第2回 AI/Analyticsカンファレンス

データを事業に活かすために必要なデータ基盤とは

TOSHIBA

東芝デジタルソリューションズ株式会社
新規事業開発部 シニアエキスパート 望月 進一郎
2021年8月27日

東芝デジタルソリューションズ

2021年4月14日現在

（株）東芝

代表執行役社長
CEO
綱川 智

エネルギーシステムソリューション事業領域

東芝エネルギーシステムズ（株）

インフラシステムソリューション事業領域

東芝インフラシステムズ（株）

ビルソリューション事業領域

東芝エレベータ（株）　東芝ライテック（株）　東芝キヤリア（株）

リテール＆プリンティングソリューション事業領域

東芝テック（株）

デバイス＆ストレージソリューション事業領域

東芝デバイス&ストレージ（株）

デジタルソリューション事業領域

東芝デジタルソリューションズ（株）

- 東芝デジタル＆コンサルティング（株）

(株)東芝　執行役上席常務
東芝デジタルソリューションズ(株)
取締役社長
島田　太郎

東芝デジタルソリューションズ　会社概要

名称	東芝デジタルソリューションズ株式会社(英文名　Toshiba Digital Solutions Corporation)
本社所在地	神奈川県川崎市幸区堀川町72番地34
設立年月日	2003年10月1日
取締役社長	島田 太郎
事業内容	**システムインテグレーション**及び**IoT/AIを活用**したICTソリューションの開発・製造・販売
資本金	235億円（東芝100%）
売上高	2,217億円（連結／2021年3月期）
関係会社	8社（国内7社、海外1社）
従業員数	(単独) 3,816人 (連結) 8,247人　(2021年3月現在)

【関係会社】

- 東芝デジタル＆コンサルティング(株)・・・ デジタルビジネス戦略コンサルティング
- 東芝情報システム（株）・・・・・・・・・・ SI、組込、半導体エンジニアリング
- 東芝ITサービス（株）・・・・・・・・・・・・ IT系保守、運用サービス
- 日本システム（株）・・・・・・・・・・・・・・ SIソフト開発
- 中部東芝エンジニアリング（株）・・・・ 半導体エンジニアリング
- 九州東芝エンジニアリング（株）・・・・ 地域対応SI、半導体エンジニアリング
- イー・ビー・ソリューションズ（株）・・・・ 各種コンサルティング
- 東芝瀋陽情報システム社・・・・・・・・・ 中国システム販売

本日のアジェンダ

- ✓ データを事業に活かす上での課題
- ✓ データを活かす手法
- ✓ データ基盤に求められる要件
- ✓ OSSを使ったデータ基盤
- ✓ データ基盤向けデータベースの例（GridDB）
- ✓ データ基盤とAutoMLを連携させたAIソリューション（ケーススタディ）
- ✓ これからのデータ基盤

属人的な勘や経験からデータに基づく経営へ

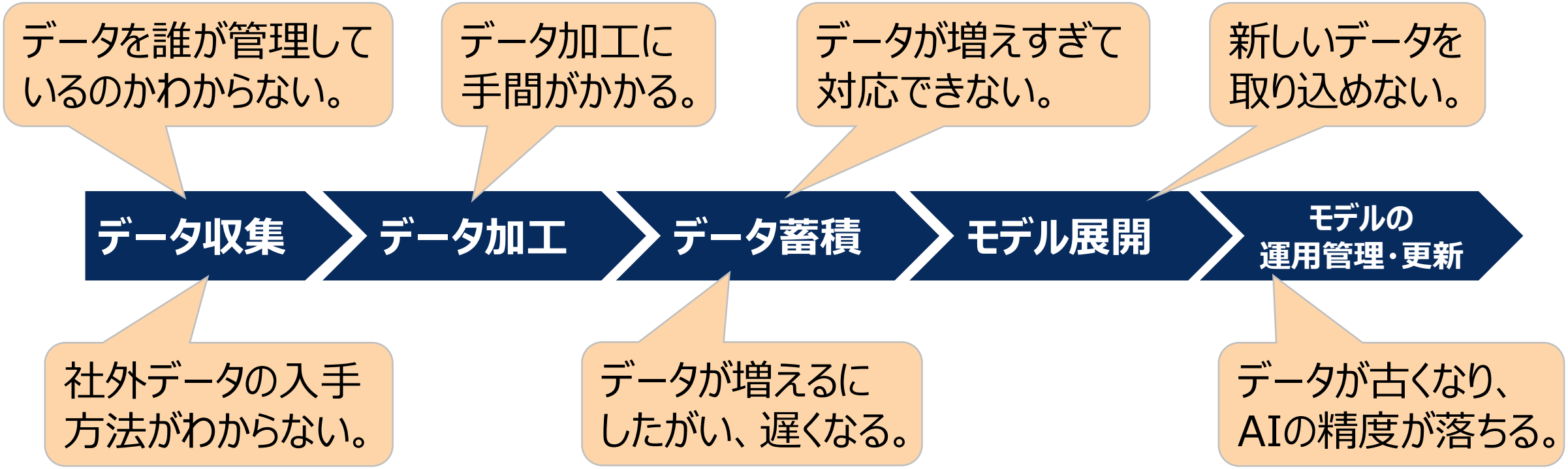
AIやビッグデータ分析を使って、データを事業に
活かす試みが盛んに行われている。

しかしそのような試みは短期的にはうまく行っても、
長続きしないといった話もよく聞く。

なぜだろう？

AIプロジェクトにおけるデータ管理の課題

とりあえずデータを集め、AIを活用したシステムを稼働させることはできた。
しかしそれが続かない。



継続的に精度よく運用するには、半自動化されたデータ基盤が必要！

データを活かす手法

レポーティング分析	データを抽出・分類して、データの並べ替えやグラフ化（見える化）を行う。BIツールによるダッシュボードなどもこの一部。
アドホック分析	定期的かつ定型的なレポート分析と違い、そのときに必要な分析を行う。事前にデータが整備されていると素早く対応できる。
AI・機械学習	機械学習やディープラーニングなどのAI技術を活用して分析を行う。トライアンドエラーによるモデル作りを行うことが多い。
モニタリング・監視	ログデータやストリーミングデータを使って、リアルタイムに異常検知などの監視を行う。
ビッグデータ分析	大量かつ多様なデータをクロス集計やクラスター分析といった手法を使って分析を行う。

集めるべきデータ

業種や予測したい事象により集めるべきデータは異なってくるが、共通して言えること。

分散と偏りのないデータが必要

たとえば全国の分析を行うときに、特定の地域に偏ったデータを使うと分析の精度が落ちる。全国のデータをバランスよく用意するべき。

季節性・社会性を常時反映するための継続的なデータの収集を行うべき

継続的にデータを集め、最新のデータを使用することで、大きな社会変動にも対応可能になる。

自社のデータだけでなく、外部のデータも必要

自社特有のデータに、他社のデータや人流データ、気象データなど外部のデータを組み合わせることにより、精度の高い分析が可能になる。

データ基盤

集めたデータをデータ活用にスムーズに導くのがデータ基盤

データ基盤を用意するにあたって

- データ基盤は長期にわたって使用するもの。
- 分析手法ごとにデータ基盤を用意するのは非効率。
- 今必要な仕様を満たすだけでなく、将来の活用も想像しながら検討する必要がある。
- できるだけシンプルな構成が望ましい。

どんなことにデータを活用したいだろう？

データはどれくらいの規模になるだろう？

どんなツールと連携させることになるだろうか？

運用コストはどれくらいが適切だろうか？

データ基盤に求められる要件

1. 様々なデータを扱える。
2. 高速に分析できる。
3. 各種収集・加工・分析ツールを包含する、または連携できる。
4. データが急増しても取りこぼしなく収集できる。
5. 蓄積するデータが増えていっても柔軟に拡張できる。
6. 運用・メンテナンスが簡単。
7. クラウドとオンプレミスに対応できる。
8. リアルタイムで分析できる。

最近、強まっているニーズ

OSSで実現するデータ基盤

データ基盤に関係するOSS

©日本OSS推進フォーラム

日本OSS推進フォーラム作成「OSS鳥観図 2021年版」から抜粋。
http://ossforum.jp/index.php/choukanzu-wg/

OSS活用で注意すべきこと

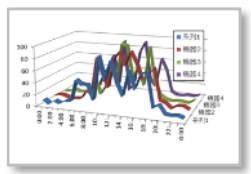
- OSSを使いこなすだけの技術力が必要。
- OSS間をつなぐ開発や検証が必要。
- 最新技術が次々に出てくる。逆に開発が停滞するOSSもある。OSSの見極めが必要。

データ基盤向けデータベース GridDB

ビッグデータ・IoTシステムに最適な超高速スケールアウト型データベース。
従来のデータベースでは不可能だったビッグデータのリアルタイム分析が可能に。

5つの特長

- IoT指向のデータモデル
- 高性能
- スケーラビリティ
- 高い信頼性と可用性
- NoSQLとSQLデュアルIF




分析アプリ　BI/BA　他のシステム　他のデータベース

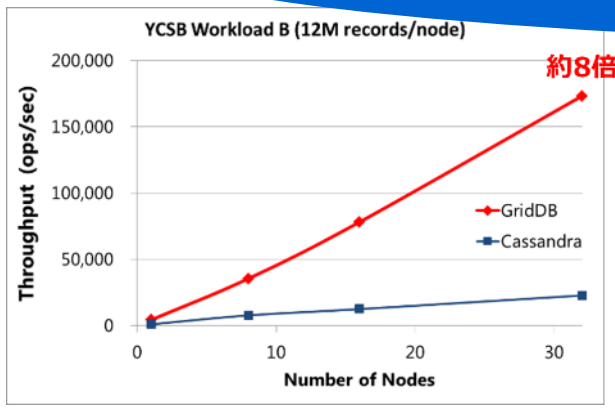
SQLインターフェイス … 分析や他システムとの連携にはSQL

GridDB

GridDB Server　GridDB Server　GridDB Server　GridDB Server

NoSQLインターフェイス … 大量高頻度のデータ収集にはNoSQL

ビッグデータ
IoTデータ



データ基盤向けクラウドデータベースサービス GridDB Cloud

高性能スケールアウト型データベースGridDBをコアに持つ
AIやビッグデータなどの分析を支えるデータベースサービス

POINT 1

パブリッククラウドで稼働するマネージドサービス

データ量や処理量の変動に柔軟に対応。
運用・監視は当社が一括して実行。

POINT 2

クラウドネイティブアプリと簡単・高速に連携

JDBCやWebAPIを介して簡単にデータにアクセス。
アプリを同じクラウドに配置すればオンプレミスと同様な高速アクセスが可能。

POINT 3

データ収集やデータの見える化機能が充実

FluentdやAzure IoT Edgeと連携したデータ登録や、Grafanaによる
見える化が可能。SQLを介して分析ツールとの連携が可能。

データ基盤の要件と GridDB Cloud

データ基盤の要件	GridDB Cloudの場合
1. 様々なデータを扱える。	**GridDB Cloudは構造化データの扱いに優れています。 非構造化データはクラウドストレージを併用します。**
2. 高速に分析できる。	**インメモリ処理を主としていますので、高速に分析できます。**
3. 各種収集・加工・分析ツールを包含する、または連携できる。	**FluentdやEmbulk、Azure Functionsなどとの連携機能が用意されています。またSQLを介して様々なツールと連携できます。**
4. データが急増しても取りこぼしなく収集できる。	**高速なNoSQLを使うことにより、大量のデータも取りこぼしなく収集します。**
5. 蓄積するデータが増えていっても柔軟に拡張できる。	**スケールアップ、スケールアウト両方に対応できます。**
6. 運用・メンテナンスが簡単。	**収集から蓄積、分析ツールとの連携までを担い、データ基盤の構成がシンプルになります。**
7. クラウドとオンプレミスに対応できる。	**パブリッククラウドはもちろん、オンプレミスでの構築も可能です。**
8. リアルタイムで分析できる。	**収集と分析を1つのデータベースで対応でき、リアルタイムで分析できます。次ページ参照。**

要件8. リアルタイムで分析できる。

従来のデータ基盤

収集データ → データ蓄積用DB → バッチによるデータ整形 → データ分析用DB → 分析

ODS※1 → ETL※2 → DWH※1

2つのデータベースを管理するのは面倒だ…。
リアルタイムに分析できないじゃないか!?

※1 ODS (Operational Data Store)：収集データを一時的に保持するデータベース
※2 ETL (Extract/Transform/Load)：データの整形、統合を専門に行うツール
※3 DWH (Data Warehouse)：情報を時系列に整理して保管するデータベース

GridDB Cloudの場合

収集データ → 蓄積＆分析用DB → 分析

POSデータ / 気象データ / 店舗データ / 人流データ …

NoSQLによる高速な収集

GridDB

SQLによる高度な分析

リアルタイムの分析が可能に！

クラウドデータ基盤とAutoMLを連携させたAIソリューション

DATAFLUCT cloud terminal. × GridDB Cloud

■概要

DATAFLUCT cloud terminal.

マルチクラウド環境で最適なAutoMLモデルを簡単・スピーディーに構築できる機械学習プラットフォーム。

- 簡単に高精度の機械学習モデルをつくれるAutoML
- マルチクラウド対応
- 簡単デプロイ機能（オンライン処理、バッチ処理）
- 信頼性の高いプラットフォーム
- 月額固定のシンプルな料金メニュー

詳細は p.18 へ

AutoML

高度な機械学習モデルを、専門家なしに簡単に構築できる。

×

時系列データベース

様々なリアルタイムデータを処理・蓄積できる。

- 店舗ごとに最適な予測モデルを構築できるサービス。
- DATAFLUCTと東芝デジタルソリューションズのもつサービスを組み合わせて共同開発。

TOSHIBA GridDB

時系列データベースサービス
『GridDB Cloud』

人流データや気象データなど大量のデータを蓄えつつ、オンデマンド分析を可能にするクラウドデータベースサービス。

- 時々刻々発生するデータを効率よく蓄え、分析する時系列データモデル
- 大量のデータに対応可能な高い処理能力と拡張性
- 多くの社会インフラシステムで使用された実績が示す高い信頼性
- 運用を気にしないですむクラウドマネージドサービス

ケーススタディ：食品廃棄物の発生量を半減させよ！

悩みを抱えている人　**出板（デイタ） 克洋（カツヒロ）さん（42歳）**

- 現在、大手食品スーパーマーケットの本部で発注業務を担当。
- 食品廃棄物の発生量を数年以内に半減させるプロジェクトにアサイン中。
  各店舗のID-POSデータを分析し、日々対策を検討している。
- その一方で、惣菜市場の成長に合わせた売上向上施策の立案も期待されており、
  両立に頭を悩ませている。

どのような問題を解決する必要があるか？
└ **各店舗で発生している食品廃棄物の発生量を50%削減と惣菜市場における売り上げ20%アップ**

ビジネス上で測定すべき事象は何か？（品質、コスト、顧客満足度…etc）
└ **各店舗の生鮮食品・惣菜の「販売数」「売り上げ」「値引きした商品数」「廃棄した商品数」**

機械学習は適切なアプローチか？（分類、予測、クラスタリング…etc）
└ **時系列予測により、1時間単位の来客数予測する　（ID-POSは導入済）**

どのようなデータを利用できるか？（利用可能なデータで十分か？）
└ **店舗毎のID-POSデータ、気象データ、地域イベント情報、人流データ…**

何を目標とするか？（ビジネス目標、顧客価値の最大化、予測精度…etc）
└ **当日内の来客数予測精度90%以上**

Step1. データ準備

まずはすぐに利用可能なID-POSデータを利用して来客数の予測モデルを作成に着手。
特性の異なる以下の店舗を対象に、ID-POSのデータから時間帯別の来客数情報を抽出。

■ 店舗データ（店舗ごとの特性情報を含むデータ）

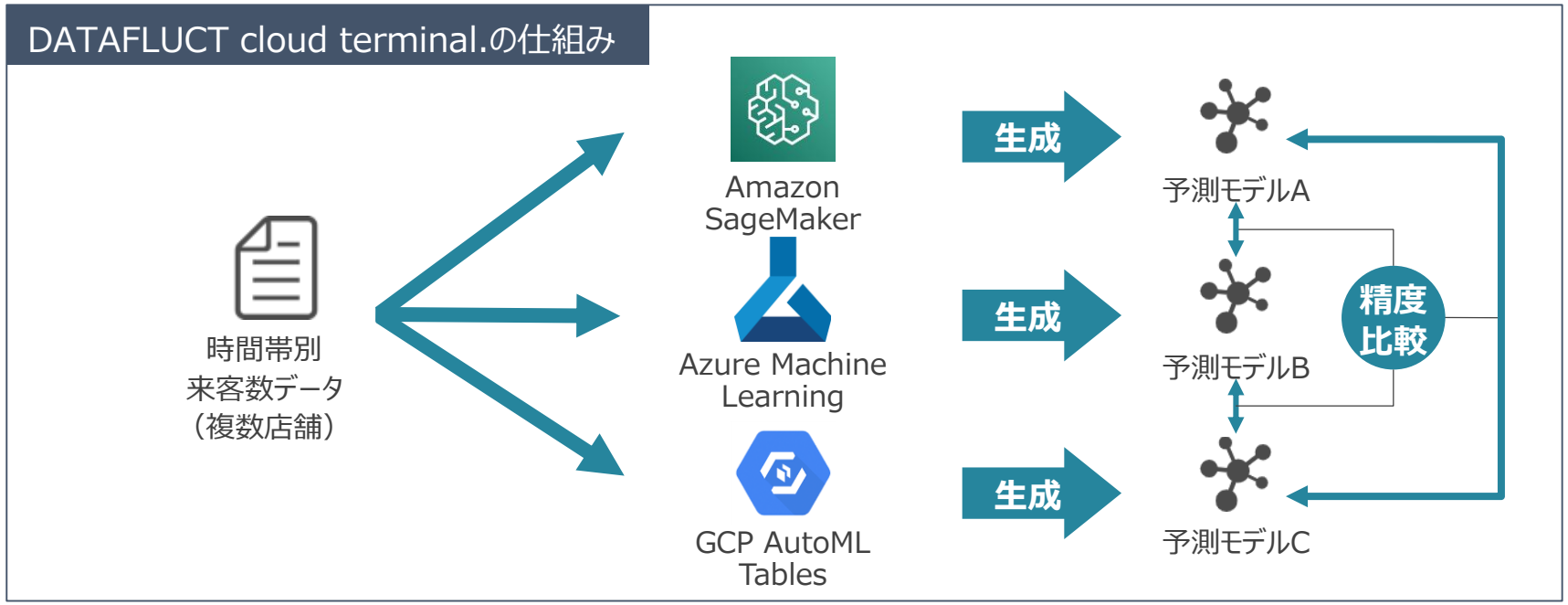
店舗	気象観測の位置	常住人口	昼間人口比率	駅からの距離	客タイプ	店舗面積
店舗1	東京	多い	増	普通	近所の住民、高級志向	中
店舗2	東京	普通	増	近い	若い主婦、学生	小
店舗3	東京	普通	減	普通	主婦	中
店舗4	東京	少ない	減	遠い	近所のお年寄り、買い溜め	大
店舗5	大阪	普通	減	普通	主婦	中

■ ID-POS データ（2021年2月の1ヶ月分のデータを時間帯別に抽出）

date	time	customerID	storeID	itemCategory	itemName	price
2021/2/16	9	294	1001	6	非食品3	200
2021/2/16	9	294	1001	3	惣菜4	360
2021/2/16	9	294	1001	5	一般食品5	600
2021/2/16	9	294	1001	5	一般食品5	600
2021/2/16	9	294	1001	4	日配品4	130

Step2. モデル構築

マルチクラウドAutoMLサービス『DATAFLUCT cloud terminal.』に
データセットをアップロードし、来客数を予測するモデルを作成。



作成モデルを使って“時間帯別の来客数を予測した結果”と“実際の来客数データ”を比較してみた結果、
人の経験則と比べても見劣りするレベルの精度（**予測誤差：±30%～50%**）に。

なぜ、このような結果になってしまったのか？

Step3. モデル評価

予測精度が低かった理由について考察し、精度改善に向けた対策を検討。

担当者の考察

- 1日単位・1週間単位で分析した場合、予測誤差は±10%程度だった。
- 時間帯別の来客数は平日でも曜日によって異なる上、店舗によっても傾向が異なる。
- 人の店舗前の通行量は、天候によって変動する。
- イベントやSNSなどの外部要因による影響も考慮した方がよい。

データサイエンティストからの助言

- データ量と特性が異なる店舗データ混ぜてしまったことによって、教師データに分散と偏りが生じている可能性がある。
- 東京の店舗のデータ量が多すぎるため、大阪の店舗の予測が影響を受けている。
- 同じ年代の顧客でも駅近の店舗の方が人数が多いため、郊外の店舗の予測に悪影響が出ている。
- 2021年2月のデータで作成したモデルは、季節性の違いや新型コロナなどの社会情勢の違いによって、3ヶ月後、半年後、1年後には更に精度が下がる可能性が考えられる。

機械学習の予測精度を高く維持し続けるためには、以下の対応が必要に。

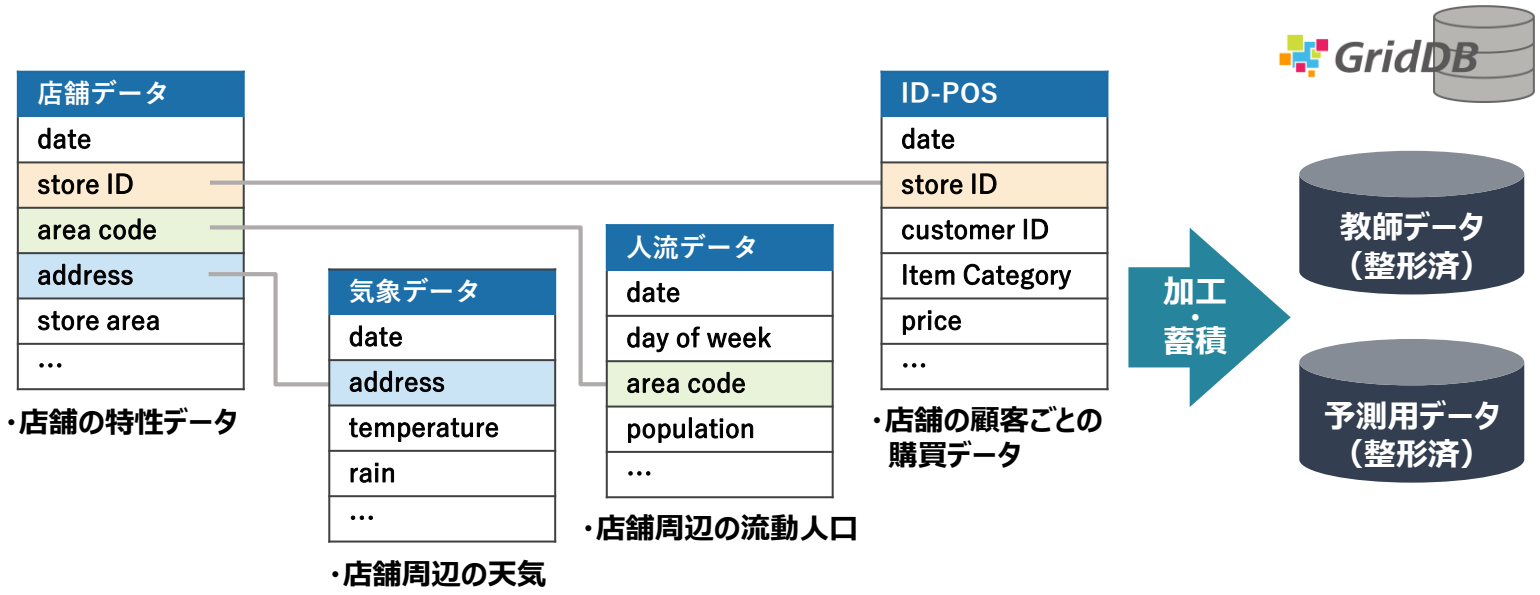
- 「自社のデータと外部のデータを組み合わせた必要なデータの準備」
- 「季節性・社会性を常時反映するための継続的なデータの収集・加工・蓄積できるデータ基盤」

Step1. データ準備（改善）

データ構造や更新頻度が異なるデータを全てGridDB Cloudに集約し、店舗ごとに最適なモデルを構築できるようにデータを整形。

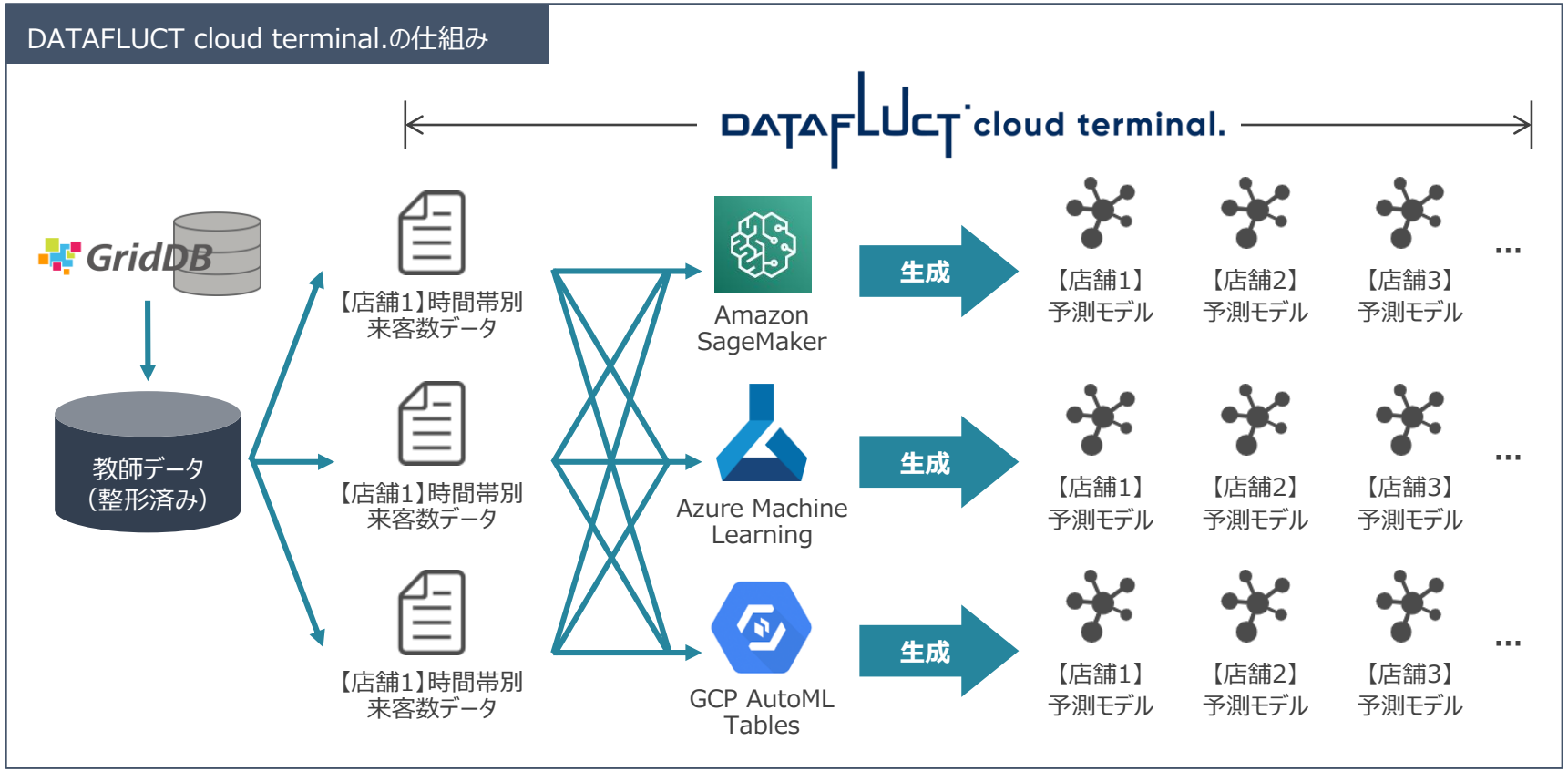
GridDB Cloudへのデータ集約・データ整形

自社データ	ID-POS、店舗データ	外部データ	気象データ、人流解析データ



Step2. モデル構築/評価（改善）

GridDB Cloud上に用意した店舗ごとに整形済のデータセットを使い、AutoMLで予測モデルを作成。

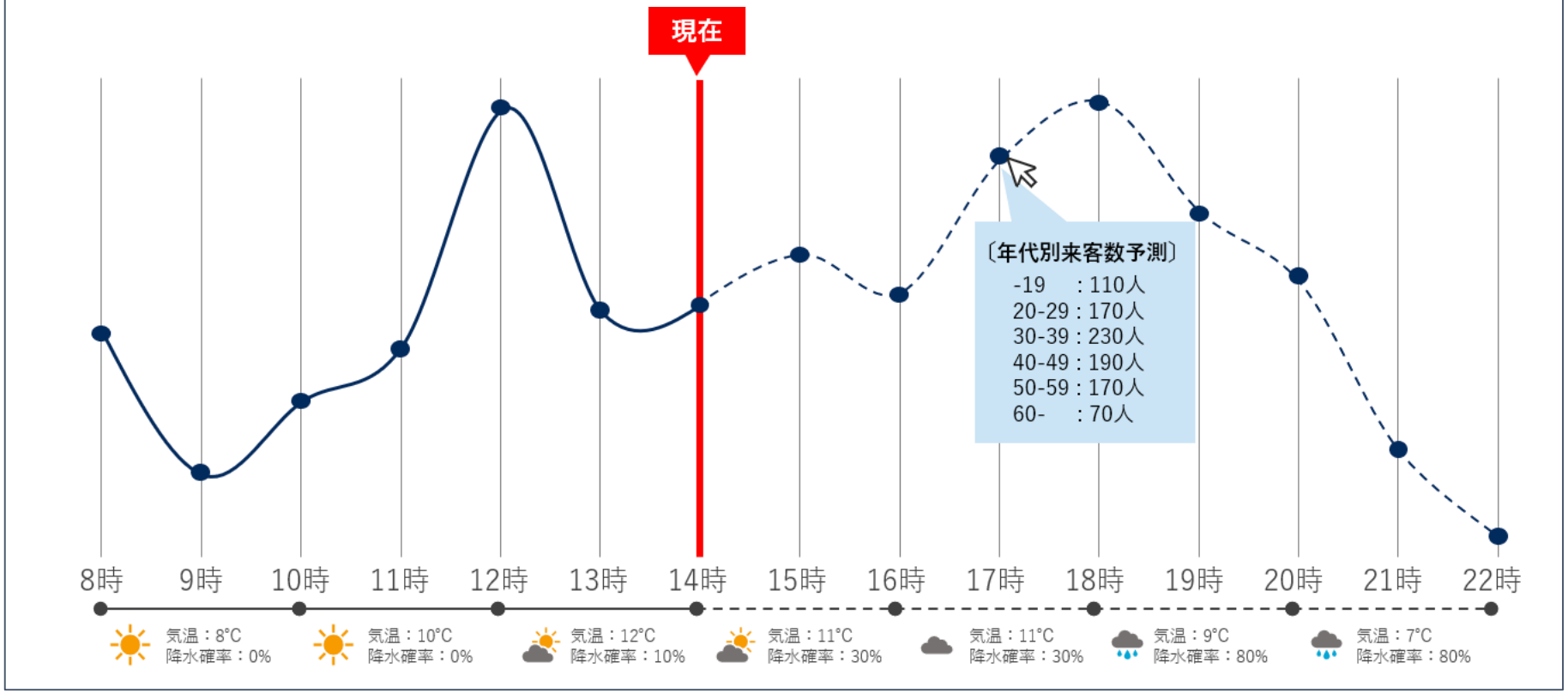


各店舗毎に最適化したモデルを用意することで、当日内の来客数予測精度90%以上 を達成する精度に。

Step3. モデル展開・運用管理

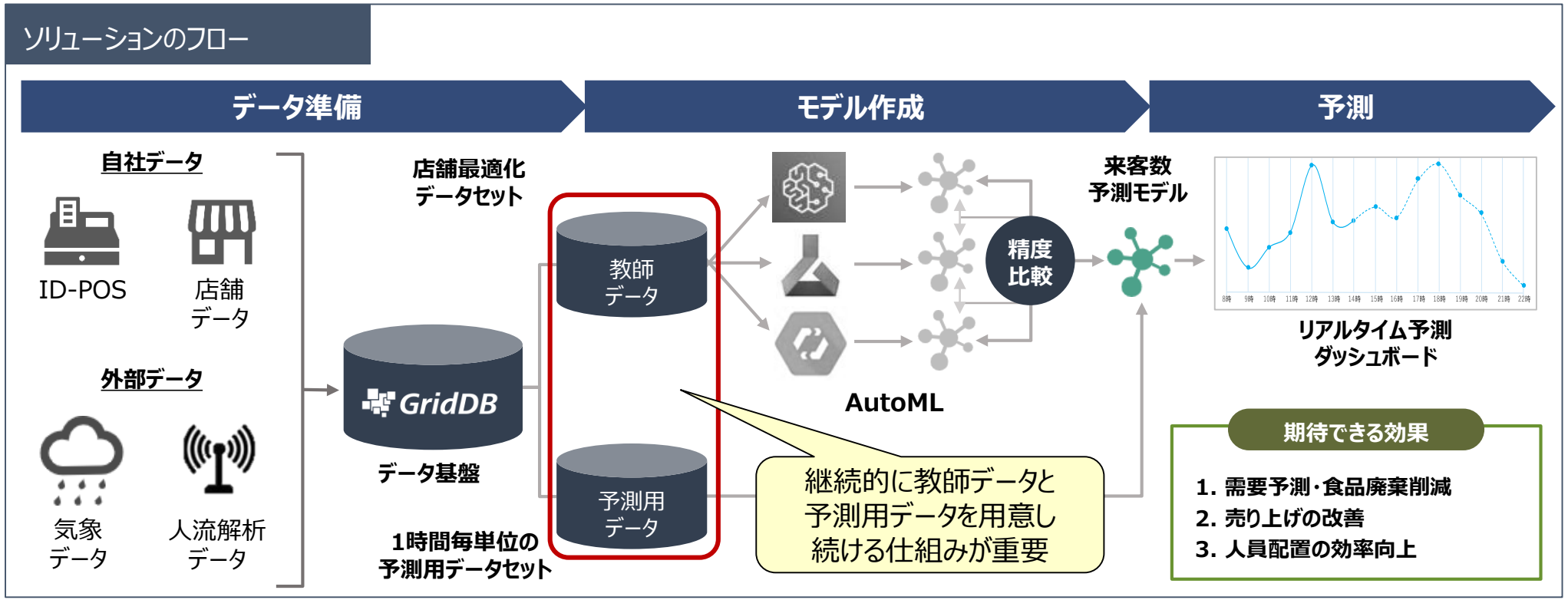
店舗毎に最適化したモデルをアプリケーションから利用できるようにデプロイ。
年代別来客数の予測結果をリアルタイムに表示するダッシュボードを用意し、
惣菜品の調理タイミングや棚出し計画のリアルタイムな意思決定が改善に成功。

ダッシュボードのイメージ図（年代別来客数予測）



本ケーススタディのポイント

- 各店舗に導入しているID-POSデータを集約するデータ基盤GridDB Cloudを活用。
- 外部要因を考慮するために外部データ（気象データ、人流データ）をリアルタイムに収集・蓄積・加工。
- 収集した店舗別データと外部データを基に、店舗ごとに最適な予測モデルをAutoMLを使って構築。



本ケーススタディのホワイトペーパー配布中

http://resources.griddb.com/whitepaper/pdf/whitepaper-sales-forecast-datafluct-griddb.pdf

これからのデータ基盤

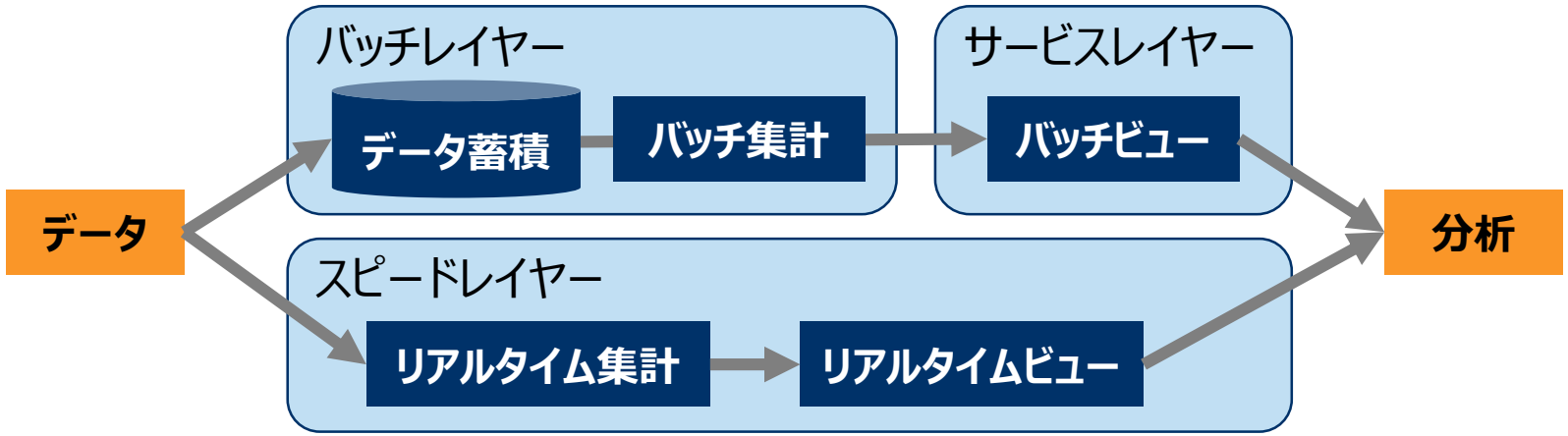
「過去の分析」から「将来の予測」へ。最新のデータを使ってリアルタイムに分析・予測。

DWHを使ったデータ基盤の弱点

リアルタイム分析は不可！

※1 ODS (Operational Data Store)：収集データを一時的に保持するデータベース
※2 ETL (Extract/Transform/Load)：データの整形、統合を専門に行うツール
※3 DWH (Data Warehouse)：情報を時系列に整理して保管するデータベース

リアルタイム分析を可能にするラムダアーキテクチャ

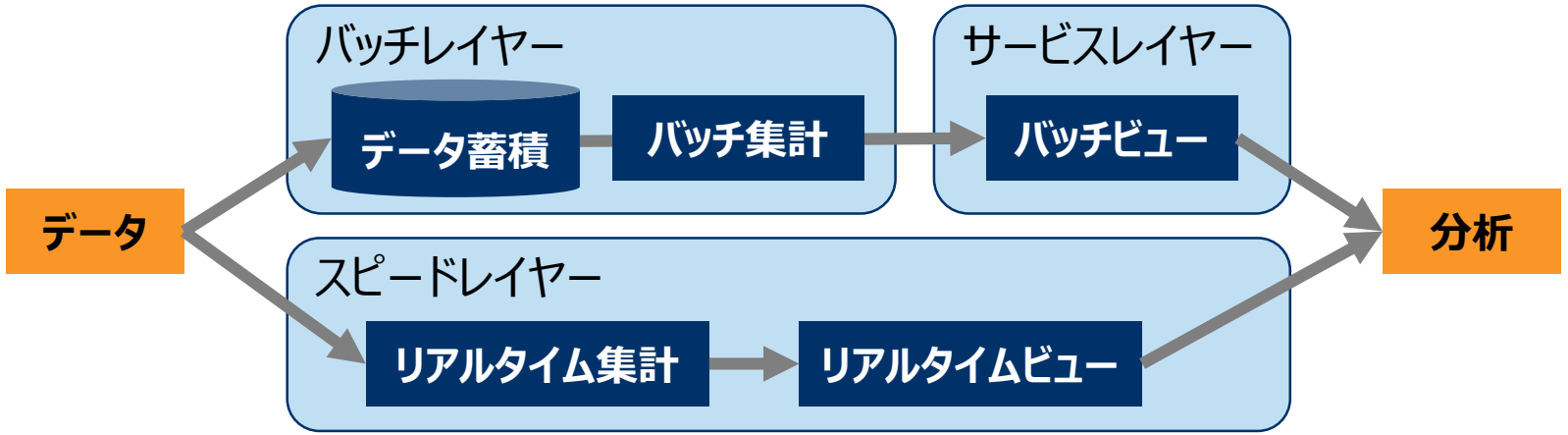


リアルタイム分析が可能。

これからのデータ基盤

「過去の分析」から「将来の予測」へ。最新のデータを使ってリアルタイムに分析・予測。

リアルタイム分析を可能にするラムダアーキテクチャ



欠点
- システムが複雑。
- 問題の把握が難しい。

今後は蓄積と分析をリアルタイムで行えるデータベースを活用したデータ基盤に

DBの高性能化がポイント

- シンプルなアーキテクチャ。
- 運用・メンテナンスの負荷を軽減。

これからのデータ基盤

今後、企業が分析の対象とするデータはどんどん増えていく。

ビッグデータに対応していくにはクラウドの活用が効率的。

メリット	デメリット
リソースネックになりにくい	データの取り出しに費用がかかる
分析のためのサービスが豊富	クラウドベンダーロックインになりやすい

クラウドには出せないデータも依然として存在しており、オンプレミスも必要。

クラウドとオンプレミスのハイブリッド型へ

これからのデータ基盤

効率的かつ俊敏なデータ活用のための 「DataOps」

DataOpsとは

> 組織全体のデータ管理者とデータ利用者の間のコミュニケーションの向上と、データフローの統合と自動化の改善に焦点を当てた共同作業によるデータ管理の手法。（ガートナー）

データ基盤といったツールだけでなく、担当者、プロセス、テクノロジーを連携させることで、
効率的かつ俊敏なデータ活用を目指す。

企業文化や組織的なアプローチも必要。

> DataOpsの実現方法は、業種や組織の構造、利用技術などにより最適解が異なる。
> 「こうすればDataOpsが必ず実現できる」という万能解は残念ながら存在しない。
>
> Web記事「データ活用の障壁をDataOpsで回避する（第1回）」から引用。
> https://dcross.impress.co.jp/docs/column/column20201007/001806-2.html

まとめ

- データを継続的にビジネスに活用していくためには
  半自動化されたデータ基盤が必要です。
- 長期間使用するデータ基盤は、将来を見越した設計を
  行うことが重要でしょう。
- これからはリアルタイムで分析可能なデータ基盤が
  主流になっていくでしょう。
- DataOpsについても注視していく必要があります。

TOSHIBA

ご清聴ありがとうございました。