IBM's Complete Test Data Management Family

IBM

Virtual Database Copies

IBM InfoSphere Virtual Dat
PIPELIN

IBM InfoSphere Optim
DATA PRIVACY

IBM InfoSphere Optim
DATA PRIVACY
for Unstructured Data

Near instant virtual database copies with central management and self-service (roll back) refresh

Database Subsetting & Data Masking

IBM InfoSphere Optim
TEST DATA MANAGEM

IBM InfoSphere Optim
DATA PRIVACY

IBM InfoSphere Optim
DATA PRIVACY
for Unstructured Data

Database subsetting of "complete business object" for leaner test data environment

Synthetic Data Fabrication

IBM InfoSphere Optim
TEST DATA FABRICATION

Production-like, synthetic test data when production data isn't available or is too sensitive

Tester Portal for Data Coverage & Assignment

IBM InfoSphere Optim
TEST DATA ORCHESTRATOR

Complete test data coverage, automated data assembly, and DevOps integration

Dev Ops

IBM’s Complete Test Data Management Family

Virtual Database Copies

IBM InfoSphere Virtual Data PIPELINE

Near instant virtual database copies with central management and self-service (roll back) refresh

Database and Unstructured data Masking

IBM InfoSphere

Optim

DATA PRIVACY for Unstructured Data

Synthetic Data Fabrication

IBM InfoSphere Optim Test Data Fabrication

Production-like, synthetic test data when production data isn’t available or is too sensitive

Tester Portal for Data Coverage & Assignment

IBM InfoSphere Optim Test Data Orchestrator

Complete test data coverage, automated data assembly, and DevOps integration

What is Unstructured Data?

- **Structured** - Information with a high degree of organization, such that inclusion in a relational database is seamless and readily searchable by simple search operations.

DB2 . Oracle . SQL Server . Informix . Sybase . Teradata …

- **Unstructured** - information that either does not have a pre-defined data model or is not organized in a pre-defined manner.

Scanned Images, PDFs, Web logs, JSON, XML, Office docs …

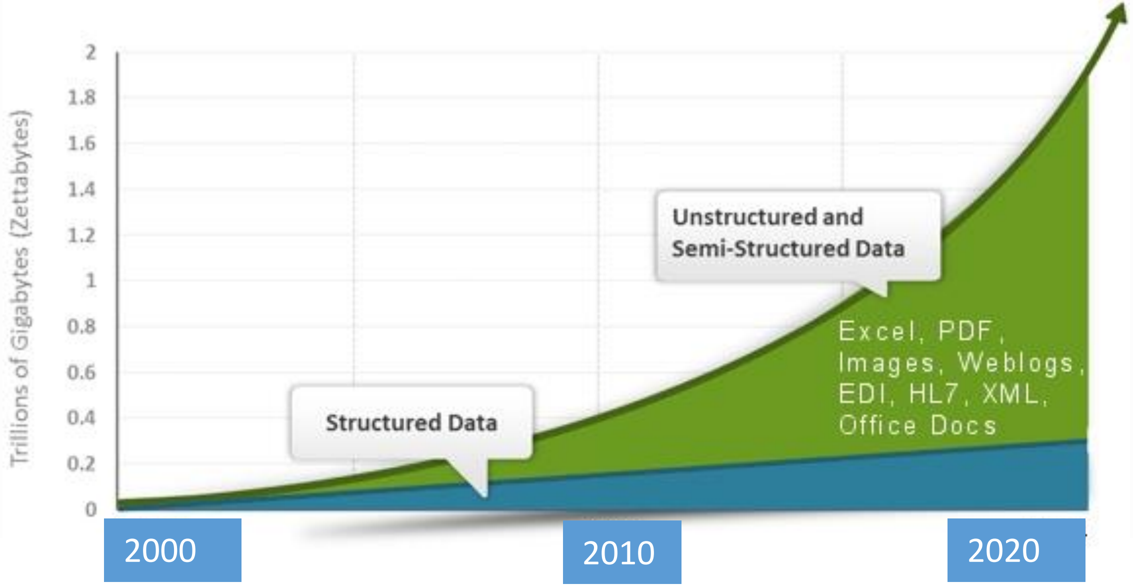
Unstructured Data Growth Challenges

Growing and Maintaining customer base is key – leveraging new services and offerings to elite patrons
Regulatory agencies are requiring more cooperation between consumers, businesses, law enforcement agencies, and legislators.

At the same time, more data is being kept exponentially each year to capture customer habits, historical and transactional data, and trending analysis.

- Marketing
- New Service offerings

❑ Over **80%** of data currently generated is unstructured.



Unstructured Data Management

The **worldwide shift from structured to unstructured data is well known**. It isn't that structured data is going away. In fact, structured data continues to grow. It's just that **growth of unstructured datasets is faster – much faster.**

When you consider the analyst predictions that **80 percent of new data will be unstructured**, it becomes clear that we need to understand and prepare for how this affects us.

Unstructured datasets are growing quickly. The **typical growing 23% annually**, which means it will **double every 40 months**. Roughly **one fourth cite growth rates in excess of 40%**, **where total unstructured data doubles every 24 months**.

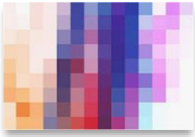
82% respondents manage **1 billion+ files and objects** and **52%** respondents manage more than **10 billion files**.

40 percent of their organization's value comes directly from their data. ***"Our data is much more valuable than we've ever thought…"That's why we save so much of it!"***

IGNEOUS - 2018 STATE OF UNSTRUCTURED DATA MANAGEMENT REPORT

Complexities of masking unstructured data

IBM



Images stored as pixels – JPEG, TIFF, PNG, GIF.
Insurance claims, scanned images, bank checks, patient records…

Adobe

PDF's containing freeform text, Adobe proprietary compressed streams, embedded images…

MS Office Proprietary formats , XLSX, XLS, XLSB, XLSM Excel 2007, 2013…

'Flat files'

CSV, fixed width, freeform log files, XML, HL7, EDI, NACHA, JSON…

010010101
100100100
…

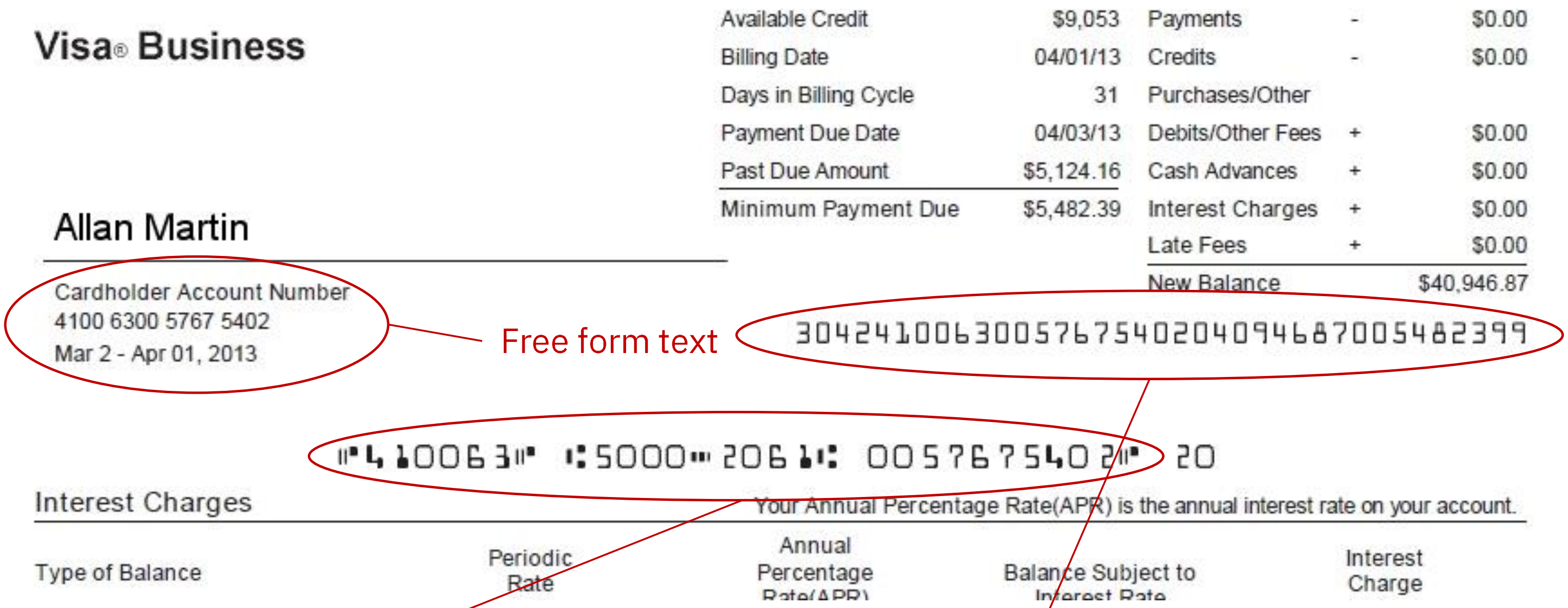
X9 Image Cash Letters, X9.37 and X9.100-187 MS Exchange, Raster file formats…

IBM InfoSphere

Optim

DATA PRIVACY for
UNSTRUCTURED DATA

Challenges of PDF's



Visa® Business

Available Credit	$9,053
Billing Date	04/01/13
Days in Billing Cycle	31
Payment Due Date	04/03/13
Past Due Amount	$5,124.16
Minimum Payment Due	$5,482.39

Payments	-	$0.00
Credits	-	$0.00
Purchases/Other		
Debits/Other Fees	+	$0.00
Cash Advances	+	$0.00
Interest Charges	+	$0.00
Late Fees	+	$0.00
New Balance		$40,946.87

Allan Martin

Cardholder Account Number
4100 6300 5767 5402
Mar 2 - Apr 01, 2013

30424100630057675402040946870054823 99

⑈410063⑈ ⑆5000⑉2061⑆ 00576754022⑈ 20

Interest Charges

Your Annual Percentage Rate(APR) is the annual interest rate on your account.

Type of Balance	Periodic Rate	Annual Percentage Rate(APR)	Balance Subject to Interest Rate	Interest Charge

Free form text

bitmap images/hex

stored as Adobe compressed streams

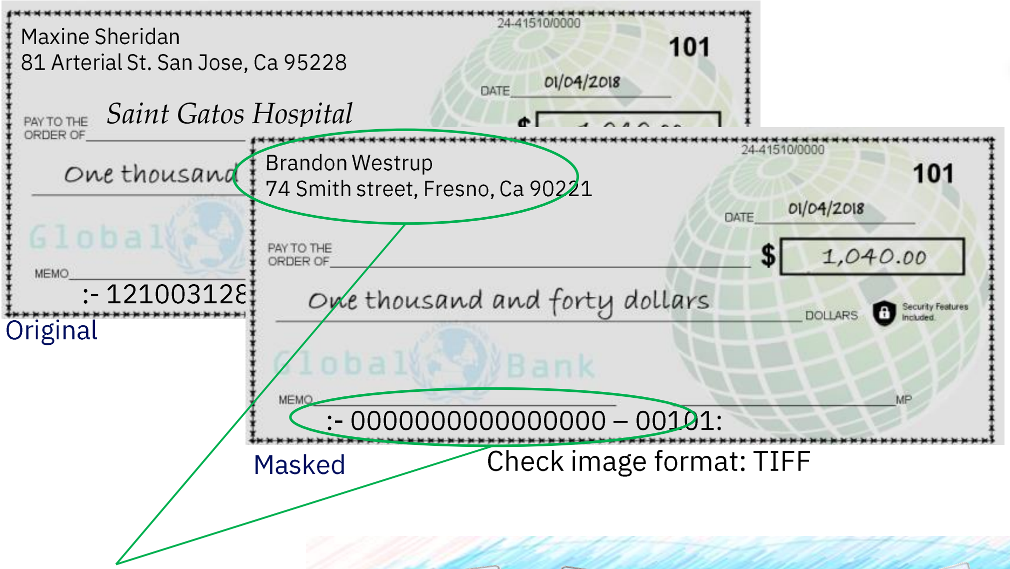
Introducing: InfoSphere Optim Data Privacy for Unstructured Data (DPU)

New IBM InfoSphere

Optim

DATA PRIVACY for UNSTRUCTURED DATA

Check Images, Insurance Claims, Statements, Web Logs, Documents…



24-41510/0000

101

Maxine Sheridan
81 Arterial St. San Jose, Ca 95228

DATE 01/04/2018

PAY TO THE ORDER OF Saint Gatos Hospital

One thousand

Global

MEMO

:- 121003128

Original

24-41510/0000

101

Brandon Westrup
74 Smith street, Fresno, Ca 90221

DATE 01/04/2018

PAY TO THE ORDER OF

$ 1,040.00

One thousand and forty dollars

DOLLARS

Security Features Included

Global Bank

MEMO

MP

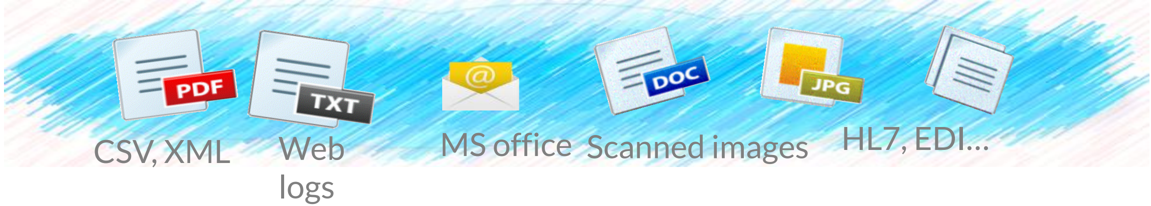
:- 0000000000000000 – 00101:

Masked

Check image format: TIFF

Masked

- ✓ Mask unstructured data across the enterprise
- ✓ Replace like-for-like, maintaining consistency
- ✓ Mask over 65 different file types
- ✓ Stand alone or with Optim, StoredIQ, and other governance tools



Maintain referential integrity

Structured

ACCTNO	FIRST	LAST	SSN
10010	Sally	Jones	619-34-8499
10020	Tony	Calvert	654-39-0800
10030	Bruce	Mercante	342-25-9246

ACCTNO	FIRST	LAST	SSN
10010	Angie	Smith	812-54-8099
10020	Tony	Martin	350-03-5618
10030	Jack	Wayne	100-03-3578

Optim™

& other IBM governance tools …

Protected!

protected?

Unstructured

CSV, XML

MS office

Scanned images

web logs

HL7,
EDI…

Completing the Compliance circle

Masking process: Rule Types

□ Direct: Find and Replace

Bruce Mercante | Wayne Andrews

*To: **Lt. Bruce Mercante** from the U.S...*

*To: **Lt. Wayne Andrews** from the U.S...*

□ Regular Expressions

\d{3}-\d{2}-\d{4}$ | 000-00-0000

*My Social Security number is: **619-35-9800***

*My Social Security number is: **000-00-0000***

□ Specific Region or Area

www. | .com | websiteredacted

<site>WWW.NHISSEC.COM</site>

<site>WWW.websiteredacted.COM</site>

□ Optim Consistent

^\d{3}-\d{2}-\d{4}$ | OPTIM..NID..

*My Social Security number is: **619-35-9800***

*My Social Security number is: **428-34-2391***

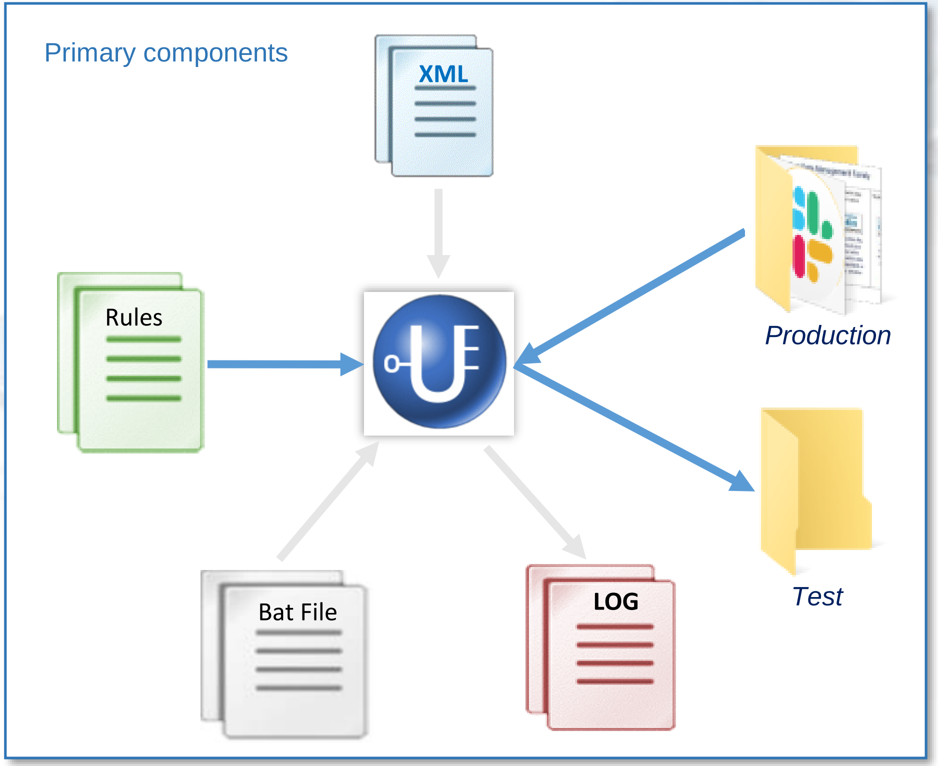
Masking Steps: Basic Process Flow

1. Create or select rules (what to find/how to replace)

2. Select source and target files/folders

3. Execute either through GUI, Command Line, or Batch

4. Validate using log process report



Masking Steps: Build/Select Rules File

Rules

- Text based files containing 1 or more RegEx rules
- Rules are formatted as Find, Delimiter, Replace
- Use standard text editors e.g. Notepad.exe
- Can contain Optim ODPP masking function calls

C:\IBM\DPU\REGRULES\SAMPLE.TXT

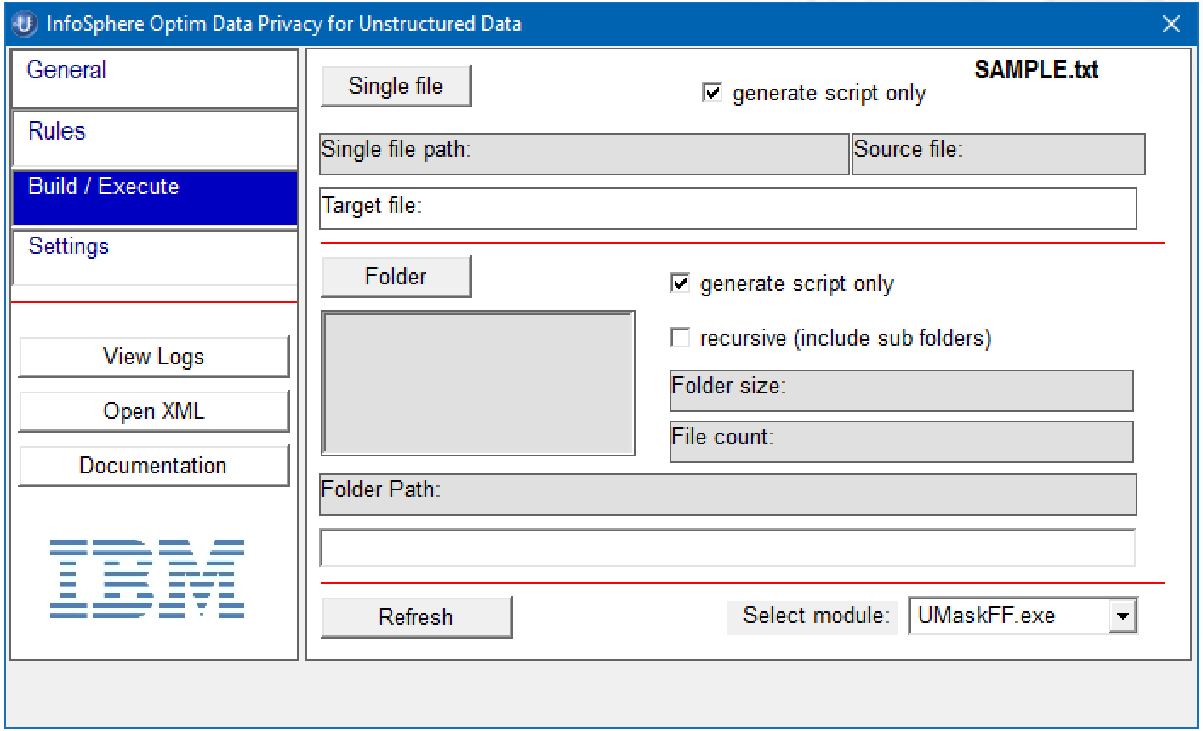
```
# Basic Email mask.
[^@]+@[^\.]+\..+ | Redacted@anywhere.com

# replace SSN with Optim function maintaining RI
\d{3}-\d{2}-\d{4}$ | OPTIM..=NID..

# replace values between 2 places
<from> | </from> | Name Redacted
```

This example is using the pipe symbol as the delimiter between find and replace

Lines with multiple delimiters indicate replace values between 2 areas

Masking Steps: Select source File/Folder

- Use GUI to select individual files or folders containing files
- Select Option to recurse sub folders if required
- Execute Immediately or generate a batch file only
- Generating batch file will also generate an associated XML containing selected options



SAMPLE.txt

☑ generate script only

Single file path:

Source file:

Target file:

☑ generate script only

☐ recursive (include sub folders)

Folder size:

File count:

Folder Path:

Select module: UMaskFF.exe

Masking Steps: Execution

- Direct CLI call or through batch script
- Batch script can point to XML file containing processing instructions
- Batch script can contain parameter overrides
- Can perform hybrid parameters and XML execution, parameters in batch file supersedes XML file contents

XML

C:\IBM\DPU\Modules\UMFF.XML

```
<S N="ProcessType">FOLDER</S>
<S N="RuleFile">C:\IBM\DPU\REGRULES\SAMPLE.txt</S>
<S N="Recurse">True</S>
<S N="SourceFolder">\\PRODP039\PROD\</S>
<S N="TargetFolder">\\QATESTLAPP03\TEST\</S>
<S N="OPTIMEnable">True</S>
```

Bat File

C:\IBM\DPU\Modules\UMFF.BAT

```
Rem this script will retrieve all paramters from a specified XML file
Rem and will also use paramters overrides superseding some paramters
Rem located in the XML file
REM ---------------------------------------------------------------------------------

C:\IBM\DPU\modules\UMFF.exe –xmlfile C:\IBM\DPU\Modules\UMFF.XML
-Targetfolder \\DEVTESTLAPP99\DEVTEST -Recurse True
```

This example reads an xml file to obtain the parameters. It will also override the Targetfolder and Recurse parameter

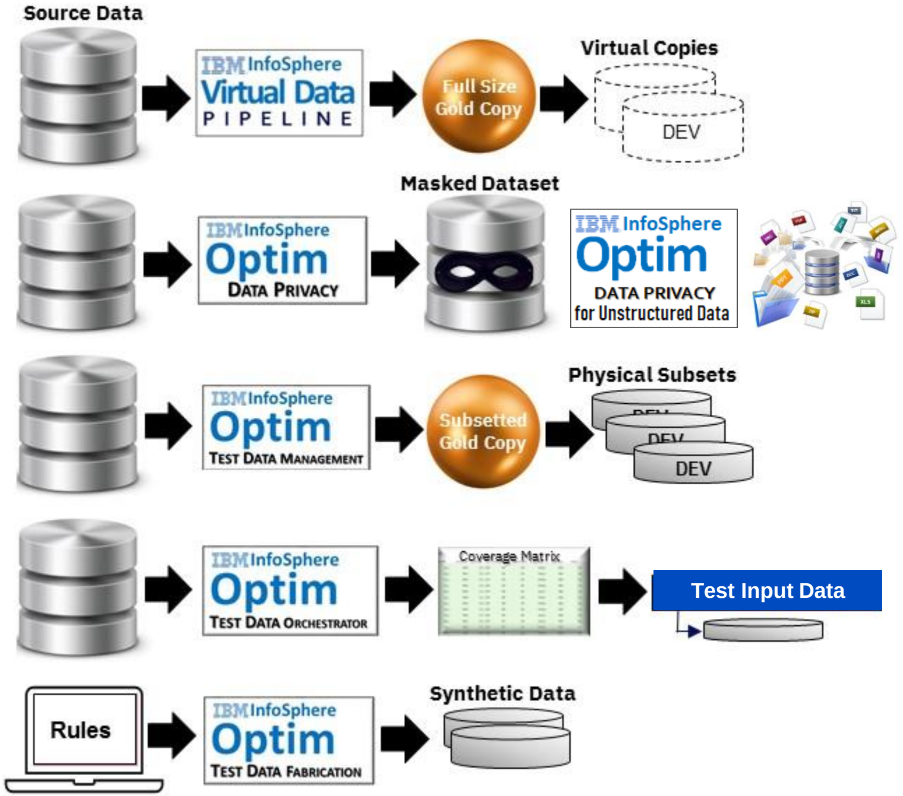
Post processing: Log files

- Logs generated at completion of processing
- DPU provides tools to summarize processing times and bytes processed for multiple log files

C:\IBM\DPU\LOGS\UMLOG_02082020102325.LOG

```
2020/01/03 09:20:11 Running on Server: "Blade“
2020/01/03 09:20:11 Username: "Allan Martin“
2020/01/03 09:20:11 ----------------------------------------
2020/01/03 09:20:11 Rules file used: "C:\IBM\\REGRULES\\Rule_ACORD_COI.txt“
2020/01/03 09:20:11 Delimiter: "|“
2020/01/03 09:20:11 Process Type: "FOLDER“
2020/01/03 09:20:12 Files Processed: 112
2020/01/03 09:20:12 Total Bytes processed: 18045328
2020/01/03 09:20:12 Elapsed time: 561.6781ms
2020/01/03 09:20:12 ----------------------------------------

ERRORS: none
```

IBM’s Complete Test Data Management Family



Source Data

IBM InfoSphere Virtual Data PIPELINE

Full Size Gold Copy

Virtual Copies

DEV

Masked Dataset

IBM InfoSphere Optim Data Privacy

IBM InfoSphere Optim DATA PRIVACY for Unstructured Data

IBM InfoSphere Optim Test Data Management

Subsetted Gold Copy

Physical Subsets

DEV

DEV

DEV

IBM InfoSphere Optim Test Data Orchestrator

Coverage Matrix

Test Input Data

Rules

IBM InfoSphere Optim Test Data Fabrication

Synthetic Data