvement. He sighed and shook his head. *The web is just getting more and more complicated. Feel like I'm trying to plug holes in a dam that was built 50 years ago.*

Paul knew that there had to be a better way to measure everything. He kept telling his boss that the end-user's perspective was just as important as measuring system response time, as measuring latency in request responses, as everything they have been measuring for years. He knew they needed a more holistic approach that could tell everyone in the organization, from marketing to IT, what users really felt about the company's digital experiences.

The digital world is too fast for us to just react to it all the time, he thought as he shut his laptop. *We need a way to get ahead of everything so that users feel that our website is a great experience. So they don't feel like they have to wait to use it, or worse, for us to fix it.*

He smirked as he shut off the lights.

Maybe I can dream up a good solution to that!

Tip:

It goes without saying that people usually don't like to wait, especially when it comes to the digital world. In fact, Google researchers discovered that when one organization's site is 250ms slower than a competitor's site, users would gradually migrate to that site. And just to put that into perspective, 250ms is about how fast we blink our eyes!

Paul isn't alone...

Today everyone feels the pain—websites are getting more immersive, more interactive, and more complex and yet we are still measuring "performance" with the same tools and techniques that we did in the 1990s. But more than the pain, the "performance" of digital experiences is becoming a larger business concern. Where once it may have been just the purview of technical folks—network efficiency, design, interface, bandwidth, the user's computer, and more—we are slowly recognizing that other aspects of digital experiences, like how they are designed, impact the end-user perception of performance as well.

It might seem that the question to ask, then: "what's the best way to assess and measure the performance of digital experiences?"

Before we can answer that question though, we need to understand the current state of performance—how it fails to meet the needs of assessing today's digital experiences, and, more importantly, how the way we apply and use acceleration technologies will never provide a truly good end-user experience.

What is Performance?

The first step in understanding the "bigger picture" about performance is to define what performance means. Is it how fast a webpage loads? Is it how responsive a web server is in processing a request? Unfortunately, that is exactly how most organizations define performance—a measure of how fast the website loads according to server-side data. This definition of performance is very limiting because it doesn't take into account what's happening with the end user. Are they on a mobile phone? Is their computer running slowly? In short, defining performance based purely on measuring delivery speeds doesn't present a whole picture of the digital experience.

Today, performance is synonymous with just speed.

But performance is more than just the speed of systems responding to a request for content. In fact, performance-as-speed is really only one factor in measuring the digital experience. Yes, it's critical to understand how quickly internal systems are responding. But it's just as critical to understand how users are interacting with that content when it's delivered. You could have the most blazing fast delivery but find out that objects are loading slowly in web browsers because they aren't optimized. Or, the wrong things are loading at the wrong times.

Predictive Performance

There are two primary categories of performance: Server-side and Client-side.

The **server-side** attempts to improve the delivery of content from first mile/middle mile to the last mile where it can be handed off to an access network for final delivery.

The **client-side** attempts to improve the delivery of content over the last mile by optimizing how content is rendered… and sometimes from where it is requested.

Performance, from a technical perspective, is really about responsiveness. How fast does an object render in the browser? How fast does DNS resolve a request? In short, what is the difference in time between when a user requests something and when a system responds? But from a user perspective, performance is more about "wait". How long did I have to wait before I could start that video? How long did I have to wait for that button to appear before I could click on it? Together, though, these two perspectives represent a measurement of performance that is more akin to today's Web.

And yet that is probably the biggest issue—in most organizations, these two perspectives don't see eye-to-eye. Technical folks claim the website is optimized. Users claim that they have to wait for the website to appear. In order to get to a shared perspective, organizations need to understand both the technical and end-user components of performance:

- **Technology**—elements we can add to webservers, web pages, video players, and more to speed up the request and response process as well as the way that content renders upon delivery.
- **Consistency**—the reliability, availability, and resilience of the delivery system itself.
- **Methods**—the human processes that businesses employ to continually check and modify the technologies involved in improving performance.
- **Authenticity**—the use of real-user monitoring (RUM) to capture the end-user perception of digital experiences more realistically.

When organizations understand these components and can effectively measure both the impact of the technology and the real-user perception of the experience, they can get a sense of the True Delivery Experience (TDE)™.

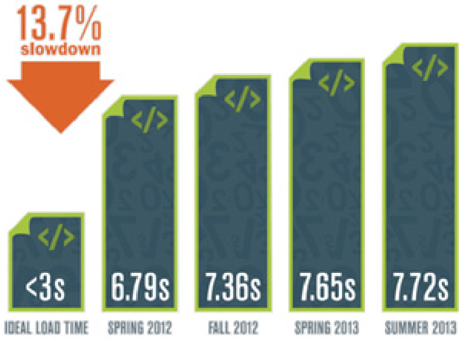
A changing world demands new perspective?

In today's world, practitioners like Paul aren't thinking about a True Delivery Experience measurement. They are thinking, and trying to capture, the measurements from their systems and make changes that they believe may have meaningful impact on how fast their website loads.

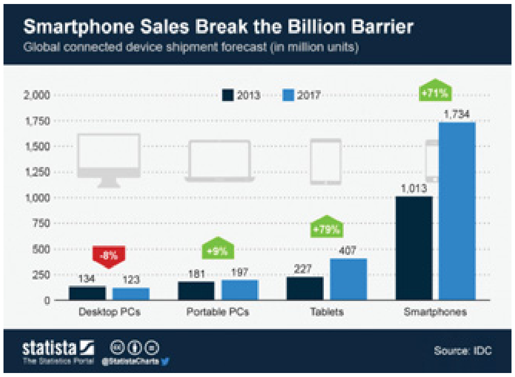
Site loading slowly? Improve how quickly it renders through caching and webserver tuning. Turn on or improve memcache. Videos buffering? Push content through a CDN or deploy adaptive bitrate. Employ a more efficient codec for video transcoding.

Smart and capable people like Paul believe that all of these, and more, will help make a digital experience more pleasing to the end-user. And that is definitely true. But as we pointed out, measuring solely against those factors and implementing changes to improve the results is really only one aspect of improving the holistic digital experience.

In fact, it can sometimes seem like Paul and others are living in a time warp. Let's face it, the web has changed. In fact, you can look at it in terms of the "old" web and the "new" web. In the old web, pages were largely static. There was little video. Measuring the user experience really was all about how fast things were because, for the most part, the elements of a website were entirely in the control of the folks behind the webserver. But in the new web, digital experiences involve many more components—external scripts (like social media integration), dynamic data-driven personalization software, and more. Now there is integration with third-party services and code that may not be under the control of Paul at all. If you look at the graphs below, you'll see that websites have not only gotten bigger, but also slower. Complexity is dragging everything down.



http://www.webperformancetoday.com/2013/07/23/report-ecommerce-page-speed-web-performance-summer-2013/



Which brings us to the big question, "can the old views of digital experiences, and the way to measure them, apply in this new world or are people like Paul measuring with the wrong criteria?" There is precedent for answering "no" to that question. In the heyday of the digital camera market, quality was measured in megapixel. But the number of pixels doesn't really determine the quality of the picture. That simply determines its depth. And back in the days of the desktop PC wars, the speed of a computer was often measured by its CPU clockspeed. Of course, that mattered very little if you had a slow hard drive and sub-standard memory.

It's like we are applying 1990s philosophies to the current environment.

But even if this gets all fixed, is Paul any better off?

"This isn't getting any better, Paul," his boss said. He looked at the report that Paul had just handed him and shook his head. "Why are people still dropping off? You've optimized every part of the HTTP request response?"

"Yup. It's like bleeding a stone. I can't get any more speed out of it."

His boss continued to stare at the report. Paul knew that it was now or never.

"But I've got an idea. I've been reading about this. I think that if we start measuring what people are actually experiencing, like how long it takes for certain parts of the website to load, rather than just the whole site, we'll get a better picture of performance from the user perspective. Kind of like True performance."

"Can we do that?"

"Yeah. We can employ another system that will order what gets loaded by the browser first. So we can talk to marketing about what's really important that users see first so they can interact with the website. Then we can see about bounce rates and conversions. Maybe we are loading the wrong things first, maybe…"

His boss looked up as he crumpled the report and tossed it into the bin.

"Make it happen. Make it happen today."

The underlying issue remains un-addressed

So let's say Paul gets all that fixed—a new understanding of performance that aligns server-side and client side, measurements about how the end-user is really experiencing performance, etc.

Great.

The organization recognizes that they need to implement technologies that address the end-user (i.e., client-side). By doing so, they also understand that measurement is more holistic. They are now capable of generating a True Delivery Experience measurement.

But they haven't solved the ultimate problem. That TDE is still lacking as a solution to improve performance. Why? Because it's reactive. And in a real-time world, reactions are always late, no matter how fast.

Reacting to what's already happened

The best description of how organizations address the issue of their digital experience performance is proactive and reactive.

Proactive refers to the technologies and processes that organizations put in place to hopefully provide the best possible end-user experience with digital assets. A proactive organization isn't waiting for something to break to improve performance. They are constantly looking for ways to "head off trouble" by implementing technologies that may improve performance regardless. Reactive, on the other hand, refers to dealing with a situation that has already happened. Reactive performance follows a pretty standard process:

- **First detection**—where did the issue first manifest?
- **Resolution**—what do we need to do to resolve it (and get the system back to operating as it was before)?
- **Analysis**—what was the manifestation of the issue and why was it there in the first place?

The Transaction Journey

The Transaction Journey represents all the systems touched-by and involved with handling and responding to a user-generated transaction (like clicking on a link or opening a webpage).

So a transaction journey might include systems like a browser, a web server, a database, DNS, routers, switches, application servers. And within each of those systems might be a myriad of functions that can all impact how quickly content is delivered or an experience responds to the user.

As digital experiences become more complicated, so too does the transaction journey.

Although both of these processes—proactive and reactive—are necessary, they are ultimately guesswork. They require an organization to imagine what might have caused a problem and then troubleshoot it when it happens or guess at what might cause issues in the future and install potentially costly solutions. In either situation, an organization must dedicate resources to solving problems which may not ultimately have any bearing on the whole transaction journey.

Thankfully it won't stay this way forever. Fundamental changes are already happening which portend a future that won't be about proactivity or reactivity. Instead, it will be about predictability.

How Cloud Computing is transforming performance

The issue with measuring and improving performance today is people. Before you blow your stack, we aren't suggesting that people are incompetent in improving and troubleshooting digital experience performance issues. Employees like Paul are the rock stars of the digital age. Rather, it's the scale of the problem.

As digital experiences have become more complex, they require analysis of more data points to understand how and which technologies can improve performance. Combine that with an increase in digital content consumption and it's suddenly a herculean task to analyze the data in order to understand what's happening.

That's where cloud computing comes in. The elastic availability of compute, storage, and processing resources is enabling organizations to tackle the challenge of insight into the performance of their digital experiences in real-time.

Big data processing

There are lots of challenges associated with processing and analyzing the data that represents an organization's digital presence. There is simply so much data being produced by both servers and clients at such a frenetic rate (from a myriad of places and devices) that it may seem impossible to get an actual picture of what's happening. Especially when an organization is trying to tackle that itself. Of course, even when an organization implements technologies to handle data processing and analysis with their own hardware (even using datacenters to distribute it), they can't keep up with the scale. A spike in usage can clog up the system turning real-time or near real-time into next week-time.

One of the biggest developments in cloud computing is open-source data crunching platforms such as Hadoop. Hadoop and other software (offering techniques like MapReduce), implemented by cloud computing providers across distributed architectures, enables businesses to quickly tap into resources that allow them to process an endless flow of data. And as the flow increases, the organization can expand resources to handle the load thereby maintaining real-time analysis.

Are we nearing a global computing mesh?

The world is definitely coming closer together. Well, at least from a network perspective.

The expansion and growth of the Internet Exchanges around the world (Virginia, Amsterdam, San Jose to name a few) have enabled access networks to come together to create peering fabrics. These fabrics mean traffic can move more easily from network to network, for example from say Limelight to Verizon. But the value of the growing connections between providers is more than just facilitating traffic.

As these networks begin to incorporate compute resources (using open-source options like OpenStack) and services (i.e., CDN), it's possible that deeper integration could enable organizations to more easily move applications across network boundaries.

Predictive Performance

What is RUM?

Real user monitoring, or RUM, is an approach to performance/ end-user experience monitoring that employs "agents" which passively monitor and "watch" what actual users are doing. This is in contrast to Synthetic Monitoring that utilizes software to simulate what users do.

Monitoring actual user interaction with a website or application is critical to truly understanding the end-user perception of the digital experience. Is the element that users are expecting to load not available until the website has completely rendered? Are there errors occurring? Is there a business process (i.e., sales funnel) failing?

Synthetic Monitoring can't measure this. And there are technologies coming into the market that make it easier to carry out RUM like Crazyegg.com and Soasta.

Compute resources

At the same time that data processing and analysis is adding a new dimension to the value of cloud computing resources, the availability of raw compute power is growing exponentially. Although the supply isn't endless, the elasticity of the resources empowers an organization to implement performance-improving technologies for digital content (such as transcoding software to render video into mobile formats on-the-fly) that may augment their existing infrastructure.

Global distribution

Of course, neither big data processing or compute resources would be much help if there were just in our backyard. Cloud-based services are becoming global. And not just because a single provider is expanding around the world. Providers are peering with each other. Clouds are connected to clouds. Innovative companies are creating software that enable organizations to utilize multiple clouds (and local compute resources) as one giant cloud.

Infrastructure as a platform?

Together, these three cloud happenings have all come together as a platform upon which organizations can extend the resources to enable and deliver their digital experiences. But more than that, this platform can also be used as a home for the processes and technologies needed to ensure that the delivery of those digital experiences meet consumer expectations.

Making True Delivery Experience (TDE)™ a Reality

What cloud computing really enables, though, is that True Delivery Experience measurement. The compute resources, data processing and analysis, and global distribution provide a means to measure the technology, methods, and authenticity (i.e., real-user measurement or RUM) in realtime. Cloud computing becomes the focal point of data coming from multiple systems and, in the case of RUM, from thousands or millions of end-users. For the first time, organizations will be able to get a picture of how all the technologies to improve performance (technology) and the individual configuration changes across delivery or storage systems (methods) is actually being experienced by real end-users (authenticity).

But a measurement like True Delivery Experience isn't much better than what we have today if it remains static, if it relies upon proactive and reactive processes. What cloud computing truly portends is a time when an organization doesn't have to do anything about their performance. The technologies and processes are doing all the work by predicting where performance issues will happen…and resolving them before they ever happen!

Predictive Performance

Go ahead. Imagine a time when your system is ensuring that it is performing at peak without you having to do anything. Sound too good to be true? Well, maybe it is today, but we are rapidly approaching the possibility of that scenario thanks to the technologies enabling cloud computing and a new way to think about performance.

Combining past, present, real, and synthetic

With that data processing and compute infrastructure just a provisioning request away, organizations can finally look at all the performance vectors in real-time. Traditionally, analyzing the performance of a digital experience has usually been about what's happened. But once a performance-impacting event happens, it becomes history. It's logged away. Out of sight, out of mind. Why? Because it's been fixed. But that's a fallacy. It's only been fixed for a specific point in time. What if the digital experience changes? What if end-user clients change? Could that problem crop up again? Absolutely.

But in tomorrow's world, an organization can keep historical data "in the hopper" ready to be analyzed against actual data to determine patterns in system and user behavior that might trigger events. This is the first half of predictive performance. An organization will be able to measure what is actually happening on the servers that are delivering the digital experiences, the user experience through synthetic and real-user monitoring, and the history of what's happened.

This is the True *Real-time* Experience Measurement™. But it's what you can do when you have that measurement that will change everything we think about digital experience performance.

Seeing the breakdown in the transaction journey before it happens

What does all that data reveal? Patterns. And through patterns, changes can be made based on the likelihood of events transpiring again. So if the system can analyze historical and actual data from servers and users in real-time it's possible to understand the transaction journey as it's happening. If the transaction journey is a combination of data points related to responsiveness and environment then real-time data could help determine if a combination of factors may lead to performance issues.

It's a new system of constant measurement—patterns of what's happened against patterns of what's happening.

It's a Happy Ending

The silence bothered Paul.

There were no alert warnings. No emails. No blips or beeps from his smartphone indicating a performance issue.

He clasped his hands behind his head and leaned back, looking at the ceiling. Ever since they put together the predictive performance solution, the system has been taking care of itself. He thought about the last report he'd sent to Sam, his boss. Arguably the best report he'd ever sent. Each line in the activity report simply repeated the same word:

Addressed.

Crackling his knuckles, he sat back up.

Now it's finally time I can get to that application, Paul thought knowing that his end users were always experiencing the best possible performance…without him having to lift a single finger.

Take Action to Optimize Performance

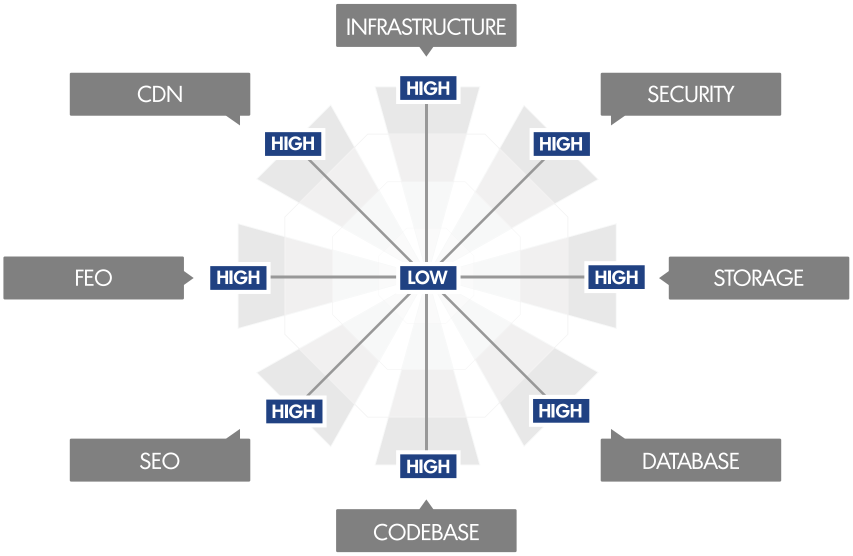
It's clear that the old way won't do for measuring the new way. Organizations and IT practitioners like Paul need a new method through which to generate measurements about the actual digital experience. That's the True Delivery Experience (TDE) measurement that we have suggested before.

But measuring alone doesn't get an organization to performance improvement if there's no way to act against the measurement. That's why the approach organizations need to take to improving their performance has to be more of a model than a method.

One example of a model to optimize TDE is achieved by measuring two values (the level of optimization implemented and the level of organizational understanding of the optimizations available) for eight dimensions of online systems:

- **Infrastructure**—the hardware and hosting solutions that together make up the primary interface, or "origin" of the system.
- **Security**—all systems designed and implemented to protect the system.
- **Storage**—the mechanism for storing the physical files that are required by the system (typically used to describe files that end users need to access).
- **Database**—the database product and structure that is used to store data by the system.
- **Codebase**—the code that makes up all aspects of the system.
- **SEO**—Optimizations made to the outputs of the system, designed to improve the ranking of the system within search engines.
- **FEO**—Optimizations made to the outputs of the system, designed to improve the speed of presentation with browsers.
- **CDN**—the use of CDN services to deliver y the outputs of the system.

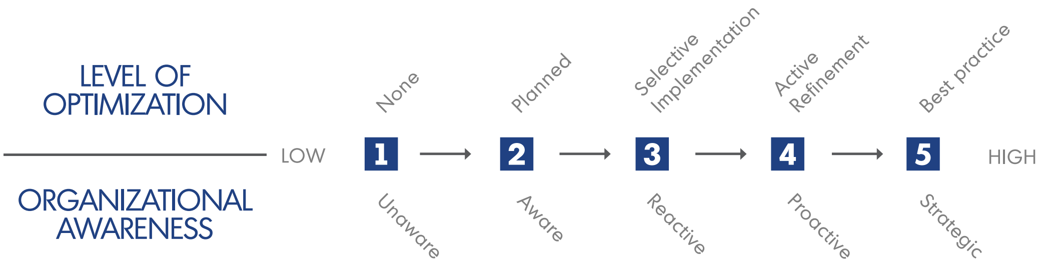
The eight dimensions are then presented in a circle, each with its own axis and scale moving from "low" (L) to "high" (H).



Predictive Performance

How to measure

The eight-dimension model, measured on implementation and awareness, provides a framework against which to assess both server-side and client-side strategies. To measure within the model, each dimension is assessed on a straight-line scale with 5 positions on each scale.

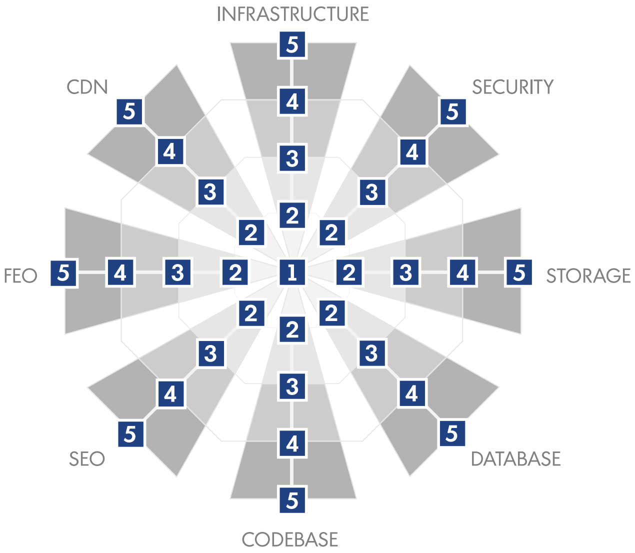


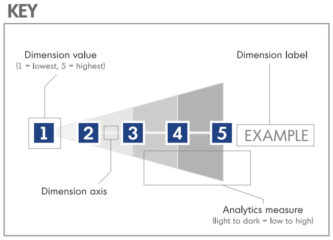
Integrating Analytics

But the model as it stands isn’t complete because it doesn’t take into account the methods by which the dimensions are measured. Those are the analytics systems.

Only we can’t measure analytics like we do the other dimensions because the application of analytics within each dimension may be different (i.e. the analytics tools used to measure database optimizations may be separate from those measuring SEO optimization, for example, and be implemented separately). To account for that, this model applies analytics as an overlay layer to all the dimensions. The overlay is split into segments for each dimension, and measured on the same “level of optimization” scale as the primary dimension.

It is assumed that online businesses understand the need for statistical analysis of their performance. The critical thing that this model measures is the level to which analysis tools and techniques have been implemented. This level may also imply the how well analysis is understood, within the organisation being measured.





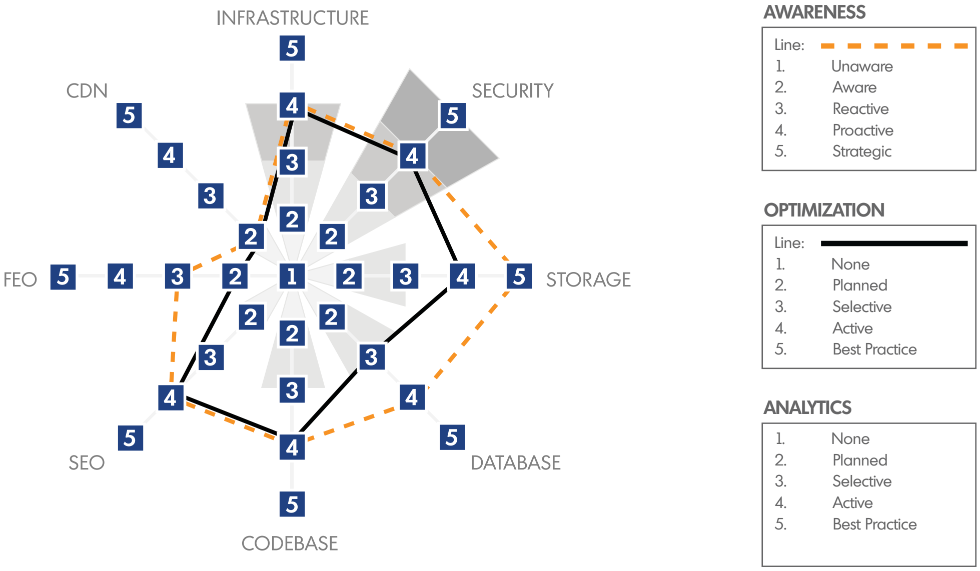
Predictive Performance

Seeing the model in action

So what does this model look like in action? Let's take the following fictitious example:

> *MyCo are a growing online media service (music and video download and streaming portal) who have been delivering their service using largely their own infrastructure and expertise to date.*
>
> *They have a regional presence and are expanding, but are having trouble developing audiences in new markets.*
>
> *MyCo believe that they need to improve the performance of their website and applications to reach these new markets, but are unsure where to focus their optimization efforts.*
>
> *MyCo have undertaken an exercise to model their systems using the TDE Optimization model.*

The results of their analysis are captured in the radar graph below.



What does it mean? Ultimately, that this model is capable of identifying where investments in optimization are most needed and if the organization can meet them. This can help build business cases for strategic investment and determining what dimensions should be considered when addressing growth plans and improvement of the overall digital experience.

But the model isn't perfect yet. What it doesn't show is how to understand on which attributes an organization should focus. That is primarily because the applicable attributes of each dimension will change from industry to industry, and within industry sectors.

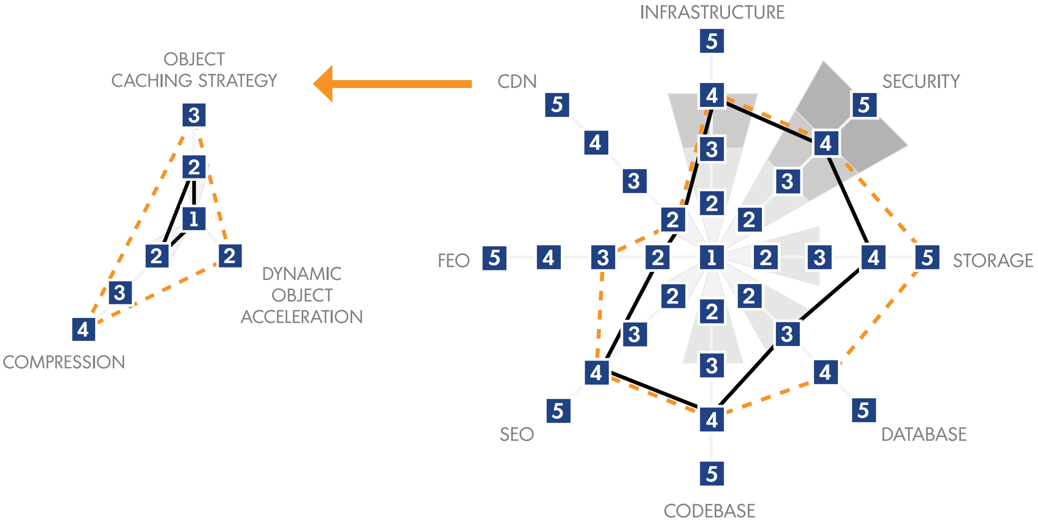
Expanding the Model

The model is inherently flexible, enabling an organization to include and measure sub-dimensions that can help identify how to implement the recommendations implied by the primary 8 dimensions.

Sub dimensions are modelled using the same approach and measures as the primary dimensions, however it is up to the modeller to determine what the sub dimensions are. General guidelines might be:

- These should be industr y specific.
- These should be mapped to the systems primary purpose.
- Specialist knowledge may be needed to identif y the most appropriate attributes to measure.
- Sub-dimensions should be modelled against primary dimensions which are easily identifiable as needing investment.
- Sub-dimensions can be modelled against all primary dimensions to show at a granular level how an organisation can improve its performance overall.

Below is an example of sub-dimensions (object caching strategy, compression, and dynamic object acceleration) for CDN using the MyCo example we explored previously.



Ultimately, this model can help an organization make decisions on how and where to improve performance. Understanding their business (and the profile of their online audience) can help establish the why.

The Future of the Future

With a model and predictive performance we are only one step away from nirvana—automation.

The ultimate vision of predictive performance is when the system is monitoring the system. Imagine an architecture of systems to deliver digital experiences that are all connected via APIs. Then imagine a system on top of that busy with monitoring all that data, of looking through the historical and real-time patterns, connected to all those other systems via API. Would it be too hard to imagine the end-game? The system on top making configuration level changes (turning knobs, flipping switches, pushing buttons) in an automated fashion based not on reaction but on prediction? If there were a likelihood of a negative performance event (probably based on thresholds set by the organization using industry baselines), the system would make changes to one or more other systems that have an effect on that transaction journey.

There is no need for human intervention in a system where data and inter-system connectivity can ensure real-time modifications to make the holistic system run optimally.

Conclusion

The future isn't here yet. Although cloud-based technologies promise significant opportunities to increase data gathering and analysis through the transaction journey, we are still some way off from the True Real-Time Digital Experience measurement. Still there is a lot organizations can do today to get ready for this future. First, they can start deploying cloud computing resources to process digital experience data points both from their own servers and from end-users (using a combination of synthetic and RUM). Second, they can start looking at performance holistically. It doesn't do any good just to focus on one part of a digital experience (like a website). Performance needs to be tuned across all elements of the transaction journey and that's why organizations need to start looking at as "digital experience optimization" rather than "web performance optimization". Third, organizations need to start deploying real-user measurements. Synthetic measurements are only one part of the client-side performance picture. Because when it comes down to it, performance is a perception. And synthetic measurements can't capture what users are actually feeling about your digital experiences…the feelings they express on your Facebook page and Twitter feed. Finally, organizations need to look at performance not as the purview of just IT or just marketing. When the digital experiences that enable users to experience the product are as important to an organization's success as the product itself, everyone has a stake and a role to play. It's time to involve everyone.

The future of performance may be tomorrow. It may be next year. But it's coming. And it's coming fast. Get ready.

[1] http://www.nytimes.com/2012/03/01/technology/impatient-web-users-flee-slow-loading-sites.html

Contact us for more information about Limelight Networks and our services.

Global Headquarters
+1 866.200.5463
+1 602.850.5000

Europe, Middle East & Africa
+44 203 728 6300

Asia Pacific Region
+65 6829 7125

info@limelight.com ▪ www.limelight.com