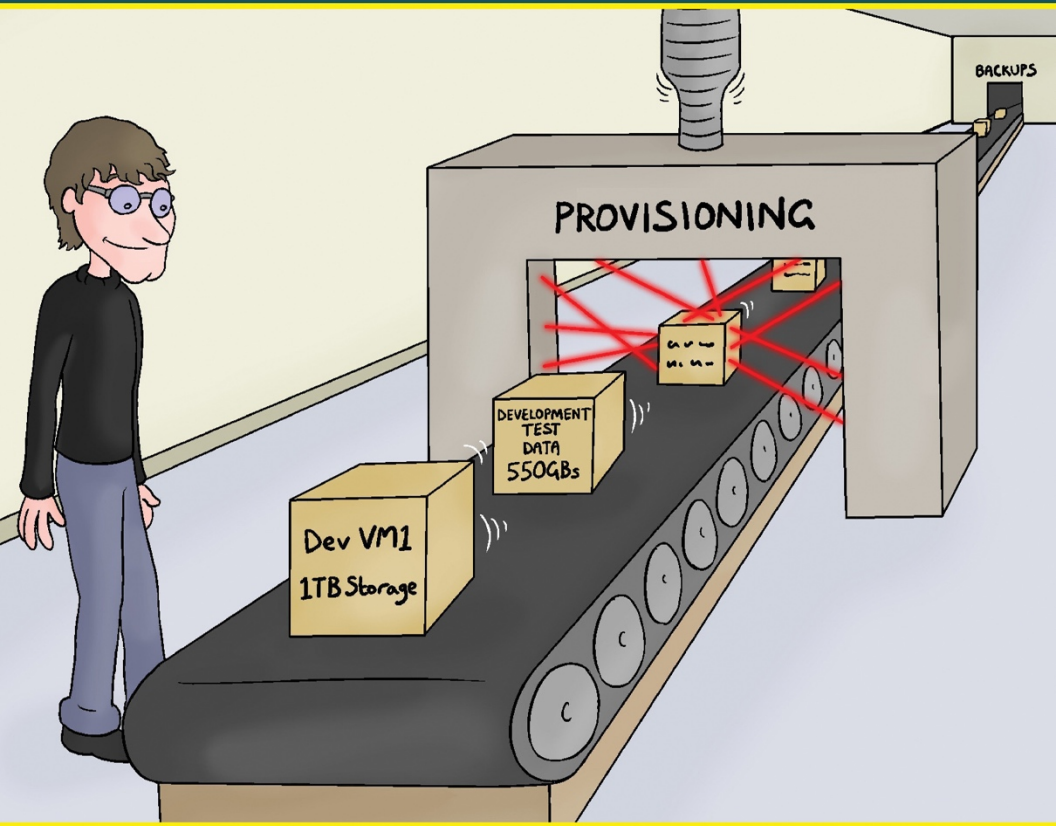
ConversationalGeek®

Conversational Test Data Management

By Mike Vizard (Longtime observer of all things DevOps)



In this book, you will learn:

- Why IT struggles to meet the DevOps pace of provisioning data
- How mass data fragmentation exacerbates the problem
- How IT can quickly accelerate application development

Sponsored by

COHESITY

Sponsored by Cohesity

Cohesity ushers in a new era in data management that solves a critical challenge facing businesses today: mass data fragmentation. The vast majority of enterprise data — backups, archives, file shares, object stores, and data used for dev/test and analytics — sits in fragmented infrastructure silos that make it hard to protect, expensive to manage, and difficult to analyze. Cohesity consolidates these data silos onto one web-scale platform, spanning across on-premises, cloud, and the edge, managed through a single UI.

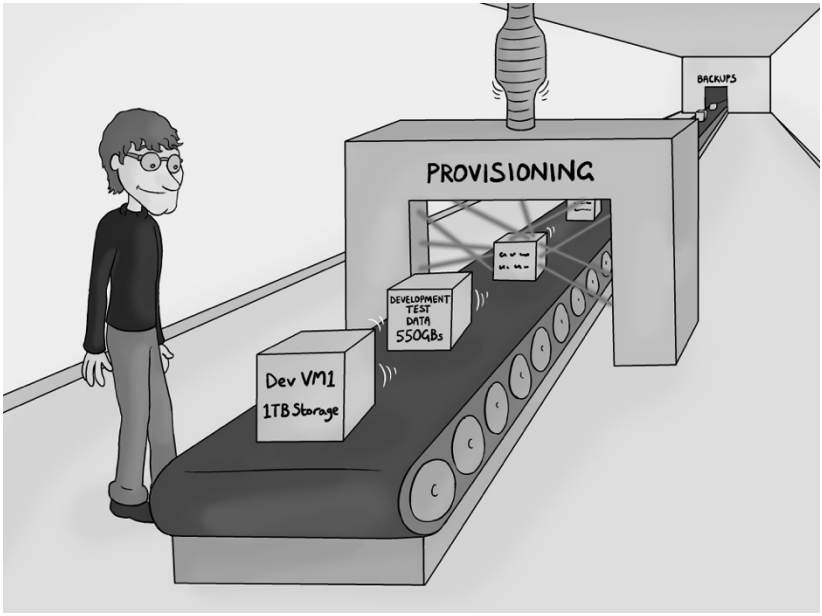
Cohesity is uniquely empowering organizations to run apps on the same platform, and makes it easier than ever to backup and extract insights from that data. Cohesity dramatically simplifies backups and reduces TCO, makes instant recovery possible, and ensures business continuity. Cohesity is a 2019 CNBC Disruptor, Leader in The Forrester Wave™: Data Resiliency Solutions, Q3 2019, and Visionary in the 2019 Gartner Magic Quadrant for Data Center Backup and Recovery Solutions.

COHESITY

For more information, visit
www.cohesity.com

Conversational Test Data Management

Mike Vizard



ConversationalGeek®

Conversational Test Data Management

Published by Conversational Geek® Inc.

www.conversationalgeek.com

Additional Information

For general information on our other products and services, or how to create a custom Conversational Geek book for your business or organization, please visit our website at ConversationalGeek.com

Publisher Acknowledgments

All of the folks responsible for the creation of this guide:

Authors:	Mike Vizard
Project Editor:	Pete Roythorne
Copy Editor:	Pete Roythorne
Content Reviewers:	Raj Dutt
	Srini Sekaran

IT's Challenge of DevOps, Test Data, and Mass Data Fragmentation

"You want this by WHEN??!?"

As application development continues to evolve to address the requirements of digital business initiatives, the rate at which applications are being developed and updated has dramatically accelerated.

Today, applications are being developed faster, along with the number of teams building applications. To sustain this agile and rapid

workflow, DevOps teams are relying on IT organizations to continuously feed them with data. With the legacy approach of data management and application development, this exponentially increases the amount of data being copied across multiple infrastructure silos.

The Need for Lots of Test Data

With each additional software build comes the need for high-quality test data where fidelity is of critical importance. Developers need this test data as quickly as on-demand, making the legacy approach of a "request-fulfill" model truly outdated. DevOps teams need test data that represents current operations, so utilizing older copies of data isn't entirely viable.

Data provisioned from backups is designed to deliver just-in-time data to developers. The need for data provisioning has quickly morphed from an occasional need to a continuous need.

The Problem of Mass Data Fragmentation

Mass data fragmentation, or the proliferation of data across locations and silos, is a significant obstruction to quickly provisioning test data to developers. This proliferation of data across the organization is caused by several factors:

1) Data is Siloed – Particularly with enterprises, infrastructure, workloads, and data sources are varied and scattered. The possibility of further silos and inefficiencies arise when there is a misalignment between DevOps teams and infrastructure. The potential is there for every build of an application to be yet, another, silo.
2) Copies are Multiplying – Over time, as developers create new builds and versions – each needing test data – data sets are being copied or reprovisioned from IT.
3) Data Resides Everywhere – Copies of data used by DevOps teams can end up residing on-premises, in the cloud, and on multiple storage environments.

The result is continual growth, increased storage costs, and the potential for security and compliance issues – all the while IT has no visibility and loses control of that data. And it's IT that ultimately is responsible for the data.

Why isn't IT on top of the issue of mass data fragmentation?

Why IT Isn't Meeting the Need

The challenge organizations of all sizes now face is finding a new approach to data protection that has been designed from the ground up to support the transition to DevOps teams without slowing down the rate at which applications are built or deployed, while at the same time helping to avoid mass data fragmentation.

According to Cohesity's Market Study, IT spends an average of 16 **extra weeks per year** managing data, such as the test data used by DevOps teams.

There are three primary data protection challenges I'll discuss that need to be addressed in order to find that balance between IT's ability to deliver data protection and DevOps teams' need to evolve all while working to keep data fragmentation under control.

1. Infrastructure and Data Sprawl

DevOps teams need their own copies of production data to develop and test an application's functionality and performance. In a legacy world, this is done essentially using two infrastructures – one for dev and one for production. The result is a need to maintain duplicative

data sets, systems, and applications – copying production data to developers as needed.

A recent survey of more than 900 senior IT decision makers conducted by Cohesity, found that 87% of respondents believe all their data is either already fragmented across silos and that it is, or soon will be, at the point where it will be nearly impossible to manage long-term[1].

In an ideal world, the backup and recovery platform employed within the organization should enable IT operations teams to leverage backed up data and infrastructure to support application development and testing using the same data, within the same environment. This will eliminate the need to maintain a separate environment for dev/test, while also reducing unnecessary data copies. This issue breaks down into a few specific points of contention as IT teams attempt to address the issue of Infrastructure and data sprawl:

- Number of Copies of Data: The first issue IT teams that support application development and testing teams need to resolve is finding a way to make data more accessible. Modern backup systems can now create clones of data by creating a reference to the original data rather than having to create an identical, full-size copy of that data hundreds of times over. These clones typically have little to no impact on storage footprint.

  DevOps teams can gain access to what essentially amount to virtual instances of production data in a few minutes. Best of all, they gain access to more realistic copies of production data rather than having to work with synthetic copies of data. The result is a better application experience that reflects a real-world use case.

[1] Cohesity, Market Study (2019)

- The Size of the Data: Just as importantly from a cost perspective, the amount of storage space the clones consume is comparatively minimal. We live in an age where even though the cost per GB of storage continues to decline, the volume of storage being consumed continues to grow exponentially. It's expected that by 2025, approximately 463 exabytes of data will be created each day globally[2]. Most of that data is never going to find its way into any form of persistent storage. However, enough of it will be stored that it will not be uncommon for organizations of all sizes to find themselves managing petabytes of storage, most of which will reside in non-latency-sensitive storage systems.

- Total Cost of Storage: Spending on data storage is expected to increase to $144.33 billion by 2027, representing a 12.5% annual compound growth rate[3]. As such, there's clearly a major economic incentive to reduce the number of storage platforms required. In many cases, simply eliminating all the ongoing license fees for existing storage systems that would be saved by consolidating storage platforms would justify the cost of acquiring a more efficient modern storage system alone. Add on top of that the savings in physical space and energy costs, and it is obvious many legacy approaches to managing secondary storage are more trouble than they're worth.

- Increased Use of the Cloud: Of course, development is no longer confined to on-premises IT environments. Data used by DevOps teams can now reside in public or private clouds, further exacerbating data sprawl and infrastructure inefficiencies.

- Multiple Data Types and Stores to Support: If the expansion of the number of platforms on which data resides is not

---

[2] Jeff Desjardins, "How much data is generated each day?", World Economic Forum, 30 March 2020, https://www.weforum.org/agenda/2019/04/how-much-data-is-generated-each-day-cf4bddf29f/

[3] Credence Research, *Next-generation Data Storage Market Report* (2019)

challenging enough, IT teams now also find themselves contending with multiple types of databases and data stores. The days when IT organizations standardized on a single relational database are long over. Developers have shown a marked preference for open source database technologies such as document databases based on the open source MongoDB project. Those databases are not only a lot simpler to set up and manage without the aid of a database administrator, but they also don't require the approval of anyone in the purchasing department to acquire.

More than 60% of organizations feel there are too many data sources and inconsistent data, with half (50%) struggling with disorganized data stores and a lack of metadata[4].

All this usage of multiple repositories to store data leads to what is known as mass data fragmentation. IT operations teams now need access to backup and recovery tools that can work across multiple database formats.

2. Developers' Need for Speed

DevOps teams can't wait days or week to get the data they need for testing. Most IT teams, however, still tend to prioritize requests based on the service level agreement (SLA) in place. In the absence of a specific SLA, the data will "get there when it gets there". DevOps teams also typically need to refresh data at a rate that internal IT teams can't always satisfy.

The result is IT is only slowing developers down.

IT needs to find ways to be able to respond to provisioning requests at a rate that aligns with DevOps teams and doesn't overburden IT in the process.

---

[4] O'Reilly Media, *The State of Data Quality in 2020* (2020)

To accomplish this, IT first needs to look a bit deeper at why the speed of provisioning is an issue:

- Misalignment Between the Objectives of IT and Developers: Historically, the SLAs attached to provisioning data requests by developers have been relatively low. Of course, no IT leader should be put in a situation where they have to choose between reducing downtime and supporting application development, as both are crucial to the business.

  However, in the case of application development and testing, IT operations teams will quickly discover that given the cost of application developer downtime the speed of provisioning is always going to be of the essence. Because of that issue, a test data management platform should be able to support multiple workloads without impacting the IT environment while meeting the specific SLAs of DevOps teams.

- Snapshot Management: Modern application development and testing environments make extensive use of snapshots of production environments to create up-to-date copies of data. Often, the most difficult challenge is the limited number of snapshots and clones that are supported by a backup solution, making it difficult for application development and testing teams to provision the data needed.

- Automation: In an ideal world, every time there is a build during the application development process a backup process would be initiated. The challenge is the length of time it takes to back up a build. Not all builds are of equal value, so IT operations teams need to be able to automatically apply backup policies based on the relative value of the updates being made to a specific build. In many cases, a build may require hours or processing to complete overnight. The IT operations team is going to want a backup of that build to be kicked off automatically at, for example,

three in the morning rather than waiting until nine o'clock when every other application in the IT environment starts to come to life as employees show up for work. In other words, in today's world, manual operations are untenable. If test data provisioned to developers needs to be up to date with production, then a policy-driven approach to data management is a necessity.

As builds become both larger and more frequently updated in the age of DevOps, the pressure on test data infrastructure is increasing. Test data needs to be provisioned in a manner that's quicker and self-service. Developers need the latest data or data that is in sync with production. In other words, the challenges in test data management today is not simply a challenge of quantity and speed. The quality and relevance of test data is critical.

- Developers Need Self-Service: As corporate data continues to evolve, the idea of developers leveraging stale data is definitely not best practice. Developers need a better way to access fresh data on demand. The legacy process of IT responding to developers' requests no longer fits the working model of a modern DevOps team.

The same provisioning issues also apply to the public cloud. Organizations can now take advantage of the elasticity and economics of the public cloud, while reducing time to market for the new applications. However, the public cloud is not a panacea. In fact, public cloud environments don't just carry over the data fragmentation, operational inefficiencies, and dark data challenges from on-premise environments but also introduce several new challenges.

For the majority of enterprises, infrastructure footprint straddles multiple public clouds and on-premise environments. The result is a need for data mobility between these various environments. A lack of which can lead to dramatic inefficiencies. With modern test data management, the advent of dev/test in the cloud adds additional

roadblocks, including the misalignment of formats among on-premise and public cloud VMs.

This produces a schism between on-premise and cloud environments, leading to manageability strains, a severe impediment to application mobility, and dramatic cost challenges –more copies need to be created and stored. It's clear: the public cloud is not a cure-all, but rather introduces several modern challenges for the enterprise.

3. Ensuring Security and Compliance

Whether data is used in production or in dev testing, if it's valuable to a cybercriminal (and internal threat actors) or is subject to regulation, it needs to be protected. And when IT provisions a copy of, say, customer data containing personally identifiable information (PII) to an environment accessible by contractors and offshore teams, there is both a security and compliance concern about how that data will be accessed and used. IT needs to find a way to meet developers' provisioning needs without losing control of the data.

A recent survey of 2,600 IT professionals conducted by F5 Networks discovered that half are transitioning toward incorporating DevOps practices to some degree.

As IT works to find ways to meet developers' specific backup needs, it's going to need to adopt DevOps practices, particularly when it comes to backups, to, in essence, meet the needs of their "customers" – the developers.

Security and compliance each have their own set of concerns and resulting requirements when it comes to backups. These include:

- Compliance Management: Once the ideal data protection strategy is created, many IT operations will soon discover they have overlooked one crucial aspect. It turns out that allowing developers to see data that contains personally identifiable information (PII) is, by general rule, strictly verboten, especially in highly regulated industries such as financial services and healthcare.

  IT operations should be able to define access policies and revoke access privileges from within the confines of the administrative console. That is especially critical given the amount of turnover that typically occurs within a DevOps team.

  It's also crucial to make sure all PII data that might be accessed by developers is automatically masked. There should be no need to conduct a full audit if the DevOps team never had any access to sensitive data in the first place!

The newly released California Consumer Privacy Act is a great example of how costly non-compliance can be. In the event of a data breach, organizations can be fined between $100 and *$750 per record, per incident*!

- Security: It's also worth remembering the overall security posture of the organization improves dramatically when there is less data strewn across the enterprise. By reducing the number of storage systems, IT organizations reduce the

number of entry points that cybercriminals can exploit to gain access.

At the same time, the cost of securing the environment declines because there's simply that much less physical infrastructure to secure. The sad fact of that matter is that in most organizations storage systems, both in the cloud and on-premises, are the soft underbelly of IT environments that cybercriminals exploit all too frequently.

In an ideal world, IT operations teams would also be able to leverage tools integrated with their storage systems to scan for vulnerabilities in the production environment so that known cybersecurity issues are not simply copied and then pasted all across the enterprise.

IT operations teams should also make certain the file system being employed is creating immutable copies of builds. Any anomalous behavior of usage patterns should generate an alert. Administrators should be able to perform a provisioning of data when required for whatever reason.

Meeting Developers' Provisioning Needs

In the case of DevOps, it's the developers that will very much be dictating the what, where, when, how, and why of data provisioning and environment recovery, as they operate in a far more specific and fast-paced mode – one that IT will need to align with.

In order to meet developers' data provisioning needs, while simultaneously meeting the data protection needs of the organization, IT will likely be required to change a few things about the way they support DevOps teams:

Remotely Manage Data

Developers are notorious for working at odd hours, so IT operations teams that need to support developers should make sure they have access to data management platforms that can easily be managed remotely using, for example, a tablet computer.

Use Developer-Friendly Data Protection Platforms

IT operations teams should also make sure the provider of the data protection platform offers a level of support and service that aligns with their needs. Developers are not going to want to stand by while the IT operations team waits for support from an IT vendor to become available at a time that is convenient to that vendor. Infrastructure should enable DevOps teams. It's critical to provide self-service access to test data.

Use Modern Platforms That Reduce TCO

We've established that meeting the ever-increasing pace of innovation in modern companies is a tall order. In test data management, as enterprises grow their business and teams, provisioning large amounts of test data is a concern. This is a valid concern; after all, data consumes resources, no matter where it resides. Modern platforms, however, have a solution for reducing the amount of storage consumed even if the amount of data provisioned increases. In looking to reduce TCO and increase the return on investments in innovation, IT teams should look for solutions that provide zero-cost clones, making it possible to provision large quantities of test data without consuming much storage.

After reading this book, you should have some context around how vastly different DevOps teams' provisioning needs are from that of standard IT operations. It's imperative that you begin to take steps to determine what your organization's development backup needs look like and what IT needs to do differently with backups in order to meet those needs.

If IT operations teams keep the following issues top-of-mind they should be able to weather any DevOps storm:

- Make it simple to provision data quickly in a way that minimizes storage costs.
- Make certain the underlying storage systems can scale to support thousands of virtual machines.
- Determine how many cloud and on-premises platforms will need to be supported today and tomorrow.
- Ascertain the number of types of databases and data stores that need to be supported.
- Evaluate the system performance to ensure one workflow doesn't impact the other, for example, a recovery workflow impacting application development process.
- Automate as many processes as possible to eliminate manual tasks.
- Determine the right approach to DevOps for your specific organization.

Some of this may be possible by utilizing existing data protection platforms and solutions, modifying the way backups are created, how data is stored, and how provisioning is performed. But, in the end, this may require looking to new platforms that are designed around meeting the varying needs within an organization – including those of developers. This requires a mindset shift. To support DevOps teams' needs by leveraging backups requires IT leaders to think of backups as more than an insurance policy.

The Big Takeaways

Almost every critical business process is now driven by software. As a result, there is a lot more focus these days on developer productivity.

As these teams grow and become more critical to the success of the organization, meeting their needs to deliver quality test data and protect development builds are becoming more important for IT. If left unchecked, mass data fragmentation will take over, increasing the cost and inefficiencies of storage, management, security, and compliance.

Legacy approaches to provisioning data and development recovery needs to be accomplished while minimizing storage costs, reducing data fragmentation, and improving speed of delivery – all while ensuring data security and compliance. It's a tall order; but one that is possible to achieve with modern data protection platforms.

It's not likely an organization is going to be able to simply rip and replace legacy data protection platforms wholesale. Application development and testing, however, provides a use case where the justification for beginning the process of transitioning to a modern data protection platform can be easily made.

The challenge now is to find a way to meet the specific provisioning and data protection needs of DevOps teams – even if it means needing to make that use case for a modern approach to data protection.

COHESITY

Data Management Redefined

One platform, one UI, run apps

www.cohesity.com

Quickly become conversational about test data management.

DevOps lives in a world where fresh test data is needed to accurately develop critical applications. But IT faces challenges when trying to meet the needs of DevOps. In this book, we'll take a deeper dive into why this problem exists and what IT can do about it to both simplify the task of provisioning data while improving the speed of application development.

About Mike Vizard

Mike Vizard is a seasoned IT journalist with over 25 years of experience. He regularly contributes to DevOps.com and previously to IT Business Edge, Channel Insider, Baseline and a variety of other IT titles. Previously, Vizard was the editorial director for Ziff-Davis Enterprise as well as Editor-in-Chief for CRN and InfoWorld.