

地味にスゴイ「データ仕様可視化技術」 ～その効用と適用事例～

2017/04/20

株式会社日立製作所 システム&サービスビジネス統括本部
アプリケーションサービス事業部
サービスソリューション本部 サービス統括部

清水 薫

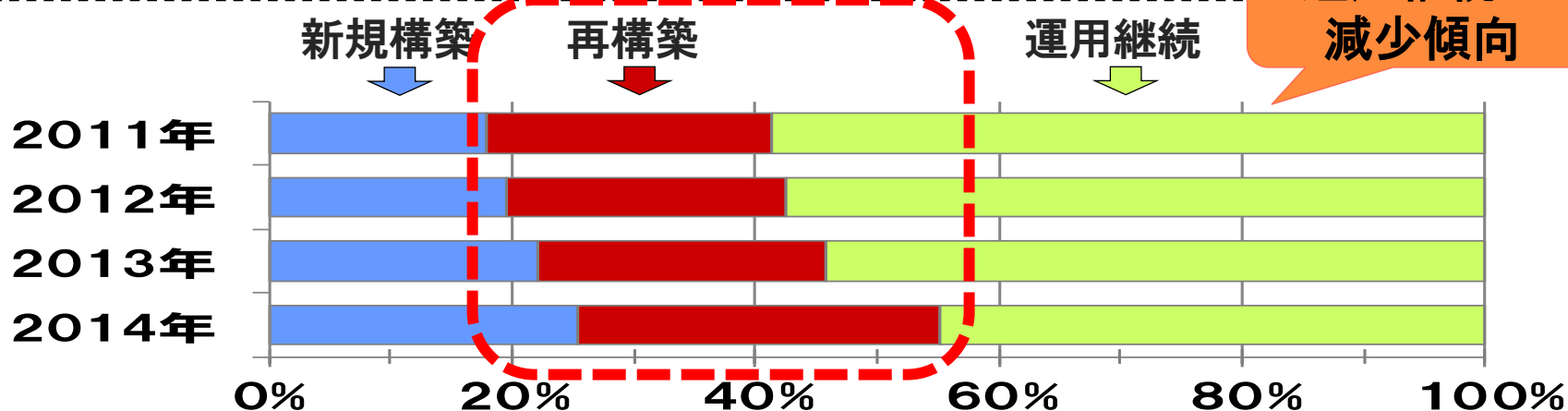
Contents

1. ITモダナイゼーションにおけるデータの問題
2. データ現状調査の課題
3. データ仕様可視化技術
4. 「レガシー」を次の時代へ ～ データの利活用～
5. まとめ

1. ITモダナイゼーションにおける データの問題

- 企業が勝ち残っていくためには
「戦略性の高いシステムへの投資を続けていかなければならない」

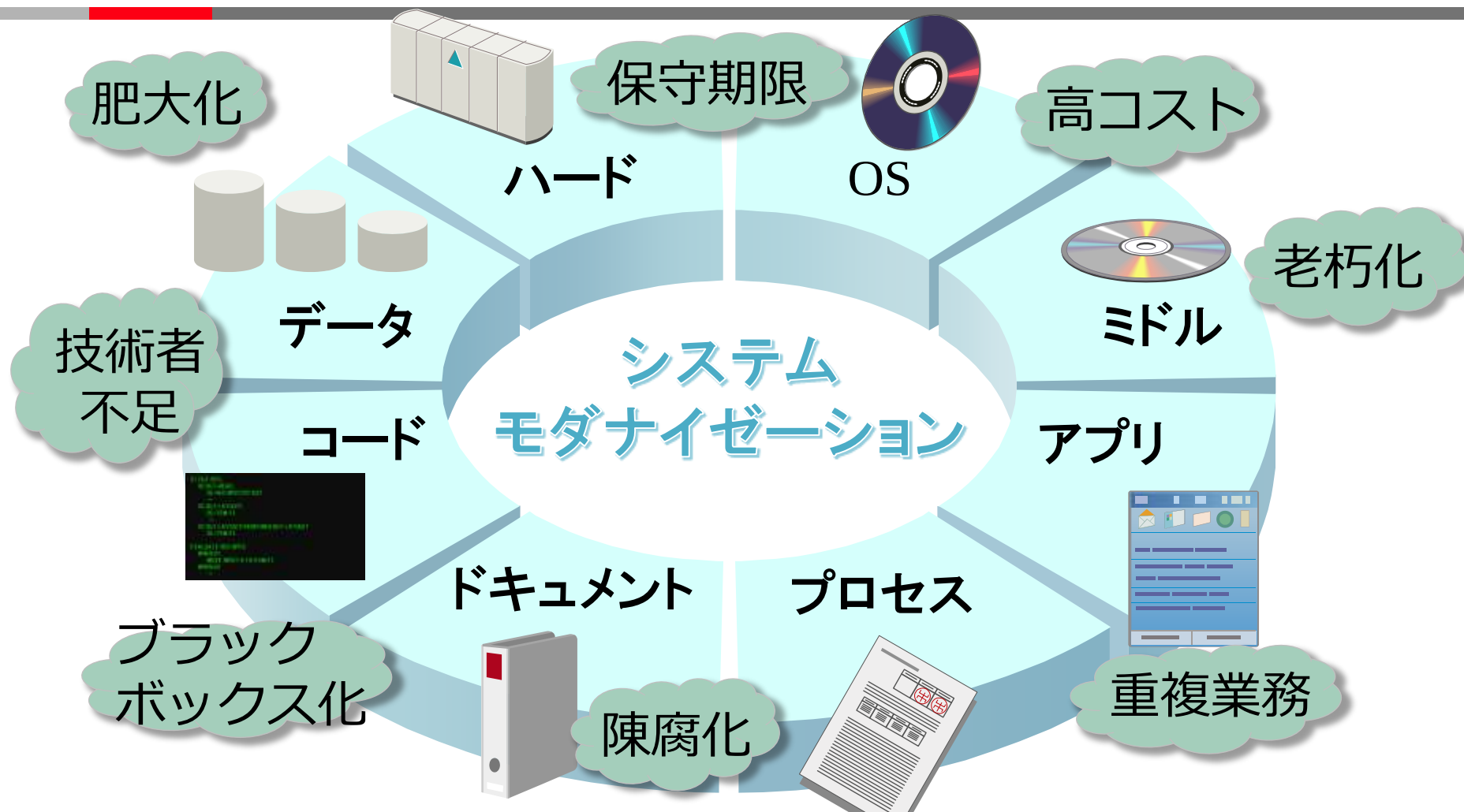
情報システムの取り組み状況 件数ベース



出典: 経済産業省 平成26年情報処理実態調査結果報告書 情報システムの取り組み状況

システム再構築への投資(取組み)が増加

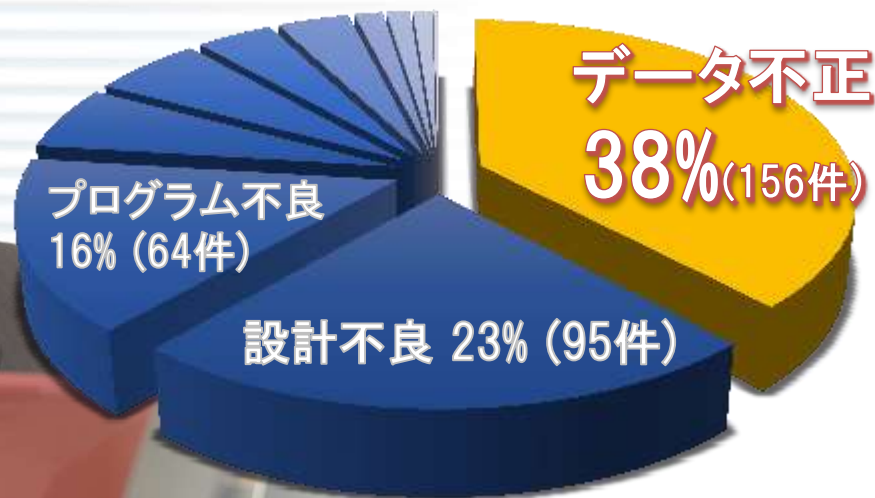
戦略的なシステム投資として再構築案件は拡大しており、
変化に対応できる再構築が重要性を増している。



これらの課題を克服し、ビジネス環境や法制度対応など、変化に対応できる強いIT基盤に変えていくことが必要

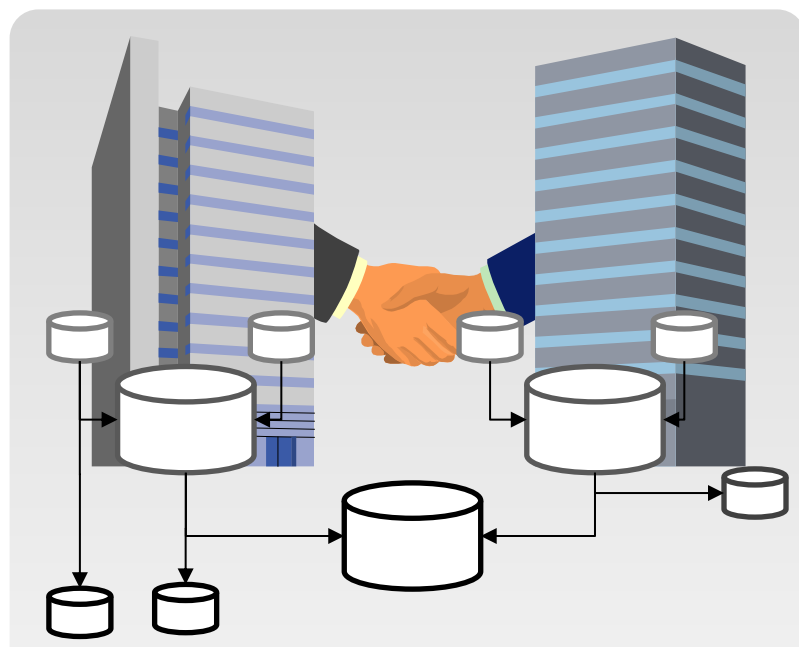
どのような問題が特に多く発生しているか？

あるプロジェクトのシステム更改後に発生した問題の
原因分類別割合 (全409件)



どの使用言語(COBOL, Java, etc.)でも、プログラムとは別に
データ不正がITモダナイゼーションの大きな阻害要因

長年運用されてきたシステムのデータは
複雑化し、異常を含んだ「汚れたデータ」となっている



度重なる統廃合による
仕様の複雑化



手順外運用による
異常データの発生

本番稼働後、
業務中にトラブルが多発

原因は…

- データに不備・不足が存在
- 現新比較テストが不十分

何故か売上の
集計が合わない…

なんで〇〇商事さんの
注文だけ通らないんだ！？



顧客ID	顧客名	売上
C01-123	(有)△△	¥123,456
C01-123	異常データ	¥789,123
<NULL>	〇〇商事	¥456,789

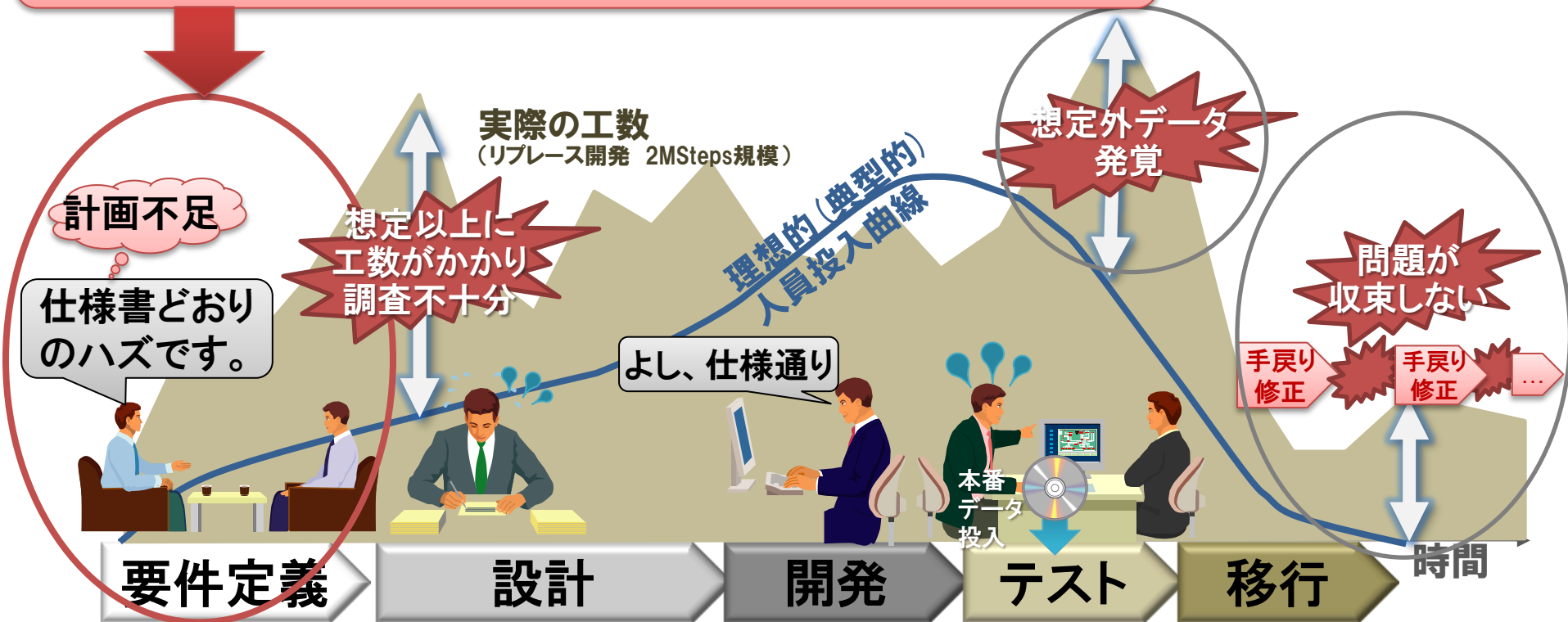
レコード漏れ

長期間システムが使えず手運用で業務を回し、
大幅にコストが増加。社外にも謝罪する事態に。



元をたせば...

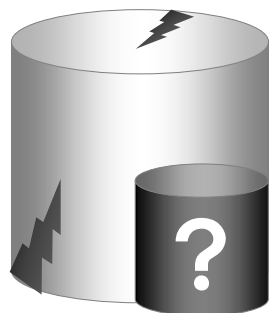
データ品質確保計画が甘かった



データ品質確保に必要な作業を十分計画・実施できず、
プロジェクトの後半にもぐら叩きが発生。

トラブルを未然に防止するには、
データ品質を確保しておくことが重要

データ品質確保の3ステップ



現状調査

方針検討

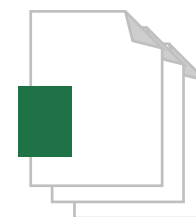
クレンジング
名寄せ



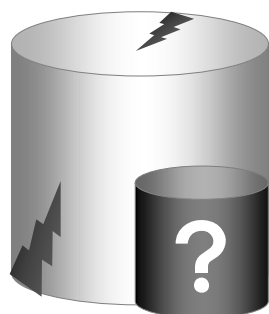
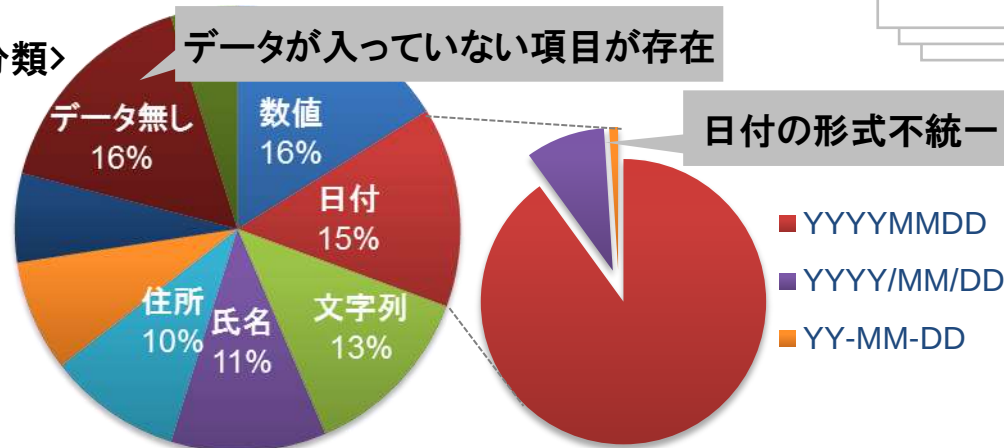
業務データの現状を確認し、疑わしいデータや 対策が必要なデータを検出

データ品質分析結果

- ✓総項目数:100 うち主要3カラムを中心に調査。
- ✓総レコード件数: 12,006,546 (約1,200万レコード)



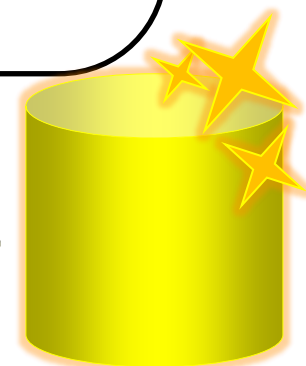
＜データ分類＞



現状調査

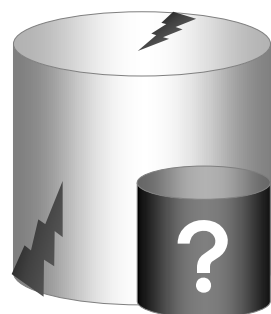
方針検討

クレンジング
名寄せ



現状調査で洗い出した懸念点・問題点に対し、クレンジングなどの対策方針を検討

#	項目名	懸念点	クレンジング 要否	方針(案)
1	発送種別	分類フラグに空白値がある	不要	空白も1つの種別として利用しており問題なし
2	品目コード	コード体系が不統一	場合により要	新規にコードを割り当てるためクレンジング不要
3	受注日付	不正な形式の日付データ	必要	長期保持する必要があるため変換が必要



現状調査

方針検討

クレンジング
名寄せ



検討した方針に従ってクレンジングや名寄せを行い、データの整合性を確保

#	企業名	都道府県	市区	電話番号
1	ABC	東京都	新宿	0312345678
2	株式会社ABC	東京	新宿区	03-(1234)-5678
3	(株)ABC	東京都	新宿区	03-1234-5678

↓ ツール等を使用して変換

#	企業名	都道府県	市区	電話番号
1	(株)ABC	東京都	新宿区	+81312345678
2	(株)ABC	東京都	新宿区	+81312345678
3	(株)ABC	東京都	新宿区	+81312345678

#	企業名	住所
1	株式会社ABC	東京都 新宿区 新宿ビル
2	株式会社ABC	東京都 新宿区
3	株式会社ABC	東京都 文京区

↓ マッチング候補を洗い出し、統合

#	企業名	住所
1	株式会社ABC	東京都 新宿区 新宿ビル
2	株式会社ABC	東京都 新宿区

候補について、目視で一致／不一致を最終判断

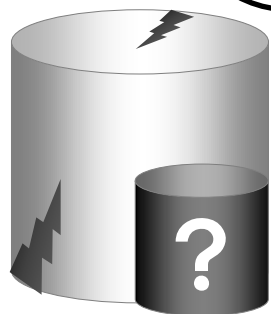


現状調査

方針検討

クレンジング
名寄せ

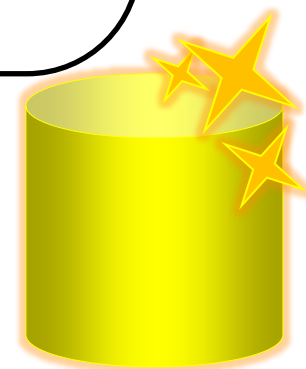
**現状を把握しないことには
データ品質は確保できない**



現状調査

方針検討

**クレンジング
名寄せ**





2. データ現状調査の課題

頼れない
仕様書

データ量が
膨大

調査計画が
立てづらい

頼れない 仕様書

仕様書の記載が曖昧。実態との乖離も。

仕様書に記載があったとしても、多くの場合不十分。
長年の運用で**例外的なデータ**が作成されている場合も。

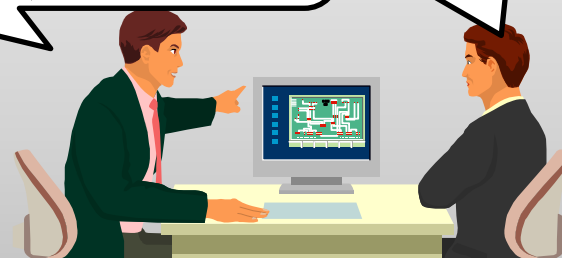
- ・そもそも残っていない
- ・記述が曖昧
- ・記述が分散
- ・履歴管理が不十分



仕様の全体把握が難しい

お客様の場合
氏名欄をこう工夫して
担当に知らせます

もはや仕様の
一部だな...



例外的なデータが存在

仕様書のみを基にした調査観点の積み上げでは、
抜け漏れ・誤認や実態との乖離を生じ、後工程での手戻りに。

頼れない
仕様書

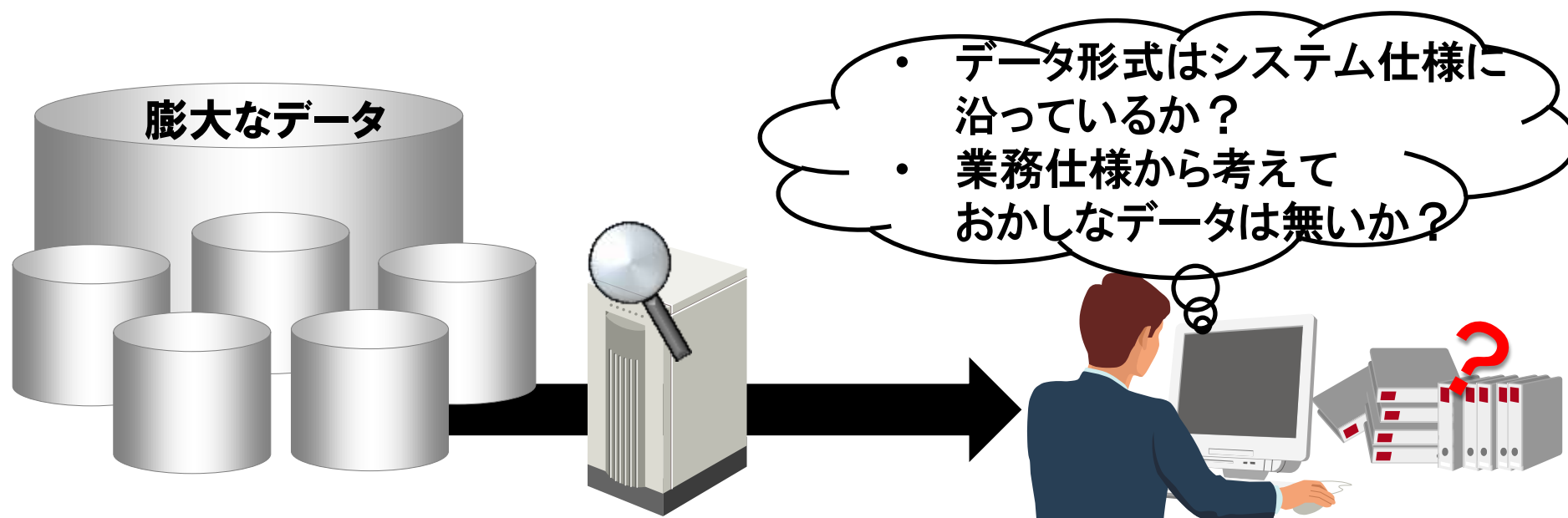
データ量が
膨大

調査計画が
立てづらい

データ量が
膨大

膨大なデータを手掛かり無しに漏れなく調査するのは困難

- ・膨大なデータは全体を俯瞰できない。
- ・検索や集計を何度も繰り返さなければならない。
- ・想定できなかった異常には気づきようが無い。



膨大なデータを手掛かり無しに抜け漏れなく調査するのは困難。
調査漏れに気づくのも難しい。

頼れない
仕様書

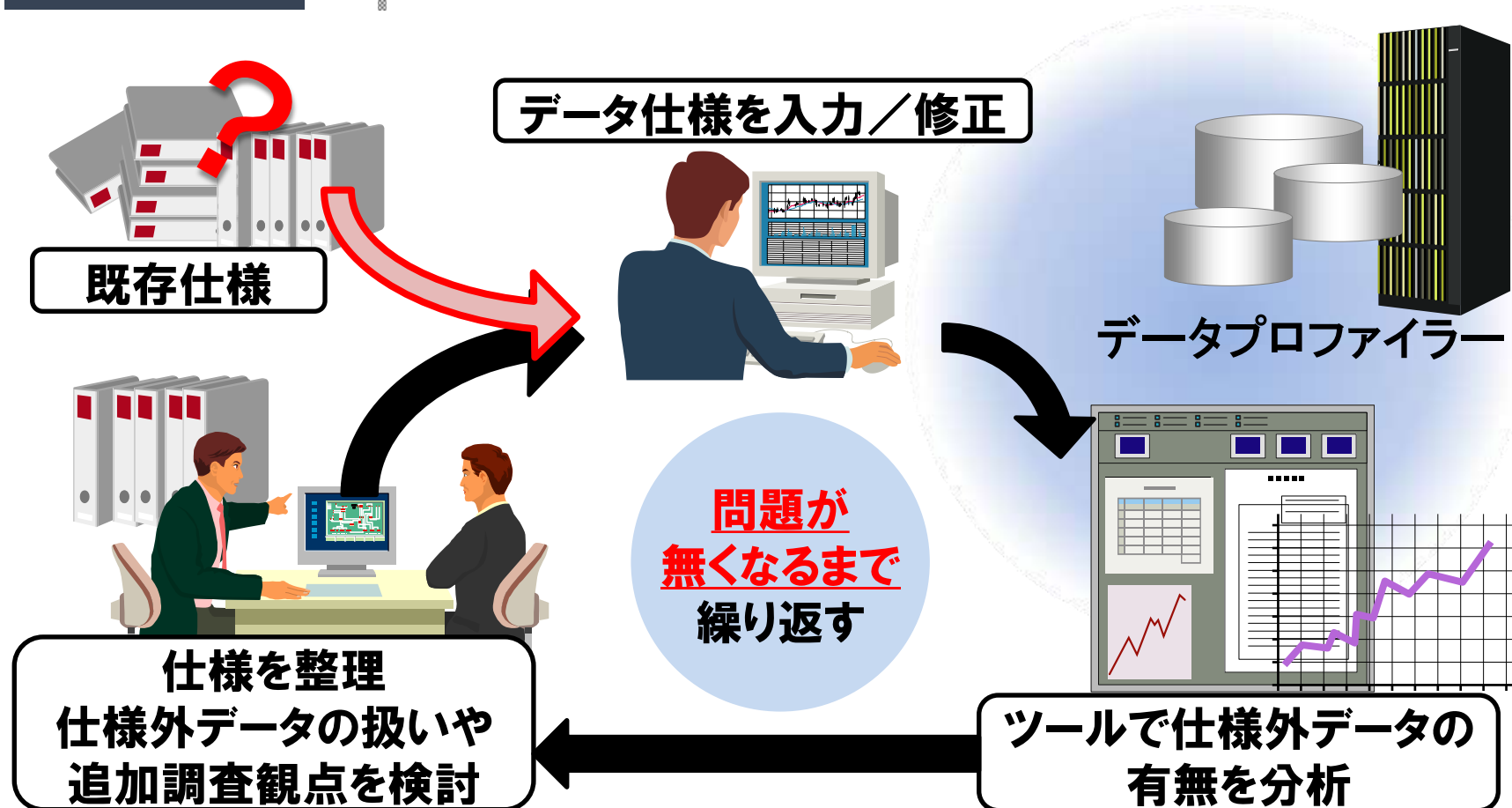
データ量が
膨大

調査計画が
立てづらい

調査計画が
立てづらい

データ仕様調査は地道な繰り返し作業

- ・ツールを使う場合でも最初の仕様入力に時間がかかる。
- ・一度で全貌が見えないので、作業の全体見通しが立たない。



- プロジェクト早期は、人手も時間も限られている。

平均的な 工程期間	要件定義 18日(※)	基本設計～総合テスト 102日間(※)
		(※) IPA ソフトウェア開発データ白書2014-2015 P.225 図表8-1-10 要件定義工程の工程比率の中央値0.146 (N=139)より、 プロジェクト期間6か月(120日間)の場合として算出
1回の調査に 要する期間 (弊社実績値)	5～20日 程度	

1回調査するだけで
要件定義期間を使い果たしてしまう

頼れない
仕様書

データ量が
膨大

調査計画が
立てづらい



データの全体俯瞰ができないため、
十分な計画が立てられず、
リスクが残存する／モグウたたき作業に陥る

**業務仕様が分からなくても、お任せで素早くそれなりに校閲してくれる。
そんな地味にスゴイ技術が欲しい。**

データ仕様可視化技術

3. データ仕様可視化技術

頼れない
仕様書

データから
仕様を推測

データ量が
膨大

大量のデータを
素早くサマリー

調査計画が
立てづらい

計画に使える
分析レポート

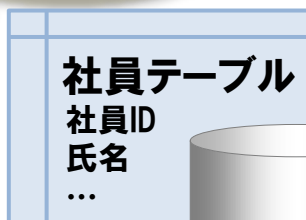
ポイント① データの特徴から仕様や例外を自動推測 (1/2) **HITACHI** Inspire the Next

■ 項目ごとのデータ仕様 (書式・制約など) を**自動推測**し、例外箇所を抽出

- どんな書式・制約があるか? 《例: 英字+8ケタの数字である》 (データ仕様抽出)
- 明らかに外れたデータがないか? 《例: 9ケタの数値も含まれている》 (異常データの抽出)

合併の影響で社員IDの体系がバラバラだ…
データ量も多いし、昔の仕様書を
全部調べるのも時間がかかるな…

社員IDは8桁または9桁で確定だな！
9桁や空データは把握していなかったので
さらに詳しく確認しよう！



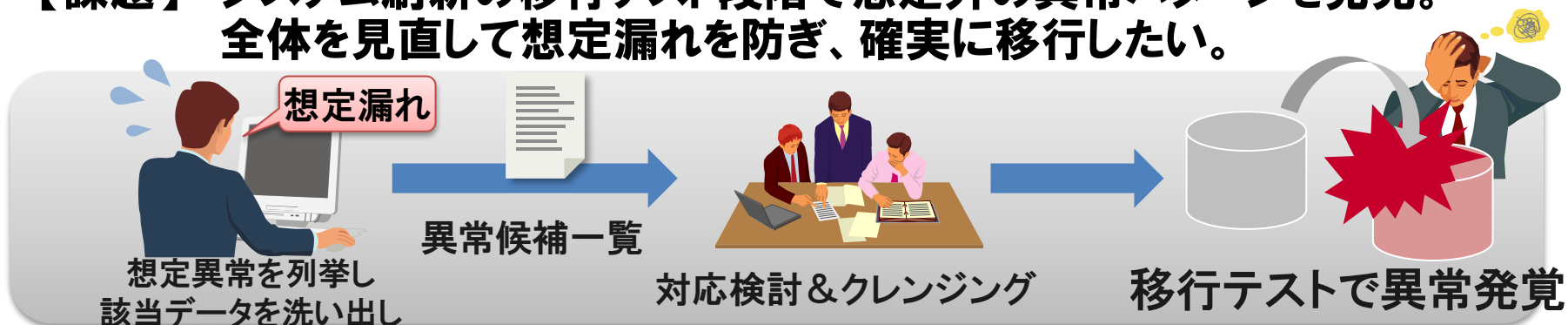
[結果イメージ]

大半は英字で始まる8桁のコード

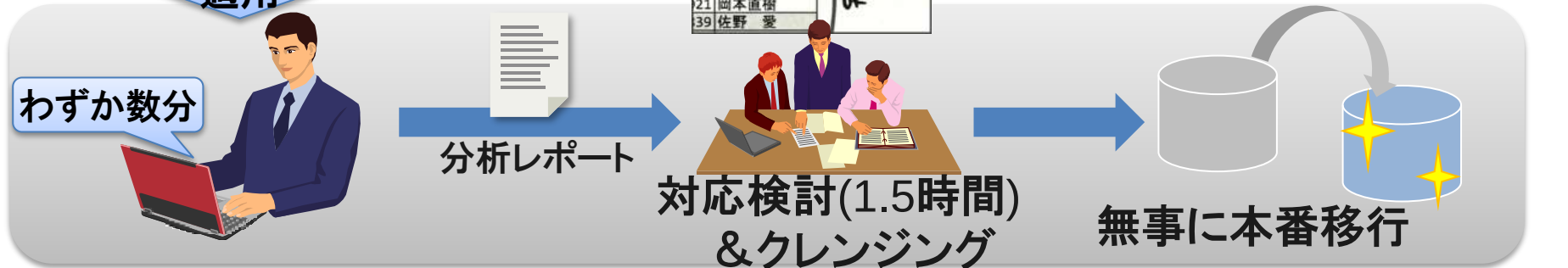
#	カラム名	パターン	正規表現	例1	例2	頻度	実数	種数
1	社員ID	コード系	^[A-D][0-2][0-9]{6}\$	A0123456	C0123456	87.0%	16116	84
			^[1-7][0-2][0-9]{7}\$	101234560	201234561	8.7%	1606	18
			<<空>> ⚠			4.0%	730	1

9桁の数値コードや空データも存在

【課題】 システム刷新の移行テスト段階で想定外の異常パターンを発見。
全体を見直して想定漏れを防ぎ、確実に移行したい。



データ仕様
可視化技術を
適用



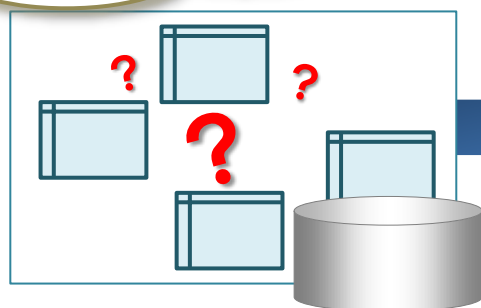
シンプルなレポートの確認だけでデータの全体像を把握。
ごく短時間でクレンジング方針を立案。

確実に
移行実施!

■ 実データからデータ構造を自動推測して、ER図や一覧形式で可視化

- どのデータが意味のある塊か／どのデータ間に関連があるか？ (データ仕様の抽出)
- 関連が途切れてしまっているデータ (マスタ不足など) はないか？ (異常データの抽出)

数千テーブルもあるのに
構造不明ではどう手を付けてよいか...



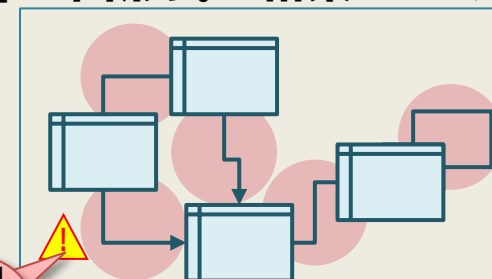
概ねこういう構造か！
マスタに不足があるようなので
詳しく調査したほうが良いな！



[一覧形式の結果イメージ]

親テーブル		子テーブル		関連の状態		
テーブル名	カラム名	テーブル名	カラム名	件数	種数	一致割合
PJWORK	WorkId	T_PRIVILEGE	WORK_ID	1489	22	100%
PJWORK	WorkId	PJVISIBLEOUTPUT	WorkIdToSee	201	40	98%

[ER図形式の結果イメージ]

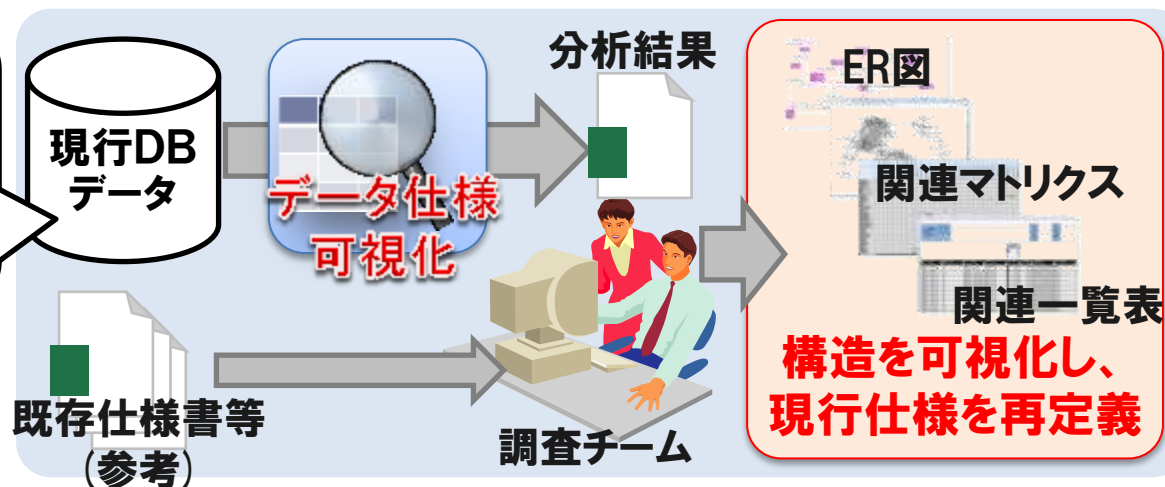


不足
データ

【課題】データの経年増加により運用中に性能問題が発生。
至急性能対策したいが、現行ドキュメントが整備されておらず、仕様把握が困難。

⇒現行DBデータをお預かりし、データ仕様可視化技術で可視化。

No	項目	合計
1	カラム数	約1万8千
2	レコード数	約2000万



現行システムのデータ構造を
わずか**1週間**で可視化し、リドキュメント。

性能対策
実施!

頼れない
仕様書

データから
仕様を推測

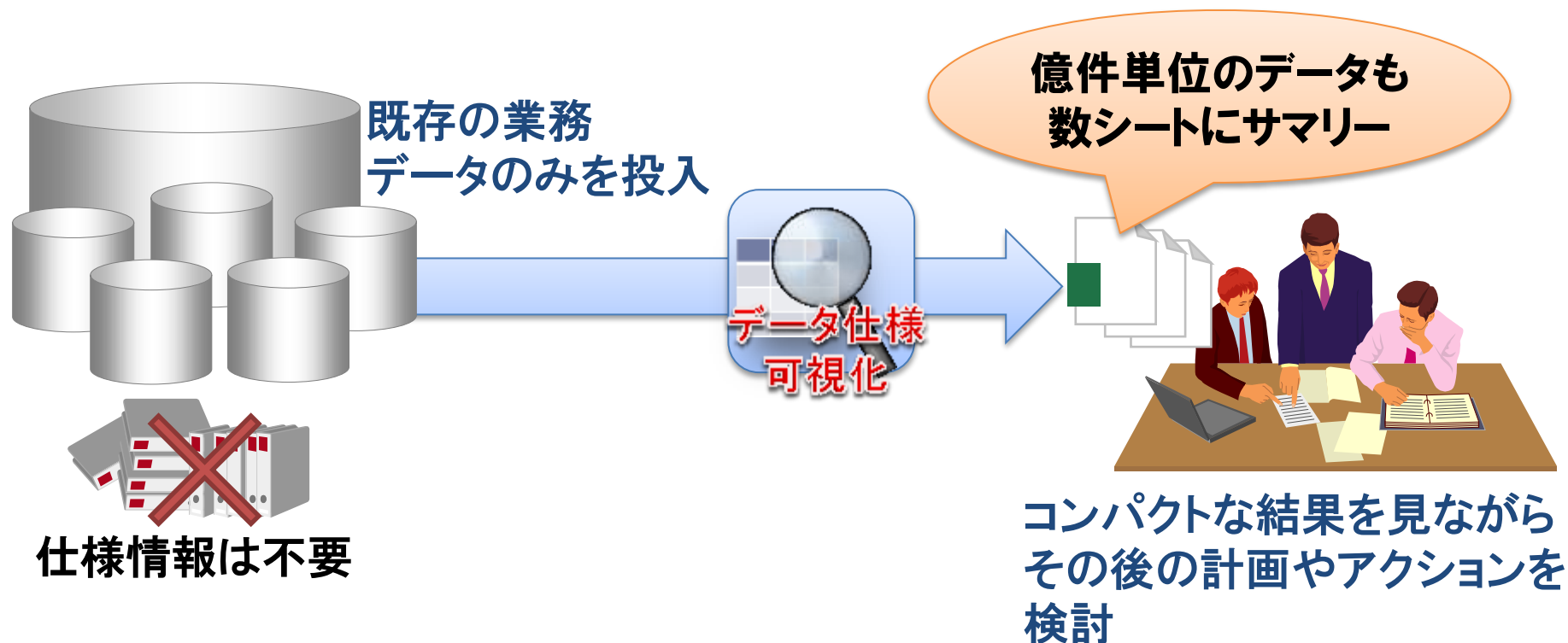
データ量が
膨大

大量のデータを
素早くサマリー

調査計画が
立てづらい

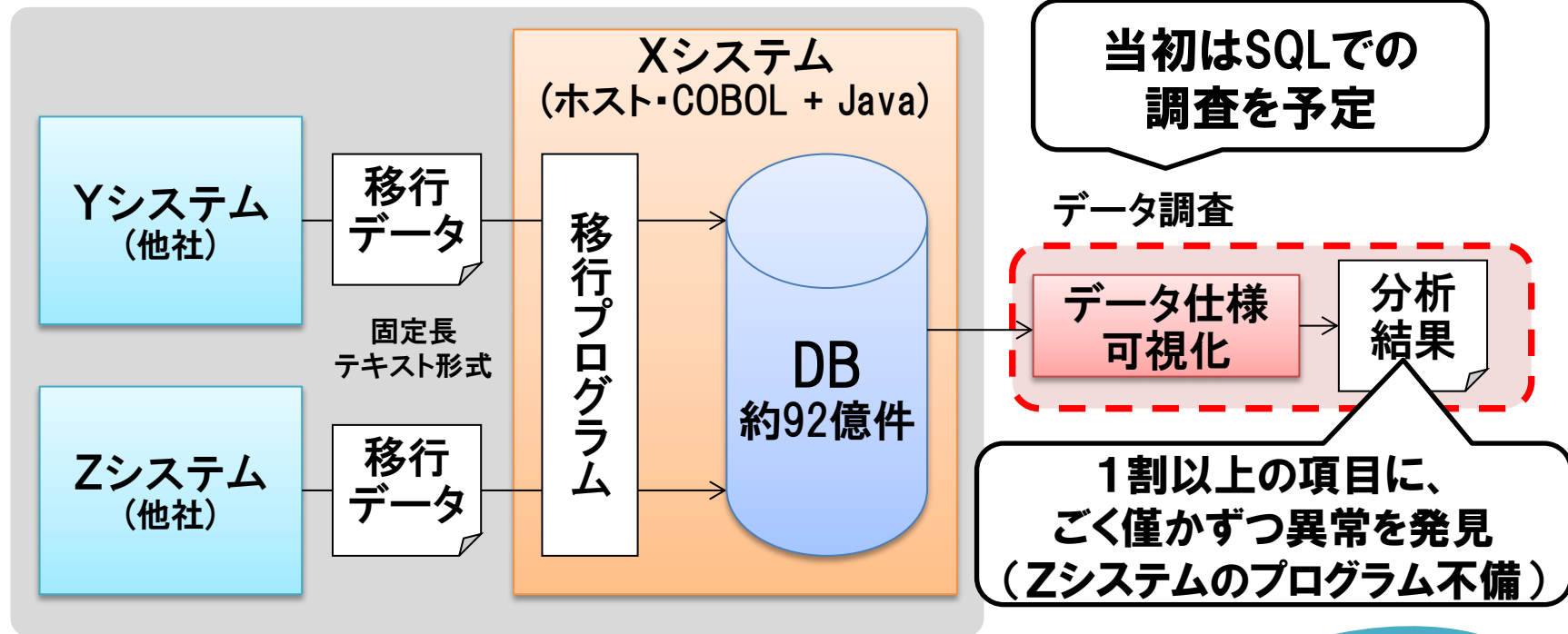
計画に使える
分析レポート

ポイント② 大量データを素早くサマリー



大量のデータもすぐに、コンパクトなレポートにサマリー。
全データを網羅的に分析し、僅かな異常データも可視化。
情報が集約され、短時間での全体俯瞰が可能に。

【課題】 他システム（ホスト）からの移行データに想定外データが含まれていると運用開始後に不具合が生じる恐れがある。
しかし、データ量が膨大で人手による洗い出しが困難。



92億件のデータに対する分析結果を
僅か8時間で確認。

本番トラブルを
未然に防止!

頼れない
仕様書

データから
仕様を推測

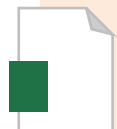
データ量が
膨大

大量のデータを
素早くサマリー

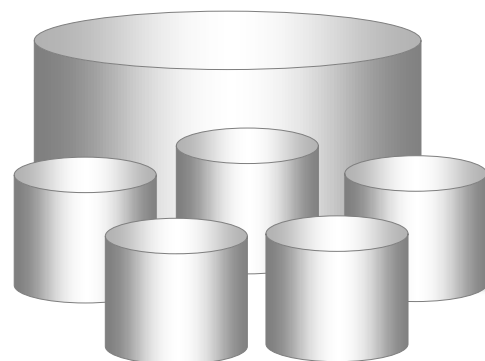
調査計画が
立てづらい

計画に使える
分析レポート

サマリー(問題点と推奨アクションの一覧)



- ✓ 「社員ID」項目に0.1%だけデータ形式の異なるデータがあります。
⇒仕様通りか確認して下さい。
- ✓ 「社員」マスタに不足が見受けられます。
⇒不足一覧を参照し、必要なデータを追加投入して下さい。



サマリー

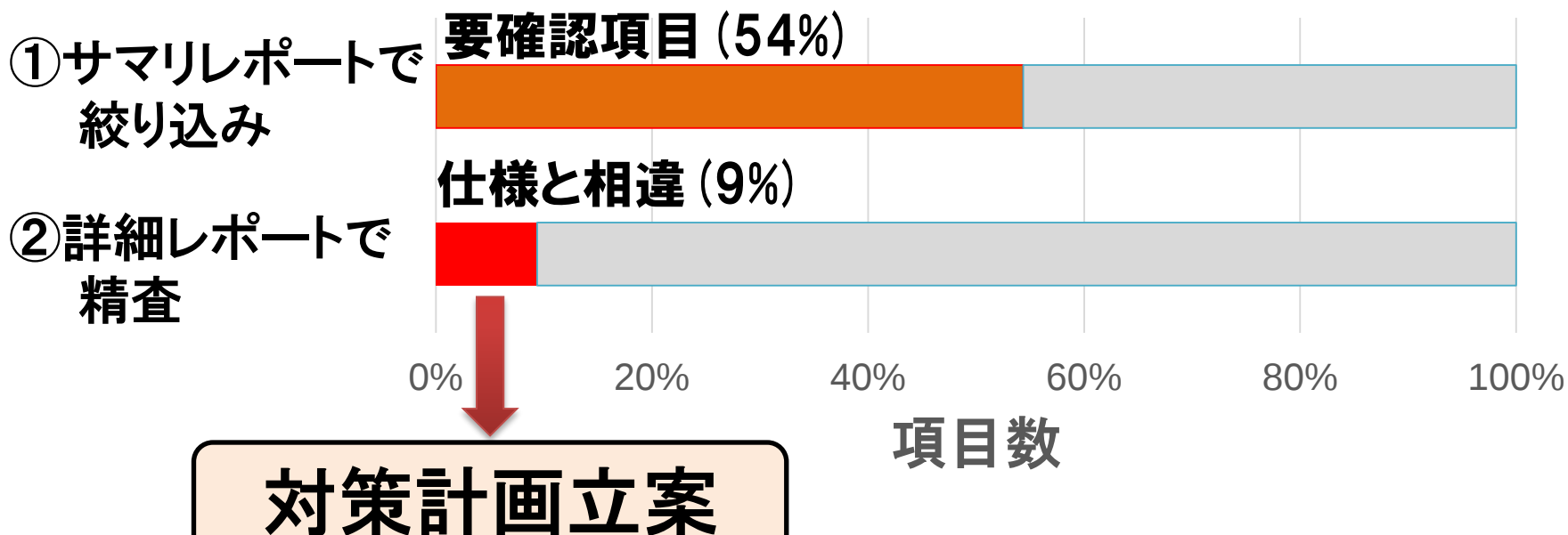


個別詳細



**確認すべき問題点や必要なアクションを一覧形式で明記。
TODOの全体像が分かり、計画立案やゴール設定が容易に。**

- 2システムの統合プロジェクト。
- 計画段階で移行データを可視化し、移行計画・移行仕様に反映。



僅か**1日**で移行データの可視化～確認まで実施。
1割近くの項目に仕様書との相違を発見。

トラブルを
未然に防止!

頼れない
仕様書

データから
仕様を推測

データ量が
膨大

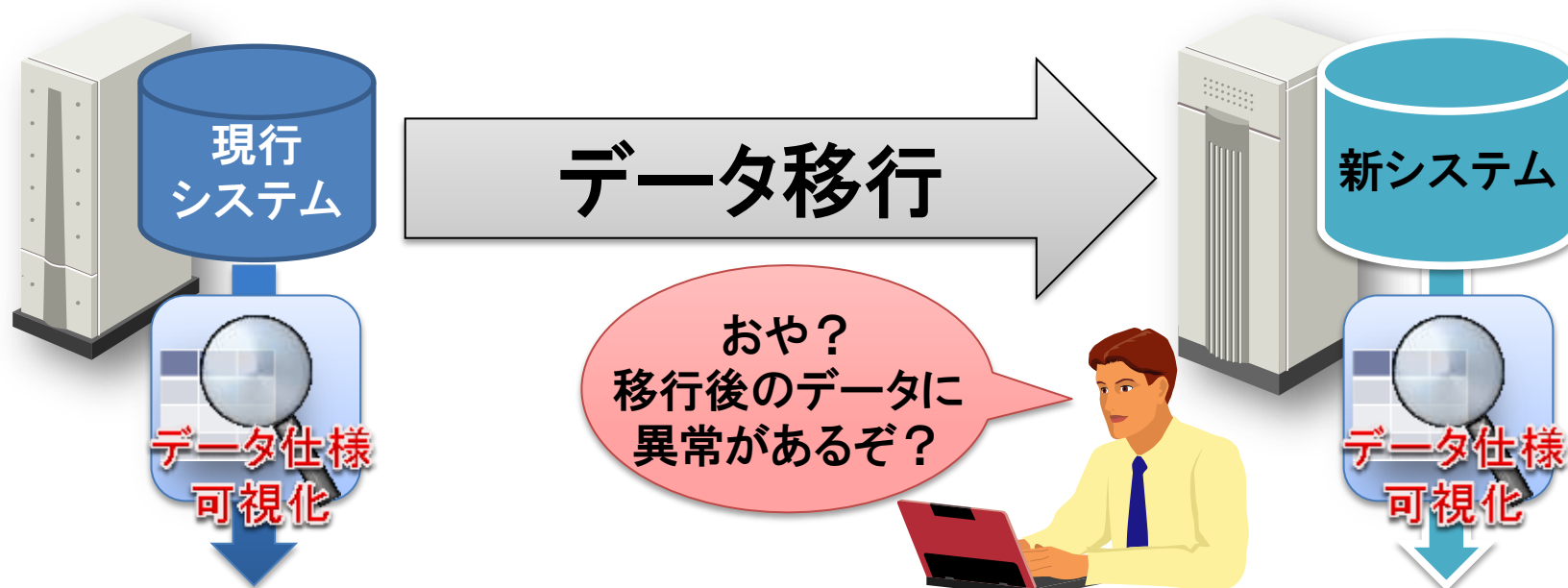
大量のデータを
素早くサマリー

調査計画が
立てづらい

計画に使える
分析レポート

現行データの全体像を素早く俯瞰でき、
計画や設計の抜け漏れ、移行時のもぐら叩きを防止

■ 現新比較テストの確認材料として活用



項目名	パターン	正規表現	頻度	実数	種数
社員番号	コード	^[0-9]{9}\$	71.0%	4145	4145
		^[0-9]{10}\$	29.0%	1710	1710

比較

項目名	パターン	正規表現	頻度	実数	種数
社員番号	コード	^[0-9]{9}\$	71.0%	4145	4145
		^[0-9]{10}\$	27.7%	1694	1694
		<<空>>	0.3%	16	1

10桁のコードデータが
一部、空に置き換わっている

実は...

移行プログラムが想定外データを
空に置き換える仕様だった！

テストでの活用例② テストデータ自動生成

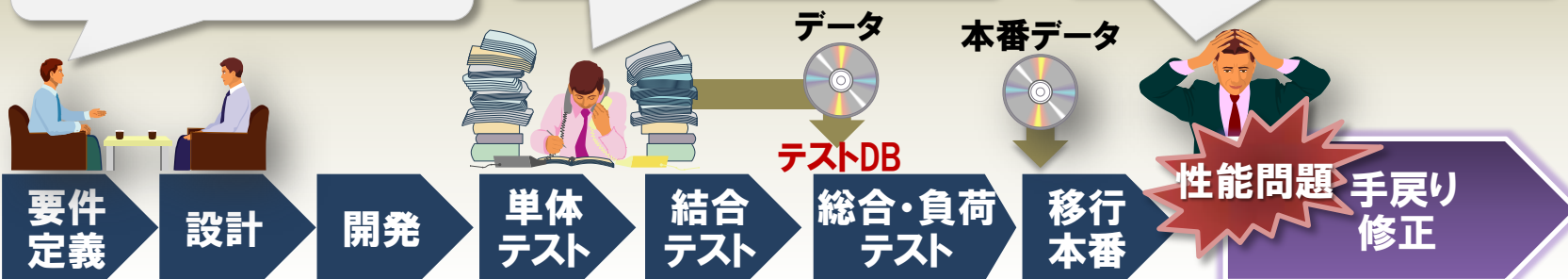
- データ仕様可視化 …既存データを分析し、データ仕様と異常データを可視化
- テストデータ生成 …可視化結果から逆に、データを自動生成してテストに活用

ありがちな例

総合テストまでは仮データで。
仕様書に従い作成して下さい。

バリエーションや制約が多いなあ
仮データの用意も一苦労だ…

インデックスが効かない！
性能が全然出ないぞ！？



データ仕様可視化 テストデータ生成

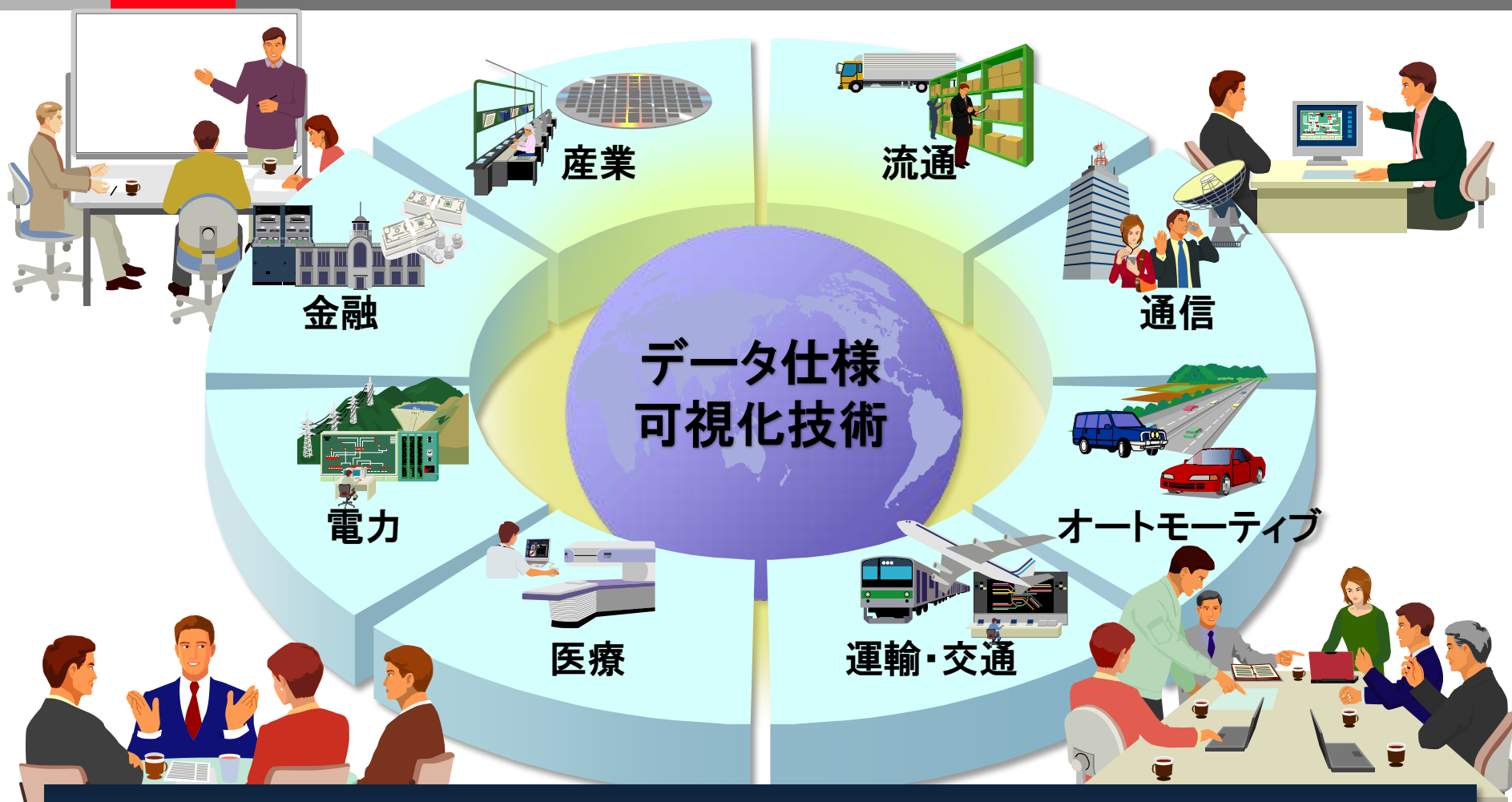
異常データと思われる
パターンが少なくともこれだけあります。

擬似データ検証で早期に
不良を検出、手戻りを防止！

認識していなかった…
別途、対策を計画しましょう。

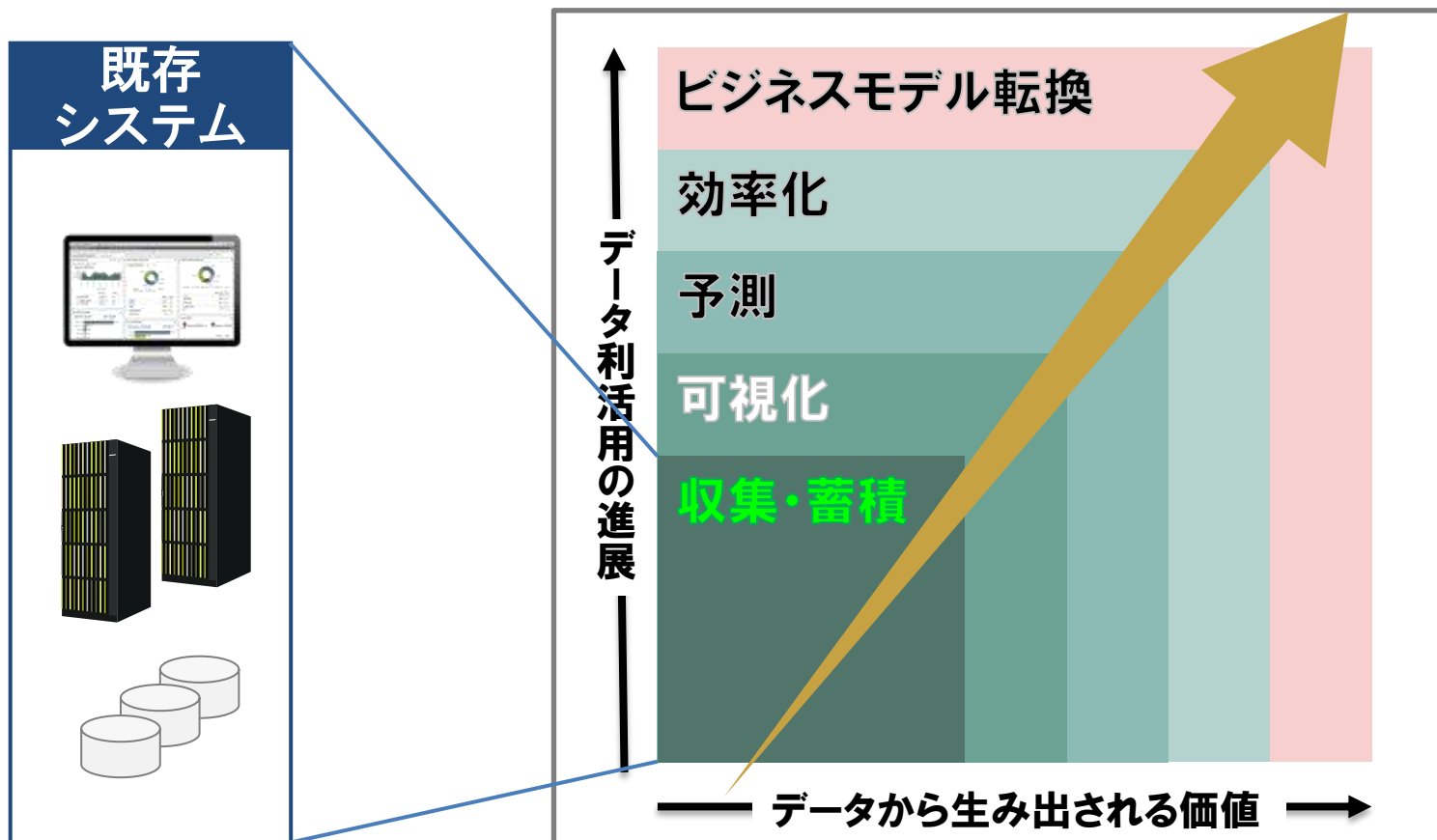


早期実態把握 &
実態に即した
テストで
「モグラ叩き」の
リスクを低減



高度な可視化技術を専門技術者以外でも
使いこなせるようにし、様々な業種のプロジェクトで広く活用

4. 「レガシー」を次の時代へ ～ データの利活用～



(出典)総務省「IoT時代におけるICT産業の構造分析とICTによる経済成長への多面的貢献の検証に関する研究調査」(平成28年)

**長年蓄積されてきたデータから価値を引き出し、
ビジネスに活かすことが重要**

コスト削減

生産性向上

収益増加

新たな事業価値を創造する・ビジネスモデルを変革する

モデル化・仮説設定 >

ナレッジ

分析・予測 >

インテリジェンス

検索・可視化 >

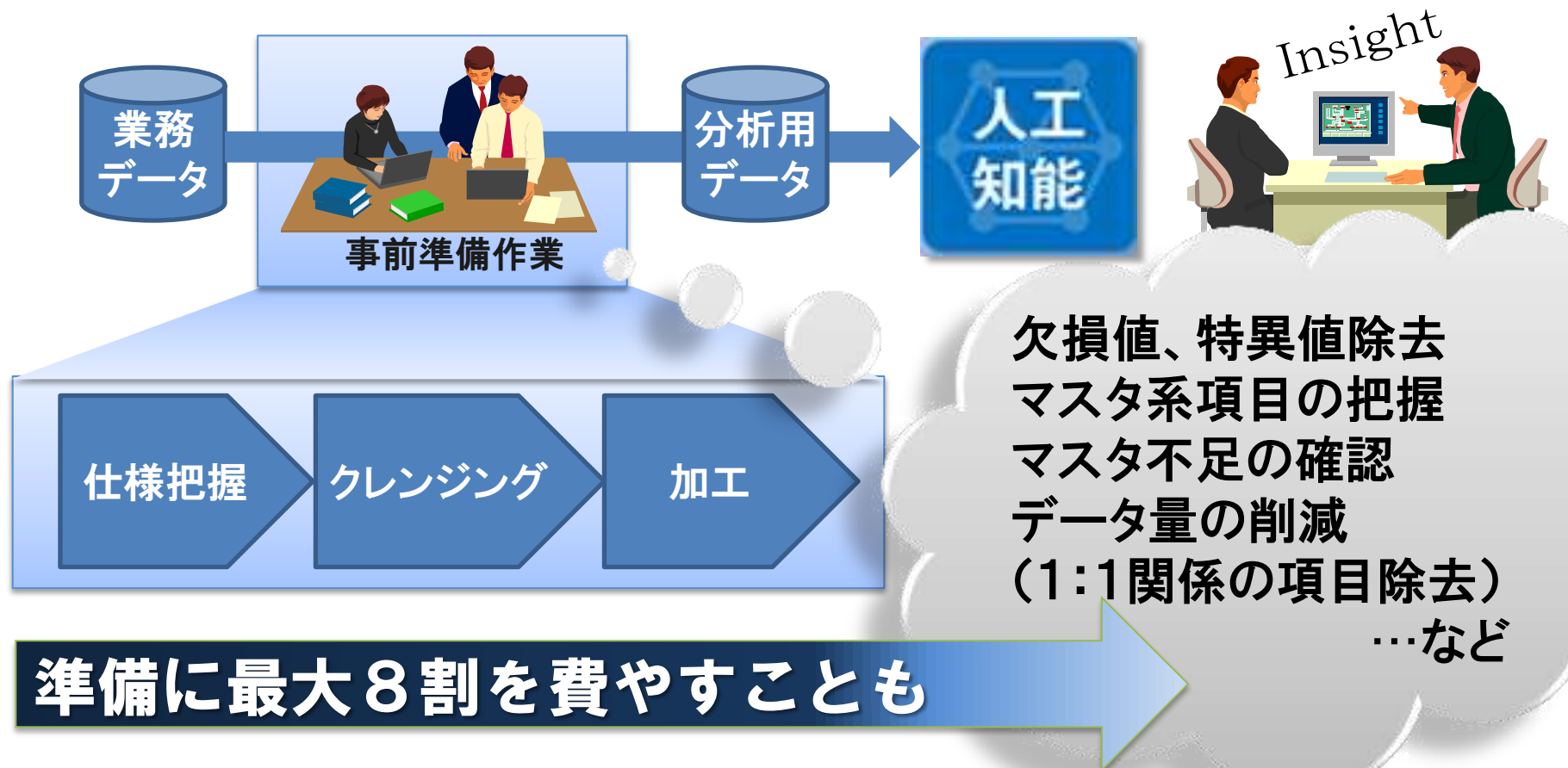
インフォメーション

データ収集・蓄積 >

データ

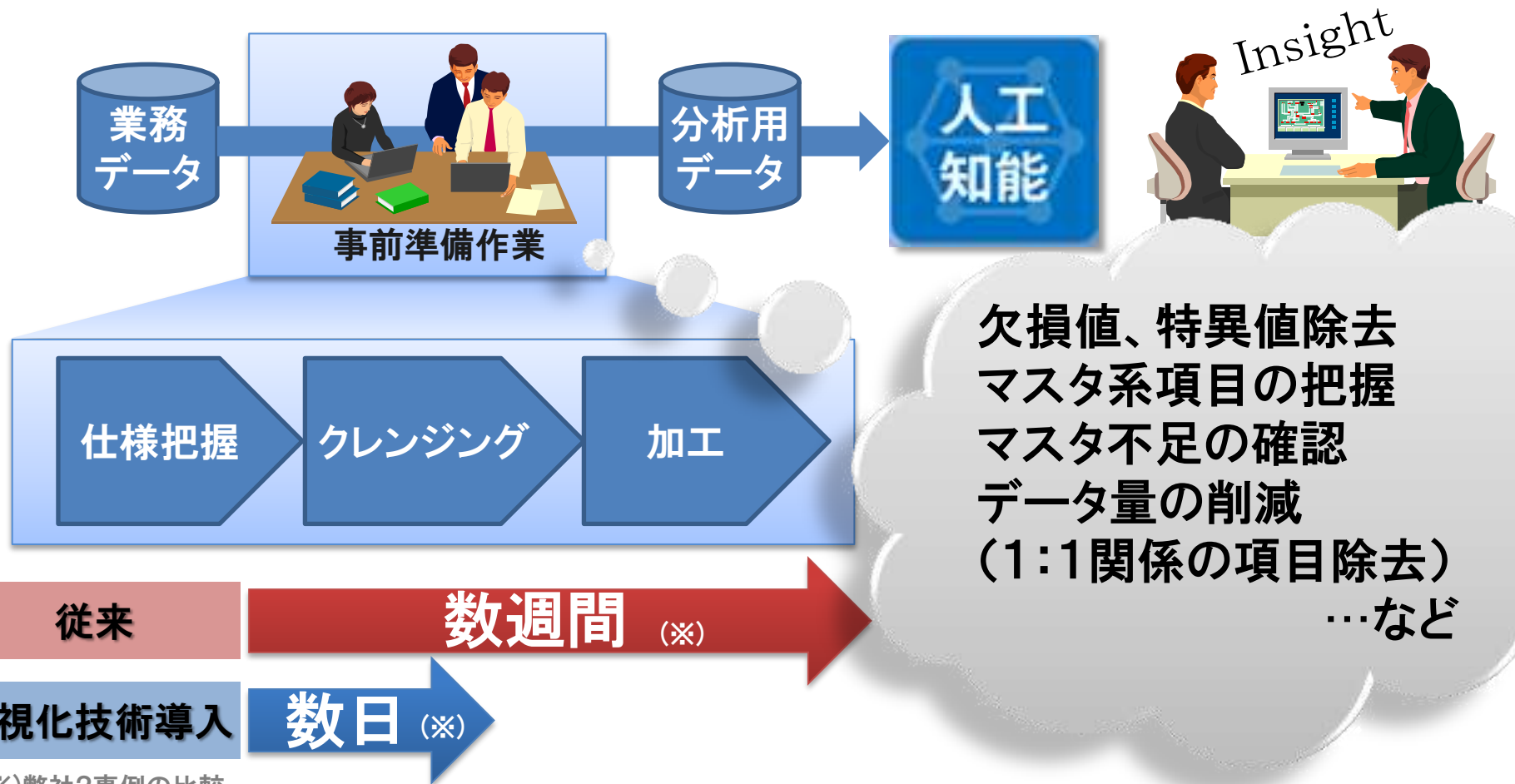
人の
スキル
ノウハウ
技術

既存データの利活用には長期間の準備が必要



既存データを利活用可能にするには地道な準備作業が必要。
特にデータの仕様把握は大きな負担。

事例 データ利活用へのデータ仕様可視化技術適用



人工知能へのインプットに必要な事前作業を大幅に効率化。
アクションに向けた議論をより素早く開始。

膨大な量のデータが 世界中でつながる IoTの世界

5.5M

2016年
1日あたりに新しく
インターネットに
つながるモノの量
(Gartner)

6.4B

2016年
世界中でネットワークに
つながるモノは2015年よりも
30%増加
(Gartner)

1T

2025年
インターネットにつながっ
ているセンサーの数
(WEF)

20.8B

2020年
インターネットにつながる
モノの量
(Gartner)

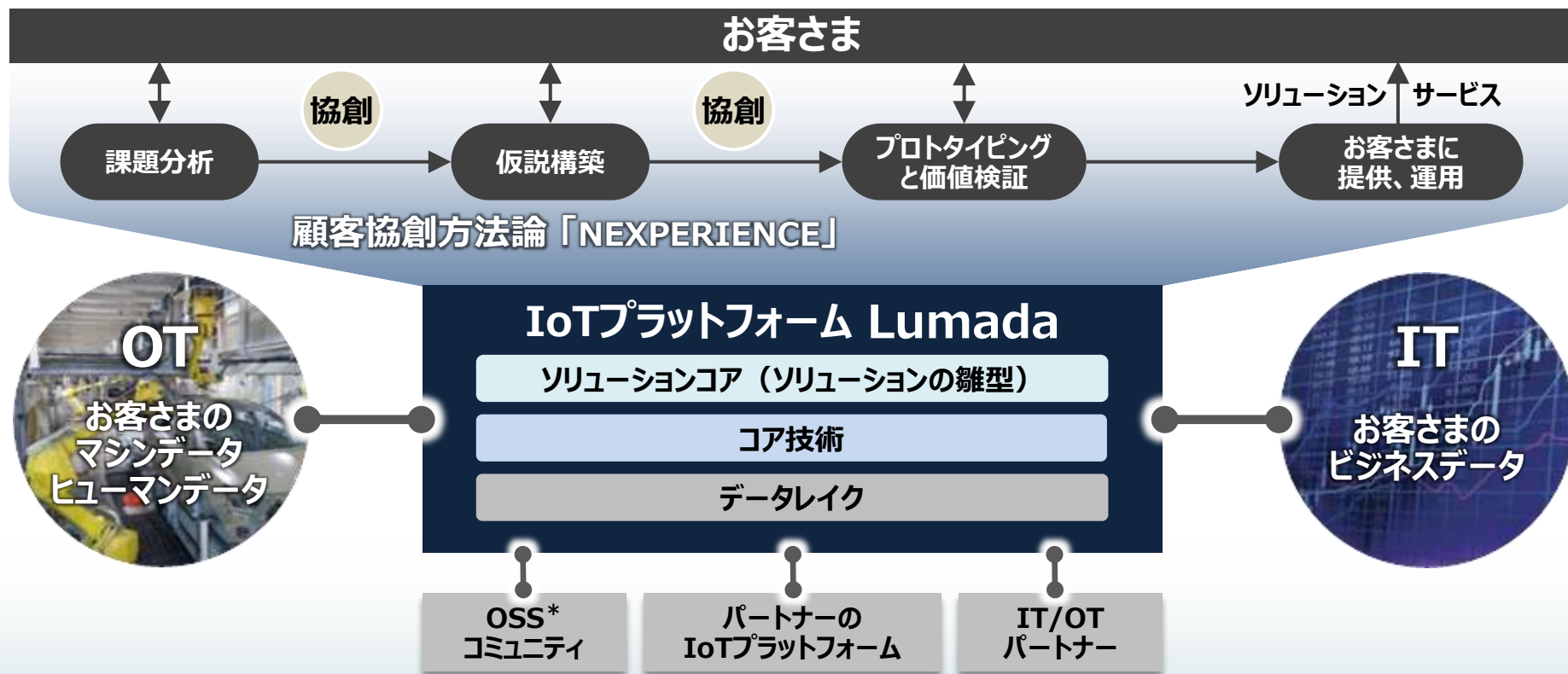
\$1.3T

2019年
世界でのIoT消費予測額
(IDC)

IDC, Worldwide Internet of Things Forecast Update, 2015-2019, Doc #US40983216, Feb 2016.

Gartner Press Release, November 2015, <http://www.gartner.com/newsroom/id/3165317> Gartner, Inc. forecasts that 6.4 billion connected things will be in use worldwide in 2016, up 30 percent from 2015, and will reach 20.8 billion by 2020. In 2016, 5.5 million new things will get connected every day.

お客さまやパートナーのシステムとつないでソリューションを迅速に協創



*OSS: Open Source Software



5. まとめ

保有する資産を活かしたい場合、モダナイゼーションはシステム再構築の有効な選択肢です。

しかし、その上で大事なのは、システム再構築後に

『事業発展に貢献できるシステム』

へと成長させていくことです。

そのためには、資産／業務の特性に合わせた適切な方法を選択してモダナイゼーションを行うことが成功の鍵です。

付録. 関連サービス紹介

ITモダナイゼーションの技術の分類

カテゴリ	説明
リファクター	機能をそのままに現行のコードを整理する 性能や保守性を向上させる為に、現行アプリケーションの設計や実装を見直すことです。マイグレーションの文脈においては、必要な資産の範囲を見極める「棚卸」が重要です。
リドキュメント	現行システムを可視化、ドキュメントを再整備 現行システムの資産分析やヒヤリングを通じて、現行システムの仕様、設計情報を可視化し、ドキュメントを整備する技術です。
リビルド	既存システムの業務仕様を再実装する方式 既存システムの仕様を生かし、再実装する方式です。移行方式としては最も自由度が高く、非機能要件に対し、最適な設計／実装の選択が可能です、最もコストがかかります。
リライト	変換ツールでコードを他言語に書き換える方式 既存のプログラムを変換ツールで他言語に書き換えて、移行を実施する方式です。開発・保守の環境含めを新しい言語に統一したい要望に応えることができます。
リホスト	既存システム資産をそのまま利用する方式 互換性のあるミドルウェアやエミュレーションを利用して既存アプリケーションをそのまま利用する方式です。移行費用を押さえつつ短期にインフラを刷新したい要望に応えることができます。

リドキュメント系のサービス

本ご紹介

■ データ仕様可視化サービス

データ仕様可視化ツールにより、システムで使用されているデータを分析。この情報をもとに、設計／開発作業の品質向上やテスト効率向上に向けたデータに係る課題解決を支援します。

■ プログラム仕様可視化サービス

10年以上の実績をもつ資産解析ツールにより、現行システムの資産を解析し、可視化情報、関連情報を作成します。可視化情報は、専用ブラウザで参照できる環境を提供します。

■ 業務仕様可視化サービス

プログラム仕様可視化をベースに情報収集、ドキュメント分析を実施、業務レベルの情報を関連付け、業務レベルの仕様を作成するサービスです。業務とシステムを関連付けながら調査を進めることで、抜け漏れを防止し、整合性の確保します。

リファクター系のサービス

■ 棚卸サービス

現行システムのプログラム資産を解析して得られる資産間の関連情報をもとに、不要なプログラム資産を洗い出すサービス。

■ プログラム類似性分析サービス

共通の「語」の並び(コードクローン)を、アプリケーション全体から抽出することにより、設計／実装の改善をサポートするサービス。

リHOST系のサービス

■ マイグレーションサービス

日立COBOL2002をはじめ、互換性を提供する各種ミドルウェアの組み合わせにより移行を実現するサービスです。

リライト系のサービス

■ 言語リニューアルサービス

言語変換の技術を用い、低コストに移行を実現したサービスです。
(PL/I→COBOL、COBOL→Java等)

リビルド系のサービス

■ COBOLマイグレーションサービス

現行分析から構築までサポートすることにより、COBOLからJavaへの再構築型の移行を実現します。BRMSの導入で保守性を向上すると共に、利用状況进行分析し業務上の課題を抽出することで、移行後の継続的な業務改善を支援します。

その他のモダナイゼーション技術

■ 個別にお問い合わせください。

モダナイゼーション技術についての問合せ

- 株式会社 日立製作所 サービスプラットフォーム事業本部
事業推進本部 ミドルウェアソリューション窓口

E-mail: midsol-p@itg.hitachi.co.jp

関連するURL

- アプリケーションライフサイクルマネジメント

http://www.hitachi.co.jp/mid-sol/search_s/modernization/alm/

- ITモダナイゼーション ソリューション

http://www.hitachi.co.jp/mid-sol/search_s/modernization/

- COBOL2002

<http://www.hitachi.co.jp/Prod/comp/soft1/cobol/index.html>

- COBOLマイグレーションサービス

<http://www.hitachi.co.jp/cobol-migration/>

- HITACHIは、株式会社 日立製作所の商標または登録商標です。
 - OracleとJavaは、Oracle Corporation 及びその子会社、関連会社の米国及びその他の国における登録商標です。
- その他記載の会社名、製品名は、それぞれの会社の商標もしくは登録商標です。

END



地味にスゴイ「データ仕様可視化技術」 ～その効用と適用事例～

2017/04/20

**株式会社日立製作所 システム&サービスビジネス統括本部
アプリケーションサービス事業部
サービスソリューション本部 サービス統括部**

清水 薫