

Why testers love production data & what to do about it.

Gary Hallam | November 2025



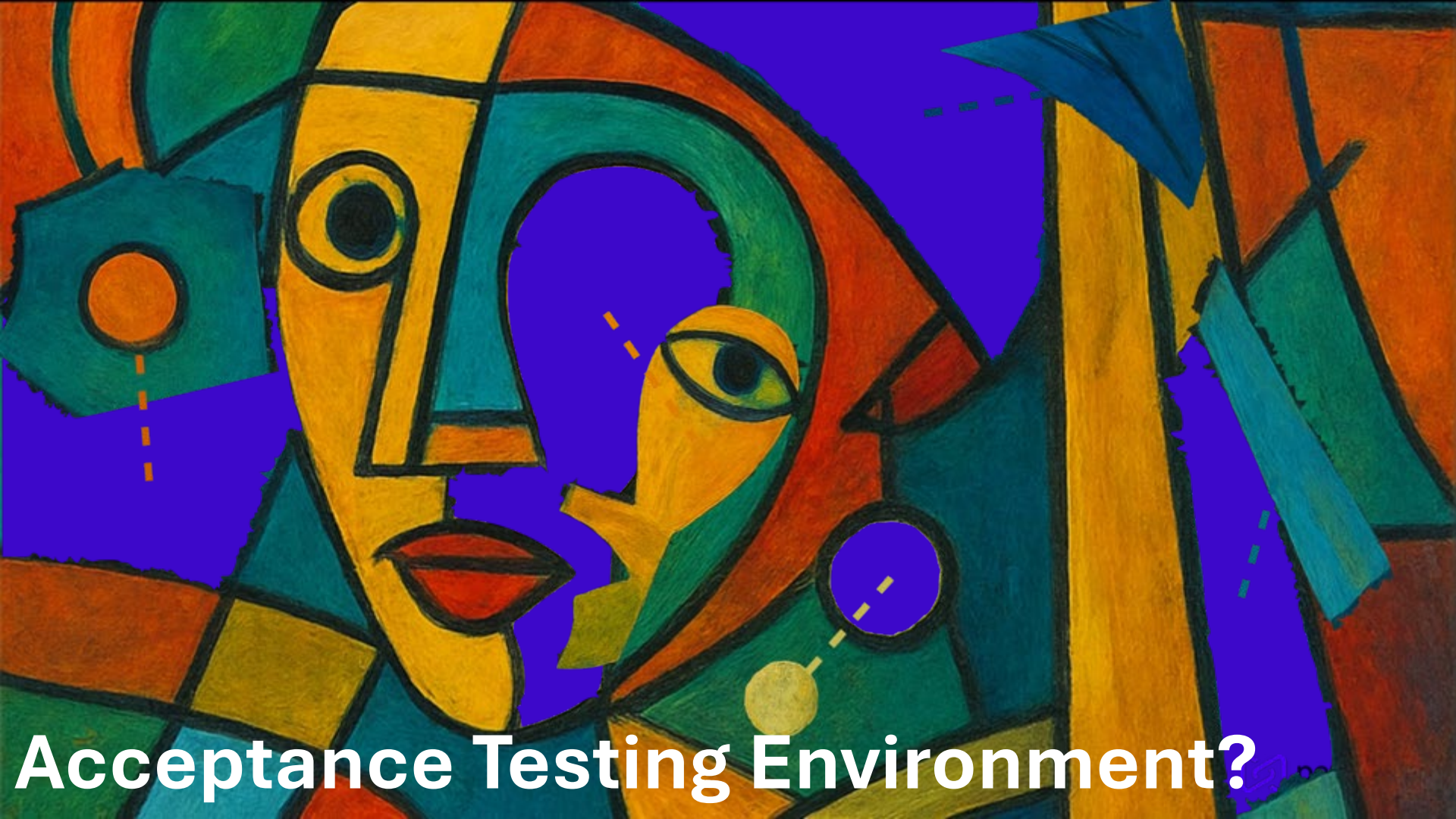
Production Environment



Pre-Production Environment?



System Testing Environment?



Acceptance Testing Environment?



Development Environment!!



Production

(Johannes Vermeer)



Testing

(Wassily Kandinski style - ChatGPT)



Development

(Pablo Picasso)



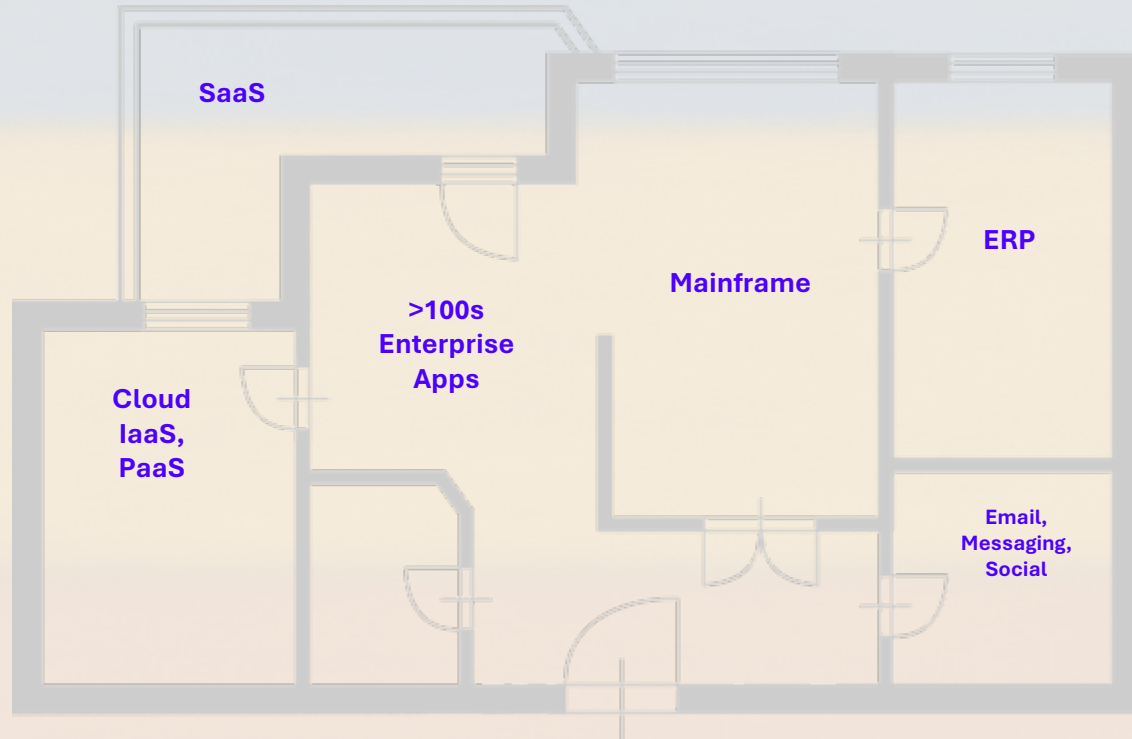
Masked

(Mark Rothko style – Copilot)

Production Data

- Rich, realistic and representative values
- Contains nuance and edge cases to ensure full testing
- It contains all the flaws associated with multi-generational architectures developed over time
- In the AI Language Model world, it is also the information source that we need to train on and augment with.
- It's hard to synthesise
- Full volume and the full complexity of often very mature business systems, developed over years
- Specific scenarios, patterns and structures of data forged through the application of system logic
- Those data structures can be difficult to recreate without the time-based application of that system logic

Real Application Architectures are Multi-generational





Masking Challenges

- Applications are large and complex
- Databases rely on referential integrity
- Applications validate data for consistency, relational integrity, application and integration integrity.
- Interconnected across the enterprise and externally with partners, industry bodies and government entities.
- Adherence to security standards
- Data has relations, time and date order
- Data can be dirty
- Dispersed data across hybrid cloud and enterprise data centres, locked in SAAS applications and esoteric data formats.



The Artificial Intelligence Masking Imperative

- Previously for Test data management – when it couldn't be masked you could put production-like controls around the data, create synthetic data or simply deal with faults created in production or reduce innovation and change.
- AI requires training datasets producing Language Models generated from your data –to ensure that sensitive data doesn't appear as part of the result sets then it's imperative that the data is masked
- It's also imperative that the data retains its utility and value and is continually refreshed.

A regulatory imperative?

South Africa's Joint Standard on Cybersecurity and Cyber Resilience (JS2)



BLOG July 15, 2025

How South Africa's Joint Standard 2 Changes the Data Compliance Landscape



Gary Hallam

DATA MANAGEMENT, SECURITY & COMPLIANCE



South Africa's Joint Standard on Cybersecurity & Cyber Resilience (JS2) is reshaping the regulatory landscape. Financial institutions must now rethink how they manage sensitive data. For data compliance leaders, this marks a critical shift where failing to adapt could bring serious consequences.

<https://www.perforce.com/blog/pdx/south-africa-joint-standard-2>



Taming Masking Complexity





Simplifying Masking Complexity

Automate Inventory Creation

- **Highly tuneable and granular sensitive data discovery**
- **Composable, Reusable**
- **Exportable for reporting and bulk actions**
- **API-driven for AUTOMATIC Creation and**
- **Integration with Business-level Data Governance tooling (like Collibra)**

Sensitive Data Discovery

Portable Inventory Profiles

- Database inventory columns/fields are matched to data domains using classification techniques
- Meta data profiling
 - Path (table & column name)
 - Data type (and field length)
- Data Level Profiling
 - Regular Expression Matching
 - List Matching
 - Checksum matching – Luhn (Mod 10 - used in credit cards & IMEI numbers), Mod97 (IBAN numbers)
- Inventory Profiles should be portable and able to be manipulated externally and by API

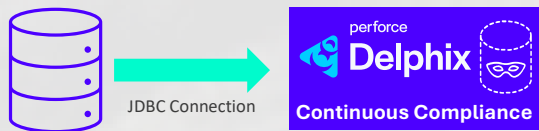
Simplifying Masking Complexity

Connect to ALL data sources

- **Relational Databases**
- **NoSQL Databases**
- **Cloud SAAS Applications (Workday, Salesforce...)**
- **Generic Connectors, upload-able JDBC drivers**
- **Hybrid Multi-cloud end-point support**
- **File Masking – multi-format (Delimited, XML, JSON, Parquet, JBASE, VSAM...)**

2

Database Profiling and Masking



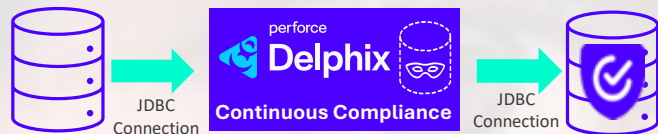
Profiling

Continuous Compliance connects to supported databases via JDBC to read metadata and sample data to discover sensitive information.



In-place Masking

Masking Engine connects to database via JDBC to update data in-place using SQL Select/Update operations.



On-the-Fly Masking

Masking Engine connects to source database and updates target database via JDBC using SQL Select/Insert operations.

Simplifying Masking Complexity

3

Mask deterministically, consistently

- Provides **REFERENTIAL INTEGRITY**
- Key or “seed” rotation – i.e. “change password”
- Algorithm federation
- One-way, permanent and
- Securely Reversible

Reversible Algorithms

Secured re-identify jobs

..	FIRST...	FIRST...	FIRSTNAME_M2	LASTNA...	LASTN...	LASTNA...	..	ADDRESS	CITY
1	Ronald	Vqnrli	TT4faIr9XT/rdvxnD3RPbQ==	Reagan	SMITH	Huldjw	..	49 Mxuuuzjzcxix Lznyn	Wmfp Nlsinb
2	George	Ynptpq	vkWRJhlZZYpL76bJvBgPhw==	Bush	JONES	Ikeg	..	4950 XKU Bpedtyaog	Wyrdfv
3	Barack	Riggso	nWH7DiiDRvvvGL8EcjNNJQ==	Obama	WILLIAMS	Vddes	..	5283 Yopgkxcfmjlk Pch	Dmqrrkbsur ER
4	Abraham	Enrdgup	6/HvSIFSaf7vp7nektwTVg==	Lincoln	TAYLOR	Dvklwiv	..	G 8gc Zwfwip	Yhmjxuwypkn
5	Thomas	Xywuhg	rz93IVU78nI30AQUB5Q==	Jefferson	BROWN	OvfhqzfkI	..	Kdagvdfsh Dslw	Shyuu Avjvnf

Character
Mapped

Tokenized

Mapped

Character
Mapped

Character
Mapped

Character
Mapped

..	FIRST...	FIRSTNAM...	FIRST...	LASTNA...	LASTNAME_M1	LASTNAME_M2	..	ADDRESS	CITY
1	Ronald	Ronald	Ronald	Reagan	Reagan	Reagan	..	40 Presidential Drive	Simi Valley
2	George	George	George	Bush	Bush	Bush	..	2943 SMU Boulevard	Dallas
3	Barack	Barack	Barack	Obama	Obama	Obama	..	1600 Pennsylvania Ave	Washington DC
4	Abraham	Abraham	Abraham	Lincoln	Lincoln	Lincoln	..	N 6th Street	Springfield
5	Thomas	Thomas	Thomas	Jefferson	Jefferson	Jefferson	..	Arlington Blvd	Falls Church

Simplifying Masking Complexity

4

Offer realistic, usable fictitious “masks”

- **Format-preserving**
- **Retain DATA UTILITY**
- **Provide data that meets validation rules**
- **Maintain data relationships (time order, location)**
- **Preferably without coding**
- **Reusable across data sets and sources**

Simple, comprehensive, intuitive

2. Algorithm Framework

1. Ready to use out of the box algorithms

Un-Masked Data → **Masked Data**

Transform sensitive data values into fictitious, yet relevant data values

Name:	Jane Smith	Name:	Andy Ron
SSN:	555-12-3136	SSN:	659-53-7945
Credit Card:	4356-3245-3242-7437	Credit Card:	4356-5132-2794-5816
Address:	145 Brown Swallow Ct. CA	Address:	1280 O'Gerald St, IL
DOB:	05-24-1963	DOB:	03-10-1973
Email:	j.smith@email.com	Email:	andy.ron@gmail.com

Select Framework

- Secure Lookup
- Character Mapping
- Payment Card
- Date
- Dependent Date Shift
- Name
- Full Name
- Email
- Segment Mapping
- Mapping
- Binary Lookup
- Tokenization
- Min Max
- Data Cleansing
- Free Text Redaction
- Numeric Expression**
- Extended

Create Numeric Expression Algorithm [Learn More](#)

Algorithm Name:

Description:

Expression:

Input Type:

Replacement Value for Nonconforming Data:

Constants [Learn More](#)

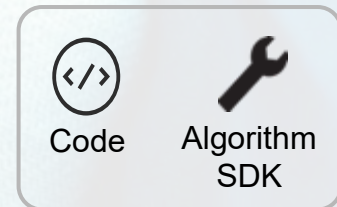
Name	Value
Java variable	Java expression

3. Algorithm Framework Builder SDK

Referential Integrity
Mask consistently within and across databases using deterministic algorithms

Data Delivery
Integrate masking w/ virtualization to deliver masked data in minutes

Business Semantics
Mask data while preserving business/application semantics and validation rules



Algorithm Framework Concept

- Binary Lookup – replaces a file like a pdf
- Character Mapping – each character replaced
- Data Cleansing – list of value replacements
- Date Replacement – removes date outliers
- Date Shift – shifts date element day, month, year etc
- Dependent Date Shift – shifts a date based on another date
- Email – specify name & domain and fallback option
- Free Text Redaction – search and replace based on list or regex
- Full Name – define algorithm for name components
- Mapping – define mapping list/table
- Min Max Date – removes outlier dates
- MinMax Number – removes outlier numbers
- Payment Card – valid card replacement
- Regex Decompose – multi-algorithm assignment
- Secure Lookup – list of replacement values
- String Algorithm Chain – apply multiple algos
- Tokenization – encryption based, reversible

Solve Masking Challenges Quicker

An Algorithm Framework should have the following capability built-in

- **Date Formats**

Automatically handle date format differences (including the difference between a text format date and a date format date)

- **Date and Number Relationships**

Maintain date and number relationships (i.e. before/after, higher/lower, 3 weeks later/earlier, £100 more/less)

- **Validation Compliance**

Maintain business rules and validations (i.e. under/over 18, under/over £12,570, Luhn checks, ID Number validation checks...)

- **Composite Fields**

Very simply extend masking consistency into composite fields (i.e. names in “full-name”, dates in ID Numbers, etc)

- **Database Constraints**

Automatically drop/rebuild/recreate indexes, constraints and triggers

- **Uniqueness**

Maintain uniqueness (where logically and mathematically feasible)

- **Data Scale & Distribution**

Perform at scale and run where your data is (federate algorithms to multiple engines across data centres and hybrid clouds)

- **Data Structure Change-proof**

Automatically handle database structure changes (for instance, new columns can be profiled by pressing a Ruleset refresh button).

Simplifying Masking Complexity

Be fast, scalable and secure

5

- **Millions of rows per minute masked**
- **Instant masked data cloning, delivery and**
- **Incremental replication**
- **Hyperscale for parallel processing**
- **SAAS – Deployed in Cloud**
- **Secure keys, AD Group Access Controls, Separation of Duties**

The Magic of Block Sharing

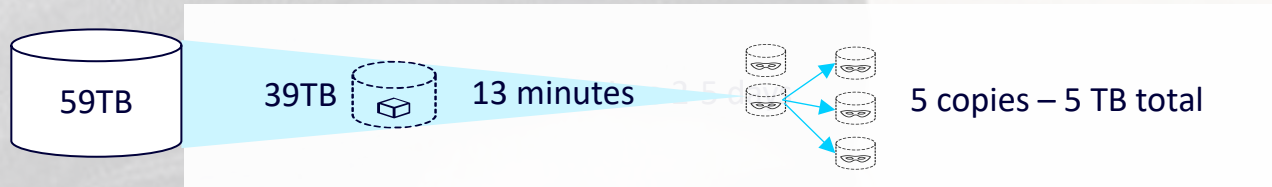


Simplifying Masking Complexity

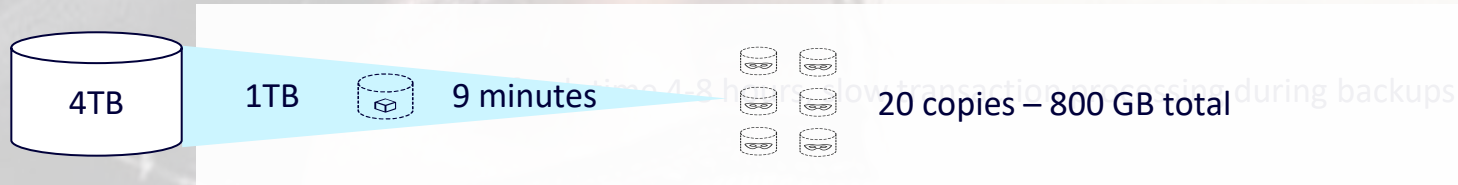
Block Sharing time and storage savings



Insurance Customer



Financial Customer



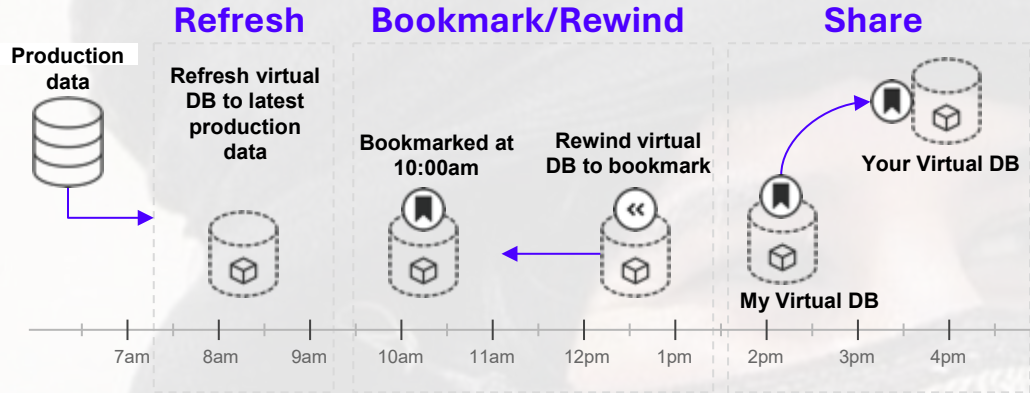
Banking Customer



Retail Customer



Block Sharing - Developer & Tester Autonomy



Save continuous history of data changes to enable self-service refresh, rewind, bookmark, and branch of data



Enable multiple versions of data for

- Destructive testing
- Alpha/Beta experiments
- Parallel application development
- Sharing data with different users

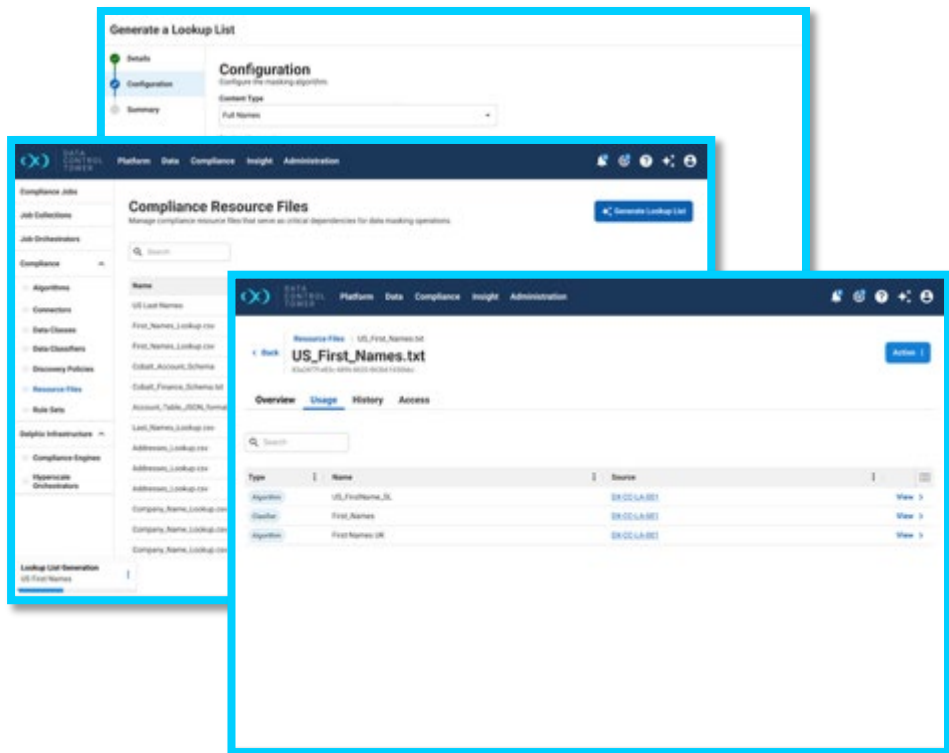
Give end users power to control data

- Version point-in-time data down to the second or transaction
- Enable unlimited independent data environments
- Deliver data to unique personas up to 100x faster

AI-Augmented Masking

Delphix in-product AI Augmentation

- Create synthetic datasets to replace (mask) non-prod, analytics, and AI data based on real business needs.
- Customize datasets based on region, vertical, and any other variables.
- Apply datasets selectively to mask non-prod and analytics PII/PHI, retaining analytical meaning without exposing data.



Questions?



Gary Hallam

Senior Account Executive
Delphix

P: +447951756650

E: gary.hallam@perforce.com