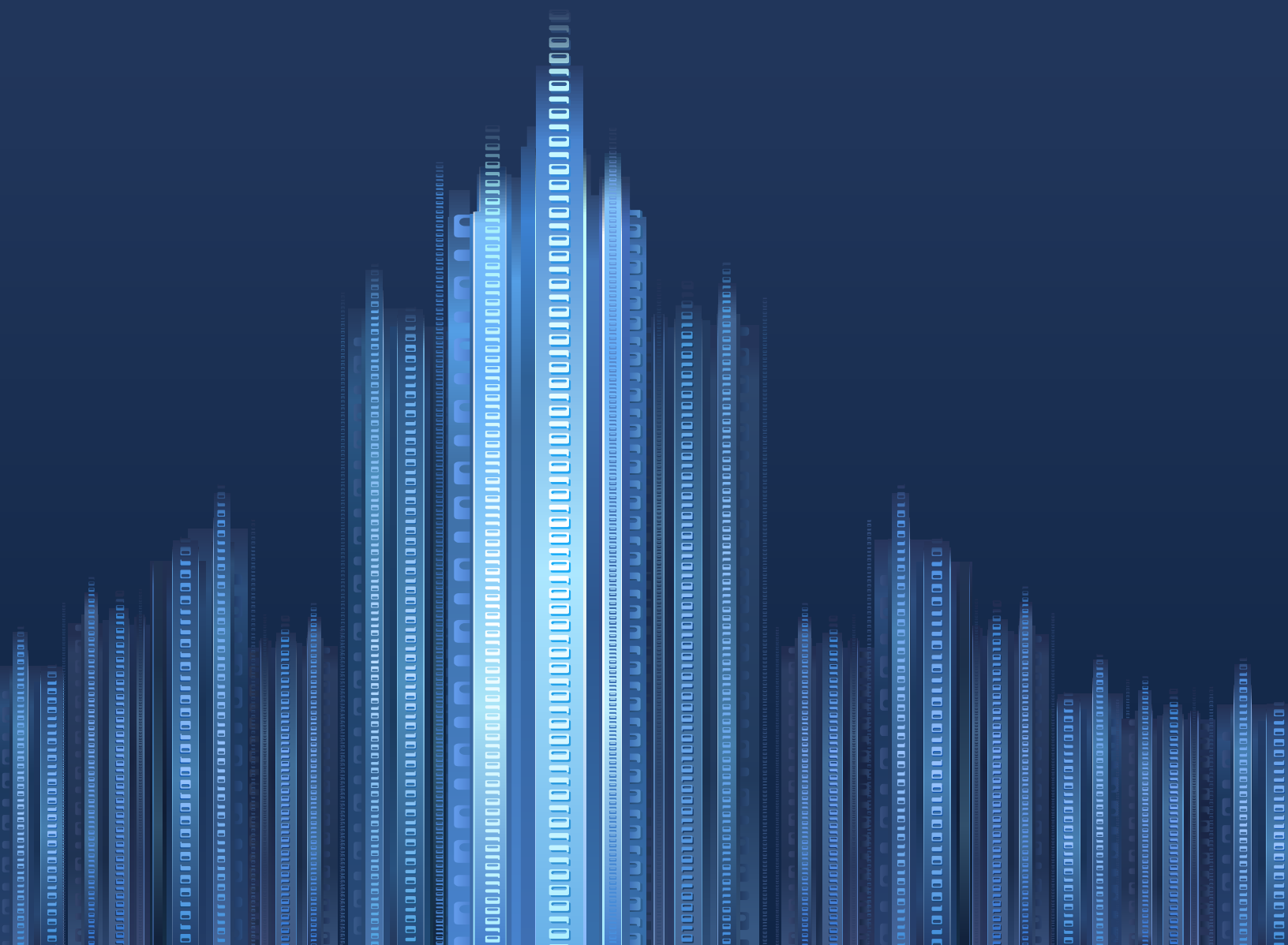
Accunet

Building a Scalable Big Data Infrastructure for Dynamic Workflows



INTRODUCTION

Organizations of all types and sizes are looking to big data to help them make faster, more intelligent decisions.

Many efforts involve the ingestion, storage, and sophisticated analysis of new or richer datasets. Unfortunately, most IT infrastructures have shortcomings when it comes to big data. They are not well suited to handle the volumes of data involved. And most cannot efficiently support the variable and unpredictable analysis and data mining workloads. Simply stated, they do not scale well.

Due to its early reliance on the analysis of large datasets, the field of the life sciences provides a prime example of the challenges companies in any industry face. In the life sciences, speed-to-decision is critical. Researchers need to quickly analyze and evaluate the results of many experiments to make choices as to which new drug candidates to move forward.

Each new generation of sequencers, mass spectrometers, imaging microscopes, and other lab equipment produces a richer, more detailed set of data. A single experiment can produce hundreds of gigabytes (GB) of data. As a result, any organization running hundreds of experiments a month quickly finds itself with data management problems and infrastructure issues.

Like big data efforts in other industries, life sciences research typically relies on automated, multi-step analysis workflows. Such workflows place higher performance demands on the servers, storage systems, and network elements.

Simply throwing raw processing, storage, and network capacity at the problem can help. But it is not the best solution. As elements are added the challenge becomes how to meet growing performance demands while minimizing the burden of managing the infrastructure.

Fortunately, lessons learned from some early life sciences big data efforts, such as those undertaken at The Broad Institute and the National Cancer Institute (NCI), show that smart infrastructure choices are the key to meeting today's challenges. Specifically, if the right infrastructure is put into place, it will provide the performance needed now and the scalability and flexibility to meet future requirements. The lessons learned by these life sciences organizations are applicable to big data endeavors in any industry.

INFRASTRUCTURE CHALLENGES WITH BIG DATA

In any big data project, storage capacity to accommodate the datasets must increase. However, simply adding raw capacity, without taking other infrastructure issues into account, can lead to problems and inefficient use of resources. The field of the life sciences provides examples of the potential impact of big data on an IT infrastructure. They include:

Interdependencies of infrastructure elements: Life sciences organizations are increasingly turning to virtualization to reduce operating costs, consolidate servers, and simplify the deployment and management of applications.

Server cluster nodes based on multi-core processors are now commonly used in conjunction with virtualization software to enable dozens or more applications to run as virtual machines on each physical server. Open source software, such as Hadoop and NoSQL, gives companies a way to leverage these clusters to run big data analytics.

EXECUTIVE SUMMARY

- Big data is placing new demands on IT infrastructures
- The challenge is how to meet growing performance demands while minimizing management burdens
- An ideal solution tightly integrates servers, storage, and network elements to optimize performance and reduce administrative chores

As this architecture becomes more widely used, organizations must address several other infrastructure issues because new performance issues crop up. A single server accessing a single storage device generates predictable workloads. But matters become much more complex in a cluster running dozens of virtualized applications.

The key issue becomes how to best integrate servers, storage, and network elements. The numerous applications running on the cluster all need simultaneous access to the data on storage devices. That means the storage solution will have to accommodate multiple concurrent workloads without degradation. Additionally, the network switches and adapter cards must offer the throughput and IO to sustain the required performance levels.

This places new demands on both the storage solution and the network. In particular, big data analytics requires that storage be flexible and capable of being dynamically grown to meet varying capacity and performance requirements. Because virtualized applications can be quickly and easily set up and torn down, the associated storage must support easy, dynamic provisioning. Additionally, provisioning and addition of new storage capacity must not involve taking systems offline.

From the networking perspective, server virtualization and big data analysis can change the dynamics of traffic flow within the data center network. Network links to the servers can become congested, impacting network performance and throughput. A common solution is simply to add more links. But this increases the number of switch ports needed and adds to the administrative burden on the IT staff. What is needed is a network that offers high performance scalability.

> "Like big data efforts in other industries, life sciences research typically relies on automated, multi-step analysis workflows."

Unpredictable workloads: Another infrastructure issue to consider relates to the change in the way data is accessed in a big data workflow. Efforts to derive decision-making information from big data sources typically use a number of analysis tools, applied at different stages of a computational workflow. For example, in the life sciences, newer lab equipment, such as next-generation sequencers, produces a much richer set of data. This makes the raw output from today's lab equipment of interest to a more diverse group of researchers.

Each group of researchers subjects the data to a wide variety of analysis tools, all with different IO and throughput requirements. Depending on the type of analysis being performed, the workflows and applications will likely have diverse performance requirements. All of this makes big data workflows highly unpredictable.

What is required is an infrastructure that can support diverse workflows, offering high sustained throughputs. Specifically, the infrastructure must be able to handle large sorts, which are quite common in big data workflows. With large sorts, the files are typically larger than system memory and therefore cannot be retained in local cache. As a result, large sorting workloads require a file system and storage solution that can deliver high throughput and IO. The storage system must also be able to provide low latency access to file system metadata.

Data management: Life sciences research has become more multi-disciplinary and more collaborative. This complicates data management and makes computational workflows more complex. As noted above, the richness of data from newer lab equipment makes it of interest to more types of researchers. Some groups might need instant analysis of the data in early stage research to determine which new drug candidates to move along and evaluate further. Other groups might need to re-examine the original data months or years later when a candidate moves into clinical trials or is being studied for potential adverse effects.

An additional implication of the multi-disciplinary and more collaborative nature of life sciences research is that data increasingly must be shared. From a storage perspective, older data must be moved to lower cost storage after its initial analysis. However, when new analysis is required, the data must be easily located and automatically migrated to higher performance storage. What is needed is a solution that offers a variety of price/performance storage choices so data can be placed on appropriate tiers. The solution should have a robust file system, capable of managing the explosive amounts of data as a single volume. Furthermore, a smart solution will help organizations manage and protect the data.

NCI AS AN EXAMPLE

Addressing these issues can help an organization determine the type of infrastructure needed for its big data challenges. That was the case with the National Cancer Institute (NCI).

In 2011, the NCI had outgrown its existing datacenter at the Frederick National Laboratory for Cancer Research (FNL), and decided to build a new facility. It put out a request for proposal (RFP) that left it up to the integrators to determine the best solution for the lab's computational work.

IT systems integrator Accunet Solutions proposed an end-to-end solution, won the bid, and was awarded the $15 million contract. The company had lots of experience in developing infrastructure for life sciences organizations, having previously worked with the Broad Institute and other organizations. (About half of Accunet's clients are life sciences organizations.)

Accunet proposed a scalable backbone for FNL's next-gen computing needs. The proposed design included specifications for switching and routing elements, a new storage platform, wireless access, virtualization, and data optimization. The plan offered the flexibility for NCI to move to a private, public, or hybrid cloud model.

The proposed solution is highly scalable in BOTH performance and capacity, makes efficient use of resources, is easy to manage, and tightly integrates all elements to help optimize workflow performance. The selection of specific vendors and technologies for the infrastructure was guided by the desire to keep things simple.

> "An infrastructure for big data must support diverse workflows, offering high sustained throughputs."

Accunet's Key Technology Partners for the NCI Big Data Backbone

Partner	Solution
Cisco	Servers and core network elements
EMC	Storage and SANs technology
VMware	Virtualization
CommVault	Data management

EMC/Isilon was selected as the storage solutions provider. The original plan called for 2 Petabytes (PB) of storage, but the lab is already approaching 4 PB. Storage solutions from EMC/Isilon offer the sustained high throughput and IO for the lab's computational workflows, a file system that accommodates the large amounts of the data used by the lab, and a variety of price/performance devices to allow data tiering.

Strategic partner Cisco was selected not only for its fabric technology, but also for its Cisco UCS (Unified Computing System) platform. The elements from Cisco include servers, VoIP (Voice over IP) gear, core network elements, and wireless networking equipment. Key attributes of the Cisco solution are that it offered lossless networking and non-blocking switch technology. These elements help make the NCI switching fabric fast, efficient, and highly available.

Other solutions selected to complete the infrastructure included EMC VNX storage for the SAN; VMware for virtualization; and CommVault for data management. The inclusion of CommVault allows storage administrators to track when files are created, where they are stored, and how many times they are touched. This information is used to create policies to automatically move data from one storage performance tier to another.

The characteristics of this infrastructure architecture, while created to meet NCI's specific needs, are ideal for big data efforts in any organization.

> "The key issue becomes how to best integrate servers, storage, and network elements."

ACCUNET AS YOUR TECHNOLOGY PARTNER

For the past 15 years, IT systems integrator Accunet Solutions has steadily expanded its business. The company is working with an impressive roster of life sciences organizations. It also works with financial services, higher education, and government organizations.

The company has a great deal of experience using leading-edge technologies to help companies become more efficient, while retaining their competitive edge. For big data efforts, Accunet Solutions brings expertise in data storage, networking, virtualization, and security.

Accunet Solutions offerings include complete design, implementation, and optimization services. In a typical engagement, Accunet Solutions combines the use of certified engineers with project management professionals to provide complete IT solutions for an enterprise. The company provides real-time project management services including resource allocation and progress reports, to ensure IT projects are completed on time and within budget.

For more information about building an infrastructure for big data, visit:
http://www.accunetsolutions.com/

Accunet