JOURNAL OF OBJECT TECHNOLOGY

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

Vol. 8, No. 7, November - December 2009

Cloud Architecture

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

1 THE NEED FOR AN ARCHITECTED SOLUTION

“What is often overlooked in the gleeful rush to cloud computing is the difficulty in molding the early technologies in the space into a truly bulletproof (or even bullet-resistant) business infrastructure. You see it all over the Internet; the push and pull between innovation and reliability, the concerns about security, monitoring and control, even the constant confusion over what entails cloud computing, what technologies to select for a given problem, and how to create an enterprise-class business system out of those technologies.

The truth is, cloud computing doesn't launch our technical architectures into the future. It is, at its heart, an economic model that drives the parameters around how you acquire, pay for and scale the infrastructure architectures you already know. Its not a question of changing the required problems to solve when utilizing data centers, just a change to the division of responsibilities amongst yourself, your organization, your cloud providers and the Internet itself.” – Principles of Cloud Oriented Architecture... James Urquhart Blog

As we have been postulating over the last few articles, a successful cloud journey requires a good understanding of the enterprises’ needs driving the value that can be derived from cloud computing, along with an architected design of the solution to deliver/consume cloud services, and the ability to grow the cloud capabilities and value delivered over a series of iterations. This article focuses on the cloud architecture needed to design cloud solutions that can deliver value.

Let us summarize the requirements leading to the need for the design of a cloud solution, as shown in Figure 1. Simply put, cloud computing is a new service delivery and consumption model with the characteristics of on demand self-service consumption of services through ubiquitous network access, service delivery over elastic location independent resource pools that can rapidly grow or shrink based on demand, and a flexible pricing models to charge for usage. As shown in Figure 1, services can range from infrastructure services which provide access to IT resources, to platform services which provide access to middleware and development/test platforms, to applications, software and business processes that can be provided directly as services.

Mahesh Dodani: “Cloud Architecture”, in *Journal of Object Technology*, vol. 8, no. 7, November-December 2009, pp 35-44 http://www.jot.fm/issues/issue_2009_11/column3/

Another significant requirement are the models for delivering cloud services. Though the initial focus of the cloud capabilities were in supporting private and pubic clouds, the realization that enterprises will have to manage multiple cloud delivery models has shifted the focus on capabilities to support such a mixed delivery environment as shown in Figure 1. There are at least five different delivery models available, three for private clouds, and two for public clouds. For private clouds, the resources are dedicated for the enterprise, and the three models differ on who owns the resources to deliver cloud services and who manages the cloud service delivery. In a private cloud, the enterprise owns the resources and manages the cloud services. In a managed private cloud, the enterprise owns the resources and a 3rd party manages the cloud services. In a hosted private cloud, a 3rd party owns the resources and manages the cloud service delivery. In the private cloud service delivery models, the pricing is flexible depending on the model, and ranges from fixed price in a private cloud to support chargeback, to the addition of time and materials pricing for the management services in managed private cloud, to the addition of pay-as-you-go pricing in the hosted private cloud. For public clouds, the two models differ on whether the resources are shared or dedicated and how the cloud services are accessed. In the shared services model, the resources can be a mix of shared and dedicated, while the access can be through both secured virtual private networks (VPNs) or through the public internet. In the public cloud, all resources are shared, and access is through the public internet. In both cases, the pricing models are based on different types of pay-as-you-go, including usage based pricing, subscription based pricing, and user based pricing. The cloud architecture should define capabilities to support all the delivery models used in an inter-connected cloud delivery environment including the support for federated identity; an “enterprise service bus” capability to allow the different service providers to work together and to handle service interconnections, flow and mediations; an event infrastructure to manage the events that are generated as cloud services are instantiated and used; and a data pipe that allows (potential large) workloads and data to flow between service providers.

As a foundation, the cloud architecture needs to support the basic requirements of any cloud solution, including:

- Delivering cloud services: where applications, data, and IT resources are rapidly provisioned and provided as standardized offerings to users over the web in a flexible pricing model.
- Managing cloud services: where large numbers of highly virtualized resources are managed such that, from a management perspective, they resemble a single large resource. These resources can be managed in such a way to facilitate rapid elasticity dependent on the demands for the cloud services.

The final consideration for the cloud architecture is to handle different types of cloud workloads. As we have discussed in earlier articles, workload analysis is an important step in determining what can run effectively in a cloud environment. It is important to understand the characteristics of the workloads in the context of a cloud delivery to determine its suitablity to be delivered as a cloud service. We want to avoid workloads where risk and migration cost may be too high, including those that are database intensive, have complex transaction processing, packaged application (e.g. ERP) workloads, and highly regulated workloads. We want to focus on workloads which can be standardized and take advantage of running as a cloud service, including web infrastructure applications, collaboration infrastructure, development and test workloads, and high performance computing workloads. Finally, we should consider new workloads that have been made possible by cloud including high-volume low cost analytics, collaborative business networks, and industry focused “smart” applications. The cloud architecture needs to provide the support for different types of workloads, especially related to their qualities of service (QoS) requirements, the different types of images that are needed to deliver the types of services implied by the workloads, and any special business and compliance policies that they need to adhere to.

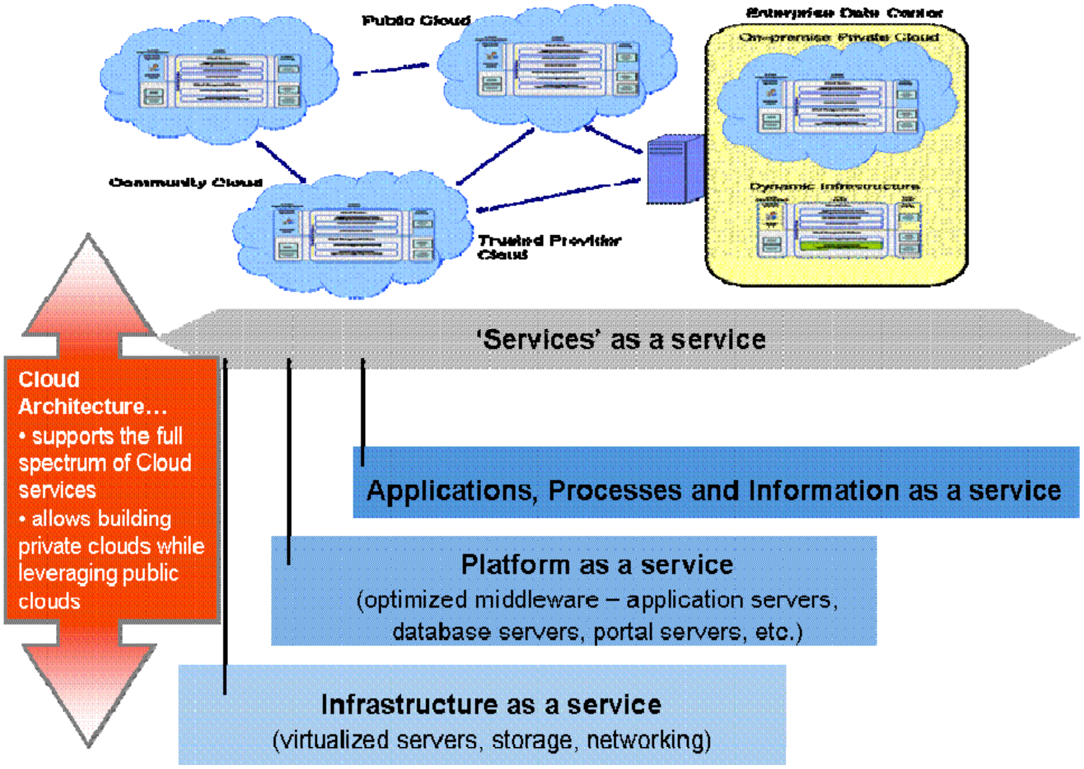
Public Cloud
Enterprise Data Center
On-premise Private Cloud
Community Cloud
Dynamic Infrastructure
Trusted Provider Cloud
'Services' as a service
Cloud Architecture…
• supports the full spectrum of Cloud services
• allows building private clouds while leveraging public clouds
Applications, Processes and Information as a service
Platform as a service
(optimized middleware – application servers, database servers, portal servers, etc.)
Infrastructure as a service
(virtualized servers, storage, networking)

Figure 1: Cloud Service Delivery

2 THE CLOUD REFERENCE ARCHITECTURE

The cloud reference architecture depicted in Figure 2 shows the separation of concerns among the service requestor, service provider and the service creator, and shows the capabilities required to deliver the different types of cloud services – infrastructure, platform and applications.

Note the definition of a common cloud management platform that delivers the business support systems and operational support systems needed to deliver the different types of cloud services. The sophistication of these BSS and OSS capabilities depend on the level of characteristics needed to deliver the cloud services. For example, to support flexible pricing models, a public cloud service provider would need all of the BSS capabilities along with the OSS metering capability. On the other hand, an enterprise that has chargeback mechanisms in place, will need the BSS billing capability along with the OSS metering capability.

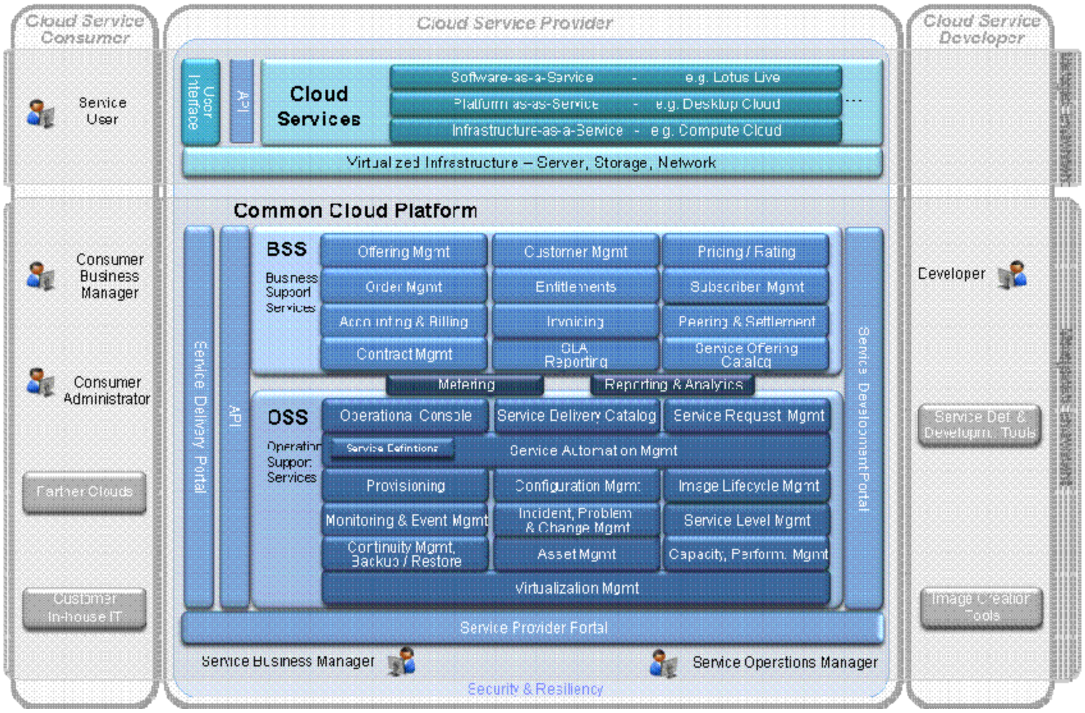

Figure 2: Cloud Architecture

Business requirements drive the cloud service offerings, and shows the range of offerings that are needed to support the different requirements, including customers using cloud computing to supplement traditional IT. Note that the cloud architecture will need consistent capability to monitor and control heterogeneous components across traditional IT and cloud. Furthermore, with the different loosely coupled workloads emerging, the cloud architecture will need to provide support for workload focused offerings, including analytics, application development/test, and collaboration/e-mail services.

Technical requirements drive the underlying IT management patterns, including a focus on handling the top adoption factors influencing cloud services – i.e. trust, security, availability, and SLA management. Figure 3 summarizes the main capabilities in the operational support systems. The architecture must focus on handling the major concerns of enterprises by facilitating internal/external cloud interoperability. This requires the architecture, for example, to handle licensing and security issues to span traditional IT, private and public clouds. Additionally, the architecture must support a self service paradigm to manage clouds using a portal which requires a robust and easy to use service management solution. A portal is key to access the catalog of services and to manage security services. Of course, all of these services must be provided on top of a virtualized infrastructure of the underlying IT resources that are needed to provide cloud services.

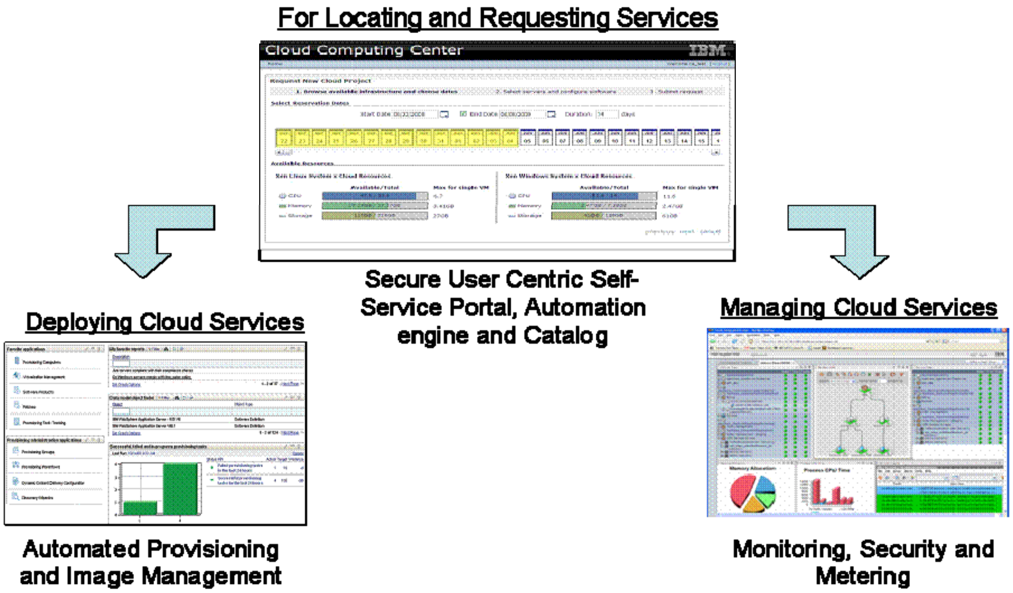
For Locating and Requesting Services
Cloud Computing Center
IBM
Secure User Centric Self-Service Portal, Automation engine and Catalog
Deploying Cloud Services
Automated Provisioning and Image Management
Managing Cloud Services
Monitoring, Security and Metering

Figure 3: Cloud Service Management

As described above, 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 summarized in Figure 4. 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 and reporting resource usage data by applications and users, and 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 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. 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.

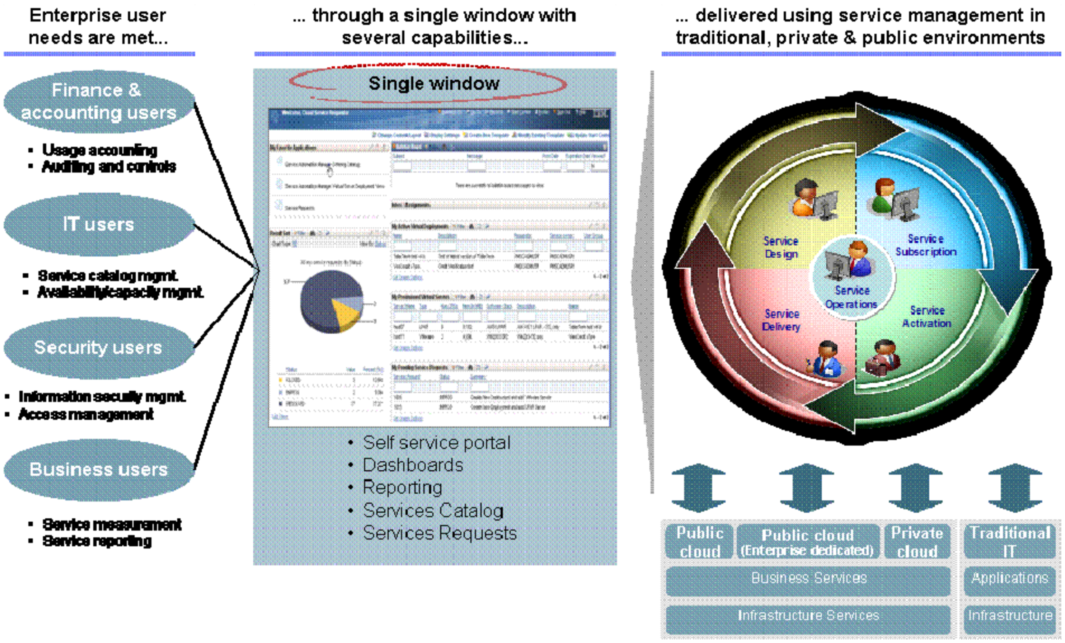

Figure 4: Cloud Service Management Capabilities

3 DESIGNING A SOLUTION USING THE ARCHITECTURE

Having understood the problems associated with cloud adoption, we can set up a program that can tackle the issues associated with the virtualization mindset, manage the transition to the cloud mindset (along with the associated behavioral changes), and help with keeping the organization entrenched in the cloud. Figure 5 summarizes the Five-Step program for cloud adoption. Step 1 is establishing the roadmap for transforming the IT infrastructure – moving from the traditional data center approach to a centralized and consolidated infrastructure, through virtualization and automation, and finally realizing an optimized IT infrastructure capable of delivering on the promise of cloud. Step 2 establishes the architecture that will provide the capabilities needed to deliver cloud services effectively and efficiently, considering the service consumer, provider and creator. Step 3 focuses on analyzing the workloads that are feasible to move to the cloud. Step 4 focuses on determining the right mix of delivery models to deploy and use the cloud services. Finally, Step 5 lays out the implementation approach for deploying the cloud services. Note that each step focuses on the transformation of the entire enterprise into the cloud mindset. Let us discuss each step in detail.

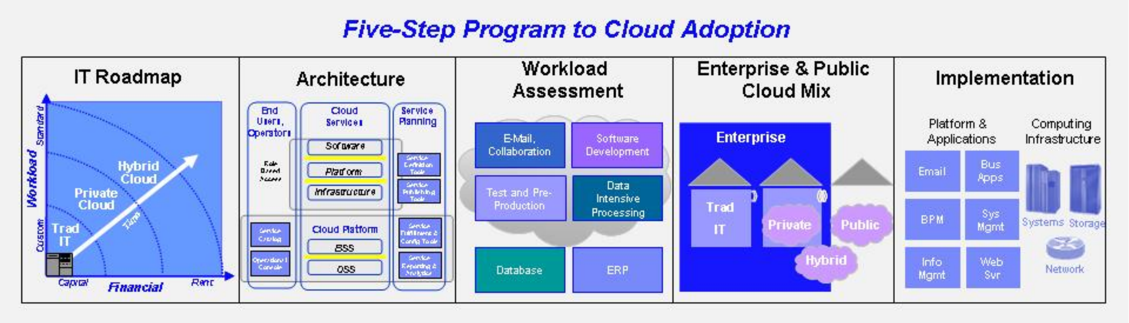

Figure 5: The Cloud Journey

As you can see from the program, the cloud solution is architected, designed and implemented. In the following we will focus on steps 2 and 5 to show how the cloud architecture shown above can be leveraged in the design of the cloud solution and the implications on the implementation of the solution.

Step 2 focuses on the cloud architecture, and provides a guide to understanding the capabilities required for an enterprise to derive value from implementing cloud. The architecture follows the separation of concerns priniciple, and subdivides the capabilities of cloud into the three main “parties” – the consumer of the cloud service, the provider of the cloud service, and the creator of the crowd service.

The cloud service creator component has capabilities targeting all aspects of the lifecycle of the “image” that is used to bundle the service that is accessible by the consumer. This “image” can include the IT resources (e.g. server, storage, network), the operating system, the middleware, and the applications. The capabilities support the need to design and build the image, store it in the library of images that can be accessed by the users, the deployment of the images, and the management of the images through the entire operational lifecycle.

The cloud service consumer serves both the end users and the operators that manage the infrastructure. The capabilities in this component ensure that the images that can be accessed are defined in a catalog, and have appropriate role-based user interfaces to access and manipulate the images.

The key capabilities of the reference architecture are defined in the service provider component. The lowest layer of the architecture defines the capabilities of the virtualized infrastructure. These capabilities facilitate virtualization of all IT resources: server, storage and network. These virtualization capabilities can handle all types of IT resources, e.g. both mainframe and distributed servers. The next layer provides an optimized middleware with capabilities for image deployment, integrated security, workload management, and high-availability. The optimized middleware is used as the way to deliver services and information built according to well defined SOA and Information architectures. The central piece of the component is service management which provides the capabilities to manage a cloud service. These services include capabilities to handle user requests: managing the self service requests made by users, the lifecyle of images, and the provisioning of images based on the request. The capabilities also handle many of the qualities of service associated with delivering images, including availability, backup and restore, security and compliance, and performance management. To facilitate delivery through flexible models, the capabilities also support usage accounting and license management.

The reference architecture provides a comprehensive set of capabilities to ensure that cloud services can be built, deployed, accessed, delivered and managed. Each of these capabilities are supported by the appropriate standards, technologies and tools – all integrated to work together to deliver cloud computing.

Step 5 focuses on the implementation and deployment of the cloud service to derive business value for the enterprise. Each cloud implementation must focus on providing the capabilities for the service consumer, the service provider, and the service creator as were articulated in the cloud architecture described in Step 2. The following are key considerations in any implementation:

- The implementation should provide an easy to access, easy to use service catalog that is used by users to request services, and by administrators to publish available services.
- The implementation should hide the underlying complex infrastructure from the user and shift the focus to services provided.
- The implemenation enables the ability to provide standardized and lower cost services.
- The implementation facilitates a granular level of services metering and billing.
- The implementation should ease complexity through workload standardization.

The implementation should consider use cases for both users and administrators, ensuring capabilities for requesting/using/maintaining/releasing cloud services and for monitoring/managing/maintaining the cloud services and the underlying virtualized infrastructure.

To summarize, ensuring the success of your cloud journey and the value from moving to a cloud service delivery model is highly dependent on designing an appropriate solution that can address your specific requirements and value drivers. It is important that you base this solution design on a well thought out cloud architecture, that is capable of supporting the capabilities that you need to address the requirements of your cloud solution. You can base your architecture on the comprehensive IBM Cloud

Architecture, that defines a comprehensive set of capabilities that can support the delivery and management of different kinds of cloud services in a unified service management environment, and facilitate the delivery of these services in a hybrid environment. Over the next few articles, we will take examples of designing solutions and show how the cloud architecture can be appropriately leveraged in the design.

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.