

増大するデータ量 への対応



⑦増大するデータ量対応

IoT／ビッグデータなどで絶えず増加するデータの保持を効率的に実施する

関連する主要サービス

S3

Kinesis

Glacier

Glue

Athena

EMR

QuickSight

Redshift



データ量への対応

効率的なデータ蓄積とIoTなどの大量のストリームデータ処理
や解析の方法が必要不可欠となる



ビッグデータに必要な技術

ビッグデータに対応したデータ蓄積・処理技術が必要不可欠

Volume
大量データ

- ✓ 大量のデータを効率的に蓄積可能なデータベース技術

Variety
多様なデータ

- ✓ 多様な形式のデータを蓄積可能なデータベース技術

Velocity
速い処理

- ✓ 高速処理が可能なデータ処理ソフトウェア／ハードウェア

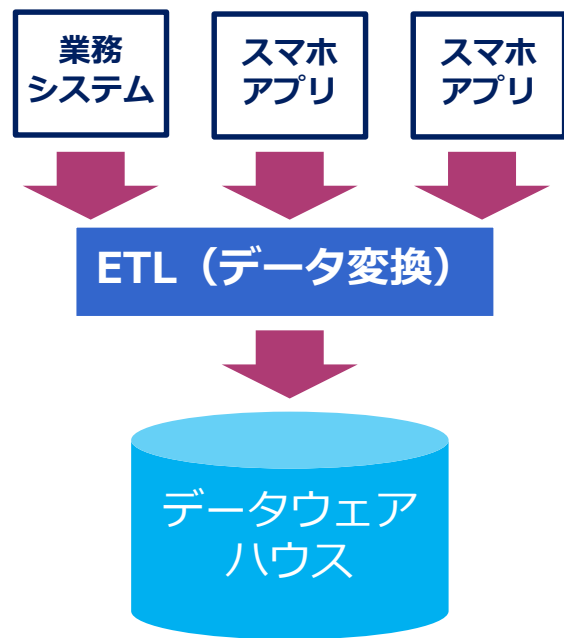


データレイクの活用

ビッグデータ活用の中心はデータレイク型データベース

データウェアハウス中心

利用用途に応じたデータを貯めて活用する
データウェアハウス



データレイク中心

出来る限り生データをほぼ全データ保存する
データレイク



データレイクの活用

データレイクでは全データを生データのまま保存する

	データウェアハウス中心	データレイク中心
データ収集	<ul style="list-style-type: none">✓ 目的別データ ⇒必要なデータのみを抽出・収集✓ 構造化データ中心	<ul style="list-style-type: none">✓ 生データ+目的別データ✓ 構造化／半構造化／非構造化データ
蓄積	<ul style="list-style-type: none">✓ 必要なデータのみを抽出・蓄積	<ul style="list-style-type: none">✓ 変換しないで生データ形式で保存✓ もしくはエッジ処理したデータを保存
処理・加工	<ul style="list-style-type: none">✓ 関連するデータ構造（スキーマ）に変換・蓄積✓ SQLによる操作	<ul style="list-style-type: none">✓ 事前にスキーマ（データ構造）を定義しない✓ SQL/SAS/MapReduce/R/NoSQLなどで操作
可視化分析	<ul style="list-style-type: none">✓ 利用者がデータ分析／レポート内容などの利用目的を事前に特定し構築	<ul style="list-style-type: none">✓ 事前に目的を定義せず、ユーザーがデータ群から新たな価値を抽出しデータを解釈・活用



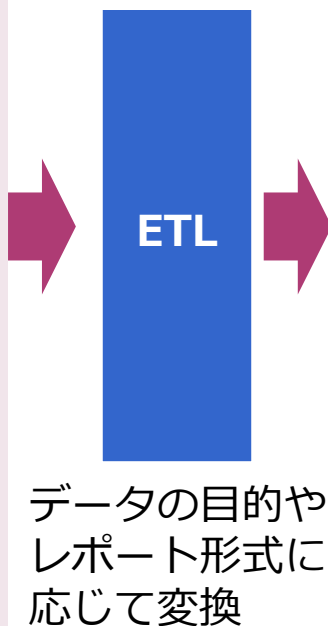
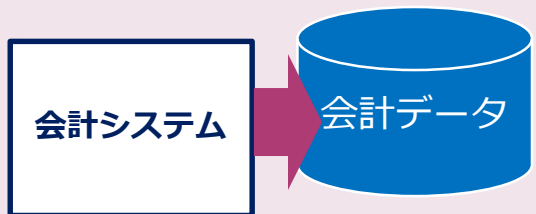
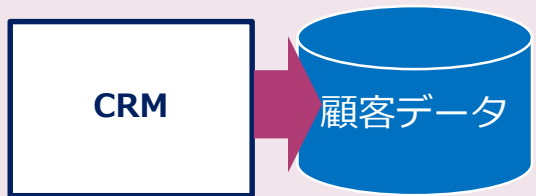
データウェアハウス型のデータ処理基盤

データ収集

データ加工・蓄積

データ活用

業務システムの現在データ



過去から現在
までのデータ



目的別に抽出したデータ

OLAP



データウェアハウス型のデータ処理基盤

データ収集

データ加工・蓄積

データ活用

業務システムの現在データ

事前に保存する
データとその構
造を決めて変
換・保存する

CRM

顧客データ

会計システム

会計データ

ETL

データウェア
ハウス

過去から現在
までのデータ

データの目的や
レポート形式に
応じて変換

データマート

データマート

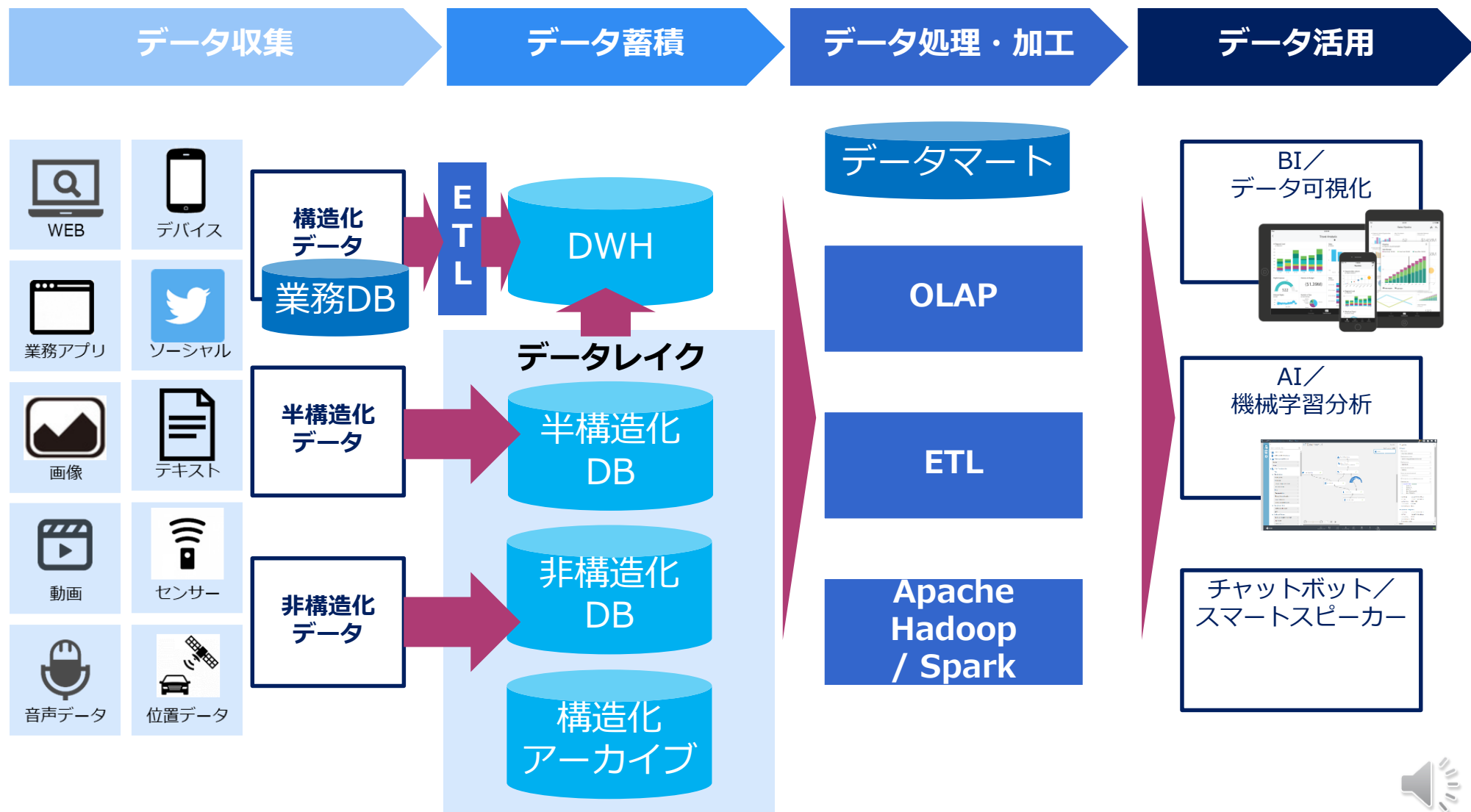
データマート

目的別に抽出したデータ

OLAP



データレイク型のデータ処理基盤



データレイク型のデータ処理基盤

データ収集

データ蓄積

データ処理・加工

データ活用

生データ形式／
様々な種類の
データを蓄積
↓
あとから加工

ETL

DWH

データレイク

半構造化
DB

非構造化
DB

構造化
アーカイブ

データマート

OLAP

ETL

Apache
Hadoop
/ Spark

BI/
データ可視化



AI/
機械学習分析



チャットボット/
スマートスピーカー



Apacheシリーズ

ビッグデータ分散処理向けの代表的な仕組み（ミドルウェア）がApacheシリーズ

大量データバッチ処理向け

Apache
Hadoop

ストリーミング処理向け

Apache
Spark



AWSのデータレイク構成

データ収集

データ蓄積

データ処理・加工

データ活用

