JOURNAL OF OBJECT TECHNOLOGY

Online at www.jot.fm. Published by ETH Zurich, Chair of Software Engineering ©**JOT, 2009**

Vol. 8, No. 3, May-June 2009

Get Your Head In The Clouds!

Mahesh H. Dodani, IBM, U.S.A.

1 BAILING OUT ENTERPRISES WITH CLOUD COMPUTING

"Jeremy Geelan: What are the main business drivers for Cloud Computing - for this overall technology trend?

Dr Kristof Kloeckner: In the end, it's all about money – how much do you spend for just maintaining the status quo, and how much on supporting truly differentiating business initiatives. This drives an imperative for dynamic infrastructures, increasing resource utilization and reducing labor costs, and for more flexible economics in the consumption and delivery of IT based services." – Cloud Computing Journal Interviews IBM's Enterprise Initiatives CTO, Dr. Kristof Kloeckner

As I promised in my previous column, this article marks the start of our journey into the cloud. I want to start our journey by focusing on the business imperative of cloud computing – what is driving enterprises to embrace it, what capabilities it promises to deliver, and what value can enterprises derive from it.

In today's economy, many businesses are faced with the challenge of reducing the cost of running their enterprises while continuing to deliver new, innovative business services. The facts surrounding the costs of IT infrastructure are stunning, highlighting the fact that inefficiencies are prolific and changes are inevitable:

- In distributed computing environments, up to 85% of computing capacity sits idle.
- Consumer product and retail industries lose about $40 billion annually, or 3.5 percent of their sales, due to supply chain inefficiencies.
- 70% on average is spent on maintaining current IT infrastructures versus adding new capabilities.
- There is an explosion of information, e.g. 54% growth in storage shipments every year.
- 33% of consumers notified of a security breach will terminate their relationship with the company they perceive as responsible.

Clearly, enterprises' IT departments need a bail-out – enter cloud computing as a disruptive approach for consuming and delivering IT based services. In simple terms, cloud computing provides anytime, anywhere access to IT resources delivered

dynamically as a service. It is both a business delivery model as well as an infrastructure management approach:

- The business delivery model provides a user experience by which hardware, software and network resources are optimally leveraged to provide innovative services over the Web. IT resources (including servers, storage and networks) are provisioned based on service requirements using advanced, automated tools. The cloud then enables the users, service creators, and program administrators to use these services via a Web-based interface that abstracts away the complexity of the underlying infrastructure.
- The infrastructure management approach enables IT organizations to manage large numbers of highly virtualized resources as a single large resource. This comprehensive management approach facilitates visibility of the entire environment (applications to physical resources), the ability to control these components to meet specified quality of service requirements and established service level agreements, and the capability to automate many of the manual tasks associated with running the infrastructure. This approach allows IT organizations to increase the traditional server-to-administrator ratio, allowing data centers to increase their resources significantly without having to increase the number of people needed to manage the environment.

Cloud computing provides the enterprise the capabilities to lower the cost of delivering IT services optimized through flexible delivery models, while at the same time making IT more responsive to business needs and allowing greater visibility of IT usage to support billing and chargeback. Cloud computing can be a key catalyst to transform the enterprise to be able to innovate to gain competitive advantage in a fast changing environment by providing reliable IT services that can quickly and flexibly adapt to meet business requirements. Furthermore, the visibility and transparency of delivering IT services facilitates better control, accounting, and governance.

Figure 1 shows the different levels of IT services that can be delivered through cloud computing and the associated delivery models. The cloud-based service can be “public,” “private” or a combination of the two, sometimes referred to as a “hybrid cloud.” In simple terms, public cloud services are characterized as being available to clients from an external service provider via the Internet. The service provider sets the standards, agreements and terms for usage of the service, and provides the user with the necessary mechanisms to easily access the service, use the service, and get billed for the usage. The other model of cloud computing, called a “private” cloud-based service, offers many of the benefits of a public cloud computing environment. The difference is that in a private cloud-based service, data and processes are managed within the organization without the restrictions of network bandwidth, security exposures and legal requirements that using public cloud services across open, public networks might entail. In addition, private cloud services can offer the provider and the user greater control, improving security and resiliency as user access and the networks are restricted and designated.

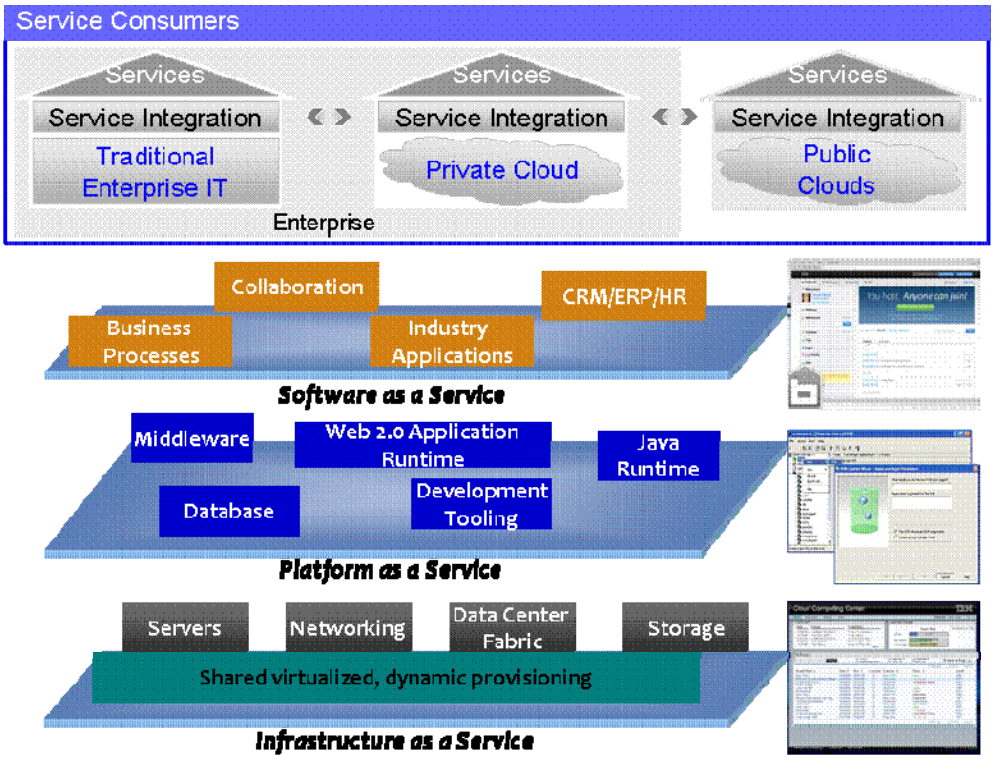


Figure 1: Delivering IT Services With Cloud Computing

The figure also shows the different services that can be provided. There are three basic categories of cloud services: infrastructure services which provides access to a virtualized pool of resources, platform services that provide middleware or an application stack (hardware, operating system, software), and application services that provide access to a specific application or business process (e.g. HR services, email services, collaboration service.) As an example of a infrastructure service, consider IBM's Compute on Demand (CoD) offering, "the leading cloud computing enterprise solution - provides flexible computing power - by the hour, week or year, global access to CoD centers, and the security you can depend on. By off-loading transactions to CoD, you can scale your infrastructure without further capital investments helping to reduce costs and improve your competitive advantage." For platform services, consider IBM's Software Amazon Machine Images, which is IBM's "new agreement with Amazon Web Services (AWS), a subsidiary of Amazon.com, Inc., to deliver IBM's market leading software to clients and developers. The new "pay-as-you-go" model provides clients with access to development and production instances of IBM DB2, Informix Dynamic Server, WebSphere Portal Server, Lotus Web Content Management, WebSphere sMash and Novell's SUSE Linux operating system software in the Amazon Elastic Compute Cloud (Amazon EC2) environment, providing a comprehensive portfolio of products available on AWS." Finally, as an example of software as a service, consider Lotus Live which "provides essential collaboration services to simplify and improve your daily business interactions with customers, partners and colleagues. Work with people seamlessly inside and outside

your organization and streamline communications. LotusLive helps you bring people and information together quickly and simply in an easy-to-use environment, designed with security in mind."

2 DELIVERING ON THE PROMISE

Figure 2 summarizes the key service components in the cloud computing framework. The physical hardware layer is consolidated into pools to provide a flexible, adaptive platform to improve resource utilization. The keys to cloud computing are the next two layers: virtualization and service management. The combination of these two layers ensures that resources in a data center are efficiently managed and can be provisioned, deployed and configured rapidly. This cloud computing environment is designed to handle a mixture of workloads. We will assume the consolidation of physical IT resources is available (as this is a common effort in most data centers) and focus our discussion on the two main layers of virtualization and service management.

Virtualization is the application of the "separation of concerns" principle to abstract the use of resources (by applications, services, etc.) from the underlying physical resources. This abstraction improves agility and flexibility, reduces costs and thus enhances business value of the resources. Untethering the physical resources from specific applications allows virtualized computing environments to be dynamically created, expanded, shrunk or moved as demand varies. Virtualization is therefore a cornerstone of a dynamic cloud infrastructure, because it provides important advantages in sharing, manageability and isolation (that is, multiple users and applications can share physical resources without affecting one another). Virtualization has been a part of IT infrastructures since IBM pioneered it in the 1960s as a mechanism to create logical partitions (LPARs) and virtual machines (VMs) to effectively utilize the vast computing power of the mainframes. Virtualization technologies have continued to evolve to its current form where it can be applied to different IT resources with increasing levels of management and control.

Let us look at server virtualization in a little more depth. The most commonly used server virtualization technology is a hypervisor that serves as a logical representation of the underlying physical servers. The hypervisor allows many virtual machines to run on the underlying hardware, allowing each VM to run a guest operating system and function as if it were solely in control of the hardware. Each VM and the associated operating system is protected from the other VMs and is therefore unaffected by any problems/issues that a particular VM may be experiencing. There are two major types of hypervisors: bare-metal and hosted hypervisors. A bare-metal hypervisor runs directly on the underlying server hardware and provides virtual machines with fine-grained timesharing of resources. A good example of the firmware-based bare-metal hypervisors is the IBM System z® Processor Resource Systems Manager™ (PR/SM™.) All System z models come equipped with the PR/SM hypervisor that provides the ability to divide physical system resources (dedicated or shared) into isolated logical partitions. Each

logical partition operates like an independent system running its own operating environment. On the latest System z models, you can create up to 60 logical partitions running z/VM®, z/OS®, z/OS.e, Linux®, Transaction Processing Facility (TPF), or z/VSE™ on a single system. PR/SM enables each logical partition to have dedicated or shared processors and I/O, and dedicated memory (which you can dynamically reconfigure as needed.) Logical partitions with dedicated resources own the resources to which they are assigned. Logical partitions with shared resources appear to own the resources to which they are assigned, but the resources are actually shared by many logical partitions. Hypervisors are also available for IBM System i® and IBM System p® hardware. Most x-86 based servers use a software-based bare-metal hypervisor such as VMware ESX Server, Microsoft® Hyper-V™ and Xen Hypervisor. The overhead of firmware-based hypervisors is generally less than the overhead of software-based hypervisors. As a result, virtualization implemented at the server hardware level can provide the highest efficiency and performance. A hosted hypervisor runs on a host operating system and uses operating system services to provide timesharing of resources to virtual machines, e.g. VMware Server and Microsoft Virtual Server.

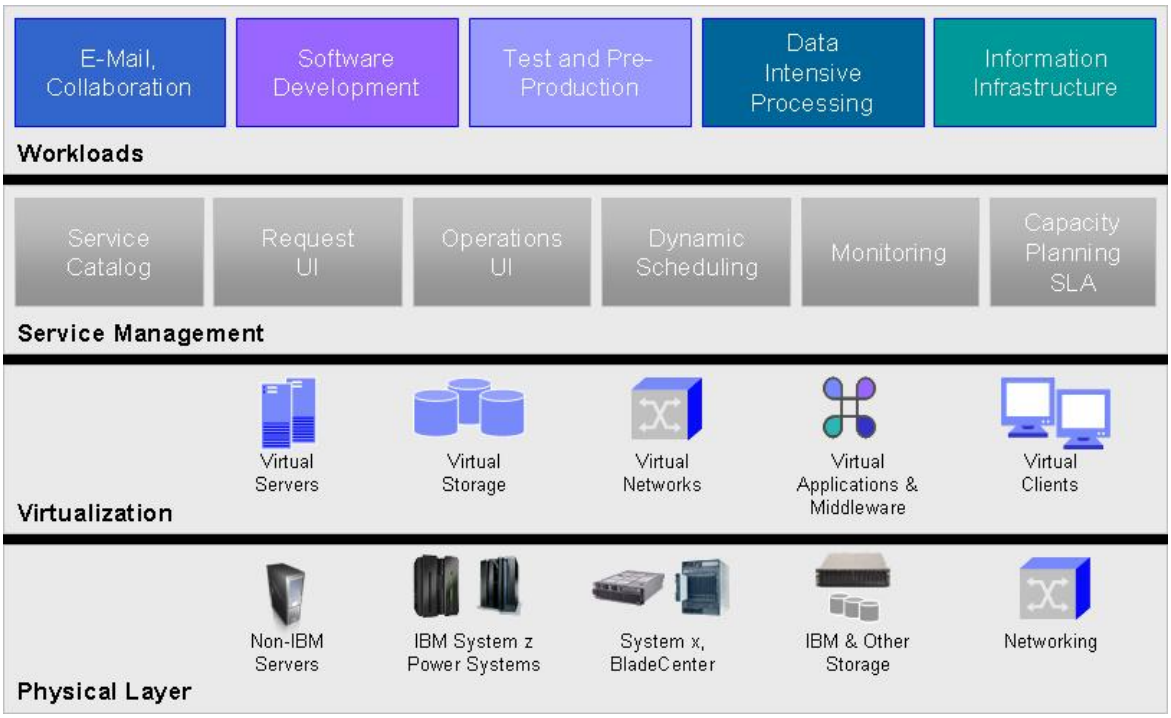


Figure 2: Cloud Computing Services Framework

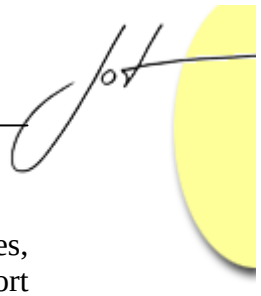
Once virtualization is in place, the IT infrastructure will require a service management layer that is able to visualize, control and automate the IT services to efficiently manage the resources and deliver value to the business. As a first step, rapid provisioning of the resources required for a VM is an immediate benefit of using virtualization. Furthermore, since a VM image and configuration files can be stored on the file system, these VM images can be run on one physical server and moved or copied transparently to another.

Provisioning of a new server or cloning of an existing server can be accomplished by simply creating a virtual server on an existing physical system and copying the previously saved VM images. In other words, a server can be provisioned without reinstalling the operating system or the applications running on it.

Administering the underlying virtualized environment is a major challenge for managing the cloud deployment. It is critical that a cloud be equipped with appropriate management tools and technologies that facilitate, simplify and enable management of the virtualized environment. Automation is an important technique that is applied to two of the most frequent tasks performed in managing a cloud environment: application onboarding and offboarding. Onboarding is the process of installing and configuring the operating system and additional software required by the application. Offboarding refers to the steps necessary to automatically reclaim the IT resources used by an application so that it is available for other purposes. In traditional data centers, both these tasks are done manually, and is time consuming and error-prone. Furthermore, applications typically have unique installation and configuration steps, exacerbating the risk from human errors. Mitigating that risk is possible through automation, by which the many complex tasks can be carried out automatically and consistently. The underlying technologies enable administrators to design workflows that automate the installation and configuration of new servers, middleware and applications, thereby making the task efficient and consistent.

As is evident from the description of the required service management capabilities, a cloud implementation should make it easy for the different roles to interact with the IT environement. Therefore, having both a request and administrator user interface to the IT environment becomes a key component of the service management layer. A self-service portal provides the mechanism for both service requests (through an established service catalog), and for administering the requests. A request-driven provisioning system should be implemented to take user requests for new services or change requirements for existing services. The portal empowers users to do many of the tasks on the systems allocated for their use, and removes the burden typically associated with IT administrators. Users can change their reservation times, add or remove resources, and manage their virtual environments (e.g. starting, stopping or restarting the servers.) Administrators are able then to focus their efforts more on monitoring the entire environment, managing workloads to ensure performance and efficient utilization of the resources, and planning for capacity based on usage trends.

Monitoring resources and application performance is an important element of any environment, and gets more harder in a virtualized environment. Monitoring is needed for effective management as it provides the basis for responding to the requirements of the applications, for reporting on resource usage for costing and accounting purposes, and for collecting data to plan for future capacity requirements. Monitoring applications, virtual machines and physical resources allows administrators to react quickly to unexpected changes in resource needs, immediately detect and solve application problems, and ensure adherence to established service level agreements. Administrators can manage



their cloud environments by moving application workloads to different resources, acquiring further resources (e.g. through infrasturcture-as-a-service offerings) to support their needs, and use managed services to handle problems. Management technologies allow monitoring ofresource usage by applications and users, the ability to allocate costs based on the usage and generate appropriate accounting information. Critical to administering computing resources is the ability to understand what the current and future capacity is to accommodate new requests. Without this understanding, one can neither accurately forecast how many customers can be supported, nor ensure that a steady pipeline of applications can be maintained.

The cloud computing model reduces the need for capacity planning at an application level. An application can simply request resources from the cloud and obtain them (typically in less than an hour) in accordance with dynamic demand. In a cloud environment, it becomes the data center manager's responsibility to predict the average or total resource requirement of all the applications and to order enough hardware in advance independently of the input from application owners. As we indicated above, the enterprise can also acquire these resources "as-a-service" from external service providers. The basis for capacity planning, then, lies in monitoring existing usage and keeping track over historical time periods. Long-term trends can be projected based on previous activity and adjusted without any knowledge of business plans.

The final layer focuses on different workloads that can run effectively in a cloud environment. Through our current experiences, we find the following workloads as feasible for current public cloud capabilities:

- Development, test and pre-production systems.
- Mature packaged offerings, like e-mail and collaboration (see above discussion of Lotus Live as an example.)
- Batch processing jobs with limited security requirements
- Isolated workloads where latency between components in the application is not an issue.
- Infrastructure as a service, including compute resources and storage as a service.
- Managed services for the IT infrastructure, including backup and restore, virus scans, etc.

It is important to note that we have many more workloads that can run in a private cloud environment, as more security and control is available over the workloads within the confines of the enterprise. As the cloud capabilities mature, public clouds will be able to run more complex workloads and address higher qualities of service.

3 THE RETURN ON INVESTMENT

Let us start by looking at the two major dimensions of "cloud-onomics" – the economics of reducing cost and optimizing the business. Cloud computing reduces the cost of running the business by leveraging virtualization, standardization and automation to free up operational budget for new investment. There are two primary levers to achieve cost reduction – operating expense and capital expense, and for many businesses it is not just a question of lowering costs it is also important to strike the right balance between operating and capital expenditures. With virtualization the enterprise has the ability to pool the IT resources to reduce capital expense of hardware, software and facilities. Standardization provides common software stacks, operational policies, and improved service management. Automation removes the many error-prone manual steps and leads to more efficient use of human resources to manage the environment. The more standardization and automation of the IT infrastructure, the more you reduce operating expense – like labor and downtime – which is by far the fastest growing piece of the IT spend. Similarly, the more you leverage virtualization within your IT infrastructure, the greater the economies of your capital expenditures will be. Of course, the cost reduction that you achieve allows you to re-focus your IT spend on optimizing the business by making your IT aligned to the business, more agile and flexible in responding to ever changing requirements while conforming to industry standards.

Determining the value that can be delivered by cloud computing requires an analysis similar to other Return On Investment (ROI) analysis as shown in Figure 3. The first step is to determine the Key Performance Indicators (KPIs) that will be used as the metrics for the value analysis. The second step is to gather the data for these metrics in the as-is environment and to-be environment to show the potential improvements. Also in this step the one-time implementation costs are determined. Finally, the analysis is done on the data gathered to determine the overall ROI and associated characteristics (e.g. payback period.)

Let us do a deep dive into an ROI analysis of cloud computing using IBM's Technology Adoption Program (TAP) as a case study. TAP is IBM's internal "AppStore" where 2,500 IBM innovators deploy new technologies which 100,000 early adopters can use and provide feedback on. The feedback and metrics are then used to assess the business value of the application and determine which applications will be put into production. TAP typically supports 120 projects per year, and needs to provide infrastructure for the innovators as well as the environment for the early adopters to access, use and provide feedback on the applications that they use.

TAP managed a traditional IT infrastructure where deployment of infrastructures was mostly manual, slow, tedious, labor intensive and error prone. It took a long time for IT resources requested by the innovators to be made available, and the "provisioning" process was costly and slow, making it difficult to respond to rapidly changing business needs. It took a typical project team up to 3 months to procure and build the infrastructure that they needed. In this environment, TAP would need 562 servers and 15 administrators to support 120 projects a year, which would require the purchase of 488 additional servers.

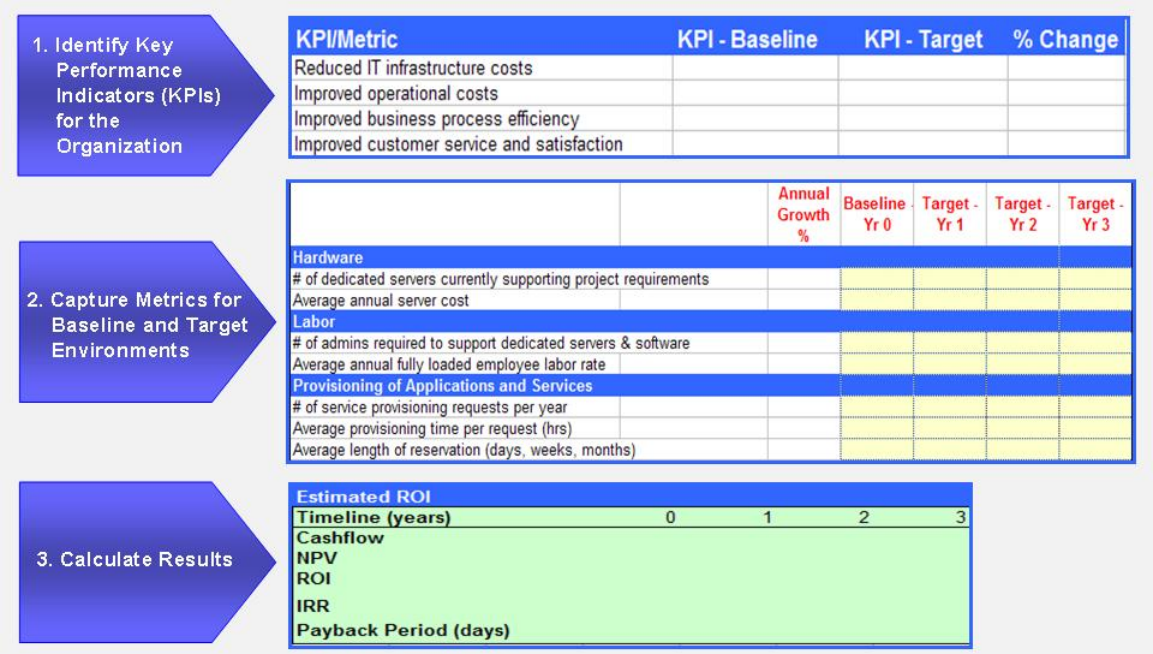


Figure 3: ROI Analysis of Cloud Computing

The TAP transformation into cloud computing started with server consolidation and virtualization to allow the physical servers to be consolidated into a smaller number of more fully utilized physical servers. The next step was to add the ability for innovators to request resources that they needed through a well defined workflow, and to provision these resources for use automatically. The main technologies that were used for the solution included WebSphere Portal to provide an interface for users to participate in the streamlined workflow for innovators' request developed using WebSphere Process Server, and Tivoli Provisioning Manager (TPM) which provided the capability to provision IT resources on demand.

The cloud computing solution enabled the TAP team to support 120 projects with an additional 55 servers rather than the 488 new servers that would have been required before the deployment. This equates to an annual hardware savings of $1.3M and annual power savings of $69,677. The consolidation, virtualization and automation led to a significant reduction in administration costs. In the cloud computing environment, TAP was able to reduce the number of administrators required from 15 to 2, an annual savings of $1.9M. This freed up valuable labor resource to work on other high-value activities. Additionally, the installation and configuration time of TAP solutions was reduced by 75%. TAP was able to reduce the time required to procure a requested innovator environment from weeks to hours, thereby speeding up the time to market of the new technologies. Altogether, these savings reduced operational costs by 83.8%, freeing up funding for new development, acquisitions, reducing debt or paying dividends.

In this article, we have taken a deep dive into the business imperative of cloud computing, showing the key drivers for enterprises to adopt cloud computing, the

capabilities that it provides, and the value that can be derived. Over the next few columns, we will continue our journey by taking a deep dive into the architecture and technical capabilies provided by cloud computing.

About the author

Mahesh Dodani is a software architect at IBM focusing on Cloud Computing. His primary interests are in enabling communities of practitioners to design and build solutions that address complex business needs and deliver value. He can be reached at dodani@us.ibm.com.