WHITE PAPER

Driving Forces for Taking Analytics into the Cloud

Stephen Brobst, Henry Elliot, Frank Freund,
Jake Le, Jay Suryamurthy

04.22 / EB9961 / CLOUD / WHITE PAPER

teradata.

Contents

Introduction

More and more analytic ecosystems are making their way into the cloud. The opportunities for increased agility and the ability to reap benefits of extreme elasticity are compelling. When making the leap up into the cloud, it is important to do so with both eyes wide open. The physics of the infrastructure deployment will be different from what organizations are used to with on-premises deployments. Moving data, storing data, and accessing data will come with a different set of physics when in the cloud as compared with traditional deployment of analytical ecosystems. The laws of physics should not be considered in isolation; the laws of economics are also critical in understanding the overall effectiveness of analytics in the cloud.

This white paper provides a framework for evaluating both the physics and the economics of different options for ecosystem deployment in the cloud.

Laws of Physics

There is a frequently occurring attitude of "I don't have to care" about the underlying physics of the platform infrastructure when considering the deployment of workloads to a cloud infrastructure. Nothing could be further from the truth. While it is true that deploying to the cloud relieves you from the responsibility of building out the platform infrastructure to support your workload, it does not mean that you shouldn't care how the platform infrastructure is configured and operated.

Think of acquiring a car service for transportation from point A to point B. For a short journey, you would probably be willing to hop into any reasonable-looking cab that you could flag down on a street corner. But for a longer journey, or one where time to destination is very important, you would certainly want to make a choice between different options based on factors such as safety, performance, and comfort. Not all vehicles and not all operators of those vehicles are created equally. Specifications of the vehicle and quality of driving matter.

Similarly, not all cloud infrastructures are created equally. The plumbing matters. And how the cloud infrastructure is operated also matters. Most public cloud infrastructure comes with multiple choices for deployment. The underlying configuration might be with direct attached storage or network attached storage. Storage might be spinning hard disk drives (HDD) or might be solid state disk (SSD) drives. There are different amounts of CPU and memory that can be configured into platform infrastructure to be used for workload execution. All of these (and more) can vary widely within a single cloud vendor's offerings and across different cloud vendor offerings. These factors can and will ultimately impact the performance and reliability of the infrastructure that you use to deploy your workload. Physics matters.

It is important to understand the characteristics of workloads and the platform configurations that you are considering for deployment. A typical multi-user workload for analytics is characterized by random I/O patterns of 80% read and 20% write. For such workloads, the use of solid state disk storage will typically deliver close to a 50 times better I/O bandwidth versus electromechanical disk drives.

This performance difference is a moving target based on the latest and greatest in both SSD and HDD technologies. Other factors have influence as well. With network attached storage, it may be the performance limitation when configured with SSD is not the storage devices, but with the network.

Moreover, we observe that the service level specifications for network bandwidth delivered by virtual machines vary across public cloud platform providers. For example, in the case of AWS, the network specification of bandwidth delivered is a unidirectional limit of 1250 MB/second. However, when reads and writes are happening at the same time with an equal ratio, there is a 50% "bonus" that is achieved whereby actual bandwidth delivered is typically 150% of the published specification. In contrast, the Microsoft® Azure limit of 960 MB/s is bidirectional and 50-50 read/write ratio will achieve 100% of the 960MB/s limit (as will any combination of read/write). Of course, as the cloud platform providers evolve, their offerings regarding these numbers are subject to change.

In most cloud platforms, the storage is presented as Just a Bunch Of Disks (JBOD) and the underlying protection mechanism is opaque. The cloud vendor will provide Mean Time Between Failure (MTBF) specifications and I/O processing rates that can be used to understand availability and performance characteristics of the platform.

Direct attached storage I/O delivery rates are more likely to be limited by the number of disk drives and the aggregate of the per drive I/O performance capabilities up to the limit of the I/O adapter (e.g., approximately 3.5GB/s on an AWS i2.8xl instance with JBOD). How the underlying storage is protected from data loss also impacts performance. For example, RAID-1 carries a 100% I/O overhead for writes. However, smart software can extract extra performance for read-oriented workloads from a RAID-1 configuration by spreading I/Os across the primary and mirror copies of data. The algorithms in use for compression and decompression can also impact storage requirements, CPU consumption, and performance for analytics.

To assess network performance, it is important to understand not only the raw specifications of network infrastructure, but also how the software interfaces to the network. For example, some cloud implementations make use of accelerated networking software, enabling single root I/O virtualization (SR-IOV) as an extension of the PCIe specification that allows an adapter to present a virtual representation of itself directly to the virtual machine. Benchmark testing has shown that this approach yields a significant performance improvement versus a less optimized implementation when interfacing to the network via Hypervisor calls.

While there are definitely circumstances where cost savings will be significant for some use case scenarios, the true value of cloud deployment is more likely to be found in the competitive advantage associated with greater agility.

Understanding the point-to-point latency and bandwidth delivered by the network is important for assessing expected performance. Different workloads will have different sensitivities to network performance characteristics. Teradata Vantage™, for example, is less sensitive to network latency and (relatively) more sensitive to network bandwidth for delivering high-performance analytics at large scale.

The ability to consume the data provided by the storage and network infrastructure depends on the amount of memory and CPU available for workload execution. It is not straightforward to compare CPU processing power across cloud platforms because custom processors are often deployed. SPECint ratings extrapolated to similar processors with known benchmarks are likely the best that one can do in the form of a paper exercise.

Memory sizes can also vary significantly across cloud instance configurations. Bigger is better when it comes to memory, but there are declining returns relative to cost as sizes become very large (e.g., more than 30% of data size).

Figure 1 depicts a hypothetical comparison between a cloud-based deployment of analytics versus a custom configured appliance designed specifically for high-performance analytics. Performance penalties that one might expect from cloud deployment are annotated as a percentage of total performance sacrificed according to differences in design choices in comparison to a purpose-built appliance.

Among all components of an analytic ecosystem, software has the most influence on the efficiency of data consumption. It is well known that better algorithms for processing data will speed up workloads much more effectively than better hardware configurations—often by orders of magnitude for large data sets. The difference between an efficient and an inefficient join algorithm can easily be the difference between linear resource consumption and exponential resource consumption.

On small volumes of data and fast machines you might not notice the difference. However, when data volumes explode the difference between linear and exponential resource consumption is huge.

With small data sets that fit 100% into memory, it is relatively straightforward to create in-memory data structures and simple algorithms to join data. But when data sets get larger than what can cost effectively fit into memory, the challenge grows exponentially. The differences in file system design, optimizer capabilities, parallelization capabilities, and corresponding join algorithms differentiate the toy databases from high-performance databases. Further capabilities in the areas of resource management and workload prioritization, advanced indexing, and failover all relate directly to software maturity and readiness for production workloads.

Immutable Laws of Computing

Laws of Physics

- IOPs
- I/O MB/s
- CPU
- Memory
- Network speed
- Platform software capabilities

Laws of Economics

- Purchase price
- TCO
- Labor
- Data movement
- Licenses
- Maintenance

For all of these stated reasons, it is critical to understand both the hardware and software capabilities of the platform for deploying your analytic workload. It is not enough to take cloud vendor service level promises at face value. If you read the fine print, you will soon understand that the service level commitments from major cloud vendors cannot be contractually enforced in any meaningful way. Their lawyers are bigger than your lawyers and the fine print in the contracts is rarely negotiable. It is essential to understand and have confidence in the ability of your chosen software and hardware stack to deliver on the service levels that you have committed to your business knowledge workers. Robust scalability and elasticity are critical for meeting these SLAs.

For a large enterprise, horizontal scaling is required. Once processing is allocated across multiple virtual machines, which themselves are spread across multiple physical servers, constraints imposed by the underlying physics of the platform will tend to reveal themselves. For example, network latency and bandwidth characteristics, robustness of the underlying I/O architecture, and management of processing locality will influence scalability of analytics in the cloud.

Similarly, the ability to expand and contract analytic processing capability on-demand can differentiate cloud platforms. The speed and efficiency by which resources can be applied to analytical workloads distinguishes solutions in their ability to effectively handle peak processing periods. Elasticity can play a role when cloud bursting workloads are shifted from on-premises platforms to expand processing capability in a hybrid execution environment. Elasticity is also essential for quickly scaling up processing capability in

teradata.

circumstances where the cloud infrastructure takes over in a disaster recovery situation.

Another aspect of the physics related to workload deployment is data gravity. Where data is born influences where analytics are best deployed. There is a "gravitational pull" of analytical processing to where large data already exists. In the past, most data for an enterprise was created in the data center of that enterprise. This means that the gravitational pull has historically influenced analytics to also be in the corporate data center. To move the data elsewhere has performance and network costs. For this reason, it is more effective to analyze data in the data center environment where it is created. However, as more source systems move into the cloud, the associated gravitational shift of data will pull analytics into the cloud as well.

It is not always easy to determine which cloud infrastructure is most suitable for deploying analytics. Many organizations run workloads on multiple public clouds. For example, an enterprise may operate a combination of SaaS solutions (e.g., Saleforce.com), packaged solutions (e.g., Microsoft Dynamics on Azure), and custom-built solutions (e.g., on AWS). This increasingly common scenario generates multiple directions of gravitational pull.

The best place to deploy a workload is a balance of two factors. First is the robustness of a vendor's cloud solution for an analytic ecosystem. Second is the locality of the majority of data created within the enterprise. Note that it will sometimes be more effective to push data away from the gravitational pull in order to deploy on a better engineered analytic platform infrastructure. All factors must be considered.

Laws of Economics

Having very high performance and very high availability will not count for much if the economics do not make sense. There is widespread belief that migrating from on-premises infrastructure to a cloud platform will save money. For 95% or more of all businesses in the world, there is absolutely an improvement of performance, reliability, and cost when migrating onto a cloud platform. These improvements are most obvious for small enterprises with less than 100 employees—which also represent the majority of enterprises worldwide.

These businesses clearly do not have the scale to cost effectively acquire and manage compute infrastructure. A small business is much better off leveraging the capabilities of Amazon®, Microsoft, or Google® for cloud compute infrastructure rather than trying to build out its own on-premises capability in this area.

Large enterprises operate in a very different economic framework. A big company will usually have its own economies of scale for acquiring technology, building out data centers, and managing infrastructure. The economic benefits of public cloud are not always so obvious for these mega major enterprises. It is very difficult to be as efficient or exploit scale as well as Amazon, Microsoft, or Google. However, it may be the case that advantages of data gravity and internal cost structures versus acquired services will allow a large enterprise to be more effective than outsourcing these capabilities.

In fact, some (large) enterprises have been building out their own private cloud infrastructures rather than using public cloud infrastructure. This approach, while it may work well for some, can easily be underestimated in terms of its true cost. Hiring and retaining the skill sets to operate a private cloud infrastructure is no small undertaking. To be effective in private cloud deployment at scale, automation becomes critical. Manual processes lack scalability and cost effectiveness.

The best-practice public cloud providers have invested significant effort and capital to reduce total cost of ownership (TCO) to a bare minimum via automation and attention to resource management and security controls. Moreover, the public cloud option facilitates a financial model with pay-as-you-go expense as opposed to the private cloud option with significant upfront capital costs for establishing the environment.

Pay-as-you-go models, however, can vary significantly across cloud providers. The differences are most likely in terms of the cost of resources, the granularity at which the metering is performed, and in the minimum length of commitment that must be made in order to be granted specific price points. For example, three-year commitments to minimum resource consumption will yield lower cost per terabytes of storage and/or units of I/O and compute consumption as compared to a by-the-hour commitment.

The economics of getting data in and out of a public cloud must also be considered. It is often the case that costs for data ingest and egress are asymmetric. Public cloud providers will often "subsidize" the cost of bringing data into their cloud platforms in order to capture additional workloads. To make it easy and efficient to bring data into their clouds, most public cloud platforms provide tools for compression, encryption, and data movement.

The methods and tools for initial bootstrap loading of large scale content into cloud environments will often be different than what will be utilized for ongoing incremental acquisition of data. Be aware that the economics of getting data out of a public cloud are very often much higher than getting data into the cloud.

Just like many mutual funds, there may be back-end loading of costs for extracting content even though there was no charge for bringing data into the cloud.

It is essential to not get overly locked into any particular cloud provider. Any robust economic model for cloud deployment must also consider the eventual cost of exiting the vendor's cloud service. Eventually, a different vendor will offer a cloud service that provides a better economic model, more desirable functionality, or better service levels. You do not want to be trapped with the incumbent cloud provider by high exit costs. Creating a realistic cost model for migrating out of one cloud and into another should be factored into the TCO for any cloud vendor. Availability of ecosystem tools and portability of software licenses across cloud platforms are critical to understand when assessing barriers to exit from one cloud provider to another.

Price per terabyte of storage is one of the least useful metrics to rely on when assessing the economics for analytic deployment across various cloud platforms. Making storage price primary is the equivalent of buying a house based solely on the price per square meter without paying attention to the quality of the house or its location. A prudent buyer would not make such an important investment using such a naive metric. Similarly, an evaluation of cloud platforms also needs to consider the quality of the infrastructure, ability to deliver service levels, and availability of tools for creating a robust analytic ecosystem.

Value comes from querying data, not from storing data. To align costs with value creation, the metrics need to be more sophisticated than just storage costs. If querying the data is not a requirement, then storing the data to /dev/null/ would be a lot more efficient than any possible cloud solution—but this is not going to meet the criteria for deploying analytics within an enterprise. The

Cloud Design Pattern Performance Trade-offs

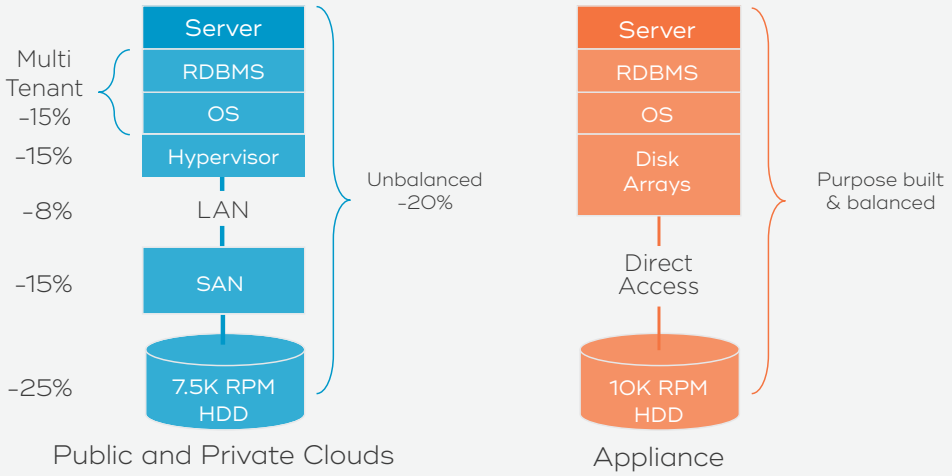

Figure 1. Design Pattern Performance Trade-offs

most useful metric for evaluating cloud platforms for analytic purposes is not cost per terabyte of storage, but rather cost per query.

Pure Play Cloud versus Hybrid Cloud

There are some brave souls jumping into the cloud with both feet. Not only are they migrating the analytic ecosystem—but also billing systems, customer care, general ledgers, and so on. This "all in" approach facilitates getting to the value of cloud deployment as quickly as possible. The pure play cloud approach also has the potential to exploit data gravity if cloud alignment is obtained between operational source systems that provide data and the analytic ecosystems that will create value from the data. However, this is tricky because even if an organization is 100% in the cloud, it does not mean that a single cloud infrastructure will meet all of the demands of the enterprise. Most organizations will deploy across multiple clouds. Gartner predicts that by 2025, more than 90% of enterprises will pursue a multi-cloud infrastructure and platform strategy.

Aligning the analytic ecosystem to whichever cloud platform has the most data gravity can yield some potential benefits in getting data to where it needs to be in a cost-effective way. However, data gravity advantages also have to be weighed against the ability of the cloud platform to efficiently handle analytic workloads.

For many organizations, a phased transition into the cloud is a more effective strategy. Project risk and migration effort are often better managed in phases. One of the most common strategies for migrating into the cloud is to start with the implementation of disaster recovery in the target cloud platform. Disaster recovery has particularly attractive economics when implemented using cloud infrastructure because a disaster recovery system is something that you need to have and yet hope never to use.

Cloud infrastructure has the property that you only pay for what you use; a utility model. In contrast, a disaster recovery system implemented on-premises is a large capital expense as well as a significant operating expense. Of course for a cloud-implemented disaster recovery system, you must pay for the data to be stored—but you only pay for the CPU and I/O that is actually consumed. Typically, the ingestion of data to keep the disaster recovery copy of an analytics repository synchronized will be less than 15% of the total workload on the production system. This means that as long as you are not going live with production on the disaster recovery

teradata.

solution, you pay only a fraction of the costs that would have been paid for a traditionally implemented on-premises hot standby system.

The economics of a disaster recovery solution that leverages cloud infrastructure can be quite compelling.

> Disaster recovery has particularly attractive economics using cloud infrastructure because a disaster recovery system is something that you need to have and yet hope never to use.

Once a disaster recovery system is in place, it can provide additional benefits. It is possible to "cloud burst" read only workloads to offload peak workloads from the primary platform. This approach obviates the need to oversize the primary production system to be big enough to handle spikes in workloads that occur rarely (e.g., end-of-year processing for regulatory reporting). The result is a solution that allows the primary system to be right-sized to a lower cost. The primary production system needs only to handle the steady state workload. Occasional surges in demand for capacity can be directed to the disaster recovery solution in the cloud.

The hybrid approach with an on-premises system optimized for day-to-day analytic processing combined with the cloud-based solution for handling highly elastic workload requirements yields the best of both worlds. Additional scenarios where the elasticity provided by cloud deployment is quite advantageous include test and development systems as well as data labs. Test, development, and data lab systems all have the characteristic that they are needed for a particular duration of time, and then they are no longer needed.

In many organizations that deploy on-premises test and development systems, the capacity of these systems is woefully underutilized for most of the time. However, during certain periods of intense test and development activity, their capacity tends to be woefully under-sized. It is exactly this type of situation where the use of private or public cloud deployment is hugely advantageous.

The flexibility to, on-demand, allocate exactly the resources required for a heavy period of test and/or development and then have these resources de-allocated just as quickly when they are no longer needed is very attractive. An added advantage for development and testing is the ability to use next generation versions of software for feature exploitation and/or testing before the software has been approved for deployment in a highly governed on-premises environment.

Similarly, data labs are used by data scientists to experiment with new data sources and/or new algorithmic approaches to uncover value in data. Cloud implementation has two significant advantages for data lab deployment. First is that elasticity of demand for resources in the projects undertaken by data scientists is typically very high. The flexibility in resource allocation in the cloud provides the perfect working model for a data scientist who cannot easily anticipate what the resource requirements will be for working with the next unknown data set or algorithm.

The second, arguably even more important, advantage of cloud deployment for data science work is agility. The internal bureaucracy for bringing in a new piece of software on-premises for most organizations creates great friction in data discovery. Exploration often requires deployment of an innovative data platform or maybe a new algorithmic software library; following enterprise standards for bringing in new technology slows down progress. In a public cloud environment, the desired tools are much more likely to be easily accessible, without going through onerous IT architecture governance and standards committees, than in an on-premises deployment scenario.

Neither test, development, nor data lab workloads are particularly sensitive to performance or availability service levels. However, all of these environments benefit significantly from the elasticity provided by a cloud environment. These use cases also benefit from the greater agility in deploying software in a cloud environment. It is not unusual for an enterprise to choose cloud implementation for its test, development, and/or data lab deployments even if it has its production analytic environment deployed on-premises. These types of hybrid cloud implementations provide a risk-managed path for getting into the cloud with high value to the business while not initiating the dramatic kind of transition required by a pure play approach.

Conclusions

The promise of reduced cost often attracts organizations to cloud deployment of analytics. While there are definitely circumstances where cost savings will be significant for some use case scenarios, the true value of cloud deployment is more likely to be found in the competitive advantage associated with greater agility. Unless an enterprise believes that platform infrastructure will provide a specific source of competitive advantage, there is a good argument for focusing smart people within the organization on more strategic issues.

In the past, there has been a lot of concern associated with cloud deployment of sensitive data. Primary areas of apprehension include security, sovereignty, and loss of control related to SLA management. These days, concerns about the security of data in a public cloud is more related to fear mongering from stakeholders with a vested interest in keeping an on-premises status quo rather than a technical reality. The major public cloud providers have invested far more in security than any single enterprise could ever afford. The fact is that corporate data centers are far more vulnerable to data breeches than public cloud infrastructure.

The issue of sovereignty is more political than technical. Government regulators will sometimes mandate that sensitive data be held only in computers housed in data centers on local soil. In some cases, this is protectionist behavior motivated by the desire to defend local IT data center industry players. In other cases, it is a naïve attempt at retaining control of data assets—as if the physical location of data will make any difference to someone attempting to hack into a data center over the internet.

However, there is a legitimate concern related to jurisdiction when protecting citizen data from access by foreign governments. The jury is still out on exactly how well data is protected from a foreign government's meddling when the data is in a data center operated by a commercial enterprise of that foreign government—even when the data center resides on national soil.

Loss of control related to managing service levels for mission-critical applications is a valid concern for enterprises that use analytics for competitive advantage. Not all clouds are created equally when it comes to delivering high performance and high availability for analytics. Stakeholders must expend the ir sources and engage expert help to determine which cloud platform offers the best solution for analytics. Both the laws of physics and the laws of economics must be evaluated to determine what is "best" for a particular enterprise. This paper provides a framework for making good decisions regarding choice of cloud platform for deployment of analytics.

All organizations, whether now or in the near future, will find themselves in the cloud for at least some applications. As data gravity shifts to the cloud, more and more analytics will also shift to the cloud. If your organization does not already have some kind of presence in the cloud, now is the time to start. Cloud deployment does not have to be a big bang initiative. Look for opportunities to introduce hybrid on-premises and cloud deployments to manage risk and deliver high value sooner rather than later.

About Teradata

Teradata is the connected multi-cloud data platform company. Our enterprise analytics solve business challenges from start to scale. Only Teradata gives you the flexibility to handle the massive and mixed data workloads of the future, today. Learn more at **Teradata.com**.

17095 Via Del Campo, San Diego, CA 92127 Teradata.com

teradata.