「ビックデータでここが変わる、ここが変わらない」
– いま、データアーキテクトにとって知っておかなくては ならないデータ活用のポイント–

2013年2月27日

株式会社データアーキテクト
真野　正

概要

- 「ビッグデータ」はもはやバズワードではなく、企業内情報システムでも日常的に語られるようになってきました。しかし、現実に取り組んで成功している企業は、まだまだ少ないようです。
- 活用が進んでいない原因を探ると、実は企業内情報システムに問題があるのではないか。企業内情報システムとビッグデータ処理系システムとはいかなる関係にあり、企業内情報システムのデータベースとして備えておくべきことは何か。
- さらに、ビッグデータ以前に、企業内情報システムでのデータ管理はうまくいっているのか、データ活用を阻害している要因をデータモデリング観点を中心に探っていきます。
- ビッグデータ時代にデータアーキテクトとして知っておくべきことは何か。
- そして、ビッグデータを活用するための新たなデータアーキテクチャとデータガバナンスの必要性を提言します。

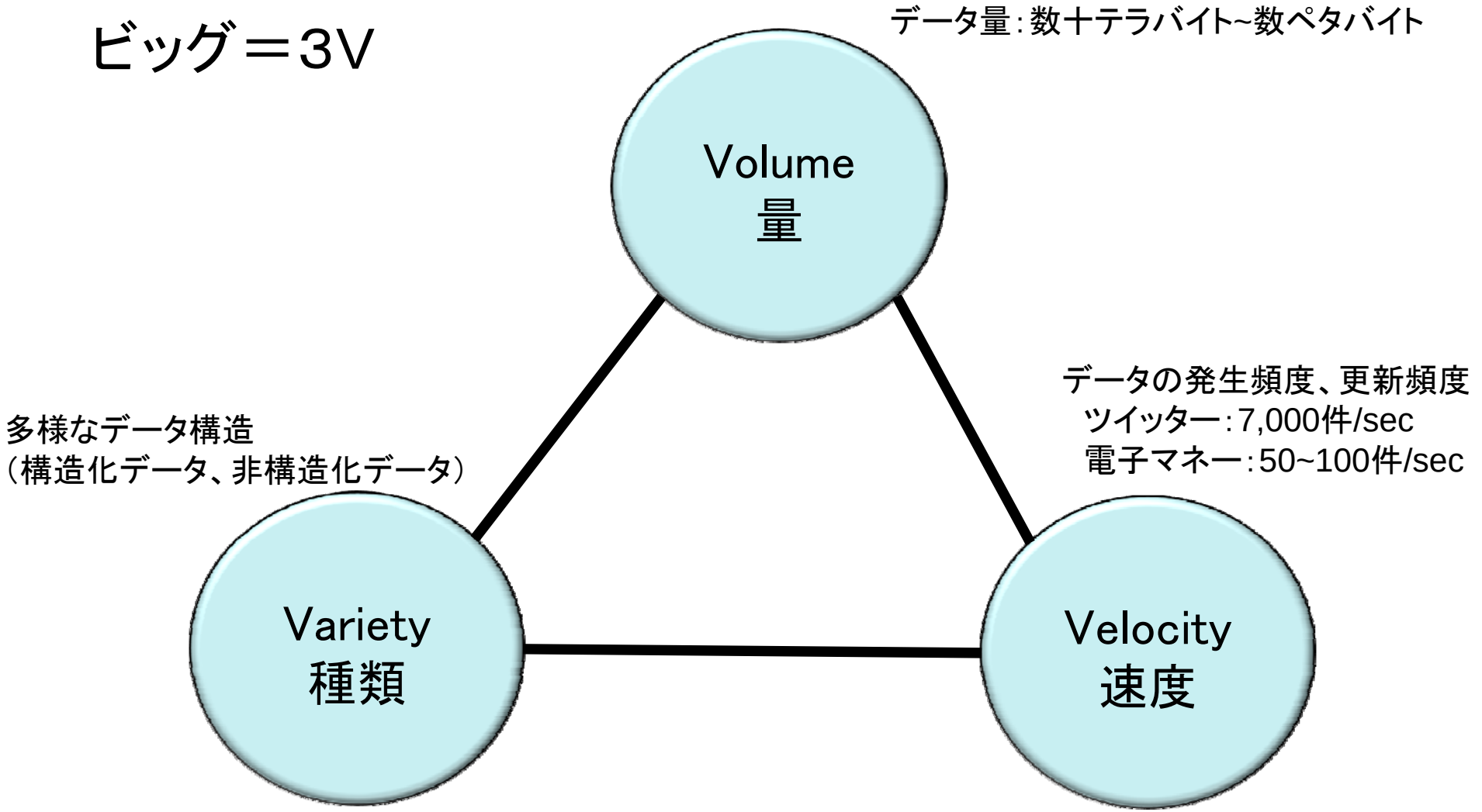
Contents.

1. ビッグデータとは何か

- 3V（量、種類、速度）
- 構造化データvs非構造化データ
- 活用事例：ログ解析、BI、基幹バッチ高速化
- ビッグデータ関連ビジネス
- DaaS

ビッグデータとは何か

ビッグ＝3V



ビッグデータとは

- Webのアクセスログやストリーミングデータのようなテキストデータで構造化できないデータ：非構造化データ
- 従来からRDBに格納していた売上データなどをより粒度を細かく、多期間にわたっての分析要求からデータ量が増えてきている
- ツイッターに見られるように、データの発生頻度が肥大化している

ビッグデータ関連ビジネス例①ーデータ流通

- 富士通は企業が保有する大量の電子情報、いわゆる「ビッグデータ」の取引市場を今春開設する。インターネット通販での購入履歴、タクシーに載せたセンサーから得られる渋滞情報などを解析し、商品開発や新ビジネスの創出に生かそうと考える企業が増えている。富士通は2016年までに参加企業を千社程度まで増やし、取引仲介サービスを主力事業の一つに育てる。（解説企業2面に）
- ビッグデータの取引市場は、この分野の先進国である米国を含め、珍しい。一方、データの元となる個人情報について、プライバシーの保護に抵触しない形で活用するための制度づくりも急ぐ必要がありそうだ。
- 富士通が開設する取引市場は「データプラザ」。リストから欲しいデータを選び、ダウンロードする。データはすべて個人情報を匿名化した上で取引してもらう。価格はデータ量や中身により異なるが数万～数千万円。このほか月数万円の会費がかかる。データプラザで売買できるのはこのほかスマートフォン（スマホ）の位置情報、交流サイト（SNS）への投稿など。流通業や製造業などと現在、参加に向けて交渉している。
- 日本マーケティング・リサーチ協会によると、企業によるデータ売買関連の国内市場は年間約2200億円。これまでは市場動向などを専門調査会社から購入することが多かった。

＜日経新聞　2013年1月30日朝刊＞

ビッグデータ関連ビジネス例②

- 日立製作所と博報堂は「ビッグデータ」と呼ぶ膨大な情報の解析事業で提携し、販売促進策などの助言まで一貫して請け負うサービスを４月に始める。日立が技術やシステムを提供してデータを解析。博報堂が独自データを加味し、それぞれの商品に合わせたマーケティング手法を提案する。顧客企業が求めるサービスを低コストでまとめて提供することで、ビッグデータの活用に消極的だった中堅・中小企業の需要を取り込む。
- ビッグデータは朝刊数十万年分に相当するような膨大なデータのこと。IT（情報技術）の進化により、ネット上で個人が発する膨大な情報を解析することが可能になり、販売促進や商品企画などに生かそうとする動きが広がっている。
- 両社は新サービスを始める受け皿組織として「マーケット・インテリジェンス・ラボ」を発足させ、それぞれ10人程度の社員を派遣。解析に使う専用ソフトウエアの開発などを進め、消費者行動の予測精度を高める。
- 新サービスではどの時間帯に、どの店舗で、どんな属性の人が何を買ったかといった基本データを顧客企業が日立に提供。日立はこれとは別に交流サイト（SNS）に消費者が書き込んだ情報から、特定の商品に対する購買意欲や満足度などをきめ細かく分析する。
- これらの情報を付き合わせることで、購買意欲があるにもかかわらず売り上げが伸び悩んでいる地域や商品を特定。博報堂は独自に蓄積してきた販売促進ノウハウを加味し、顧客企業に商品開発やマーケティング戦略の見直しを提案する。
- ビッグデータの活用は大手メーカーなどで広がりつつあるが、中堅・中小企業は解析結果を生かすノウハウが乏しく、利用しにくい面があった。日立と博報堂は中堅・中小企業が負担できる水準に利用料金を抑える方針で早期に詰める。３年後をめどに300億円程度の売上高を目指す。
- 調査会社などによるとビッグデータ関連の国内市場は2020年度に１兆円規模と、11年度の５倍以上になる見通し。

＜日経新聞　2013年2月13日朝刊＞

ビッグデータ関連ビジネス例③

- ローソンやサンリオなど日本の流通・サービス企業が投資ファンドを立ち上げ、米シリコンバレーのIT（情報技術）企業に出資する。ファンドの規模は総額約230億円をめざす。日本企業が持つ膨大な販売情報や顧客情報といった「ビッグデータ」を米企業の最先端のデータ解析技術と結びつけ、効果的な商品開発や販売促進につなげる。日本の産業界でビッグデータを活用して競争力を高める取り組みが本格化する。
- すでにローソン、サンリオ、リクルートグループなど4社が出資。出資額は5000万ドル（約47億円）前後とみられる。2013年度中に衣料品チェーン、食品メーカーなども含め出資企業を15社程度に増やし、総額2億5000万ドル（約230億円）規模に拡大する。
- ローソンは会員数約5000万人のポイントカードを通じて個人の購入履歴も含めた詳細な販売データを収集している。米企業の先端技術で新商品の購入頻度を解析し「ヒットの芽」をいち早く見いだすなど的確な販売計画につなげる。サンリオは最新の交流サイト（SNS）を使った自社キャラクターのファン層の拡大、リクルートは企業と消費者を結びつけるマッチング事業に新たな技術を取り入れる。
- ビッグデータ活用で先行するのは橋やトンネルなど社会基盤の保守だ。NTTデータは昨年開通した東京ゲートブリッジに大量のセンサーを設置。圧力や振動などのデータを絶えず集めて異常の早期発見に役立てる。
- 一方で販促などの分野への応用は始まったばかり。NTTドコモが東急百貨店などと組み、スマートフォンの位置情報や来店履歴を基にクーポンを配信するなどのサービスを始めたが一部にとどまっているのが現状だ。

＜日経新聞　2013年2月23日朝刊＞

2. ビッグデータの背景技術

- CAP定理
- NoSQL
- DBMSバリエーション
  - リレーショナル、Key-Value、カラム指向、ドキュメント指向
- Hadoop
  - 大量データ並列処理
- 統計解析
  - 統計処理技術者が人気

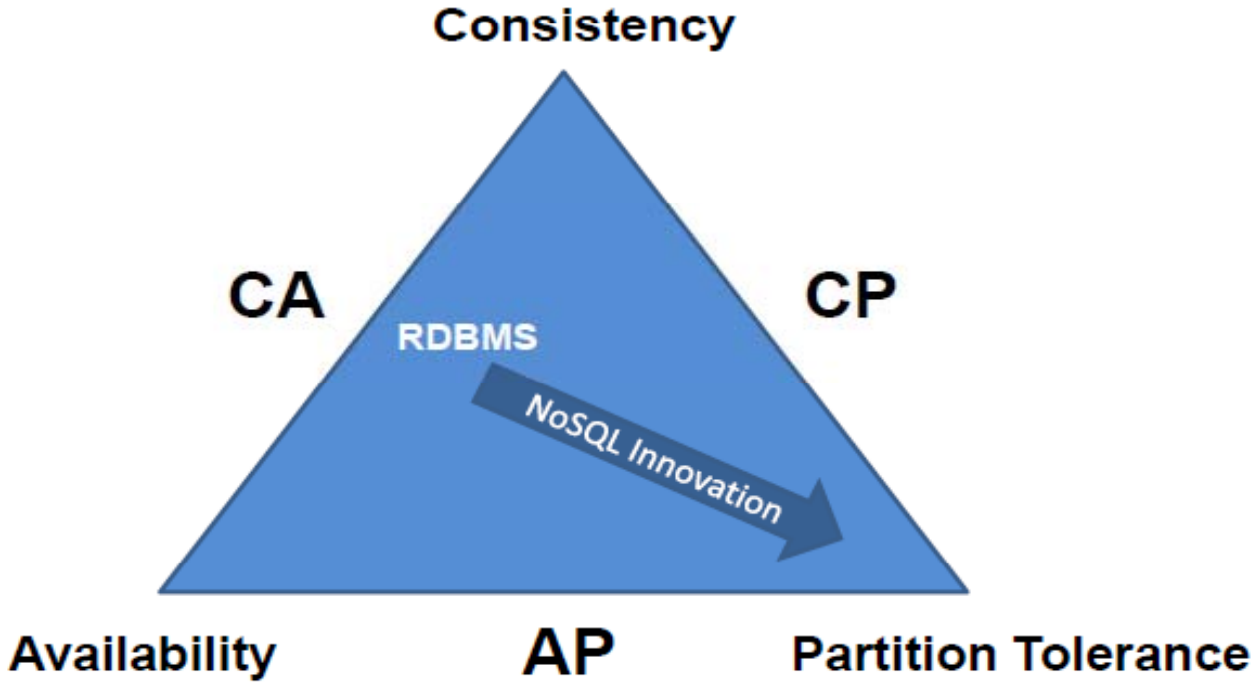
CAP定理

- 分散コンピューティング環境においてCAPを同時に保証することは困難

- Consistency：整合性
- Availability：可用性
- Partition：分断耐性

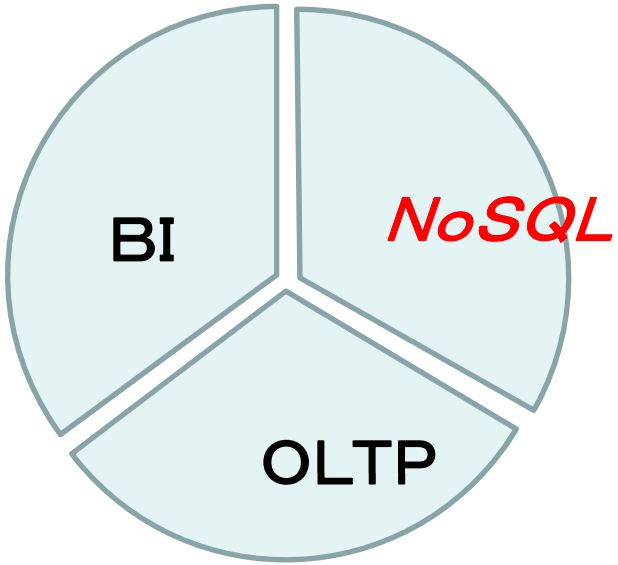
⇔

- Atomicity：原子性
- Consistency：整合性
- Isolation：独立性
- Durability：永続性

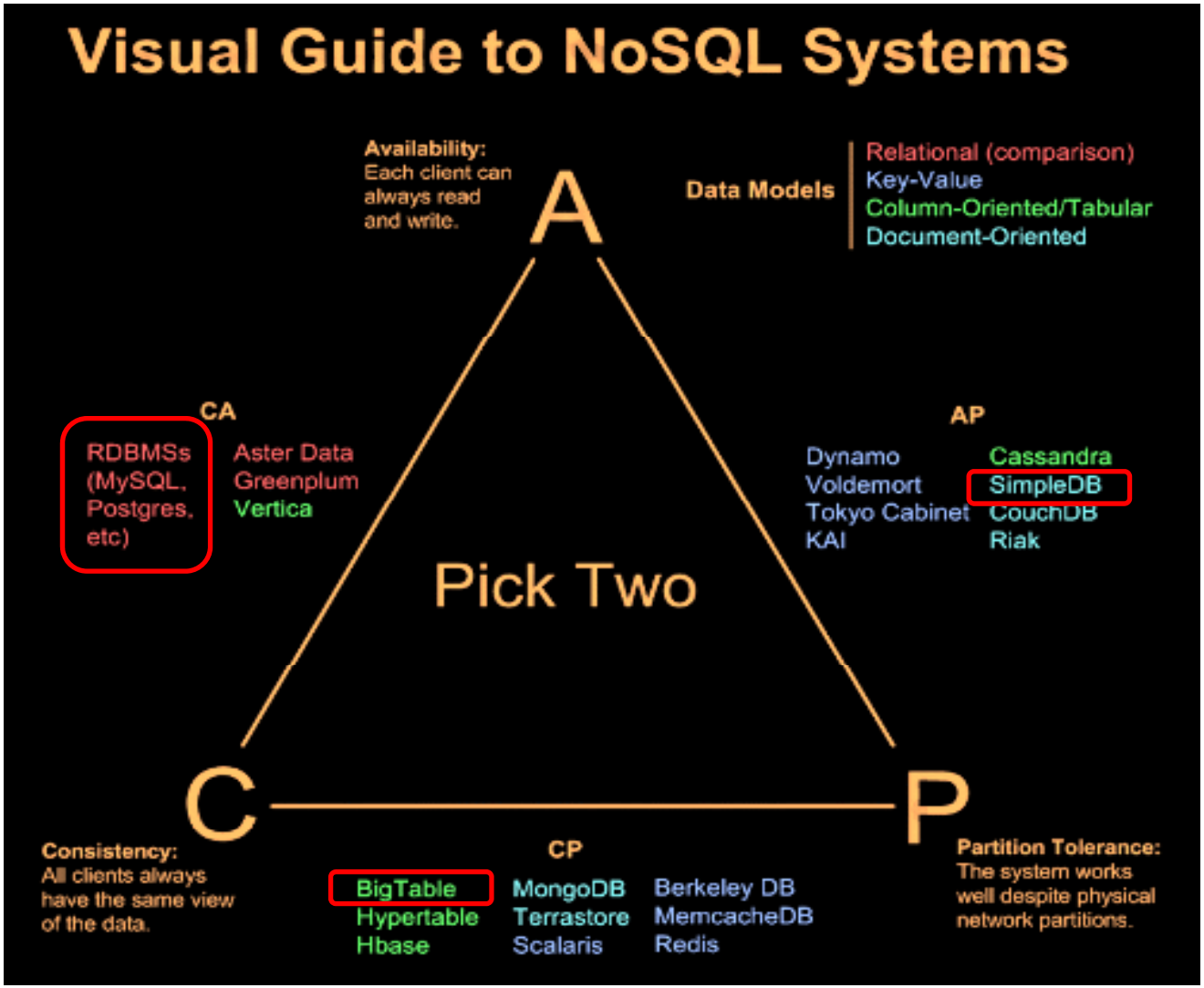


NoSQL

- NoSQL=NO!　SQLに非ず、Not Only SQL
  - SQL(RDB)を補完する
- スケールアウト可能なDB
- データモデル
  - リレーショナル
    - RDB
  - Key-Value
    - スキーマレス
    - プログラムでの構造定義で変化対応
  - カラム指向
    - Sybase-IQ
  - ドキュメント指向
    - MongoDB
      - JSON形式でのドキュメント格納



データモデル（物理構造）



データモデル分類

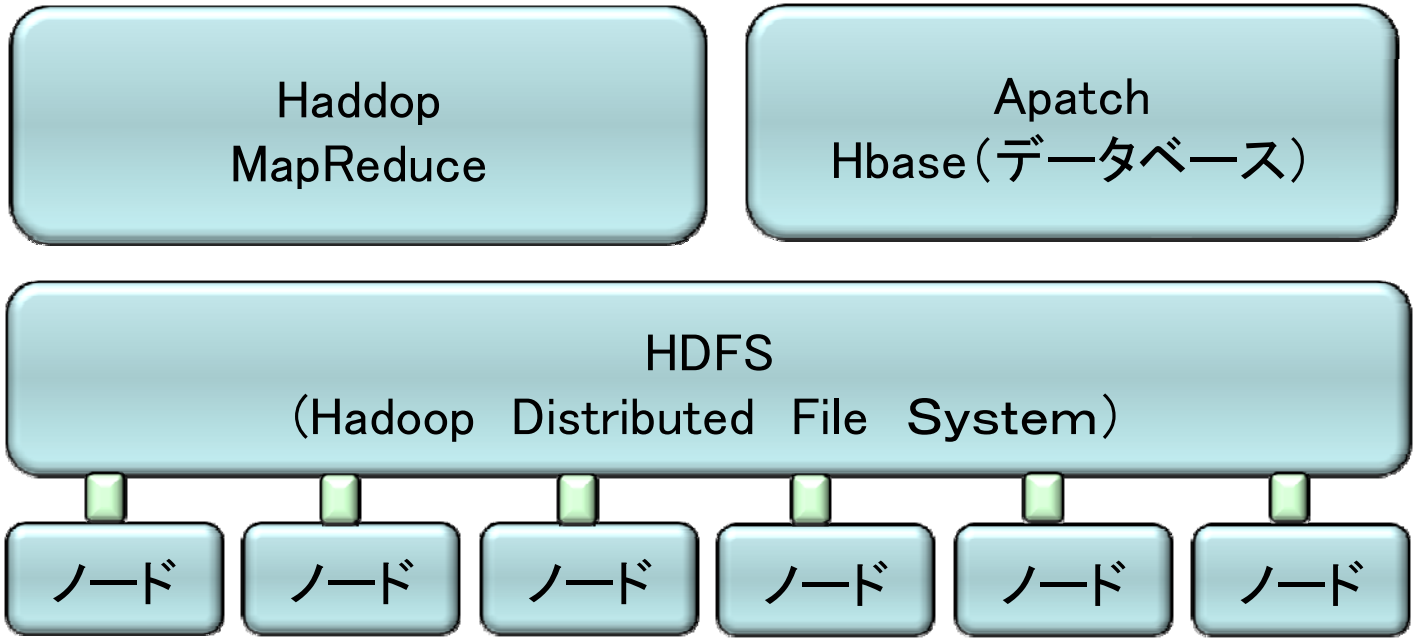
- Relational
- Key-Value
- Column-Oriented
- Document-Oriented

※リレーショナルデータベースの多くは、複数台で構成されているときは、ネットワーク分断耐性が犠牲になることが多い。Amazon SimpleDBなどは、データの一貫性を犠牲にしていて、データベースに書き込まれても、それを読めるようになるには1秒以下の時間がかかる。Google File System上のBigTableは可用性が保証されていない。

出典：Nathan Hurst's Blog

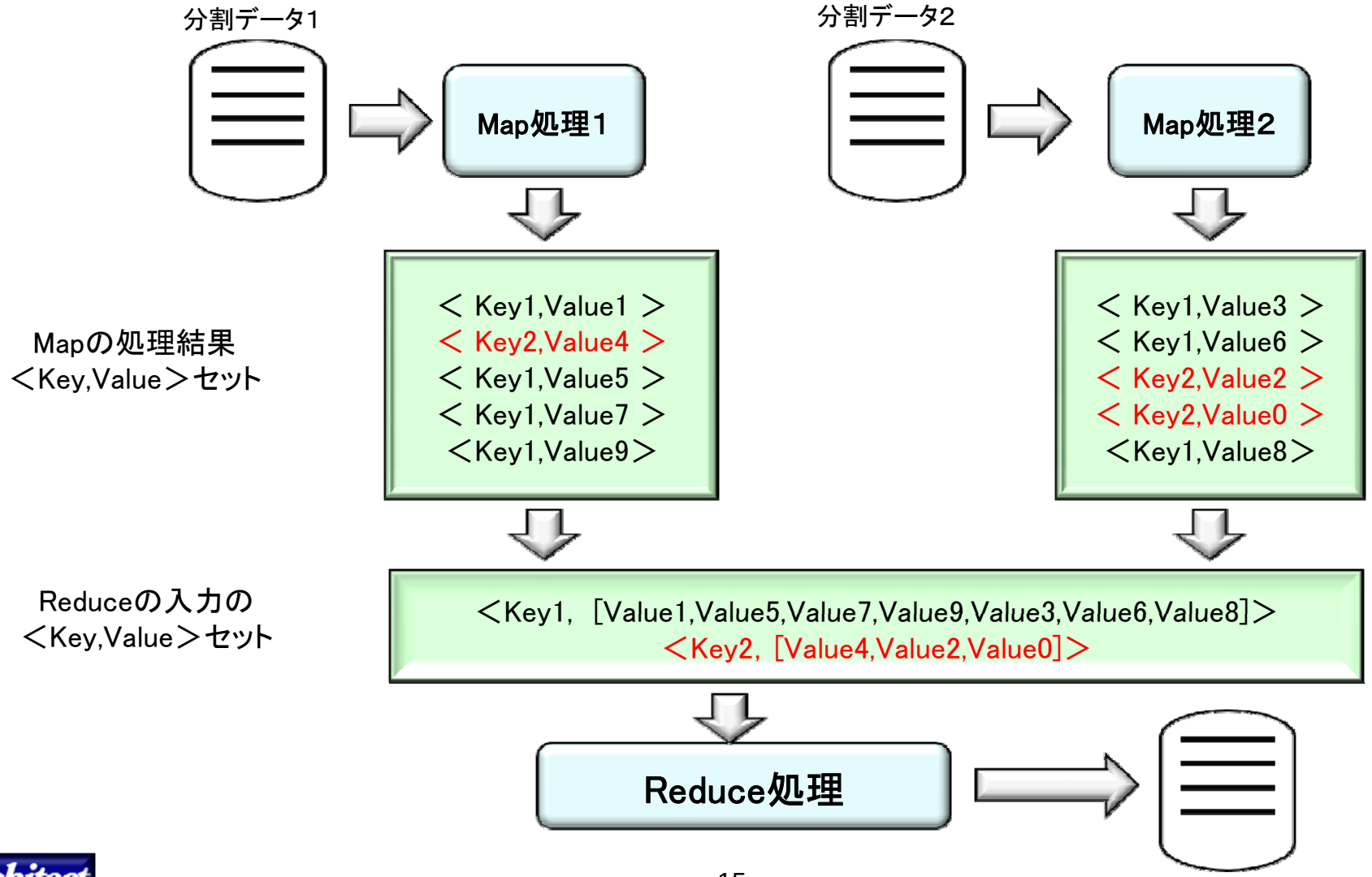
Hadoop

- オープンソース
- データの並列処理による時間短縮
- 並列化可能な大量データの処理に向いている
- 基幹業務での長時間バッチ処理にも適用可能
- 更新処理には向かない



MapReduceに向いているデータ

- Hadoopが扱うデータは、キーが重複しているデータ
- パラレル処理で同一キーデータを集約（Map）してReduceする



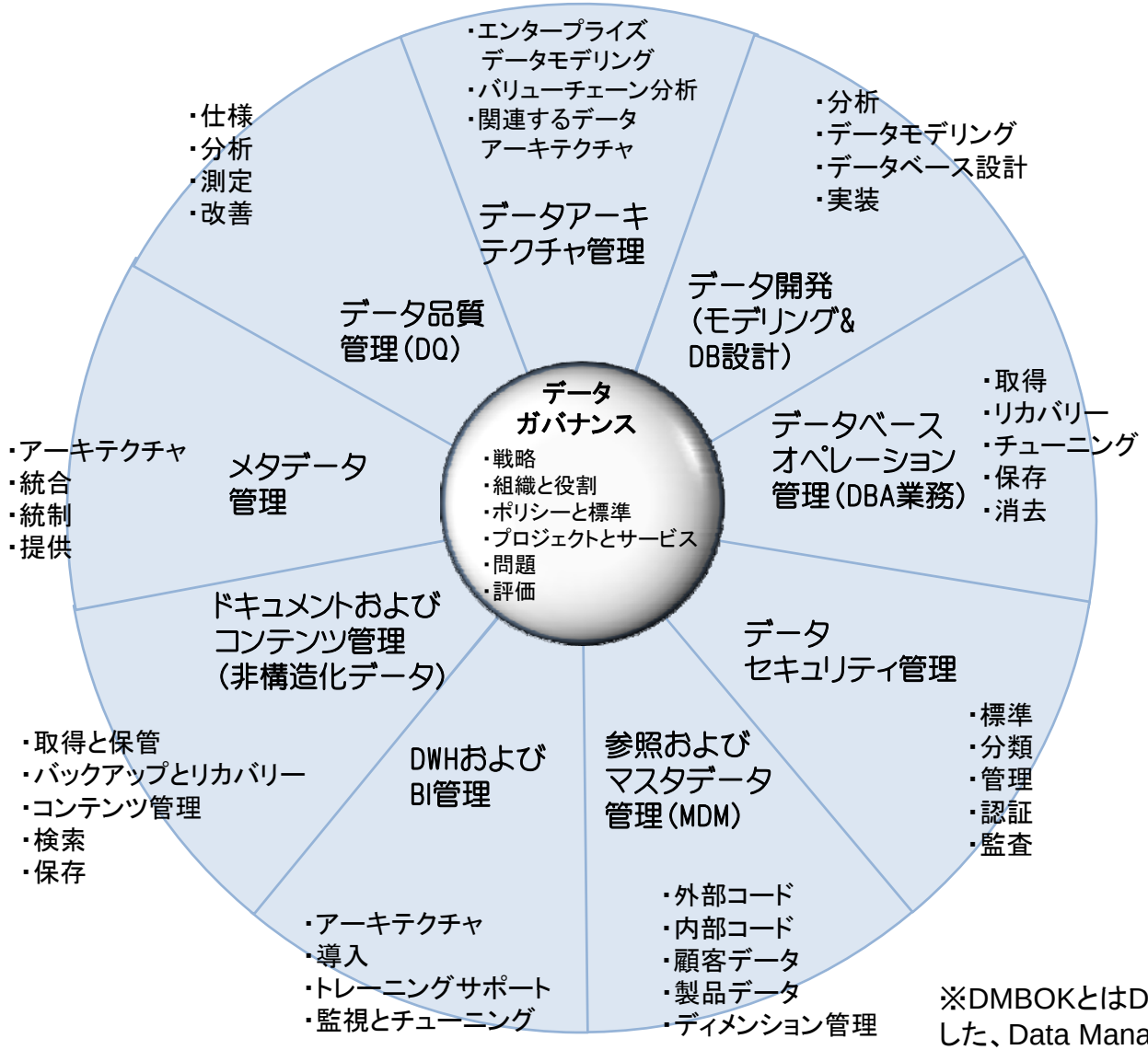
NOSQL、Hadoopプロダクツ例

- Hadoop/HBase
- MongoDB：Opensource、Json、documentDB
- MarkLogic：XML-DB
- CouchDB：Apach、Jason
- Cassndra：Appach、FaceBook
- Riak：Json、ErLang（並列処理指向プログラミング言語・実行環境）
- Hypertable：Google BigTable
- Memcache
- VOLT

３．ビッグデータとエンタープライズシステムとの関わり

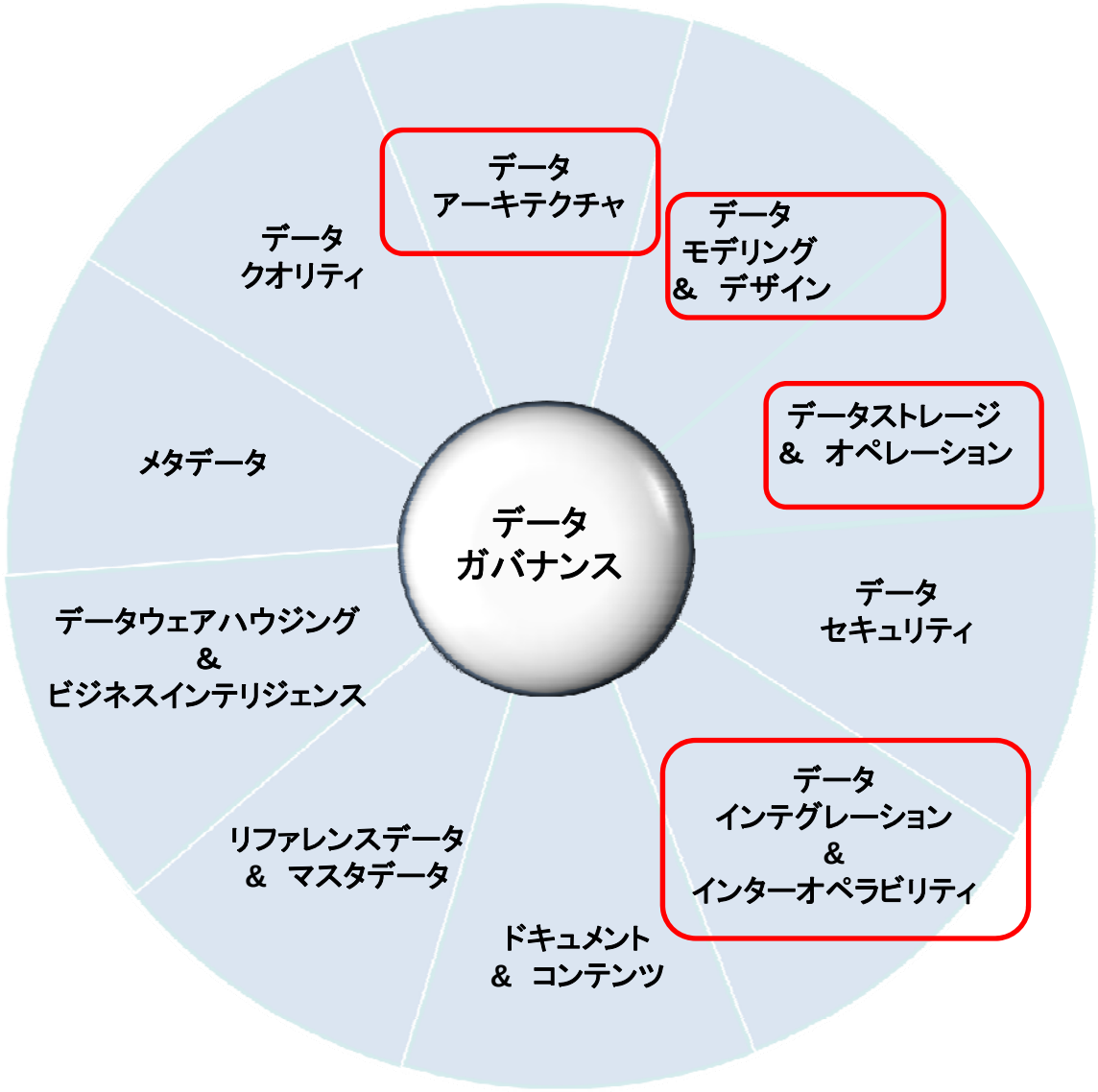
- DAMA-DMBOKに見るビッグデータの取組み
- ビッグデータ活用上の課題
- 非構造化データと構造化データの接点
- 量の増大：企業内でのトランザクション量増大（構造化データ）
- 種類の増大：
  - ◆ SNSなどWeb系データの取込み（非構造化）
  - ◆ 外部データの購入

DMBOKでのデータマネジメント10機能



※DMBOKとはDataManagementAssosiationが策定した、Data ManagementBody of Knowledge

改訂中のDMBOK機能（１１）



DMBOK-2から

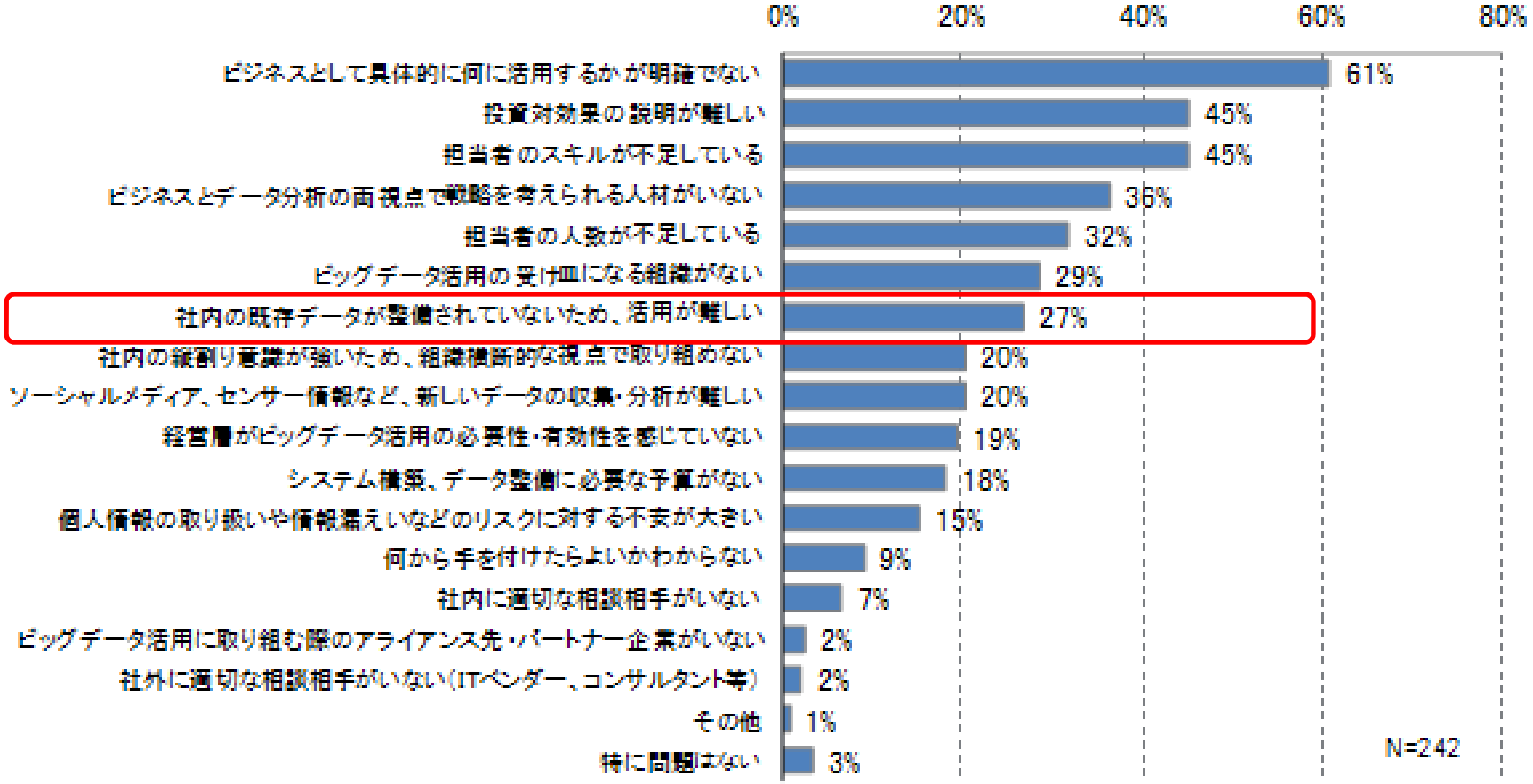
- DMBOKの改訂版では、ビッグデータに関する取込みが大幅に行われている
- データアーキテクチャ
  - エンタープライズ・データアーキテクチャ
  - データアーキテクチャ実装
  - ビッグデータアーキテクチャ
- データモデリング＆デザイン
  - モデリングテクニック
  - 概念・論理モデリング
  - 物理モデリング
  - データモデリングとデザインのガバナンス
- データストレージ＆オペレーション
  - DBMSアプリケーション
  - ファイルストレージシステム
    - Hadoop、NoSQL
  - メンテナンス
  - ガバナンス

DMBOK-2から

- データインテグレーションとインターオペラビリティ
  - アプローチ：統合・相互運用
  - データ獲得（外部データ取込み）
  - データ移動／サービス：構造化データ／非構造化データ
  - データインターオペラビリティ
  - ガバナンス
- リファレンスデータ＆マスターデータ管理
  - 共通アクティビティ
  - リファレンスデータ
  - マスターデータ

ビッグデータ活用上の課題

図5:「今後、貴社でビッグデータの活用を進めていく場合、どのようなことが問題・課題となりそうですか」（複数回答）



出典：～ビッグデータの利活用に関する企業アンケート結果～
2012年12月25日
株式会社野村総合研究所

データ活用への高度な要求

データにはより高度な要求が突きつけられている

	従来	ビッグデータ時代
データ鮮度	月次・日次	瞬時・数時間おき
データ取得間隔	数時間～1日	秒・分
データ粒度	集約データ	明細データ
データ範囲	部門・事業単位	グループ企業単位
データ精度	低	精緻
データ種類	構造化データ（RDB）	テキスト、画像データ

短い間隔でより多くのデータの取込みが必要

リアルタイムに明細レベルのより精緻なデータを提供

非構造化データと構造化データの接点

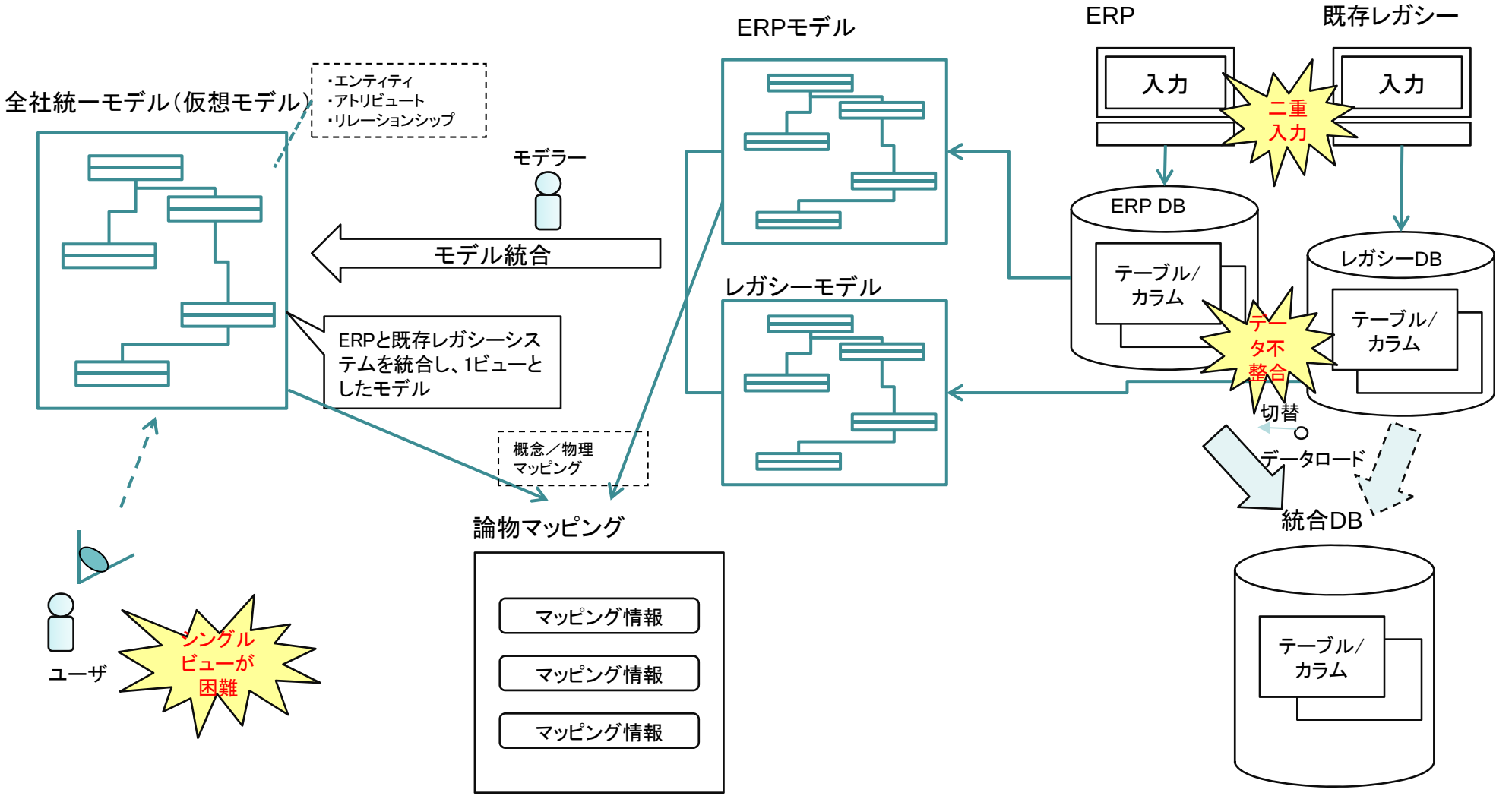
- マスターで関連づけられる
  - ◆商品、顧客
- Key-ValueのKeyの一部
- 分析対象属性の抽出
  - ◆構造化すべきデータ属性だけを抽出する

４．エンタープライズ系DBシステムの抱える課題

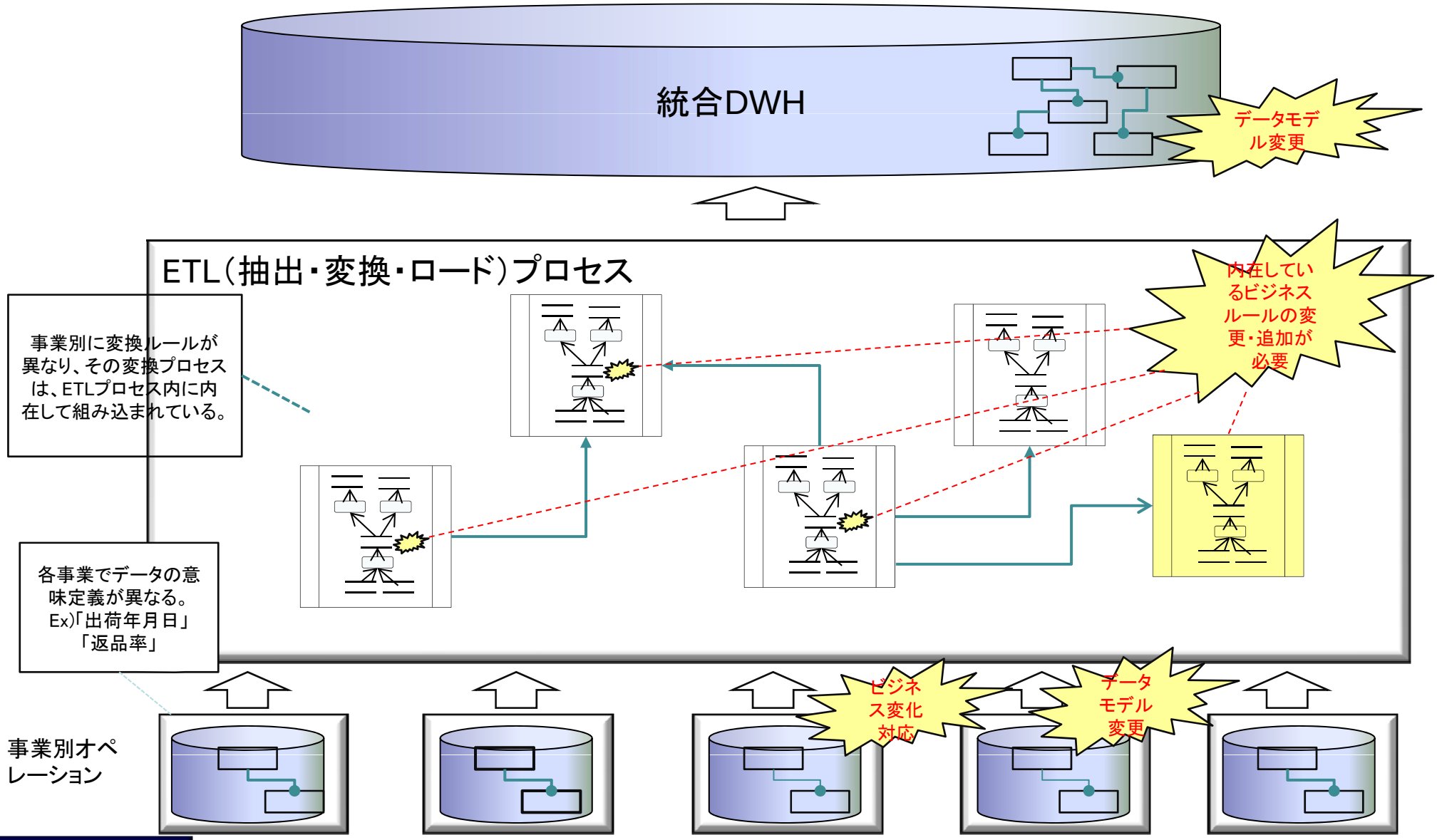
- コンプレックス・システムによる二重入力、データ不整合
- データの源泉、加工ルールが見えないため欲しいデータが容易に取得できない
- 基幹系、情報系システムでのデータ遮断
- マスターデータ管理
  - ◆古典的な命題ではあるが
  - ◆グローバル化、ビジネススピードの変化に追随していくために必要に迫られている
- データ品質
- データ管理要員不在

コンプレックス・システムによる二重入力、データ不整合

- 二重入力が発生➔結果としてデータ不整合の温床となる
- 既存レガシーシステムとERPシステム（新システム）の併存するデータを1ビューで見ることができない。



データの源泉、加工ルールが見えない



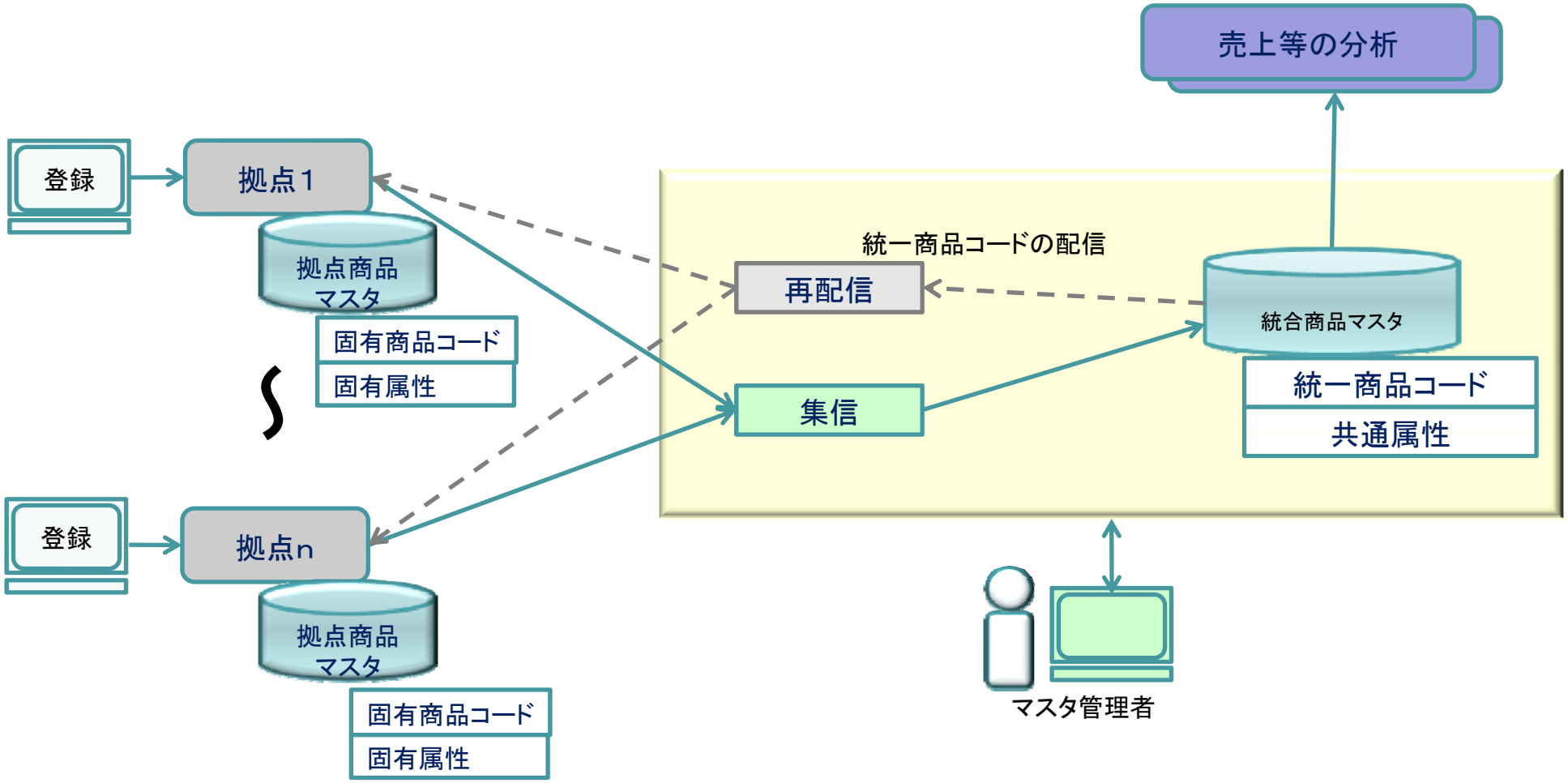
グローバル企業での統合DWHにおける課題

- 事業別に変換ルールが異なり、その変換プロセスは、ETLプロセス内に内在して組み込まれているため、ビジネス変化対応時のビジネスルールの変更、追加に追従するのが困難（俊敏に対応できない）。
- 事業別オペレーショナルDBの変更をDWHにいかに反映したら良いか。
- 各事業別DBではデータの意味定義が異なるため、グローバル視点での把握が容易にできない。

<<共通課題>>
•統合データモデルに基づくDWHは作成した
•DWHが充分活用されていない
•汎化された統合データモデルでは、欲しいデータの所在が解らない
•目まぐるしく変わる経営サイドのデータに対する要求
•ビジネス変化への迅速かつ廉価なシステム対応

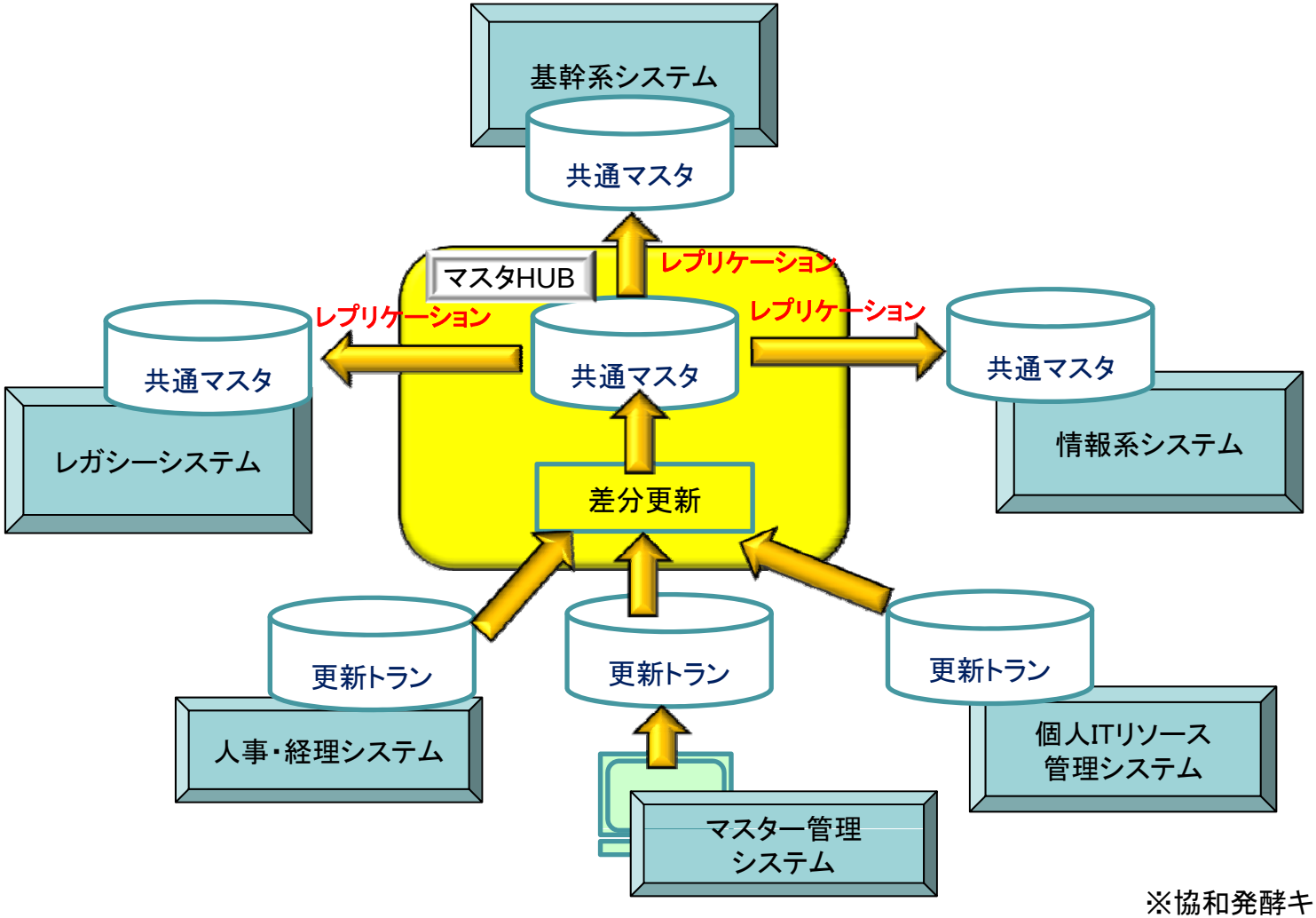
マスターデータ管理：商品マスタ

- 同一商品に対して拠点毎に異なったコードが発番されている
- 管理属性も拠点毎に異なる
- 全社レベルでの売上分析ができない



MDMを構築し成功している例

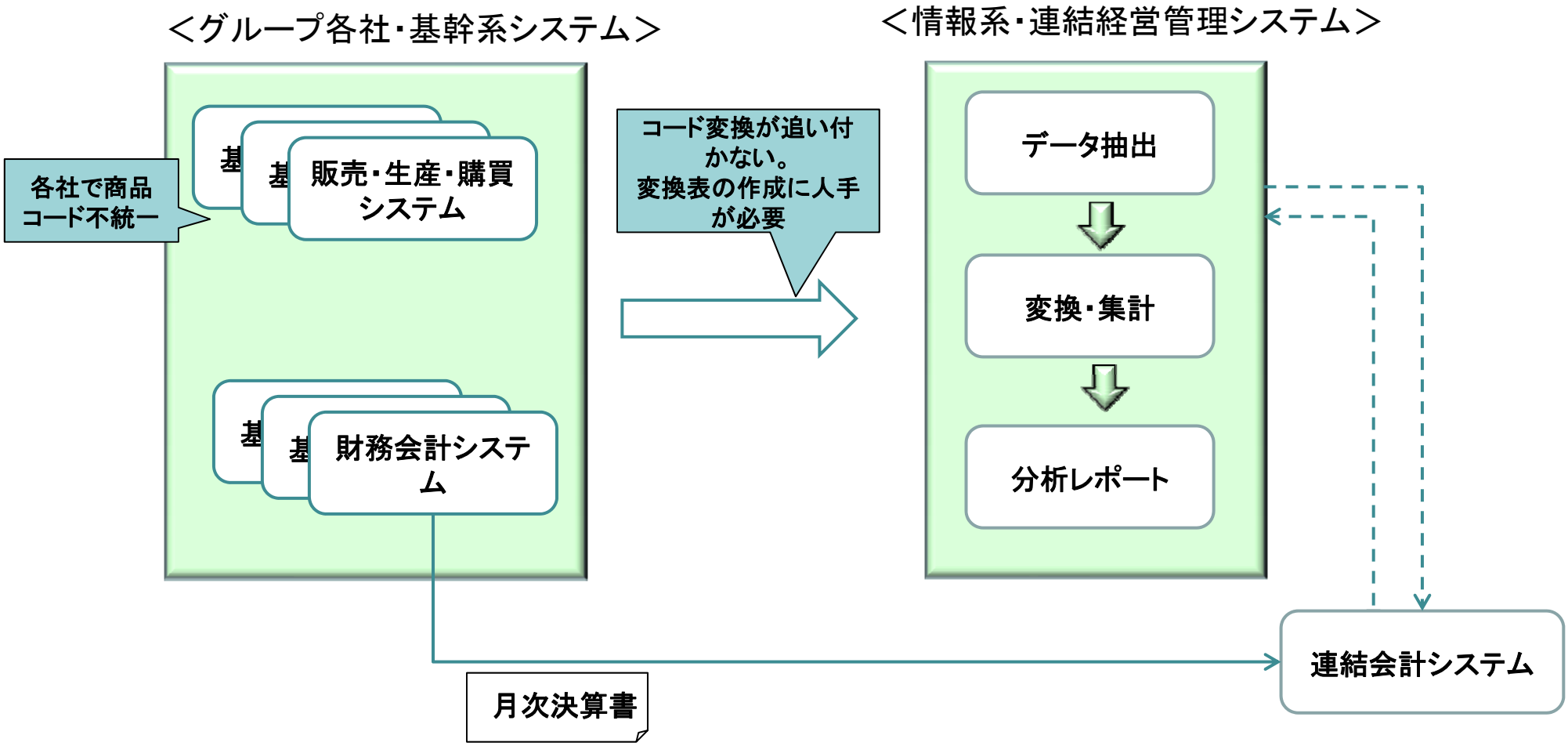
■ マスターの発生源を絞り込み、更新情報を利用システムに伝播



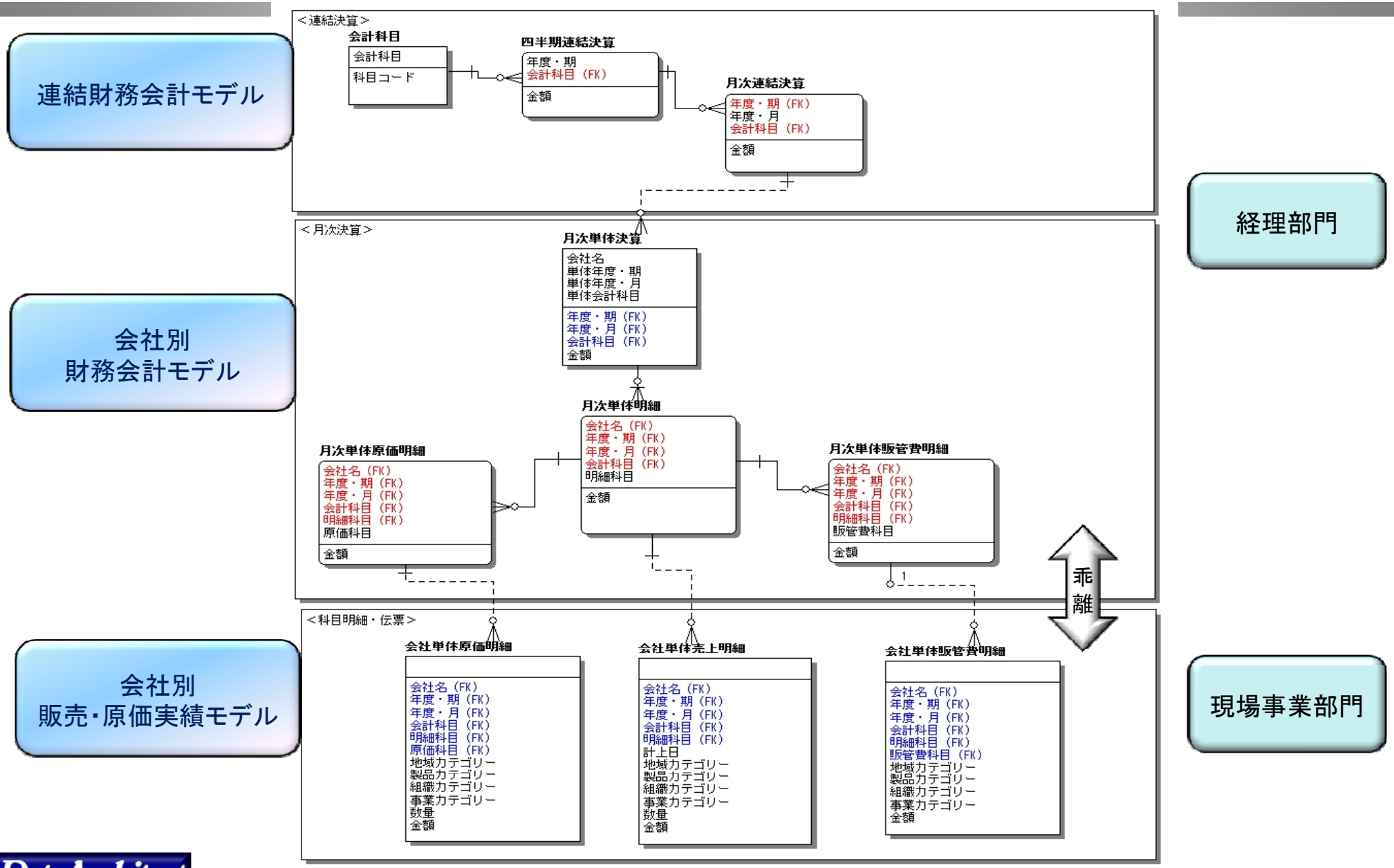
※協和発酵キリン社事例

基幹系と情報系システムでのデータ遮断

- グループ会社毎での商品コードの異なった体系
- 連結財務会計はできているが、連結経営管理ができていない！
- 集約データと明細データでの隔離



財務諸表から売上実績データが紐づかない！



５．ビッグデータ活用のために何をなすべきか

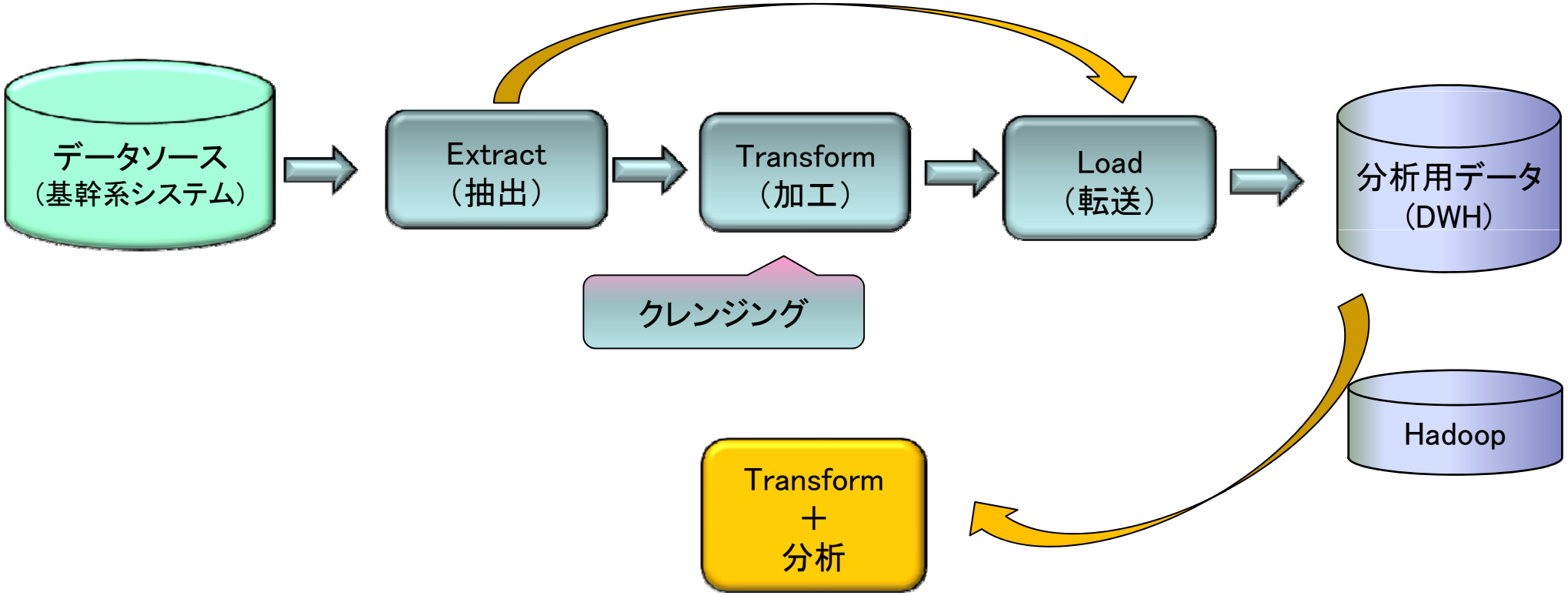
- データ品質の確保
  - データモデルによる品質向上
  - 正規化を忘れていないか
- データを活用するための土壌
- データ戦略部門・人材の配置
- ビッグデータを取込むためのデータアーキテクチャ策定

データ品質

- データ品質を担保しないまま、やみくもにビッグデータを分析しても、統計学的には有意なデータ分析結果が得られたとしても、ビジネスには有効なデータとはなり得ない
- 品質を何処で作り込むか（データを何処できれいにするか）
  - ◆DBに投入する前に
  - ◆とにかくDBに投入してその後で．．．
- 要求品質を見極める
  - ◆顧客が望まない名寄せもある
- 統計学
  - ◆有意性n%（仮説検証）
- 製造業での品質管理は日本のお家芸
- データ品質を高めるためのモデリング

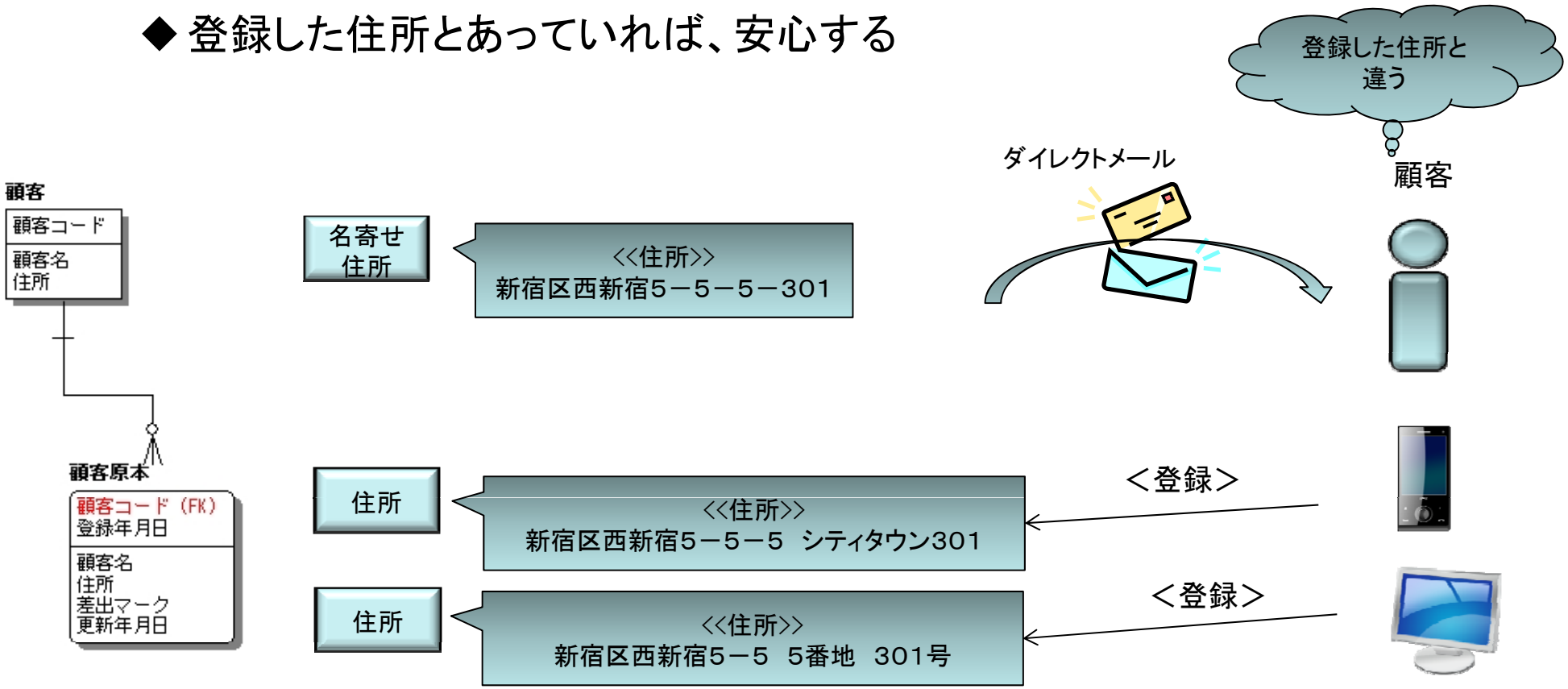
品質を何処で作り込むか

●ETL：データをクレンジング・加工して分析用データとして加工
●ELT：分析用データの器にひとまず投入してから加工処理を施す



要求データ品質：何が求められているか

- 要求品質を捉える
- むやみな名寄せはビジネス上有益とならない場合もある
  - ◆ 微妙な住所の違いに顧客は不信感を抱く
    - ● 個人情報が漏えいしているのではないか
  - ◆ 登録した住所とあっていれば、安心する

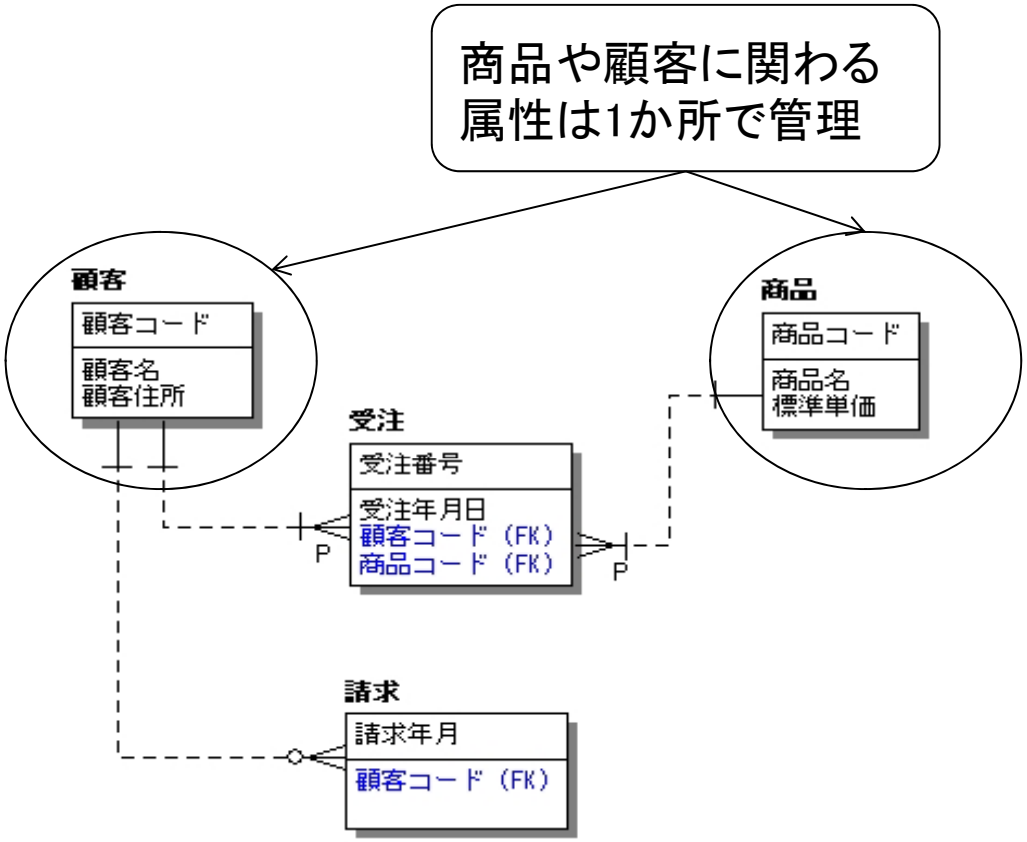


データモデルによるデータ品質向上施策

■ 誤ったデータが混入しないための器（データモデル）を用意する

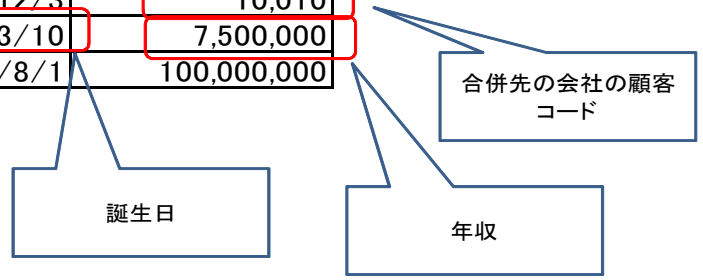
データ品質が悪い状態	データモデルでの防御策
誤ったデータが混入している	掛け持ちデータ項目を作らない NULL項目を回避する 主キーだけで統合しない
データが重複している	One Fact in One Place原則 移行データは期間限定で保持 重複データの導出元の明示
必要なデータがない	誤った正規化 適正なマスターの断面管理

大原則：One Fact in One Place



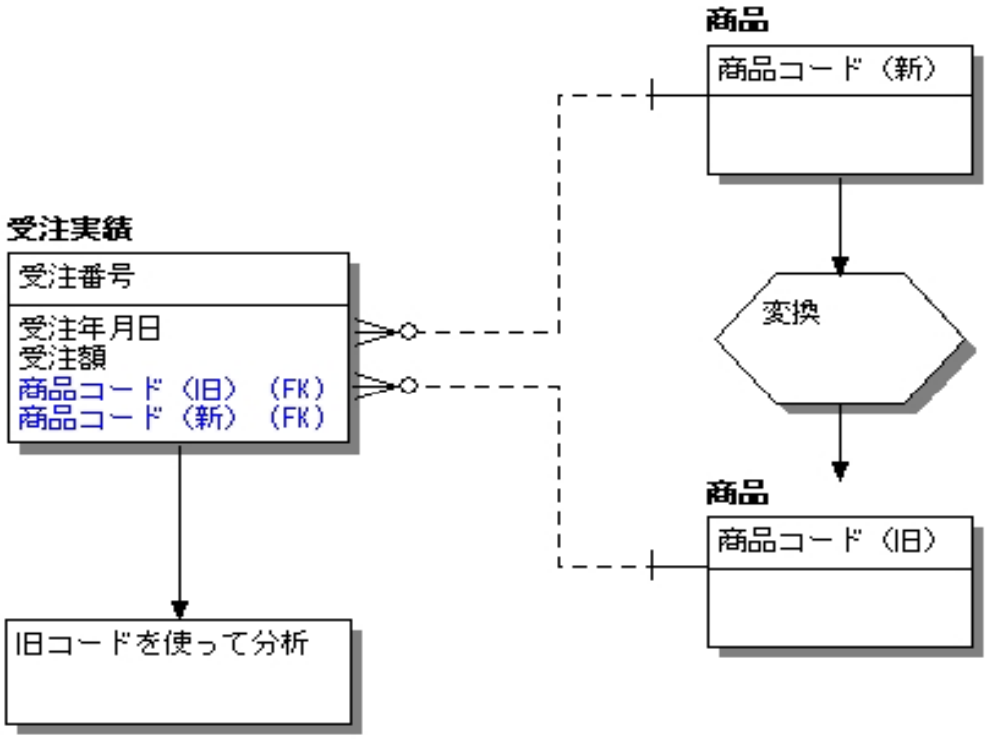
掛け持ちデータ項目排除

顧客コード	会社名	創立年月日	資本金
10010	ABC商会(株)	1980/1/10	10,000,000
10020	(株)データアーキテクト	2005/12/3	10,010
90008	山田太郎	1960/3/10	7,500,000
10030	(株)DEF工業	1975/8/1	100,000,000

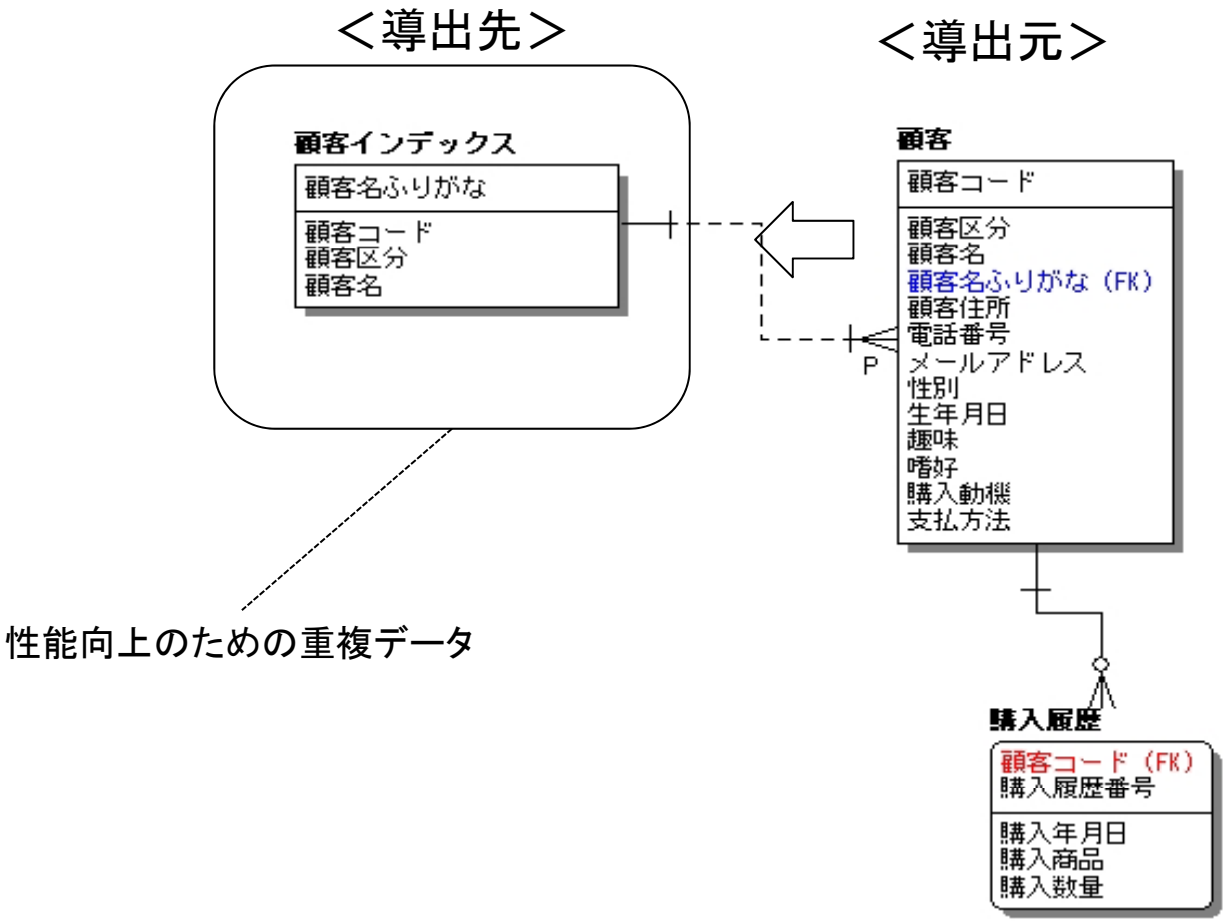


移行データは期間限定で

■ 期間を過ぎたらデータモデル洗浄を行う



重複データの導出元



準備すべきこと

- 組織・人材の整備
  - データ戦略部門を設置してCDO(chiefDataOfficer)を配置
    - データアーキテクト
    - データ管理者
      - DBAを育成
  - データサイエンティストの育成
  - データスチュワートの現業への配置
- アーキテクチャ策定
  - ビッグデータを付加した新たなデータハイウェイ
  - エンタープライズデータモデルの整備は大前提
  - ビッグデータ設計上の留意点など
- エンタープライズシステムの整備
  - MDM：マスターデータの統一

準備すべきこと

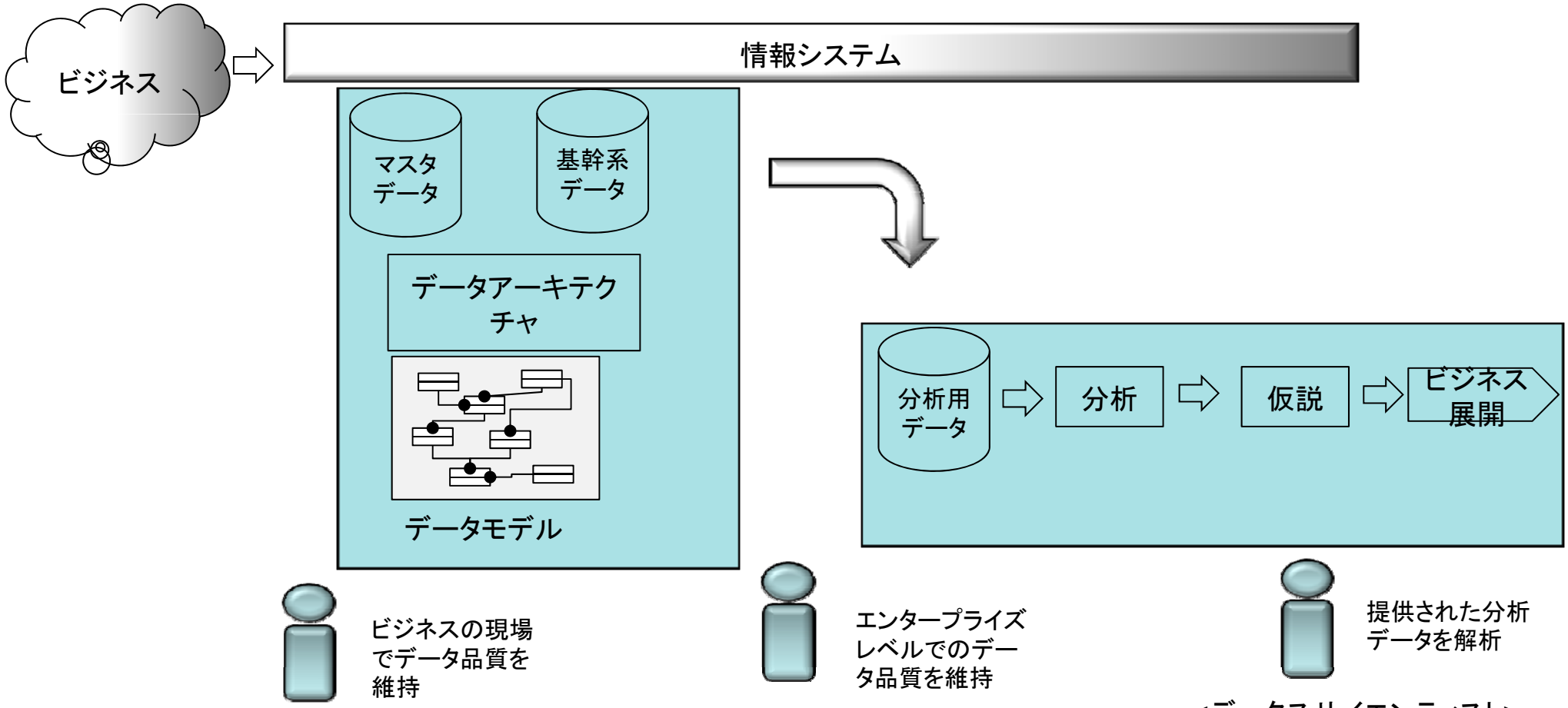
- ガバナンス
  - データセキュリティ
    - 適切なセキュリティとアクセス権限の管理
  - データ品質の維持
    - 一貫性と正確性の確保
  - メタデータ整備
    - 定義の標準化、所在や所有者の明確化、
  - 分析ニーズに足るだけのデータの種類とデータ量、鮮度の維持

エンタープライズシステムに習うべきところ

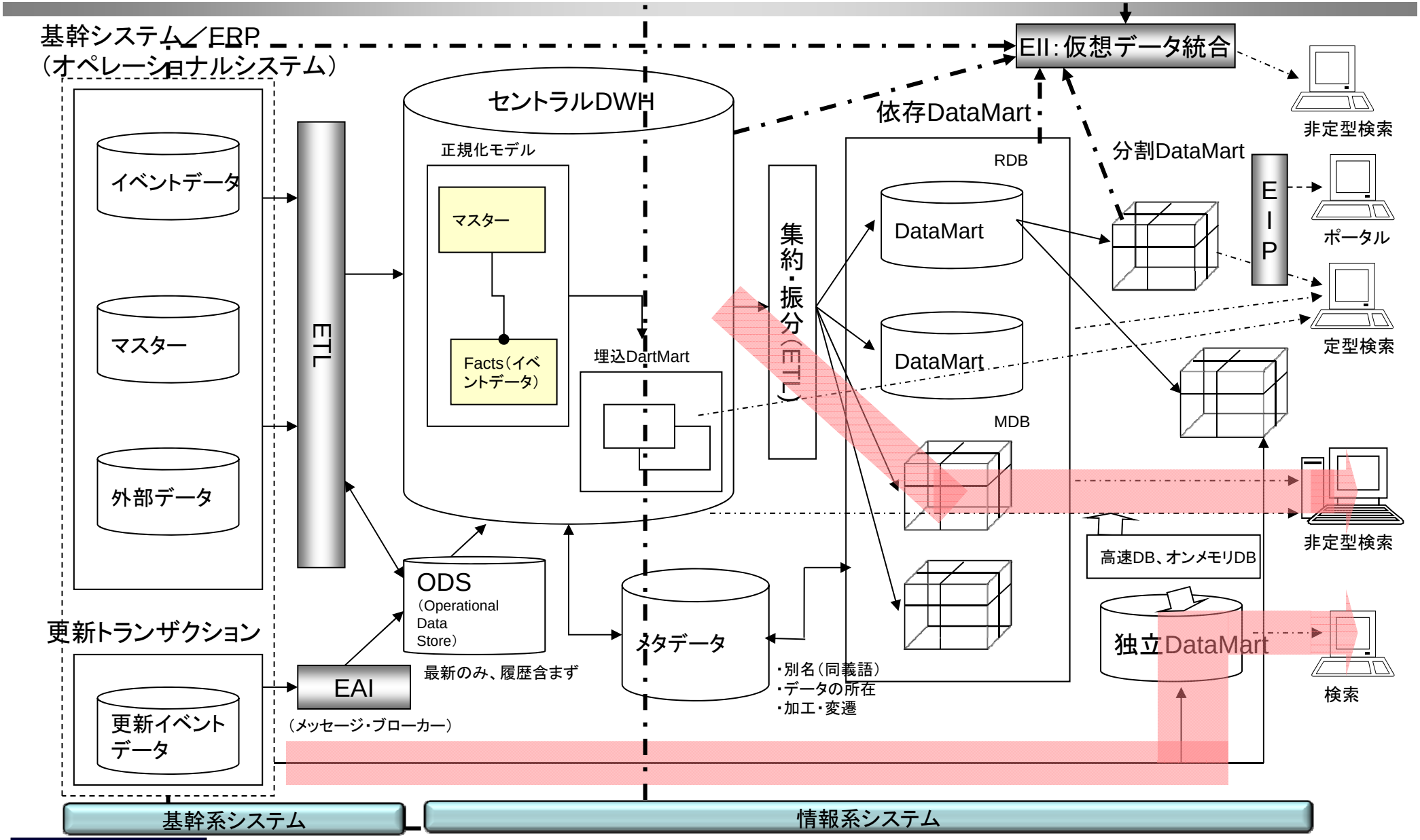
- インターネット上のさまざまなメディアやチャネルから収集したデータは、フォーマットや品質がまちまちで、一貫性を欠いている
- 収集したデータをビジネスで有効活用するには、精度が高く、かつ一貫性のある「データモデル」を構築し、仮想データ統合することが得策
- エンタープライズシステムが抱えているマスタデータに対するMDM課題解決と同等の考え方

データ戦略部門・人材の配置

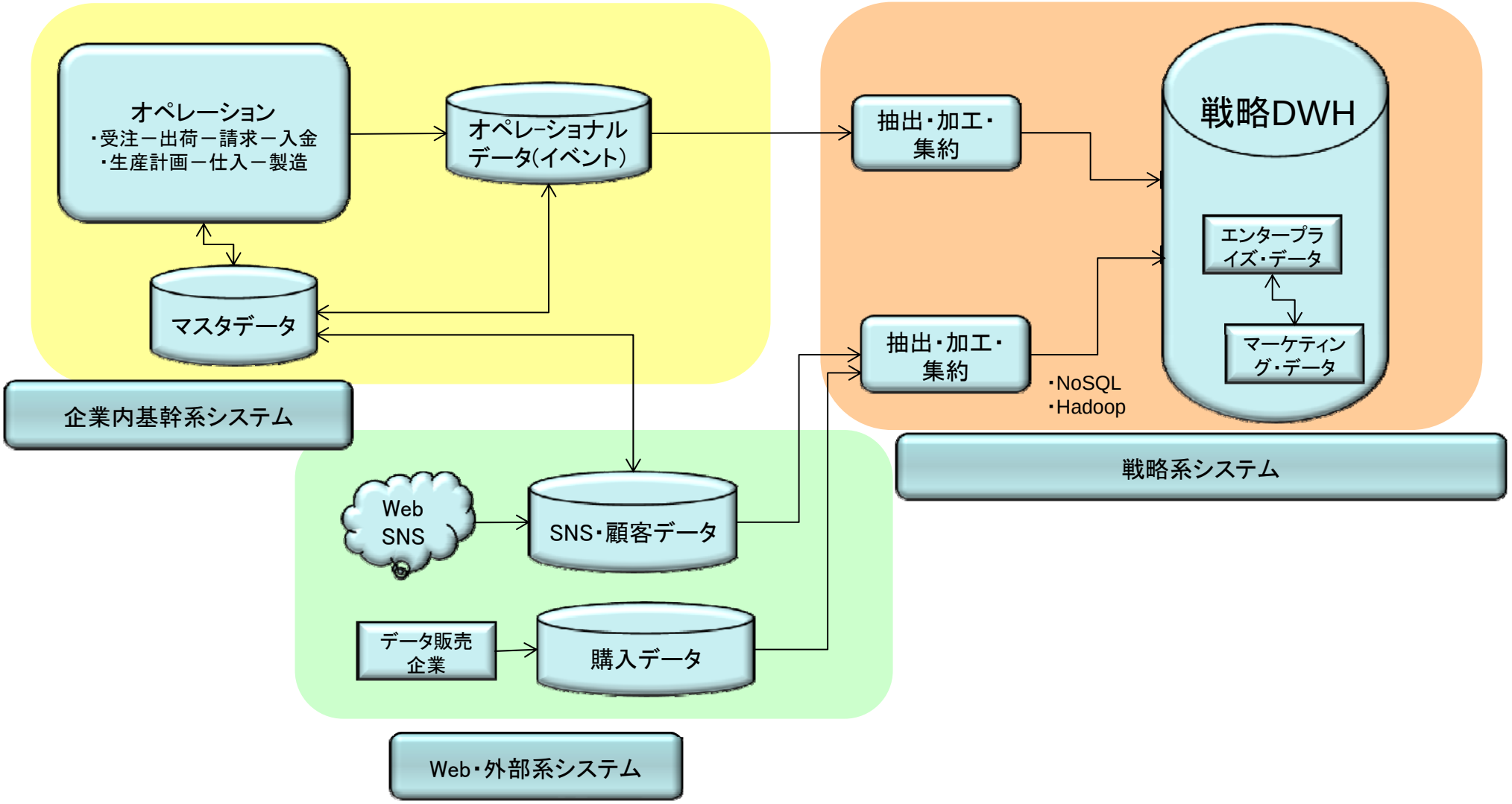
- データ戦略部署の新設
  - CDO(chief Data Officer)、データアーキテクト、データサイエンティスト、データスチュワート
- 企業内外のデータをクローリングして正規化、クレンジングしてデータサイエンティストに託す
- データサイエンティストに渡すまではデータ管理者（データアーキテクト）、データスチュワートの役割



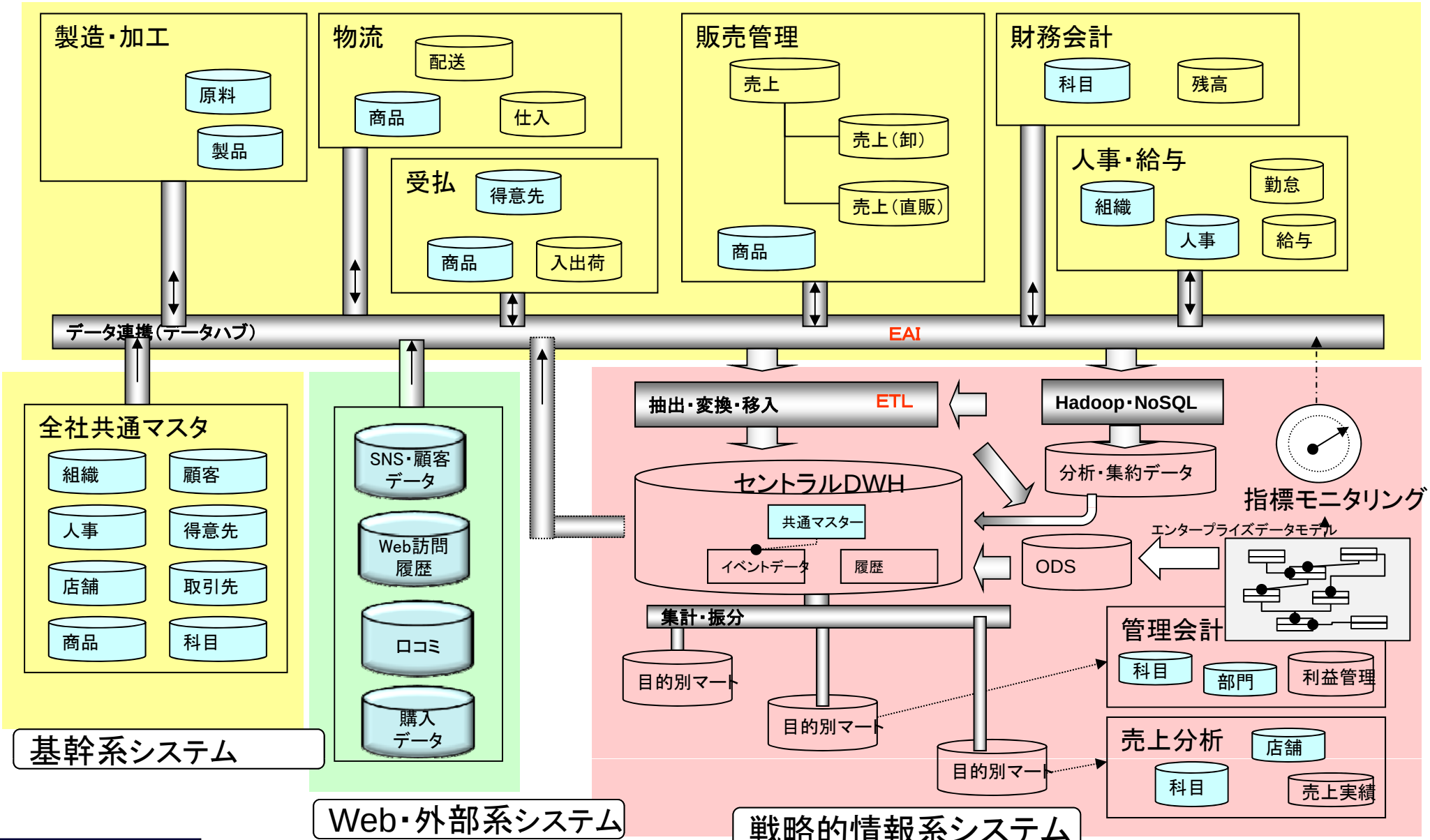
エンタープライズシステム・データアーキテクチャ(as-is)



ビッグデータを考慮したデータアーキテクチャ



ビッグデータを組込んだ理想的なデータアーキテクチャー例



6. データ戦略

- 経営戦略に則ったデータ戦略をプランニング
- 経営戦略に基づいたIT戦略
  - 今後はデータ戦略
  - CDOの配置
- IT戦略は、クラウド、コモディティ化のため差別化が無くなってきている。
- データ戦略：
  - コンテンツ、モデル、アーキテクチャのデータコンテンツマネジメントが重要となる
  - データガバナンス
- ビッグデータを扱うには、データ品質の担保が必要
  - 企業内から集積するデータは、メタレベルでの品質強化を図る必要がある
  - エンタープライズ・システムのデータ整備ができていない状態で新たなビッグデータを取り込めない

データ駆動型企業

■ データ駆動型企業への脱皮

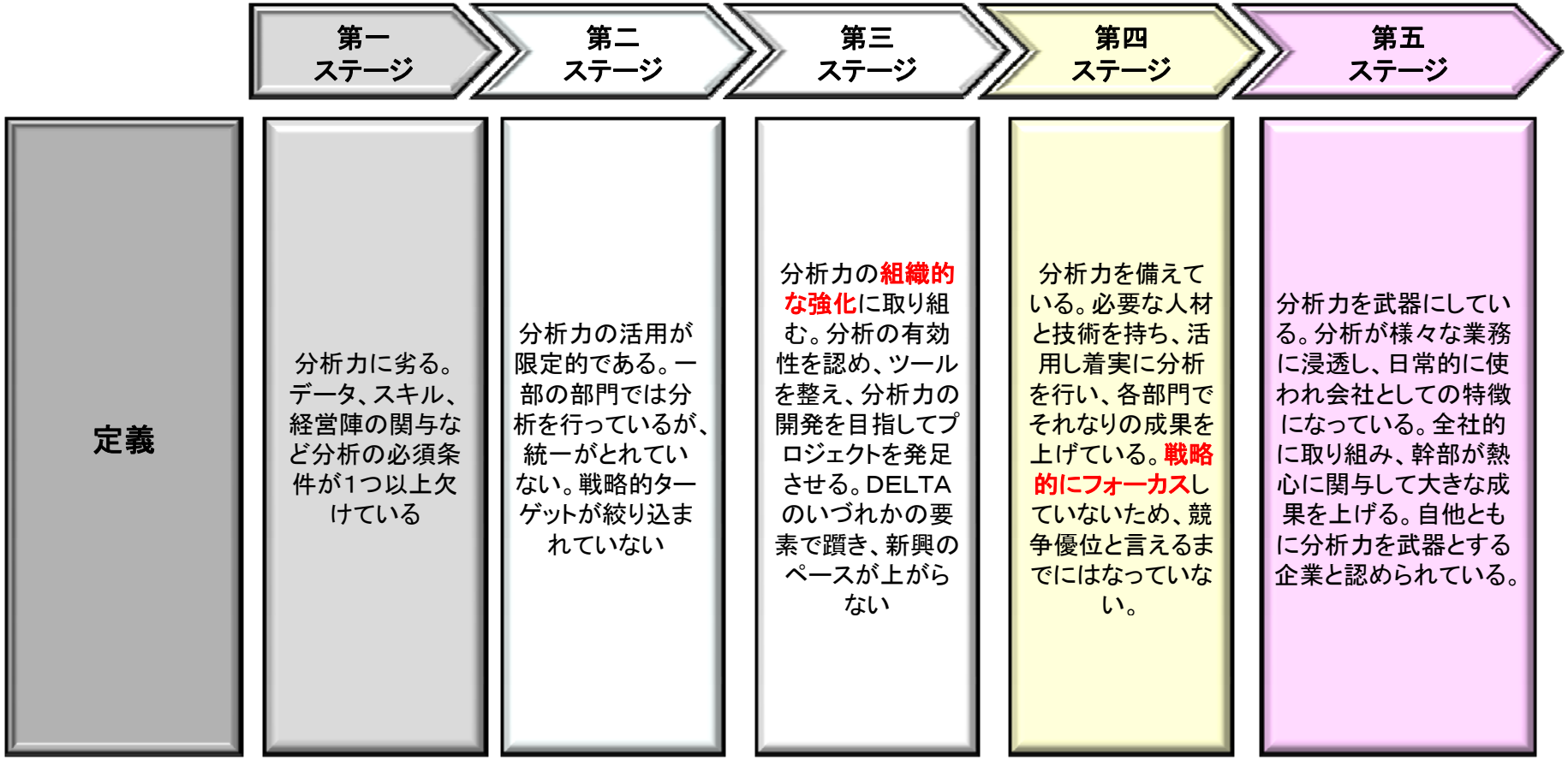
◆分析力を武器とする企業

- ダベンポート氏によるステージの考え方

■ データ戦略

	社内データ	外部データ
戦略的	マーケティングデータ	他社データ
オペレーショナル	基幹システムデータ	定性流通データ

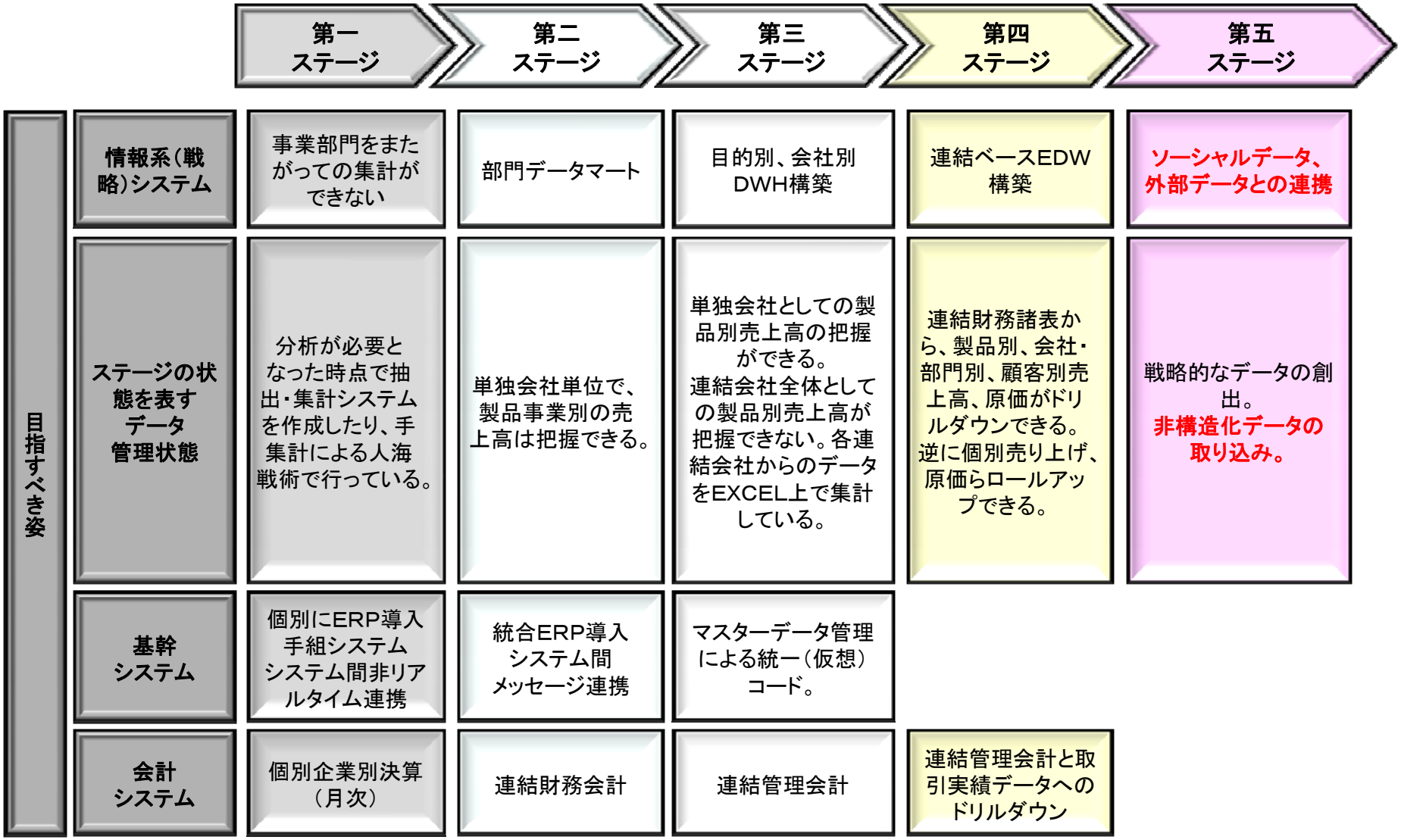
データの戦略的活用化のステージ（例）



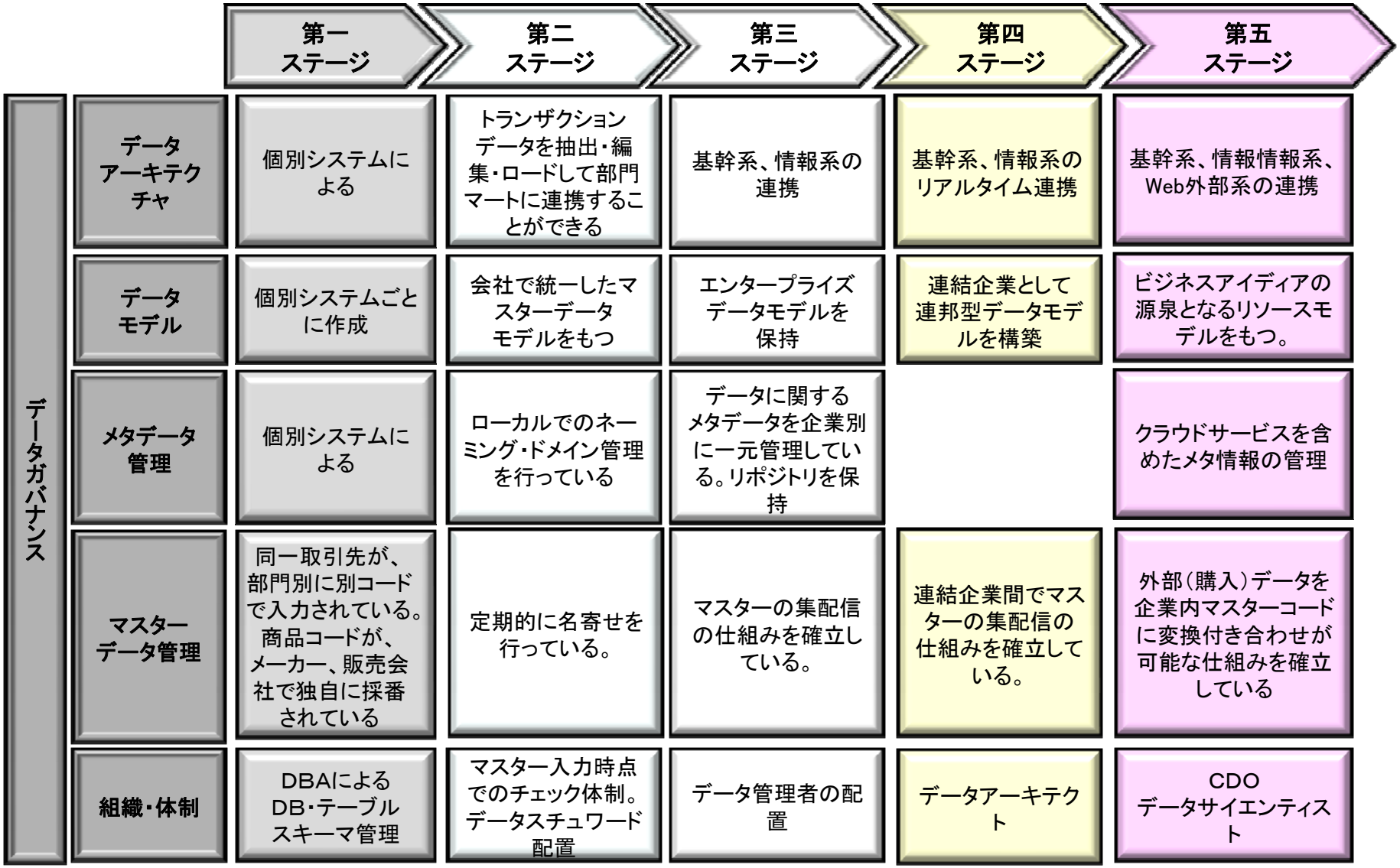
分析力を支える5つの要素をDELTAであるとしている。
・データ
・エンタープライズ
・リーダーシップ
・ターゲット
・アナリスト

出典：分析力を駆使する企業
トーマス・H・ダベンポート著

データの戦略的活用化のステージ（例）



データの戦略的活用化のステージ（例）



謝辞

ご静聴有難うございました。

講演内容に関してご質問がございましたら、下記までお問い合わせください。

mano@dataarch.co.jp