Putting Your Enterprise Data in the Cloud? Top Considerations
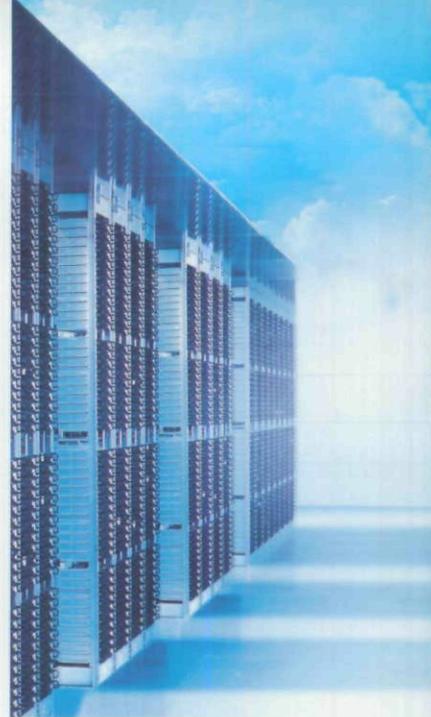
Most discussions regarding the possible transition to the cloud tend to center around transition of on-premise applications to the compute cloud (e.g. Amazon EC2 or Google App Engine). In this article we will focus on transition of on-premise data to the storage cloud – specifically public storage clouds offered at the infrastructure layer (e.g. Amazon S3) – and the factors one should consider before making the switch.

Primacy of Secondary and Tertiary Use

Latency issues will keep the cloud from becoming the primary storage for most on-premise applications. On the other hand, the cloud is increasingly becoming a destination for secondary and tertiary storage (Fig 1). Instead of writing to a tape and shipping tape to an off-site secure location, many businesses are now sending their backup data and archives to a remote storage cloud. In the process, they acquire the risk profile of a much larger organization, such as Amazon, which spends millions of dollars to ensure robustness and accessibility of their data centers.

The recent downtime of clouds has grabbed the headlines, but fortunately, most secondary and tertiary applications of storage don't require always-on service level agreements. In other words, if the storage cloud is not accessible right now, backup software should be intelligent enough to make it up in the next backup run. Of course, you may hit Murphy's Law if you happen to be recovering a file simultaneously while your storage cloud vendor is down. Fortunately, sending data to multiple clouds is a far easier option than trying to send physical media to two different locations.

Just like the answering machines gave way to voicemail, tape libraries will eventually give way to cloud-based storage.

Over-provisioning: Expensive; Under-provisioning: Hara-kiri

A key reason to move to a storage cloud is provisioning, one of the most hairy problems for storage administrators. Storage capacity needs, especially for backup applications, can gyrate widely over time, making it nearly impossible to accurately provision physical storage for a particular application. Since under-provisioning tends to have significantly dire consequences, most storage subsystems are over-provisioned.

A storage cloud should be able to scale, based on changing needs. Let's say, for example, that data retention policy changes, due to a new compliance requirement, requires an organization to maintain data for an extended
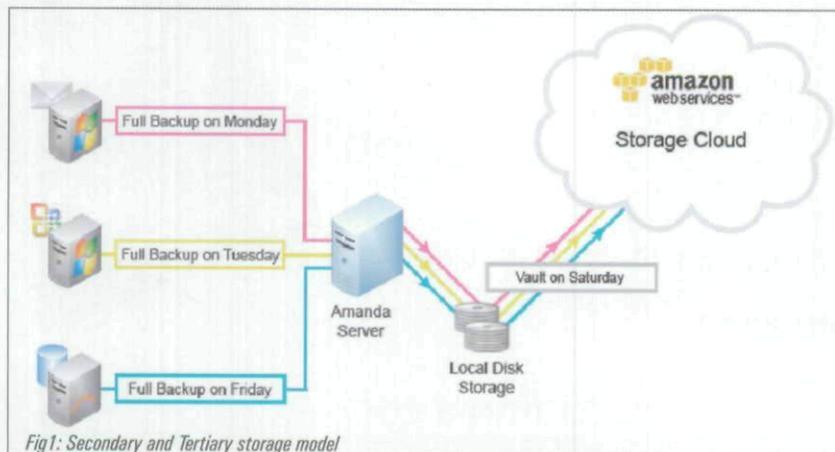

Fig1: Secondary and Tertiary storage model

period of time, the storage cloud being used for your backup archives should be able to easily scale to accommodate the increased need for storage capacity. On the flip side, if an organization decides to purge a set of backup archives that are no longer needed, the storage capacity and associated costs should reduce immediately.

The storage cloud should be elastic to your needs, both in terms of scalability and cost.

Security and Privacy

Our research suggests that security and privacy of data is a key concern for moving data to the cloud. A cloud vendor should provide documentation on the security and privacy measures that are taken in their data center. Ideally, if sensitive data is being transferred, it should be encrypted even before it leaves the data center. The U.S. government agencies, including the National Security Agency, consider Advanced Encryption Standard (AES) algorithm with 256-bit keys suitable for top secret documents. So, this is a good choice for encryption before sending sensitive data to the cloud.

Application of these practices makes most practitioners think of electronic keys to be stronger than physical keys.

Compliance

Data management is where regulatory compliance obligations meet IT. The European Commission's Data Protection Directive 95/46 restricts European Union (EU) businesses from shipping several categories of data to 'third countries', defined as any country outside of the EU. So, as a storage administrator you need to make sure that storing your data in a particular storage cloud meets the compliance needs of your industry or regional jurisdictions.

Robustness

If you choose a storage cloud as the destination of your crown jewels, first and foremost it needs to be at least as robust as storing it on an in-house RAID array. A good storage cloud vendor should at least create redundancy within their main data center. Ideally, there should be geographic redundancy to protect against data center-wide meltdown or connectivity issues. Any single failure must be tolerated without causing downtime for the end-user application or inhibiting your ability to access your data.

Seamless Retention

Your storage cloud vendor should seamlessly support your retention policy — however long it may be. While underlying technologies of storage will change over the years, your data should always be accessible at the same URI. This is in fact a key advantage of leveraging a storage cloud for archives and getting rid of the hassles of moving from one media type to another every few years.

Location Matters

Another factor, which makes the location of the cloud a key decision factor, is availability of network bandwidth. For some geographies reliability, speed, and cost of transferring data to a cloud in another continent can be prohibitive. In this case, you may need to choose a storage cloud which is geographically nearer. The cloud vendor should also do asynchronous replication to a remote location in the background to provide geographic redundancy.

Advantage of Open

Locking up your data in proprietary formats and proprietary clouds comes with a huge cost and pain. When sending data to the cloud you need to make sure that you will always have access to your data, regardless of whether you are still using a particular vendor's technology. This is where open source software plays a key role. Amanda, the world's most popular open source backup software, is integrated with the storage cloud. Amanda uses an open format regardless of the media it is writing to (e.g. disk, tapes, or cloud).

In addition, Amanda supports up to 2048-bit keys with public-key cryptography as well as 256-bit AES encryption. Amanda's scheduler is intelligent enough to gracefully pick up the backup cycle if you missed a cycle because of unavailability of the storage cloud. With a single Amanda server you can choose to store your backup archives to multiple media at the same time. You can maintain a repository of archives on a local disk for near-term recovery, and a set of archives on the storage cloud for long-term archiving and disaster recovery. ◼

About Author

Chander Kant,
Chief Executive Officer,
Zmanda