弊社における Hadoop基盤の活用事例

2014年11月28日

自己紹介

- 氏名

  山本　直人（やまもと　なおと）

- 所属

  アクセンチュア株式会社　テクノロジーコンサルティング本部
  マネジャー

- 担当領域

  テクノロジー・アプリケーションアーキテクチャ

  ビッグデータ基盤技術

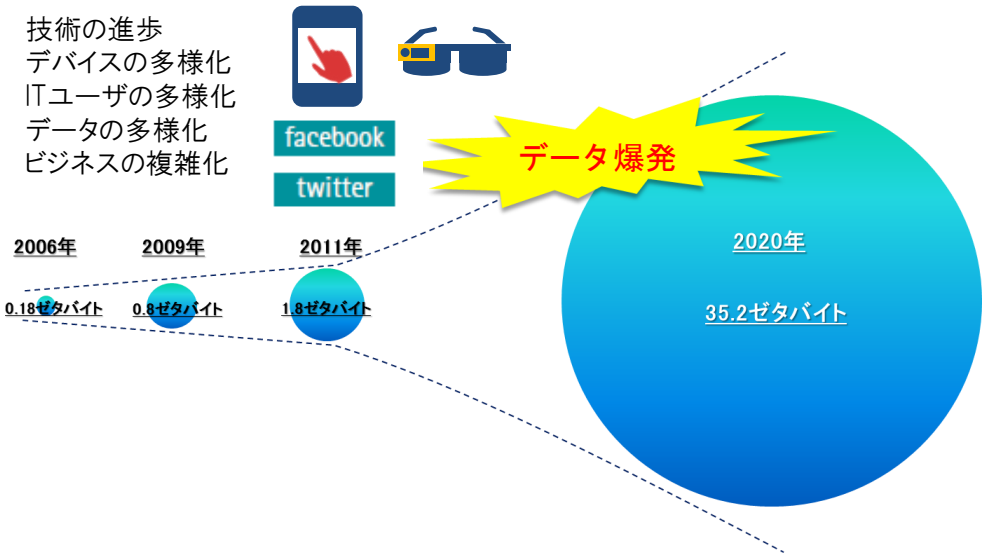
- その他

  前職は、ノーチラステクノロジーズの前身のあの会社

  好きな言語はHaskellとAsakusaDSL

（今更ながら、、）データ爆発・ビッグデータ？

- 技術の進歩
- デバイスの多様化
- ITユーザの多様化
- データの多様化
- ビジネスの複雑化



出典：IDC Digital Universe Study

- Googleは、2004年に1日あたり100テラバイトのデータを処理、また2008年には1日あたり20ペタバイトを処理するまでに成長
- Facebookでは、毎週60億のコンテンツをユーザが共有、毎月30億の写真がアップロード、毎秒100万枚の写真が参照、Facebookのサーバでは、毎秒5000万件の処理を実行
- NYSEでは、1日ごとに1テラバイトの新たな取引データを生み出している
- The Internet Archiveでは、およそ2ペタバイトのデータを保有しており、その容量は1ヵ月あたり20テラバイトのペースで増加
- CERN（The European Organization for Nuclear Research（欧州原子核研究機構））は、毎日40テラバイト、年間で15ペタバイトのデータを生成

◆ 既存テクノロジーの限界

- バッチ処理時間長期化により、実行頻度に制限、もしくはそもそも処理不能
- バッチ実行頻度が限られることによる、データ鮮度の低下（価値の高い情報が得られない）
- SQLによる多様なデータ構造の取り扱いの限界
- 今後増え続けるデータに対応するためのハードウェアスケールアップコスト高騰
- ・・・

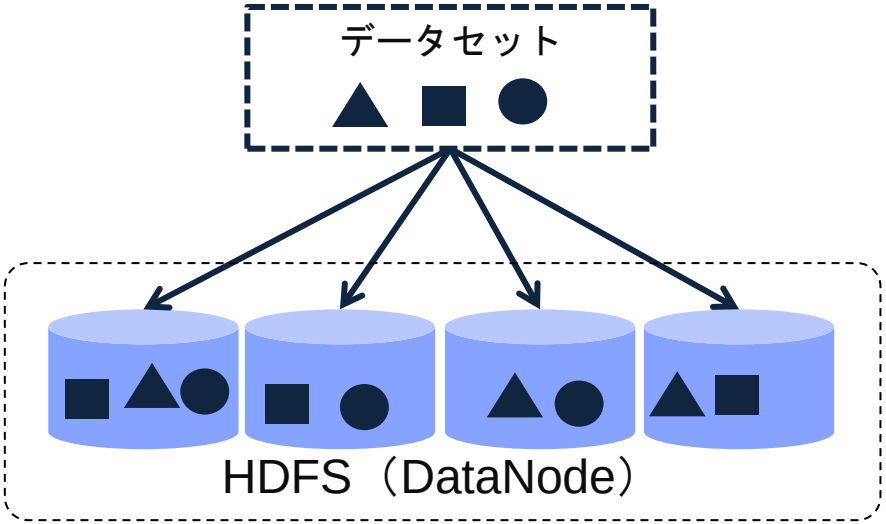
戦略的なデータ活用ができない。。。

超概説：そもそもHadoopって？

- TB・PB以上の大規模データに対する分散処理基盤
- 大量データを確実かつ効率的に処理するために、大きく以下の要素で構成

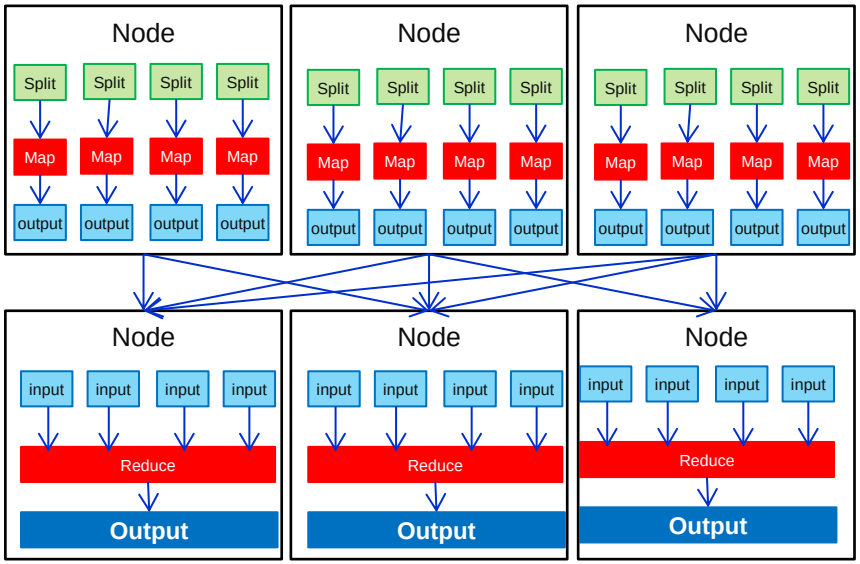
HDFS

- ✓ 高信頼性ファイルシステム
- ✓ 大規模データをブロックに分割し、複数サーバに分散して配置（3つのレプリカを作成）



MapReduceフレームワーク

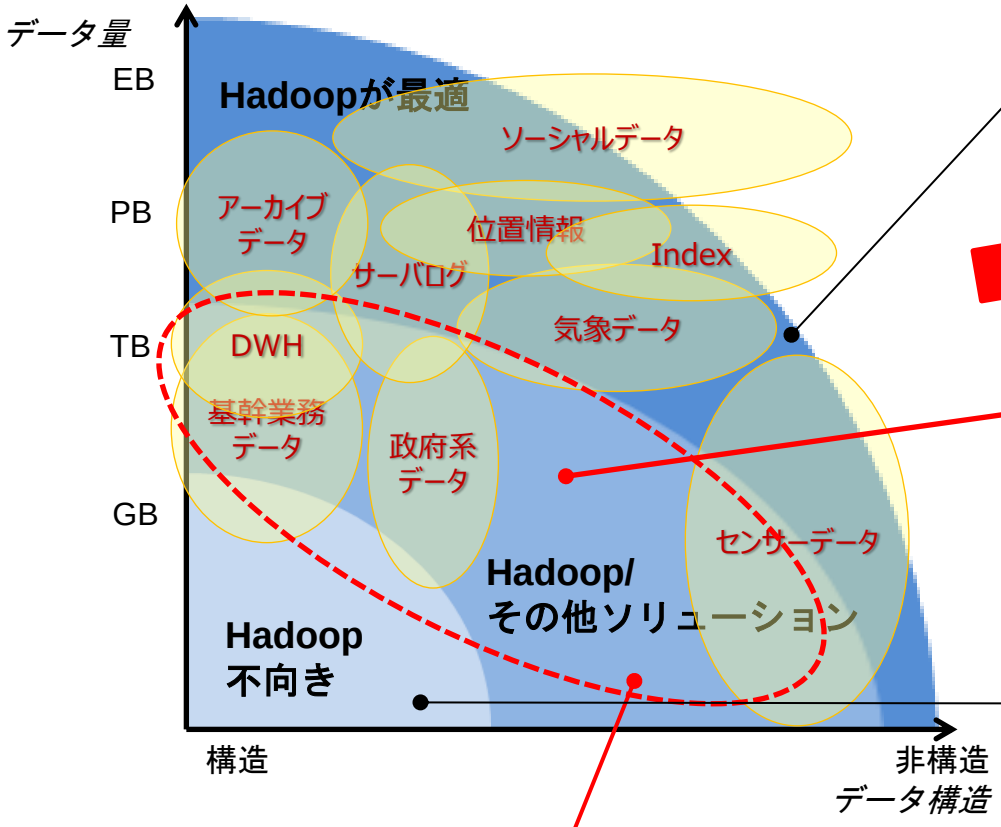
- ✓ 並列分散処理のためのフレームワーク
- ✓ HDFS上に分散配置されたデータを並列処理（Map処理、Reduce処理）することで高いスループットを実現



大規模データに対して、信頼性を担保しつつ高スループットな処理を実現

Hadoopを使った基盤構築の狙いどころ

HDFS/MapReduceの処理特性として、バッチによる大規模データ、非構造データに対する集計処理等がターゲットとなる領域と考えます。



Yahoo、Amazon、Facebook、Twitter等の先進企業での分析系システム

- 超大量ログ分析（行動履歴解析、傾向分析等）
- インデックス作成
- ページランク作成
- 位置情報・天候情報等分析

SIの主戦場

大企業、中央政府機関の基幹システム

- 大規模なバッチ処理（大量トランザクションレコード処理・複雑なJOIN・ネスト構造をもつSQLの置き換え、DWHに対するオフロード処理等）
- 大量レコードの集計・分析処理

企業等の小規模システム

- 普通のSQLでも処理は可能であるため、Hadoopは向いていない
- 非構造データや複雑な集計処理においてはピンポイント導入も可能

✓ 大量データを扱うバッチ（集計処理・分析処理等）でHadoop活用可能
✓ しかし、Hadoop以外の選択肢あり

Hadoop導入のポイント

弊社の主戦場は企業や政府の基幹システムであり、そこにHadoop導入していくにあたってのポイントを以下の通り考えてみました。

	ポイント
目的を定めた導入	**既存技術（SQL等）でボトルネックとなる大規模基幹バッチ・集計処理の置き換え** ✓ 大規模基幹バッチ処理への適用（大量レコード、複雑なSQLが必要） ✓ ログ等の非構造データの分析処理への適用 ✓ 大規模システムのバッチ基盤を置き換え ✓ スケールアウトモデルへの転換
評価・試験的導入	**成長規模が読めない状態での新規ビジネス開始時の評価・試験運用として導入** ✓ 安価に導入可能（HadoopはOSS）であるため、試験的に導入 ✓ AWS等Cloud利用で余計なコストを更にカット ✓ （KPIを定め）実績を評価することでスケールさせることが可能

そこで、実際に各提案でHadoop基盤導入をトライ！

Hadoop導入の提案にチャレンジ

これまでもシステム開発の提案において、対象データや処理がHadoopに適していそうな案件にHadoopのソリューションを提案してみたものの。。。

提案システム名	処理対象のデータ	Hadoop上での処理
Webシステムのデータを使って顧客360度ビューで分析する基盤	• 超大量のログデータ • アクセス履歴 • オープンデータ	外部が提供するオープンデータとWebサイト内で蓄積された履歴データでユーザ個別のリコメンドを提供
基幹系業務における大規模バッチ処理の基盤	• 大規模なトランザクションデータ	日次バッチ処理で、最大5000万件以上のトランザクションデータを処理
画像データの解析用基盤	• 非構造化状態のテキストデータ	画像ファイルを専用ロジックでテキストに変換し、そのデータから、画像を解析

Hadoopが採用されなかった理由(弊社視点)

どうしてHadoopがソリューションとして採用されなかったかを、弊社独自の視点でみると、以下の理由が挙げられます。

提案システム名	採用されなかった理由	
Webシステムのデータを使って顧客360度ビューで分析する基盤	ソフトウェア会社とのアライアンスの壁	大量データを処理するBigdata基盤のシステム構築で選定されるソリューションは、アライアンスの関係で、まず、SAP HANA、TERADATA、EXADATAが候補として上げられる。
基幹系業務における大規模バッチ処理の基盤	UIによって分析の効果を訴求	大量データを簡単なHadoop基盤上で複雑なロジックを用いて分析したとしても、スマフォ、タブレットのアプリ上やTaburauやPower BI等のビジュアルなUI、グラフで結果を表現した上で、効果を語れないと受け入れられない。
画像データの解析用基盤	Hadoopを知らない、または既存ソリューションを選択	バッチ基盤にHadoopを選択せず、既存ソリューションを選択する。そのため、バッチ基盤に「Hadoop」が利用されない。提案、営業する人がHadoopのソリューションで何ができ、何に向くのかを知らない。

Hadoop＋UIで、効果が見える事例を作り、Hadoopを普及していくことが不可欠

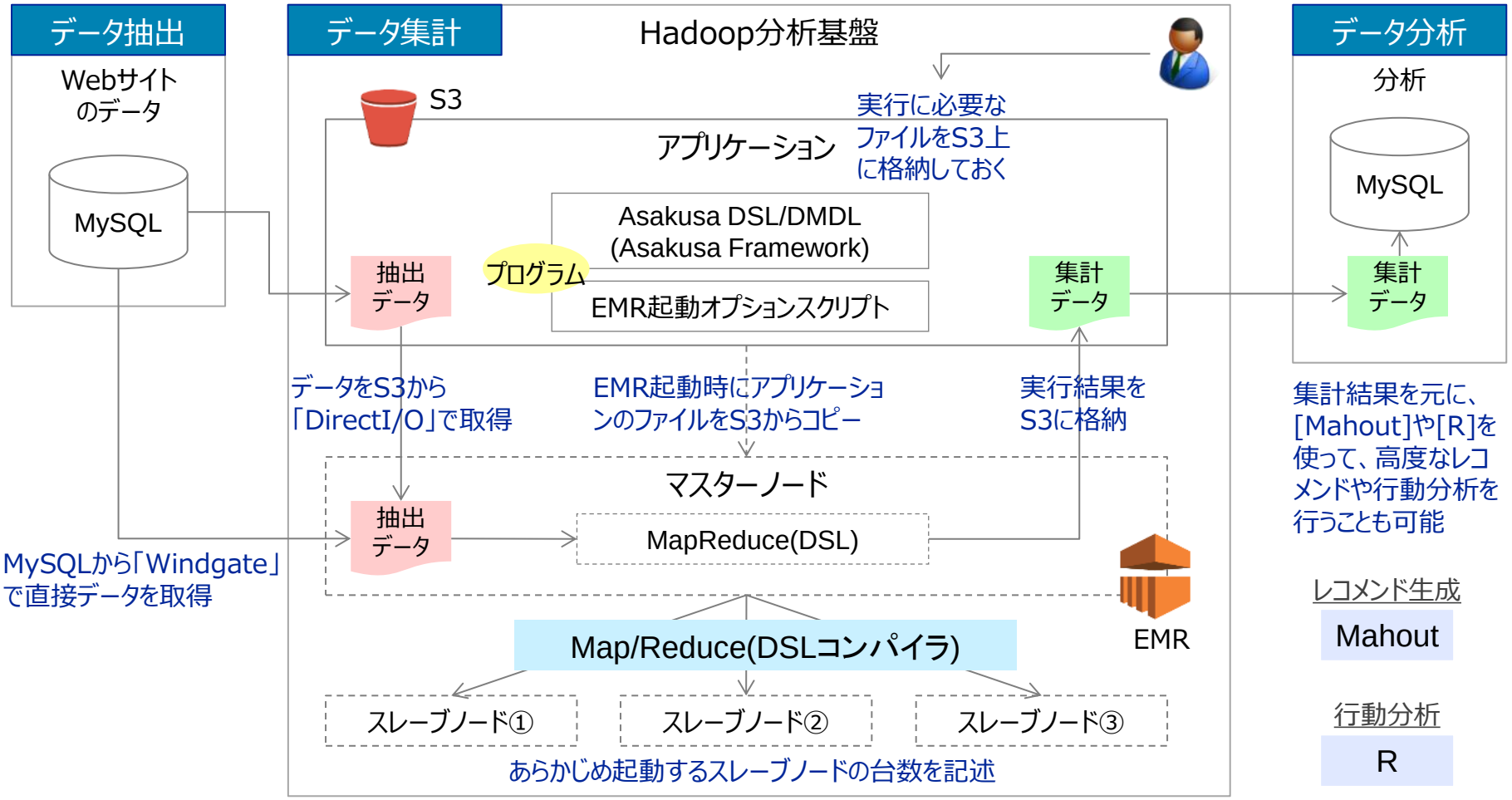
弊社グローバルでのHadoop活用事例

弊社のグローバルの事例をみると、大規模データを対象としたシステム基盤の構築で、既存のソフトウェアからHadoopに置き換えています。

クライアント	処理概要	ビジネス上の課題	課題解決のためのソリューション
大手ネット銀行	オンラインバンキングの取引で、アプリケーション上の操作をトラッキングし、RDBMSで分析、ユーザの導線やアプリケーション改善に活用	RDBMSで処理を実行していたが、スケールアウトが限界で、1秒間に5000メッセージを処理することができなくなった。	リアルタイム処理をNoSQLで、分析基盤をHadoop＋MapReduceで構築し、大規模データをスケールアウトで対応可能とした。
ホスティング会社	顧客が利用するプラットフォームのサービスや品質を改善するために、モニタリングしたデータをBIツール(Infomatica)で分析	顧客の要望やプラットフォーム要件の増加で、扱うデータも増え、既存のBIツールだけでは、対処しきれなくなった。	Hadoop基盤でHiveベースの集計アプリと既存BIツールを統合し、要件にあうボリュームのデータを分析可能とした。
保険・資産運用会社	顧客の資産運用で、アナリティクスの情報や蓄積データからプロアクティブに提案を行ったり、法改定に対応したサービスを提供	SASベースでシステムを構築していたが、SASは機械学習よりはデータ変換でのみ使われていた。	Hadoopにより、イテレーションの分析を行え、ユーザにアドホックに情報を提供できるようにした。

HadoopによるWebサービスのデータ分析基盤

Asakusa＋Amazon Web ServicesのEMRで、Web内のエンドユーザの行動を把握するための分析基盤を構築してみました。



弊社リコメンドサービスの紹介

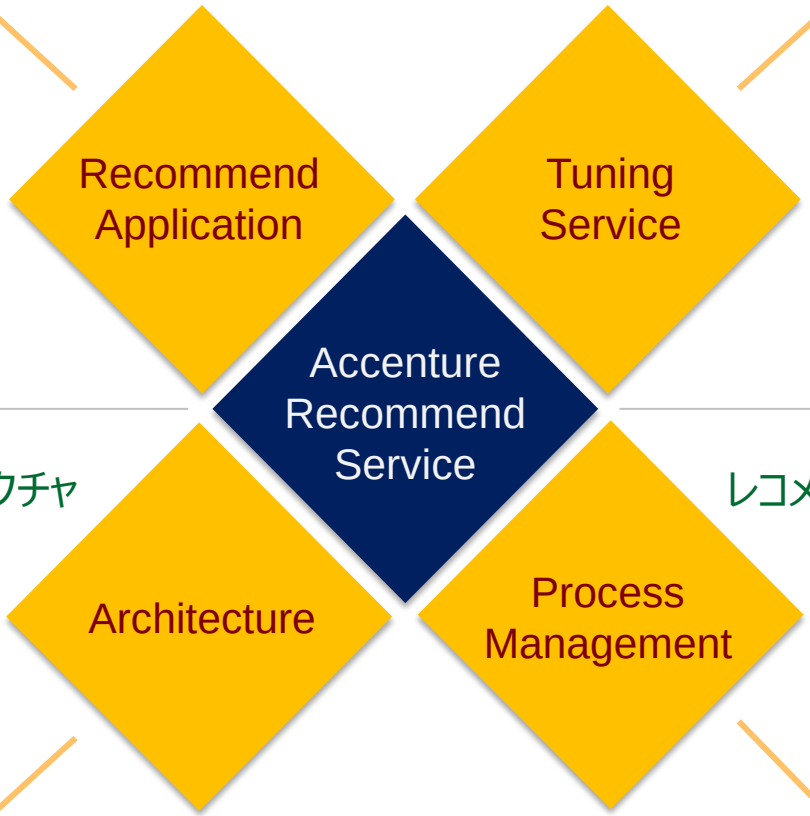
アプリケーションやアーキテクチャといったシステムの構築だけではなく、データ分析に基づくエンジン最適化、レコメンド業務の設計から定着化までトータルでサポートします。

幅広い適用可能領域

ARSアプリケーションは、ECサイトでのクロスセル・アップセルのみならず、メールマガジン中の商品広告やキャンペーン情報送信対象選定やニュースサイトにおける関連記事への誘導、クーポン送付による実店舗への誘導など、幅広い領域への適用が可能です。

専門家によるエンジン最適化

一般的にマーケティングに用いられる顧客属性データやPOSデータといった構造化データを始めとし、アクセスログやセンサデータ、ソーシャルメディアなどのビッグデータから価値を生み出す、弊社データ解析のエキスパートがレコメンドエンジンのチューニングを行います。



スケーラブルで柔軟なアーキテクチャ

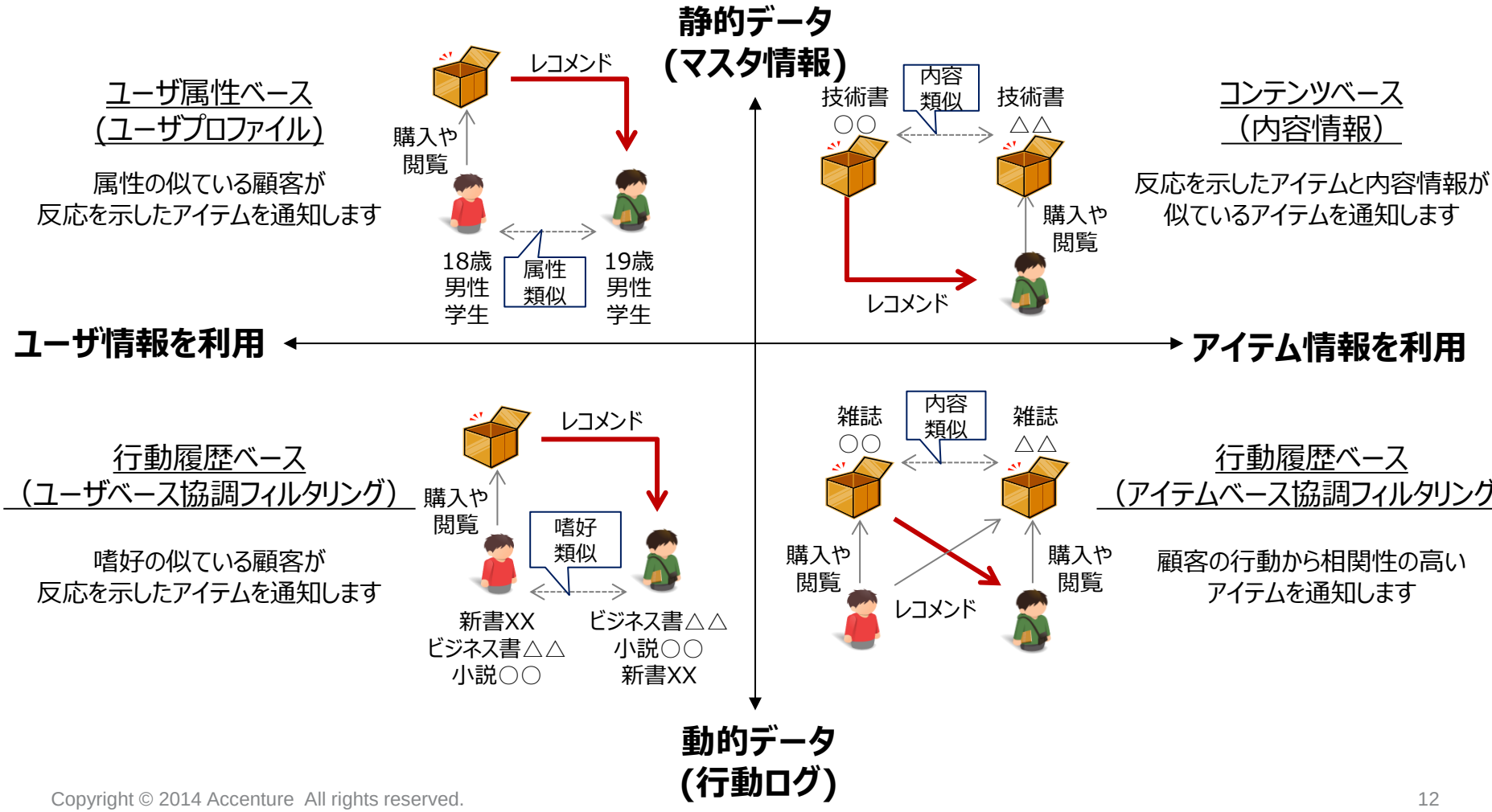
ARSは安価なIAサーバでスケールアウト可能なアーキテクチャを採用しており、急激な需要の増大に対しても迅速な対応が可能です。また、各モジュールは独立した設計となっているため、モジュールの追加や変更に対して柔軟な対応が可能です。

レコメンド業務の設計・定着化支援

レコメンドはシステムを導入するだけで全てを解決できるわけではありません。
分析・ターゲティング・効果検証を始めとしたレコメンド独特の業務プロセスについて、弊社の豊富な経験・実績から設計・導入・運用まで一気通貫したサポートが可能です。

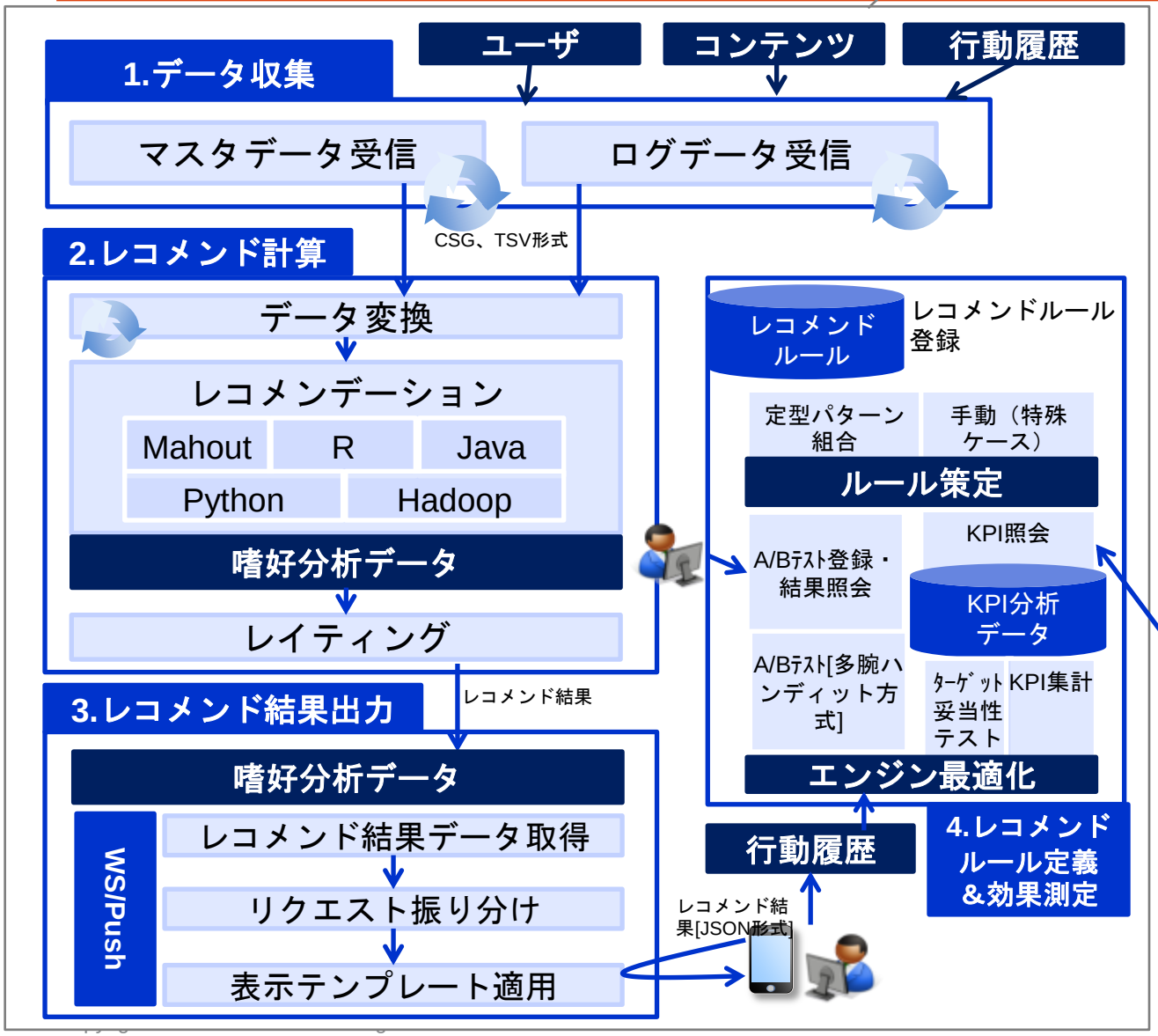
レコメンドアルゴリズムパターン

例えば、ECサイトにおいて、以下の4パターンのレコメンドアルゴリズムが利用可能で、4つのロジックパターンを駆使して、確度の高いレコメンドを実現可能です。



ARS - サービスの論理構成

ARSはレコメンドだけではなく、「異常検知アルゴリズム分析」、「最適品揃え分析」など、汎用分析基盤としても活用できます



1. データ収集
テキストファイル（CSV、TSV）形式によるデータ受信が可能

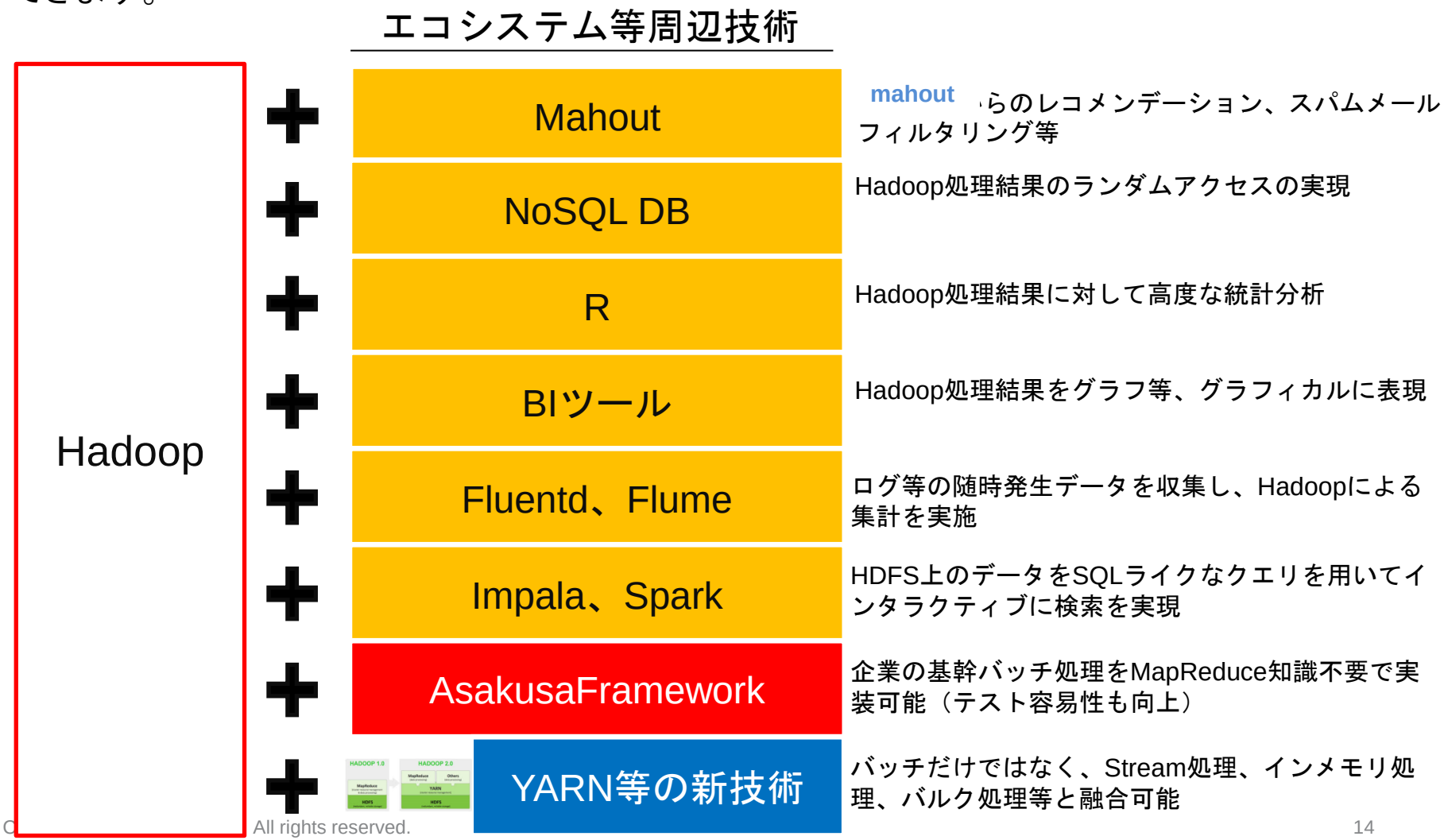
2. レコメンド計算
JavaとHadoopをベースとし、Mahout、R等のOSSを使用した柔軟なカスタマイズが可能レコメンドエンジン

3. レコメンド結果出力
WebServiceによるレコメンド結果のPull、および外部WebServiceへのPushが可能

4. レコメンドルール定義&効果測定
ルール設定：WebUIによるレコメンドルールの登録、及び編集が可能
エンジン最適化：WebUIによるKPI（照会・編集やﾀｰｹﾞｯﾄ妥当性テストの結果照会、A/Bテスト（多腕ハンディット方式）の登録・結果照会が可能）

エコシステムの活用でいろいろできる！

Hadoopとエコシステムを組み合わせることで、大規模データを様々な用途に活用することができます。



まとめ

前述した「Hadoop導入のポイント」を踏まえ、弊社内のHadoopの導入実績からみて、以下のような観点を抑えた上で、Hadoopを推薦することが重要ではなかろうかと考えます。

Hadoopの特性	大規模のバッチ処理には向くことを証明	大規模なデータを扱うには、データ分割・パーティションときてスケールアウトが当たり前。データボリュームが増えてもスケールアウトで対応できる、テクニカルなHadoopの良さを説明
エコシステム等との組み合わせ	経営層にも受けるようなUIを用意	バッチ処理を実行するところだけみせても受けない。経営層には、バッチ処理の結果、何がどう変わるのかを説明 Hadoopの処理結果をコンテンツとした、ビジュアルなプレゼンテーションの実現
スモールスタートからのスケール	Hadoopを知らなくても業務的に何ができるかを証明	ECのリコメンド用のデータをHadoopで分析していることを知らなくても、目的に対して（マーケティング施策等）、どう効果測定できるかを安く作って証明

ご清聴、ありがとうございました。