MarkLogic

将来のニーズがわからない時代のデータガバナンス

マークロジック

シニアセールスエンジニア

ダニエル・ホルゲート

© COPYRIGHT 2016 MARKLOGIC CORPORATION. ALL RIGHTS RESERVED.

MarkLogic

アジェンダ

- 自己紹介
- データガバナンスとは?
- 本来のデータガバナンス実現の例
- MARKLOGICによるデータガバナンス
- データガバナンスの機能
- データガバナンスに関するMARKLOGICその他の機能
- まとめ

© COPYRIGHT 2016 MARKLOGIC CORPORATION.ALL RIGHTS RESERVED.

MarkLogic

自己紹介

- ダニエル　ホルゲート (Daniel Holgate)
- シニアセールスエンジニア
- MarkLogic、オーストラリア・シドニー所属
- MarkLogicに入社する前、JAVAによる開発、システム設計、構築の経験
- リレーショナルデータベース上の開発　＆　システム統合プロジェクト

© COPYRIGHT 2016 MARKLOGIC CORPORATION.ALL RIGHTS RESERVED.

データガバナンスとは何か？

データガバナンス（DG）とは、エンタープライズ内のデータに関する可用性、ユーザビリティ、整合性、セキュリティに関する包括的な管理のこと。エンタープライズは、正確さ、アクセシビリティ、一貫性、完全性、データの格納方法、アーカイブ化、バックアップ、事故/盗難/攻撃に対する防御などに関する説明責任がある。また政府による規制を順守している必要がある。

http://searchdatamanagement.techtarget.com/definition/data-governance

- DGはテクノロジー的というよりもプロセス的

しかし。。

- セキュリティ
- トラッキング
- 信頼性
- 照合
- 品質
- ストレージ
- 監査
- コンプライアンス

© COPYRIGHT 2016 MARKLOGIC CORPORATION. ALL RIGHTS RESERVED.

MarkLogic

「実用的な」データガバナンスとは？

- われわれ全員が、すでにデータガバナンスの作業を行っている
- しかし今のところ現実は、理想のデータガバナンスからは程遠い

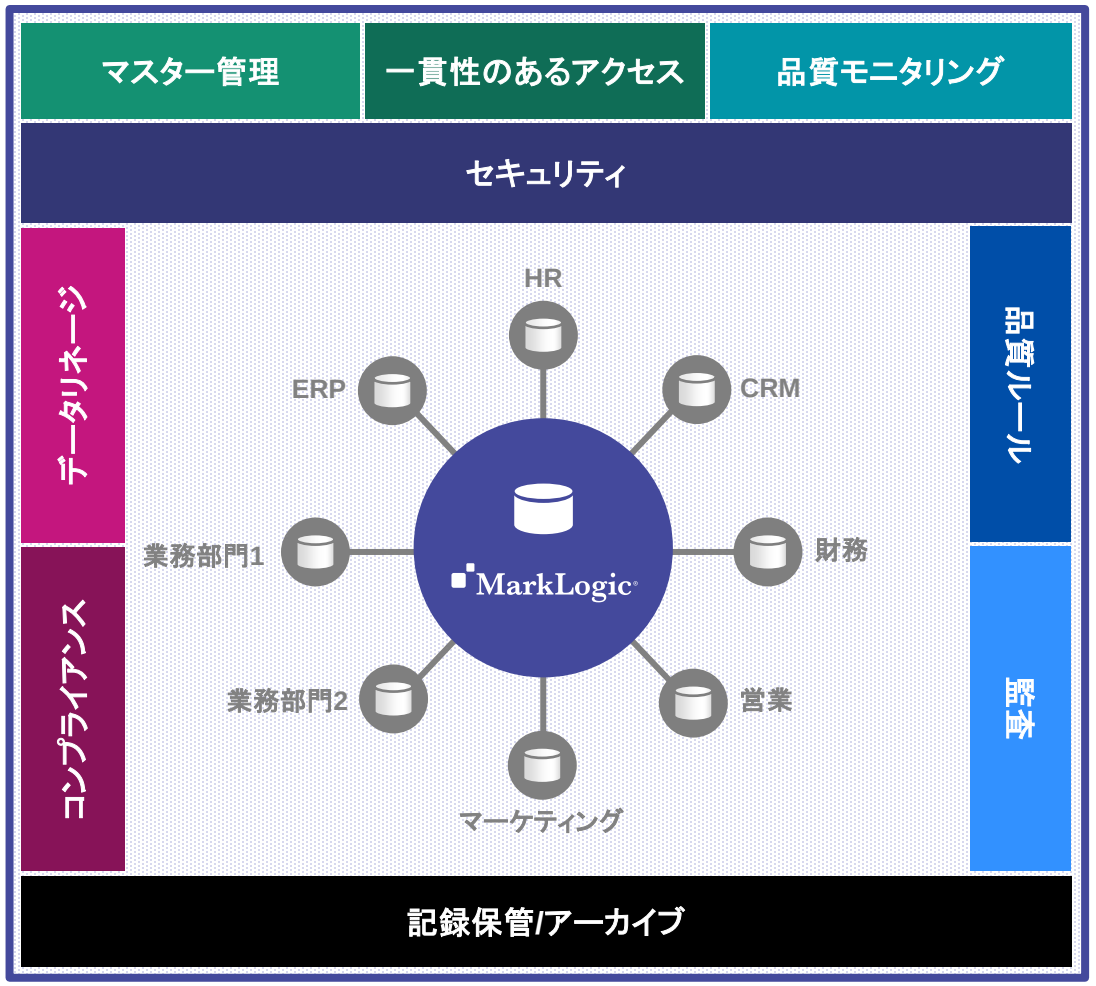
© COPYRIGHT 2016 MARKLOGIC CORPORATION. ALL RIGHTS RESERVED.

MarkLogic

テクノロジーに基づくデータガバナンス

- 避けなければならないこと
  - データ漏洩・流出
  - 罰金や罰則
- 実現しなくてはならないこと
  - 安全で一貫性のあるデータアクセス
  - トラッキングと信頼性を保証
  - 単純さとデータガバナンスプロジェクトの成功

© COPYRIGHT 2016 MARKLOGIC CORPORATION.ALL RIGHTS RESERVED.

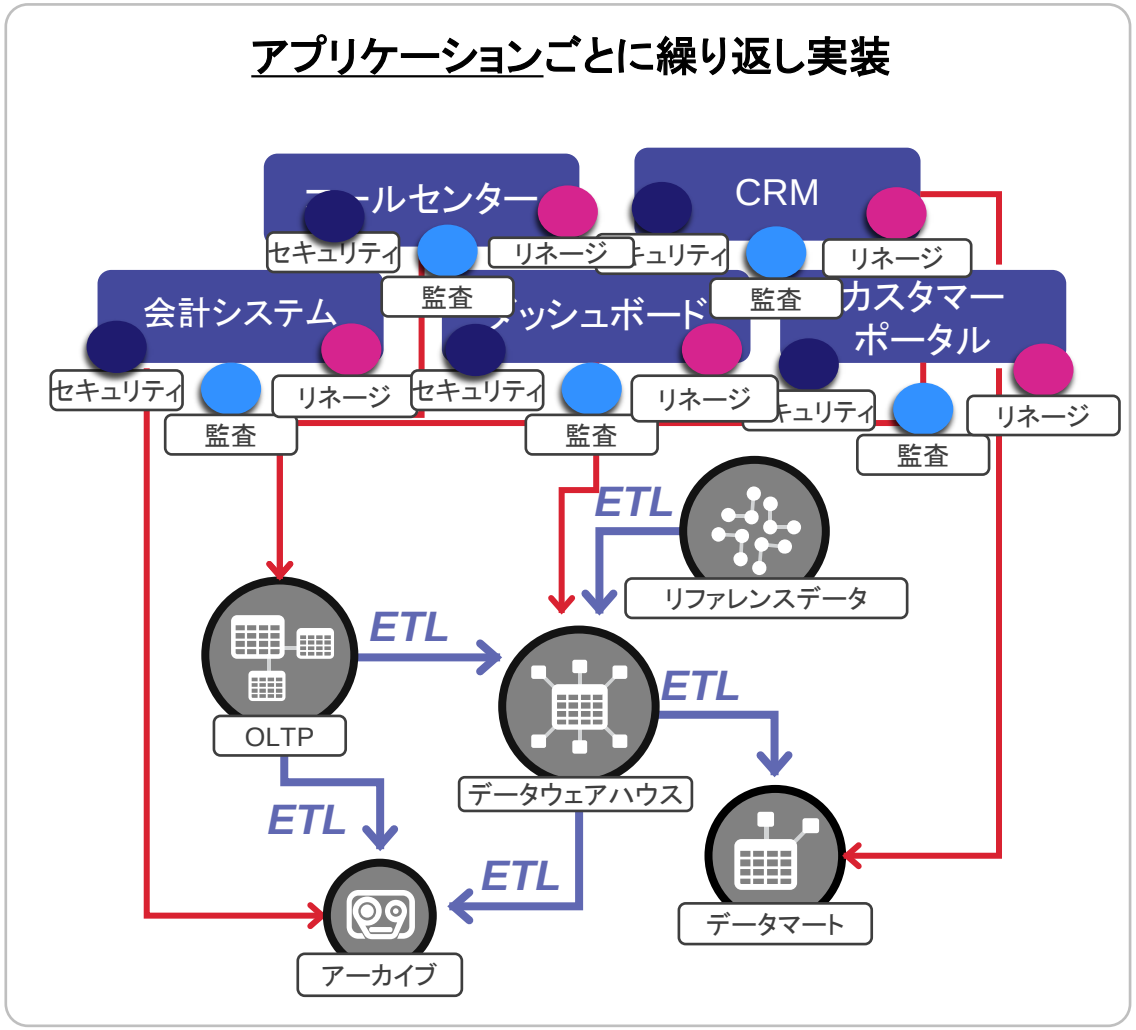


MARKLOGICのアプローチ

データハブモデルによるデータガバナンス

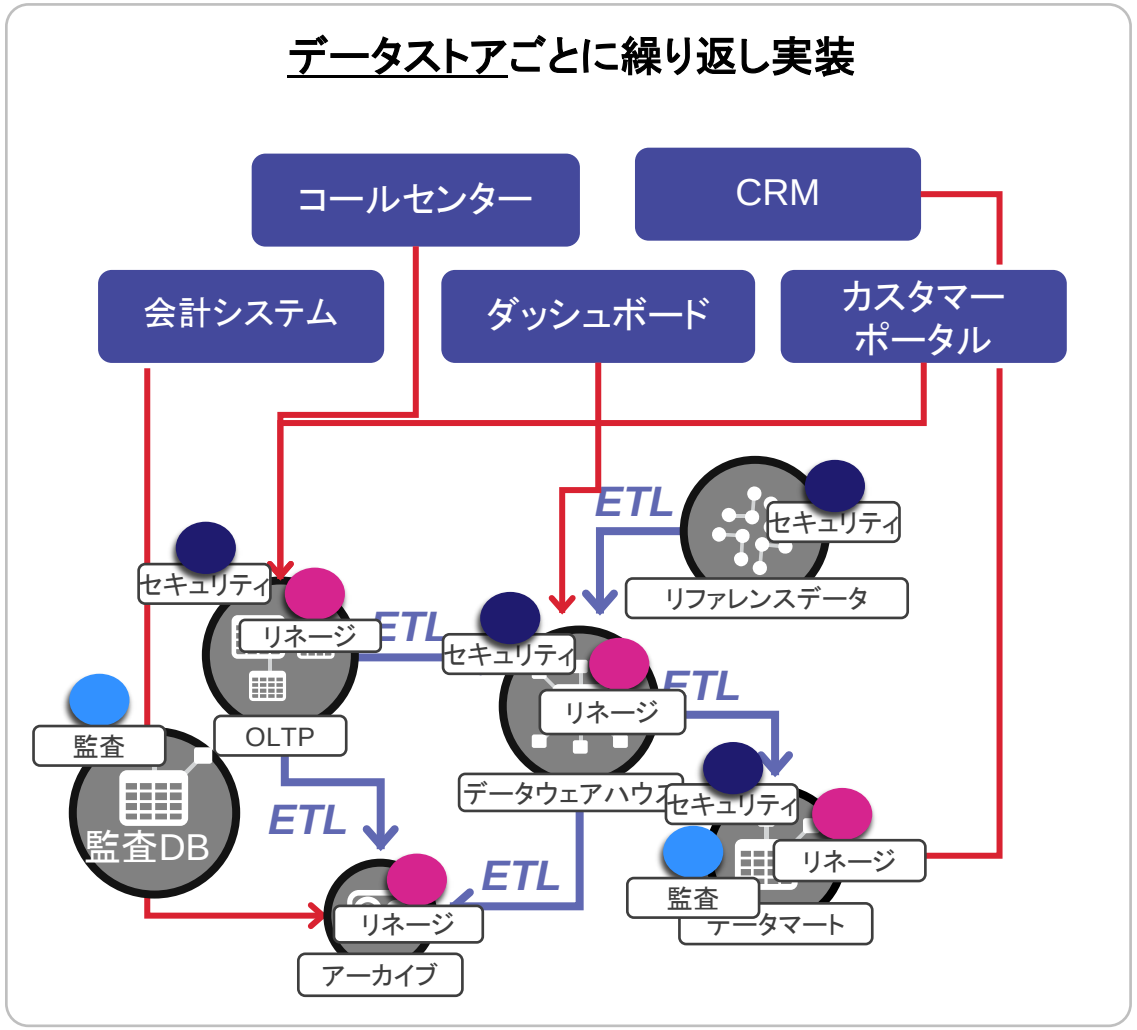
© COPYRIGHT 2016 MARKLOGIC CORPORATION. ALL RIGHTS RESERVED.

データガバナンスの実現　1



© COPYRIGHT 2016 MARKLOGIC CORPORATION. ALL RIGHTS RESERVED.

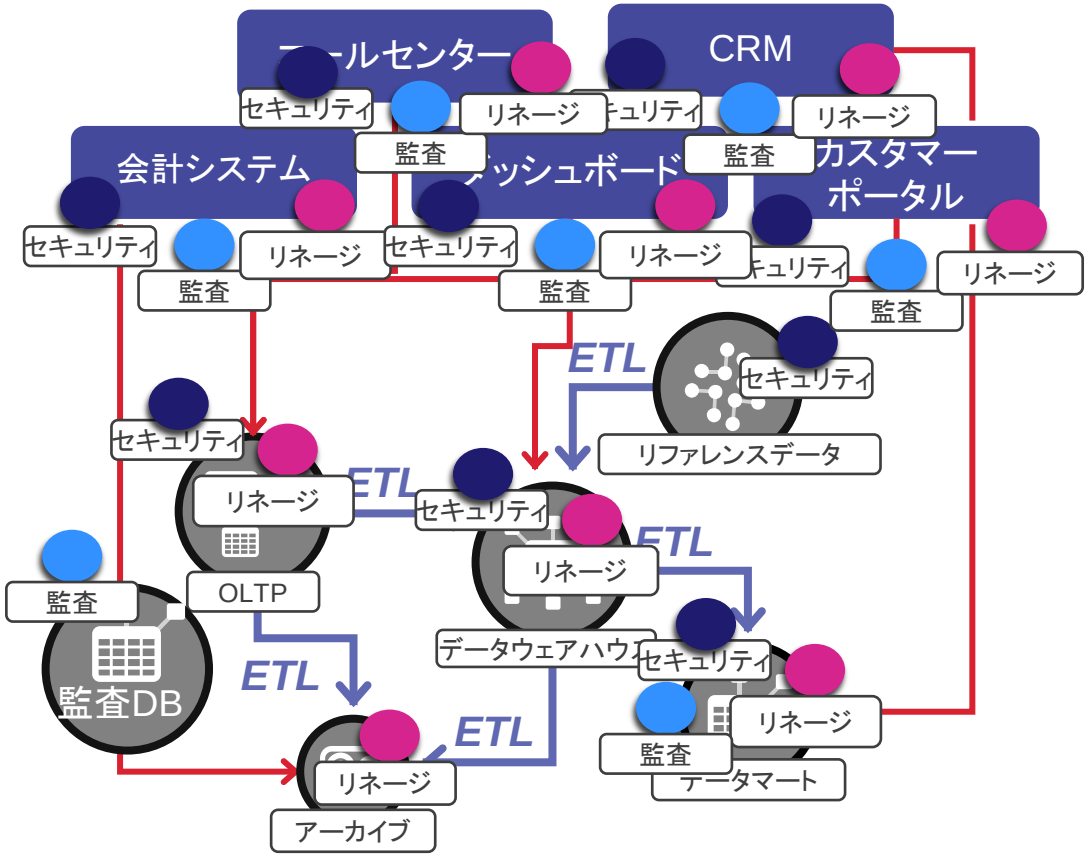
データガバナンスの実現 2



© COPYRIGHT 2016 MARKLOGIC CORPORATION. ALL RIGHTS RESERVED.

データガバナンスの実現 3

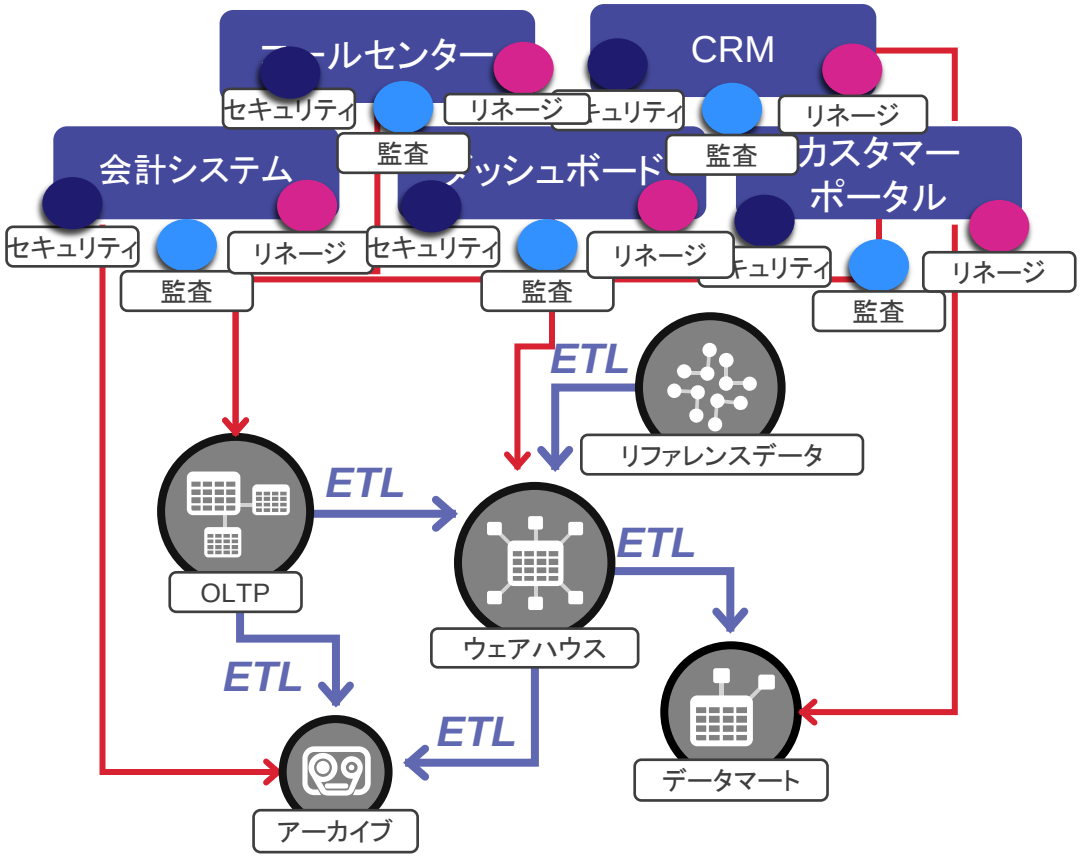
アプリケーション層、データ層両方で実装



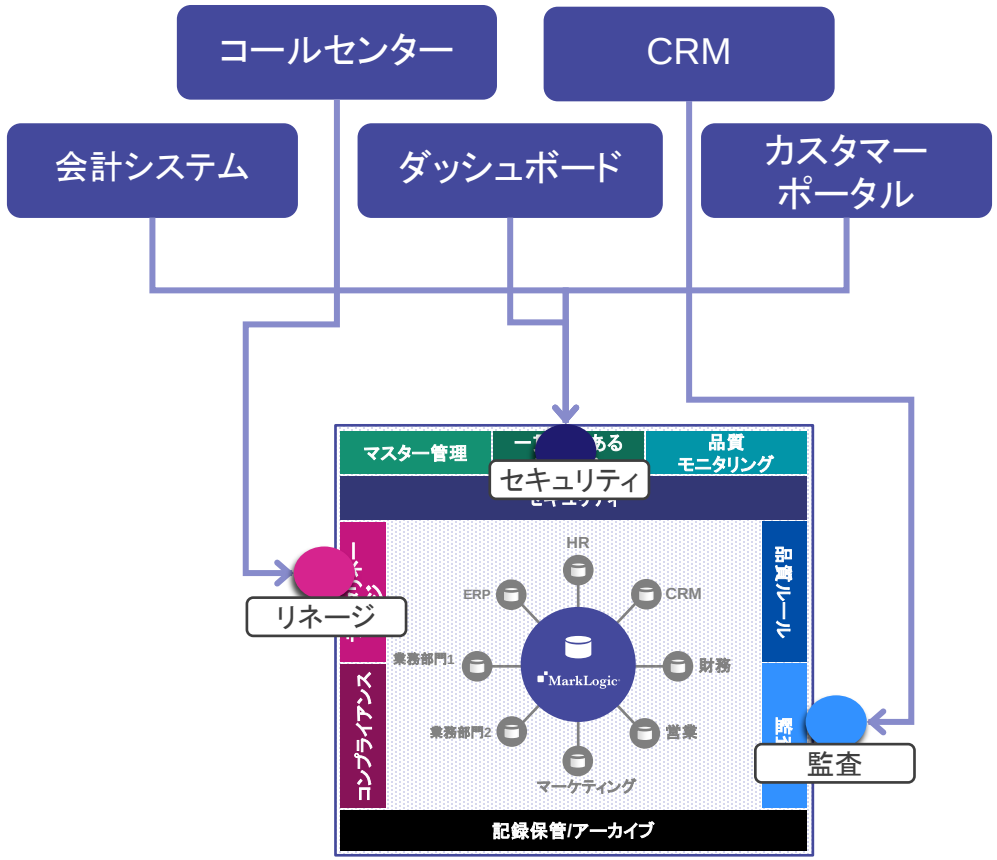
© COPYRIGHT 2016 MARKLOGIC CORPORATION. ALL RIGHTS RESERVED.

データハブによるデータガバナンスの実現

アプリケーションごとに繰り返し実装



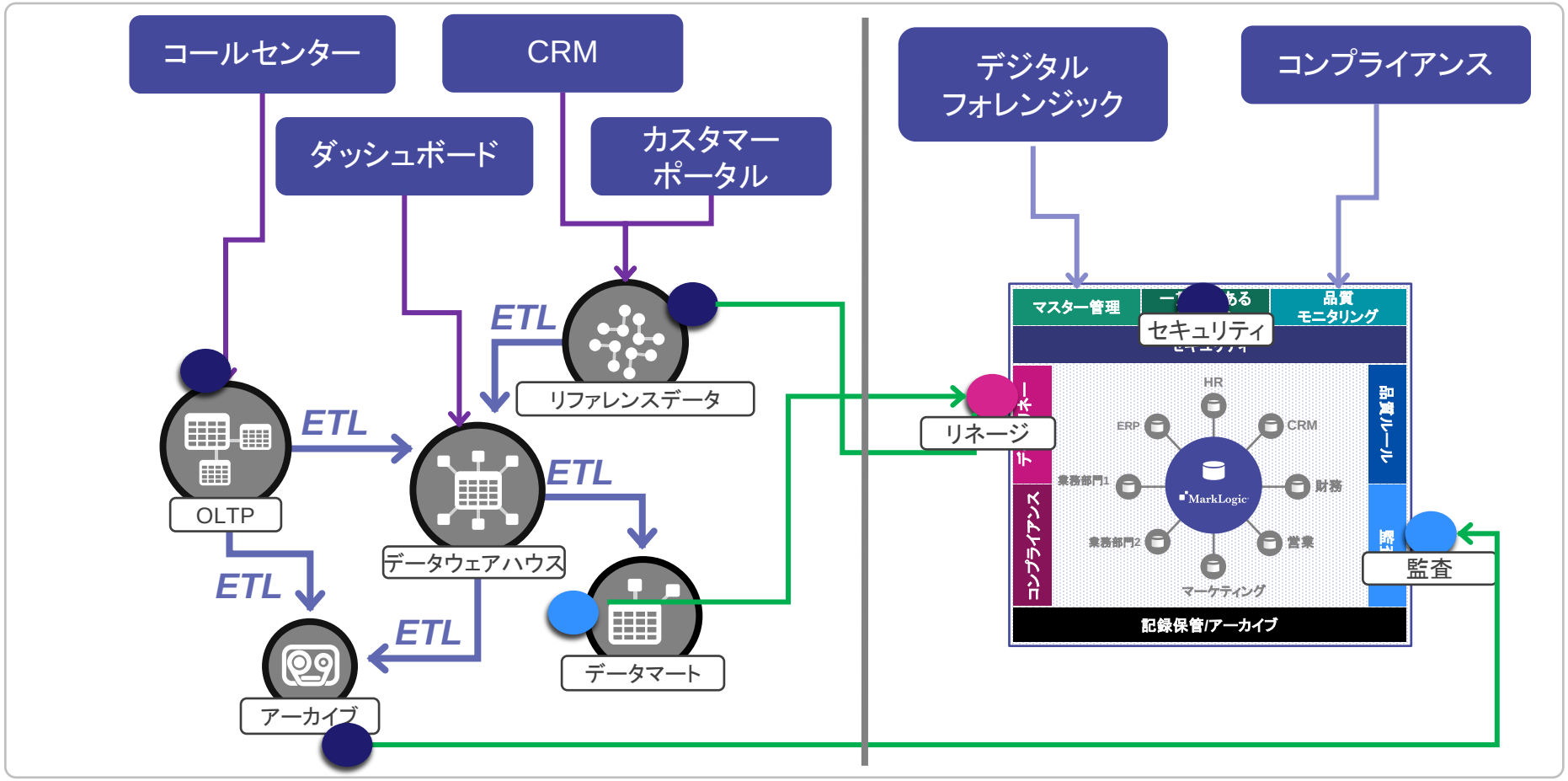
データハブ/データ層で一元化



© COPYRIGHT 2016 MARKLOGIC CORPORATION. ALL RIGHTS RESERVED.

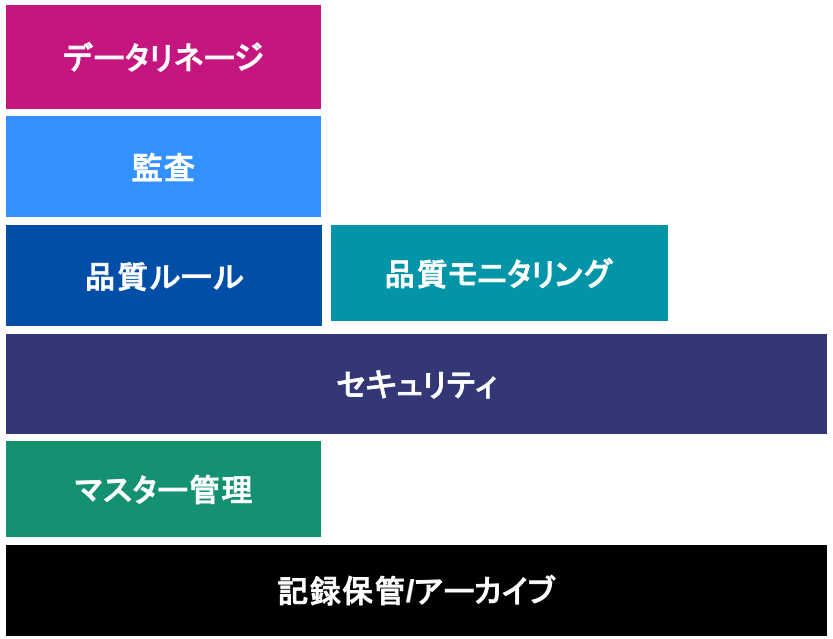
データガバナンスハブ

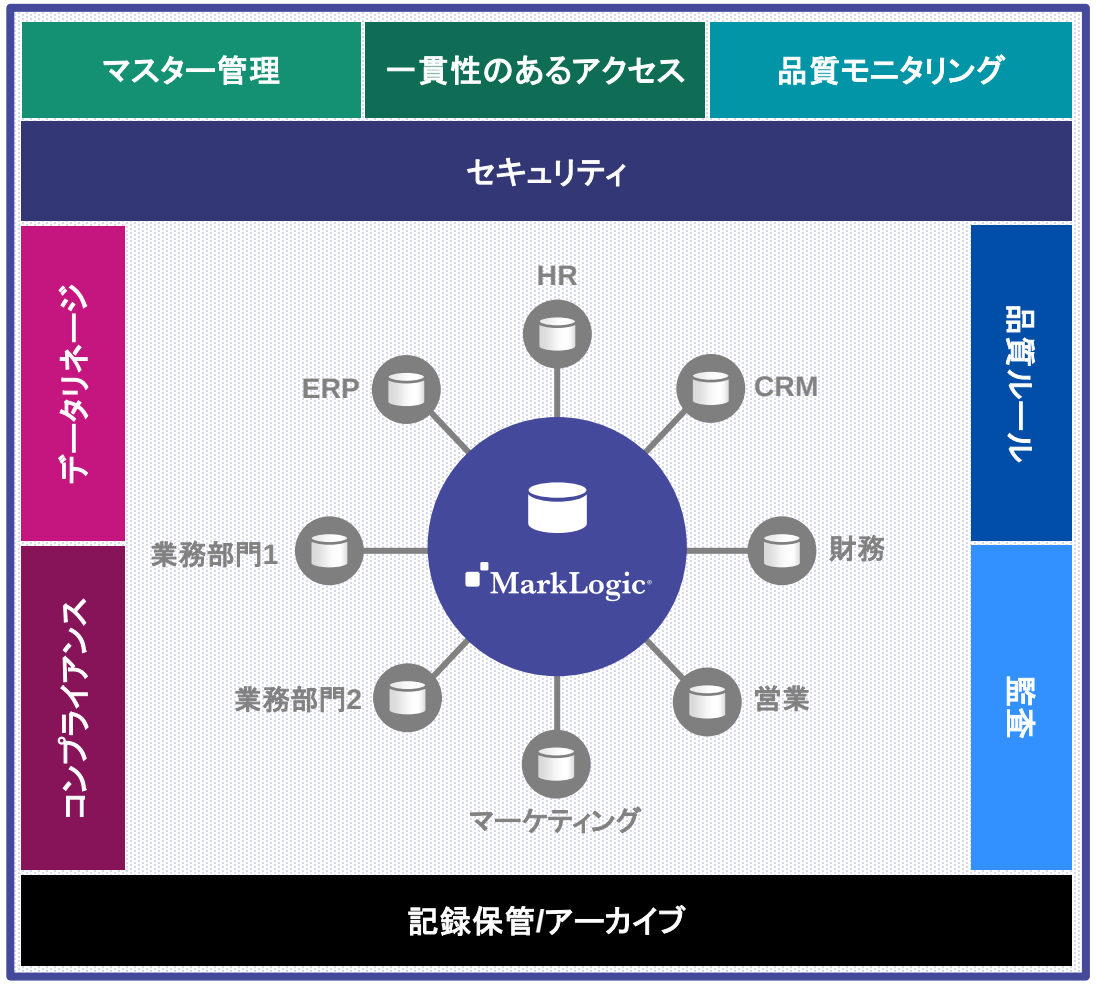
データガバナンス用のデータ(データ系統、監査履歴)を保管



© COPYRIGHT 2016 MARKLOGIC CORPORATION. ALL RIGHTS RESERVED.

データガバナンスの機能





© COPYRIGHT 2016 MARKLOGIC CORPORATION. ALL RIGHTS RESERVED.

MarkLogic

データリネージ（データの出自・系統）

DATA LINEAGE

ドクター：「異常がないことを確認するために、もう一度MRIを受けてください」

- なぜ？　インポートされたカルテ情報には、すでにMRI検査済みとある
- しかし単純で説明のないデータ（*リネージ＝出自がないもの*）は信頼性に欠ける

カルテ

検査種別：MRI
　詳細種別：軸位断（axial）、T2強調MRI、頭部・頸部
　異常なし

リネージ：
　依頼者：Saul Yakenflaster, MD
　内容確認：Saul Yakenflaster, MD

インポート詳細：
　LIMS Import System / August 22, 2014 / Rec no: LIMS-992-F47b
　Quest Diagnostics社 / MRI-HN-9929331

© COPYRIGHT 2016 MARKLOGIC CORPORATION. ALL RIGHTS RESERVED.

MarkLogic

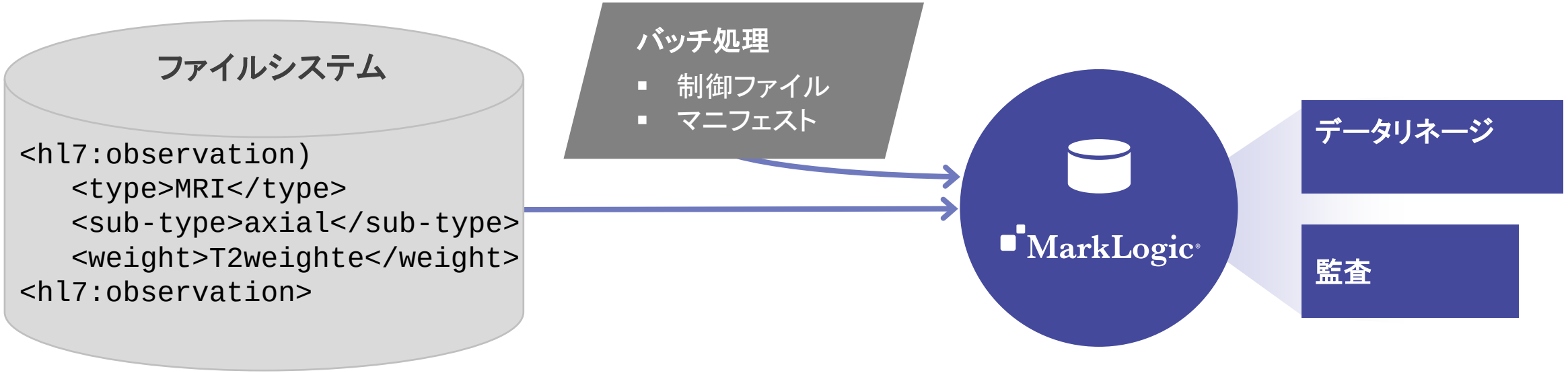
MarkLogicにおけるデータリネージ技術

DATA LINEAGE

MarkLogicがスキーマレスなのでデータを「As Is」で格納するでけでなく。。

- WebServiceの場合、入力のSOAPリクエストのXML
- バッチのインポートジョブ用のバッチ制御ドキュメント
- バッチのマニフェスト

。。というデータもリネージ情報として保管できる



© COPYRIGHT 2016 MARKLOGIC CORPORATION. ALL RIGHTS RESERVED.

MarkLogic

データリネージ

データリネージのメリット

- データの来歴の明確化
- 矛盾する情報の修正
- データの理解を助ける
- データの追跡や取り消し作業

```
<loan loan-id="96583">
    <date>2010/02/22</date>
  <applicant>
    <name
      <full>Robert Jones</full>
    </name>
    <address>123 Main st </address>
    <name
      <full>Bob Jonas</full>
    </name>
    <address>123 Main st</address>
[ ...]
```

??

??

MarkLogic

データリネージ

- リネージはメタデータ
- 注釈のようなもの
- XMLやJSONで対応ドキュメントごとあるいはフィールドごとに追加

```
<loan loan-id="96583">
    <date>2010/02/22</date>
  <applicant>
    <name
      source="loan-application" id="L-9322">
      <full>Robert Jones</full>
    </name>
    <address>123 Main st</address>
    <name
      source="cust-svc-phonecall">
      <full>Bob Jonas</full>
    </name>
    <address>123 Main st</address>
[ ...]
```

MarkLogic

データリネージ

- セマンティックデータも利用可

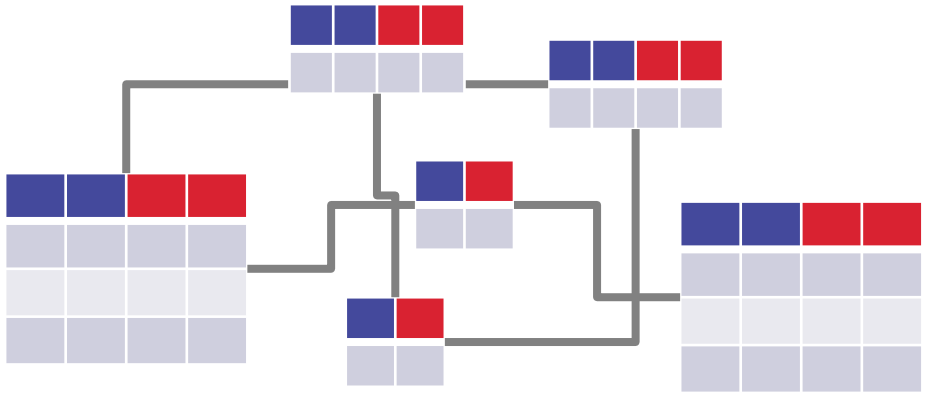
```
<loan loan-id="96583">
    <date>2010/02/22</date>
  <applicant>
    <name
      id="LOAN-142934">
      <full>Robert Jones</full>
    </name>
    <address>123 Main st</address>
    <name
      <full>Bob Jonas</full>
    </name>
    <address>PO Box 9922, Chicago, IL<>
[ ...]

loan isA fin:loan
  hasName foaf:name#BobJones.

foaf:name#BobJones
  hasLable "Bob Jones"
  fromSourceType fin:loanApplication.
```

MarkLogic

RDBMSにおけるデータリネージ

1. バッチ処理に関するメタデータ及びリネージ情報を保管する場合
   - バッチ情報用の新しいテーブルやデータモデルを設計・作成する作業が発生
2. 要素ごとにデータリネージ又はソースシステムの情報も保管する場合
   - データモデルにソースシステムを識別するためのカラムを追加
   - アプリケーションが利用するすべてのインサート文の書き換え
3. データの移動履歴、ソースシステム、変更に関する他のデータを保管する場合
   - さらなるデータモデルを設計・作成する作業が発生

Address	Name	Address_Src	Name_Src
123 Main st	Robert Jones	Call Center	Call Center



© COPYRIGHT 2016 MARKLOGIC CORPORATION. ALL RIGHTS RESERVED.

MarkLogic

データリネージの使用法

- ユーザーに表示できるのが最も理想的なケース
- ソースシステム、人、プロセスに関する品質上の問題の追跡
- デジタルフォレンジック（デジタル鑑識）
- データの取り消しや修正作業用に、データの点と点を結ぶ

検査種別：MRI
　詳細種別：軸位断（axial）、T2強調MRI、頭部・頸部
リネージ：
　依頼者：Saul Yakenflaster, MD
　内容確認：Saul Yakenflaster, MD
　インポート：
　LIMS Import System / August 22, 2014
　Quest Diagnostics / MRI-HN-9929331

© COPYRIGHT 2016 MARKLOGIC CORPORATION. ALL RIGHTS RESERVED.

コンプライアンス用のデータリネージ

- **コンプライアンス規制の例**
  - US 行政命令12333号

    概要：インテリジェンスとしての価値がない国民データを削除する義務
  - EU eプライバシー指令

    概要：本人の同意なしでの個人情報の収集・保管を規制
  - バーゼル銀行監督委員会の「実効的なリスクデータ集計とリスク報告に関する諸原則」(BCBS 239)

    ドイツ銀行がMarkLogicで取引・金融商品に対するリスクを管理

  - データスチュワード(データ管理責任者）は、しばしば「システム・オブ・レコード」のみで修正作業を行う
  - ➢ その後の取り消し・修正作業は下流のストレージまでトレースできる必要がある
  - ➢ リネージ情報がないと困難

© COPYRIGHT 2016 MARKLOGIC CORPORATION. ALL RIGHTS RESERVED.

監査

- データにいつ、誰がアクセスしたのか？
- 時間の経過に伴いデータがどのように変化したのか？
- 誰が変更したのか？
- どのようなエラーが発生したのか？　修正されたのか？

© COPYRIGHT 2016 MARKLOGIC CORPORATION. ALL RIGHTS RESERVED.

MarkLogic

監査の事例

米国HMO（健康維持機構）は、患者本人でない家族メンバーにクレーム報告書を誤って送付した。米国HHS（保険福祉省）の市民権局はヘルスプランのコンピュータシステムに問題があり、2000家族の保護されるべき健康情報が危険にさらされたということを把握した。このため保険会社は**6ヶ月間におけるすべてのトランザクションを確認し、誤った患者情報をすべて修正しなくてはならなかった**

- 影響を受けた人は誰？
- 該当トランザクションすべてをレビュー可能？

上記のことを把握するためには監査が鍵になる

© COPYRIGHT 2016 MARKLOGIC CORPORATION. ALL RIGHTS RESERVED.

MarkLogic

監査の事例　HealthCare.govのデータのクリーンアップ

- HealthCare.govは米国の医療保険市場・Eコマースサイト。ユーザ数数百万人単位、2014年から運用開始
- 外部システムから不正確なデータがMarkLogicに流れ込むことによってデータクオリティ上の問題が発生

- データクオリティ関連の修正作業は全てクリーンアップ（修正）タスクとして実装され、監査記録を残す
- クリーンアップタスクの手順
  1) 新旧のドユメントを両方格納し、それぞれのURIを記録
  2) クリーンアップ作業のサポートチケット番号と修正日付を記録
  3) 管理用のツールで上記の2）のデータを取り、標準化された形式で格納
- クリーンアップ情報は、内部利用の参照用ツール上リンクとして公開されている

© COPYRIGHT 2016 MARKLOGIC CORPORATION. ALL RIGHTS RESERVED.

MarkLogic

監査の例 日付付きレコード

- ドキュメントのほうが簡単

  ドキュメント全体に対して日付あるいは期間を追加

- バイテンポラルドキュメント機能も監査に利用可

  = イミュータブルデータモデル （変更不可）

```
{"loan":{
  "loan-id": "96583",
  "date": "2010/02/22",
  "applicant": {
    "name": "Bob Jones",
    "address": "123 Main St"
  },
  "status": "approved",
  "issues": [
    {"credit-risk-issue": "resolved"}
  ]
 }
}
```

MarkLogic

履歴データのトラッキング

- ドキュメントとトリプルのほうが簡単

  ドキュメント全体に対して日付あるいは期間を追加
- バイテンポラルドキュメントを使用

  = イミュータブルデータモデル （変更不可）
- 履歴の利用目的:
  - ユーザに表示
  - ユーザーナビゲーションに便利
  - 監査
  - 履歴
  - 矛盾の解消

```
{"loan":{
  "loan-id": "96583",
  "last-modified": "2010-04-18",
  "modifier": "sallyjones228",
  "status": "history"
  "applicant": {
    "name": "Bob Jones",
    "address": "123 Main St"
  },
  "status": "approved",
  "issues": [
    {"credit-risk-issue": "resolved"}
  ]
 }
}

{"loan":{
  "loan-id": "96583",
    "status": "current"
  "applicant": {
    "name": "Bob Jones",
```

MarkLogic

RDBMSにおける監査

- 監査用にスキーマを新規設計・作成する作業が発生
- オブジェクト/サービス層を追加、データマッピングも必要になる
- ビジネスエンティティ（発注書など）を元の形に戻すために複数のテーブルをJOINしなければならない
- 監査の内容は大きく異なるため
  - アジャイルでも効率的でもない

© COPYRIGHT 2016 MARKLOGIC CORPORATION. ALL RIGHTS RESERVED.

MarkLogic

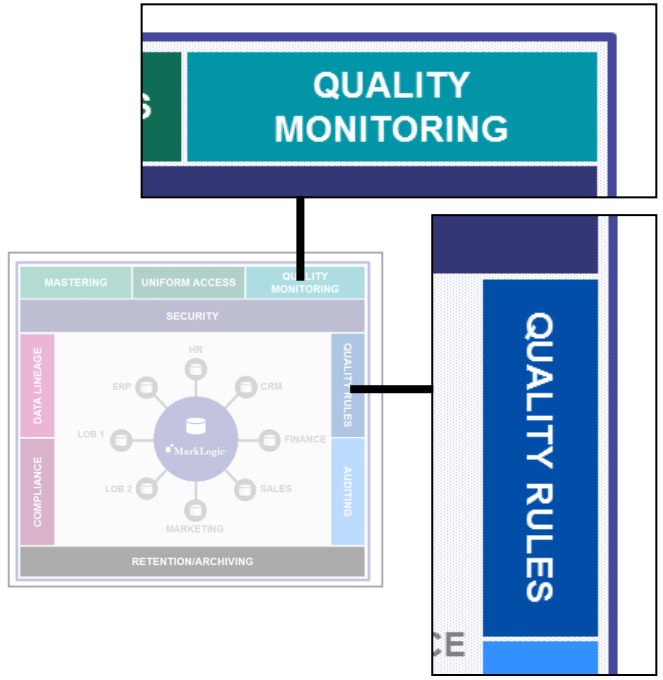
MarkLogicにおける監査の主要用途のまとめ

- データのクリーンアップ(修正）
  - 変更データを即座にログできる – データサイズ・データタイプ・ドキュメント内の位置を問わず
- レコードの監査
  - 「As Is」でデータを書き込めるのは大きなメリット
  - コンプライアンスとベストプラクティスにおいては、監査情報が必須
- MarkLogicのログファイルにも監査情報が記録される
  - データアクセス、セキュリティアクセス、変更などの操作情報を記録

© COPYRIGHT 2016 MARKLOGIC CORPORATION. ALL RIGHTS RESERVED.

データの品質

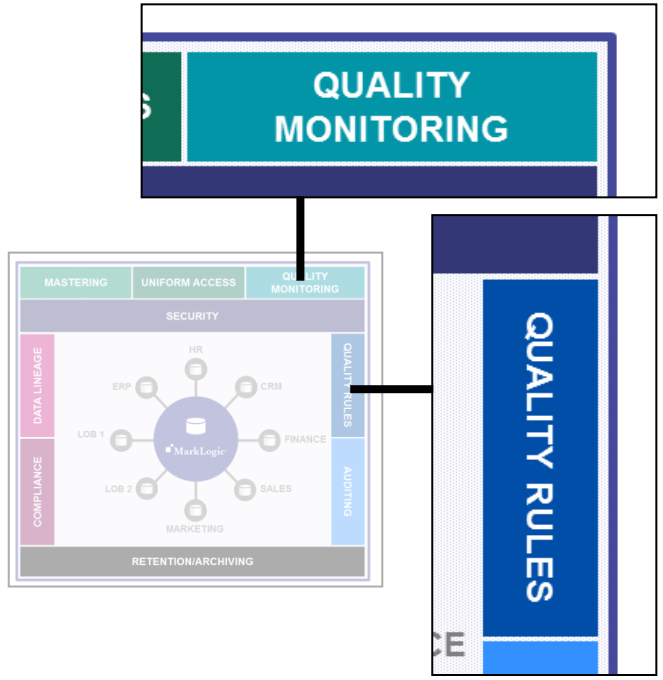
- データ品質のルールによって、問題が重大化する前に悪いデータを把握
  - HealthCare.govのプロジェクトにおいて、ある州の保険機関が親と子のデータを逆にして送付してしまった
    - ルールを作成した:　　親の年齢＞子の年齢
  - 無効な住所データもよくあった
  - 人間が介在し、定期的にデータをサンプリングして、チェックと修正を行う



© COPYRIGHT 2016 MARKLOGIC CORPORATION. ALL RIGHTS RESERVED.

MarkLogicにおけるデータ品質

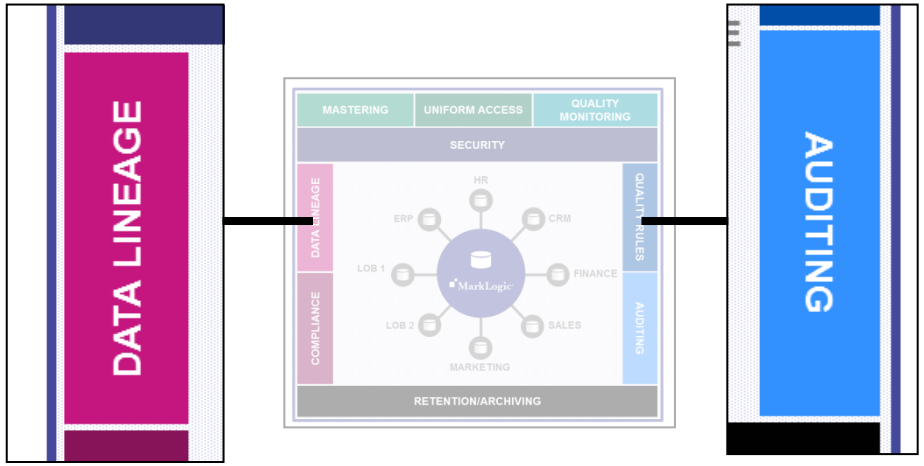
- 統一データによって統一ルールが可能に
  - 必要に応じてマークロジック内で必要に応じデータを標準化
  - 同じフィールドと構造
  - 統一テクノロジー（XML,JSON）
- ルールも一箇所で管理
- 「データハブ」アプローチにより、さらにデータをまとめる
- 優れた監査、履歴、リネージ情報によってさらにデータのチェックが可能
  - より優れたデジタルフォレンジック
- 問題をその源まで遡る



© COPYRIGHT 2016 MARKLOGIC CORPORATION. ALL RIGHTS RESERVED.

トレースとリネージの事例

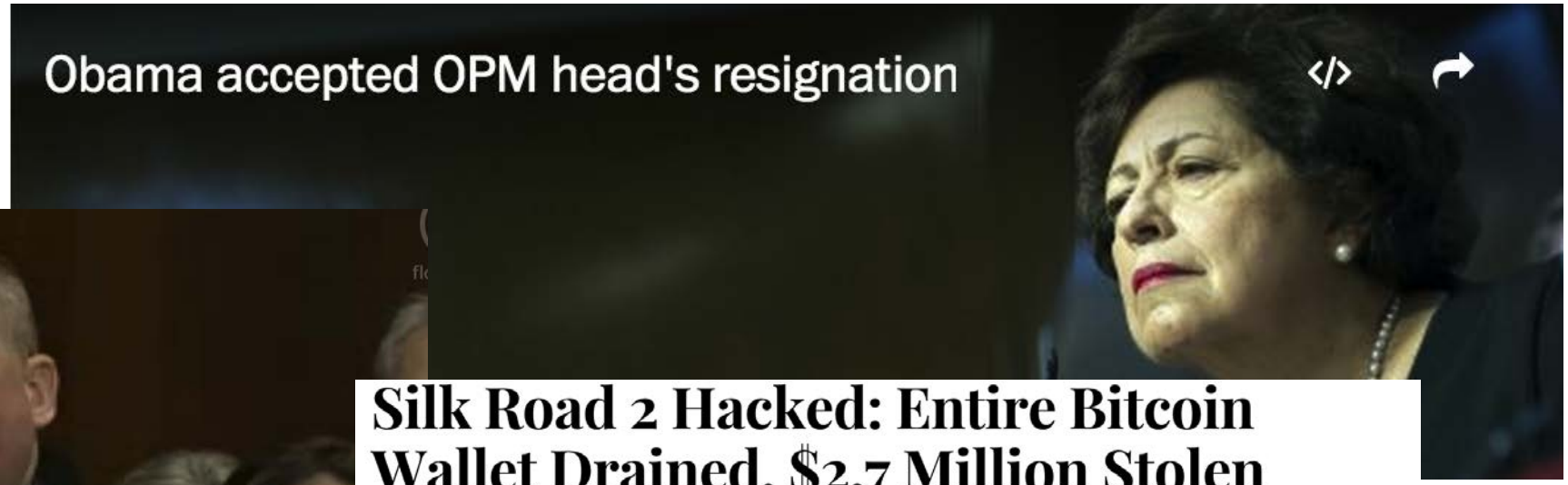
- 重要かつセキュアな情報機関のシステムにおいて、MarkLogicは全データリネージを3段階で管理、複数コピーを保持
  - **レベル0**
    - ソースデータは元の形式で保管
      - テキストメッセージ
      - バイナリのWordドキュメント
      - バイナリ のMySQLダンプ・バックアップ　等々
  - **レベル1**
    - ソースデータの「ロスレス（Loss-less）」バージョン
  - **レベル2**
    - 対象システム準拠形式となる、調整済み/標準化されたバージョンのデータ



© COPYRIGHT 2016 MARKLOGIC CORPORATION. ALL RIGHTS RESERVED.

MarkLogic

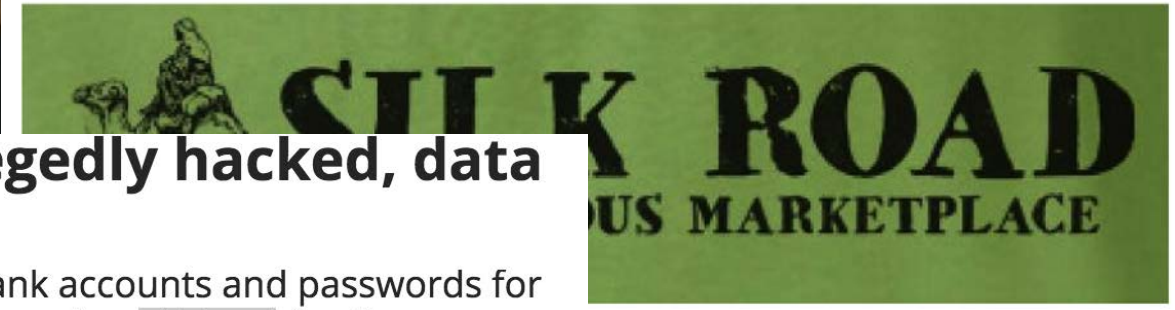
セキュリティ



米大手量販
被害者40C

Silk Road 2 Hacked: Entire Bitcoin Wallet Drained, $2.7 Million Stolen

BY RYAN W. NEAL ON 02/13/14 AT 8:43 PM

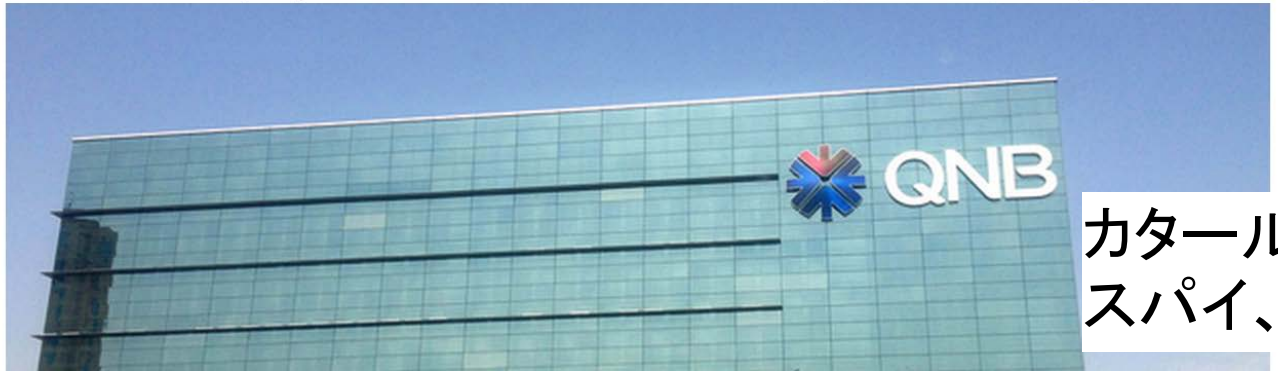


5人

ッキング事件
分のビットコイン盗難

Qatar National Bank allegedly hacked, data of 1,200 entities leaked

The data breach includes the personal bank accounts and passwords for Al Jazeera employees and members of the ruling al-Thani family



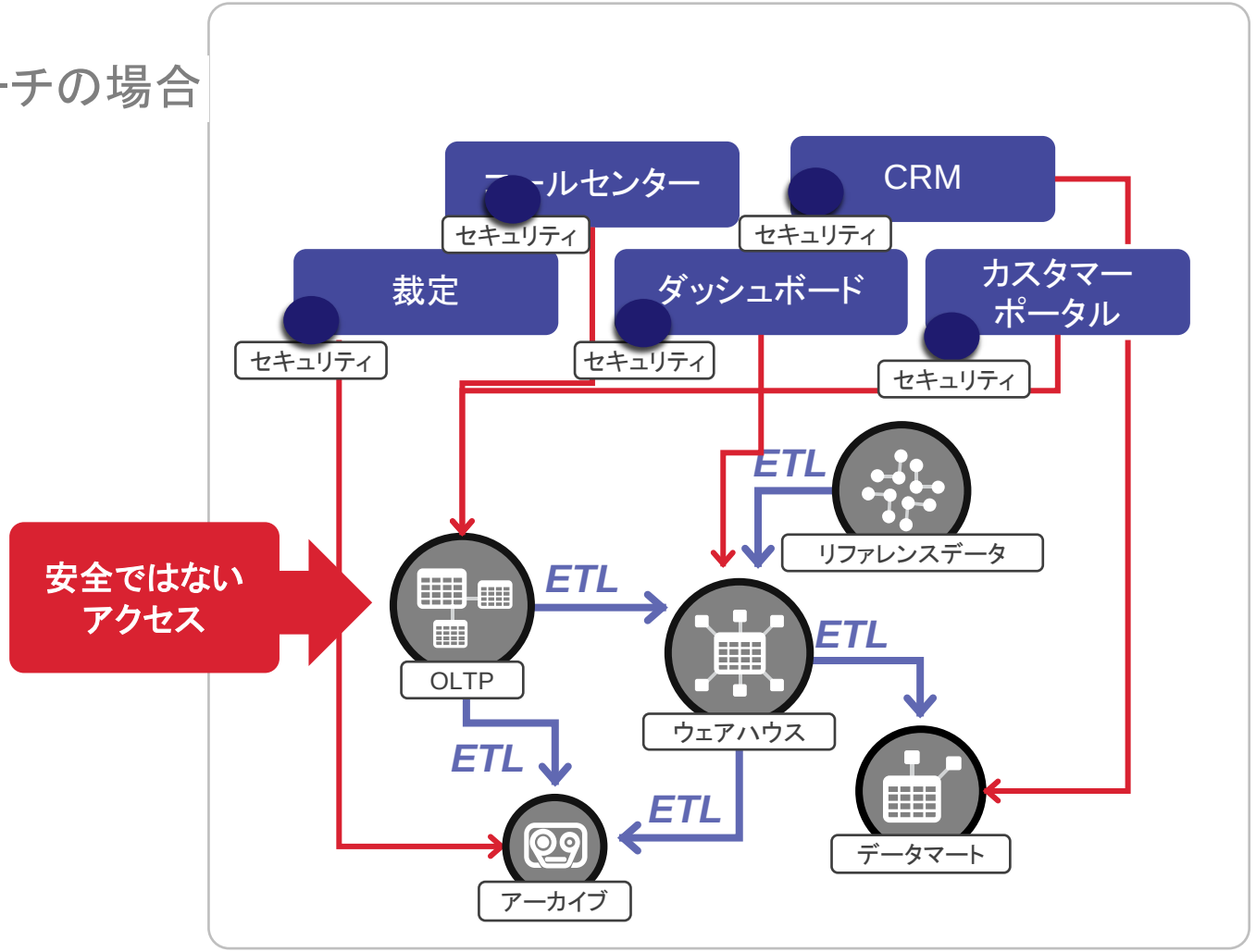
カタール国立銀行がハックされ王族、スパイ、アルジャジーラの情報が流出

COPYRIGHT 2016 MARKLOGIC CORPORATION. ALL RIGHTS RESERVED.

セキュリティ

「アプリケーションごとに繰り返し実装」アプローチの場合

- セキュリティは横断的関心事
- メンテナンスは悪夢
- セキュアで安全なデータアクセスは、特定のアプリケーションからのみ可能
- データ層自体がセキュリティ上のリスクとなる
- 「データレイク」に注意



© COPYRIGHT 2016 MARKLOGIC CORPORATION. ALL RIGHTS RESERVED.

MarkLogic

セキュリティ

- MarkLogicが誇る最高レベルのセキュリティ
  - コモンクライテリア認証済み（NoSQLデータベースで唯一）
  - 米国の複数の情報機関ならびに国防省においてATO（Authority To Operate=運用認定）取得済み
  - FIPS-140認定済み（暗号化）
- ロールベースのアクセス制御
  - 同一ユーザーは複数のロールを持つことが可
  - ロールごとにドキュメント/コレクションのパーミッションを設定
- コンパートメントセキュリティ
  - ロールを複雑なAND/ORで設定可能

© COPYRIGHT 2016 MARKLOGIC CORPORATION. ALL RIGHTS RESERVED.

MarkLogic

ベストプラクティス – データの照合

- すべてのデータが移動されていることがデータ品質の主要KPI
- 完全に信頼できるデータ移動は存在しない
  - バグが全くないソフトウェアが存在しないのと同じ
- データセットが同期していることを定期的にチェック
  - 典型的な4つの手法
    - 合計数の定期チェック
    - マニフェストに基づくバッチ件数のチェック
    - アイテムごとの個数の定期チェック
    - より詳細なチェック:アイテム＋ハッシュあるいはアイテム＋日付　でチェック

© COPYRIGHT 2016 MARKLOGIC CORPORATION. ALL RIGHTS RESERVED.

データガバナンスに関する
MARKLOGICのその他の機能

MarkLogic

MarkLogicの機能

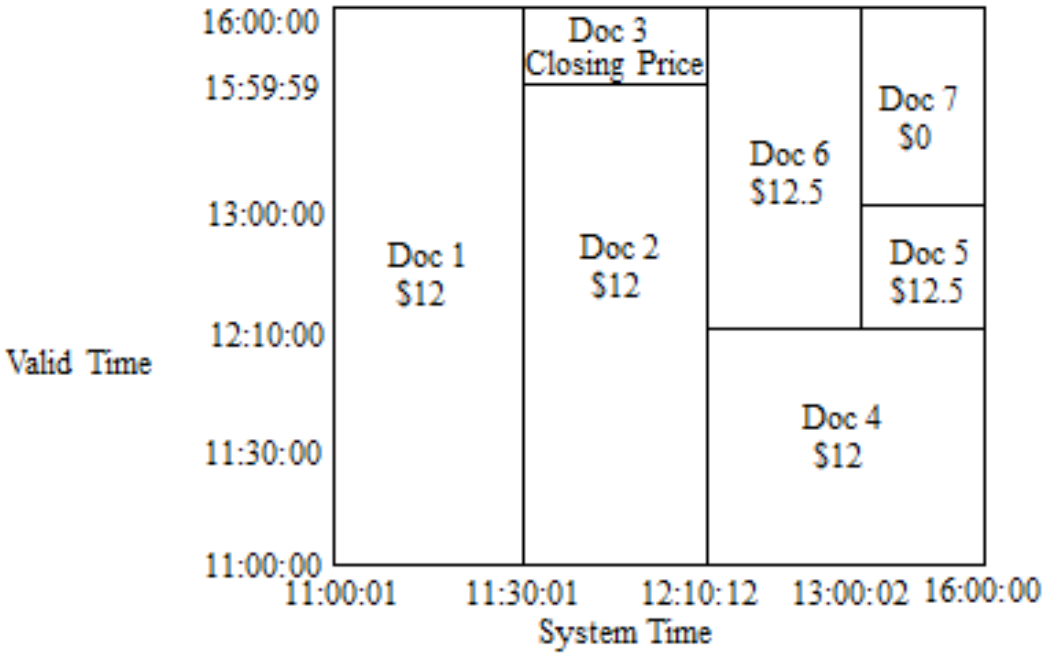
- ここまで取り上げたのは、
  - 統合されたガバナンス機能
  - セキュリティ
  - 「As Is」で格納 - 監査、生データ、リネージ用に
  - 日付付き/履歴データ用のドキュメントモデル
- 他に考慮すべき機能は…

© COPYRIGHT 2016 MARKLOGIC CORPORATION.ALL RIGHTS RESERVED.

MarkLogic

バイテンポラルドキュメント

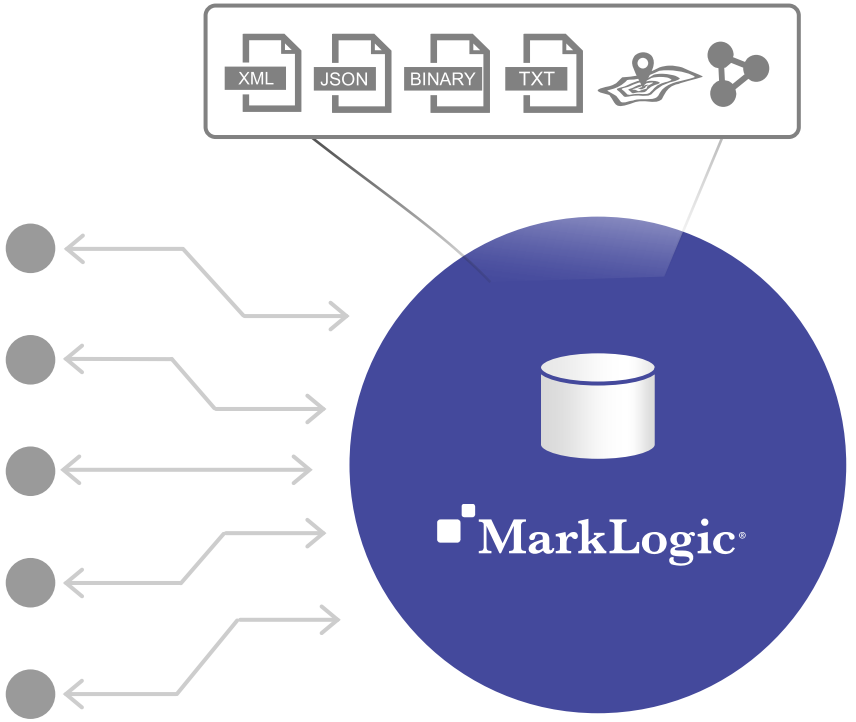
- 「システム時間」と「ビジネス時間」の両方に基づいてデータをトラッキングする
- あらゆるタイムスタンプに関して、包括的な時間像を再構築する
- 履歴データや監査に利用



© COPYRIGHT 2016 MARKLOGIC CORPORATION. ALL RIGHTS RESERVED.

包括的データ形式に対応

- 構造化データ – XML、JSON
  - あるいは何らかのリレーショナルデータも可
- テキスト、非構造化データ
- バイナリデータ
- セマンティックデータ
- あらゆるデータを1つのシステムでガバナンス可



© COPYRIGHT 2016 MARKLOGIC CORPORATION. ALL RIGHTS RESERVED.

MarkLogic

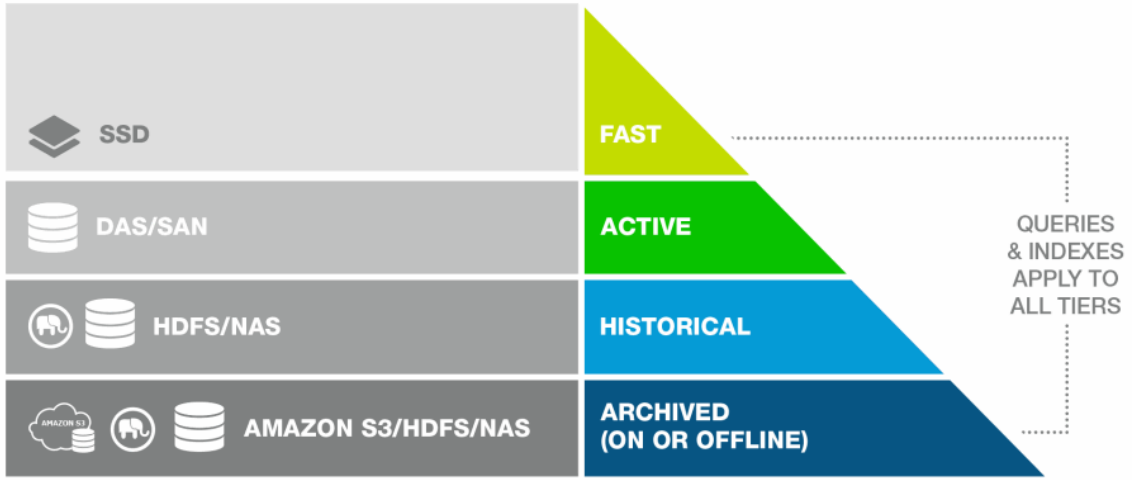
階層型ストレージ

- ストレージポリシーを定義によって各ドキュメント内の値でドキュメントを自動的にストレージ階層に振り分ける

```
{"loan":{
  "loan-id": "96583",
  "date": "2010/02/22",
  "applicant": {
    "name": "Bob Jones",
    "address": "123 Main St"
  },
  "status": "approved",
  "issues": [
    {"credit-risk-issue":
"resolved"}
  ]
 }
}
```

ストレージポリーシの例：

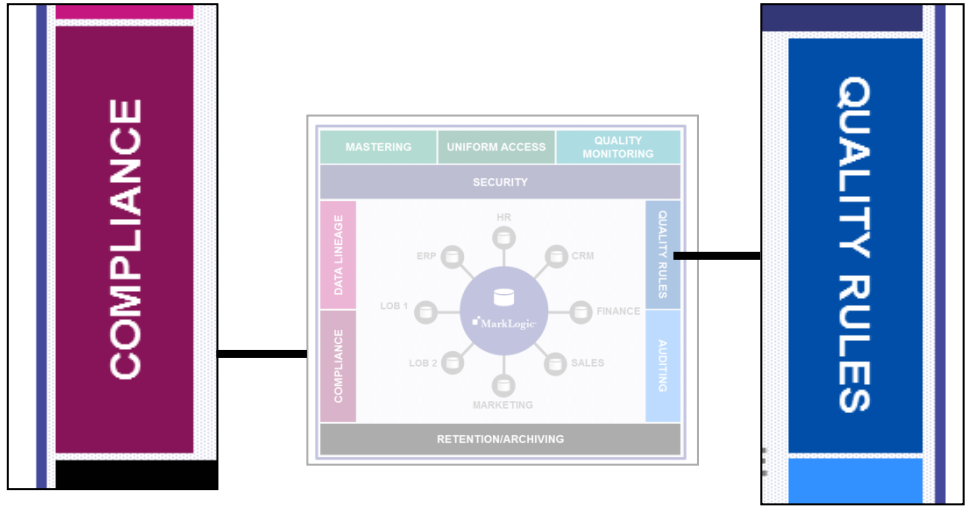
- 一年以内の新規データはSSDに保管
- １年以上５年以内のデータはNASに保管
- 5年以上のデータはS3やHDFSに保管
- ストレージのコストダウンも図る



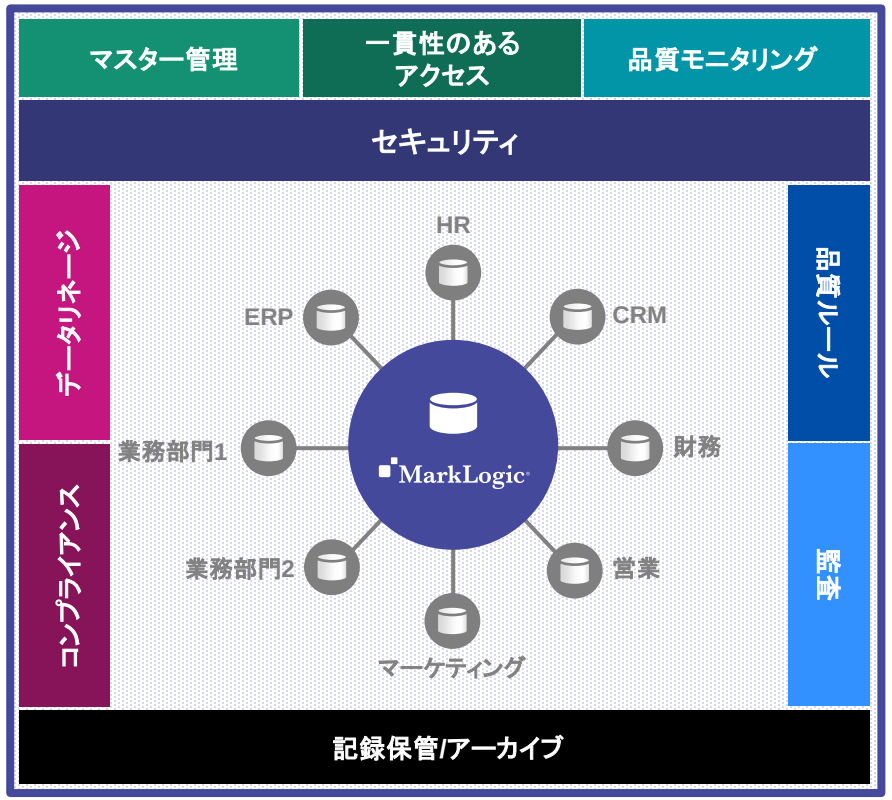
© COPYRIGHT 2016 MARKLOGIC CORPORATION. ALL RIGHTS RESERVED.

セマンティック

- コンプライアンスは複雑で、オントロジーによって表現されることで管理がより簡単に
  - 金融関連オントロジー FIBO
- 通常、リンクと関係付けが必要とされる
  - セマンティックはスキーマ非依存
  - RDF三つ組（トリプル）で定義
- セマンティックの更新は軽く速い
  - 既存レコードの「注釈付け」に理想的



© COPYRIGHT 2016 MARKLOGIC CORPORATION. ALL RIGHTS RESERVED.



まとめ

- 将来のニーズに対応出来るのはデータハブ
- 我々は常に「ガバナンス」を行っている
- 統合されたデータでのガバナンスの方が簡単
- MarkLogicはデータガバナンス用に設計されている

メリット

- データハブ、データガバナンスハブを利用
- データとその履歴、リネージ、監査情報を一元管理

© COPYRIGHT 2016 MARKLOGIC CORPORATION. ALL RIGHTS RESERVED.

Q&A