Data Analytics OnAir

BigQuery の強みを活かした エンタープライズ基盤のクラウド化実践

2022年6月7日

株式会社ブレインパッド

BrainPad Analytics Innovation Company

スピーカー紹介

ビジネス統括本部

アライアンス開発室長

筧 直之

データエンジニアリング本部

ソリューション開発部 グループマネジャー

西尾 陽子

本日お伝えしたいこと

- 最近の BigQuery 利用シーンの変化
- Enterprise DWH クラウド移行のリアル
- クラウド移行のメリット最大化のためには

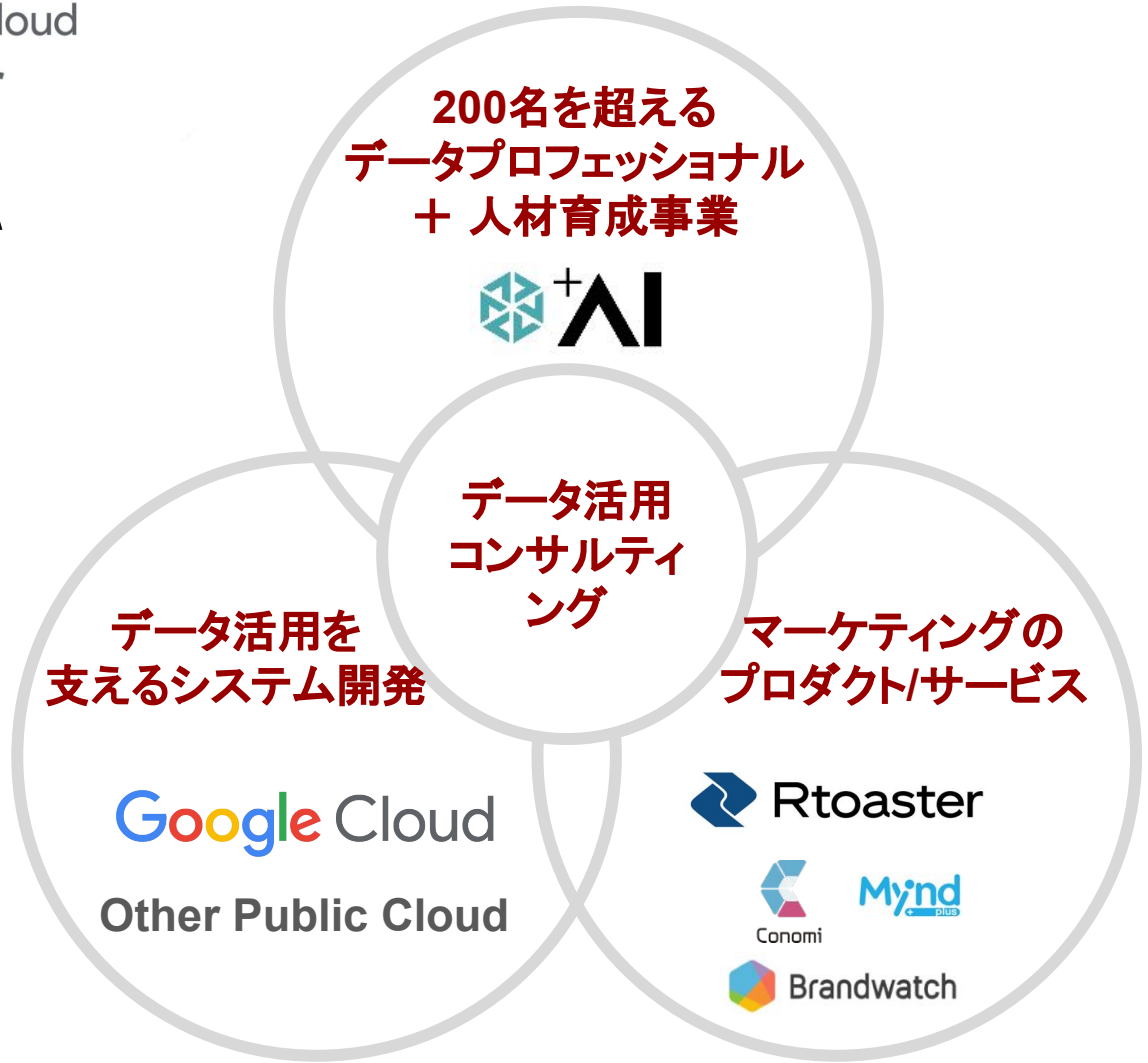
ブレインパッドとは？

BrainPad

Google Cloud Partner

データの利活用を通じて会社を変革したい企業を支援する2004年創業のリーディングカンパニー
（東証プライム上場企業）

- ツール/ベンダーに制約を持たずに、目的志向で顧客のビジネス課題を解決
- データ分析/AI開発の領域を超えて、データドリブンな経営を支えるための、人材・データ活用インフラ、業務システムの高度化を全面的に支援
- Google Cloud プレミアパートナーとして、Data Analytics 領域を主軸に協業



当社が最近相談頂く DWH 案件

以前より AI/ML やデジマ案件などで BQ を使っていたが、
ここ1,2年は大規模基盤、移行系の DWH 案件でも利用するシーンが増加

AI/ML 案件での BigQuery 活用

マーケティング アナリティクス / CDP 案件での BigQuery 活用

大規模基盤/移行案件での BigQuery 活用

大規模基盤 / クラウド移行プロジェクトのリアルを2パターンでご紹介

本日は、その中でも最近取り組んだRDB系の大規模業務基盤、hadoopを用いた分散処理サービス基盤の
２つのクラウド移行をご紹介

ユースケース 1

Teradata / SAS 環境
のクラウド移行

データの移行

ユースケース 2

Hadoop / Spark環境
のクラウド移行

処理の移行

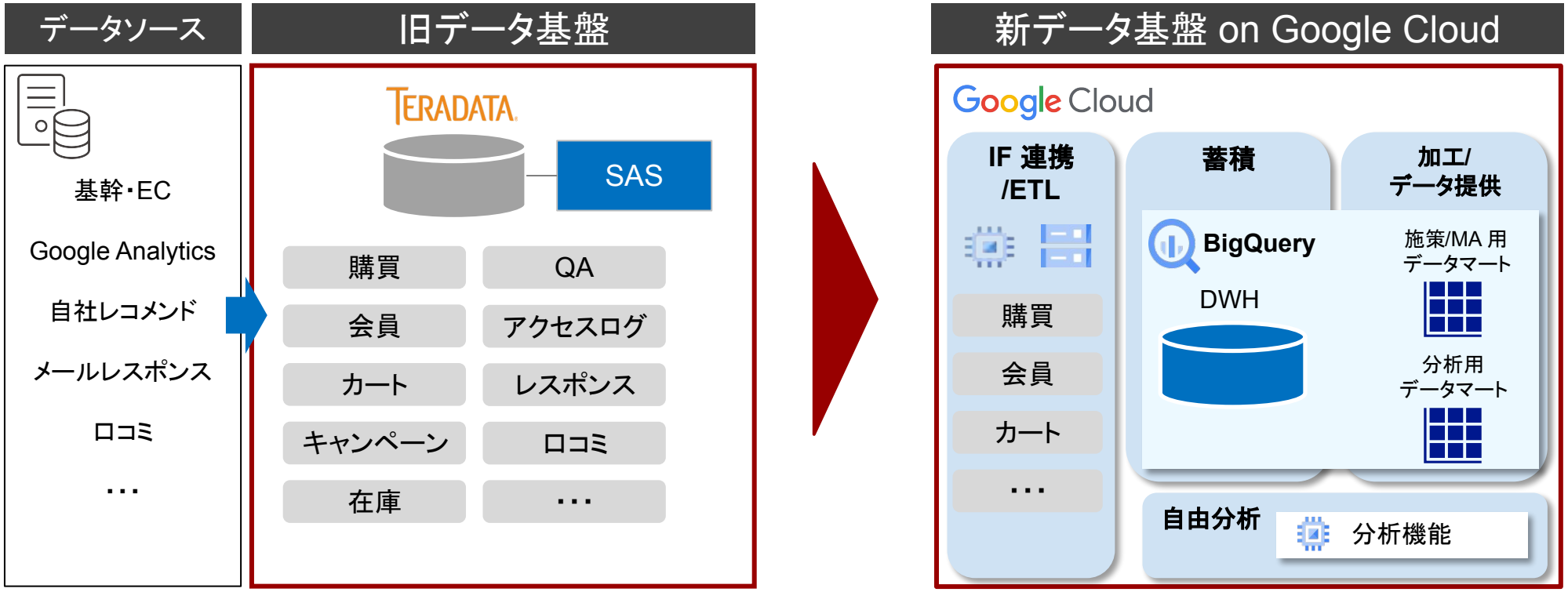
ユースケース1

Teradata / SAS 環境のクラウド移行

大手総合通販企業
カタログ事業のマーケティング施策/分析業務基盤

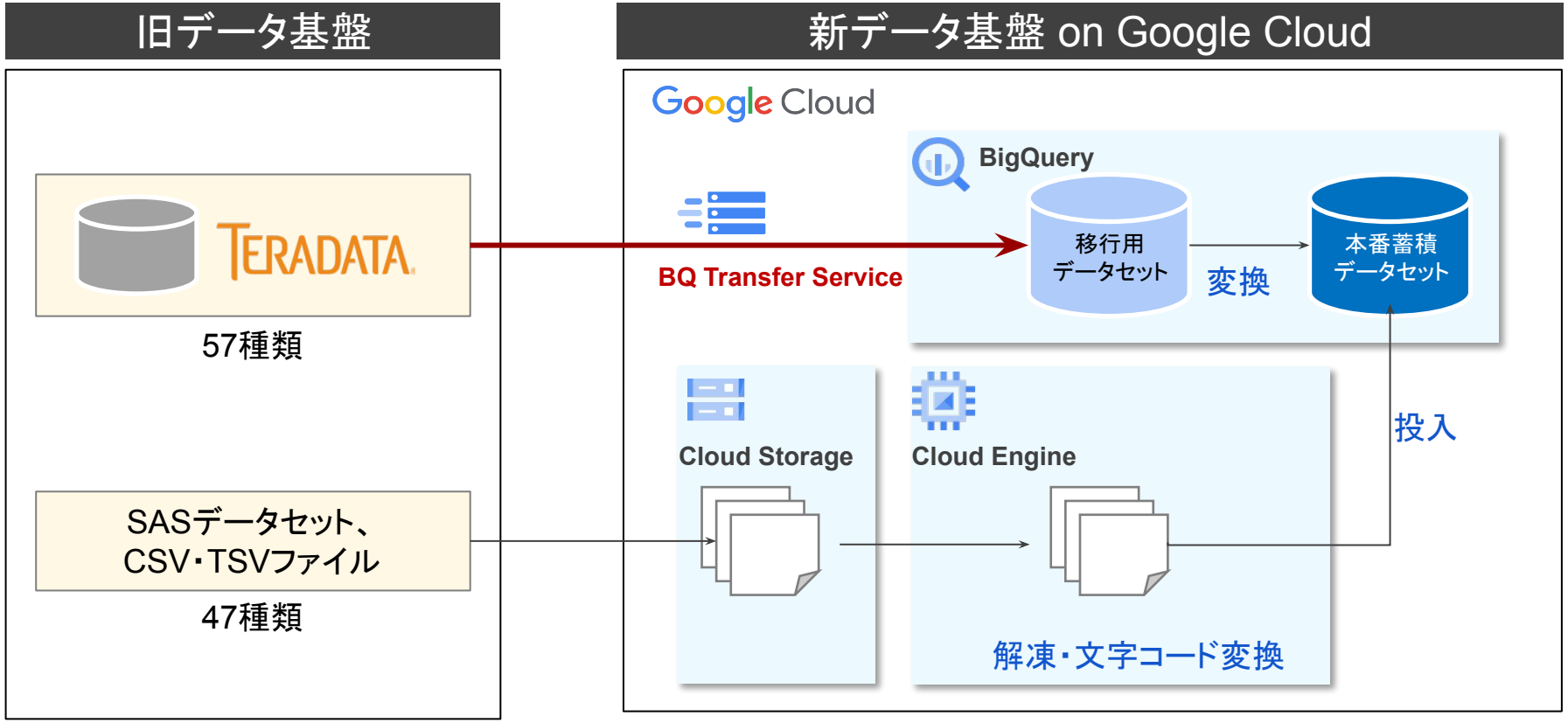
Teradata / SAS からの移行事例

- BigQuery への過去データ移行が容易に短時間で実施できた事例
- 分析や施策の高度化など実現したい将来像があるが、データベースがボトルネックとなっていた
- スケールアップできる、BigQuery の性能・利用しやすさ・契約の分かりやすさがが決めてとなり、Teradata を中心としたデータ基盤から、BigQuery を中心とした Google Cloud への基盤へ、刷新を行った



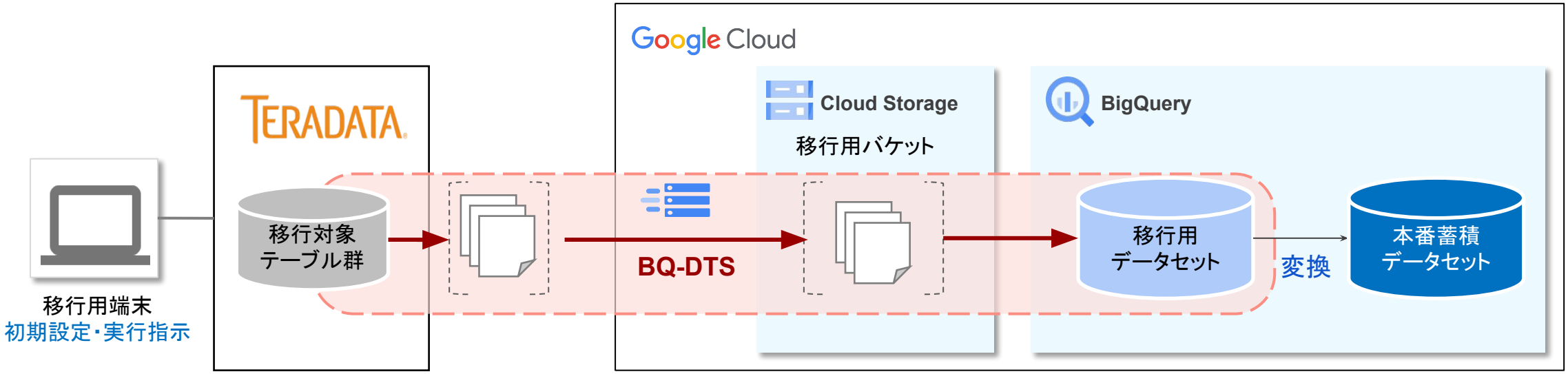
過去データの移行

- データ基盤構築においては、過去に蓄積データの新環境への移行が必要となる
- 今回も蓄積されたファイル・DBのテーブルをすべて BigQuery に移行を実施した
  - ファイル：GCSにアップロードし、BigQueryへ
  - Teradata ： BigQuery Data Transfer Service (BQ-DTS) を利用して一括で BigQuery へ



BigQuery Data Transfer Service (BQ-DTS)とは

- BigQuery へのデータ移行を自動化するマネージド サービス
- 移行元テーブルを Export → ファイルをストレージへ Upload → BigQuery でスキーマ定義・データ投入まで全自動で実行してくれる



BQ-DTS により全自動で BigQuery に取り込まれる

実際の移行をしてみて

- 移行対象テーブル約57、サイズ約340GBの移行は、約11時間であった
  - 検証時は10GBあたり5分を想定していたが、実際は10GBあたり20分（1GBあたり2分）程度
  - Export・Uploadが時間の大半であり、DB負荷・ネットワーク状況に大きく左右される

 実測結果

フェーズ	テーブル	Teradata サイズ(GB)	Export （秒）	Upload（秒）	BQ Load （秒）	合計処理時間 （秒）	1GBあたりの 処理時間 (秒)
事前検証時	X	18.6	0:03:45	0:04:16	0:01:00	0:09:01	0:00:29
	Y	9.4	0:02:35	0:00:20	0:01:00	0:03:55	0:00:25
本番移行時	A	7.1				0:17:41	0:02:29
	B	8.5				0:15:40	0:01:51
	C	71.8				2:34:43	0:02:09

手動による移行よりも

✔ **移行にかかる手間が少ない、工数・期間が大幅に削減できた**

- お客様側の作業負担が軽減
- 担当者間のやり取りにかかるオーバーヘッドがない
- 複数回の移行も容易

✔ **品質が担保しやすかった**

移行Tips

- データ移行に BQ-DTS を使用すると期間・工数としても、品質としてもメリットが大きい
- 以下の点などを留意して使用するとよい

❖ **実行手順観点**
- ➢ 移行元DBのリソース状況を加味して実行を開始すること
- ➢ 1GB未満のサイズの小さいテーブルが多数あるとオーバーヘッドがかかる
- ➢ 増分移行モード（β版）も利用可能

❖ **データ観点**
- ➢ 移行前・後で項目単位で差がないかを検証すること
- ➢ 差が発生した場合は、移行用データセットからクエリ変換して本テーブルに投入する
  - ■ 当時発生した例
    - ● 後ろスペース有無
    - ● TIME型、TIMESTAMP型のタイムゾーン

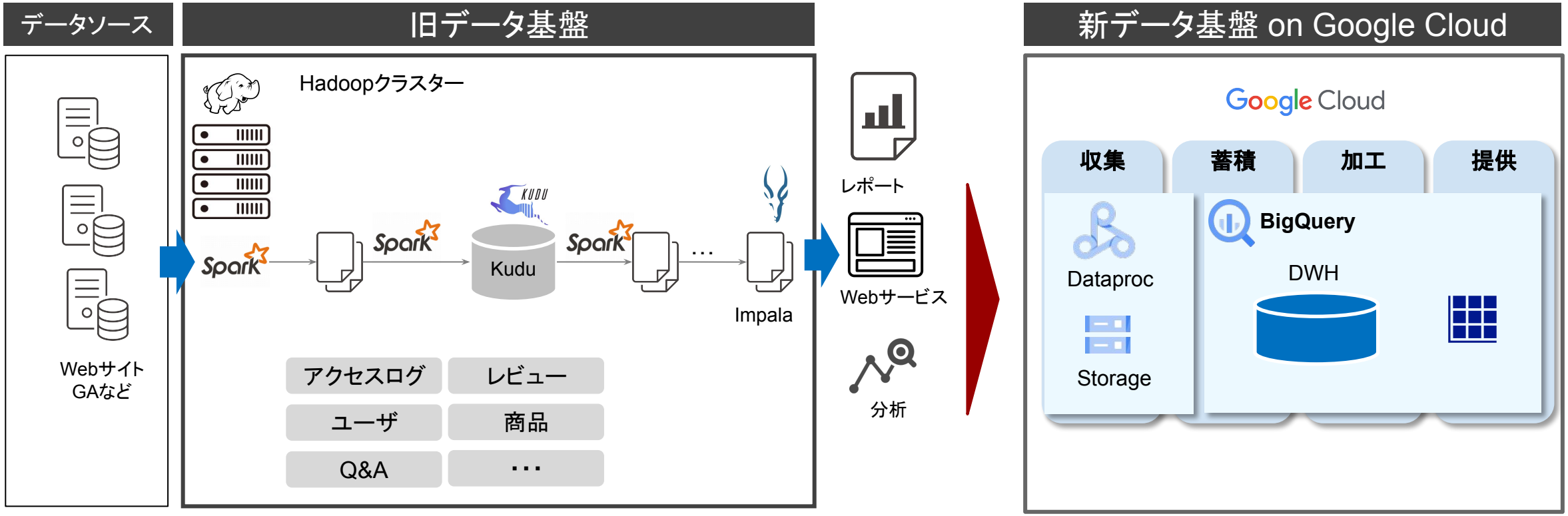
ユースケース2

Hadoop / Spark環境のクラウド移行

大手ポータル・ECサイト運営企業
メーカー向けサービス基盤

Hadoop / Spark からBigQueryへの移行事例

- オンプレHadoopのデータ基盤で、大規模データ × 複雑なデータ加工処理を行われている企業様
- データ処理に関わるパフォーマンスや運用・保守面の課題を解決するために、Google Cloud＋BigQueryでのデータ基盤刷新を実施



抱える課題とBigQueryへの移行メリット

- オンプレ＋Hadoop環境に関わる課題を多数抱えていた
- BigQueryへの移行により、インフラ管理や細かなパフォーマンスを気にせず、処理開発に専念できるように

	主な課題	解決・改善
データ・ロジック	● データ量の制約 ● ロジック複雑化	❖ 扱えるデータ容量拡大、処理速度向上 ❖ 性能制約により複雑化していたロジックがシンプルに ❖ 簡易にパフォーマンス向上が可能（パーティショニング・クラスタリング）
パフォーマンス	● 処理時間の延伸 ● チューニングの負担 （パラメータ・プログラム構成）	
インフラ	● インフラの運用管理・コスト負担	❖ SaaSのためインフラ管理が不要 ❖ 利用・蓄積した分だけのコスト抑制 ❖ 臨時環境も容易に準備可能

移行ステップ

- 今回は Spark のDataFrameによる処理を読み解いて、SQLで再構築した
- 基盤構築ではあわせて既存不具合の解消や、インプットデータの変更などの要望が出ることが多い
- 段階的に移行することで、結果的に品質担保しやすく、工数が小さく済む

1. そのままDataFrame→SQLへ移植	関数:SQLクエリ＝1:1で処理内容を維持して移植
2. 単体テスト、現新データの比較	移行ロジック、実行順序、データの組み合わせの正しさを確認
3. 最適化	クエリの改善・最適化、記述方法の統一、データフロー変更
4. 単体テスト、現新データの比較	（最適化に合わせたケース修正）正しさを確認
5. 現行と結果が変わる改修の組み込み	不具合解消、データソース変更など
6. 単体テスト	単体レベルで回帰テスト（現行データとの比較はできない）

移行Tips

- 検討・注意すべきポイントはあるものの、BigQuery・SQLに移行できない処理はなかった
- 関数も揃っており、必要であればUDFなど定義が可能

❖ **Spark Dataframe は結合の際に暗黙的な型変換がされるが、BigQueryは型の一致が必要**

❖ **BigQueryの大量データを高速に処理できる特性を活かした処理に変更するのもあり**
- ➢ 差分→全件更新に変えるなど、ロジック・フローをシンプルに

❖ **中間ファイルの出力有無は再検討をおすすめ**
- ➢ Spark：IOを減らしてメモリ上で処理させるため、中間ファイル出力は極力避けていた
- ➢ BigQuery：中間データをテーブルに書き出すことで途中データが確認でき、保守性が向上する

まとめ

クラウド移行により得られたメリット・成果

ユースケース 1

Teradata / SAS 環境
のクラウド移行

- 市場の変化に合わせて適切な取り組み（新施策の実施等）がしやすくなった
- データ量をほぼ気にせず蓄積できるようになり、必要な分析の促進につながった
- 同じデータと同じ処理基盤を皆が使えるようになり、業務の標準化・シンプル化
- index管理が不要となり、単純なデータ投入や処理のみに時間が短縮された

ユースケース 2

Hadoop / Spark環境
のクラウド移行

- 運用・保守業務として削減できた工数・負荷を新規開発へシフト
- 技術要素を BigQueryやSQLに統一できたため、対応可能なエンジニアを確保しやすくなった
- ロジックがシンプルになったため、新任者でもキャッチアップしやすくなった
- 社内での他のプロジェクトでも同じ技術を用いる前提で進められるため、将来的な拡張がしやすくなる

クラウド移行のメリットを最大化するためには・・・

- 現状のデータ・処理・構成をそのままクラウドに持っていっても（Lift & Shift）、得られるメリットは、管理面くらい
- せっかく移行するのであれば、得られるメリットを最大化するために、いろいろ整備しておくのが良い

データ項目の定義

レイヤー/データのフロー定義

ジョブの並列化

ご清聴ありがとうございました！

BrainPad
Analytics Innovation Company

株式会社ブレインパッド
〒108-0071 東京都港区白金台3-2-10 白金台ビル3F
TEL：03-6721-7002　FAX：03-6721-7010
www.brainpad.co.jp　info@brainpad.co.jp