IBM InfoSphere Virtual Data Pipeline for Test Data Management and DevOps

Highlights

- Near instant virtual copies
- Data refresh
- Self-service access
- Minimal storage consumption
- Test in the cloud

Most organizations increasingly rely on their applications and automation to achieve efficiencies, competitive differentiation increase productivity, and gain larger market share. To keep up with market pressure, organizations need to develop and release higher-quality applications faster. Adopting automated tools can help these organizations achieve continuous integration and maintain uninterrupted release schedules. Achieving success with this requires cooperation and coordination between application development teams and IT operations, and in many cases, creating new roles in the function of DevOps.

DevOps is a methodology and set of practices that unify a team including business leadership, architecture, development, testing, deployment, and operations to be responsible for the creation and delivery of business capabilities*. Currently, over 60% of Fortune 1000 companies look for automation as a primary DevOps initiative and that number is expected to increase. **Dev**elopers (Dev, QA, Build teams), and **Op**erations (UAT, pre-production, release management, application management and monitoring teams) make up the two key constituencies of **DevOps**. These teams require automated tools to achieve DevOps objectives. IBM InfoSphere Virtual Data Pipeline can help organizations achieve their DevOps goals by accelerating data delivery, test data management, application development, and deployment cycles.

Challenges in Application Development and Release Cycles

The majority of challenges in test data management can be categorized into four key areas: Quality, Speed, Control, & Cost Issues.

1. **QUALITY** Product quality suffers when Development and QA engineers are not able to test with the right data early in the development cycle. Not having the right data can lead to gaps in test coverage, potentially missing requirements. Finding critical bugs late in the development cycle leads to software quality issues and delays the release cycle. Organizations must make a tough choice between releasing a low-quality code versus delaying the release to fix the quality issues.

2. **SPEED** The typical process of development includes opening up a ticket to get a copy of production data set, the approval process (1 - 2 weeks), compute and storage provisioning process (1 week), process of creating DB clones and data masking (1 week) can introduce significant delays of up to 4 weeks before the requestor gets access to a copy of the production data set that can safely be used. Many environments such as SAP or Oracle ERP can require over ten copies. The process to create so many copies of production data sets that are multi TeraByte (TB) in size introduces significant delays.

3. **CONTROL** When organizations figure out a way to create many copies of production data sets, operations teams struggle to provide role-based access controls (RBAC) on who has access to copies of which production data sets. Neither do they typically have an audit trail mechanism. Operations also do not have easy, automated mechanisms to mask and refresh copies of production data sets.

4. **COST** When organizations decide to create copies of production data sets, they unfortunately end up creating many physical copies. This leads to massive storage expenses. Additionally, this puts a huge burden on DBAs since they are the ones who have to create DB clones, mask the data, create many copies, and then bring the DB online on all the test environments. This raises the infrastructure and operational costs significantly.

Figure 1 **Dummy Data leads to Quality Problems**

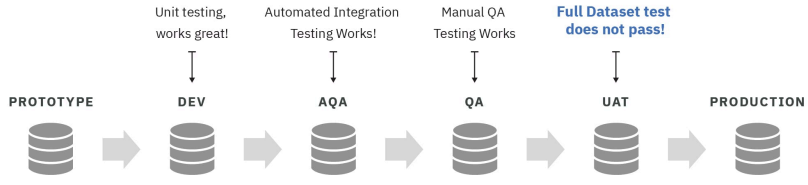
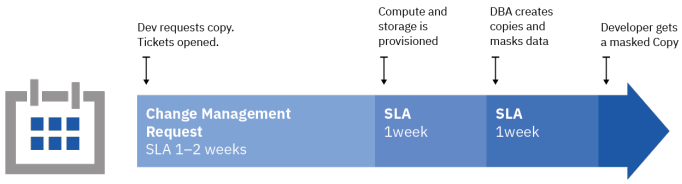


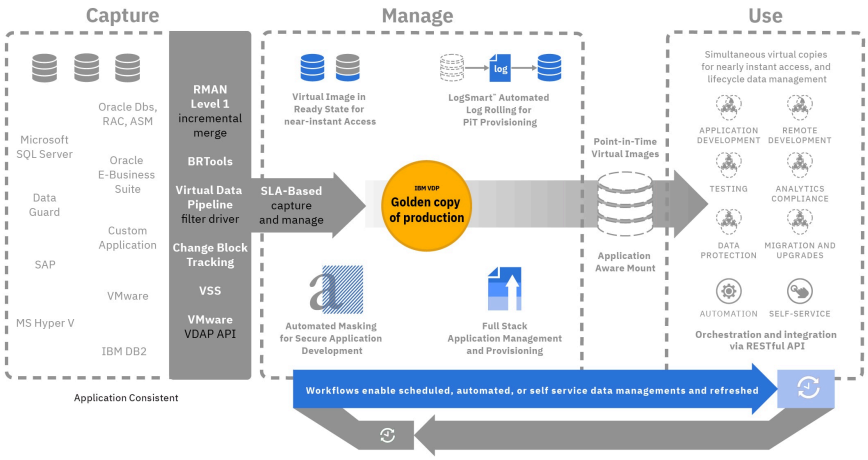
Figure 2 **The Speed Problem**



IBM InfoSphere Virtual Data Pipeline™

IBM InfoSphere Virtual Data Pipeline (VDP) technology delivers application-centric, SLA-driven automation to test data management and DevOps in which VDP decouples the management of data to be used for test/dev from production infrastructure. The result is a single, simple solution—efficiently virtualizing all copies of production data for data protection, application development and test, analytics, and more. VDP manages all of your data through a single platform, from only one golden copy of your production data. VDP allows IT administrators to access a single, masked, any point-in-time copy of your data, through a self-service interface. The approach to properly capture, manage, and use data is radically simple, speeding time to market and reducing costs.

Figure 3 **Virtual Data Pipeline for Testing and Development**



Virtual Data Pipeline brings DevOps technology automation for application development and testing with a single tool.

Capture

IBM InfoSphere Virtual Data Pipeline performs a deep discovery of applications, databases, Virtual Machines (VM), volumes, and file systems. VDP uses the most efficient way possible to capture the data. Typically snapshots along with changed-block-tracking technology are used to capture application-consistent copies of data. This data is then transported at the block level over the LAN or the SAN. This technology is the industry's fastest, most efficient and scalable method of data capture, eliminating the traditional "backup window" while delivering efficiencies previously impossible using traditional technology.

VDP is incremental-forever architecture. An application consistent “golden image” will be maintained and only changed blocks will be sent. This process is performed depending on the application and/or database type and based on administrator defined Service Level Agreements (SLAs). For example, when capturing Oracle Data, VDP will use RMAN Level 1 incremental backup. When systems are virtualized using VMware vSphere, VDP will leverage vStorage API’s for Data Protection (VADP) for Changed Block Tracking (CBT). In instances of physical environments, a lightweight VDP Connector with change tracking driver is used to track changed blocks and provide application consistency.

VDP will capture and hold a virtual image in its native format for efficient usage. All common storage formats are supported, including raw volumes, file systems, and ASM.

IBM InfoSphere Virtual Data Pipeline Captures

- IBM Db2
- Oracles DBs and applications
- Microsoft SQL
- Virtual Machines
- File Systems
- Business applications
- Volumes
- Physical or virtual
- Popular or custom databases
- SAP
- From Oracle Exadata or Dataguard

Manage

VDP manages changed blocks in such a way that a point in time full “virtual” copy can be synthesized in a near instant fashion. Based on administrator driven SLAs, VDP can retain data for a certain duration, deduplicate the data for long term retention, and replicate the data to a remote site for remote development or disaster recovery purposes. Based on the SLAs and defined workflows, administrators can determine how often to refresh the data and run the data through masking scripts or software, such as IBM InfoSphere Optim Data Privacy, automating the end-to-end process of data refresh and masking. VDP LogSmart™ gives you the flexibility of automated log roll-forward to a user-specified point-in-time during the provisioning process.

Consistency groups can even allow this to be performed across multiple databases at once for supported database platforms.

Use

IBM InfoSphere Virtual Data Platform delivers the unique capability of enabling applications to immediately use point-in-time copy data, without the need for a traditional “restore” operation. Application data from any available point-in-time can be accessed on any authorized system for development, testing, analytics, recovery or dozens of other common use cases. VDP offers vast flexibility when accessing data with options for secure, role-based control of data management for multi-tenancy, virtual images from any point-in-time, and various image types.

Automation Requirements to Improve DevOps

Scalability, consistency, data control, and ease-of-use are all fundamental needs when incorporating automation technologies into an existing environment. There are many key requirements that need to be satisfied to help DevOps improve quality and accelerate application development and release cycles.

1) Near Instant Multiple Copies

Provision dozens of near instant copies, with minimal storage consumption

2) Data Masking

Data must be **masked automatically** before DEV & QA teams get access

3) Data Refresh

Enable teams to test on most recent copies of production data-sets with **automated refresh**

4) Self-Service

Enable Dev and QA teams to **browse** and **access** point-in-time, masked copies without having to disturb IT Operation teams

5) Role Based Access Control

Allow IT administrators or application owners to set role-based access control policies on **who** can access **what** masked copies on **which** test servers

6) Bookmarks

Allow users to **bookmark** their test environments so they can roll back to a given point

7) Database Integration

Enable users to specify database **environment customization**, and bring them online with the specified customization automatically

8) Test Data Protection

Enable users to **protect** and save their Dev & QA data sets so that only the changed blocks are protected and stored in the most space efficient manner. It allows users to browse back in time to access various point-in-time data-sets, and then return if desired.

9) Compliance

Many industries have specific regulations that may require the organization to “stand-up” a database or application at a past point-in-time for analytics or forensic analysis.

10) Test in Cloud

Allow masked copies of on-premises data-sets to be made available in a remote office or in the cloud so Dev & QA users can use cloud compute to test

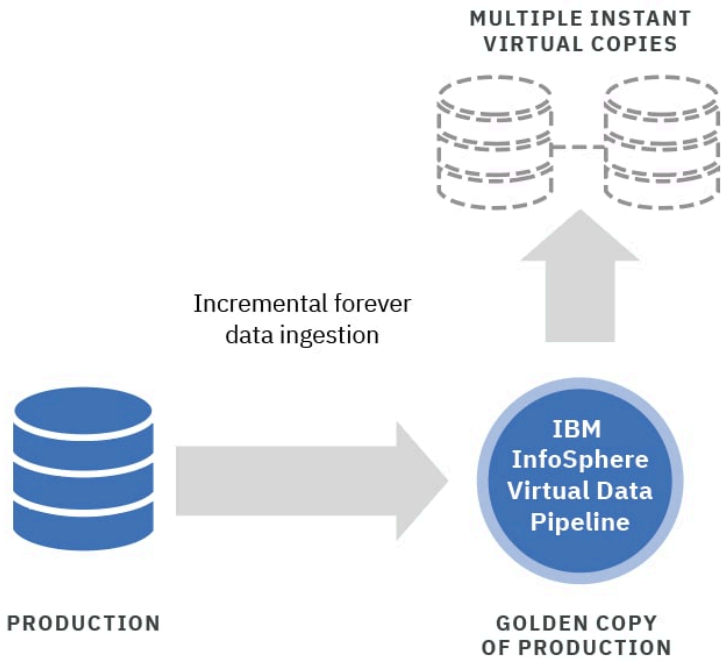
Bring Technology Automation to DevOps

To accomplish their objectives, DevOps leaders are looking more-and-more toward automation and tools that help achieve continuous delivery to their customers and stakeholders.

1. Near instant virtual copies

IBM InfoSphere Virtual Data Pipeline will ingest data, incremental forever, based on a defined SLA and will store a single, golden image tracking each point-in- time. VDP has the data virtualization technology that can synthesize and mount a 'virtual full' image nearly instantaneously to any number of devices. VDP presents the near instant 'virtual copy' to any physical server or virtual machine using iSCSI or Fiber Channel protocol. It does not matter whether it's a 1 TB data set or an 80 TB data set; VDP makes multi-TB data sets accessible in a near instant fashion.

Figure 4 **Near Instant Virtual Copies**



2. Data Masking

Within their databases, organization have sensitive data that can include anything from social security numbers and pay rates to intellectual property or customer financial information. When using data for development and testing, this can require custom logic and rules to mask the data before being presented to another system or handed over to developers and testers. Some may

have simple scripts, and some may use 3rd party tools to mask or sanitize the data, such as IBM InfoSphere Optim Data Privacy. Using the VDP workflow feature, an administrator can specify their data masking script, a data-masking server, and schedule the tasks. Based on the schedule, VDP will mount an image to the data-masking server and invoke the data masking script. Once complete, VDP marks it as 'masked'. All further virtual copies would be created off this masked copy ensuring a "safe" copy will be used.

This ensures that

- Only IT administrators or DBAs control and specify the scripts to mask the data

- Data masking is automated thus eliminating any manual intervention

- Dev & QA engineers access only the masked virtual copies.

3. Automated Refresh

Developers often request fresh or updated data when testing applications. VDP automates the process to provide an updated copy on a schedule or on-demand.

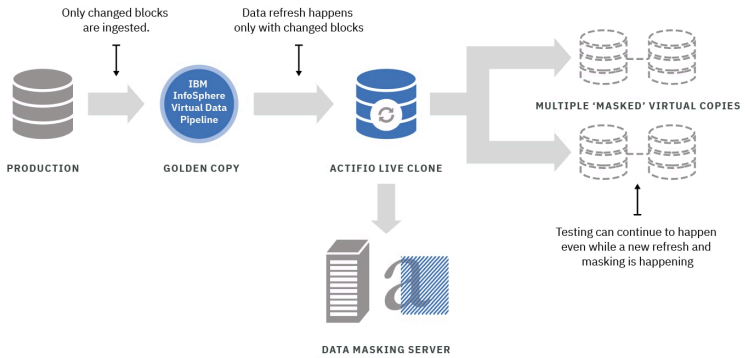
Data refreshes for developers leverage the efficiency of the VDP engine, which ensures only changed blocks are copied. The data movement is SLA driven, which means the data is immediately available when it comes time to refresh the copy being used by a developer. Data masking can even be incorporated into the refresh process, having an updated and masked copy always ready for an on-demand refresh requested by a developer.

4. Self Service

Once the administrator gives access rights, developers and QA engineers have the ability to login to the VDP user interface with their own account. They will be able to browse only the 'masked' data sets of granted databases set by administrator using role-based access controls. A masked virtual copy is selected, mounted to their test server, and they can start accessing the data set. This whole process, being self-service, is not only fast but also eliminates burden on IT staff and DBAs.

Development and QA engineers can also achieve this functionality by leveraging the VDP REST API and/or CLI.

Figure 5 **Automated Refresh**



5. Roles-Based Access Controls

To ensure proper data control, VDP administrators can specify which users can access what data sets on what server. Administrators can also specify that only the masked copies be accessible by particular users. Reporting and audit logs list all user activities, including mount and unmount operations and what data sets were accessed.

6. Bookmarks

VDP allows users to bookmark their virtual copies, and provide and search on a label name. Multiple virtual copies can be created off these bookmarks for multiple people within the team.

7. Database Integration

When mounting a database, VDP allows users to specify database customization parameters as shown below. This makes it easy for users to focus on their test cases instead of burdening DBAs with requests to customize the database environments and bring the databases online in test environments.

When users perform a mount operation, they are prompted for these provisioning options when using common databases such as IBM Db2, Oracle and Microsoft SQL Server. Additionally, users can custom define an application class to include provisioning options and scripts needed to perform application aware mounts for any application.

8. Test Data Protection and Rewind

VDP allows users to maintain multiple Points In Time (PIT) virtual copies, thus allowing them to pick and choose any of the available PIT versions and perform their testing. VDP allows virtual copies to be protected by specifying a simple SLA such as how often to protect and how long to

retain.

9. Compliance

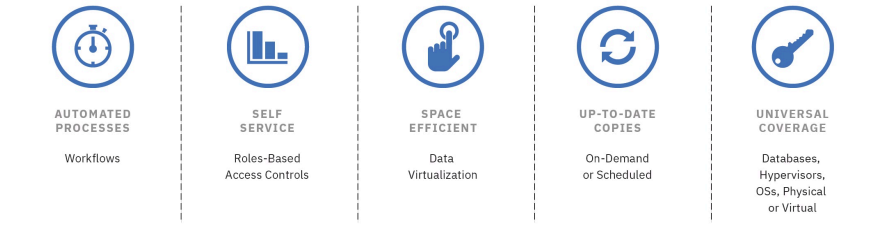
Many industries have specific regulations that may require the organization to “stand-up” a database or application at a past point-in-time for analytics or forensic analysis. VDP provides the ability to keep both a short-term history, or images deduplicated and compressed for long-term retention. This is helpful in responding to Data Subject Access Requests (DSARs).

10. Test and Development in the Cloud

VDP can be deployed as a physical or virtual machine in a remote location including a remote office, service provider, or the public cloud including Amazon AWS or IBM Cloud. Images can then be efficiently replicated from the primary datacenter to the remote or cloud location.

All data is ingested and deduplicated at primary site. Only deduplicated data is replicated to the remote VDP instance, thus ensuring bandwidth needs are minimal. Users can instantiate VMs and mount ‘masked’, near instant virtual copies of data sets from VDP to those VMs, bring databases online automatically, and start testing.

Figure 6 **IBM InfoSphere Virtual Data Pipeline Benefits**



IBM InfoSphere Virtual Data Pipeline Test Data Virtualization Benefits

Accelerated time-to market IBM InfoSphere Virtual Data Pipeline provides data automation tools for DevOps and test data management. This, along with best practices such as continuous integration and delivery, lets organizations bring new software releases faster to customers. This helps organizations add value to customers faster and stay ahead of competition.

Large infrastructure savings VDP virtual copies are storage efficient ensuring significant storage savings in your datacenter, remote office, and in the cloud.

Time savings VDP’s automation for data refresh, data masking, role based access control, database customization and integration ensures significant time savings for DBAs, storage administrators, and IT administrators.

Reliable releases Unit testing by developers, automated build testing, functionality and regression testing can be performed on full virtual copies of production data-sets.

Improved product quality VDP enables critical bugs to be caught early in development life cycle and significantly improve the product quality & predictability.

Control of sensitive data Reducing risk exposure by having a single golden image along with role-based access controls and automated masking provides for a more protected development environment.

Universal coverage A single platform for the entire physical, virtual, application, and multiple databases means flexibility, easier administration, lower costs, and avoidance of additional point-tools and infrastructure silos.

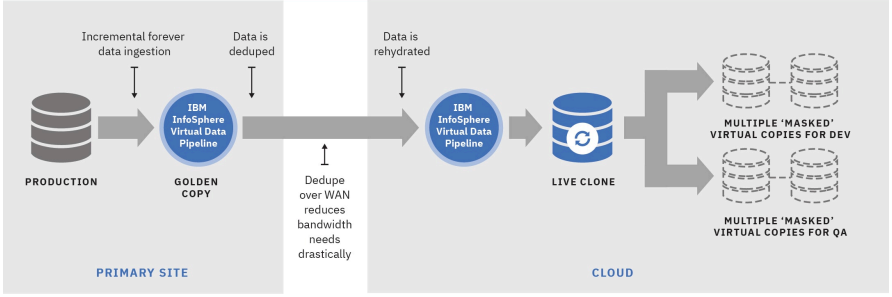
Integrates with DevOps pipeline VDP integrates easily with other automated DevOps tools through an open REST API. Moreover, IBM InfoSphere Optim Test Data Orchestrator (TDO) and VDP can work from a single VDP golden copy to give your organization capabilities to create full virtual copies of databases as well as highly subsetted data environments with only the exact data needed for testing.

Secure and efficient remote development Achieve effective remote development through replication optimization, continuous updates, and automated data masking.

Improved customer satisfaction Improved product quality and predictable roadmap execution leads to a very high customer satisfaction

Business pressures will continue to push developers, QA, and IT to move faster, release error free software more quickly and respond more rapidly to the dynamic needs of the market. The DevOps approach as well as traditional development will increase the need for automation, consistency, functional testing, and flexibility of infrastructure. There are many tools to help achieve Continuous Integration & Continuous Delivery. But, the cornerstone of Continuous Delivery is Continuous Testing – which requires fast and reliable access to a rich set of production-like test data. By utilizing IBM Virtual Data Pipeline, unit testing, functionality testing, performance testing, regression testing, and security testing can be accomplished on virtual copies of production data sets in a safe, secure, and self-service fashion.

Figure 7 **Test and Development in the Cloud**



*IDC MaturityScape: DevOps, June 2014, IDC #249471

Why IBM?

IBM InfoSphere Virtual Data Pipeline technology decouples data from infrastructure, enabling dramatic improvements in business resiliency, agility, and access to the cloud. VDP replaces siloed data management applications with a radically simple, application centric, SLA-driven approach that lets users capture data from production applications, manage it more economically, and use it when and where they need it. In conjunction with other IBM InfoSphere Optim Test Data Management Solutions, users can mask, fabricate, and ensure complete test data coverage. IBM offers the Optim suite to help organizations speed time to market and ensure compliance across the testing cycle.

VDP and Optim are part of the comprehensive, scalable Unified Governance and Integration platform and solutions—available on premises, on cloud and hybrid environments—successfully delivering trusted data for insights and compliance to businesses, governments and individuals.

For more information

IBM InfoSphere Virtual Data Pipeline integrates with the portfolio of IBM Optim solutions, which are designed to help manage data from requirements to retirement. To learn more about IBM Optim solutions, visit: **ibm.com/optim**

To learn more about Unified Governance and Integration visit **ibm.com/unified-governance-integration**. Follow us on Twitter at @IBMAnalytics, on our blog at **ibmbigdatahub.com** and join the conversation #IBMUGI.

Additionally, IBM Global Financing provides numerous payment options to help you acquire the technology you need to grow your business. We provide full lifecycle management of IT products and services, from acquisition to disposition. For more information, visit: **ibm.com/financing**