DELPHIX

Efficient Test Data Management

Marcin Przepiorowski | Senior Technical Principal | October, 2017

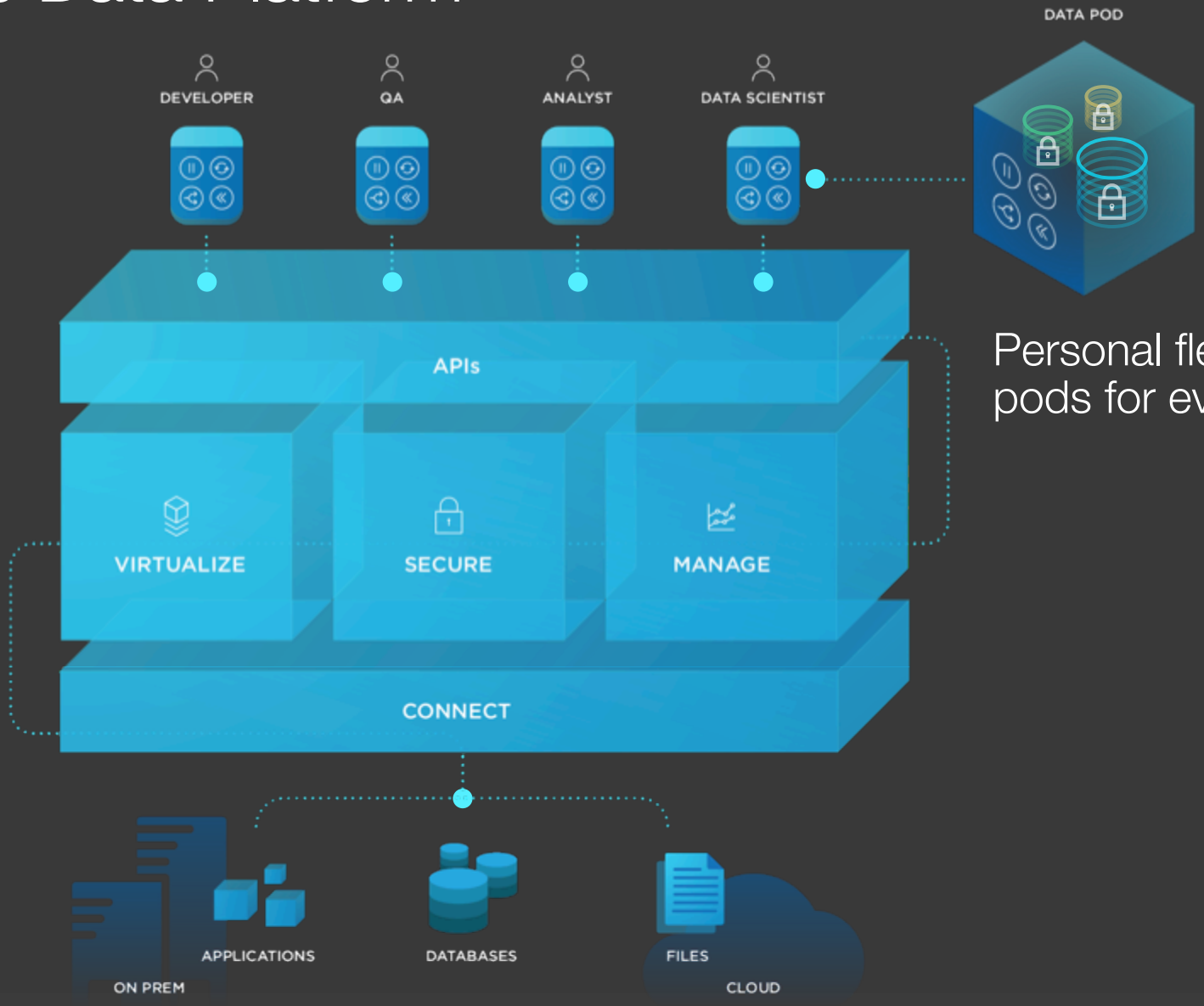
About me

Oracle consultant/DBA since 2000

co-developer of OraSASH – free ASH/AWR like repository

Blogger ???

Delphix Dynamic Data Platform



Personal flexible data pods for everyone

This session is focused on the tools and processes.

No actual database or vendor platform special knowledge is required to gain value from the session.

Real life examples

Test Data Management - Tools and processes

Security

Example 1

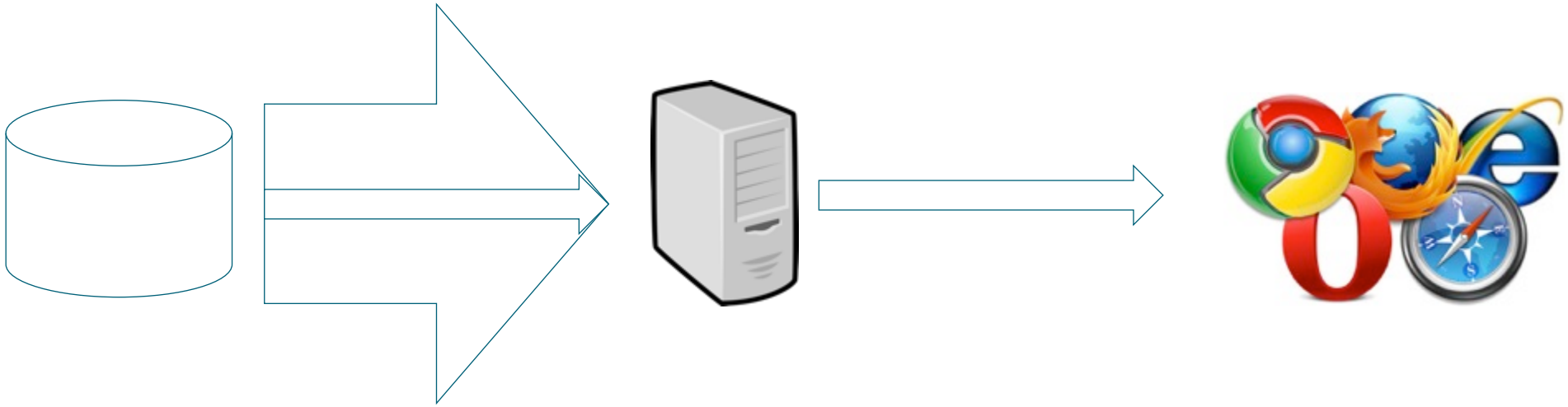
≠

Test data are not like a whiskey or wine
keeping them long doesn't increase a value

- Migration project based on development environment created 2 years before project kick off
- Risk to production migration due to data changes

https://cdn.pixabay.com/photo/2016/03/31/15/23/cabinet_1293245_1280.png https://www.flickr.com/photos/dionhinchcliffe/9505044956

Example 2



- Development project done for 18 months without tests on even 1 % of the production data

- Major issues found to close to the scheduled go – live

Test Data Management

Test Data management is very critical during the test life cycle.

Over 80% of organizations stated that RECEIVING or REFRESHING the data to perform tests was the largest consumer of testing time, (over 90%) leaving actual work to consume less than 10% of the overall testing scenario.

https://www.tutorialspoint.com/software_testing_dictionary/test_data_management.htm

Real life examples

Test Data Management - Tools

Security

What is a good TDM process

A process that assists in delivering a data sets for testing / development on time.

- Delivers a “right size” and secure datasets on time
- Has an ability to quickly isolate and deliver test cases for development to investigate.
- Has an ability to identify code and data changes by versions

Test Data Management

Test data

METHODS	PROS	CONS
Cloning	Relatively simple to	• Expensive in terms of hardware, license and support costs • e-consuming: Increases the time required to run test cases due to large data volumes • agile: Developers, testers and QA staff can't refresh the test data • ficient: Developers and testers can't create targeted test data sets for specific test cases or validate data after runs • collaborative between DBA and testing teams • scalable across multiple data sources or applications • Laborious: Production systems are typically large • Risk ised or misused (developers, testers and QA staff need real alid business reason to access sensitive data such as cor rmation)
Generating synthetic test data	Safe	• Res highly skilled DBAs with deep knowledge of the underlying data tionships that might not be formally detailed in the schema • Tedious: DBA boundary conditions within the synthetic data set to ensure a robust testing process, which adds time to the test data creation process • Challenging: Despite the time and effort put forth by the set or challenging to work with because synthetic test data do retain the proper context • Time-consuming: Process is slower and can be error-p
Subsetting production databases	Less expensive compared to cloning or generating synthetic test data	• Skill-intensive: Without an automated solution, requires y and protect sensitive data

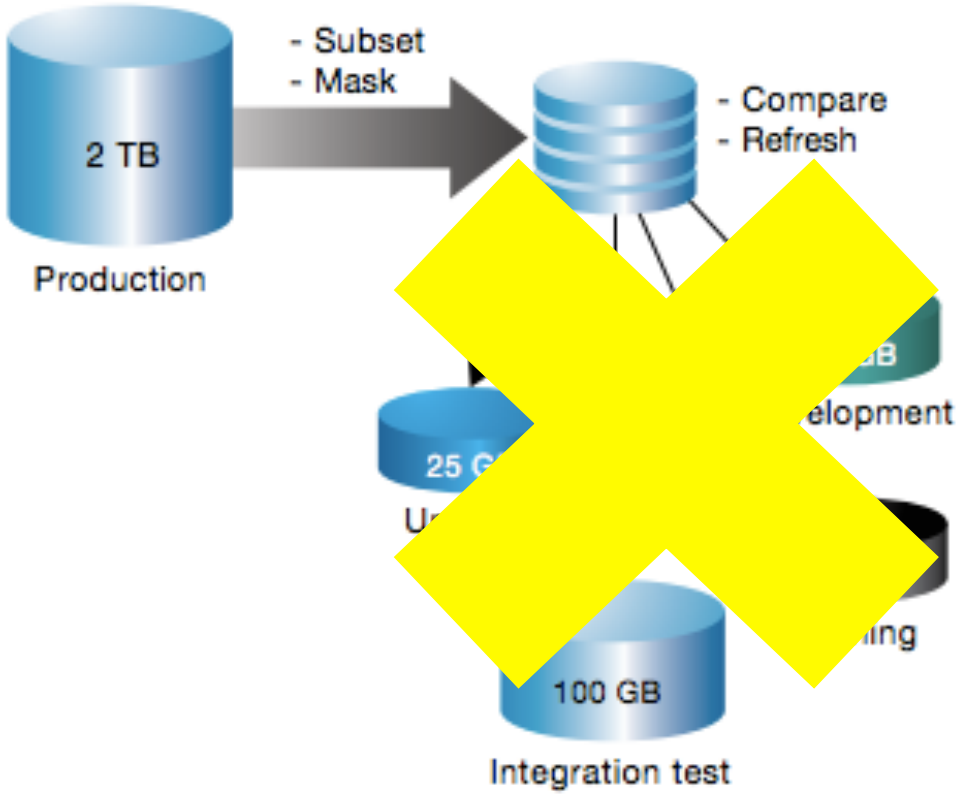
Time consuming

Resource consuming

Skills intensive

http://www.informationweek.com/pdf_whitepapers/approved/1345732672_back_to_basics.pdf

What is the “Right Size”?



Development and test on “unrealistic” amounts of data can create a code quality issues.

Production implementation can end up with scaling issues which are very costly to fix

http://www.informationweek.com/pdf_whitepapers/approved/1345732672_back_to_basics.pdf

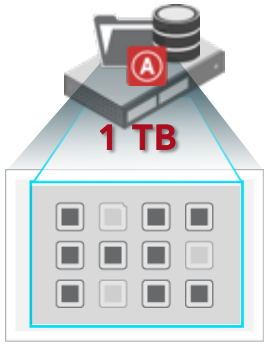
Sub-setting vs clone

- Complicated structures (ex. history of all objects in one table) – not easy to subset
- Sub-setting requires a business logic to be implemented.
- Easiest option for sub-setting is if application has a proper archiving option

Cloning issue

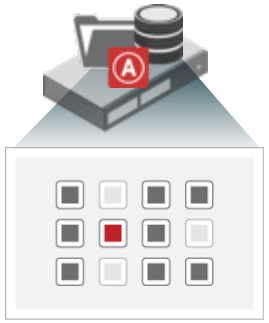
PRODUCTION

Database/App Tier



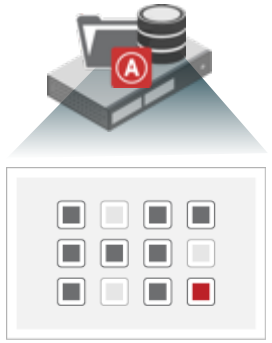
1 TB

QA



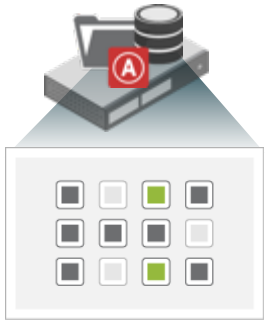
1 TB

TEST



1 TB

DEV



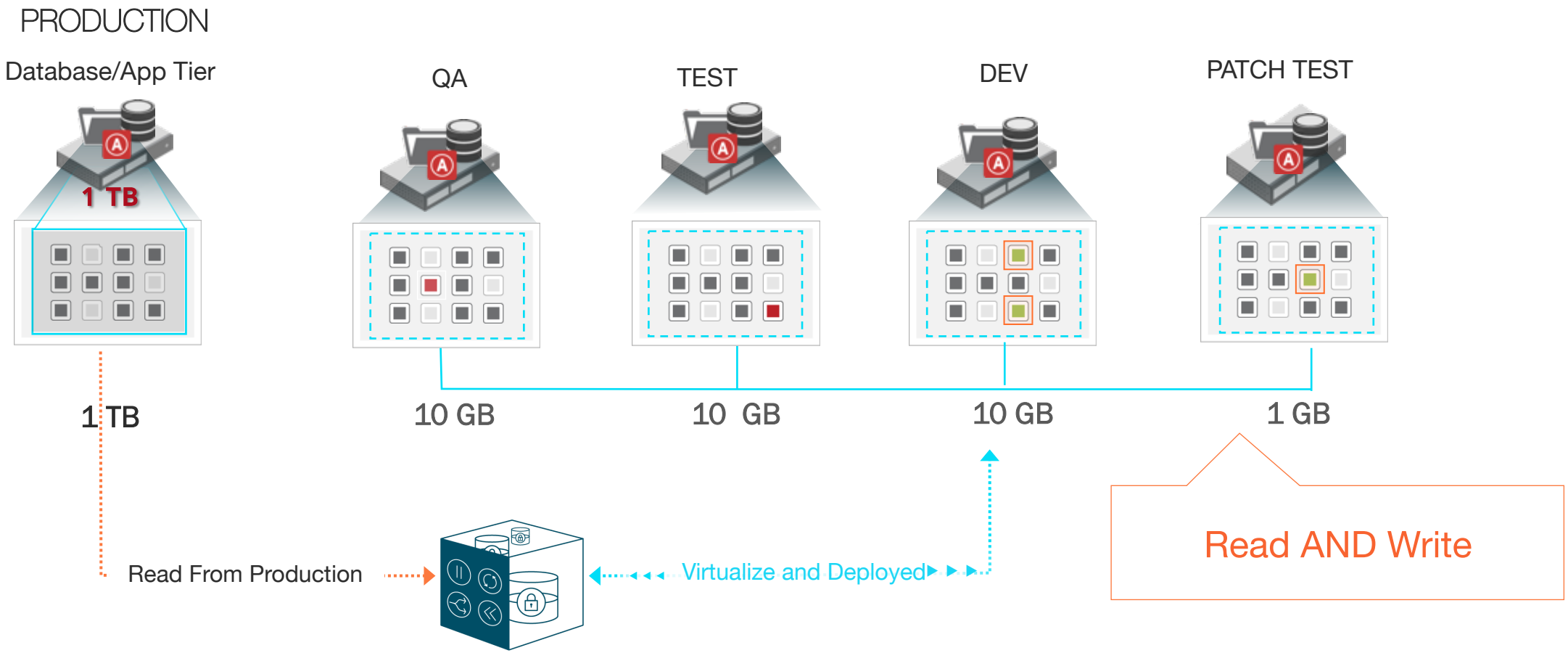
1 TB

PATCH TEST

1 TB

https://www.flickr.com/photos/torkildr/3462607995

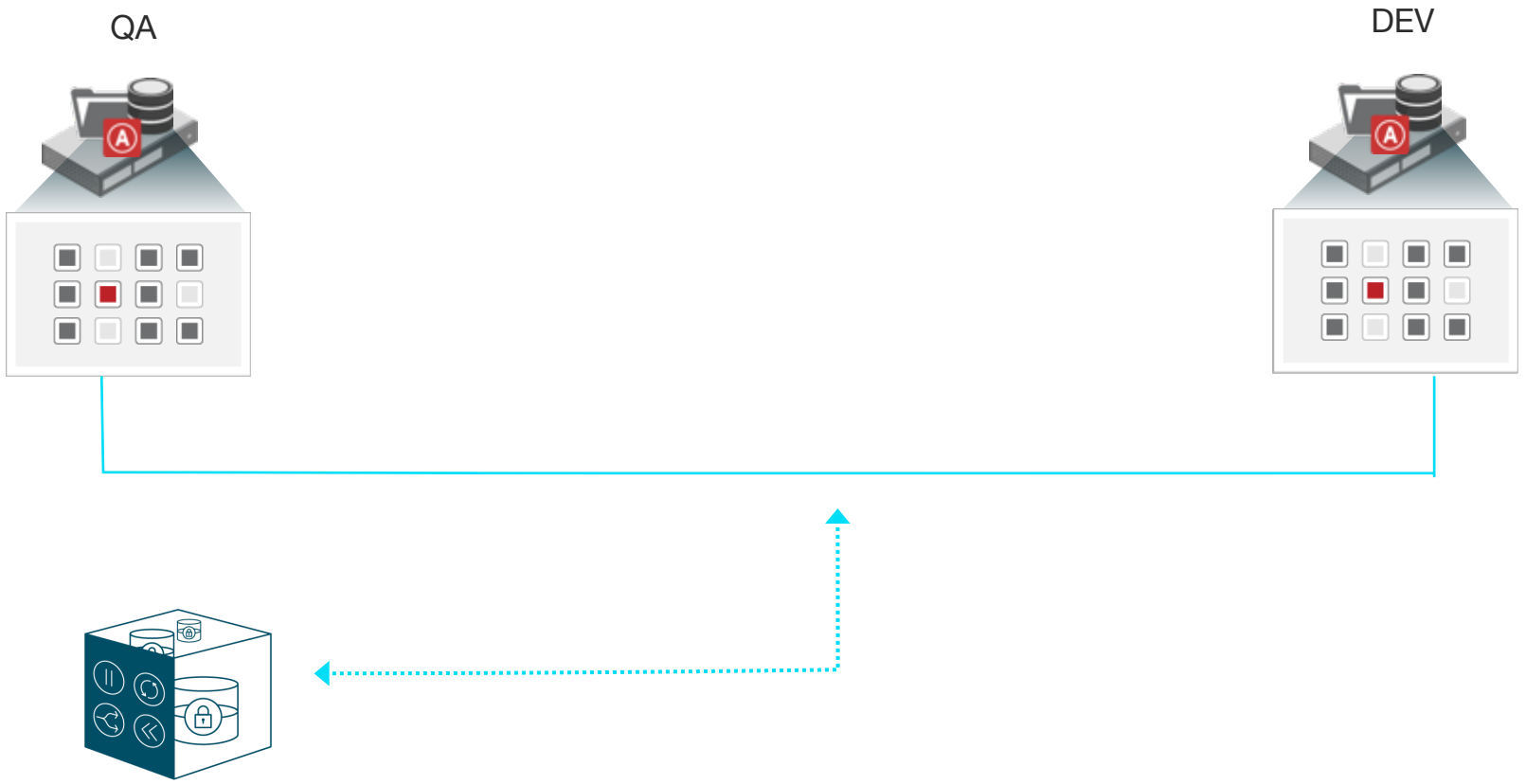
Cloning issue - solution



Isolate and deliver test data

- In case of data related issues, developers need to work on QA system to nail down the problem
- Potential tools and access issues

Virtualization – sharing data between QA and Dev

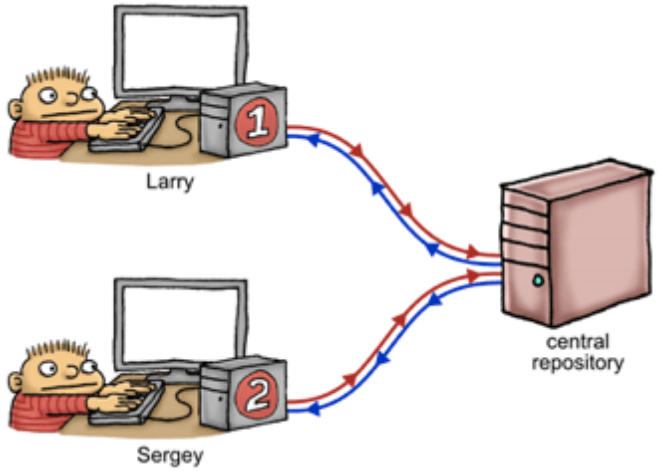


Code and data changes / versions

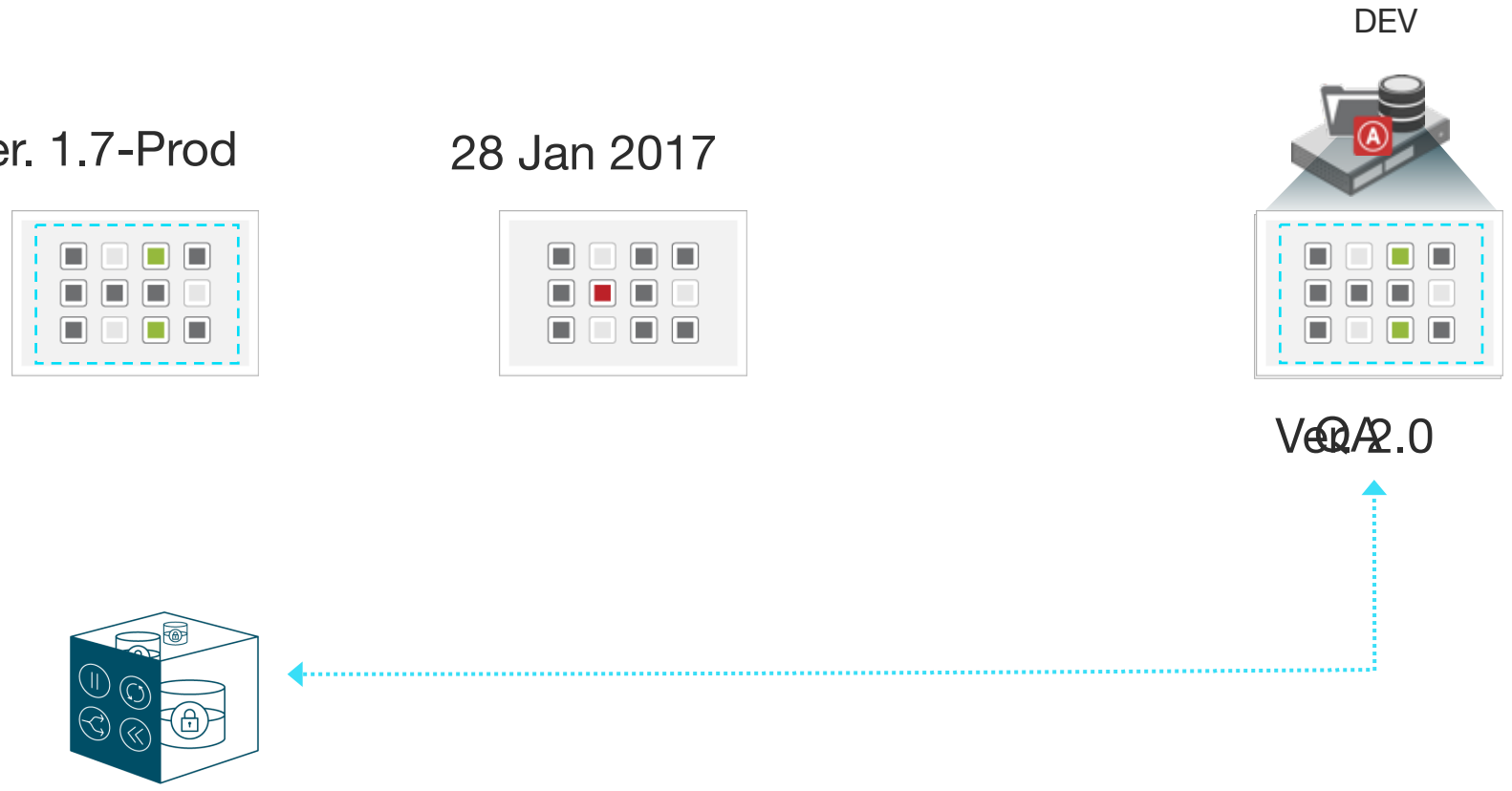
- Most of RDBMSs are not good in keeping code and data versioning
- Branching a database is a challenge
- Rolling back changes is a challenge

Code - Source Control

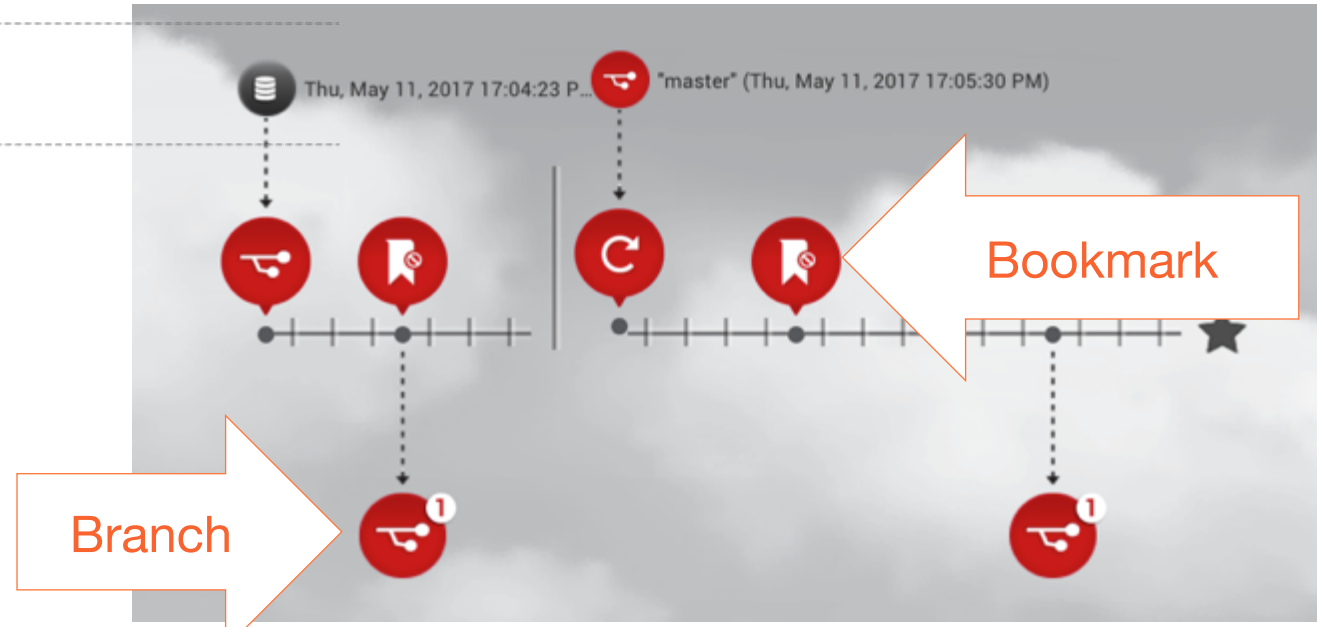
“A component of software configuration management, version control, also known as revision control or source control, is the management of changes to documents, computer programs, large web sites, and other collections of information.”



Data – “Source” Control



Data – “Source” Control



Virtualization – How

- Storage based solutions – buy or DIY
- File system based solutions – buy or DIY
- Appliance solutions – buy

Real life examples

Test Data Management - processes

Security

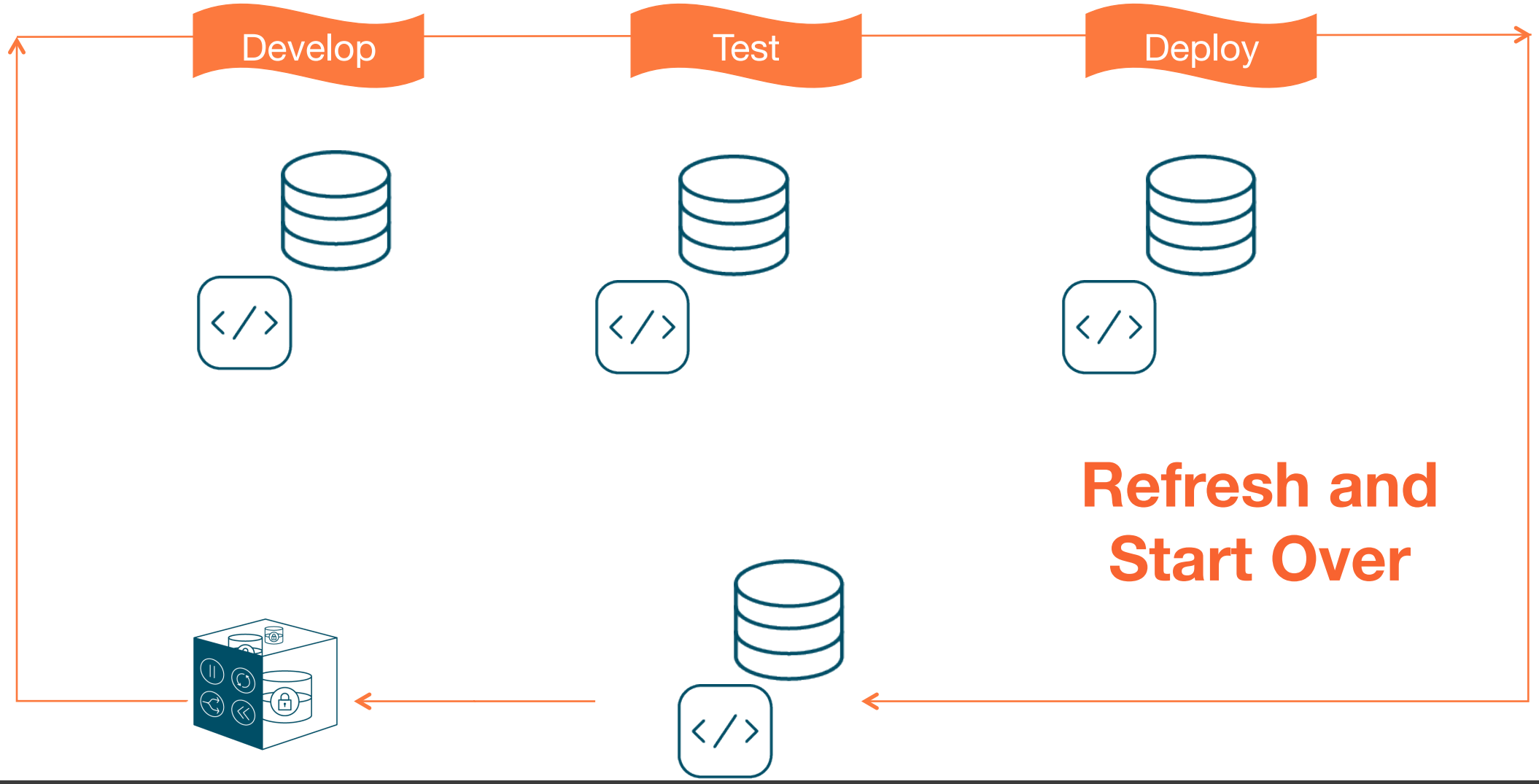
Continuous …

Data
Virtualization

puppet

TRICENTIS

Agile TDM



ORACLE

SSA_USER1

Database Cloud Self Service Portal

Create Database

Submit | Cancel

General

Service Template: Thin Copy of CRM Production

* Request Name: SSA_USER1 - Mon Sep 22 2014 16:14:52 EDT

* Zone: DbaaS_Zone

* Database SID: CRMTest

* Database Service Name: CRMTest_Srvc1

Schedule Request

If Start Date is set to "Immediately", the timezone "Eastern Daylight Time (GMT -4:00)" will be used for End Date.

Start: (•) Immediately ( ) Later (UTC-05:00) New York - Eastern Time (ET)

Duration: ( ) Indefinitely (•) Until 9/22/2014 8:17:43 PM

Deployment Input

* User Name: dbaas_crm

* User Password: ••••••••••

* Confirm User Password: ••••••••••

TIP The same password will be used for all schemas created as part of this request.

Service Instance Properties

View

Name	Value
▷ Optional	

Snapshots

Timeflow Slider

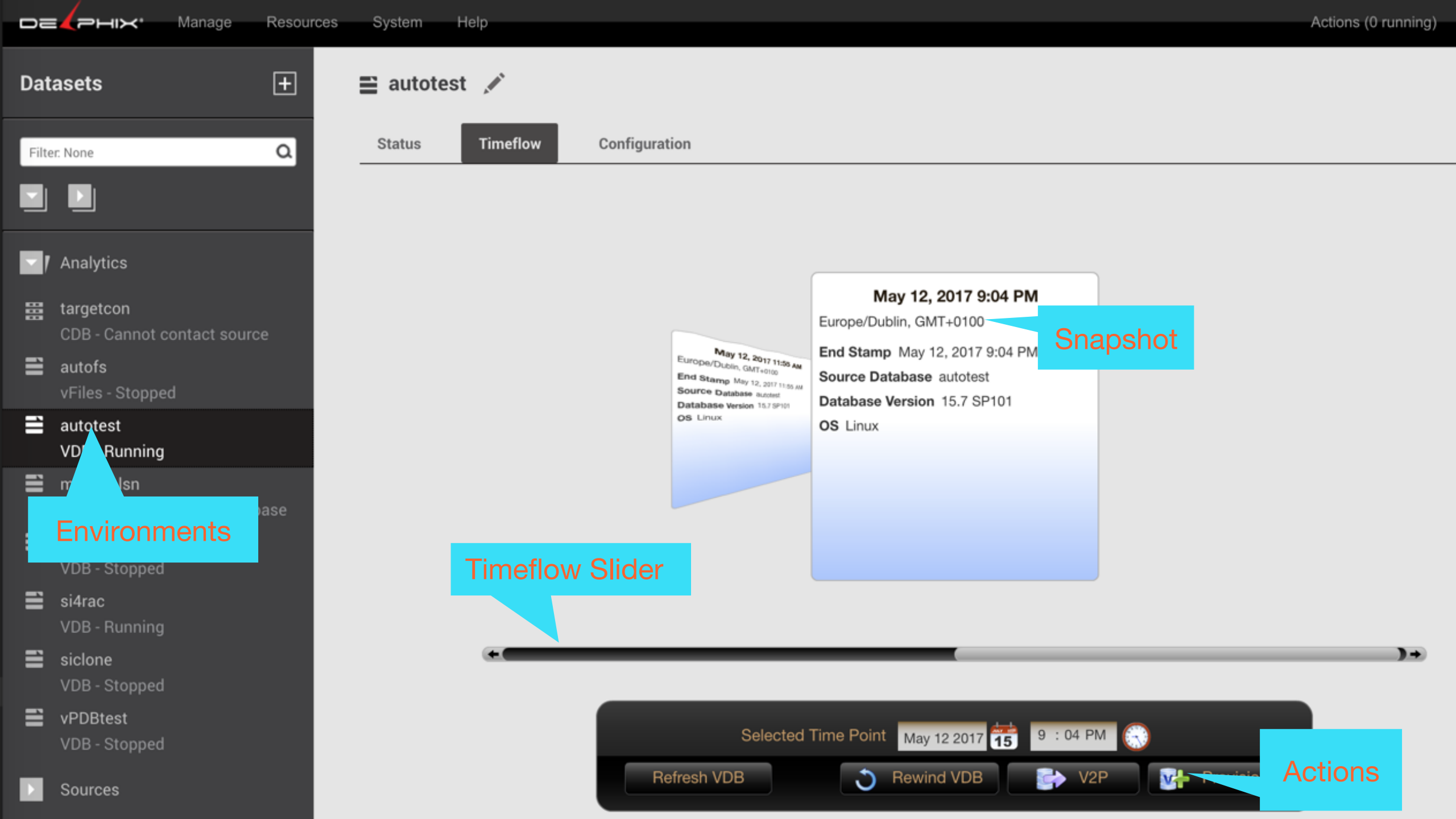
Snapshot taken on Sep 15, 2014 02:18 PM

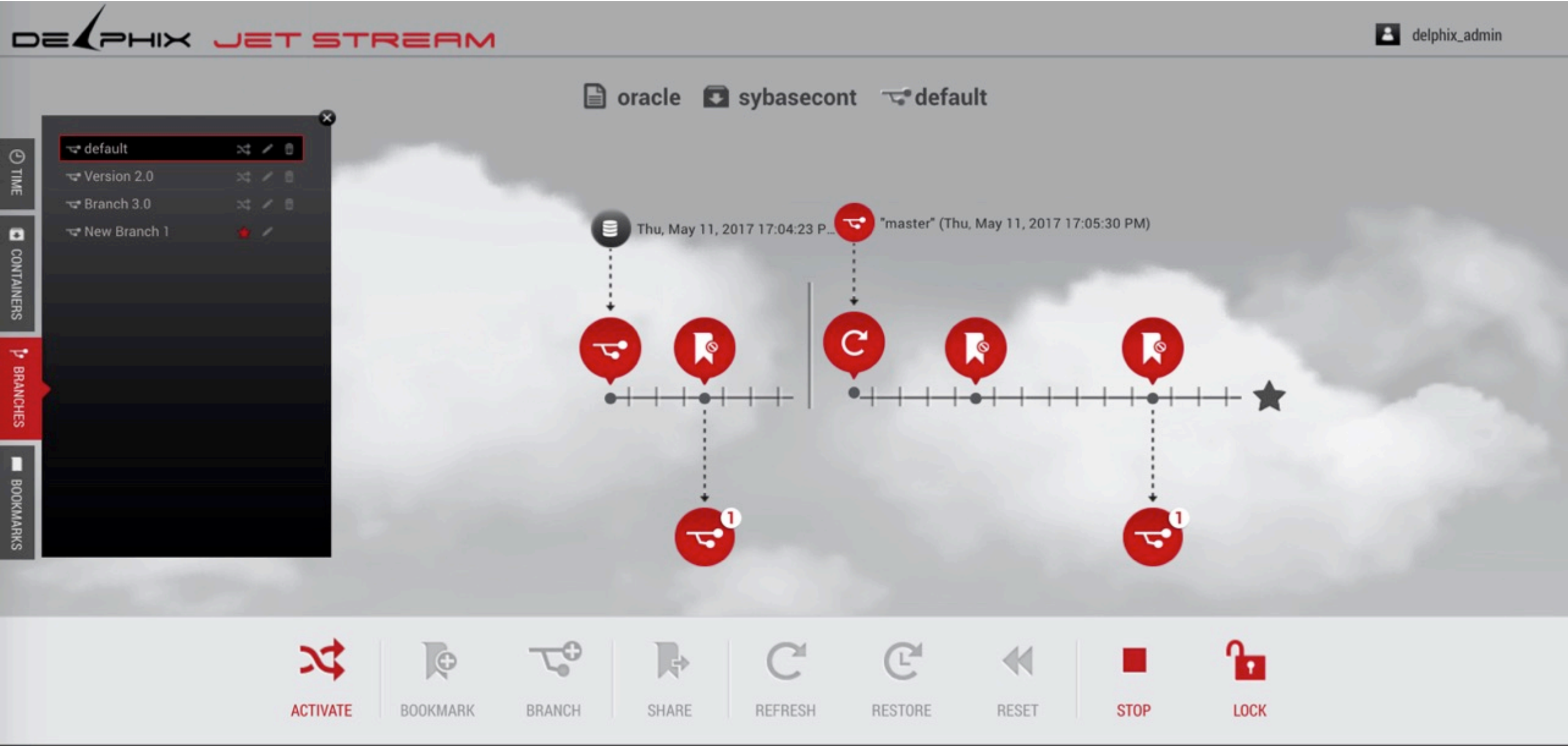
Snapshot taken on Sep 16, 2014 02:18 PM

Snapshot taken on Sep 17, 2014 02:18 PM

Snapshot

Hours Weeks


DELPHIX
Manage
Resources
System
Help
Actions (0 running)
Datasets
Filter: None
Analytics
targetcon
CDB - Cannot contact source
autofs
vFiles - Stopped
autotest
Running
Environments
VDB - Stopped
si4rac
VDB - Running
siclone
VDB - Stopped
vPDBtest
VDB - Stopped
Sources
autotest
Status
Timeflow
Configuration
May 12, 2017 9:04 PM
Europe/Dublin, GMT+0100
End Stamp May 12, 2017 9:04 PM
Source Database autotest
Database Version 15.7 SP101
OS Linux
Snapshot
Timeflow Slider
Selected Time Point
May 12 2017
9 : 04 PM
Refresh VDB
Rewind VDB
V2P
Actions



Real life examples

Test Data Management - processes

Security

Data Security

Totals for Category:	**Banking/Credit/Financial**	**# of Breaches: 36**	**# of Records:**	**26,262**
		% of Breaches: 4.3%	**%of Records:**	**0.1%**
Totals for Category:	**Business**	**# of Breaches: 375**	**# of Records:**	**2,548,225**
		% of Breaches: 44.4	**%of Records:**	**8.6%**
Totals for Category:	**Educational**	**# of Breaches: 74**	**# of Records:**	**489,376**
		% of Breaches: 8.8%	**%of Records:**	**1.6%**
Totals for Category:	**Government/Military**	**# of Breaches: 58**	**# of Records:**	**12,300,322**
		% of Breaches: 6.9%	**%of Records:**	**41.3%**
Totals for Category:	**Medical/Healthcare**	**# of Breaches: 302**	**# of Records:**	**14,400,946**
		% of Breaches: 35.7	**%of Records:**	**48.4%**
	Totals for All Categories:	**# of Breaches: 845**	**# of Records:**	**29,765,131**
		% of Breaches: 100.0	**%of Records:**	**100.0%**

http://www.idtheftcenter.org/images/breach/ITRCBreachReport_2016.pdf

Confidential data

Do I Have to Mask Data?

Type of Data	Year Passed	Ruling
Data Masking in the EU	2014	ARTICLE 29 DATA PROTECTION
GDPR	2016	Regulation (EU) 2016/679
HIPAA	1996	Health Insurance Portability and Accountability Act
PCI	2016, (Updated)	Payment Card Industry Standards
PII		Personably Identifiable Information
SOX	2002	Sarbanes-Oxley Act

Masking in the Picture

As 80% of data in a company are copies, then 80% of data won't be subject to security like a production environment. Securing this data is not just a priority, but in many cases, subject to legal ramifications, (i.e. PCI/PII)

Masking in the Security officer picture

Masking Requirements

- Masking shouldn’t be reversible
- Easy to audit
- Masking should be a simple, automated, repeatable process

Masking in the DBA picture

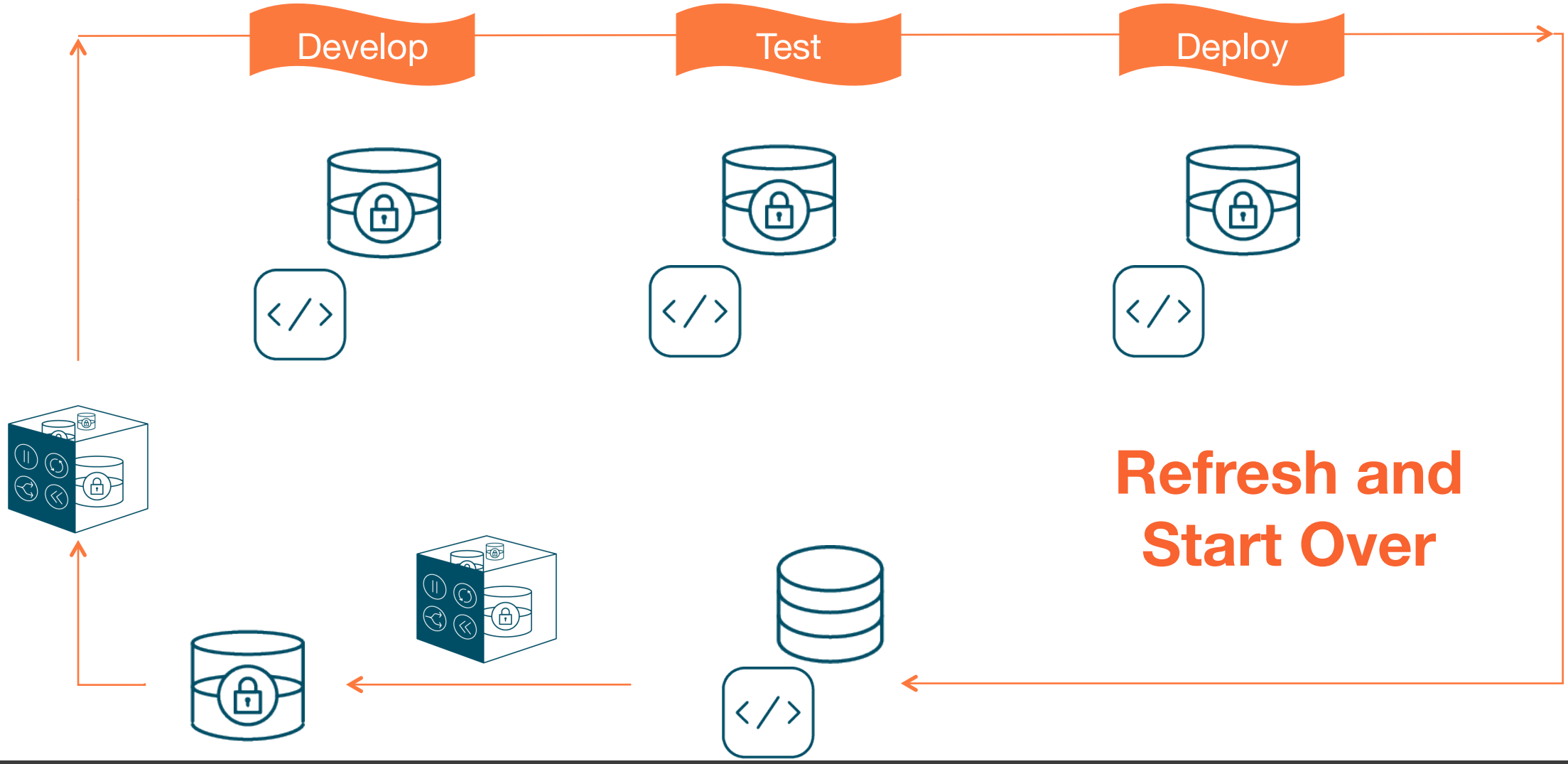
- Your data has to look same for optimizer after masking
  - Ex. age of customers
  - Ex. date of purchase
  - Ex. addresses
- Keep in mind correlation between data

- Keep in mind referential integrity

How This All Comes Together…

- Virtualization is the key to fast, efficient and FULL copies of production environments for agile and automated testing for agile shops.
- Data masking that can be done once, easily maintained with a repeatable process via a strong discovery and implementation as part of the virtualization process secures the 80% of data that is outside the control of production.
- Virtualized environments that are built with development and testing in an Agile or DevOps environments makes it simple to accomplish what may see impossible and do so at light speed.

Agile TDM



Marcin Przepiorowski
Senior Technical Principal
marcin@delphix.com
@pioro

