Cloud workloads: Mitigating risk with performance assessment

Optimizing application migration with time-tested analytical methods

IBM

Contents

Introduction

This white paper details a proactive approach to understanding and assessing end user performance risks before migrating workloads to a cloud environment. We will outline considerations and motivations for evaluating workload readiness for cloud computing, as well as determining the most appropriate method given where your organization is on the cloud adoption journey. We then describe a prescriptive application performance analysis process that you should include in any cloud migration project. And to conclude, we provide recommendations on specific tools and processes that help you develop a better cloud strategy for your organization.

Determining cloud migration readiness

Assessing applications and workloads for cloud suitability allows organizations to determine which applications and data can readily move to a cloud environment and which delivery models (public, private or hybrid) can be supported.

Generally, some applications and workloads are more suitable for cloud computing than others. Various analysis techniques are available to help drive recommendations. Often, analysis starts at a more abstract level, with qualitative questions, and narrows down to more prescriptive, detailed approaches. This paper covers the steps for a detailed approach focused on ensuring end-to-end application performance and availability for business critical workloads.

Evaluating a potential workload's affinity to cloud computing

A good starting point to quickly assess a potential workload's general affinity to cloud computing is through answering high-level qualitative questions across a range of criteria. Often, a cloud consultant facilitates this process through a brief workshop. This type of analysis can explore:

- How self-contained is the workload?
- What are scalability requirements for this workload?
- How standardized is the underlying IT infrastructure?
- Is the workload available as an application or business process on the cloud?
- How substantial is the benefit of rapid application deployment for this workload?
- Does the workload require strong controls to meet compliance or regulatory requirements?
- What are the data transfer requirements for the workload?

Analyzing an overall portfolio for a specific target cloud

Typically, the next step involves a more rigorous workload or portfolio analysis. Detailed information is collected on the existing infrastructure and software stack associated with the workloads of interest. This analysis requires detailed data collection on specific workload images, such as operating system versions and hardware specifications that include memory and storage resources, middleware and software package information. This detailed baseline information can then be combined with nonfunctional requirements and existing operational costs to evaluate the potential fit for target cloud platforms.

These types of assessments may also deploy auto discovery tools for uncovering application connections and infrastructure elements. Typically the more connections discovered or required, the less suitable the application is for cloud migration. In other words, the cost to migrate to cloud computing may outweigh the benefit.

For any workload, numerous variables and considerations exist, such as business performance requirements and transaction flows. A more detailed analysis, as we discuss in the next section, can reveal true performance impact and ramifications. This information helps your organization make the more effective decision regarding migrating to the cloud.

Performance assessment: a more detailed approach to workload analysis

The best time to undertake an application performance analysis is once a cloud strategy is in place and some degree of preliminary workload analysis has been completed. **These workloads often need further analysis with a focus on business requirements and end-to-end transaction flow.** This paper will lay out the process for evaluating aspects of these workloads in more detail, including:

- End-to-end response-time implications of migrating the workload to cloud computing
- Overall integration with other applications and data sources
- Integration complexity (each integration is also referred to as a *connection*)

An application undergoing migration to a cloud service usually has connections of various kinds with other applications and systems. In today's highly distributed environments, each user request can flow through a large number of components (possibly over 50 different servers, application containers and databases)—see Figure 1. This makes it extremely difficult to understand how changes to individual elements affect performance. Managing and tracking transactions end-to-end is necessary to link underlying infrastructure and network components to actual user experience and business requirements.

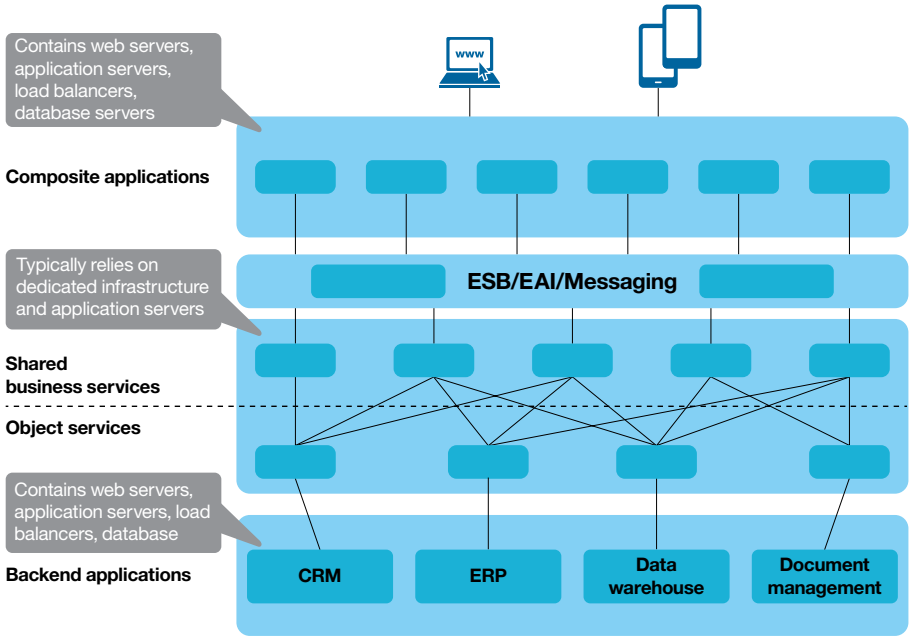

Figure 1. Example components and layers in a typical highly distributed environment[1] *(Note: ESB stands for Enterprise Service Bus. EAI stands for Enterprise Application Integration.)*

Without considering business requirements and end-to-end flow, an application can be prematurely deemed suitable or unsuitable for cloud migration if assessed by the questionnaire approach or through discovery tools alone. In short, application owners need to understand and address the impact of these components and connections, including:

- What happens if an application is migrated to a cloud service and the other applications it has connections to remain in-house?
- What protocols do these applications use to talk to each other?
- Can the same connection continue working, that is, is the protocol supported by the cloud service?
- If so, does the response time remain adequate, given the frequency of transactions, the size of the payload, and the bandwidth and latency of the connection to and from the cloud service?

To answer these questions accurately requires a detailed assessment of the performance and response-time requirements the connection must support.

As shown in Figure 2, different analysis methods deliver different levels of confidence. More due diligence produces more accurate results and reduced risk. The chart underscores that conducting both the business flow analysis and the analytic modeling detailed in this paper can provide a good return on investment—approximately 20 percent gain in accuracy by conducting both.[2] If you require a higher degree of confidence, then the additional step of conducting end-to-end performance and load testing on the workload is recommended.

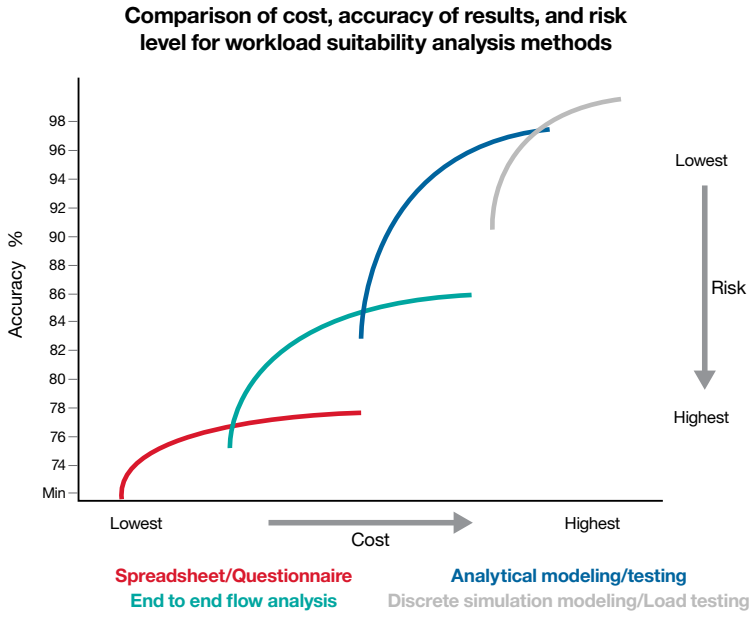

Figure 2. Accuracy versus cost considerations for various workload suitability analysis approaches

Performance assessment steps as part of a cloud migration project

This section provides a prescriptive series of steps end users should take when considering migrating existing applications to cloud computing to help meet workload performance and response-time requirements:

1. Identify business transactions and document their end-to-end application data flow regulations

2. Perform a response-time impact risk assessment

3. Perform response-time impact testing

Requirements and best practices are highlighted for each step in the sections that follow.

Step 1: Identify business transactions and document their end-to-end application data flow regulations

In this consultative approach, the business helps identify individual business transactions, documents their end-to-end flow through the IT infrastructure, and maps the business requirement and importance to the technological considerations supporting it. This includes identifying the:

- End user (or whoever starts the transaction)
- Transaction characteristics (synchronous real-time versus asynchronous batch, frequency, amount of data moved, network protocols used)
- Business importance
- Response-time sensitivity

This needs to be done for each transaction type that has a unique flow through the supporting IT infrastructure. The difficulty of migration depends on the number of interfaces, their complexity, applicable non-functional requirements and whether integration standards are adopted.

Identifying and understanding specific business transactions is the best way to assess performance requirements. A business transaction is essentially a user request. For example, an eCommerce application might include "Check out" and "Add to Cart" as two important business transactions. Each business transaction includes all of the downstream activities until the end user receives a response and perhaps more, if the application does some additional asynchronous processing not directly part of the response to the user.

As shown in Figure 3, the application may have an implementation of "Check out" that performs request validation and then stores the order data into a database. From there, it starts some asynchronous processing relating to the order by means of publishing a request to a topic in a messaging system, before responding to the user that the order is placed, without waiting for the asynchronous processing to complete. The asynchronous processing may then involve a back-end process that listens to the messaging system topic. This process may request the dispatch of items from the warehouse to the customer via invocation of an external service, and then may store data relating to the order in a Hadoop data store in preparation for deep analysis of customer buying behavior.[3]

All of these activities are grouped together into a single business transaction ("Check out") so that you can understand how every part of the system affects the end users and the response time they receive.

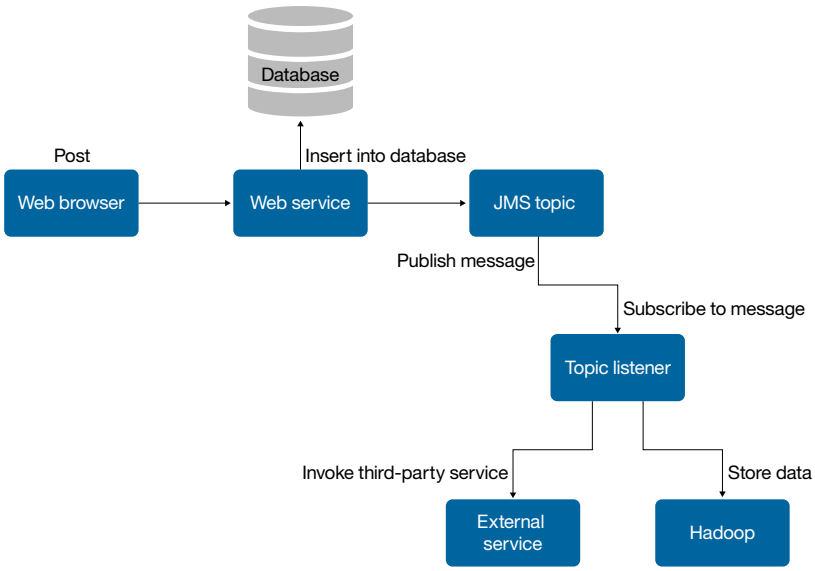

Figure 3. An example of a typical end-to-end flow for a business transaction[4] *(Note: JMS stands for Java Message Service.).*

For the sake of our discussion, it is important to understand the effect of separating any elements of the business transaction resulting from the migration to cloud computing. In the eCommerce example above, initially, the front-end application might be migrated to a cloud service while the messaging system and its downstream processing remain on-premises. What would be the response-time impact to this business transaction caused by the migration, if any? How would the end users be affected?

The best way to answer these questions with confidence is to perform a deeper analysis of the transaction and its response-time characteristics, especially at all connection points. There is a strong likelihood in the case of the "Check out" transaction that a longer latency in the processing of the published message by the asynchronous back-end process is likely to have almost no effect on end users—but this can only be understood by careful examination of all the connections.

At their core, an application and workload are comprised of various types of transactions which may or may not span multiple applications. Not all transaction types have the same performance and response-time requirements, and therefore a separate assessment of each transaction is usually necessary. After the transactions have been identified and documented, a risk assessment should be performed in order to evaluate the latency impact for each and to determine if migration to cloud computing is appropriate.

Step 2: Perform a response-time impact risk assessment

A risk assessment evaluates the business requirement and potential impacts of migrating a business transaction to cloud computing. Impacts could include:

- Sensitivity to delay
- Overall transaction characteristics
- End-to-end flow
- Response characteristics of any integration and connection points of concern

Using established criteria, each transaction should be assigned a rating based on both business importance and sensitivity to delay. Transactions rated as high importance to the business and sensitive to delay are assessed as "high risk of impact" and should be considered for response-time impact testing, as described in Step 3 below.

The approach described here discovers the business requirement, business importance and nature of each transaction—delivering ***the insights to make better strategic decisions about cloud migration.*** Another benefit is that the process requires business team participation, bringing them to the table and helping to establish their commitment to the project.

As an example, given an application with multiple connections to back-end services, one connection may support a synchronous real-time transaction that is response-time sensitive and susceptible to an increase in network latency. Another connection point may support an asynchronous batch-oriented transaction that is not response-time sensitive and is not susceptible to increased network latency, only performed once a week, and only concerned with completing processing within a determined batch window. Considering response-time impact and taking business importance into account, the first synchronous connection may not be suitable for splitting between cloud and in-house services whereas the second asynchronous connection potentially could be split.

Most business transactions in current systems span multiple applications and shared services. At the connection points, factors such as response-time requirements, communication characteristics and network protocol may be different. Using our example from Figure 3, request validation through publishing a request to a topic to the messaging system is synchronous and real-time in nature, and therefore it is considered response-time sensitive. However, from that point forward the processing becomes asynchronous, which is less sensitive to response-time delays.

Therefore, a detailed understanding of a business transaction's end-to-end flow is required. This potentially includes conducting performance testing to truly determine which services are feasible to migrate to cloud computing, and which connections can work from a cloud service to an in-house application. Earlier, we noted that for most high-level cloud affinity assessments, typically the more integrations discovered, the less suitable the application is deemed for cloud migration. ***Still, without considering business requirements and end-to-end flow, an application can be prematurely deemed unsuitable for cloud migration if purely assessed on questionnaire-driven merits alone.***

Step 3: Perform response-time impact testing

Once the performance and response-time requirements are better understood for any given application and associated connections, response-time impact testing is warranted for any proposed changes, impacting the more sensitive transactions. Response-time impact testing helps to quantify the potential impact of changes and can also potentially identify mitigation opportunities. This form of testing involves first establishing a baseline of current transaction response time. The next task is to model the impact on response time of changing application or network conditions such as application message size, available network bandwidth, load and latency.

Assuming "like for like" in regards to system performance and system type, and assuming that all components of an application would move together, assessing potential response-time impact of cloud migration involves two areas of focus:

1. Presentation and tier 1 services migrate to a cloud service and become remote (WAN) from the end users who use them where they were local (LAN) before.
2. Application-to-application and or shared services connections that were LAN based become separated across a WAN after the migration to a cloud service.

If a workload passes the initial assessment for suitability, then in all likelihood from a presentation / tier 1 services perspective it is already architected to meet performance and response-time objectives including those for both local and remote users. However, for situations where the presentation / tier 1 services are legacy-based, such as client server implementations, then application network modeling tools can be employed to establish a baseline. "What if" analytical modeling scenarios can be performed depicting the impact on response times of a network change such as increased network latency—as shown in Figure 4. This includes testing and modeling any application connections of concern.

For response-time impact testing, prioritize transactions that are real-time synchronous in nature, important to the business, and considered response-time sensitive. Start by documenting the end-to-end flow, identifying areas of concern, and then create a plan for testing. Testing should focus on capturing and analyzing the network flows of concern. Testing also has the advantage of validating transaction end-to-end flow and characteristics. Without testing, an understanding of transaction flow and characteristics may be based on incomplete information. It can be challenging for developers to know how their code, and the use of shared services affects performance.

Diagnosing performance issues in pre-production and production environments requires significant efforts from multiple teams and can result in delays in the rollout and delivery of new projects. Testing requires a much deeper review of how things work, and it can lead to discoveries that can change opinions regarding the suitability of an application for migration to cloud computing.

For example, a transaction that is considered real-time synchronous and response time sensitive, and deemed potentially unsuitable to migrate to cloud computing, may only begin that way—on the front end. Deeper inspection may find that the connection is asynchronous in nature, making it suitable for cloud migration. Conversely, the testing might reveal that response-time impact is not acceptable and mitigation steps are necessary.

Response-time impact testing is most critical for business applications that require a phased migration approach (for example, not all elements of a business application can migrate at the same time). In today's complex shared services environment, this is a common area of concern for any data center relocation or consolidation initiative. A well-planned and well-executed migration approach is required to reduce risk. A better understanding of response-time requirements and transaction characteristics help determine the best migration plan by identifying what must move together and what can be separated. Also, where a response-time impact is unavoidable, end user expectations can be set accordingly. In some cases, the business can tolerate a response-time impact for an interim period of time. But there is no way of knowing this if it is not quantified and shared.

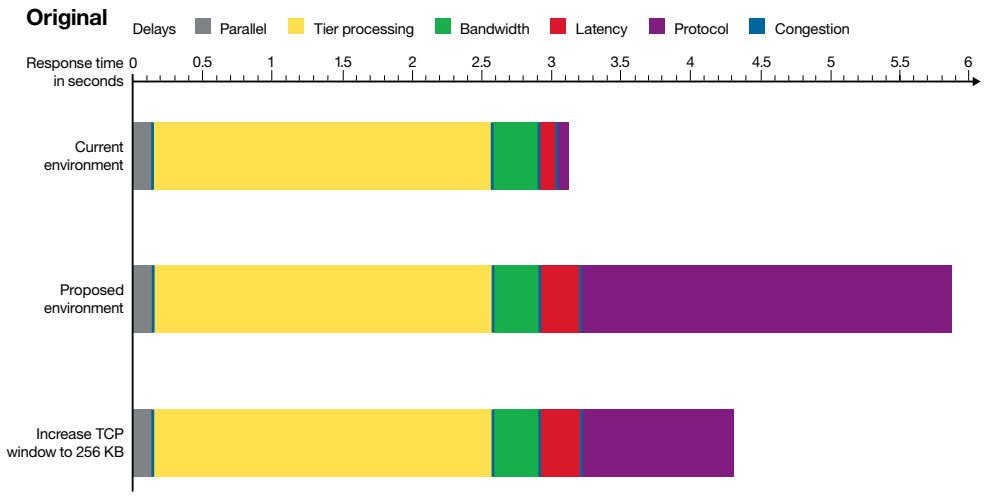

Figure 4. An analytical modeling scenario depicting response-time impact of increasing network latency on baseline application performance. *(Note: TCP stands for Transmission Control Protocol.)*

Figure 4 illustrates the impact of increasing network latency (second and third bar) on baseline application performance in the current environment (top bar). The colors identify where overall transaction time is spent. The second bar shows the effects protocol overhead has on overall response time if a component of the business application is split across a WAN connection—doubling the response time from 3.1 seconds to 6 seconds. The third bar shows the effects of increasing the Transmission Control Protocol (TCP) window size from the default 64KB to 256KB in order to mitigate some of the effects of separating the application.

With the level and detail of data collected from testing, the business can decide whether the impact of the migration to cloud services will be acceptable or not. Tools meeting these requirements fall under the Application Performance Management (APM) suite, providing performance optimization along with capacity planning, performance monitoring and reporting. Often times, these APM tools will include analysis capabilities that help to identify areas of focus if mitigation steps are necessary (for example, where time is being spent, efficiency of application and network interaction, programmatic versus systematic processing delays). By understanding the focus areas, businesses are in a better position to evaluate the cost of migrating to cloud versus the benefit.

IBM's suite of workload analysis capabilities

No matter where you are in the cloud adoption continuum, IBM provides services and tools that can help. For example, our Cloud Affinity Workshop uses our Cloud Affinity Tool to provide simple, practical criteria to understand the positioning of workloads relative to the potential benefits and effort needed to deploy them to cloud. This tool uses criteria based upon our extensive client experience and also incorporates input from IBM research. Graphical output is automatically generated upon completion of inputs, enabling quick analysis and decision making during the workshop. Ideally, this workshop is used in conjunction with building an overall cloud strategy.

The Cloud Affinity Workshop can help define a specific workload scope for a more detailed Workload Transformation Analysis (WTA.) IBM's WTA capability includes the automation of information collection for specific workload images. These images can include operating system versions; hardware specifications such as memory and storage resources; middleware; and software package information. The capability uses an analytical tool that combines the detailed baseline data with nonfunctional requirements and existing operational costs. The final report evaluates the potential fit for target cloud computing platforms.

Application Performance Optimization Consulting (APOC) performs the deepest dive and is an appropriate follow-on activity to our higher-level assessments. APOC helps you accelerate adoption of your cloud strategy by thoroughly assessing performance and response-time requirements, as we have outlined in this paper. APOC can deliver the insights you need to make better strategic decisions about cloud migration, and where it makes the most sense for your workloads and data to reside. It can give you a clear picture of your application integrations and corresponding impacts, as well as the achievability of your performance requirements and response-time goals. This capability can:

- Analyze and map application integration points to business requirements
- Uncover the true nature of each integration point and its suitability for cloud
- Determine which multiconnection application(s) can move to the cloud
- Assist with cloud-based application performance testing, optimization and lifecycle performance management activities

Using quantitative analysis from an IBM patented business transaction mapping process, APOC provides the data you need for more effective decision making around your cloud adoption initiatives.

Why IBM?

A solid strategy for cloud computing is critical to helping you deliver innovative IT services that can create new business value, and IBM Cloud Advisory Services can help. In fact, overall IBM was positioned as a leader in the IDC MarketScape: Worldwide Cloud Professional Services, 2014 Vendor Analysis.[5]

Demonstrated value

Over the past five years, the APOC methodology has helped our clients avoid USD1.2B in costs.[6]

Our approach leads our clients to more effective cloud adoption decisions and a greater understanding of a migration's impact on critical at-risk applications.

Industry-leading intellectual capital

Through literally thousands of client engagements, IBM has developed consistent processes, capabilities, knowledge and experience. This industry-leading intellectual property helps us to deliver on our commitments. Our process-oriented approach helps us to protect your critical business applications and solutions and support your IT initiatives, including data center and cloud migrations, application deployments, and/or holistic application performance analysis.

In-depth experience

IBM has more than 17 years of experience developing and maturing application performance engineering methods, empirical data and using lessons learned. This breadth of experience helps us assist you in aligning business requirements with technical considerations—ultimately avoiding negative business impacts from a cloud migration initiative.

With expertise in 17 industries and global capabilities that span 170 countries, IBM Cloud Advisory Services helps clients around the world benefit from new opportunities available on the cloud. To learn more, visit: **ibm.com**/cloudcomputing.

Note: This white paper has been adapted from *Migrating Applications to the Cloud: Assessing Performance and Response Time Requirements*, published by the Cloud Standards Customer Council.[7]

For more information

To learn more about IBM Cloud Advisory Services, please contact your IBM representative or visit the following website: **ibm.com**/cloudcomputing

Additionally, IBM Global Financing can help you acquire the IT solutions that your business needs in the most cost-effective and strategic way possible. For credit-qualified clients we can customize an IT financing solution to suit your business requirements, enable effective cash management, and improve your total cost of ownership. IBM Global Financing is your smartest choice to fund critical IT investments and propel your business forward. For more information, visit: **ibm.com**/financing

About the author and contributors

Author

Mark Houghtlin has been with IBM for 29 years in various support and services roles in the network and application performance management and optimization space. Mark owns an IBM patent for Business Transaction Mapping and has won two IBM corporate awards for technical achievement and innovation.

Contributors

Mike Edwards is a Senior Technical Staff Member at IBM's Hursley Park labs in the United Kingdom. He focuses on standards, security and privacy in cloud computing and the BlueMix PaaS environment. Mike has worked over 30 years on a variety of software projects at IBM, including OS/2, Java, Web services and service-oriented architecture (SOA).

John Gotsch has over 30 years of IT consulting experience and is currently focused on assisting clients who are using cloud to transform their IT environments by optimizing costs, improving efficiency and increasing IT effectiveness. John is currently a global developer of consulting-based services that evaluate and recommend workloads suitable for cloud migration.

Global Technology Services
Route 100
Somers, NY 10589

Produced in the United States of America
June 2015

[1] "Sharepath for SOA and shared services environment." Correlsense Inc. 2014. http://www.correlsense.com/sharepath-for-soa-and-shared-services-environments/

[2] Based on IBM client experiences.

[3] Whittle, Dustin. "Managing the Performance of Cloud-Based Applications." http://www.appdynamics.com

[4] Ibid.

[5] IDC. "IDC MarketScape: Worldwide Cloud Professional Services, 2014 Vendor Analysis." IDC #250238. August 2014. http://www.idc.com/getdoc.jsp?containerId=250238

[6] Based on IBM client experiences.

[7] Cloud Standards Customer Council. *Migrating Applications to the Cloud: Assessing Performance and Response Time Requirements*, October 2014. http://cloud-council.org/resource-hub.htm#assessing-cloud-performance-requirements

Please Recycle

WUW12371-USEN-00