XIMIX Solution Overview
Bigdata & BigQuery on Google Cloud

Nippon Information and Communication Corporation

Team XIMIX

powered by NIHC 日本情報通信株式会社

目次

1. Google Cloud Bigdata基盤 構成要素

概要

Google Cloud とは

Google Cloud とは、Googleが運営しているクラウドコンピューティングのプラットフォーム。Google検索、YouTube、Gmail、GoogleMapsなど誰もが一度は利用したことのあるエンドユーザー向けのサービスと**同じインフラが利用できる**プラットフォームです。

言わずと知れた世界No1の
検索エンジン。Googleの事業の柱

全世界で月間視聴者数15億人以上
毎日10億時間視聴され、視聴回数は数十億回にのぼる

全世界で利用者数10億人以上
全てのユーザに15GBの容量を提供

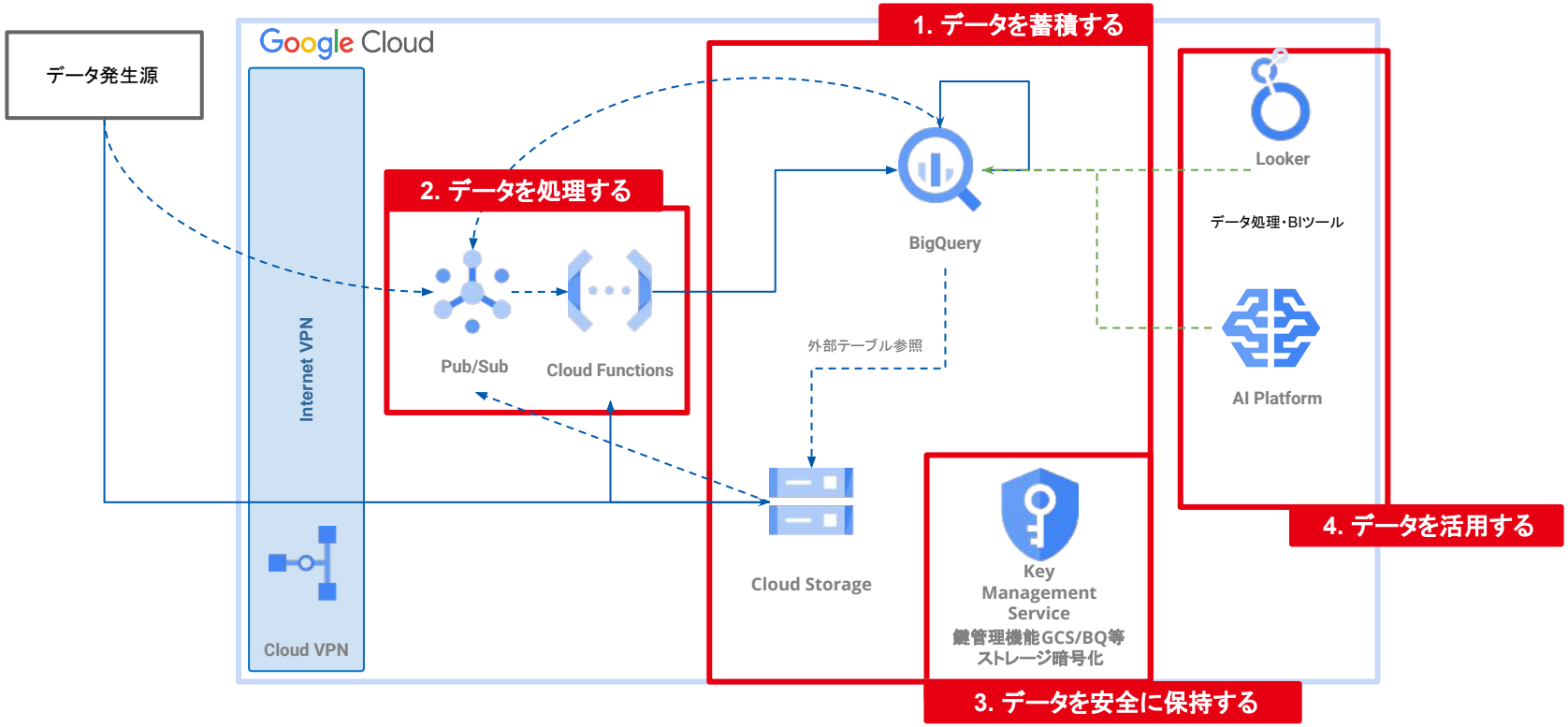
220カ国/地域の正確で総合的な地図、
1億を超える場所の詳細な情報を提供。
Google検索とも連動

全世界のユーザが利用しても安定稼働しており、これらのサービスを支えている
Googleのインフラを**エンドユーザ様も同じアーキテクチャ/クオリティでご利用いただける**のが
Google Cloud です。
（もしかしたら、購入したコンピュートインスタンスの隣はGmailのサーバかもしれません・・・）

Google Cloud 提供サービス例

- コンピューティング　：Compute Engine、App Engine、Kubernetes Engine
- **ストレージ　：Cloud Storage、Cloud SQL、Firestore、Bigtable、BigQuery**
- ネットワーキング　：VPC、Cloud DNS、Cloud CDN、Cloud Armor
- **データパイプライン　：Dataflow、Dataproc、Dataprep、Data Fusion**
- AI、機械学習：Cloud Vision、Cloud Translation

Bigdata基盤 構成要素



2. データの蓄積と BigQuery

ストレージ・DBサービス

ストレージサービス

各サービスの概要

主なストレージサービスの概要は下記の通りです。

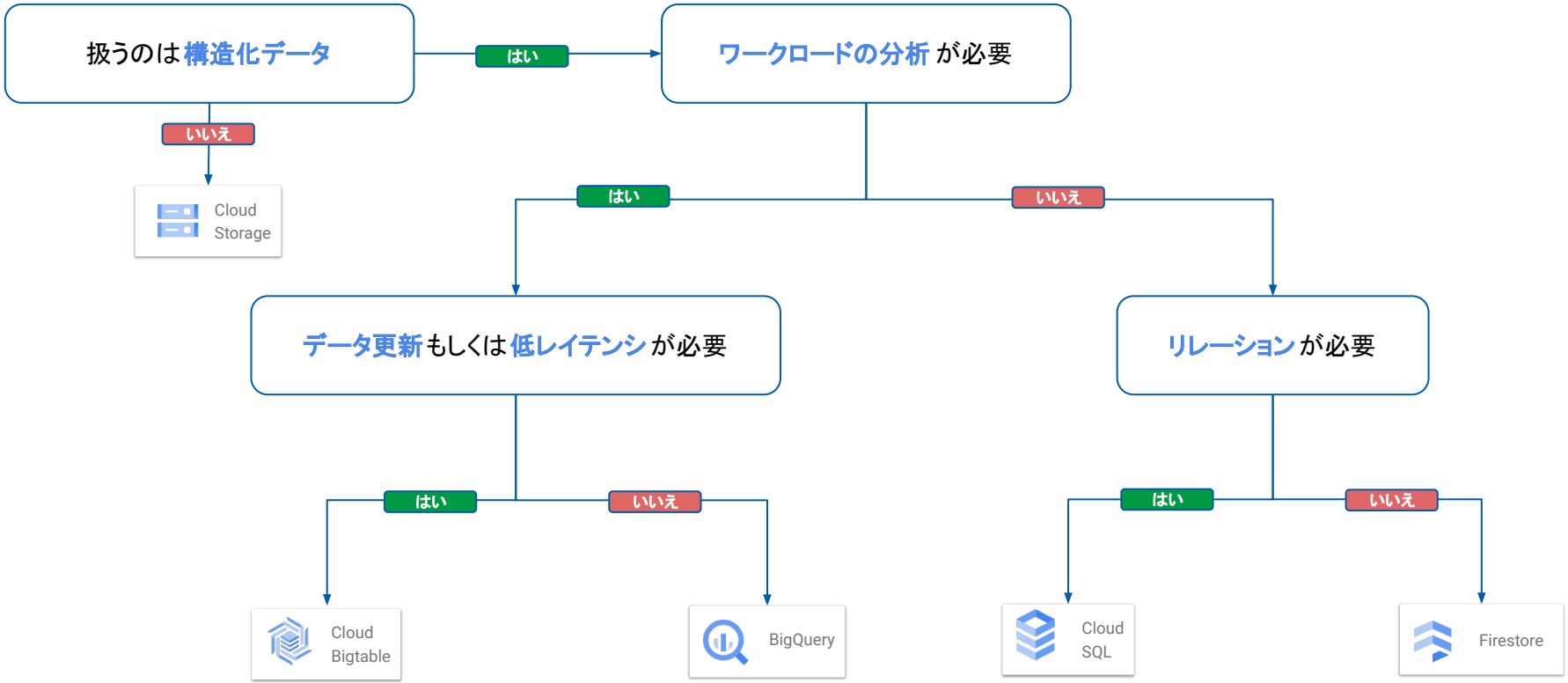
サービス名	概要
Cloud Storage	Google Cloud でオブジェクトを保存するためのサービス。音声や動画をアプリやウェブサイトに直接ストリーミングするために必要な可用性とスループットを提供します。 高速で低コスト、耐久性の高いストレージが特徴。
Cloud SQL	「MySQL」および「PostgreSQL」のPaaSデータベースサービス。リレーショナルデータベースの設定/メンテナンス/運用/管理のフルマネージドサービスを提供。
Firestore	NoSQLドキュメント指向データベース。 柔軟なデータ構造、高機能なクエリ処理、リアルタイムアップデート、オフラインサポートなどが特徴。
BigTable	大規模な分析ワークロードにも運用ワークロードにも対応できる、フルマネージドでスケーラブルな NoSQL データベース サービス。低レイテンシで高スループットであることが特徴。
BigQuery	ビジネスのアジリティに対応して設計されたサーバーレスでスケーラビリティと費用対効果に優れたクラウド データ ウェアハウスです。高速なクエリ処理が特徴。

ストレージサービス
各サービスの紹介

主なストレージサービスの項目比較は下記の通りです。

項目	Cloud Storage	Cloud SQL	Firestore	BigTable	BigQuery
容量	ペタ+	ギガ	テラ	ペタ	ペタ
メタファ	ファイルシステム風	Relastional Database	Persistent HashMap	Key-Value	Relational
読み込み	ローカルPCへコピー	Select Row	filter object on Property	Scan Rows	SELECT Row
書き込み	One file	Insert Row	Put Object	Put row	Batch / Strem
Update Granularity	An Object (a file)	Field	Attribute	Row	Field
Usage	Store Blobs	No-ops SQL database on Cloud	Structured data from App Engine Apps	No-ops, high throughput , scalable , flttented data	interactive SQL querying fully managed whrehouse

ストレージ
サービスの選択



BigQueryの特徴

1. フルマネージド（クエリーサービス）として提供
2. 費用は利用した分のみの従量課金
3. 既存DWHの課題を解消したクラウド型DWH
4. 大容量データを活用する拡張（GISとML）

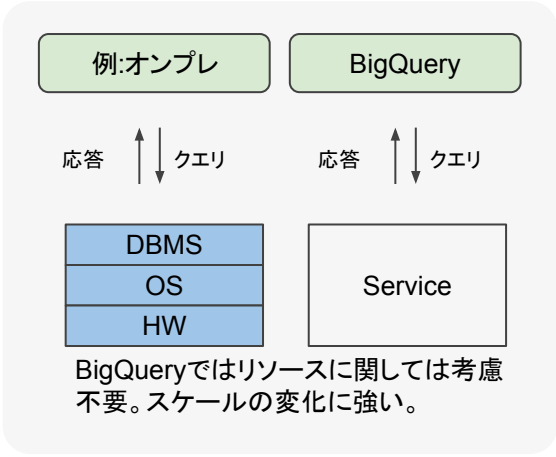
BigQueryの特徴

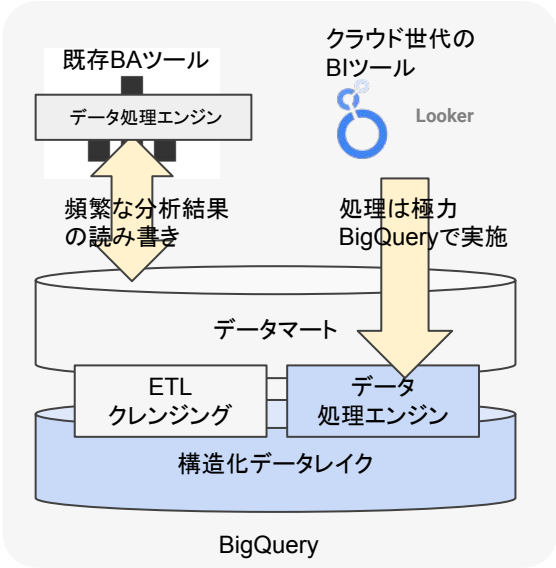
BigQueryの基本的な機能

- **フルマネージドサービス** として提供されるクラウドベースの列指向データウェアハウスであり、0バイトからペタバイトデータの処理が可能。容量無制限、ハードウェアなどの交換が不要で使い続けることが出来る。
- 費用はデータの保管料とクエリ処理量に対しての **従量課金**
- 特にBigDataに対しての **分析処理を高速** に行うことが得意（対象データが膨大になった際大きな劣化がない、同時に利用してもパフォーマンスの低下が発生しづらい）従来バッチ処理で行うような負荷の高いクエリをオンデマンドで利用可能
- 標準SQLやSDK経由で分析処理が可能、インデックス設計不要
- 外部テーブル（GCS、BigTable）のデータも直接分析可能

BigQueryからの広がる拡張機能

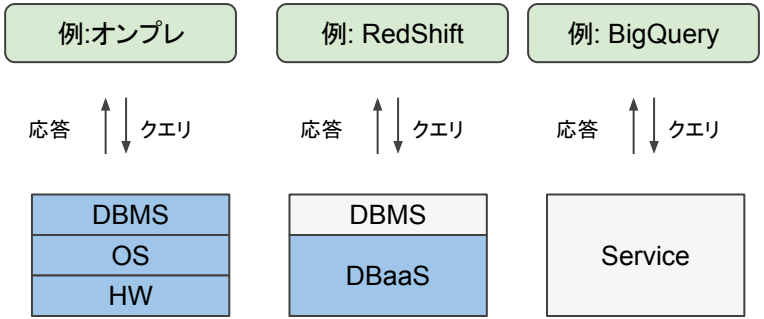
- 地理関数対応 (BigQuery GIS)
- BigQueryの膨大なデータを教師として Machine Learningの学習・推論を可能 (BigQuery ML)
- 応答速度の速さが必要な、既存 BAツールのデータマートむけキャッシュ機能の提供（BigQuery BI Engineとして提供）





サーバレス型（クエリーサービス）

- BigQueryはフルマネージドサービスとして提供されます。利用者は、BigQueryのバージョンアップはHW的な問題に対して何も考慮する必要はありません。アップグレードの際にダウンタイムは発生せず、システム パフォーマンスの低下もありません。
- 利用者は事前に、リソースに関わる見積もりを行う必要がありません。利用者からはBigQueryを実行するComputeを意識する必要がありません。



BigQueryではリソースに関しては考慮不要。
スケールの変化に強い。

AWSとの比較

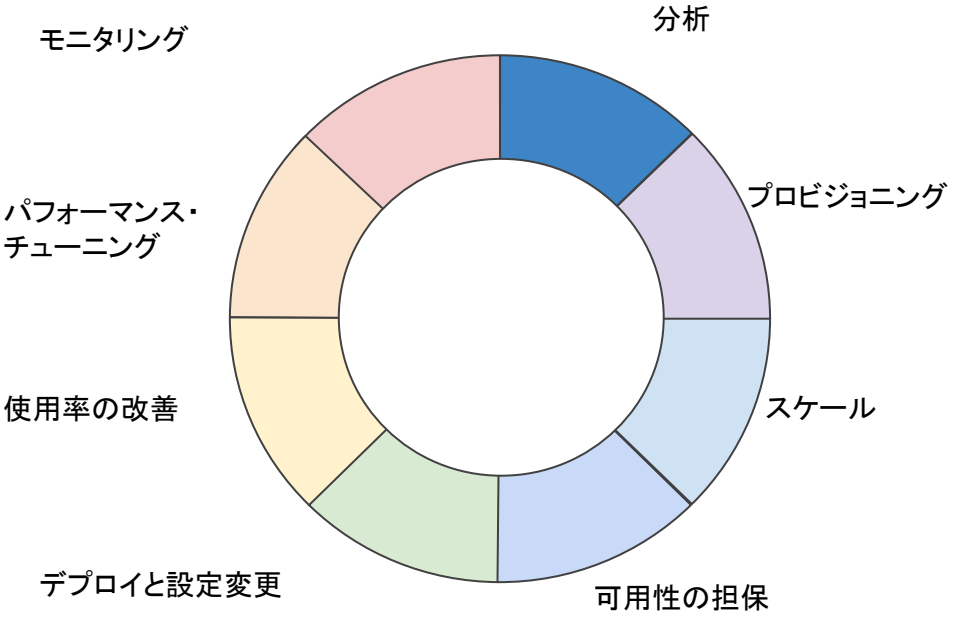
RedShiftに対してBigQueryはサーバレスであり、初期時にリソースの見積もりは不要です。1Byteから利用を始めることができます。またスケールアウトも考慮する必要がありません

従量課金

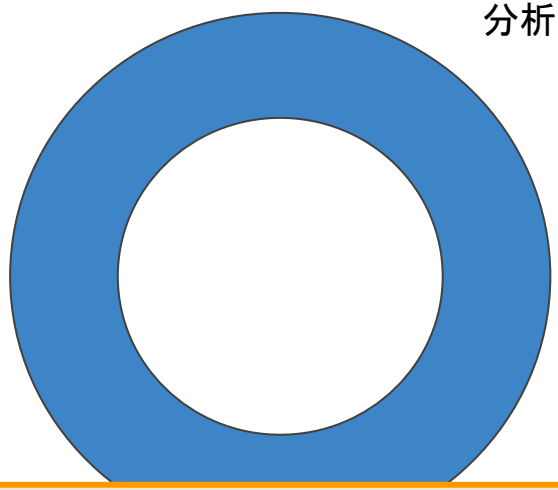
- BigQueryはクエリがスキャンするデータ量により料金が決定される。
  - 列指向であるためSELECTされた列はすべてスキャン
- パーティション列を定義したテーブルはWHERE句によるスキャン範囲を行レベルでフィルタリングできる。
  - LIMIT句はスキャン範囲に影響しない

フルマネージドによるメリット

従来のデータウェアハウス



BigQueryによる分析



フルマネージドサービスサービスであるBigQueryは従来のDBMSと比較し分析に注力することができる

BigQuery
特徴

BigQueryには以下のような特徴があります

特徴	説明	具体的な用例
BigQuery GIS （地理情報）	BigQuery GISの地理関数により **地理データを分析および可視化** が可能	ある位置情報（緯度・軽度）が入った定期的な走行データがあるときの、東京都内での最大走行速度の計測。
BigQuery ML （機械学習）	BigQuery MLにより、標準SQLクエリを利用して**機械学習モデルを作成し実行** 可能	あるデータを基に線形回帰モデルの作成、評価、予測。
BigQuery BI Engine	BigQuery BI Engineの高速なメモリ分析により**BIツールとのデータ連携を高速化** 可能	Looker studio や Looker によるBIデータ分析。
BigQuery Data Transfer Service	BigQuery Data Transfer Serviceによりテーブルへの**SaaSデータ取り込みを自動化** 出来る。	Amazon S3から定期的にデータを自動取り込む。
外部テーブル参照	**外部テーブル**を参照出来る。	GCSのファイルを外部テーブルとして参照する。

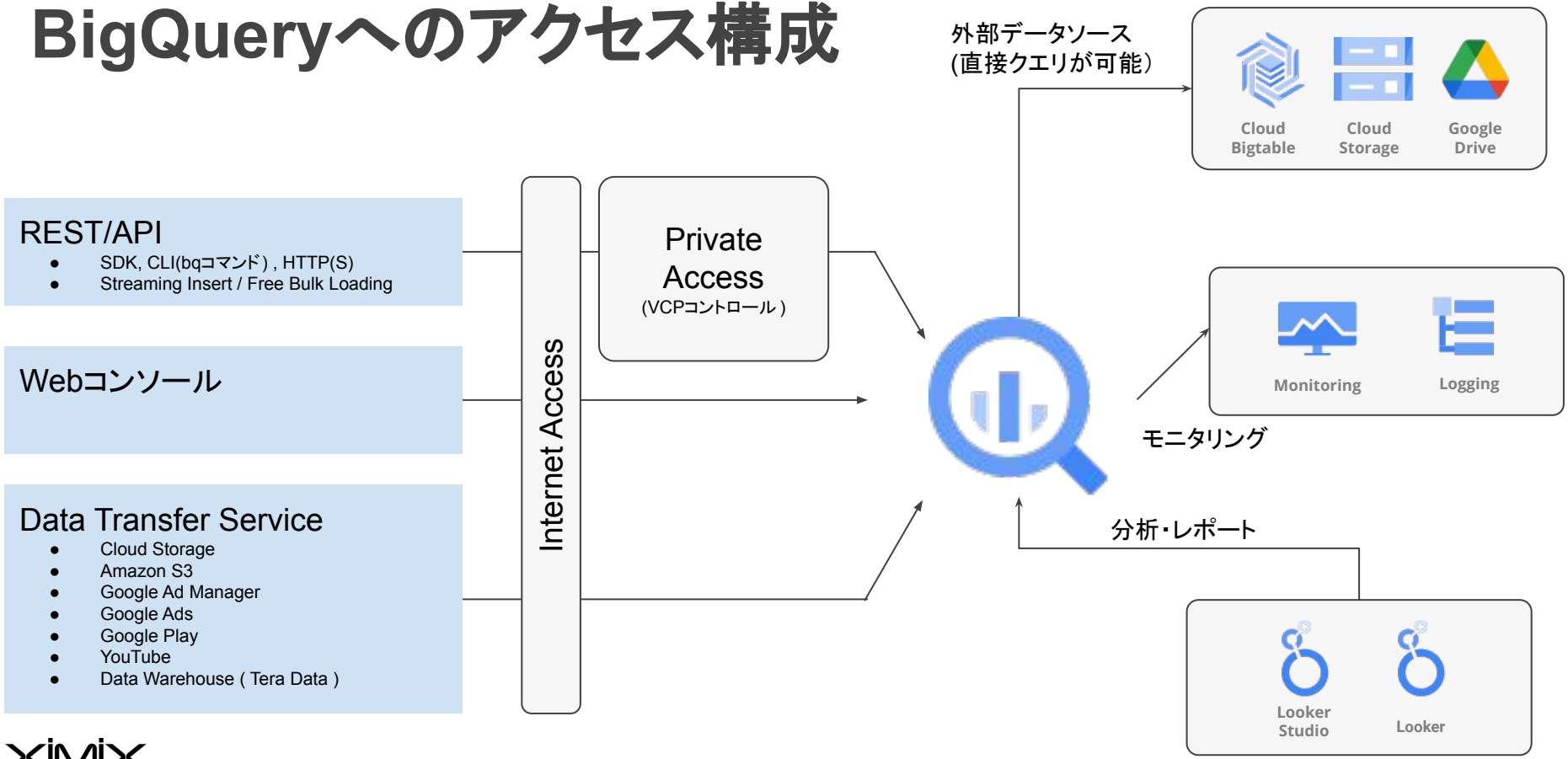
BigQuery

テーブルのデータソース

テーブルのデータは以下の方法により取り込み可能です。

取り込み方法	ファイル形式
ファイルアップロード	CSV , JSON , Avro , ORC , Parquet ※上限10MBまで
GCS	CSV , JSON , ORC , Avro , Parquet Datastore エクスポート ファイル Firestore エクスポート ファイル
Google Drive	CSV , JSON , Avro Google スプレッドシート
Bigtable	Bigtable
Data Transfer Service （定期的なSaaSデータ取り込み）	GCS、 Google広告レポートデータ、 Youtube チャンネルレポート、 Teradata、 Amazon S3 、 Amazon Redshift 等

BigQueryへのアクセス構成



数値から見る比較

BigQueryは、小さいQueryに対しても1秒以上かかる代わりに大きなデータの取扱の際に非常に早く実施でき、またパフォーマンスの劣化が少ない。

試験項目	パターン	アプライアンス	BigQuery	ソフトウェア版
実行クエリ(構造/インデックス)	100M相当	0.035	2.660	0.011
	1G相当	0.062	1.800	0.035
	2T相当	107.420	5.080	3.699
実行クエリ(キャッシュ)	100M相当	0.035	1.480	
	1G相当	0.062	1.280	
	2T相当	107.420	2.020	
実行クエリ(データ分散) 2TByteテーブルのJoin処理	パターン1	178.210	75.720	75.588
	パターン2	71.720	16.780	24.060
実行クエリ(パーティショニング)	100M相当	0.026	3.500	0.030
	1G相当	0.061	2.920	0.018
	2T相当	0.057	2.940	0.050
実行クエリ(UDF)	100M相当	0.027	6.800	0.009
	1G相当	0.064	7.080	0.009
	2T相当	73.400	9.240	0.259

2018年度時点の比較

既存アプライアンスとの比較

大手アプライアンス製品とBigQueryの特性を比較した資料となります。

	DWHアプライアンス	BigQuery	影響
構造	行指向	列指向	DMLの性能特性が異なる
インデックス	なし	なし	列指向ではデータ＝インデックスとなる。クエリ性能に差異
キャッシュ	なし	あり	繰り返し処理の性能に差異
データ分散キー	あり	なし	結合処理の性能に差異
パーティショニング	なし ゾーンマップが類似	あり 分割テーブル	Netezzaの場合にはソートキー＝ BigQuery パーティションキーとなる
マテリアライズドビュー	あり	なし	特定クエリ性能に差異
費用	機器費用＋保守費用	ストレージ、クエリ量による課金	テーブルの物理分割で費用低減

AWS RedShiftとの比較

AWS RedShift との特徴的な部分の比較（各機能については常に更新されており以下の資料との相違点がある場合があります。
実際の利用の際には各ドキュメントをご確認ください）

	RedShift	BigQuery	BigQuery側の補足事項
運用	インスタンス管理が必要	不要（サーバレス）	事前のデータ容量、パフォーマンス等の適切なプラニングが不要
費用	インスタンス＋ストレージ課金	従量課金/定額課金	ストレージ料金、検索結果の読み込み料金（クエリ課金）で請求が行われる。初期の投資が少ない。大量用途では定額プランが用意されている
チューニング	クラスタリング・インデックス	パーティション	インデックス管理は不要、パーティション（日付分割、取り込み時間、整数値）により読み込み量を制御可能(Query費用、速度）
テーブル制約	UDF(独自言語）、トランザクションあり	UDF(Javascript)、日本語カラム×、トランザクションなし、特殊な型（配列型、地図あり）	配列型があることによりJSONなどの形式との親和性が高い。UDFもJavascirptでかけることから学習コストが低い
外部データ参照（外部表）	RedShift Spectrumにより実装	標準で、GCS、Gドライブ内のデータを外部表として参照	GCSを参照することでBQストレージの料金を軽減可能。速度としてはBQストレージが早いがGCSの安いプランが利用できる
接続の制限	ODBC/JDBC	API (ODBC/JDBCはサードパーティ製)	
メンテナンス・バックアップ	メンテナンスウィンドウあり、 最大35日スナップショット保持	メンテナンスウィンドウなし 7日以内任意の地点を取得可能	BigQueryはロールバック出来ないが7日分の変更履歴を保持しておりある時刻へのスナップショットアクセスが可能
アクセス制限		データセット・テーブル単位	BigQueryを利用したすべての操作ログ、一時結果は管理コンソールよりアクセス可能
暗号化	標準	標準	データ管理の暗号化鍵の利用も可能
監視	S3(監査ログ), CloudTrail(操作ログ)	StackDriver	リソースの監視、監査ログをStackdriverに保管することが可能

各種お問合せ先

E-Mail: contact_ximix@niandc.co.jp