Inside
Bio·IT World

The Quarterly eBook of Bio-IT World's Most Trending Articles

Produced by
Healthtech Publishing
250 First Avenue, Suite 300
Needham, MA 02494
www.healthtechpublishing.com

Inside Bio·IT World

The Quarterly eBook of Bio-IT World's Most Trending Articles

Managing and Moving the Data We Produce

Bio-ITWorld.com

Produced by
Healthtech Publishing

Produced by
Healthtech Publishing
250 First Avenue, Suite 300
Needham, MA 02494
www.healthtechpublishing.com

Produced by
Healthtech Publishing
250 First Avenue, Suite 300
Needham, MA 02494
www.healthtechpublishing.com

WEKA

About Our Sponsor

Recent years have seen a boom in the generation of data from a variety of new sources: connected devices, IoT, analytics, healthcare, smartphones, and much more. In fact, as of 2020, 90% of all data ever created had been created in the previous four years. Gaining insights from this data presents a tremendous opportunity for organizations to further their businesses, expand more quickly into new markets and to advance research in healthcare or climate - just to name a few. However, the challenge of managing the sheer amount of data being generated, coupled with the need to more quickly glean insights from it, has created an infrastructure nightmare. Organizations have been reporting unstructured data growth of over 50% year over year (Gartner, 2018), while at the same time, 79% of enterprise executives agree that not extracting value and insight from this data will lead to extinction for their businesses (Accenture, 2018). This data management problem is particularly acute in the areas of artificial intelligence/machine learning (AI/ML), life sciences (including genomics and microscopy), financial analytics, high-performance computing (HPC), and anywhere there are both extreme compute requirements and the need to store and analyze massive amounts of data.

Data growth has put the underlying IT infrastructure under tremendous stress across organizations, and in turn created a huge problem for IT administrators. In order to deal with the growth in unstructured data, many organizations have turned to shared (scale-out) storage systems, where capacity can be expanded by adding more storage nodes. However, these solutions have not addressed the performance demands of their modern applications, and many organizations resort to copying and running data locally on the application servers. By 2022, more than 80% of enterprise data will be stored in shared storage systems in enterprise and cloud data centers, up from 40% in 2018 (Gartner, 2018). Legacy shared storage solutions are no longer adequate as they were designed to solve yesterday's problems and cannot handle the previously unforeseen problems now being encountered. They cannot feed data into compute resources fast enough, resulting in an inability to complete work on time, nor can they scale to petabytes of capacity when performance counts. At today's scale, just managing the metadata associated with the source data becomes a problem when dealing with billions of files within a data set.

WekaIO (Weka) provides a way forward to help solve today's and tomorrow's biggest computing problems. Weka has developed a modern file system (WekaFS™) that is designed to provide extreme performance at ANY scale. WekaFS will solve any high-performance data storage problem and is ideal for all extreme technical compute and performance-intensive applications, including machine learning, deep learning, genomics, microscopy and technical compute.

Weka: the file system for those who solve big genomics problems.

Inside
Bio-IT World
The Quarterly eBook of Bio-IT World's
Most Trending Articles

Produced by
Healthtech Publishing
250 First Avenue, Suite 300
Needham, MA 02494
www.healthtechpublishing.com

EDITOR'S NOTE

Allison Proffitt

> **"We need to be able to use, share, and work with our burgeoning data to make scientific progress."**

In 2019, when Genomics England pledged to sequence five million genomes, David Ardley knew he needed a new plan fast! There was no way the existing storage system could keep up with expected 140 petabytes of data that was coming. It turns out, this isn't an uncommon problem. As our data generation capabilities explode, our storage and file system tools must keep pace. We need to be able to use, share, and work with our burgeoning data to make scientific progress. Like never before, our data management capabilities will drive innovation.

Allison Proffitt

Allison Proffitt
Editorial Director, Bio-IT World and Clinical Research News

PRESIDENT
Phillips Kuhl

EDITORIAL DIRECTOR
Allison Proffitt (617) 233-8280
aproffitt@healthtech.com

ASSOCIATE DIRECTOR OF PUBLICATION MARKETING
Tom Conley (781) 972-1351
tconley@healthtech.com

VICE PRESIDENT, BUSINESS DEVELOPMENT, DIGITAL MEDIA
Angela Parsons (781) 972-5467

MANAGER, BUSINESS DEVELOPMENT
Katelin Fitzgerald (781) 972-5458
Lead Gen Companies A-I
kfitzgerald@healthtech.com

MANAGER, BUSINESS DEVELOPMENT
Joseph Vacca (781) 972-5431
Lead Gen Companies J-Z
jvacca@healthtech.com

CONTACT INFORMATION
editor@healthtech.com

250 First Avenue, Suite 300
Needham, MA 02494

CONNECT WITH US:

ABOUT BIO-IT WORLD

Part of the Cambridge Healthtech Institute Media Group, Bio-IT World provides outstanding coverage of cutting-edge trends and technologies that impact the management and analysis of life sciences data, including next-generation sequencing, drug discovery, predictive and systems biology, informatics tools, clinical trials, and personalized medicine. Through a variety of sources including, **Bio-ITWorld.com**, Weekly Update Newsletter and the Bio-IT World News Bulletins, Bio-IT World is a leading source of news and opinion on technology and strategic innovation in the life sciences, including drug discovery and development.

This index is provided as an additional service. The publisher does ot assume any liability for errors or omissions.

Subscriptions: Address inquires to Bio-IT World, 250 First Avenue, Suite 300, Needham, MA 02494
888-999-6288 or e-mail kfinnell@healthtech.com

Inside Bio-IT World
The Quarterly eBook of Bio-IT World's
Most Trending Articles

Produced by
Healthtech Publishing
250 First Avenue, Suite 300
Needham, MA 02494
www.healthtechpublishing.com

Building A File System To Keep Up With Genomics England's FIVE MILLION GENOMES PROJECT

BY ALLISON PROFFITT | MARCH 24, 2020

When David Ardley took over the platforms function at Genomics England, he inherited "quite a challenging storage environment," he told Bio-IT World. The UK Department of Health launched Genomics England in 2013 with an audacious goal to sequence 100,000 genomes. Since October 2018, the vision has expanded to 5 million genomes—a growth of 4,900%.

Inside
Bio·IT World
The Quarterly eBook of Bio-IT World's
Most Trending Articles

Produced by
Healthtech Publishing
250 First Avenue, Suite 300
Needham, MA 02494
www.healthtechpublishing.com

At the end of 2018, GEL already had 21 petabytes of genomic data and expected that number to grow to over 140 petabytes by 2023, when the 5 million genomes project is slated for completion. GEL's previous scale-out NAS solution had already hit its limit on storage node scaling and was experiencing performance issues.

It was time for something new. GEL kicked off the RFP process in January 2019 seeking a storage platform for a new era of genomics.

GEL needed several functionalities from the new storage system, Ardley explained. The RFP benchmarked against three primary requirements: The new storage needed a robust security and disaster recovery plan. All the data needed to be active and accessible for researchers. And the system needed to perform at scale to achieve the goal of five million genomes by 2023.

The GEL team evaluated four proposals. They rejected parallel file systems because of their complexity and lack of enterprise features; they rejected all-flash scale-out NAS because the costs wouldn't scale as GEL's needs did.

Ultimately GEL chose WekaIO's WekaFS solution. "That was primarily driven by Weka as a cache layer that required potentially low management," Ardley explained.

The WekaFS product is a distributed, shared file system running on commodity, off-the-shelf flash and disk-based technologies into a single, hybrid solution, says Barbara Murphy, VP of Marketing at WekaIO. "Our software layers on top of all those individual servers and creates those massively distributed file systems with all the underlying storage in each one of those servers combined together to present a single scale-out NAS solution."

The solution met all of GEL's priority requirements. First, storage needed to be distributed across multiple sites. Backing up 21 petabytes of data—and growing—is impractical, but GEL still needed a disaster recovery strategy and a strong security plan. "One of the key requirements of storage was to have sites 50 to 100 miles apart," Ardley explained. For security and backup, the object store is geographically dispersed over three locations all 50 miles apart, but all located within England. If a major disaster occurs in the primary location, Weka's Snap-to-Object feature allows the system to be restarted in a second location.

> **"Our software layers on top of all those individual servers and creates those massively distributed file systems with all the underlying storage in each one of those servers combined together to present a single scale-out NAS solution."**
>
> **Barbara Murphy**
> VP of Marketing at WekaIO

The existing storage was a fairly traditional tiered approach, Ardley explained, but the new solution needed to be more flexible. "Researchers are looking at the same data and they randomly access everything," he said. "Having all the data really active and performing is a requirement."

WekaFS delivers a two-tier architecture. The primary tier consists of 1.3 petabytes of NVMe-based flash storage that supports working datasets. The secondary tier consists of 40 petabytes of object storage to provide a long-term data lake and repository. In this case, the underlying object store is ActiveScale from Quantum (recently acquired from Quantum).

The whole 41 petabytes are presented as a single namespace. The two tiers scale independently; if more performance is needed on the primary tier, it can increase independently of the second tier.

"We manage the moving back and forth of data between those two tiers seamlessly to the user," Murphy

Inside
Bio·IT World
The Quarterly eBook of Bio-IT World's
Most Trending Articles

Produced by
Healthtech Publishing
250 First Avenue, Suite 300
Needham, MA 02494
www.healthtechpublishing.com

explained. "The user doesn't have to do anything, they don't have to load special software, they don't need data migration software, they don't need anything. We manage that all internally. When they say, 'I want XYZ file,' we move it from the cold tier which is on the object store right into the flash tier so it's available immediately for the user."

> **"The user doesn't have to do anything, they don't have to load special software, they don't need data migration software, they don't need anything. We manage that all internally. When they say, 'I want XYZ file,' we move it from the cold tier which is on the object store right into the flash tier so it's available immediately for the user."**
>
> **Barbara Murphy**
> VP of Marketing at WekaIO

Finally, GEL needed a system to perform at scale. For five million genomes—ignoring compression—Ardley estimates about 150PB of needed storage. And as the number of genomes increased, he needs the system to handle ingest rates of eventually 3,000 genomes per day. "The I/O requirements for storage are very high," he said. Of course there are other bottlenecks to that ingest rate, but "we wanted to design it such that storage and the high performance compute elements weren't a bottleneck," he emphasized.

"The rate that they're actually creating data is phenomenal, and in fact we're just about to do another expansion of the system," Murphy said.

Cost, Ardley said, wasn't a primary priority, but, "It just so happens that the one we picked was also the cheapest," he said.

ROLLOUT UNDERWAY

GEL chose the WekaIO solution in April 2019, installation took about two months, and data migration began in September 2019. The data migration is still going. "We literally had about 25 petabytes of read/write data in a very unstructured way," Ardley says. "You can imagine that's not going to be a very easy migration process. It's been very challenging, but we're at the tail end of that."

Some of the biggest challenges in data migration came from identifying who owned various datasets, whether they could be deleted, and what authentication was needed.

"Prior to starting this project, we did a lot of cleanup, basically just to keep the servers running. There was a lot of duplication, directories that no one knew who owned them," he said. "There were lots of directories that were very small or empty… We spent a lot of time just trying to find owners of data… knowing whether we could delete something. Who do we need to talk to say we're moving this data? If we didn't have that really locked down and managed, you're just creating a problem later on."

But even with the migration and data stewardship challenges, Ardley is pleased.

"So far the performance has been really good, though I wouldn't say it's been as stressed as it's going to be. We've had to fine tune it, but it's lived up to the performance we were expecting to see, which is quite nice!"

Inside
Bio·IT World
The Quarterly eBook of Bio-IT World's
Most Trending Articles

Produced by
Healthtech Publishing
250 First Avenue, Suite 300
Needham, MA 02494
www.healthtechpublishing.com

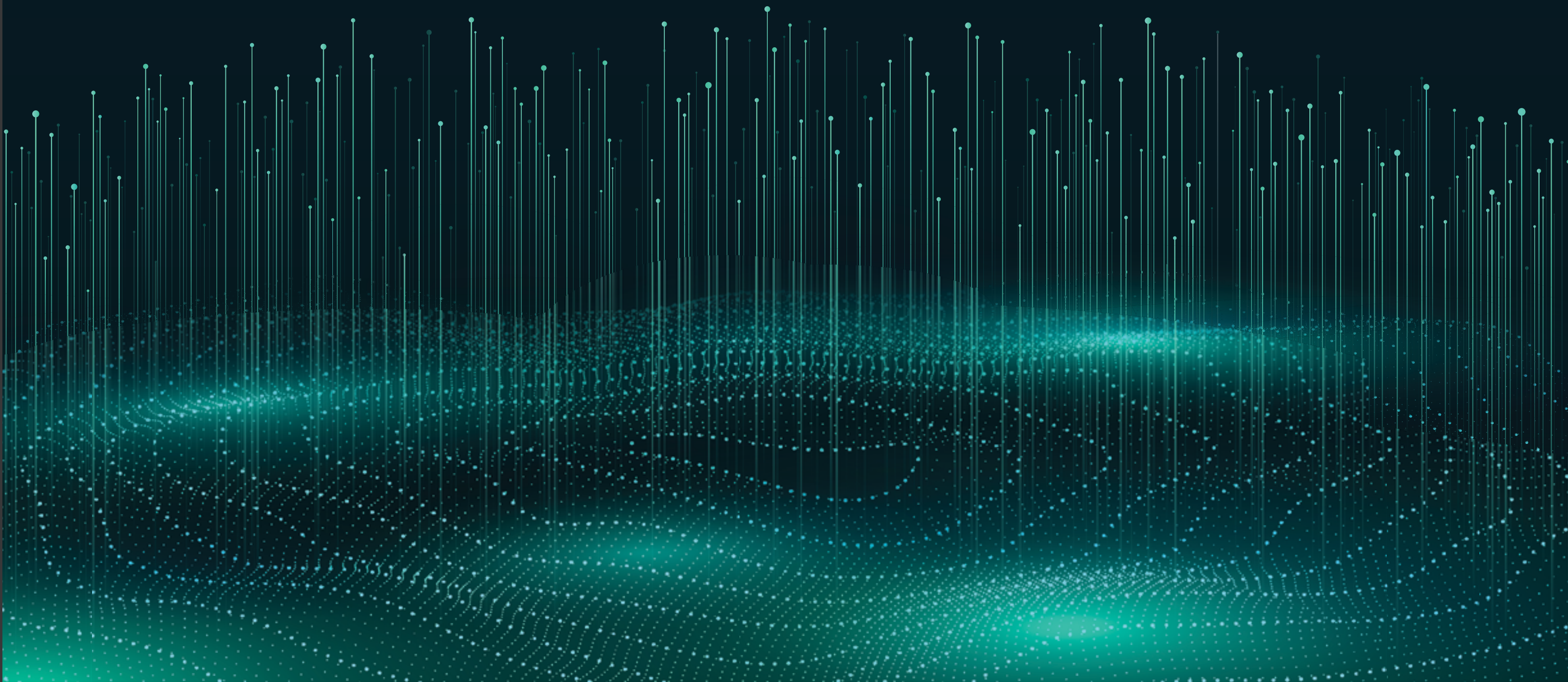
DATA STORAGE AND TRANSPORT: HEADACHES AND POSSIBLE REMEDIES

BY PAUL NICOLAUS

As organizations generate ever-growing amounts of data, finding ways to overcome the related difficulties of storing, moving, and managing that flood of information remains a central issue.

Nearly every research institution in the healthcare and life sciences space faces the need to distribute large and frequently changing data sets, explained Seth Noble, founder and CEO of Data Expedition, which provides software solutions for network performance and reliability.

The ability to send and receive large volumes of scientific research quickly and reliably is critical, but large genomic datasets can take hours or even days to transfer. Whether stored locally or in the cloud, he added, transporting data to research colleagues in a timely manner is a major challenge.



Inside
Bio-IT World
The Quarterly eBook of Bio-IT World's
Most Trending Articles

Produced by
Healthtech Publishing
250 First Avenue, Suite 300
Needham, MA 02494
www.healthtechpublishing.com

Many continue to use hybrid storage approaches where data is hosted partially in cloud environments and partially in private data centers, explained Mark Lambrecht, global director for health and life sciences at SAS, an analytics and data management software company. This is because some datasets are so large that it just isn't feasible, from a cost standpoint, to bring all data to cloud storage even if there are good business reasons for doing so, he said.

In other instances, cloud storage may be restricted. One example is patient data that cannot be moved beyond the walls of a medical facility. This calls for federated approaches where analysis is handled over different locations and only the summary results are centrally stored. "These hybrid and federated approaches require robust transport data standards that can transfer data without losing any semantic meaning to the data and how they are interlinked," he explained.

The wide world of life sciences is getting smaller, according to Steve Levine, as digital technologies close the loop between researchers discovering new treatments and medical practitioners figuring out "precisely which to provide to whom."

Upstream, research and development scientists may disagree on plenty, but one thing they all agree on is that they are inundated with data, continued Levine, the senior director of life sciences at Dassault Systèmes, a software business headquartered in France with a campus in Providence, RI, where the 3DS SIMULIA brand is headquartered. Factors like competition and the quickening pace of research have nudged organizations to explore new ways of extracting every bit of value from the data that is collected. "Downstream, with the onset of value-based care and the dawn of precision medicine, providing the best patient experience is driving the digital transformation in care centers very rapidly," he added. Unlike other industries that strive to provide the best products or services at the lowest cost, however, the healthcare industry looks to deliver the best solution at the most reasonable price.

The vagueness of what is considered reasonable has allowed the industry to move forward without placing an emphasis on data standards. This, in turn, has slowed the adoption of digital technologies and allowed "a vast diversity of data types to proliferate."

While the cloud may offer a lower entry to providing data access, Levine contends that it is not the key technology that determines utility. In the life sciences, "the key is not so much where the data is stored," he added, "but rather how the data is stored."

> "Downstream, with the onset of value-based care and the dawn of precision medicine, providing the best patient experience is driving the digital transformation in care centers very rapidly."
>
> **Mark Lambrecht,** Global Director for Health and Life Sciences at SAS

DATA MANAGEMENT RESOURCES

In recent years, there has been an incredible growth of data coming from instruments, said Vas Vasiliadis, who leads the customer team for Globus and teaches in the Computer Science program at the University of Chicago.

Next-generation sequencing has been around for a while, "but we're also seeing some very high-resolution instruments" such as light sheet and cryo-electron microscopes growing substantially over the last year or two, he said. Other examples include high-resolution photon sources like The Advanced Photon Source at Argonne National Laboratory or the UK's Diamond Light Source.

Many of these instruments deliver a continuous flow of data but have little local storage. It is important to be able to pick that data up and move it somewhere else where it can be stored more permanently and where people can gain easier access to it for downstream analysis.

"The idea is, get the data off there quickly, and very importantly, get it off there reliably," he said, because in many scenarios it

Inside
Bio·IT World
The Quarterly eBook of Bio-IT World's
Most Trending Articles

Produced by
Healthtech Publishing
250 First Avenue, Suite 300
Needham, MA 02494
www.healthtechpublishing.com

isn't possible to reuse a sample or redo an experiment.

Being able to do this at scale and in an automated fashion is essential, too, considering these types of instruments tend to be expensive—and shared—resources. There are people lined up to use them, Vasiliadis added, so the quicker you can get data off of there from the previous run and move on to the next sample, the better.

A hybrid solution for research data management, Globus enables users to move, share, and discover data using a single interface. Whether files are stored on a supercomputer or a laptop, the data can be managed from anywhere using a web browser. The storage is all owned and controlled by users as this cloud-hosted service manages the interactions.

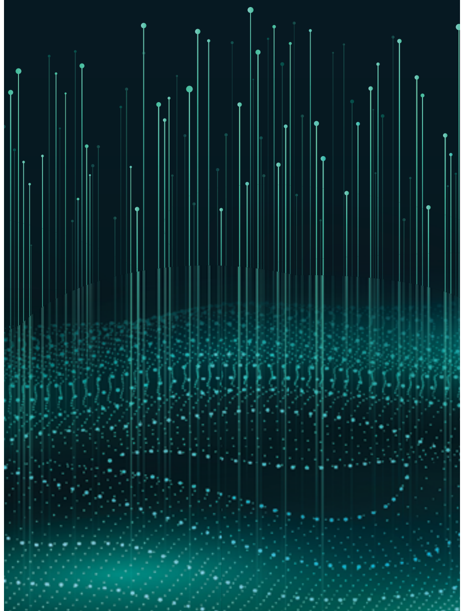
"The core of the service is a high-speed, reliable file transfer capability," said Vasiliadis. Globus uses an open protocol called GridFTP that provides multiple parallel streams between the two endpoints where the data are being moved, and within each of those streams is a further set of parallel threads used as part of the protocol.

In recent years the service has gone beyond that core focus. The ability to share data with external collaborators has been added, for example, along with additional services for data search and the automation of data flows that are common to large research institutes.

At the German Center for Diabetes Research (DZD), the data and knowledge management team uses graph technology to manage large sets of information and use it for prevention and individualized treatments for diabetes.

The DZD has a research network that accumulates large amounts of data distributed across various locations, including labs and hospitals. To address this challenge, DZD is building a master database, which is based on graph database management system Neo4j, to provide its team of scientists with a more holistic view of information.

> "Cloud is more an endpoint that just really kind of compounds your data management problems rather than alleviates them."
>
> **Mike Conway,**
> Technical Architect, National Institute of Environmental Health Sciences



In general, a graph is a mathematical structure with nodes and edges. "The nodes are connected to edges, which is—I think—the most intuitive way of designing a data model for biological data," said Alexander Jarasch, head of data and knowledge management at DZD.

In a browser-based manner, users can take a data-point such as a specific gene, patient, or experiment and see what data is connected to it. Jarasch said he started with the Neo4j technology back in 2017 «to build up the meta database from all our different sites and disciplines and also species.»

Work completed to date involves human clinical studies, which has generated human data and measurements as well as metadata like the number of patients, the number of samples, and what was measured. Other work involves the storage of basic research data on animal studies, which is more of an ongoing process.

Some new projects are extending beyond diabetes to discover potential connections with other illnesses, such as cancer or Alzheimer's disease. There are efforts to connect different databases on a meta-level, he explained, "because we see that there are side effects or long-term complications from diabetes to cancer or from lung diseases to diabetes or back and forth."

Dealing with the Data Lifecycle

Designing, managing, and scaling solutions to deal with large amounts

Inside
Bio·IT World
The Quarterly eBook of Bio-IT World's
Most Trending Articles

Produced by
Healthtech Publishing
250 First Avenue, Suite 300
Needham, MA 02494
www.healthtechpublishing.com

of research data boils down to effectively managing that data through its entire lifecycle, according to Mike Conway, a technical architect working in the realm of scientific research data management at the National Institute of Environmental Health Sciences (NIEHS) and their Office of Data Science.

It's not necessarily a question of scale, he explained. It's trying to address scale by working in upstream systems to improve the capture or recording of proper metadata that can be used as queues for data governance and policies and then automating those policies based on these properties of the different datasets.

From his vantage point, cloud isn't necessarily a data management solution. "Cloud is more an endpoint that just really kind of compounds your data management problems rather than alleviates them," Conway said.

"Our management problems really start when a sequencing run is done," he said, and there's a need to get that information to a point where it can be passed along to a bioinformatician or analyst. Those types of issues aren't addressed by cloud. It's more about gathering the metadata along the way until the data arrives at a point where it might make sense to push it out to a cloud service for later analysis or sharing.

Without the application of what he defines as data management, "what you end up with are terabytes and petabytes of data piling up on some vendor-supplied storage array," Conway said, where nobody knows what to keep and "you start getting into these governance questions."

If the raw data off the sequencer has been preprocessed, when can we get rid of that raw data? Is this human data versus animal data, and how are we differentially treating it? These are the types of questions that need to be tackled but are not addressed just based on where you're storing the data. "It's more about how we define those policies, how we automate them, how we can also get assurance that those policies have been applied," he added.

THE BIG BLIND SPOT OF DATA MANAGEMENT

Conway's background involves work with Reagan Moore, who developed the integrated Rule Oriented Data System (iRODS), an open source data management software, at the San Diego Supercomputer Center and later at the University of North Carolina's Data Intensive Cyber Environments (DICE) Center. This is the nexus from which his data management work with NSF and related DataNet projects stems.

These are examples of where the academic research side of data science is now starting to permeate government as well as biotech and pharma. In the same way that the Internet started in an academic research environment and percolated out, "I think that is also happening with data management, and it's really worth paying attention to the work that has been done on cyberinfrastructure on the NSF side," he added.

When he considers efforts to address data management challenges at NIEHS, Conway explained that the solution is not a homogeneous one. Based on his background, iRODS is being used because the platform automates the policy and provides the tools needed to handle the assessment of the application›s policies.

There is also interest in the use of pipelines and workflow languages to move data along the lifecycle.

Conway explained that much of the focus for the past year with the epigenetics core has involved the integration of a Data Commons with BaseSpace Clarity LIMS to gather information about the samples. This, he said, "allows us to apply governance on the data as it's properly ingested in the Commons with the technical metadata." Other efforts are more

> "We're also working with Synaptica on creating a metadata catalog of standard ontologies and vocabulary terms and working to get that permeated int the daily workflows of NIEHS."
>
> **Mike Conway,**
> Technical Architect, National Institute of Environmental Health Sciences

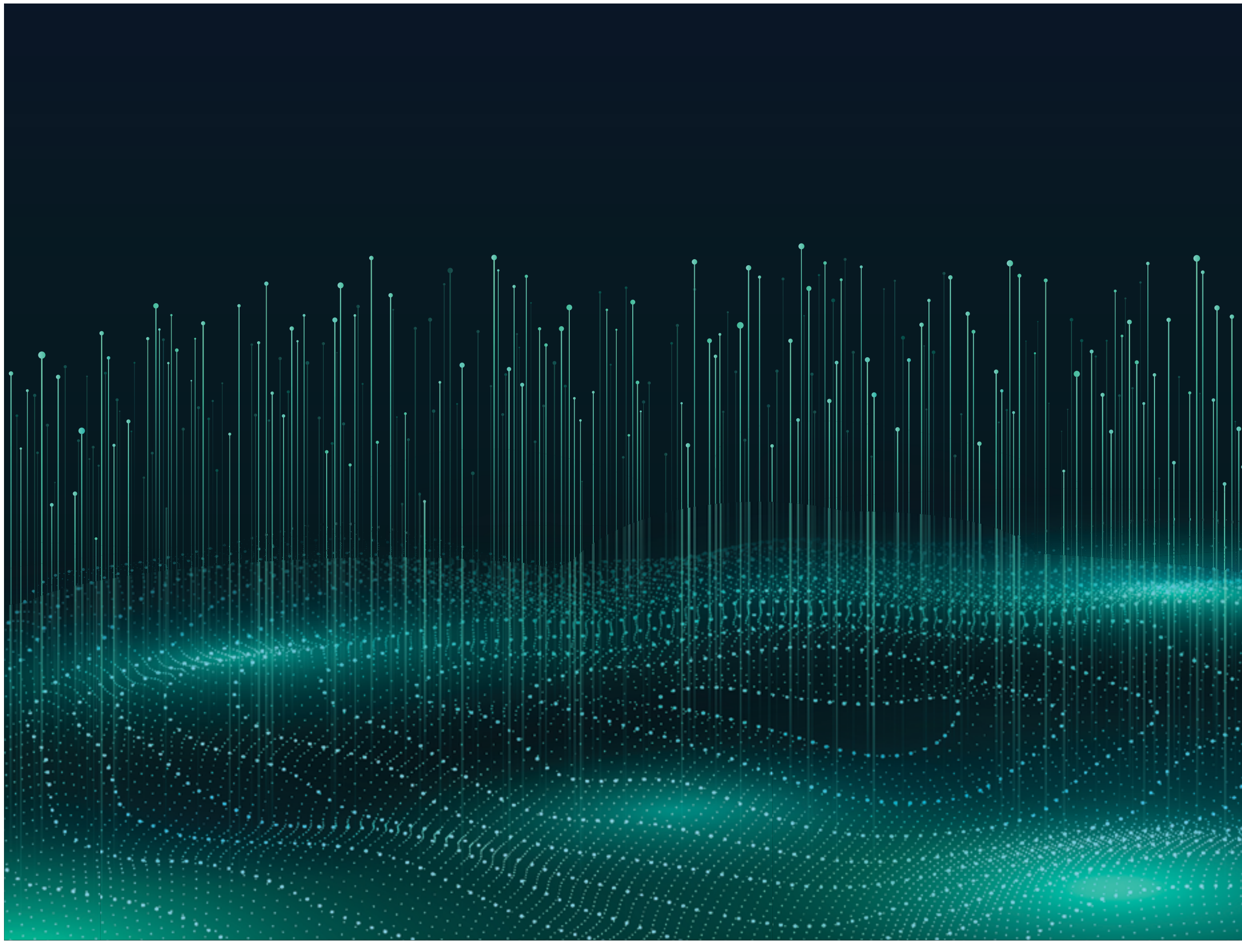
Inside Bio·IT World
The Quarterly eBook of Bio-IT World's
Most Trending Articles

Produced by
Healthtech Publishing
250 First Avenue, Suite 300
Needham, MA 02494
www.healthtechpublishing.com

organizational policy in terms of data governance and retention.

He likened this process to Sim City, referencing the city-building video game series. Within the game, "you have a flat piece of land, and you have to lay the power lines, you have to lay the plumbing, and then you start seeing buildings sprout up around that," he explained.

Similarly, cyberinfrastructure efforts can involve a combination of solutions. "We have iRODS, we have standard workflow languages for pipelines, and these sort of landing zones," Conway said. "We're also working with Synaptica on creating a metadata catalog of standard ontologies and vocabulary terms and working to get that permeated into the daily workflows of NIEHS.»



"We are definitely looking toward the cloud," he acknowledged, especially through the NIH Science and Technology Research Infrastructure for Discovery, Experimentation, and Sustainability (STRIDES) Initiative. Several cloud pilots are lined up, but all of the data management challenges he's referring to pertain to internal research projects and come before anything touches the cloud.

This is the "big blind spot" of data management from his vantage point. Most talk tends to revolve around data sharing—authentication, authorization, maintaining data sharing agreements, licensing, provenance, and creating permanent identifiers for shared datasets.

All of that is really important, Conway said, but there's a whole range of issues that are prerequisites before you even get to data sharing in the cloud. If there's junk metadata and disorganized data, there is plenty of prior work that needs to be handled before it can be published out as a shared dataset or reference collection.

"That's sort of the unsexy part about data management," he added, "but I think that's actually the more important part."

Inside
Bio-IT World
The Quarterly eBook of Bio-IT World's
Most Trending Articles

Produced by
Healthtech Publishing
250 First Avenue, Suite 300
Needham, MA 02494
www.healthtechpublishing.com

TRENDS FROM THE TRENCHES:

6 Themes (And A Few Observations)

BY DEBORAH BORFITZ

At his 2019 Trends from the Trenches address at the recent Bio-IT World Congress & Expo in Boston, BioTeam co-founder and senior director of infrastructure Chris Dagdigian shared his pearls of wisdom on the current state of scientific computing, a field where incompetence is now an "existential survival threat" to life science organizations. "We've done 'OK' entering the data-intensive science era," he says. "The hard part is managing what we have."

Dagdigian began by sharing a few general observations, notably that leadership still views scientific computing as a "cost center to be minimized" rather than a core competitive differentiator and HR recruitment and retention tool where insights and value routinely get extracted from data. The user base is also climbing and includes both seasoned scientists forced away from familiar, laptop-scale analytical methods and new hires often showing up with prior high-performance computing (HPC) and cloud expertise. Companies are "pretty bad at training," he adds, especially when it comes to helping intermediate-level users become experts.

Produced by
Healthtech Publishing
250 First Avenue, Suite 300
Needham, MA 02494
www.healthtechpublishing.com

The definition of HPC is also "being stretched in extreme ways," Dagdigian says, putting it "in danger of becoming a dumping ground for problems that don't fit on cheap leased laptops." The rate of software and tooling innovation is also happening faster than IT can curate and maintain development and execution environments.

"It may be time to start looking at things like Intel commercial compilers," Dagdigian says. A modest investment in compiler and toolchain optimizations could pay significant dividends, given the high cost of graphics processing units (GPUs), NVLink high-speed GPU interconnect and Nvidia DGX-2 platform ($400,000 list price). The performance differential between stock (Relion Cryo-EM) and even upgraded compiler developer tools (GCC-7 on CentOS/RHEL 7) is dramatic relative to Intel ICC compiler (Intel Parallel Studio), he notes.

The most prevalent Relion benchmark, favored by vendors, uses a 50 gigabyte input data set—small enough to fit in RAM—but that's an unrealistic test for anything other than getting compiler and CPU/GPU optimizations correct, Dagdigian says. Experimental data sets in 2019 are generally no smaller than 60GB in size. BioTeam is proactively seeking multi-terabyte CryoEM data organized for Relion 2D or 3D classification that it can share and publish its own results testing against.

MONSTER DATA FLOWS

Artificial intelligence (AI) and machine learning (ML) are both "awesome and ugly," says Dagdigian. "You are going to lie or die on the quality of your training data. I think there is going to be extreme pressure to get the cleanest and best data," tempting people to take "ethical shortcuts."

> **"You are going to lie or die on the quality of your training data. I think there is going to be extreme pressure to get the cleanest and best data."**
>
> **Chris Dagdigian,**
> BioTeam Co-founder and Senior Director of Infrastructure

Organizations are starting to employ creative strategies for gaining access to high-quality training data, in one case by providing access to sophisticated analysis tools on the cloud in exchange for data to feed its algorithms, Dagdigian says. The opt-in data sharing process generates 30,000 de-identified, metadata-tagged MRI scans per week that are getting dropped into medical and data workflows.

Little has changed in the networking arena over the last 18 months, Dagdigian says, the most shocking news being that Nvidia purchased Mellanox. "Moving scientific data across networks is still problematic, we're still doing a bad job of rolling out 40-gig and 100-gig networking, enterprise IT is still concentrating on the data center rather than building out edge and labs, we still need to separate science data traffic and business network traffic, and our connections to the outside world are too small. Our firewalls and security controls are also still designed for the way business consumes the internet and not for moving a massive amount of data as a single monster flow."

Cloud has been a capability play for life science organizations for the past decade, but cloud providers might start feeling a little pushback in 2019, says Dagdigian. Serverless computing is transformational, but discovery-oriented science still "relies heavily on interactive human efforts with bespoke tooling." Cloud marketing hype aside, "not everything can be distilled to an API [application programming interface] and built into a service mesh architecture. That is not going to happen in our world."

Two concerning developments are cloud efforts to build bespoke accelerated hardware for AI, ML, and inference, which will add complexity to an otherwise simple cost and capability evaluation process, says Dagdigian. The other is the consistent scarcity of GPU resources on Amazon Web Services. Not being able to get GPUs in a timely or inexpensive manner might drive some colocation and on-premise decisions.

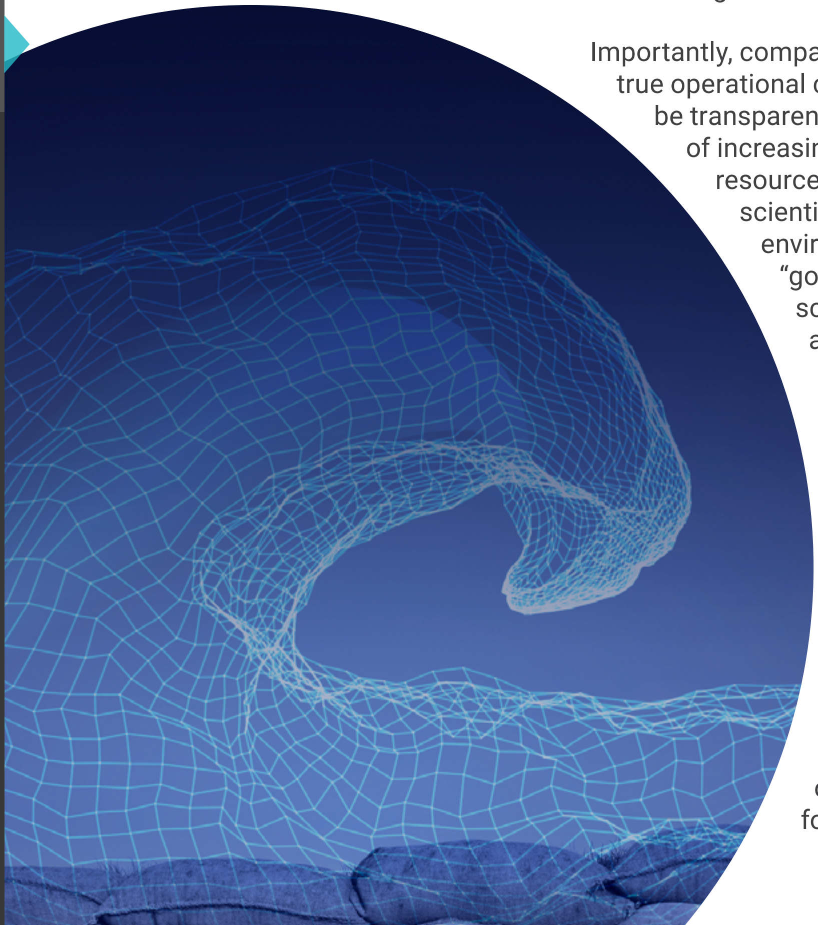
Produced by
Healthtech Publishing
250 First Avenue, Suite 300
Needham, MA 02494
www.healthtechpublishing.com

PREDICTIONS

In his latest Trends talk, Dagdigian kept the focus on six themes the industry can expect to see in 2019:

1. THE UNIT COST OF STORAGE VS. CONSUMPTION RATE WILL FORCE HARD CHOICES.

The unit cost of data storage is decreasing, but not as fast as data is being generated, Dagdigian says.



Meanwhile, the cost of running a sophisticated IT computing operation continues to mushroom—compelling some companies to cut research mission simply to sustain themselves. Governance and HPC resource allocation need to shift from IT groups to scientists and include some sort of data triage process. For open-ended moonshot projects, researchers could perhaps be handed internal credits or grants to "spend" however they see fit, he says. "We're going to have to get serious about this."

Importantly, companies need to have true operational cost data and be transparent about usage of increasingly contested resources within their scientific computing environment, with "good logging of scientific tools and codes being invoked," Dagdigian says.

"There is some value to lean times," he quickly adds. "It's the lean times that allow you to blow up silos and to de-crust your infrastructure or org cart and… force organizations to come together around APIs and middleware and common ways of doing things." Organizations will start thinking hard about how to run their technological operations in a different way on a larger scale.

2. STORAGE SELECTION PENDULUM HAS MADE A SEISMIC SHIFT TO BIG AND FAST.

There have been several fundamental changes on the storage landscape, notably that «the capacity/performance calculus has swung the other way,» says Dagdigian. "We now need very fast storage to handle machine learning, AI and image-based workflow requirements." ML training and validation also require ongoing access to "old" data. The dominant file type is no longer genome sequences but image-based (e.g., 3D ultrasounds).

"A tremendous among of deployed storage is nearing end of life or end of contract support," he says, and "really interesting next-gen storage companies have entered the market." Parallel storage systems have also become more attractive to life science companies, who may also be willing to bear more of the administrative and operational burden of getting the performance needed to support a new class of scientists and AI work. The benefits of scaling

Inside Bio-IT World
The Quarterly eBook of Bio-IT World's Most Trending Articles

Produced by
Healthtech Publishing
250 First Avenue, Suite 300
Needham, MA 02494
www.healthtechpublishing.com

out network-attached storage are less valuable in context, he says.

"The new requirements for speed plus capacity is deeply scary," says Dagdigian. "We can't trade away performance in exchange for larger capacity anymore. I think there will be lot of [enterprise IT] storage platform switches over the next couple years."

3. END USERS NEED TO START TAKING RESPONSIBILITY FOR THEIR DATA.

The level of knowledge required in the life sciences is becoming too much for a central support organization, says Dagdigian, so domain expertise will be moving to stakeholders in labs and R&D organizations. The problem is that scientists and even leadership typically don't know what data is important and therefore treat it all equally, resulting in significant operational overhead and overspending—and data hoarding.

Dagdigian admits he used to be a fan of single namespace storage, believing it would prevent scientists from doing "wildly inefficient things like storing data in three different locations," and it was the job of the IT department to provide it. "I'm over that," he says. "I think in general we've done a bad job of encouraging users to actively take responsibility for their data."

It remains "wildly inappropriate and impossible" for IT staff to make data classification decisions, or to know the value of a piece of data and if it can be moved outside of a geographic boundary, he continues. Scientific data handling—the actual work of classifying, curating, tagging, moving and life cycle managing data—needs to be pushed back to the researchers who need and use the information. A notable exception would be large-scale physical data ingest, export, and movements. The IT department can also worry about storage that meets business and scientific requirements, and providing users with self-service metrics, monitoring and reporting tools.

Expect to see greater competition between Intel and AMD, which introduced a new line of microprocessors in 2017. The advent of AI and ML has also brought a flood of new processors to the market and the emergence of five hardware vendors trying to differentiate themselves in this space by building custom silicon and software frameworks.

The landscape is also increasingly complicated for GPUs—now needed for AI and ML as well as virtual desktop infrastructure, data visualization, molecular dynamics simulations, chemical informatics and Cryo-EM—and different products and memory configurations are needed for different tasks. Different numbers of GPUs are also needed per chassis.

"It's time to resurrect the benchmark and evaluation crew," Dagdigian says. The real-world cost of GPUs, CPUs and TPUs need to be factored into cost analyses.

> "We can't trade away performance in exchange for larger capacity anymore. I think there will be lot of [enterprise IT] storage platform switches over the next couple years."
>
> **Chris Dagdigian,**
> BioTeam Co-founder and Senior Director of Infrastructure

4. COMPILERS, TOOLCHAINS AND SILICON MATTER AGAIN.

Only a few years ago, it didn't much matter where you got your hardware and chips (invariably Intel) because they were all relatively the same, Dagdigian says, outside of some GPU divergence for purposes of visualization versus simulation. "Now we have to care about CPUs [central processing unit] and GPUs and things called TPUs [tensor processing units]."

Produced by

Healthtech Publishing

250 First Avenue, Suite 300
Needham, MA 02494
www.healthtechpublishing.com

5. COLOCATION FACILITIES ARE BEING USED MORE OFTEN.

Economics still favor on-premises for 24/7 scientific workloads, Dagdigian says, at least when cost concerns aren't superseded by capability or business requirements. Cloud-based computing is "easy and well understood and serverless is… transformative, but persistent month-over-month costs in the cloud combined with petascale egress fees make the economics challenging."

One "sign of the times" is that data center and telecom developer Markley Group recently named Steve Litster, formerly global lead for scientific computing at Novartis, as its chief technology officer. BioTeam has an active on-prem to colocation project with the Markley Group, Dagdigian notes. The colocation trend has lots of drivers, including the high cost of new builds ("tens of millions of dollars and you usually have to build more than one facility") or upgrades to on-prem facilities, poor cloud economics for some workloads and use cases, consolidation activities, and the ability to aggregate cloud traffic in a colocation suite.

6. LIFE SCIENCE HPC STANDS APART FOR THE SHEER SIZE AND DIVERSITY OF ITS DOMAINS AND WORKLOADS.

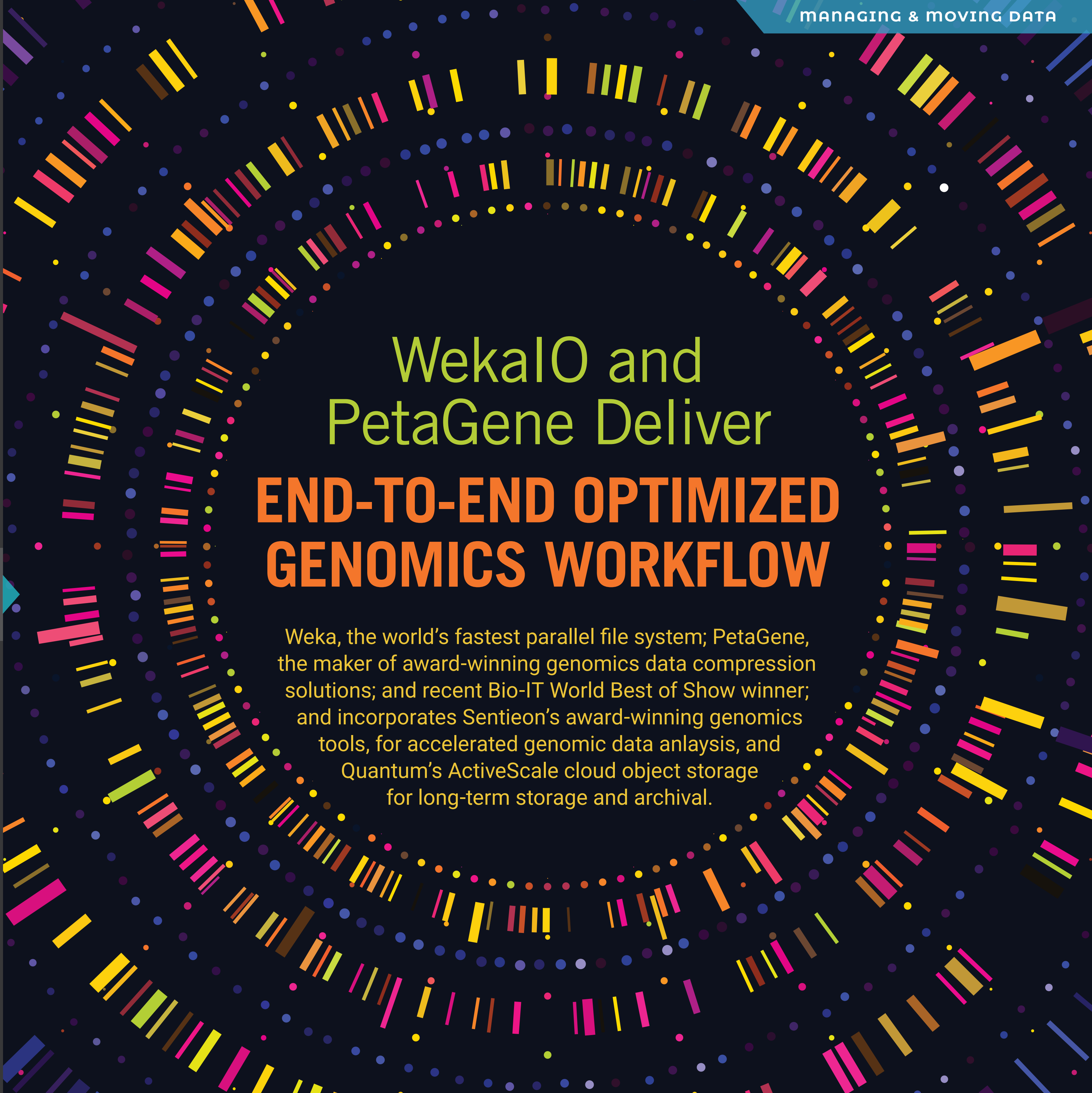
In other industries, the world of HPC and supercomputing has a modest set of dominant and domain-specific codes and «the application landscape is approachable,» says Dagdigian. Not so in the life sciences, which is characterized by "crap code," vast numbers of applications (more than 600 spanning 10 domains), highly specialized subdomains and infrastructure requiring support.

"Individual scientists can now swamp a leadership-class supercomputer with completely valid research questions," Dagdigian says. "That's not sustainable," which is why governance and service scope constraints will become more prevalent.

Dagdigian adds that he favors service-oriented scientific support organized around use cases and end user requirements rather than technological expertise. This "team of team" approach to service delivery is "a great way to blow away traditional IT silos."

Inside
Bio·IT World
The Quarterly eBook of Bio-IT World's
Most Trending Articles

Produced by
Healthtech Publishing
250 First Avenue, Suite 300
Needham, MA 02494
www.healthtechpublishing.com



WekaIO and PetaGene Deliver

END-TO-END OPTIMIZED GENOMICS WORKFLOW

Weka, the world's fastest parallel file system; PetaGene, the maker of award-winning genomics data compression solutions; and recent Bio-IT World Best of Show winner; and incorporates Sentieon's award-winning genomics tools, for accelerated genomic data anlaysis, and Quantum's ActiveScale cloud object storage for long-term storage and archival.

Inside
Bio·IT World
The Quarterly eBook of Bio-IT World's
Most Trending Articles

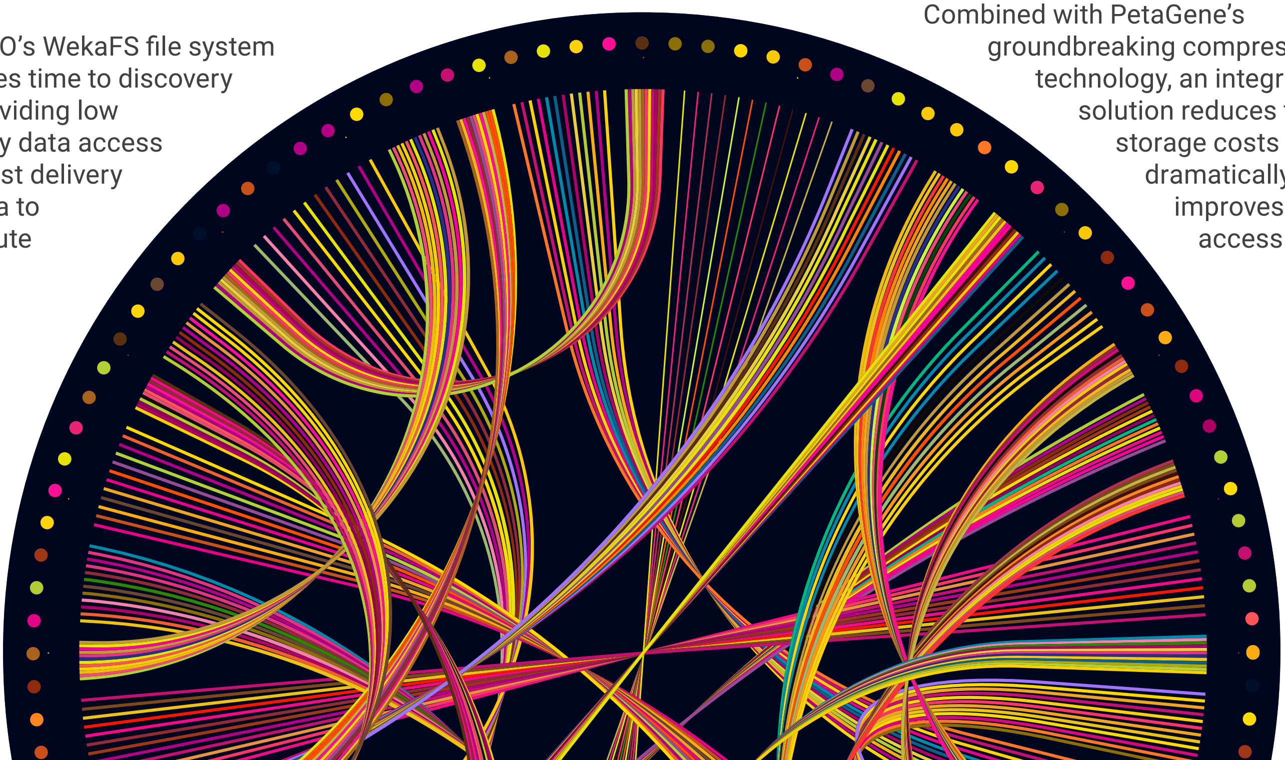
Produced by
Healthtech Publishing
250 First Avenue, Suite 300
Needham, MA 02494
www.healthtechpublishing.com

The cost of genome sequencing has dropped dramatically, resulting in an explosion of genomic data which if stored on legacy NAS storage systems can be prohibitively expensive. Current analytics platforms struggle to process these massive amounts of data in a timely manner, and storage costs dominate the budgets of large genomics applications. As storage costs escalate and money gets diverted to pay for infrastructure, the pace of discovery slows. Together, WekaIO, Quantum, Sentieon, and PetaGene offer the genomics industry a scalable, robust, and high-performance solution that delivers performance that legacy NAS systems cannot offer as well as a cost savings model that allows for the research to continue rather than investing in more storage infrastructure.

WekaIO's WekaFS file system reduces time to discovery by providing low latency data access and fast delivery of data to compute servers, eliminating the I/O bottleneck and the CPU starvation problems common to genomic and cancer research workloads. With a single namespace that spans local storage and the cloud, WekaFS delivers simplified management and data protection. Its performance is 3x that of local file systems and 10x that of traditional NAS. Together with PetaGene compression and integrated tiering and remote backup to the cloud with Quantum ActiveScale object storage, WekaIO provides unprecedented storage performance and capacity scaling for genome sequencing workloads.

"We are excited to share our work with PetaGene for the life sciences community at BioIT World," said The Business Development Team at WekaIO, in a press release. "Genomic workloads are among the most challenging for storage systems with billions of small files and intense metadata operations. Our software delivers extremely high bandwidth and IOPS performance at a fraction of the cost of NAS appliances. Combined with PetaGene's groundbreaking compression technology, an integrated solution reduces total storage costs and dramatically improves data accessibility,

> **"Genomic workloads are among the most challenging for storage systems with billions of small files and intense metadata operations."**
>
> WekaIO



Produced by
Healthtech Publishing
250 First Avenue, Suite 300
Needham, MA 02494
www.healthtechpublishing.com

helping to accelerate the pace of research and discovery."

PetaGene genomic data compression provides up to 90% reduction of BAM and FASTQ.gz file sizes, without any loss of information, resulting in greater than 50% net savings in overall storage costs. In addition, PetaGene compression technology reduces transfer times of genomic data by 60% to 90%. Whether these compressed files are stored locally or in the cloud, PetaGene's PetaLink technology provides transparent and secure access to this genomic data to all applications, tools, and pipelines without modifications to established workflows.

> **"Our work with WekaIO results in a storage solution for genomics and life sciences that is easy to manage and combines industry-leading storage density and performance with breakthrough scale and economics."**
> **Vaughan Wittorff,** CCO and Co-founder of PetaGene

"Our work with WekaIO results in a storage solution for genomics and life sciences that is easy to manage and combines industry-leading storage density and performance with breakthrough scale and economics," Vaughan Wittorff, CCO and Co-founder of PetaGene, said in an official statement. "Our ability to provide lossless compression and workflow transparency of genomic data combined with the high performance of both PetaLink and WekaFS is an infrastructure improvement that will benefit the entire genomics industry. Furthermore, the Sentieon tools offer speed and accuracy of the analysis of genomic data. PetaGene seeks to work strategically to provide the genomics industry with novel solutions and we are excited to support WekaIO with this capability demonstration."

PetaGene also announced they have become a NetApp Alliance Partner. Combining PetaGene's expertise in compression techniques specifically designed for genomic data with NetApp's leadership in data services across hybrid and multi-cloud environments offers powerful new ways to simplify genomic data management. The result is improved performance and reduced costs when conducting research using the enormous datasets created by the explosion in genomic sequencing.

In an official statement, Dan Greenfield, CEO & Co-founder of PetaGene, commented, "Partnering with a leading data services provider such as NetApp will allow holders of genomic data to access ready to deploy solutions for the storage and management of their data, with the benefits of our compression technology already built-in."

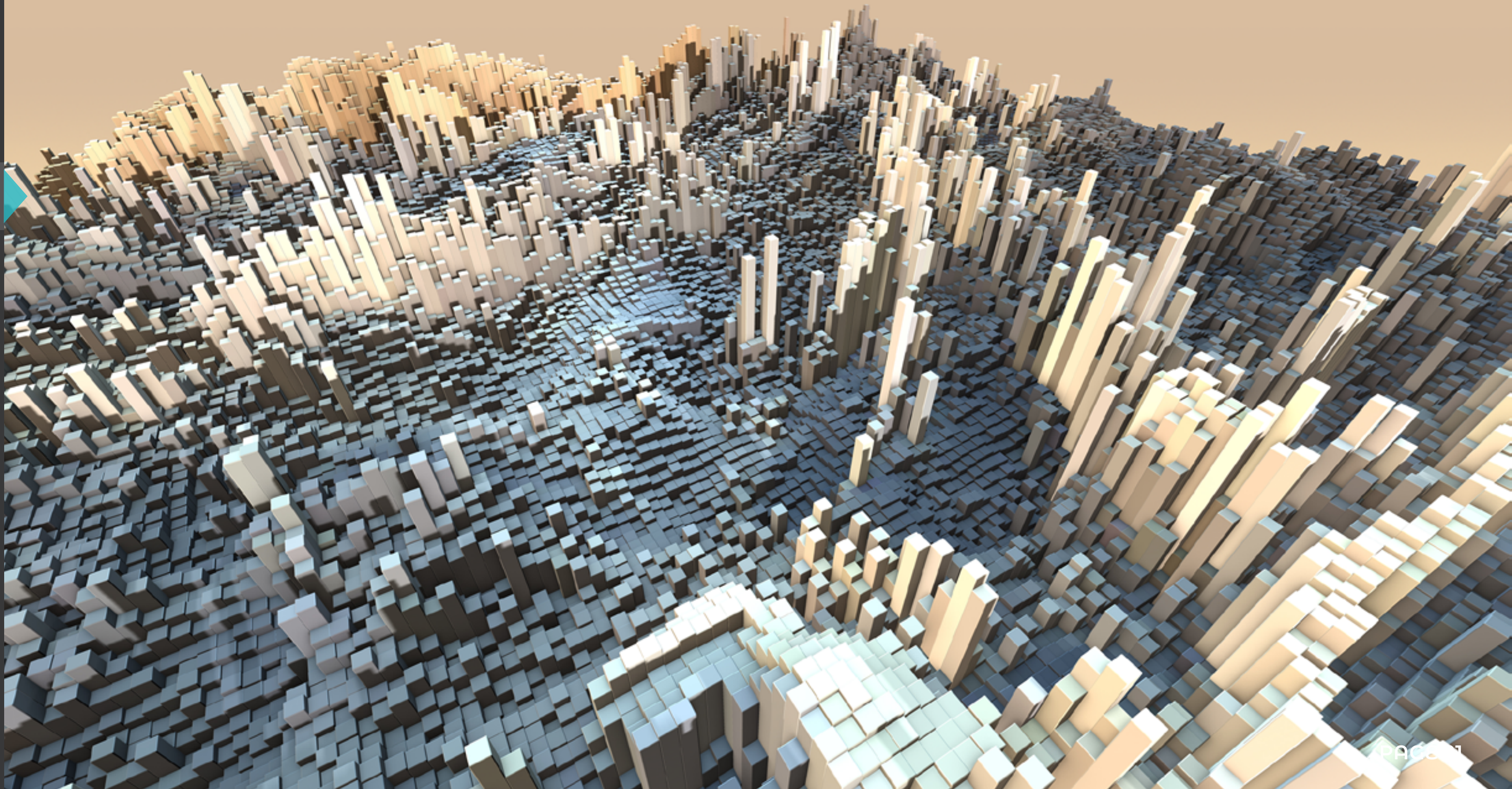
Inside
Bio·IT World
The Quarterly eBook of Bio-IT World's
Most Trending Articles

Produced by
Healthtech Publishing
250 First Avenue, Suite 300
Needham, MA 02494
www.healthtechpublishing.com

DATA is FOREVER

BY JOE STANGANELLI

For attendees at this year's Bio-IT World Conference & Expo, perhaps the biggest focal point was the near-permanence of data—particularly juxtaposed against the limited contemporaneous uses for any given dataset in healthcare and life-science IT. Genomics, medical imaging, and other healthcare and life-science data comprise some of the biggest of the Big Data in the world. Compounding the problem is that, as a matter of course, clinicians have to go back several years (if not decades) to find experiments and other records. And yet, for every decades-old dataset so accessed, many more are recorded and left to gather digital dust "just in case".



Produced by
Healthtech Publishing
250 First Avenue, Suite 300
Needham, MA 02494
www.healthtechpublishing.com

"Science still has the data-hoarding problem because so much of a scientist's career progression [depends] on their publication history," observed Chris Dagdigian, BioTeam Co-Founder and Senior Director of Infrastructure, while speaking on this year's BioTeam panel at the conference. (See Bio-IT World's coverage here.)

"Data is forever," Linda Zhou, Director of Research and Life Sciences Solutions for conference exhibitor Quantum, told Bio-IT World in summation—before following with a limitation: «When you give me a file, I don›t assume the drive is going to be there forever… We don›t hardcode data.»

PLAN AHEAD FOR MIGRATION

Consequently, Zhou explained, the best solutions lend themselves well to data migration across lifecycles. In particular, Zhou advocated for object storage for the kinds of large, unstructured datasets common to the bio-IT field—because so many parts make up the whole. Meanwhile, other exhibitors and speakers were bullish on containerization, largely because of the ease with which containers—and their data—may be migrated; additionally, they allow researchers and data scientists enough sandbox room to play without amassing too much technical debt.

"You can't just keep buying storage—because individuals may think it's infrastructure," Stewart Sherpa, an inside sales representative for DataFrameworks (another conference exhibitor), told Bio-IT World. "But it's not."

Indeed, organizations have run into trouble before by over-relying on expensively bespoke storage architectures that obsolesce and/or cloud vendors with prohibitive lock-in terms.

"The sheer amount of data that you are uploading to another provider cannot be moved overnight," Lance Smith, Associate Director of IT at Celgene, warned attendees in a breakout session he presented. "On a day-to-day basis, you're just not going to be moving clouds… and if you get the cloud wrong, you will get a bill that is insane."

For instance, life-sciences IT consultant Chris Dwan recently related to Bio-IT World a woeful anecdote of a client firm that had gone all in on cloud storage—but had so bungled its subscription finances that it wound up paying its entire budget every month just to keep the cloud provider from deleting the company›s data.

INFORMATION VS. BOXES

Dwan and others told Bio-IT World that this Hobson›s-choice scenario came to be because the company›s IT department was uninvolved in the purchasing decision—reflecting some of the potential dangers of letting researchers and bioinformaticians take the DIY approach to IT.

"The data-science group should really focus on the information, and not so much [on] the hardware," said panelist Jerald Schindler, Vice President of Biostatistics at Alkermes, at the conference's day-two plenary keynote session on defining data science. "It's about the information. And so what you want are people who are focused on the information—not the box that it's in."

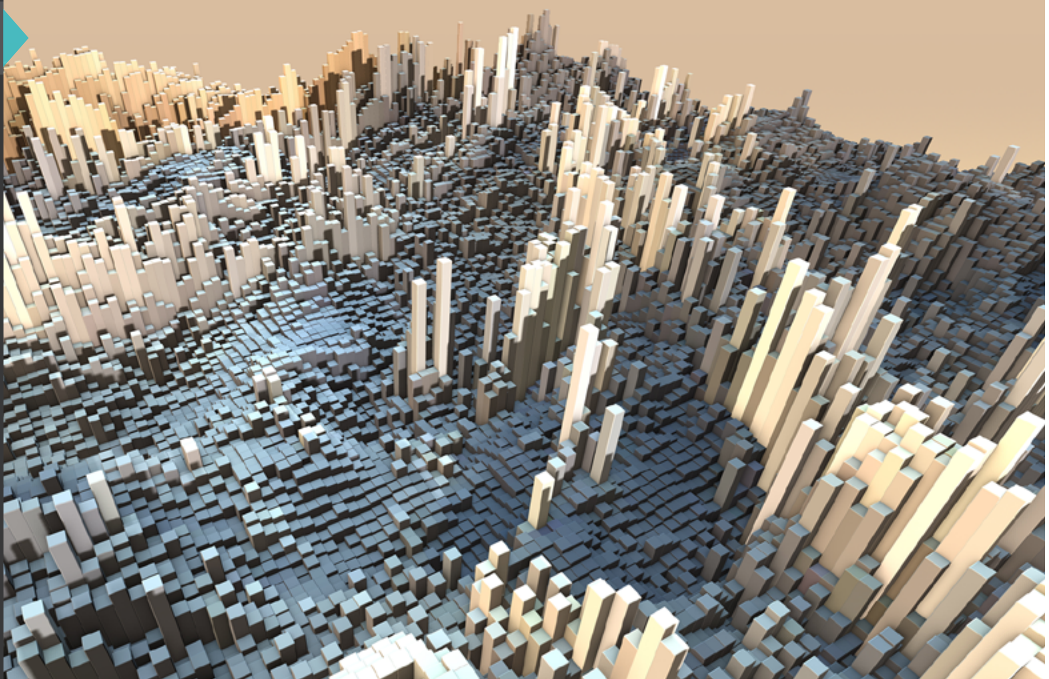
> "The sheer amount of data that you are uploading to another provider cannot be moved overnight. On a day-to-day basis, you're just not going to be moving clouds . . . and if you get the cloud wrong, you will get a bill that is insane."
>
> **Stewart Sherpa,** DataFrameworks

Produced by
Healthtech Publishing
250 First Avenue, Suite 300
Needham, MA 02494
www.healthtechpublishing.com

Not everyone at the conference, however, necessarily took the black-and-white view that data scientists must do only data science and IT must do only IT. Sherpa, for his part, explained that the one impacts the other—and that the people who put things into boxes effectively become subject-matter experts on the boxes that they need.

"I think people are asking IT to make these decisions—but they [in IT] are not the ones generating [the data]," said Sherpa of cloud- and other storage-buying processes.

> "I think there's a big revolution coming in terms of self-service. It's going to be like email. Everyone had to start learning how to use email. There [are] a lot of tools coming out for the . . . non-coders."
>
> **Tanya Cashorali,** CEO and Co-Founder, TCB Analytics



EMPOWERING DATA USERS

Indeed, Sherpa's remarks reflect the long-running trend in the bio-IT market that vendors have picked up on—that scientists demand technological empowerment without having to run to IT for every little thing. Moreover, Dwan offered the above client story as an example of how important it is for central IT to appropriately educate and empower the scientist-users in its organization—a philosophy echoed on this year's BioTeam panel that Dwan himself moderated.

"I think there's a big revolution coming in terms of self-service," said BioTeam-panel panelist Tanya Cashorali, CEO and Co-Founder of TCB Analytics. "It's going to be like email. Everyone had to start learning how to use email. There [are] a lot of tools coming out for the…non-coders."

"I think, for me, the biggest resource gap is [that] we're not applying enough human power to managing our data [or] curating our data to help whip it into shape," added Dagdigian. "It's cheaper to buy storage to keep storing crap forever."