A Field Guide for Test Data Management

Kai Stroh, UBS Hainer GmbH

UBS HAINER

Typical scenarios

Common situation

- Often based on Unload/Load
- Separate tools required for DDL generation
- Hundreds of jobs
- Data is taken directly from production
- Requires temporary space (unless cross loader is used)

Usual Complaints

- Refreshes are done too infrequently, data becomes stale
- Takes days or even weeks for a refresh, DBAs need to monitor jobs, check many return codes
- When problems occur, DBAs need to check what parts are missing, rewrite and rerun jobs
- Long running jobs with high CPU load have negative impact on four-hour rolling average MSU value
- LOB and XML data difficult to copy

Unload/Load needs temporary space

Alternatives to Unload/Load

- DSN1COPY is available everywhere
- Offers no built-in automation
- When used, is often controlled via homegrown scripts
- Very error prone
  - Incorrect usage leads to errors in Db2
  - Newer Db2 features difficult to handle (PBG, XML, Rotation)
  - Space related problems often occur

BCV5
EFFICIENT Db2 DATA MIGRATIONS

Solution for full table copies: BCV5

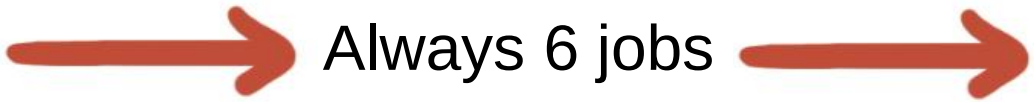
- Comprehensive: Handles structures, data, statistics, and more
- Fast: Works on VSAM level, automatic parallel processing and workload balancing, 10 times faster than Unload/Load
- Flexible: Rule-based selection and renaming process
- Robust: Many built-in plausibility checks

Easy to use, fully automated

Graphical interface

ISPF interface

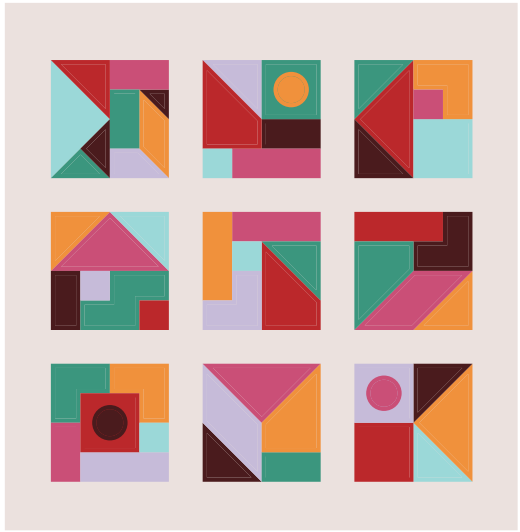
Batch interface



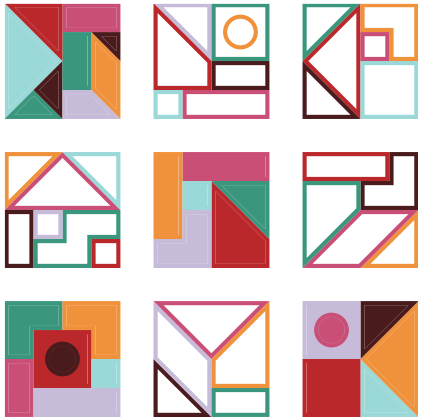
Always 6 jobs

Statistics
XML
Structures
TCP/IP copies
LOBs
Speed
Parallelism
Rebind
Space allocation
Renaming
Identity columns
Data
Image copies
Sequences

Production Db2



Relevant tables



Additional benefits of BCV5

- Can synchronize structures or work with existing target structures, even if they are different
- Automatic detection of restricted states, required rebuilds
- Automatic allocation of space in target
- Copy to a remote LPAR directly via TCP/IP
- Read from real source VSAMs or from source image copies

What speed can you expect from BCV5?

- 1 – 1.5 TB per hour throughput for file system level copy
  - VSAM to VSAM
  - Image copy data set to VSAM
- Slowdown is possible when:
  - A target index needs to be rebuilt
  - A tablespace requires Unload/Load due to structural differences

Optional BCV5 feature: in-flight copies

- Can make consistent copies without stopping source (“in-flight”)
  - Provides common point-in-time copy for a set of objects
  - No uncommitted changes in target
  - Works for tablespaces, indexes, LOB, XML (rebuild index not required)
- No impact on the availability or performance of the source
  - Direct VSAM access: does not affect Db2 buffer pools
  - Stop not required, quiesce not required
  - Can process changed pages only for very fast log apply phase

Example: BCV5

- Large automobile company from Michigan
- Environment: 23,325 page sets (VSAM clusters) to copy
  - 11,336 tablespace partitions
  - 11,969 index partitions
  - 20 LOBs
- Previous situation:
  - Hundreds of jobs
  - Required time for a refresh: ~2 weeks

Example: BCV5

- BCV5 job chain timings:
  - 91 minutes elapsed time
  - 14 minutes CPU time

Job	Elapsed Time	CPU Time
Stage 1	3m 06s	1m 03s
Stage 2	32m 56s	1m 52s
Stage 3	24m 10s	2m 08s
Stage 4	23m 15s	6m 48s
Stage 5	2m 53s	59s
Stage 6	5m 02s	1m 14s
TOTAL	**1h 31m 22s**	**14m 04s**

BCV4
FAST SUBSYSTEM CLONING

UBS HAINER

BCV4: For when size matters

- BCV4 clones full Db2 subsystems extremely fast
- Works on volume level, can exploit FlashCopy V1
- Consistent clones are possible without impacting the availability or performance of the source
- Very fast renaming of target FLQ for 10,000s of data sets
- Can also clone IMS databases

Characteristics of full clones

- Cloning process includes Db2 catalog, directory, BSDS, active logs, and certain libraries
- Overwrites everything in the target Db2
- Some renaming possible:
  - SSID and VCAT prefixes can be changed
  - Object names inside Db2 are not changed

Advantages of full clones

- The data in the cloned system is identical to the production
- Provides decoupling of production and other environments
  - Improves security, integrity of production
  - Only one well defined set of jobs take data from production
  - All other test / QA environments are populated with data from the clone
  - Copying data from the clone to another environment always “resets” the data in that environment

Considerations for full clones

- The clone contains production data and should therefore be secured against unauthorized access
  - RACF security for page sets and logs
  - Db2 security for table access privileges
- Ideally, tests should not run directly in the cloned Db2 as this will change the data

Possible reservations against full clones

- Argument: Cloning increases maintenance requirements
  - The clone might not need maintenance at all
  - Data in the cloned system does not change, eliminating the need for regular image copies, runstats, and reorgs
  - Cloning process can always be repeated if clone is damaged

Possible reservations against full clones

- Argument: Cloning requires additional DASD space
  - True, but benefits usually outweigh the cost
  - Benefit: Decoupling from production – no random Unloads or SELECTs against production tables to get test data
  - Benefit: Clone is not changed can always be stopped – copying from the clone into other environments are consistent
  - Benefit: Provides ability for acceptance tests against real production data, tests of Db2 version upgrades

Possible reservations against full clones

- Argument: PIT volume copy not possible between control units
  - True, but technologies such as space efficient FlashCopy exist
  - Create a space efficient FlashCopy first
  - Then create a logical volume copy to another control unit
  - Must be done before space on the repository volume is exhausted
  - Result: Point in time copy to a different control unit

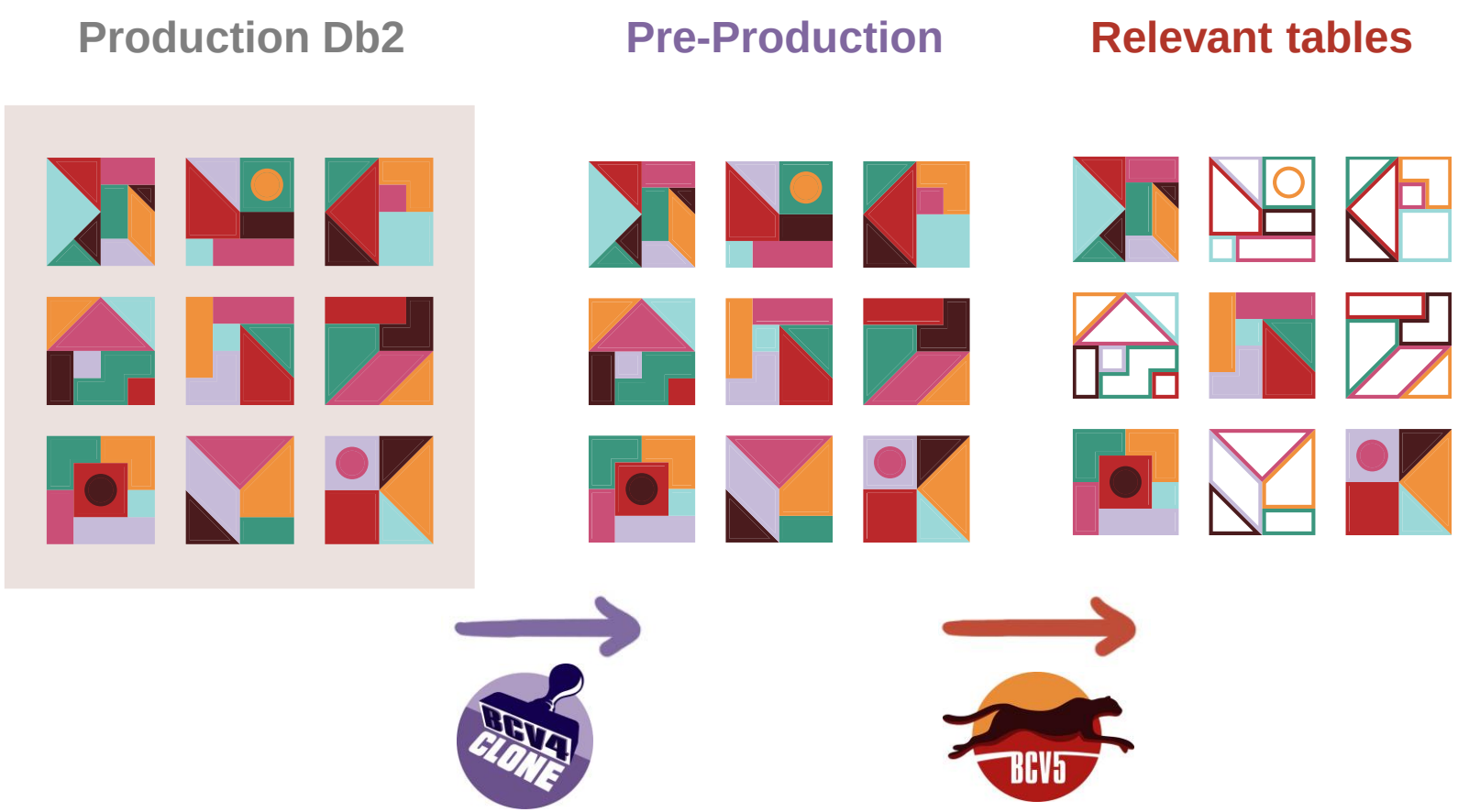
BCV5
EFFICIENT Db2 DATA MIGRATIONS

BCV4
FAST SUBSYSTEM CLONING

UBS HAINER

Combine BCV4 and BCV5

- Use BCV4 to create a pre-production environment (master copy, golden copy)
- Typical refresh frequency: every 2 – 3 months
- Use BCV5 to refresh smaller downstream environments
- Typical refresh frequency: daily, weekly

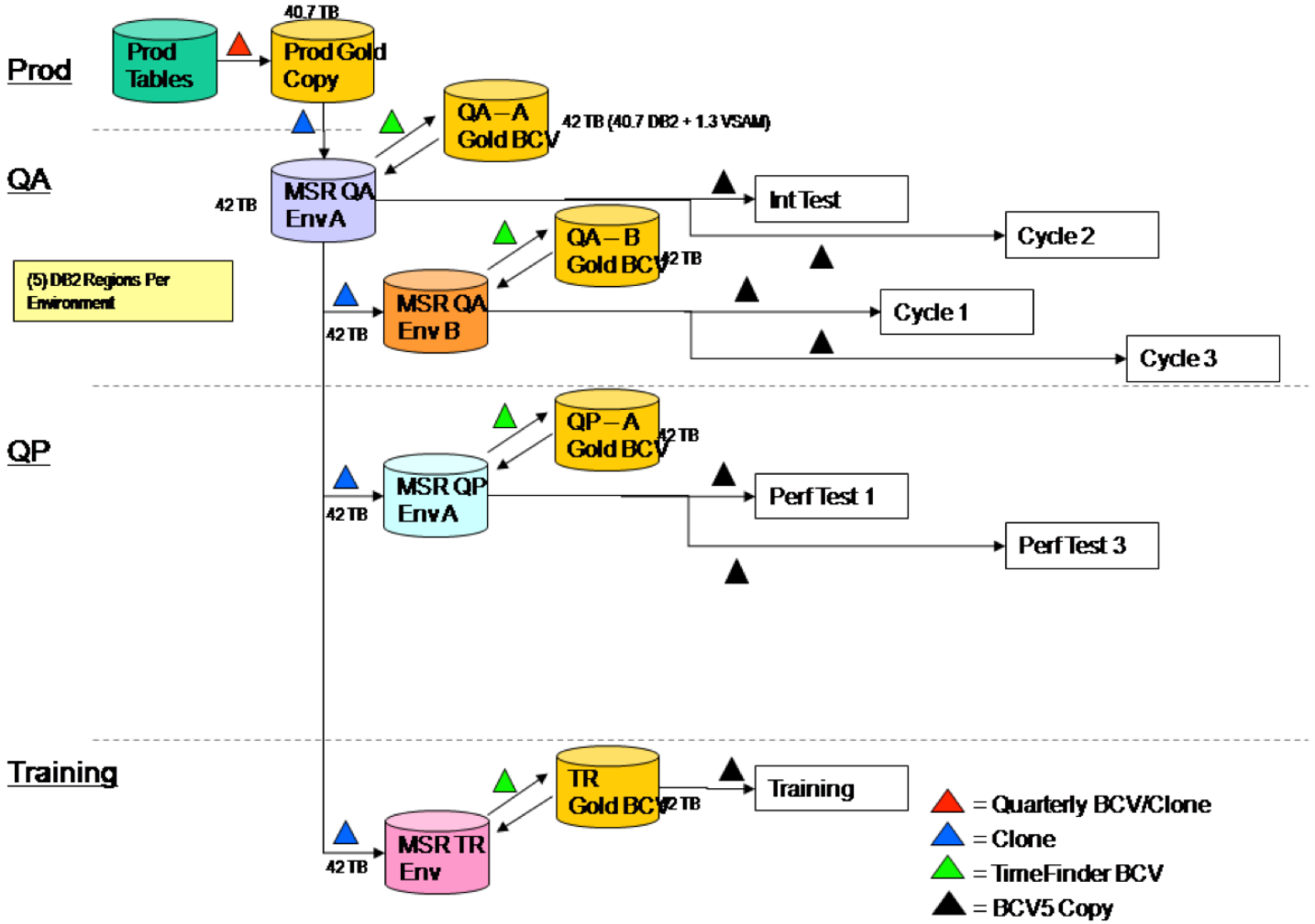


Example: BCV4

- Largest current BCV4 customer
  - 22 way data sharing group, 120 TB
  - Renaming 60,000 VSAMs on 2,225 volumes: 48 minutes
  - Rename and restore ICF catalog: 1 minute per catalog
  - Catalog, Directory and BSDS changes: 7 minutes
  - Updating 1,364 active log data sets (2 GB each): 30 minutes
  - Consistent full clone is done in half a day

Example: BCV4

- Suitable for smaller Db2 shops as well!
  - Standalone Db2, 5 TB
  - 1610 Volumes
  - Consistent full clone done in about 1 hour



BCV5
EFFICIENT Db2 DATA MIGRATIONS

BCV4
FAST SUBSYSTEM CLONING

XDM RLP
SET UP TEST CASES EASILY

UBS HAINER

Add even more flexibility

- Developers, QA are interested in subsets of rows
- Consistent subsets with respect to RI constraints to create test cases that “make sense”
- Merge new data into existing target tables that already have rows

XDM for all your row level needs

- Maximum reduction of manual work
- Scheduler based, automated process
- Platform independent
- Saves expert time
- Automatic restart of interrupted copy tasks

What makes row level copies difficult?

- RI constraints may or may not exist in Db2
- Constraints do not necessarily form a strict hierarchy
  - Self referencing tables
  - Cyclic references between two or more tables
  - Multiple different references between the same two tables
- It may not be possible to formally describe constraints in Db2
  - Constraints may use substrings of a value
  - Conditional constraints based on contents of other columns

Consequences of improper RI handling

- Data in child tables may or may not be desired
- Improper processing leads to inadequate / incomplete test data
  - Customers without invoices
  - Invoices without items
  - Items without suppliers
- Cyclic references: Must keep track which tables have been visited, by way of which other tables, and how often

Considerations for row level copies

- Works with regular SQL, not for mass data
- Copy time varies from seconds to hours
  - Complex tasks must run as a background process
  - Should not run on a user's workstation, but on a server
- Data transfer to different platforms may be required
  - FTP, FTPS, SFTP, network shares, other methods

Additional features of XDM

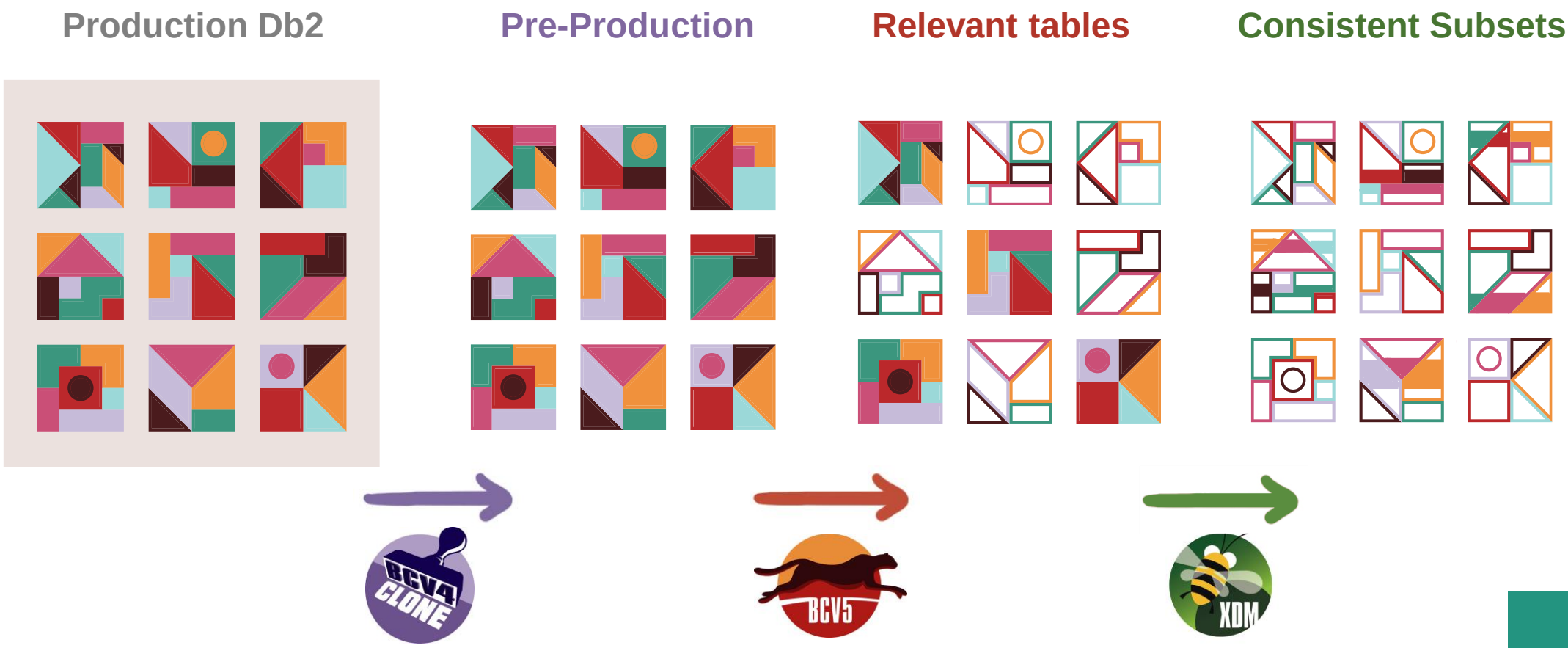
- Sophisticated data masking possibilities
  - Predefined methods
  - Extensible through custom scripts
- Graphical interface runs on Windows, Mac OS, Linux/Unix
- Web-based interface
- Compatible with many different platforms
  - Db2 for z/OS, Db2 for LUW, Oracle, MSSQL Server, IMS, VSAM

Example: XDM-RLP

- Large European insurance company (11 Million customers)
- Db2 for AIX with 2,500,000,000 rows in 40 tables, desired subset: 110,000,000 rows based on ~100 customer numbers
- Many RI constraints (no cycles), partially in the Db2 catalog
- Masking of PII, including key columns
- Proper konwledge of RI constraints is crucial
  - Without them, you will not get all the rows you expected
  - Masking of RI columns only works properly if XDM knows all RIs

Example: XDM-RLP

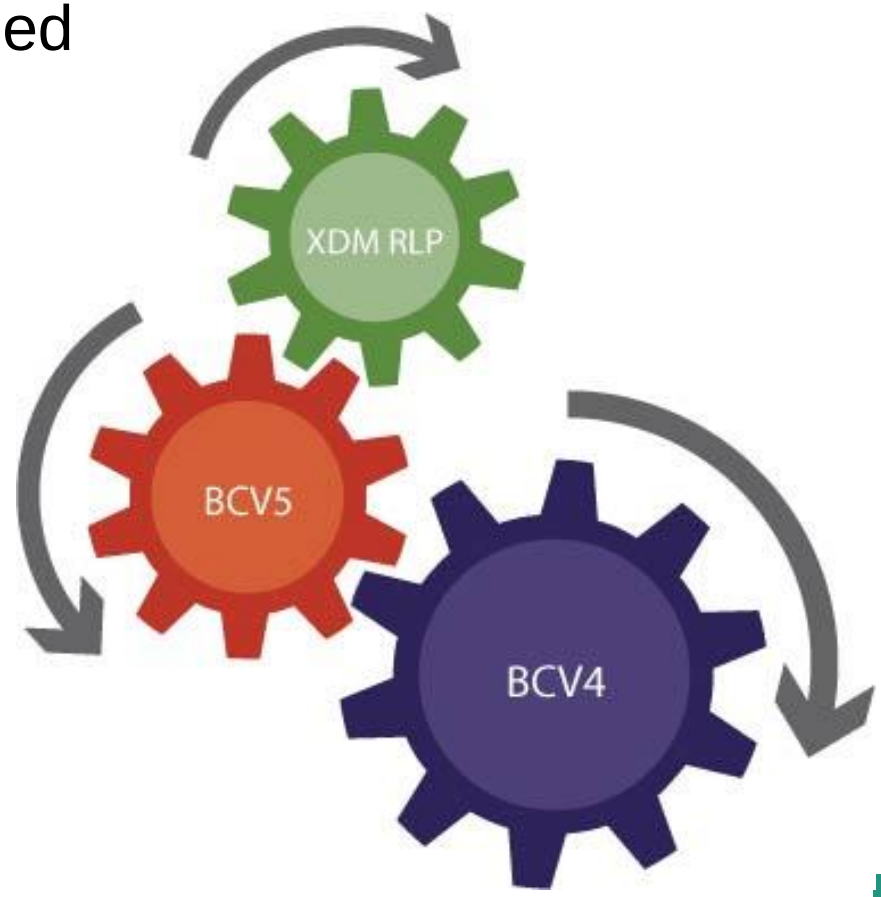
- Total processing time: 8 hours
- Had to create additional indexes on the source tables, otherwise processing time would have been 24 hours
  - Make sure to have indexes on the child columns of your foreign keys
  - XDM-RLP can print the SELECT statments that it executes so you can see the JOIN and WHERE conditions
- Proposed masking algorithm inadvertently caused unique key violations, had to make slight modifications

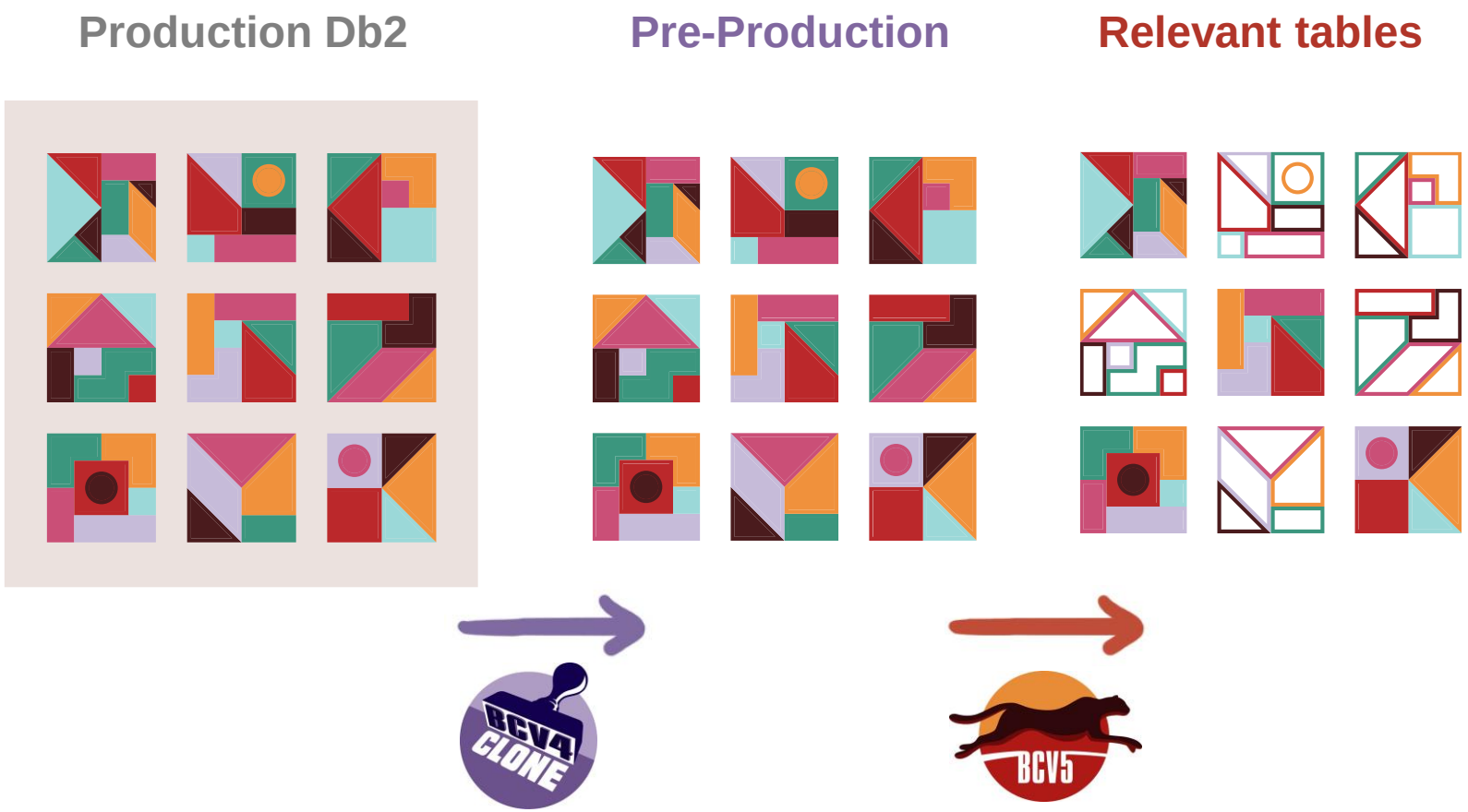


Create a refresh strategy

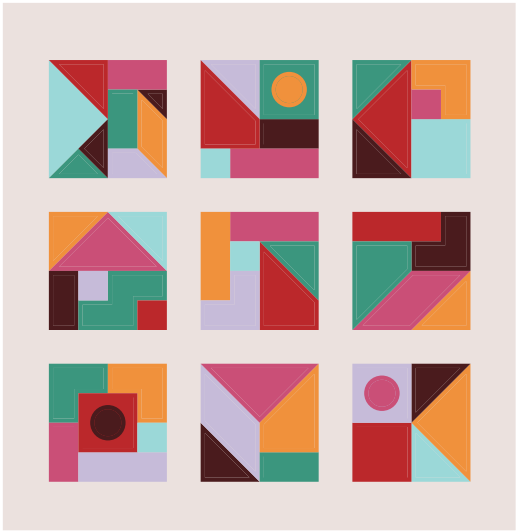
Different environments need to be refreshed at varying frequencies:

- Preproduction environments:
  - Every 2 – 3 months
- Integration environments:
  - Weekly – monthly
- Development environments:
  - As needed, multiple times per day

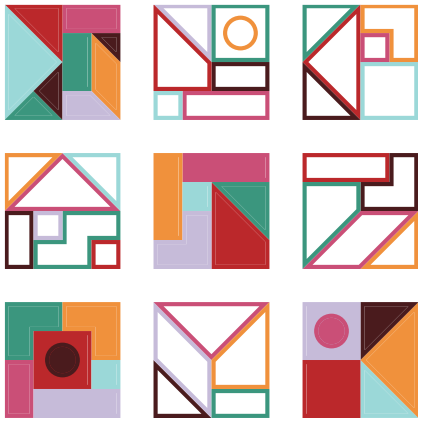


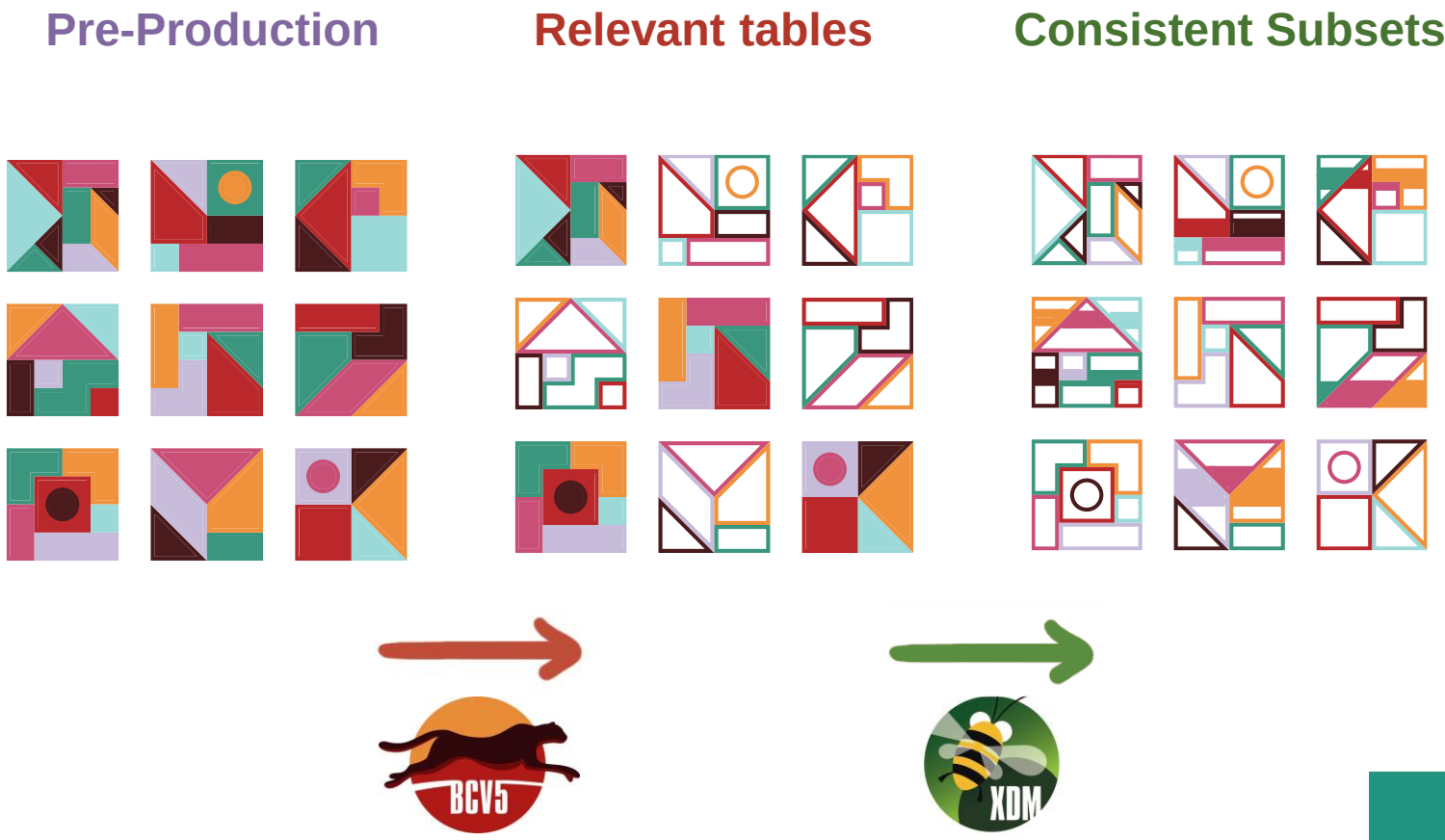


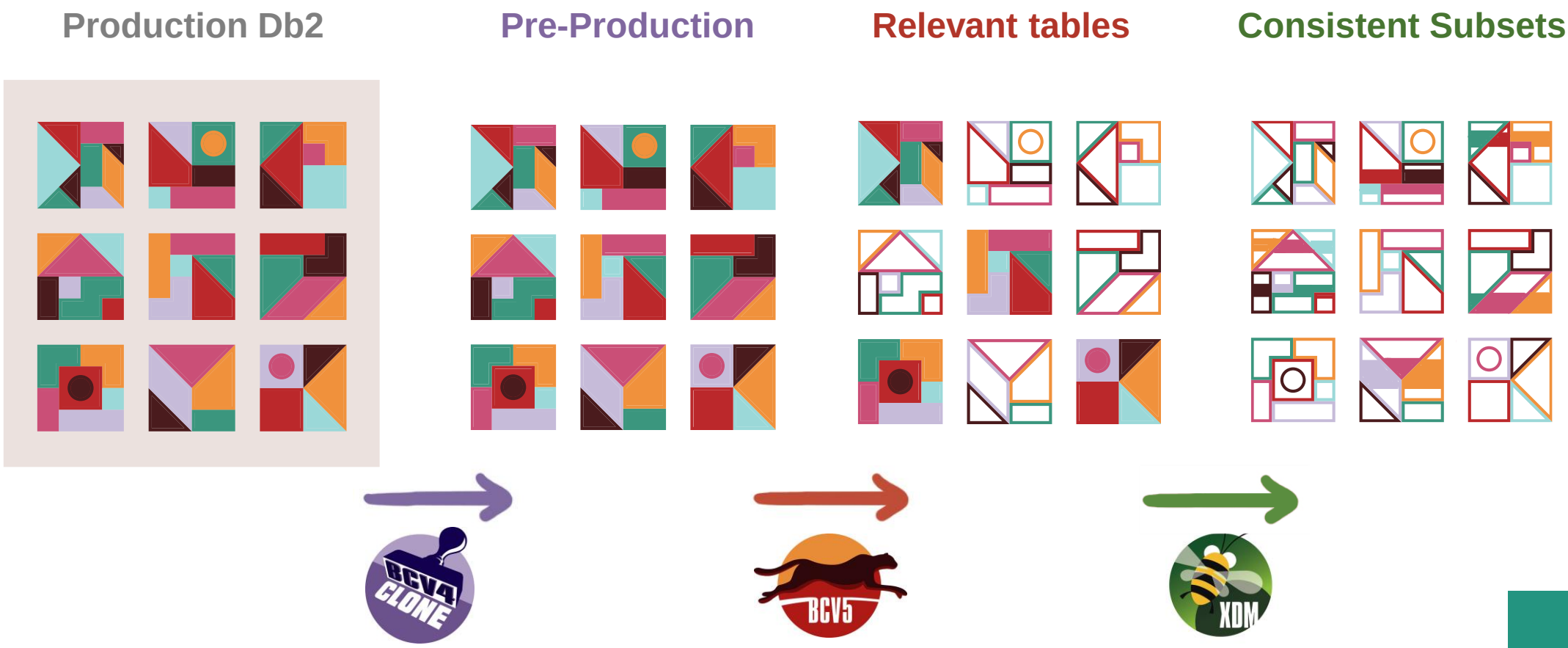
Production Db2



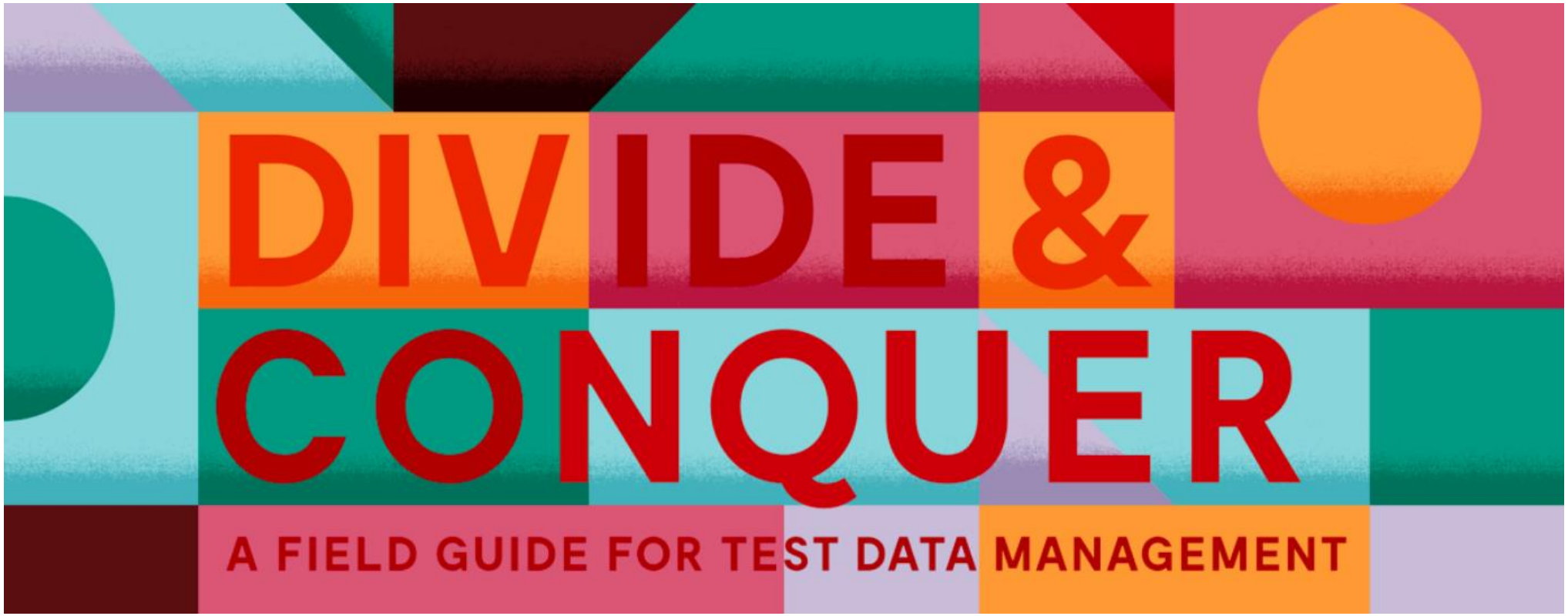
Relevant tables







Read the test data management article at:
http://tdm.ubs-hainer.com



Thank you for your attention!

For more information visit us at **ubs-hainer.com**

or send an email to **s.tursman@ubs-hainer.com**

UBS HAINER