How to be certain about your data in an uncertain future

Be a speedboat, not an oil tanker. And don't let the CEO read Forbes

28 Jan 2016 at 12:35, Danny Bradbury

If it wasn't for users, managers, or compliance execs, IT would be an easy place with goalposts that stayed put. The real world is far less predictable. The rules of play may change. So how do you design data strategies to cope?

Data regulations are a good example. The EU's Safe Harbour legislation made the rules clear when it comes to storing personal data outside the EU. Then Snowden revealed that the US was snooping on everyone's data, causing people to rush to the European courts to question the offshoring of data over the Atlantic. The whole thing unravelled, leaving companies less certain about where they stood than before.

Other regulatory factors may influence what you have to do with your data. Maybe your industry will issue new guidelines. Or your compliance officer will suddenly decide that they have to anonymise everything.

The changes don't have to be regulatory, either. A merger or acquisition might mean that your IT department eats – or is eaten by – another. Your data may suddenly have to play nicely with someone else's. Perhaps the CFO went to a conference last weekend and has since decided that hybrid cloud is the way to go, or maybe the CEO read an article about big data and real time analytics in Forbes. You can run when you see them walking towards your office, but you can't hide.

Jon Cano-Lopez is the CEO at ReAD Group, which builds vast databases of UK consumer data, and his tech team has to think constantly about how to ensure their data architectures can adapt to change.

“If you ever try to sit down and work out a solution, by the time it is delivered, the requirements will have changed slightly,” he said, adding that you have to future-proof the data architectures without endangering sensitive data. “The underlying data structures have to be flexible, as do all the onboarding processes. But you have to do that without taking shortcuts.”

What should you hedge for?

Doing this properly involves some form of risk analysis. You may not know what’s going to happen in the future, but you can at least figure out what’s more likely, and what’s going to have the biggest impact. Look at external pressures on the company, from regulators and elsewhere. What have competitors had to do?

Are companies in your sector driven towards low-latency apps that need data residing closer to the server? Is your competition migrating to the cloud and saving money? What might that mean for you? Another risk to look for is vendor lock-in, warned Artyom Astafurov, senior vice president of M2M at global technology consulting firm DataArt.

“When you’re building a data structure around a platform like Amazon Web Services, you’re highly dependent on their services,” he said, by way of example. “Vendor lock-in is the price you pay for a gain in velocity.” The same is often true when choosing particular hardware vendors for on-premise solutions. Having a list of possible external disruptors and their associated effects on your data will give you a corresponding list of desired characteristics. Being more cloud-friendly may make data portability key, depending on your cloud strategy and the likelihood that it may change, for example. If your business is ‘risk on’ and looking for growth, then maybe your board is anticipating new business products and services that will place new demands on your infrastructure. How strong is your dialogue with them?

After working out the likely changes and the kinds of capabilities that you want in your data architecture and supporting infrastructure, auditing your existing capabilities can help you baseline what you already have and then perform a gap analysis. What stands out as a problem area? Data silos might be an obvious one. Analytics are poised to transform contact centre operations, for example, but for that to work you need visibility, which implies access to data from various sources in the company. If contact centres are an important tool to your business but all your data is stovepiped in different systems and formats, unlocking it could be a focal point for your modernisation roadmap.

Abstract all of the things

One of the biggest barriers for companies trying to make their data architectures and supporting infrastructures less rigid is that the data services and the infrastructure they rely on are still tightly coupled, warned Astafurov. He argues that abstracting one from the other can make it easier to access them programmatically.

When vendors talk about software-defined anything, from storage to whole datacenters, this is typically what they’re on about. The concept promises IT departments the ability to adapt their infrastructure to support new services as necessary. Astafurov takes it one step further, though, advocating services that can be automatically run on standardised, repeatable clusters.

“Resource management and service discovery helps you to separate the infrastructure from the services,” he said, citing tools like Apache Mesos which can manage clusters of machines and run services as containers. “What we are seeing lately (and Mesophere is a good example) is building environments where the payload is a container which is dynamically scheduled to that machine,” he explained.

Separating the data and the services in this way makes it possible to control where and how the data is stored programmatically, typically via APIs. This underpins the cloud computing concept, and can free up companies to begin moving their workloads around in their on-premise infrastructure based on new business requirements, or even farming parts of them out to third party providers in a hybrid cloud arrangement. Until they build this level of abstraction into their systems, that fluidity will be difficult to achieve.

Data-aware storage

Some companies are taking the idea of abstracting services from infrastructure to the next level, with data-aware storage. This makes the storage layer more aware of the nature of the data it is holding, providing administrators with still more flexibility. Last summer, analyst firm Taneja group held a webinarabout data-aware storage, and highlighted three characteristics. It captures attributes of the storage stored on the device, looking for contextual patterns to identify sensitive data, for example, or singling out files meeting certain parameters. The storage layer might know that a file is low-latency video, and understand where best to store it based on that information.

Data-aware storage must also have real-time analytics, so that it can provide useful information about how the data is being used, ideally by particular applications. If it senses that an application has increased its IOPS significantly, it might be able to use this information, along with its understanding of the data, to solve the problem early on. Finally, it must provide services to help manage that data, such as balancing quality of service across different workloads, and it should make these available programmatically via APIs, which again highlights the ‘infrastructure as software’ idea.

El Reg has discussed some companies in the data-aware storage layer space before, such as Primary Data, which launched its DataSphere product last year. The company uses a policy engine to manage data according to objectives set by the business (it calls them ‘service level objectives’). This theoretically makes it possible for IT admins to encode some of those new business requirements as they emerge and have the storage layer do all the work behind the scenes.

This flavour of data-aware storage is also being touted as a silo-busting technology because it enables data to be moved across different storage types, ranging from NAS to fibre-connected SANs and direct-attached storage.

Consider data formats

Making storage infrastructure aware of the data it’s storing enables IT administrators to set and change policies over time, and can also make migration easier. It’s a relatively new concept in data architectures, though, and given that companies are still struggling with marketing terms such as software-defined storage, adoption will be slow. In the meantime, there’s another layer of your data architecture to consider – the application storage format. For years, transactional applications have relied mostly on relational formats, but non-relational formats are on the rise.

NoSQL databases represent a way to store and reference data in non-relational formats. In the context of futureproofing your data architecture against future requirements, their primary benefit is flexibility.

Relational databases are rigid, using schemas that are collections of tables containing columns that must be decided up front. That makes it difficult to change data structures on the fly, said Bob Weiderhold, CEO of NoSQL firm Couchbase. Schemas may have to be rebuilt entirely when changes need to be made.

“In many companies, that can take six months,” he said. “With NoSQL, it’s schema-less.”

Instead of a pre-defined schema, NoSQL data formats allow inferred schemas, which simply acknowledge new attributes as they’re added to records, he explained. You want to store the name of a customer’s parakeet? Just start adding that key value pair to records as you create them. No redesigning of tables are necessary, because there are no tables. Various NoSQL approaches do this in different ways. Couchbase stores records in nested documents, typically using JSON to represent different fields and values).

Define first, join later

What this means is that companies don’t have to be as structured about their data models, said Jon Cooke, head of data science at GFT, which designs data architectures for financial services clients.

“What we’re helping banks do is create flexible data platforms that can respond to new events in the marketplace without having to re-architect the entire data model,” he said. “What you should be doing is defining your data elements, your business entities, in isolation without forming some sort of spaghetti diagram. Join them when you actually ask the business questions.”

NoSQL may not be right for everyone. Typically, it lacks the ACID transaction capabilities necessary for drawing together lots of separate systems in a single transaction.

“Increasingly we’re moving more towards that but today you wouldn’t run your financial transactions per se on this platform,” said Sean O’Dowd, global financial services director at MapR, which provides Hadoop solutions. But then, there are lots of things that don’t need to be ACID-compliant, in finance and everywhere else.

If your business requirements and demands on your data are rapidly evolving, or if your data volumes are rapidly growing, then it may be worth looking at NoSQL as one in a range of future-proofing tools for your data architecture.

Agility rests on the ability to rapidly reconfigure the storage and structure of data. To properly set the scene for that, companies should be exploring not just virtualization but also programmatic allocation of resources. Even with all that in place, you may not be able to pivot instantly to cope with rapidly-evolving business conditions. Still, you’ll be moving in the right direction.

Original article – http://www.theregister.co.uk/2016/01/28/data_migration_strategies/