db tech showcase

ONLINE 2020

27TUE OCTOBER - 10THU DECEMBER

ようこそ！

DB tech showcase ONLINE 2020へ

DX推進で見直されるマスキングとは？

～データアナリティクスでの活用事例～

株式会社インサイトテクノロジー
プロダクト開発本部
髙橋　則行

本セッションのテーマ

- データマスキングとは
  - テストデータ生成用途
  - 属人化・個人情報の抽出・処理時間の問題
- DX推進・迅速なデータアナリティクス
  - クラウドリソースの活用
  - 昨今のマスキングニーズ
- 事例等
  - データマスキングの基本的な要件
  - オンプレやクラウドの参考製品
  - 今後のデータマスキング

データマスキングとは

データマスキングの概念

Wikipediaによると

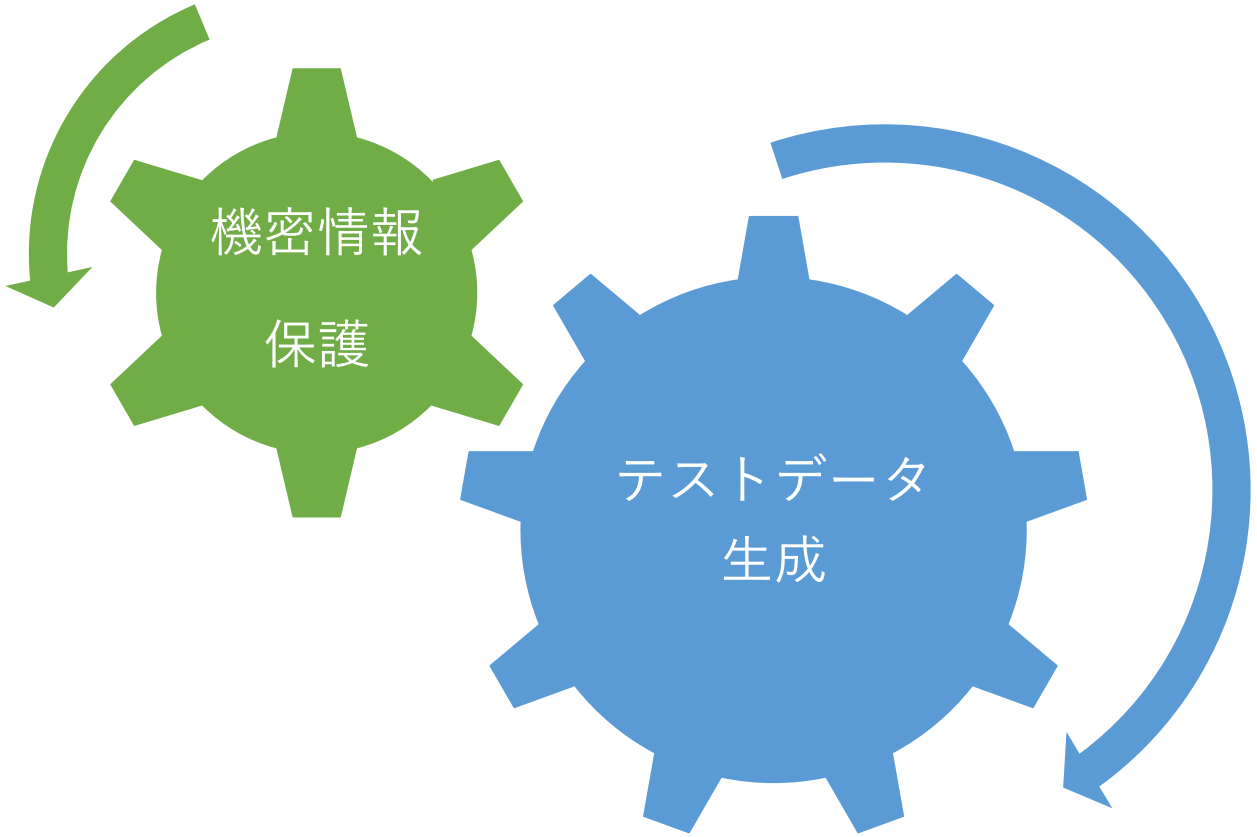
- 文字またはその他のデータを含む**元のデータを非表示**にするプロセス。
- **個人**を特定できる情報、商業的に**機密性の高い**データとして分類されるデータを保護するため。
    - https://en.wikipedia.org/wiki/Data_masking

データマスキングの概念

Oracle社によると

- 本番データベースから**非本番テスト・データベース**にコピーされる**機密情報**を、**外見は本物**であるもののマスキング・ルールに基づいて修正されたデータに置き換えるプロセス。
  - https://docs.oracle.com/cd/E57425_01/121/RATUG/GUID-2B0418D5-0D85-4F9B-9A7F-53665681BE25.htm

どうやら・・・



テスト目的

テクノロジーの分類

SDM (Static Data Masking)

- SDMではデータは使用前にマスキングされ、データベースなどで永続的に保護されます。

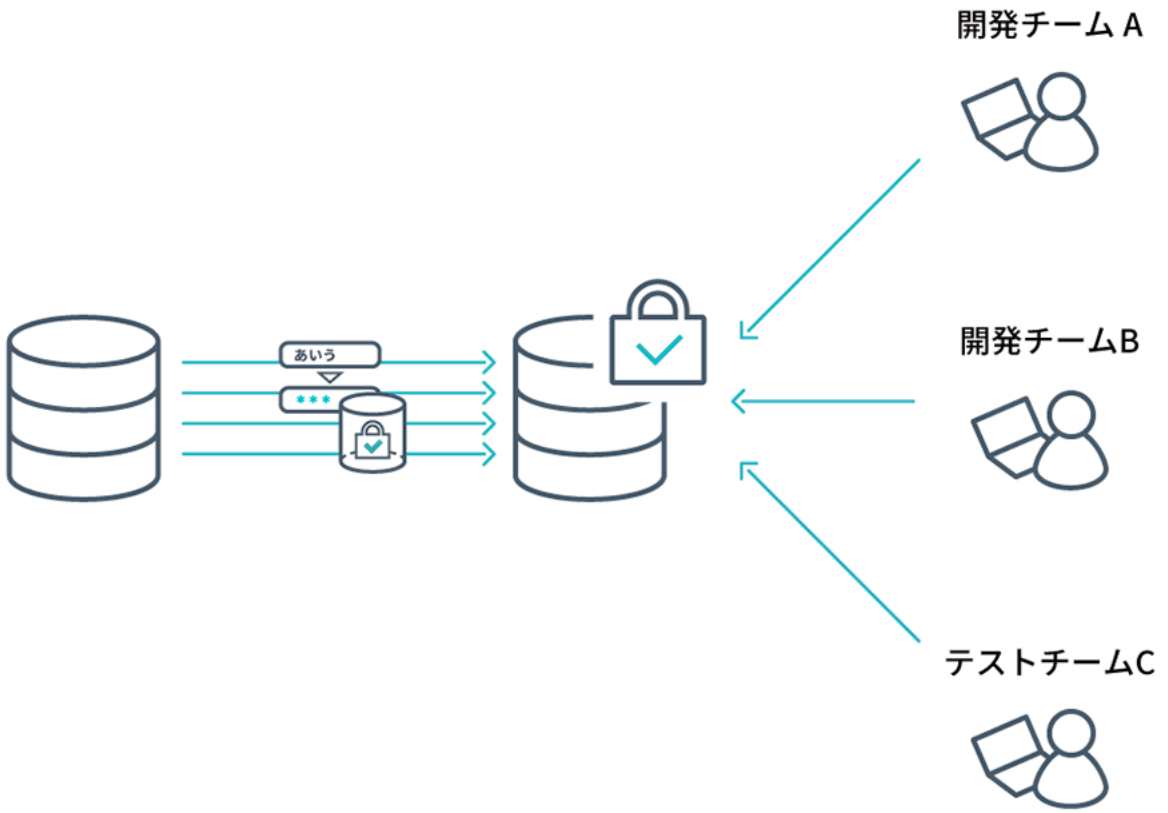
DDM (Dynamic Data Masking)

- DDMではデータにアクセス（リアルタイム）するタイミングで、ポリシーに基づいてマスキングを適用します。元データに対しては変更が行われません。

👉 クラウドでのデータ分析にも活用が進む

データマスキングの利用用途

- テストデータマネジメント
  - システム開発のためのテストデータ
    DBのkey属性や、複数テーブル間の関係性を維持しながら、マスキングデータを作成
  - 性能テスト
    DBのカーディナリティを維持し、本番同等のデータによる性能評価を行う



データマスキングの利用用途

- 分析データのマスキング
  - 各種分析用途に使用できるように個人情報の匿名化/秘匿化
  - 分析用データの加工

データマスキングの利用用途

- クラウド環境への移行
  - PoC 検証時の検証データの作成
- 限定的なデータ開示
  - 医師と医薬品メーカーで、医療情報の出しわけ

これまでの課題

属人化

- 開発者が手組みで実装
- SQLをベースにした限定的なマスキング

個人情報

- 本番稼働後は、事業部門に個人情報管理が散在
- 個人情報の特定に一苦労

対応時間

- マスキング対象の特定と承認の時間が増大
- データベースコピー、SQLでの更新など処理時間が増大

新たな課題

即時性

- 迅速なデータ分析のための処理時間短縮
- アジャイル開発

異種間

- テストDBに安価なリポジトリ
- オンプレ to クラウド

非定型

- 非定型データに含まれる個人情報の取り扱い
- 個人情報特定の自動化

DX推進・迅速なデータアナリティクス

オンプレの世界では

- 本番環境セグメントにDWH構築
- 保護された環境＝マスキング不要

オリジンデータで解析
初期コスト高い
柔軟な変更不可
複数会社の分析不可

やっぱりクラウドの活用は必須

db tech showcase ONLINE 2020

- クラウドにDWH構築
- 保護されない環境＝マスキング必要

パワフルなDWH
迅速なスタート
柔軟なリソース変更
複数社データ解析
セキュリティリスク

今更ですが、DXって

エリック・ストルターマン（2004年提唱）

1. デジタルトランスフォーメーションにより、**情報技術と現実が徐々に融合**して結びついていく変化が起こる。
2. **デジタルオブジェクト**が物理的現実の**基本的な素材**になる。
3. 固有の課題として、より本質的な情報技術研究のためのアプローチ、方法、技術を開発する必要がある。
    - https://ja.wikipedia.org/wiki/%E3%83%87%E3%82%B8%E3%82%BF%E3%83%AB%E3%83%88%E3%83%A9%E3%83%B3%E3%82%B9%E3%83%95%E3%82%A9%E3%83%BC%E3%83%A1%E3%83%BC%E3%82%B7%E3%83%A7%E3%83%B3

今更ですが、DXって

ガートナー社による定義

1. 業務プロセスの変革
2. ビジネスと企業、人を結び付けて統合する
3. **仮想と物理の世界を融合**して人／モノ／ビジネスが直接つながり、顧客との**関係が瞬時に変化**していく状態が当たり前となる
    - https://ja.wikipedia.org/wiki/%E3%83%87%E3%82%B8%E3%82%BF%E3%83%AB%E3%83%88%E3%83%A9%E3%83%B3%E3%82%B9%E3%83%95%E3%82%A9%E3%83%BC%E3%83%A1%E3%83%BC%E3%82%B7%E3%83%A7%E3%83%B3

DXで重要なのは・・・

db tech showcase ONLINE 2020

今の業務をソフトに置き換える

アナログの業務を
分析

それをソフトで
置き換え👉自動化

デジタルに
最適じゃない

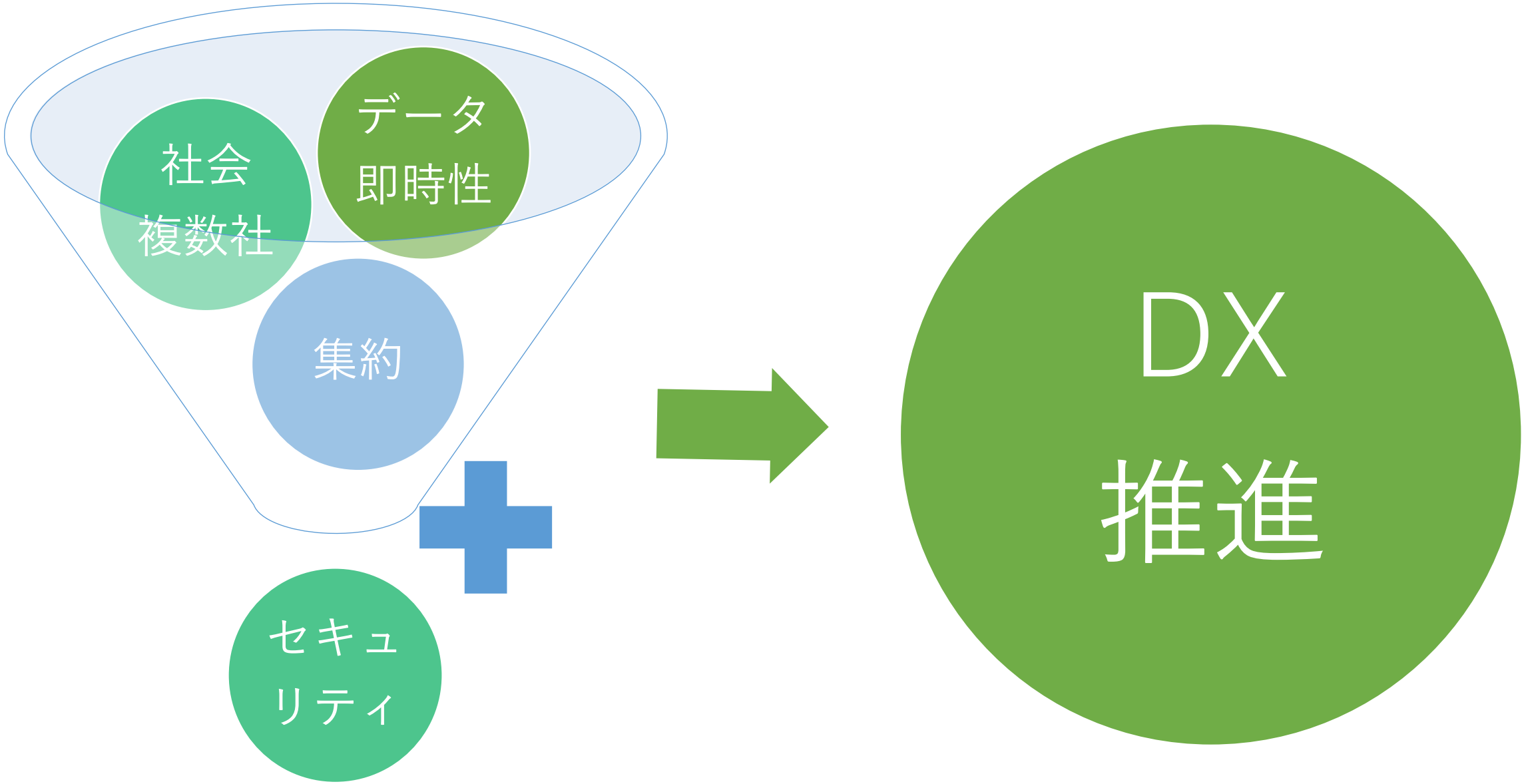
DXの定義に倣う

デジタルが基本素材
現実もデータ化

それを前提に
業務改革👉新プロセス

瞬時に繋がり変化し
ていく

DXで重要なセキュリティと即時性



昨今のマスキングニーズ

解析基盤をクラウドに移行

クラウドにあげる前にデータを保護

リアルタイムにデータ差分を同期

複数社のデータを集約

👉マスキングニーズ

活用シナリオ

解析基盤をクラウドに移行

クラウドにあげる前にデータを保護

リアルタイムにデータ差分を同期

複数社のデータを集約

👉マスキング済みデータをクラウドに保持

👉ETLツールによる転送中にマスキングプロセスを完了

👉ETLツールの差分同期中にマスキングを完了

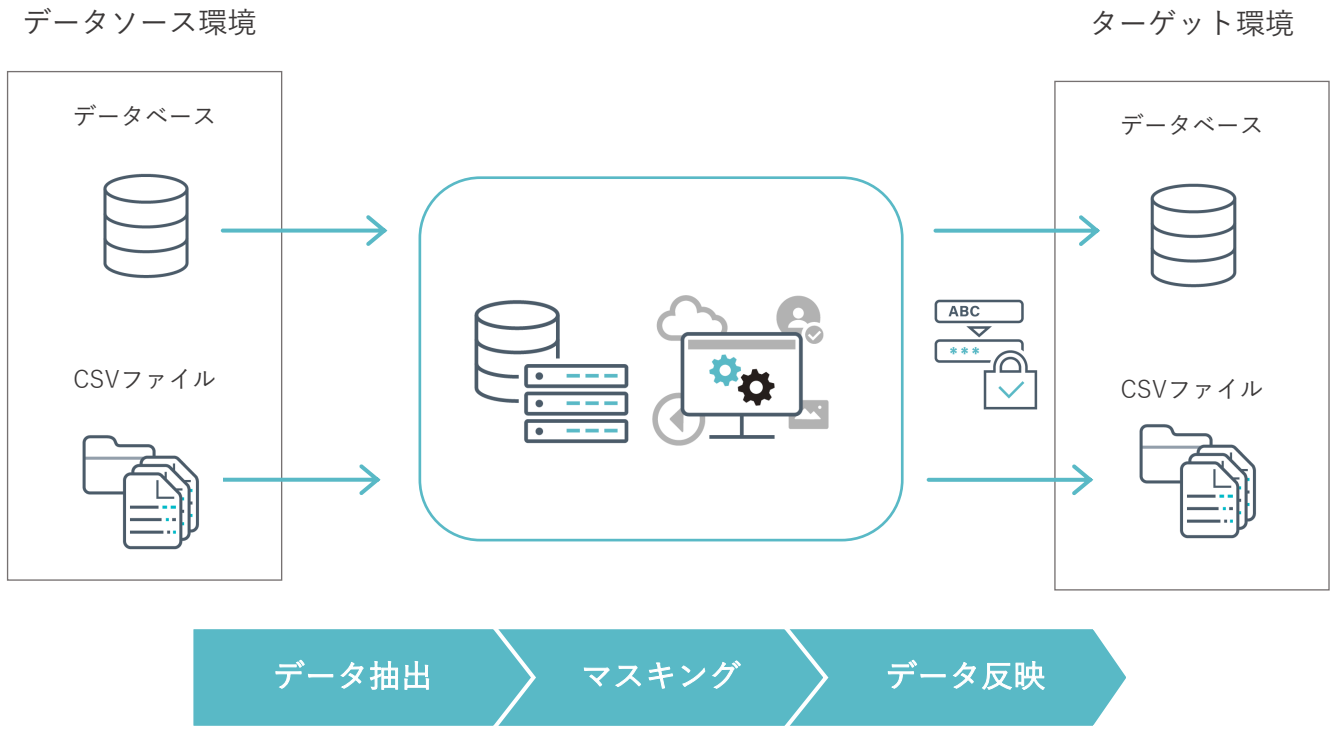
👉異種データを統合してマスキング（キーの一意なマスク）

事例等

データマスキングの基本要件①

db tech showcase ONLINE 2020

データベースやファイルをインターフェースとして、
『データ抽出』、『マスキング処理』、『データ反映』を行います。



インターフェース

- CSVファイル
- データベース
- ETL/DB仮想化ツールと連携

データマスキングの基本要件②

db tech showcase ONLINE 2020

- Encryption （暗号化）
  - 暗号化では、アルゴリズムを使用して暗号化テキストに変換、復号化キーにより読み取り可能な形式に戻せる。
- Tokenization （トークン化/トークンナイゼーション）
  - 機密情報をトークンと呼ばれるランダムに生成された代理データに置き換えて保存、機密情報を復元することはできない。
- Format Preserving Encryption（フォーマット保持暗号化）
  - 暗号文が平文と同じ形式になるようにデータを暗号化する。クレジットカード要件を満たすなど。
- Redaction （改訂、墨消し）
  - リダクションとは、機密データを削除するか別の文字で上書する。

マスキング例

フォーマット保持暗号化

文字タイプを保持して暗号化し、カーディナリティやアプリケーション動作に影響しない方式。

（元データ）

カラム1
123456
abcdEFGH
!#$%&()
これはテスト
（Null）
123abc
124-+xdf
xyzテスト

マスク処理

（マスク後データ）

カラム1
992379
CQsGKcvQ
?}^*.}?
衣励藻織ヲ盧
（Null）
992CRM
064@|Xvd
aRg猫予侑

元データと同じ文字タイプで、マスク後データを作成

元データに複数の文字タイプが混在している場合も
文字順序に従って、同じ文字タイプでマスク後データを作成

・数字 ⇒ 数字、英字 ⇒ 英字、記号 ⇒ 記号、日本語 ⇒ 日本語 などの変換
・文字長はマスク前後で変わらない
・クレジットカード番号のチェックディジットを維持するものなどがある。

マスキング例

リダクション（改訂、墨消し）

電話番号の市外局番までのフォーマットを維持し、それ以降を墨消しした例。
プログラム動作として市外局番を要求するケースなどに活用。

（元データ）

日付	電話番号	…	…
2020-01-10	03-5475-1450	…	…
2020-01-11	011-211-4886	…	…
2020-02-09	06-6359-1450	…	…
2020-02-09	052-856-3362	…	…

マスク処理

（マスク後データ）

日付	電話番号	…	…
2020-01-10	03-9345-XXXX	…	…
2020-01-11	011-689-XXXX	…	…
2020-02-09	06-9124-XXXX	…	…
2020-02-09	052-448-XXXX	…	…

上記の電話番号のマスキング概要

- 「-」を区切り文字として、3つのテキストに分割
- 第1テキストを「**変換なし**」で設定
- 第2テキストの文字列を文字種「**数字**」で設定
- 第3テキストの文字列を「X」でリダクション

マスキング例

トークナイゼーション

辞書データを用いて、本番データに近いが、全く異なるデータに置き換えるマスキング。
画面表示が崩れないようにテストするケースなど。

（元データ）

氏名	住所	...	...
東 宜昭	東京都渋谷区恵比寿...	...	...
菅野 愛佳	北海道札幌市中央区...	...	...
長田 理沙	大阪府大阪市北区...	...	...
加納 貞久	愛知県名古屋市中村区...	...	...

マスク処理

（マスク後データ）

氏名	住所	...	...
宮崎 孝市	東京都渋谷区恵比寿...	...	...
庄司 沙也	北海道札幌市中央区...	...	...
川崎 成美	大阪府大阪市北区...	...	...
大平 智和	愛知県名古屋市中村区...	...	...

（辞書ファイル）

氏名	...	...
大平 智和	...	...
庄司 沙也	...	...
宮崎 孝市	...	...
川崎 成美	...	...
⋮		

辞書ファイルからデータを抽出して、適用
（乱数を作成して、辞書ファイルより抽出）

オンプレの参考製品：Oracle

- Oracle Data Masking and Subsetting Pack
  - テスト、開発などの目的で本番データをコピーする際、データをサニタイズしたり、不要データを取り除く。
  - データベースとしての一貫性を維持。
  - 国民番号、クレジットカード番号といった個人情報を含んだカラムを検出。
  - データベース内に定義された親子関係も自動的に検出。
  - 条件に依存したマスキング、ある入力によって生成されるマスク出力が常に一致するマスキングを提供。
  - インプレースで上書き、マスクデータエクスポートの２方式。
  - Oracle EEライセンスと同契約数の購入が必要。
    - https://www.oracle.com/technetwork/jp/database/options/data-masking-subsetting/overview/index.html

クラウドの参考製品：Google

- Google Cloud Data Loss Prevention
  - 特に機密性の高いデータを検出、分類、保護するためのフルマネージドサービス。
  - 名前、メールアドレス、電話番号などを分類するinfoTypeが120種類用意されている。
  - Cloud Storage、BigQuery、Datastore の機密データをスキャン、分類。
  - 構造化データと非構造化データのマスキング、トークン化、変換をサポート。
  - k-匿名性、l-多様性といった統計指標を測定することが可能。
    - https://cloud.google.com/dlp?hl=ja

ベンダーの参考製品：Insight

- Insight Data Masking
  - データベースの専門家が、現場の声を吸い上げて作った国産マスキングツール。
  - マルチDB、オンプレ・クラウドDB、CSV連携などあらゆる局面で対応できる。
  - Qlik ReplicateなどETLに組み込める唯一のマスキングツール。
  - 日本語への対応に重点を置いたマスキング機能。
  - カーディナリティを保持し、性能試験に活用可能。
  - インメモリと並列処理による高い性能。
  - データ転送時、同時にゼロSecondマスキングすることが可能。
  - Insight Asirと連動し機密情報のAI検出に対応予定。

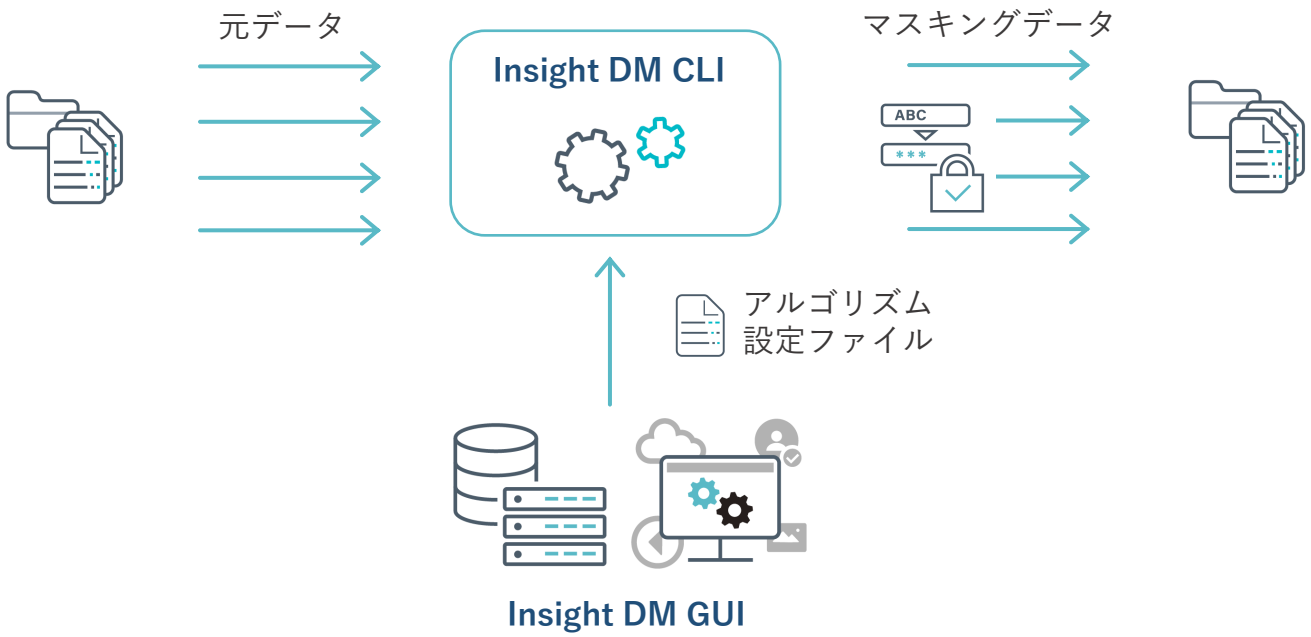
Insight Data Masking 製品概要

db tech showcase ONLINE 2020

DBやCSVをターゲットにマスキング

Command Line Interface（CLI）プログラムにより、DBやCSVファイル同士、あるいは相互変換しながらのマスキングを実行。

Windows / Linux環境に対応。



- データベース、汎用機などから抽出したCSVファイルをマスキング
- ETLツールと連携して、データ加工とデータマスキングを実施
- 並列でマスキングジョブを実行することで処理時間を短縮

Insight Data Masking 製品概要

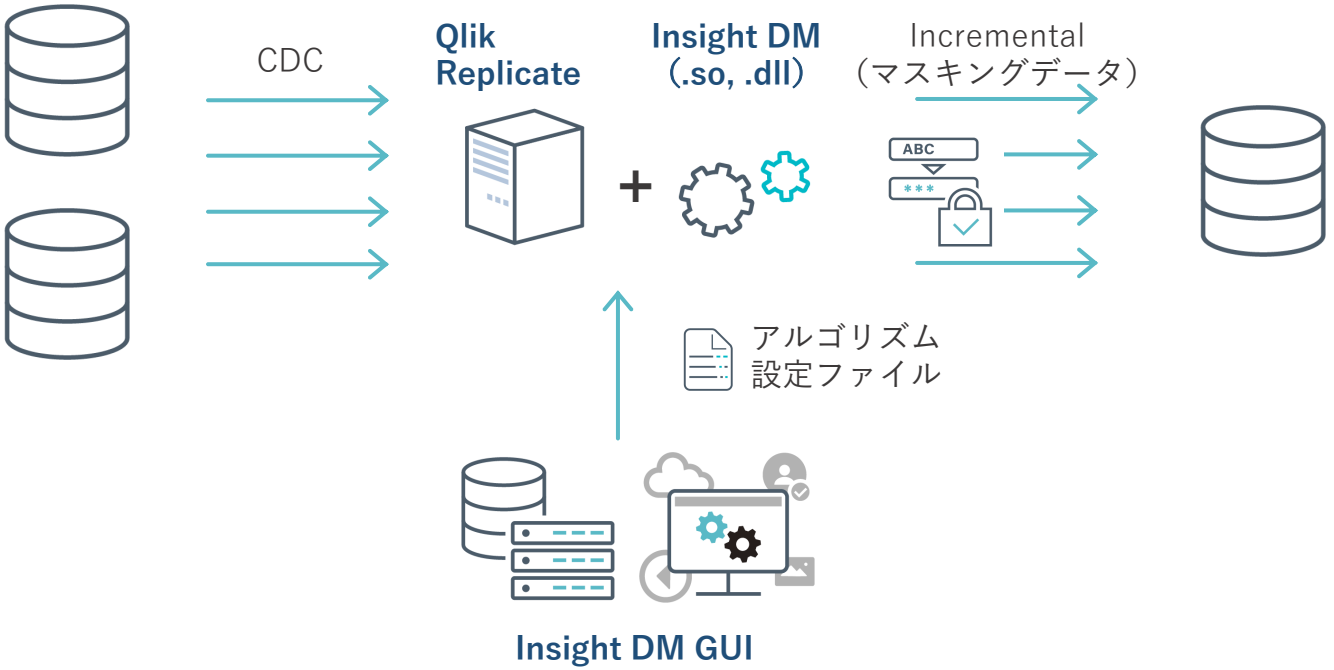
db tech showcase ONLINE 2020

Qlik Replicate との連携（Qlik Replicate Edition）

Qlik Replicate の「User-Defined Transformations」機能を用いて、

マスキングアルゴリズムをインラインで提供します。

ネットワーク転送のレーテンシの時間内でマスキングが完了し、処理時間がHidingされます。



- リアルタイムにマスキングされたデータをターゲットDBに反映
- ゼロSecondマスキング
- 差分更新に対応
- オンプレ、クラウド、異種DB、Hadoopなどの組合せが可能

ベンダーの参考製品：Insight

db tech showcase ONLINE 2020

Pros

- データベースの種別に縛られずマスキングできる。
- クラウドとオンプレに縛られずマスキングできる。
- ETLと連携し、高速に差分更新ができる。

Cons

- データベースサーバとは別のリソースが必要。
- 日本語に強いが、多言語に弱い。

本セッションのテーマ（再掲）

- データマスキングとは
  - テストデータ生成用途
  - 属人化・個人情報の抽出・処理時間の問題
- DX推進・迅速なデータアナリティクス
  - クラウドリソースの活用
  - 昨今のマスキングニーズ
- 事例等
  - データマスキングの基本的な要件
  - オンプレやクラウドの参考製品
  - 今後のデータマスキング

db tech showcase

ONLINE 2020

27TUE OCTOBER - 10THU DECEMBER

Thank You!