Evaluating Databases for Cloud Applications

Understanding what's required and how to be successful

Table of Contents

Introduction

For any business that wants to successfully compete in today's digital economy, it is not a question of *if* but rather *how much* of their business will be done with cloud applications.

A cloud application is one with many endpoints including browsers, mobile devices, and/or machines that are geographically distributed. The application is intensely transactional (high velocity reads and/or writes), always available, and instantaneously responsive no matter the number of users or machines using the application. In short, customers need cloud applications to deliver real-time value at epic scale.

The move towards cloud application use vs. traditional application delivery models is unmistakable with Gartner Group noting a current trend in the need for systems to support the "development of new digital — in most cases end-consumer facing — applications that need to be able to scale and transform in a similar agile and dynamic fashion as consumer-oriented social media applications."[1]

Not surprisingly, as modern businesses move more towards cloud computing and its application model, there is a noticeable impact on the software being used to run the business. This fact is highlighted in a Goldman Sachs technology report on the effect that cloud computing is having on legacy database vendors like Oracle and Microsoft: "Again [with] this survey, spend on databases *is expected to have the biggest negative impact over the next two years*, as more people expect to spend less on Oracle (13% more, *52% less*) and Microsoft (9% more, *38% less*)."[2]

While statistics such as these show how businesses are looking beyond traditional relational database management system (RDBMS) technology for a database that better tackles cloud applications, what isn't as obvious is the fact that these same enterprises wrestle with determining and fulfilling the database requirements needed for their cloud applications.

This paper discusses the general characteristics of a cloud application and then examines the key database requirements that make cloud applications successful.

The Multi-Faceted Cloud Application

Today's cloud applications are far from being one-dimensional in nature; instead they are intensely multi-faceted and include a variety of components, each with its own set of data requirements.

For example, a modern retail cloud application now includes various modules such as product catalogs, user profile management, fraud detection, recommendation engine, shopping cart, clickstream/log analysis, and others. Behind each one of these application units are data needs that, if unmet, dramatically and negatively impact the user's overall experience.

---

[1] Gregor Petri, Gartner Group, "Three Factors Will Continue to Impact Enterprise Cloud Playbooks:, September 2015.

[2] Goldman Sachs Equity Research, "IT Spending Survey: Sentiment moderates into 2016", January 2016, Pg. 15.

Given how legacy and Web 1.0 application's data-handling functioned in the past, the data demands of today's cloud applications appear intimidating. For instance, a recommendation engine must track every move a user makes on a web or mobile application in a transactional manner, analyze those incoming motions in real-time, search a user's historical behavior, locate potential recommendations, and circle back to the cloud application's user interface with those recommendations. If that weren't difficult enough, the whole sequence must happen within stringent performance and data security service level agreements.[3]

All of this data-related activity involves numerous pieces of data management technology that must be blended together to act as one – transactional, analytical, search, in-memory, and other similar engines – each of which must also run in a highly performant way without impacting the other's speed. In other words, the system must be functionally cohesive within a single architecture.

Lastly, the cloud application's data model is increasingly becoming multi-faceted, which means a one-size-fits-all approach will no longer suffice. Certain components of a cloud application may require a tabular data model while another might need a graph engine to function properly. Add to that the desire that each model be stored in the same persistent datastore in order to avoid the headaches associated with sharding a system and trying to keep dependent data sets in sync across multiple models.

The Promises vs. the Reality of a Cloud Application's Database

The data requirements of a cloud application can certainly be challenging. One should exercise caution when researching a database solution because there are many vendors who claim to have solved the various database issues facing cloud applications, but when examined closely, most don't fully deliver.

Anyone who has read material on cloud computing and its applications knows that a lot of promises are made, especially where data management is concerned. These promises are oftentimes prefaced with the word 'transparent'. Transparent elasticity, transparent scalability, transparent redundancy, transparent manageability, and so on.

The truth is there is no magic behind a cloud application where its database is concerned. While a software or cloud vendor may attempt to mask the architectural limitations of a legacy RDBMS with their own custom framework that tries its best to meet the requirements of a cloud application, it will fall short almost every time in terms of performance, scalability, complexity, or all three.

If a cloud application's database is going to achieve transparent <fill in the buzzword> it has to start with the right architectural foundation, one that was built from the ground up to handle a cloud application.

Database Requirements for a Cloud Application

While each cloud application is unique, there is a foundational set of technical database requirements that must be met for the business to successfully compete in its market. These include the database being:

- Continuously available
- Geographically distributed
- Operationally low latency
- Linearly scalable

---

[3] For examples of this type of processing, see Gartner's report on hybrid transaction/analytical processing (HTAP): https://www.gartner.com/document/2657815.

- Immediately decisive
- Functionally cohesive
- Operationally mature

Legacy RDBMS's will struggle with all but the last requirement in general, and even then, they will not be able to exhibit operational maturity across the other criteria. With relational technology not being a good fit, architects and IT professionals are looking to distributed database platforms like Apache Cassandra™ and DataStax Enterprise to satisfy their cloud application's requirements.

The sections that follow examine each database requirement of a cloud application in more detail and discuss how DataStax Enterprise (DSE) – a database platform built on proven open source technology that delivers the functionality, tooling, certification, and support required by today's enterprises to power cloud applications – addresses each one.

Continuously Available

A cloud application's ability to remain continuously available even in the face of the most devastating infrastructure outages can be tied directly back to both high customer satisfaction and revenue for the enterprise. Depending on the business, downtime for a cloud application either equates to tens of thousands of dollars lost per minute or, in the case of some financial institutions, tens of millions of dollars. What oftentimes is harder to measure is the loss of customer trust and reliance when a cloud application becomes unavailable.

While high availability may be achieved via legacy RDBMS failover procedures, true continuous availability for a cloud application is another matter and requires an architecture that was designed to never allow single points of failure.

Unlike traditional RDBMS's and some NoSQL databases, DataStax Enterprise builds on the core architecture of Apache Cassandra, which sports a masterless design where every node in a database cluster operates independently with respect to database operations. There is no concept of a master node or other data layer needed to accomplish the redundancy in compute and data resources required to satisfy true 100% availability.
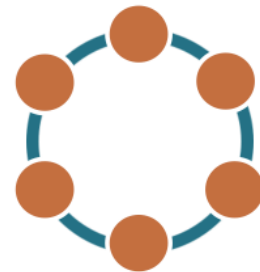
DataStax Enterprise is designed to automatically write data across multiple nodes in a database cluster and maintain as many copies of the data as the user desires. This can involve single or multiple geographical locations, as well as different cloud availability zones in one or multiple cloud providers.

Geographically Distributed

A cloud application presents a follow-you-everywhere data management challenge. A user wants access to their requested data 24x7 and expects it to be delivered fast regardless of their location.

This is sometimes referred to as *active everywhere* or *location independence*: the ability to deliver active data everywhere a customer needs it. Beyond the obvious benefits, having full geographically distributed capabilities built into a cloud application allows the business to more quickly expand into new markets and also simplifies legal compliance with geographic data distribution (or data sovereignty) laws.
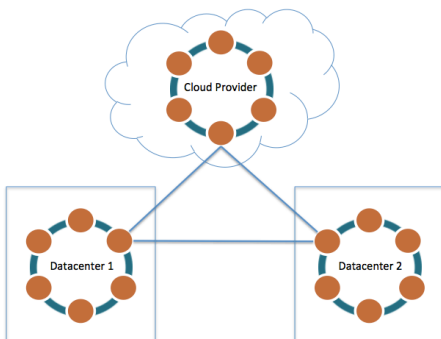
While it is widely acknowledged that the RDBMS architecture isn't well suited for active-everywhere use cases, Cassandra's design was built from the ground up to handle widely distributed data that is accessed and synchronized across multiple data centers and cloud providers. Multiple copies of data can be replicated across as many locations as needed and should a specific locale go offline, client requests are automatically re-routed to the other, nearest live location.

By leveraging Cassandra's design, DataStax Enterprise also fully supports hybrid cloud environments where an enterprise splits its design across on-premise datacenters and one or more cloud providers. A single DataStax Enterprise cluster and database can span each location, with the database appearing to the application as a single datastore.

Operationally Low Latency

A cloud application user expects the response times (i.e. the speed at which they interact with the system and data) to be predictable no matter where they are located. They also want the wait times (or latency) to be as short as possible for both read and write operations.

For a cloud application, a scale-up design is prohibitive not only because it doesn't support a geographically diverse user base, but also because it lacks the scale out, divide-and-conquer computing capabilities needed to keep performance at expected levels.

DataStax Enterprise delivers short wait times and predictable performance by utilizing a number of built-in mechanisms. For example, one of Cassandra's hallmarks is its extreme write and read performance made possible by its log-based write engine and its node-aware routing that is able to automatically direct client requests to the least busy and most closely located nodes.

In DataStax Enterprise, built-in analytics and search functionality allows analyze-and-search-in-place functionality for an operational database, which avoids the headaches and performance slowdowns associated with sharding a system across multiple RDBMS and NoSQL providers. Moreover, workload management is native to the architecture so transactional, analytical, and search workloads are able to execute in a manner where there is no competition for either data or compute resources, which also contributes to high performance.

Further, DataStax Enterprise provides an in-memory option so that frequently read data enjoys the fastest possible response times.

Lastly, DataStax Enterprise supplies automated management services that ensure the system operates efficiently without the involvement of the operations team. Through its Performance and Best Practice Services, key diagnostic metrics are automatically and constantly collected, and potential problems are proactively identified along with recommendations for rectifying them.

All of DataStax Enterprise's automatic management services may be handled via the command line or visually through DataStax OpsCenter, which is a web-based management and monitoring solution that can be utilized from any laptop or mobile device.

Linearly Scalable

It's not uncommon to hear developers of successful cloud applications say how they were taken off guard by how quickly their data and number of application users grew. Part of a cloud application's success where predictable performance is concerned is its ability to scale in a linear fashion so that the end result is a future-proofing of the application's response times – regardless of the data volume or number of users, the application continues to perform as required.

A related point to note about scalability is that it should not come at the expense of availability. The ability to increase or decrease compute capacity should be accomplished in an online fashion so that no application downtime is experienced.

Where Cassandra is concerned, linear scalability is something that has been demonstrated by numerous DataStax customers. Because scale is baked into Cassandra's core architecture, additional capacity can be added to DataStax Enterprise in an online fashion with no disruption being experienced by application users. As additional nodes come online, Cassandra evenly and automatically redistributes data so incremental growth is accomplished with little to no effort.

Where planning for the future is concerned, DataStax Enterprise supplies its Capacity Service, which lets customers forecast what additional data or compute resources will be needed down the road. For example, the Capacity Service can predict when a cluster will run out of disk space or when current RAM allocations will be exhausted. Such functionality aids both IT staff and their management as the information provided by the Capacity Service allows them to easily plan and budget for future needs.

Immediately Decisive

A cloud application needs to handle multiple types of data, varied data streams and different workloads in a simultaneous fashion. The data complexity involved is typically required to enable the application to make immediate decisions and provide insightful responses back to a web or mobile user at transactional speed.

While an enterprise architect can traverse down the system sharding route of dividing workloads and DBMS providers onto separate systems, this approach normally leads to greater expense, overhead, complication, and slower time to market for the application. An integrated, single DBMS platform is preferred if it is capable of delivering real time decisions in a performant way.

DataStax Enterprise does just this by providing a full range of data manipulation and mixed workload capabilities in its tightly integrated platform. The platform provides full data ingestion and interrogation across transactional, analytical, and search components so that no matter where data is added, modified, or deleted, it is synchronized everywhere and automatically maintained by the database.

Functionally Cohesive

No architect or operator wants to tackle a database platform that consists of a dizzying array of technologies that are clumsily bolted together. Instead, they desire an elegant, functionally cohesive database that smartly weaves together the right technologies in a way where they work together as one.

DSE does exactly this by tackling transactional, analytical, search, and in-memory workloads in a single system that also supports multiple data models. By doing this, DSE simplifies the cloud application's data dependencies and flattens the database architecture into a single layer. The end result is a platform that delivers less expense, complexity, and overhead, plus it shortens the application's time to market.

DataStax Enterprise also provides the ability to blend together transactional, analytical, and search commands into single requests so that what previously took three different commands on three different systems can now be more quickly accomplished in one. For example, a DSE Analytics query using Apache Spark™ can utilize DSE Search syntax (delivered via Apache Solr™) to create one statement sent against the database that performs both an analytics and search task at the same time. Having such capabilities in the database delivers exactly the type of hybrid transactional-analytical-search functionality and decision-making abilities that cloud applications require.

In addition, as mentioned previously, cloud applications are increasingly using multiple data models for the various components that make up the application as a whole. The use of an *adaptive data management* framework might practically equate to a tabular model being used for a couple of the application components while a graph model may be used for others.

However, while the use of innovative functionality like adaptive data management delivers very real benefits, the cloud application's developers will want it to be operationally easy to manage and come with standard best practice functionality that includes strong security, comprehensive workload coverage (i.e. run all types of workloads – transactions, analytics, search – regardless of the underlying data model), and excellent commercial software support. All of this is found in DSE.

Operationally Mature

The companies behind cloud applications face the dilemma of needing to utilize innovative new technologies that give them a competitive edge in the market, while at the same time needing those technologies to exhibit the same maturity and dependability of legacy software that have been around for decades.

A solution like DataStax Enterprise demystifies the usage of these powerful, new database technologies while also providing enterprise-ready data management for cloud applications. This ability showcases an important distinction between the open source components that comprise DSE and the platform itself.

For example, Apache Cassandra is widely known to be an innovative, fast-moving NoSQL database project driven by the open source community. On the other hand, DataStax Enterprise is a database product built on that proven open source technology (and others) that delivers the functionality, tooling, certification, training, and support required by today's enterprises to power their cloud applications.

Customer Examples

Below are two examples of customers using DataStax Enterprise for their cloud applications. For other customer case studies, visit the Customers page on DataStax.com.

eBay

eBay is the world's largest online marketplace, enabling the buying and selling of practically anything. eBay currently serves over 162 million active users and has 800M+ live listings.

One of the keys to eBay's extraordinary success is its ability to turn the enormous volumes of data it generates into useful insights that its customers can glean directly from the pages they frequent. To accommodate eBay's explosive data growth (its data centers perform billions of reads and writes each day) and the increasing demand to process data at blistering speeds, eBay needed a solution that did not have the typical bottlenecks, scalability issues and transactional constraints associated with common relational database approaches. The company also needed to perform rapid analysis on a broad assortment of the structured and unstructured data it captured.

In order to address these challenges DataStax Enterprise with integrated search and real time analytics was deployed across multiple data centers. The end result allows eBay to more cost effectively process massive amounts of data at very high velocities, and achieve far more than they were able to with the higher-cost

proprietary system they had been using. eBay is managing a sizable portion of its data center needs—250TBs+ of storage—in DataStax Enterprise clusters.

ProtectWise

ProtectWise, a provider of the industry's first Cloud Network DVR for complete visibility and detection of enterprise threats and accelerated incident response, has revolutionized the way enterprise network security is handled, giving security professionals a real-time picture of what is happening on their network even if they might have missed the initial signs of intrusion. They do this by deploying clusters of lightweight sensors at points on a network that record network data and then ship it securely to Amazon's cloud where the ProtectWise platform performs a variety of analyses to identify out of the ordinary patterns and anomalies.

When ProtectWise embarked on this journey to take enterprise network security to the cloud, they realized that they needed to build their application on a database platform that could not only handle the enormous volumes of high velocity, streaming, time-series data, but they needed to do so without taking a hit on performance and availability. ProtectWise realized that to meet these demands, they needed to shift their database approach to support the solution they were looking to build. ProtectWise had some previous experience with Apache Cassandra and felt that DataStax Enterprise was the distributed database that would allow them to store and manage time-series data in a way that legacy systems just couldn't do. With DataStax Enterprise, ProtectWise is handling millions of transactions and writes per second and without downtime or data loss.

Conclusions

With evidence consistently indicating the rise in digital commerce and consumerism and the demise of companies that fail to embrace the Internet and cloud economy, the importance that a business puts on its cloud applications cannot be overstated. Because the heart of any system is its database, a cloud application cannot be successful without the proper database infrastructure being in place to support it.

While legacy database technology is unable to meet the key criteria needed for today's cloud applications, DataStax Enterprise delivers a comprehensive data management layer with its unique always-on architecture that accelerates the ability of enterprises, government agencies, and systems integrators to power their exploding number of cloud applications. DSE well serves cloud applications that require data distribution across datacenters and clouds, through the use of its secure, operationally simple platform that is built on Apache Cassandra.

For more information about DataStax Enterprise and downloads of DataStax software, visit www.datastax.com.

About DataStax

DataStax, the leading provider of database software for cloud applications, accelerates the ability of enterprises, government agencies, and systems integrators to power the exploding number of cloud applications that require data distribution across datacenters and clouds, by using our secure, operationally simple platform built on Apache Cassandra™.

With more than 500 customers in over 50 countries, DataStax is the database technology of choice for the world's most innovative companies, such as Netflix, Safeway, ING, Adobe, Intuit, Target and eBay. Based in Santa Clara, Calif., DataStax is backed by industry-leading investors including Comcast Ventures, Crosslink Capital, Lightspeed Venture Partners, Kleiner Perkins Caufield & Byers, Meritech Capital, Premji Invest and Scale Venture Partners. For more information, visit DataStax.com or follow us on @DataStax.