「MapRによるHadoopの最新情報 ～市場動向とユースケース～」

2014/11/28

マップアール・テクノロジーズ株式会社
アライアンス&プロダクトマーケティング
三原　茂
smihara@mapr.com

アジェンダ

1. MapR社のご紹介
2. Hadoopとは？
    1. Hadoopのマーケット状況
3. Hadoopの位置付け
    1. DWH
    2. BI
    3. ストレージ
4. MapRディストリビューション
5. まとめ

MAPR®

MapR社概要（1）

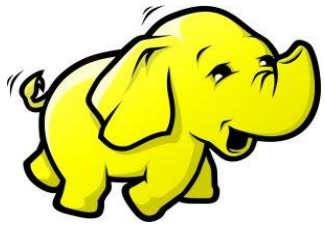
MapR Technologies Inc.：

【Founder】 John Schroeder & M.C. Srivas
【設立】 2009年、カリフォルニア州サンノゼで設立
【従業員】 約 300 人
【拠点】 13カ所（米国、イギリス、フランス、ドイツ、シンガポール、オーストラリア、日本、韓国、スウェーデン）
【URL】 https://www.mapr.com/

日本法人：

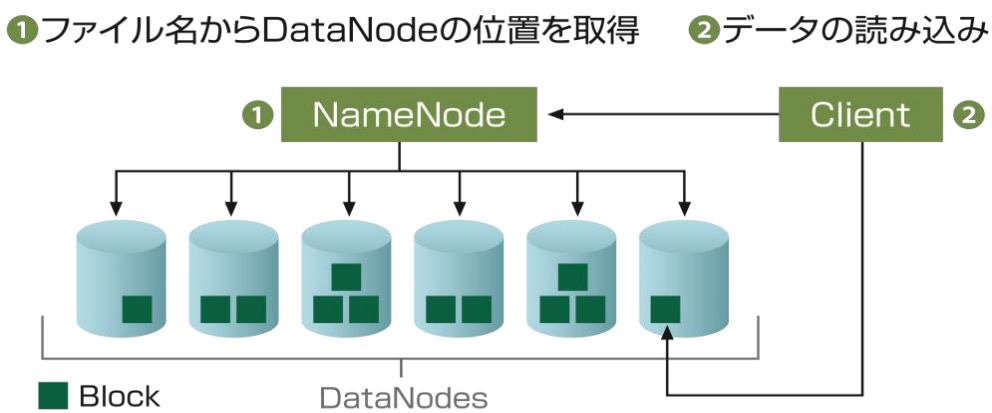
【会社名】マップアール・テクノロジーズ株式会社
(MapR Technologies K.K.)
【所在地】〒100-0005
東京都千代田区丸の内1-8-3
丸の内トラストタワー本館20F
Tel: 03-5288-5370
eMail : sales-jp@mapr.com
【設立】 2013年4月

Hadoopとは？

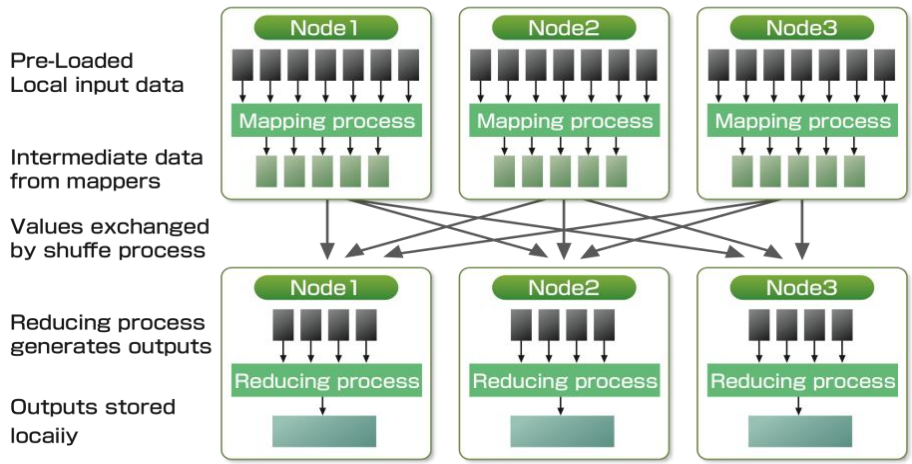


大きく**2つ**のコンポーネントで構成：

HDFS (Hadoop Distributed File System) 分散ファイルシステム



MapReduce 大規模分散処理フレームワーク



データをためる

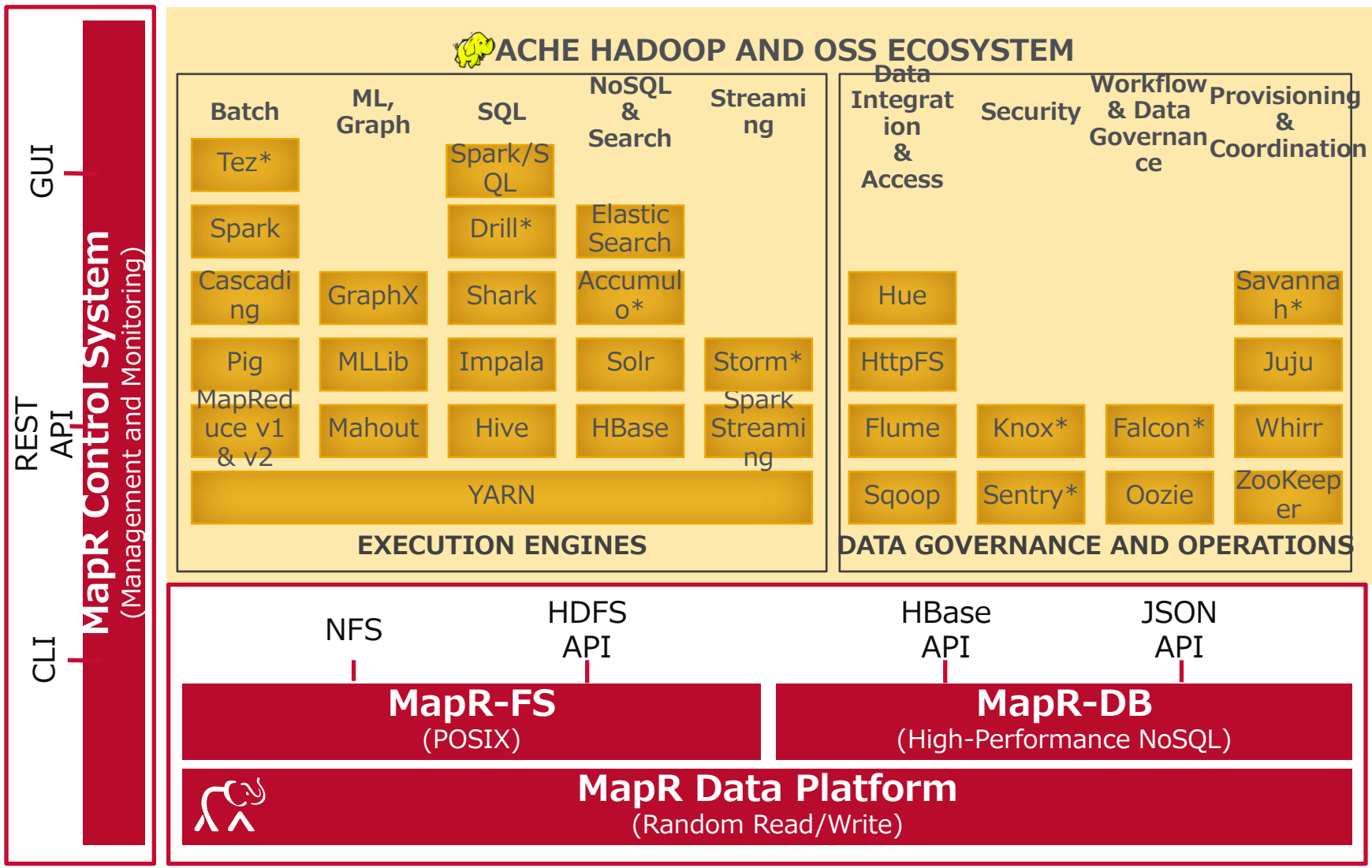
データをブロックに分割して複数のサーバに分散配置／3つのレプリカを作成

データを加工する

Map/Reduceというシンプルな処理の組み合わせで、HDFS上にあるデータの分散処理を行う汎用的なフレームワーク

- 処理の近くにデータを置く：データ（保管）と処理能力（加工・分析）をデータのある場所で！
- 設計当初から大規模、大容量、増加し続けるデータに対応（分散処理＆スケールアウト）

MapRのパッケージ



* In Roadmap for inclusion/certification

各Hadoop ディストリビューションの違い

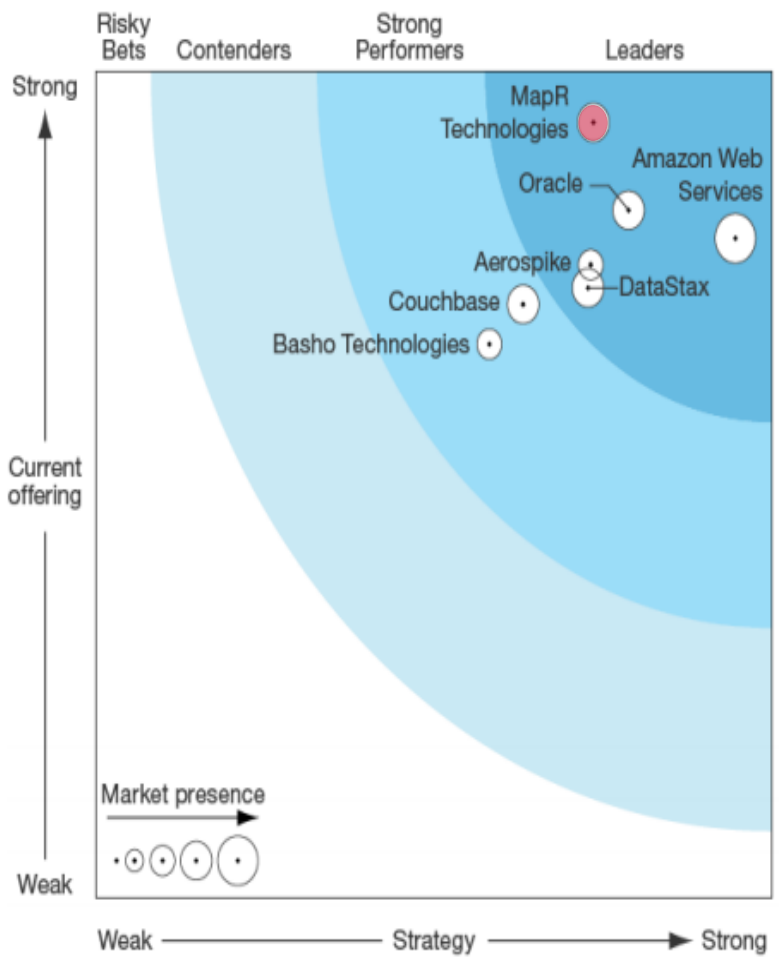
先進性や進化

Apache Hadoop	各商用 ディストリビューション	MapR
	管理機能 イノベーション	管理機能 イノベーション
OPEN SOURCE hadoop	OPEN SOURCE hadoop	OPEN SOURCE hadoop
		インフラストラクチャ イノベーション
	保守/教育サービス	保守/教育サービス

各Hadoopの評価

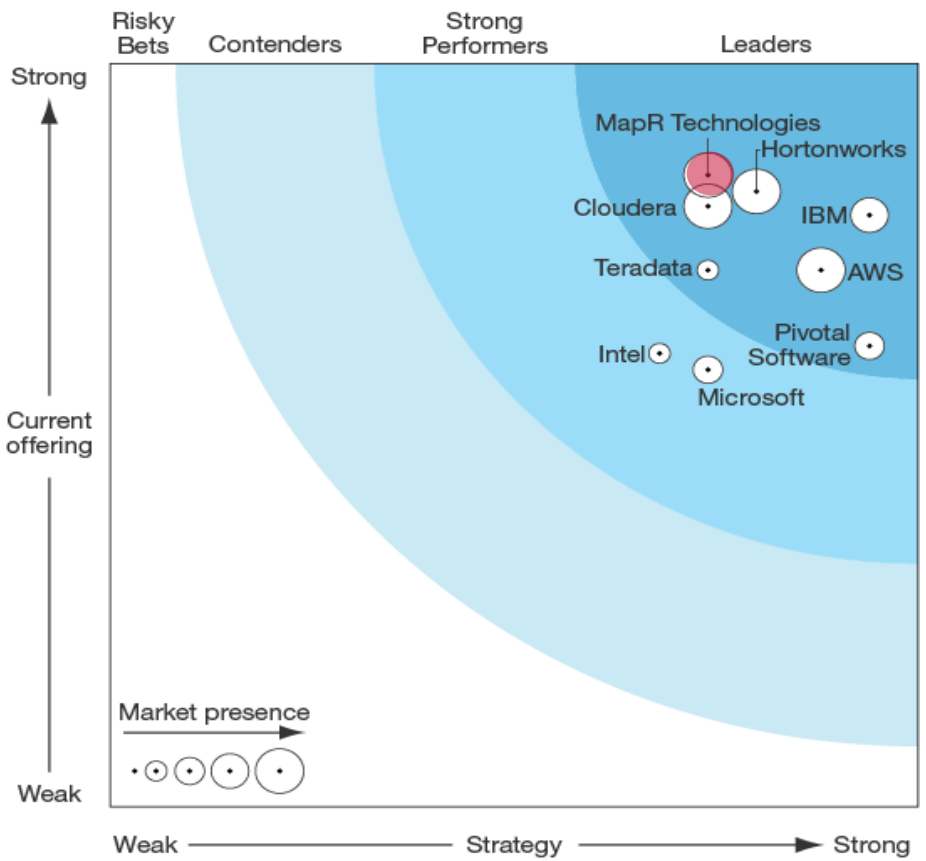
FORRESTER®

NoSQL: M7 (MapR-DB)



Forrester Wave™ NoSQL Key-Value Databases, Q3 2014

Hadoop: M5



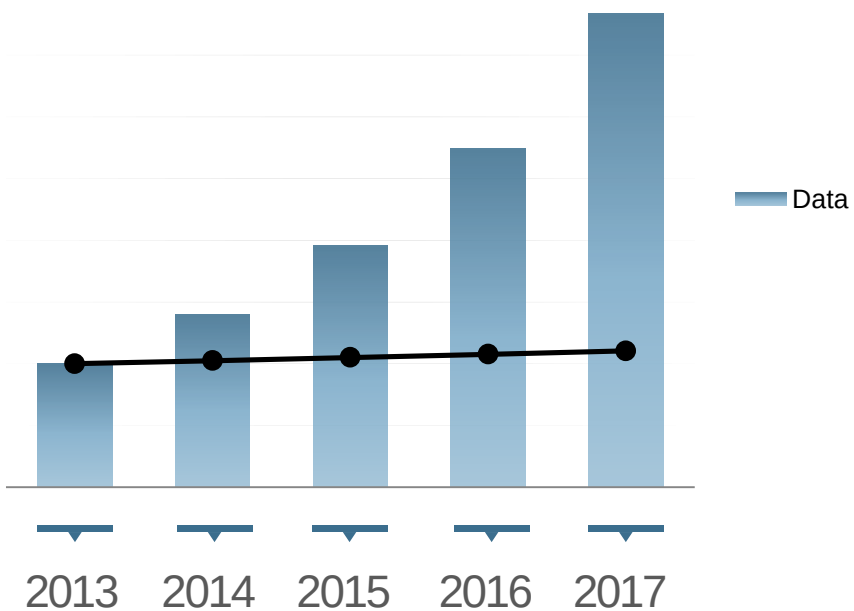
Forrester Wave™Big Data Hadoop Solutions, Q1'14

破壊的技術と採用へのモチベーション

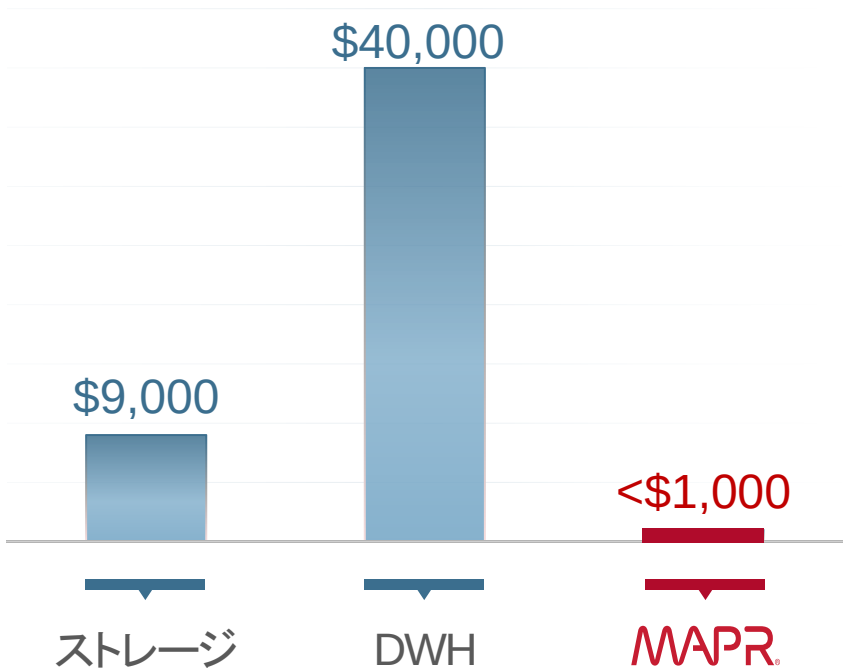
Hadoop as “disruptive Technology”

IT予算の伸び率
2.5%

データの増加率
40%



テラバイトあたりの
コスト



- Gartner, "Forecast Analysis: Enterprise IT Spending by Vertical Industry Market, Worldwide, 2010-2016, 3Q12 Update.“
- Wall Street Journal, “Financial Services Companies Firms See Results from Big Data Push”, Jan. 27, 2014

MAPR

Hadoop活用の主な目的

① セールス/マーケティング
（売上/収益向上）

② リスク回避
（予防保守/不正検知）

③ 情報系IT
（コスト削減）

④ 基幹系IT
（効率向上）

Asakusa

大手GMSチェーンのMapR活用例

Retail
Point of Sale System

Pharmacy
Information System

On-Line Shopping
Data

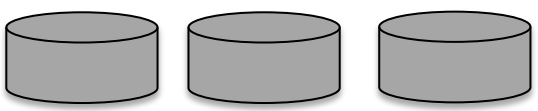
Loyalty Program System

Financial Data From
In-Store Bank Branches

Data From
Delivery Trucks

Business Problems

- 既存データはデータベースに関連無く広範囲に存在
- ロックされたバリューを簡単に解除することは困難
- 顧客統合視点では無い
- 効果的なロイヤルティプログラム
- 消費者の購買行動を部分的に理解
- 競合は顧客実装が進んでる ： 小売戦略中心
- ロイヤル顧客へのプロモーション



Financial & Logistics Data（構造化データ）

分析

ト
ティング
ンデーション

2500ノード

SNS, On-Line, POS, ポイント,位置情報ログ（非構造化データ）

コンバージョンレート

+2%

上乗せ収益

$1B

Business Benefits

- オンライン及びオフラインの行動を関
- 売れ筋商品の捕捉と理解
- どんな人（性別、年齢）が何時に何
- 動線・陳列最適化による商品クロスセ
  - レジ到達時間等の顧客のより包括的な視点活用
- クロスセリングからアップセリングを加速

MAPR

大手金融業のMapR活用例
～不正検知～

以前のシステム：RDB + DWH

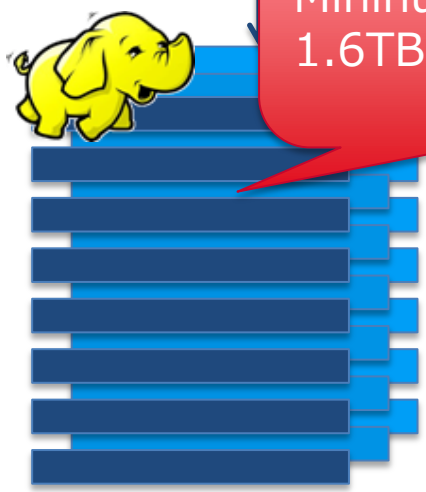
1TBデータ
抽出リピート

Data Analytics Tool

1 TB Memory Limit

100%のデータを
連続的に摂取

Terasort:45秒
Mininutesort:
1.6TB



2000ノード+

1. SASにデータ抽出 (一度に1TB)
2. 例外処理を検出
3. 前の分析を基本にクエリを改良
4. 第2の洗練されたデータサンプルを抽出 (1TB)
5. 更なる例外処理を検出
6. 必要であれば繰り返し

矛盾データ検知時間:
巧妙な場合には数時間から数日必要
巨大な違反は検出し停止させるのに1週間必要な場合がある

数分で矛盾検知

- MapReduce + Hive (データウェアハウス 構築環境) + R (統計解析向けプログラミング言語 & 開発実行環境) + Python (オブジェクト指向スクリプト言語) により数分おきに全のデータを連続的に分析
- 不正行為と違反を直ちに検知

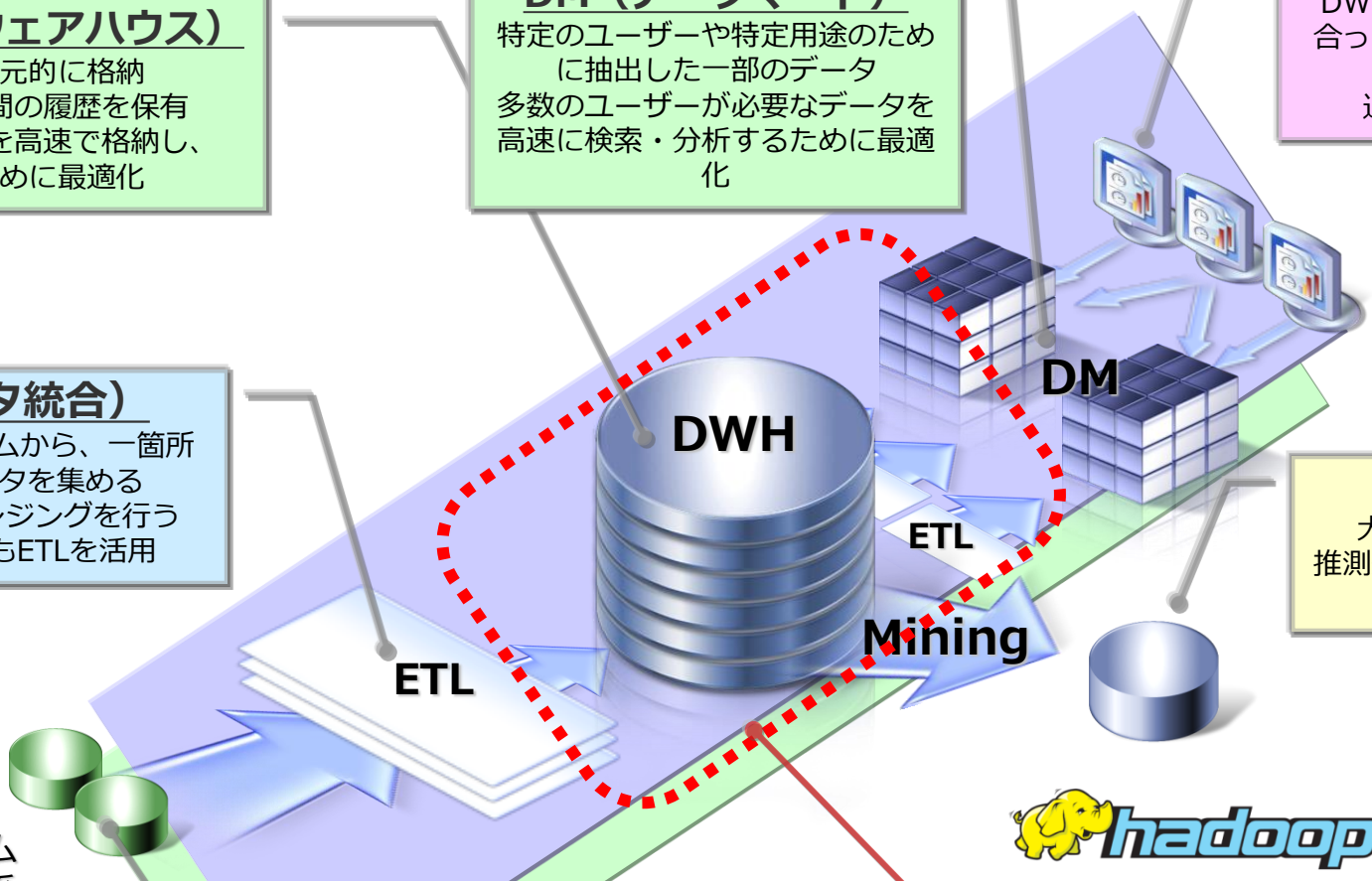
分析関連システムのどこに位置付くか？

DWH（データウェアハウス）
各種データを一元的に格納
大量の明細や長期間の履歴を保有
多数のソースデータを高速で格納し、
日々更新するために最適化

DM（データマート）
特定のユーザーや特定用途のために抽出した一部のデータ
多数のユーザーが必要なデータを高速に検索・分析するために最適化

BI(ビジネス・インテリジェンス)
DWH/DMのデータを分析ニーズに合った表、グラフ、レポートとして表示
通常Webベース(ブラウザ)

ETL（データ統合）
複数のソースシステムから、一箇所（DWH）にデータを集める
抽出、変換、クレンジングを行う
DMの切り出しにもETLを活用



マイニング
大量データから未来を予測
推測統計、パターン発見、検定などの数理を活用

ソースシステム
基幹系、業務系

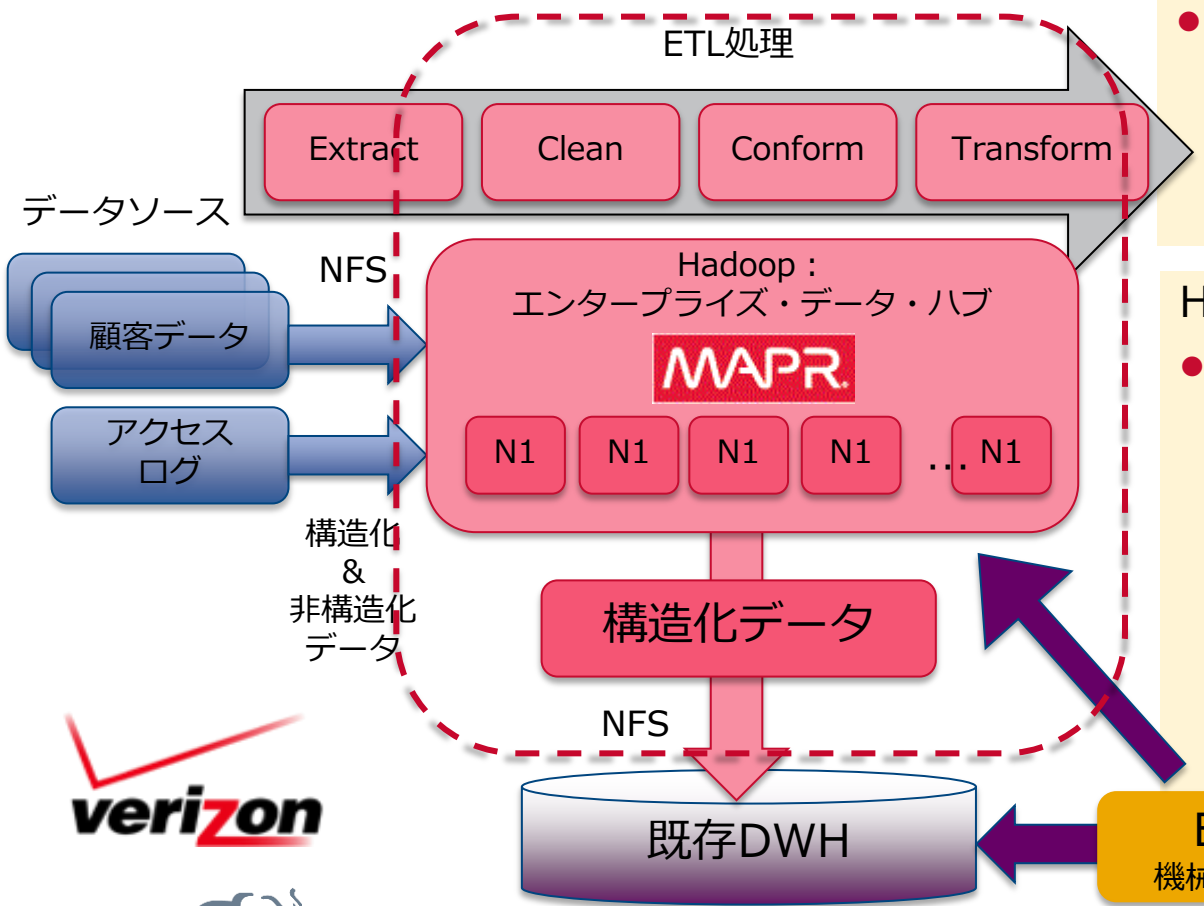
ソースシステム
基幹系、多数の業務系システム
様々な見るべき情報を保有
配置やデータ品質はバラバラ
（従来はこれを一つづつ見ていた）

Hadoop
大量データを溜め、処理、分析が
1カ所でできる新技術

Verizonのコスト削減方法

1. ベライゾンのユーザが日々どういう使い方をしているか、ユーザプロファイルや営業/請求についての詳細を知る
2. サービス品質とサービス向上への修理や増設の元データが欲しい

1. 解約率を下げる
2. DWH関連コスト削減



お客様のDWHの課題：

- 増え続けるデータに既存DWHでは対応できない
  - コスト
  - パフォーマンス
  - 非構造化データ（ログ等）への対応

Hadoop導入のメリット：

- DWHに入れるべきデータを選別でき、データ量とコストのバランスを取れる
  - DWH増設より圧倒的に低コスト（1/10に削減）
  - どこまでもスケールするパフォーマンス（3倍）
  - あらゆるデータを格納・処理
  - ソースから分析までの一環したデータフローを実現

verizon

Data Warehouse and Analytics

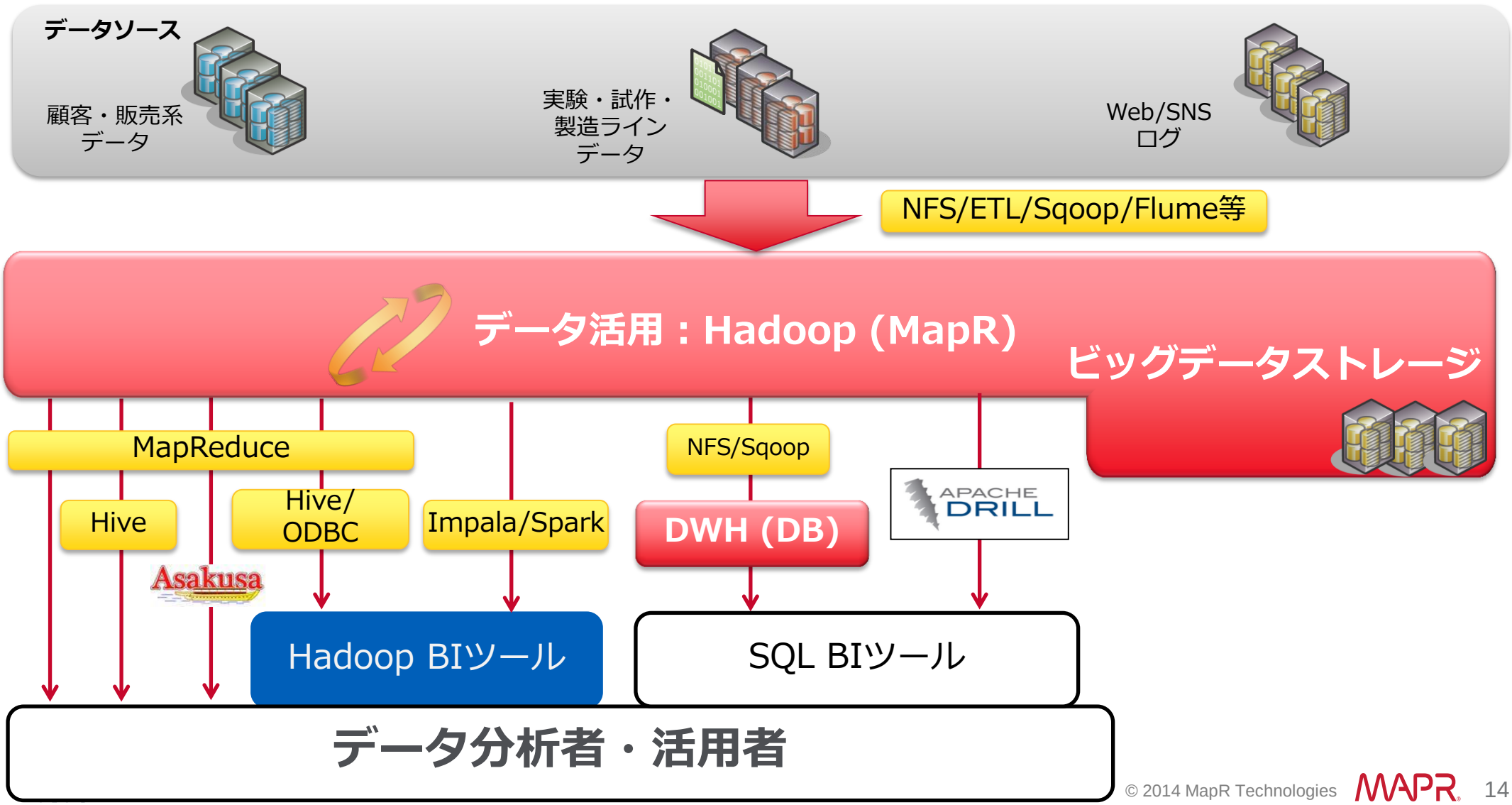
MAPR

分析・レポートのニーズと現状

1. データ量を増やしたい — DWHの弱点
2. データ種を増やし、様々なニーズを拾いたい — DWHの弱点
3. コストは掛けられない — DWHの弱点
4. 既存SQLをなるべくそのまま活かしたい
   - BI等の既存ツールを継続利用したい — Hadoopの弱点
   - MapReduceをJavaで書くのは敷居が高い
5. セルフサービス化をしたい — DWHの弱点
   - 情シス負荷は増やせない
6. リアルタイム性を上げたい — Hadoopの弱点

多様化するHadoop環境へのアクセス

データはHadoop基盤に集約
分析の結果はHadoop基盤、DWH基盤の両基盤で活用



分析システム最適化のロードマップ

DWH1.0

TERADATA

NETEZZA an IBM Company

ORACLE EXADATA

DWH2.0

TERADATA

NETEZZA an IBM Company

ORACLE EXADATA

hadoop

DWH2.5

VERTICA

hadoop

次世代 DWH

APACHE DRILL

hadoop

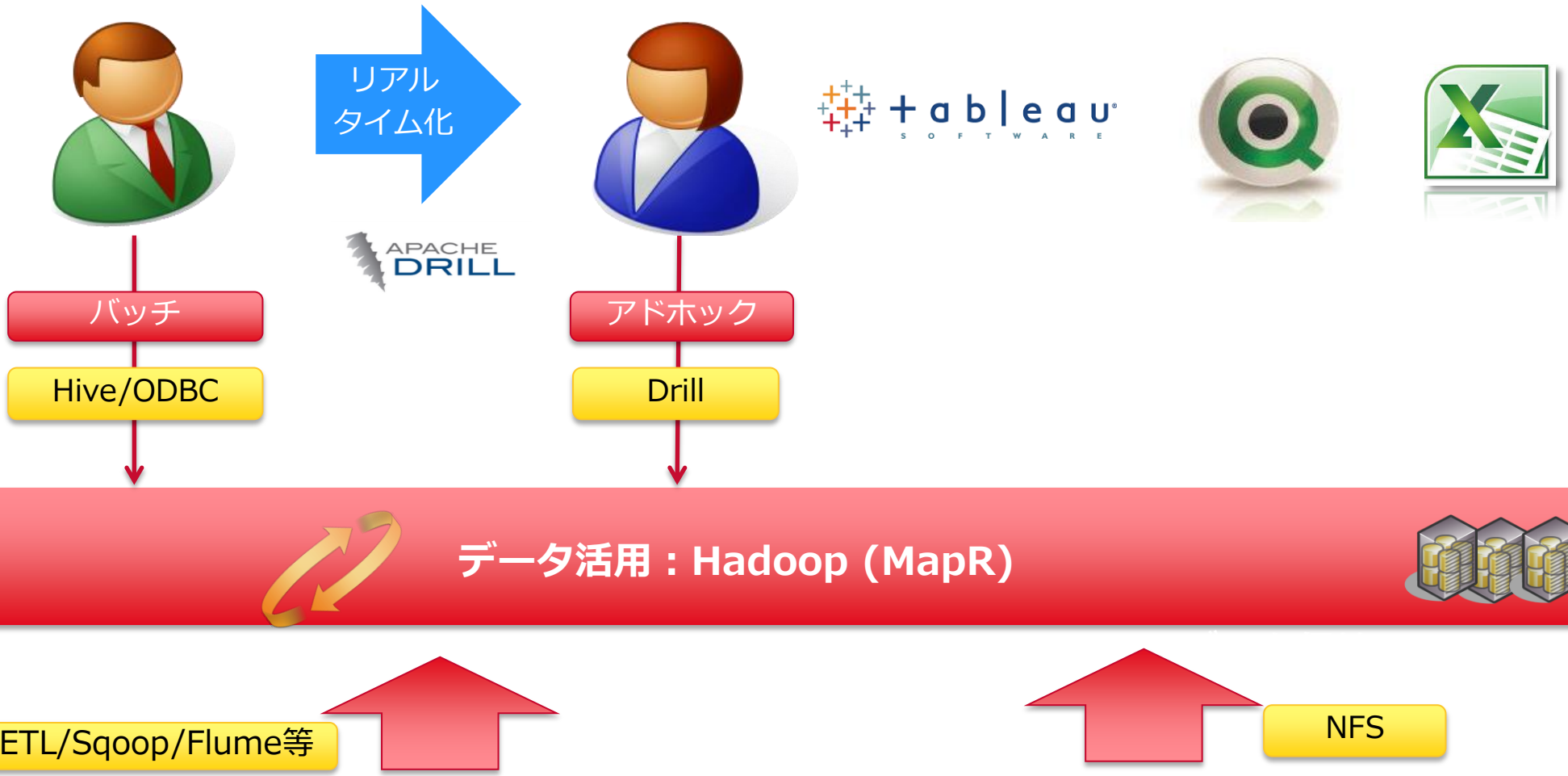
MAPR

Hadoop上のSQLソリューション

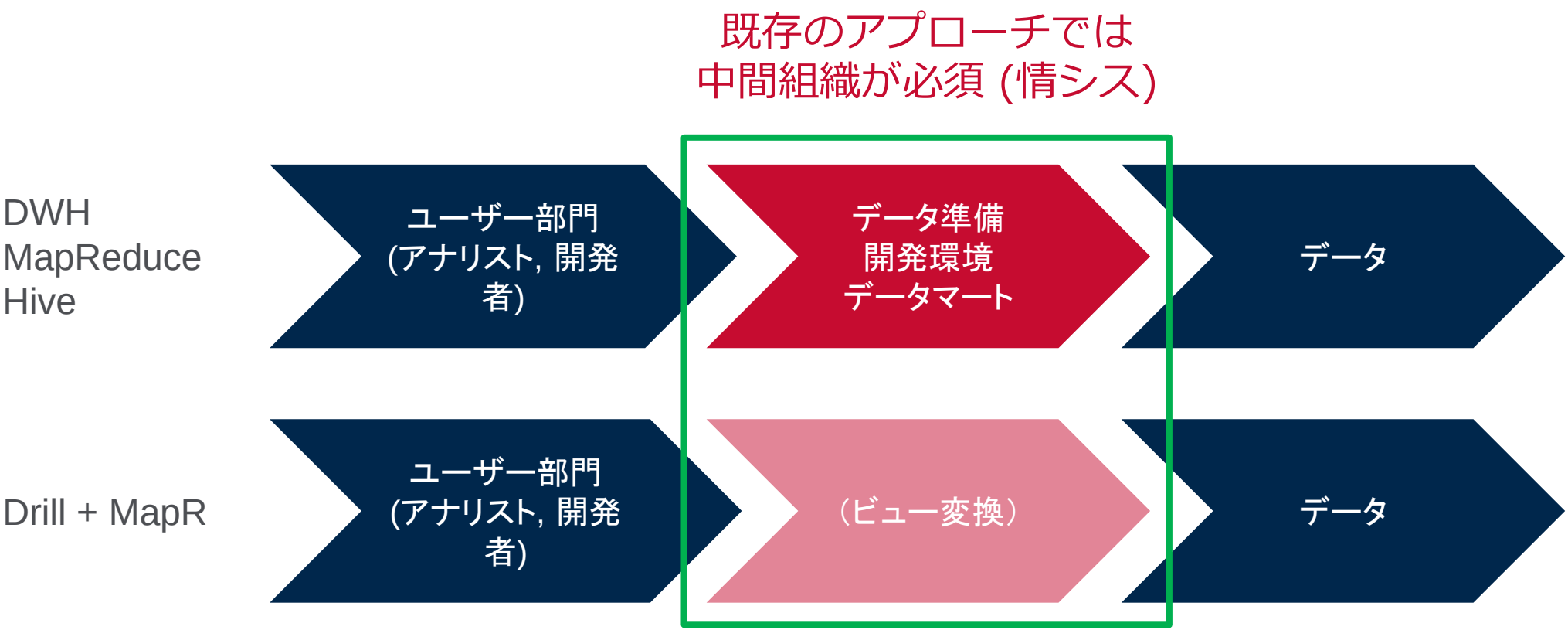
	Drill	Hive + Tez	Impala	Spark
レイテンシ	**Low**	Medium	**Low**	Medium
ファイル検索	Yes (all Hive file formats)	Yes (all Hive file formats)	Yes (Parquet, Sequence)	Yes (all Hive file formats)
HBase/MapR-DB 検索	**Yes**	**Yes**	Various issues	**Yes**
スキーマ	**Hive or schema-less**	Hive	Hive	Hive
クエリ言語	**ANSI SQL**	HiveQL	HiveQL (subset)	HiveQL
クライアント接続	ODBC/JDBC	ODBC/JDBC	ODBC/JDBC	ODBC/JDBC
大規模JOIN	Yes	Yes	No	No
階層データ	**Yes**	Limited	No	Limited
Hive UDF	Yes	Yes	Limited	Yes

APACHE DRILL

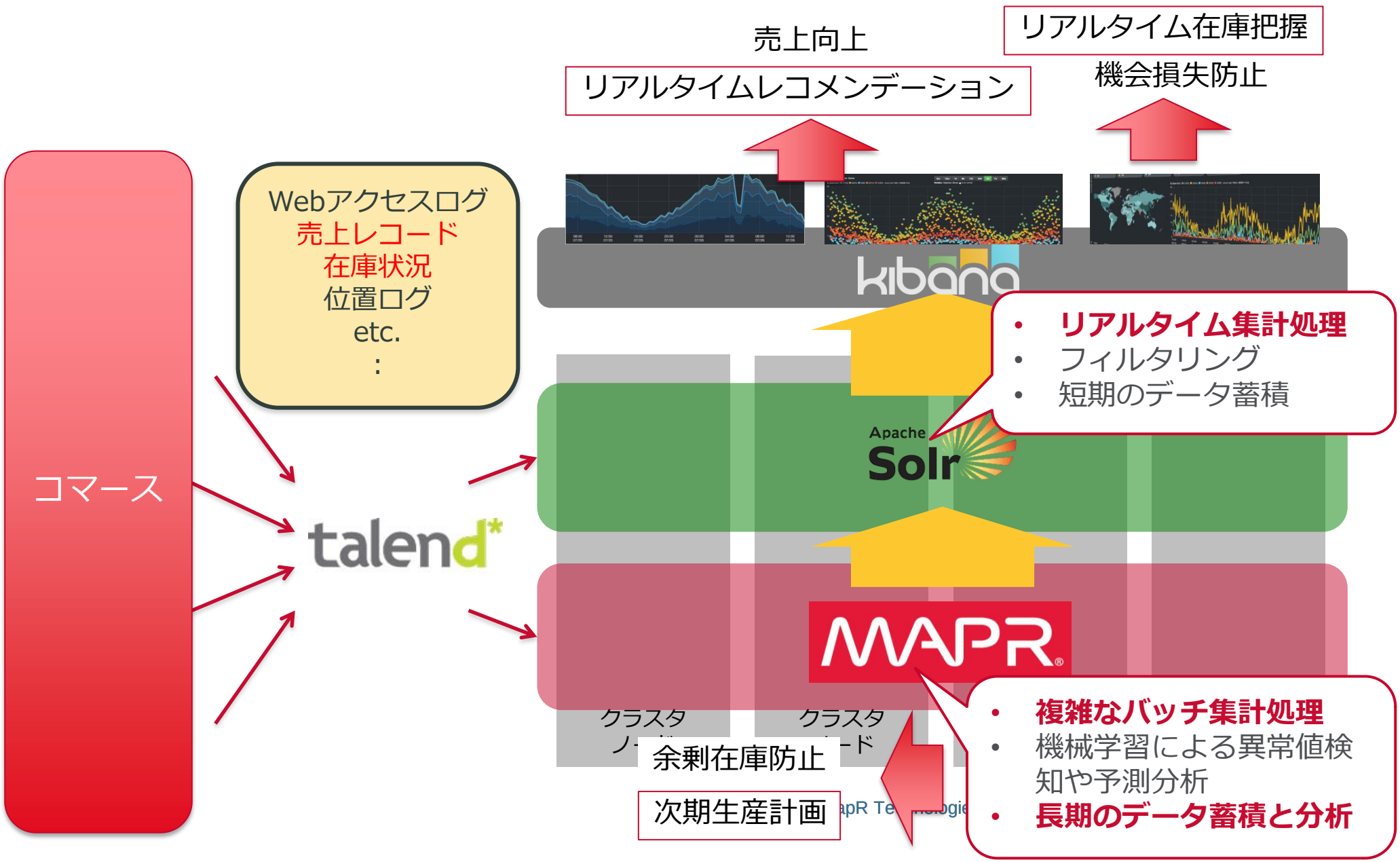
Hadoopのデータに既存BIやSQLも



俊敏性とセルフサービス化



リアルタイム分析環境例
（電通レイザーフィッシュの事例から）



Q:ビッグデータ活用していますか？

「もうやっている」（データウエアハウスとして）

ビジネスへの具体的な活用が明確で無い

ビジネスとデータ分析の両方の観点からの
戦略を考える人材が存在しない

投資効果を説明できない

MAPR

難しい（？）目的とKPI設定
～エンタープライズの悩み～

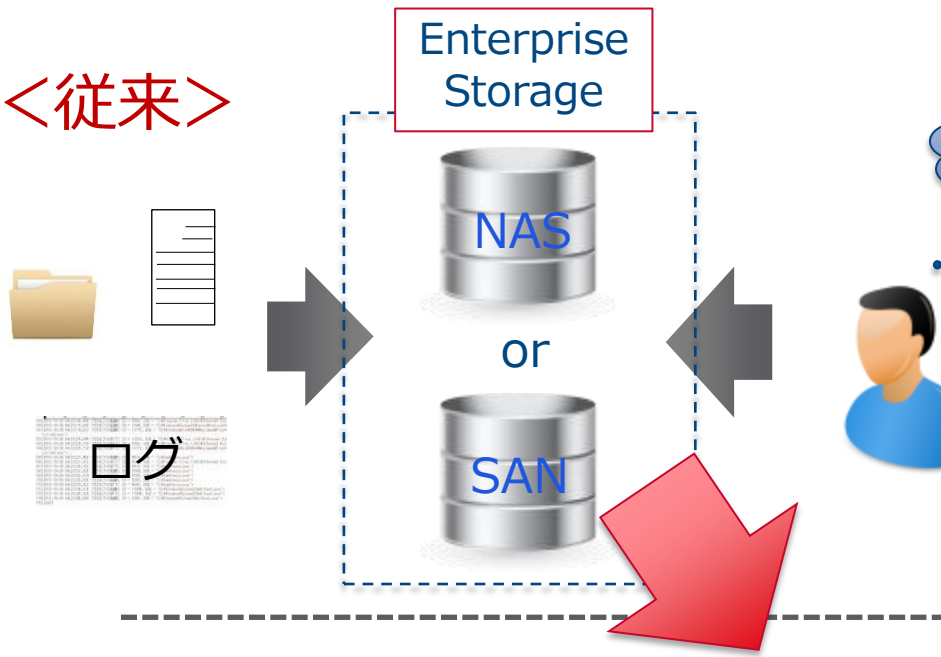
アプローチを変える

分析を意識したストレージにデータ蓄積をすることが重要
（後日、分析のためにTB/PBのデータ移動は高コスト）

売上増 or コスト/リスク削減 or 効率化？
（ただし、データは共通なハズ）

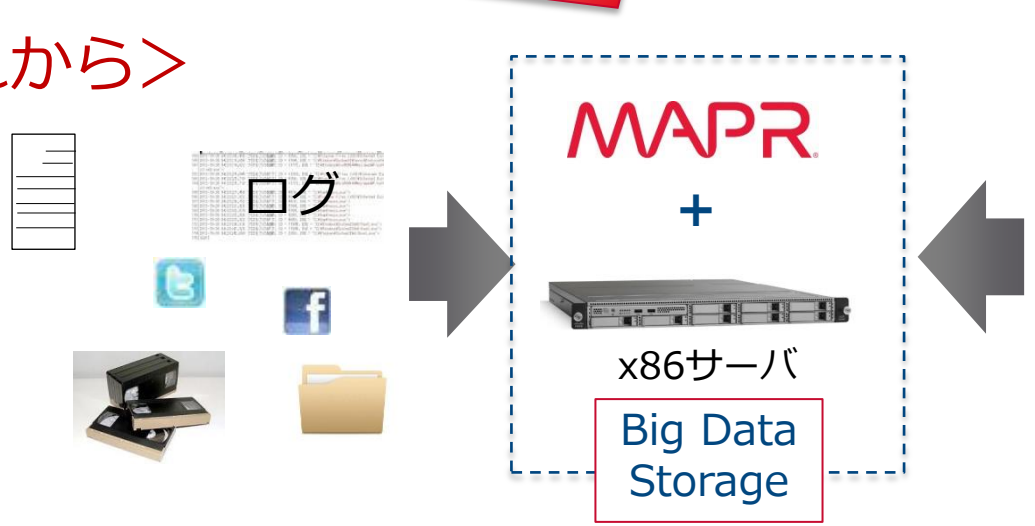
ビッグデータストレージ（analysis ready）としてのMapR

<従来>



- 膨大、かつ増え続けるログデータ等をどこに保管？
- スケールアウト型のNASは種類が少なく高価

<これから>



- MapR FS上に置くことで、コスト効率良く保管
- 分析や活用のために他にデータ移動することなく、いつでもHadoopを使ってデータ処理
- バックアップやスナップショットが取れるのはMapRだけ

More DATA beats better algorithms

**（より多くのデータは、
より良いアルゴリズムを凌ぐ）**

The Unreasonable Effectiveness of Data, published by Google

Zions Bank:

ビッグデータストレージと不正検知

データプラットフォームを統合することでコスト効果高く、セキュリティ分析と不正検知を行う

ZIONS BANK.
WE HAVEN'T FORGOTTEN WHO KEEPS US IN BUSINESS®

目的

不正を発見するチームとセキュリティ分析のチームが共同で利用するデータストアのプラットフォームを構築し、その上に統計モデリングを載せ、不正や不正につながる怪しい行動を発見する

チャレンジ

- 既存のインフラはスケールしない
- この数年レポート作成に時間がより掛かるようになっていた

MapR利用のメリット

- データストレージコストを50%削減
- 1.2PBのデータからのクエリが24時間から30分に削減
- 限界の無いスケーラビリティにより、より多くのデータを使え、より正確なモデルと洞察を得られた

Business Impact

“Zions Bankでは初めてセキュリティ分析のために全データを中央集権的に集め利用したが、不正検知にもそのデータが使えることが分かっただけではなく、不正検知に非常に役に立つことがわかった

Michael Fowkes - SVP Fraud Operations and Security Analytics

MAPR

MapR 製品

M3
COMMUNITY EDITION

- Hadoop (M5) & NoSQL (M7)
- 管理ツール
- NFS アクセス
- パフォーマンス
- ノード数の制限なし
- 無料

M5
ENTERPRISE EDITION

- 管理ツール
- NFS アクセス
- パフォーマンス
- HA
- スナップショット
- ミラーリング
- 24 X 7 サポート
- サブスクリプション

M7
ENTERPRISE DATABASE EDITION （MapR-DB）

- Hbase互換のNoSQL DB
- M5の機能＋
- HBaseの運用を簡素化
- HBaseの高速化
- 安定したレスポンスタイム、低レイテンシー
- ファイル/テーブルの統合スナップショット

Also Available through:

amazon web services™

Google Compute Engine

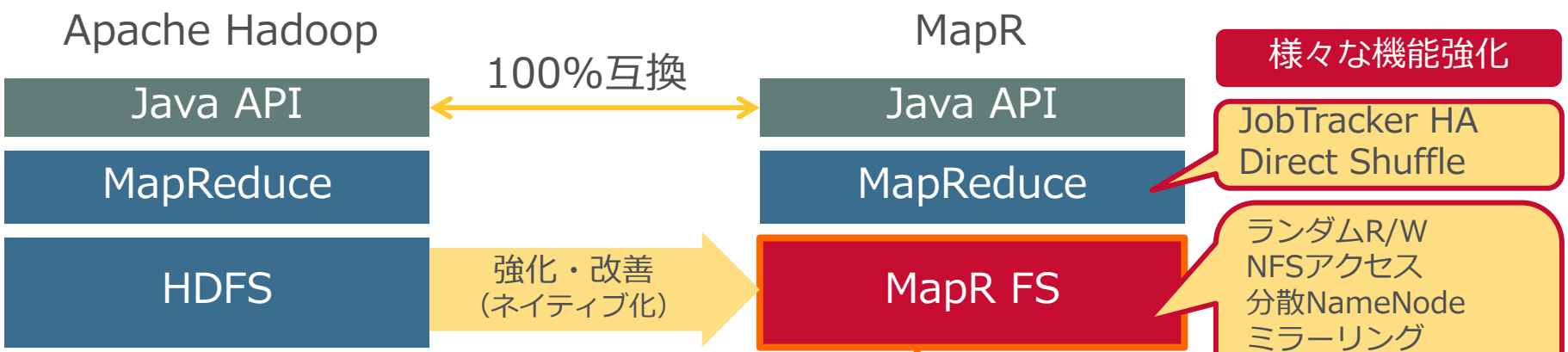
MAPR®

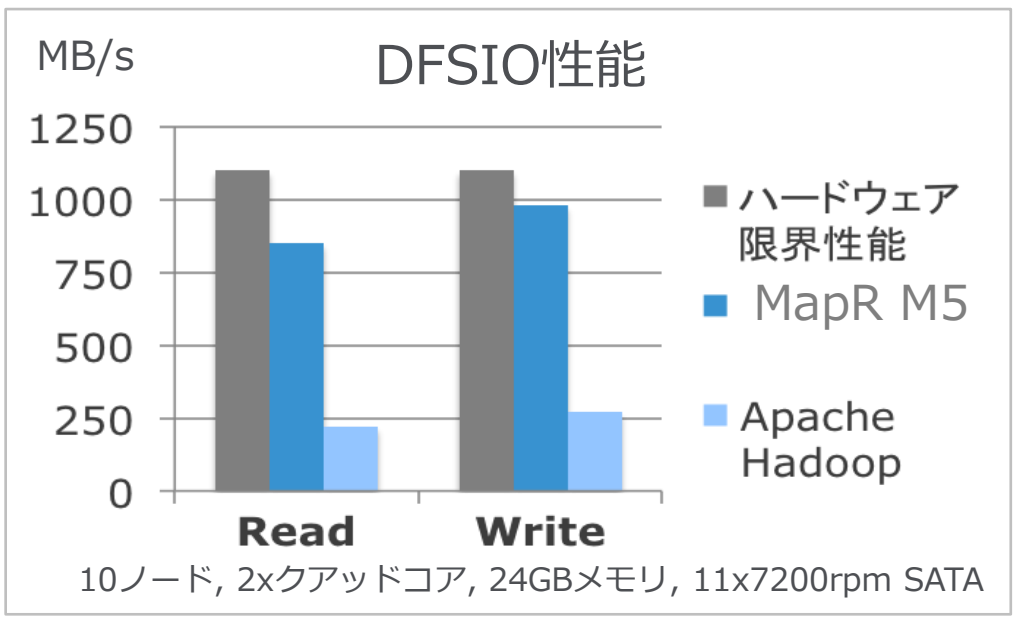
MapRによるOSS Hadoopの強化とメリット

Apache Hadoopをエンタープライズで利用する際に問題となるポイントを1つ1つ解消

効果 / メリット	強化ポイント
パフォーマンス	● ネイティブファイルシステム ● ダイレクトシャッフルによるシャッフルの最適化 ● 分散Name Node(CLDB)によるボトルネック解消
運用性	● POSIX対応NFS I/FとランダムRead/write可能ファイルシステム ● マルチテナント ● リニアなスケーラビリティ ● ノード数減
信頼性	● 単一障害点の除去(NameNode, Job Tracker) ● スナップショットによるデータ保護 ● ミラーリングによる簡単バックアップとDR

エンタープライズ用Hadoopのために！
アーキテクチャ再設計と再実装による強化

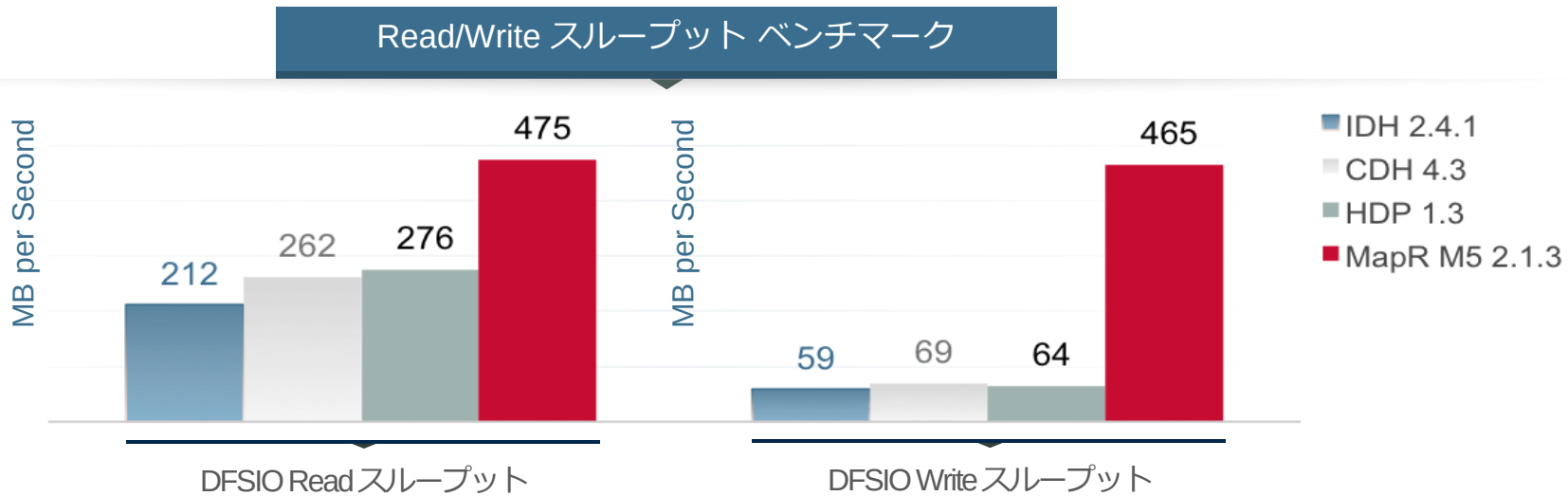




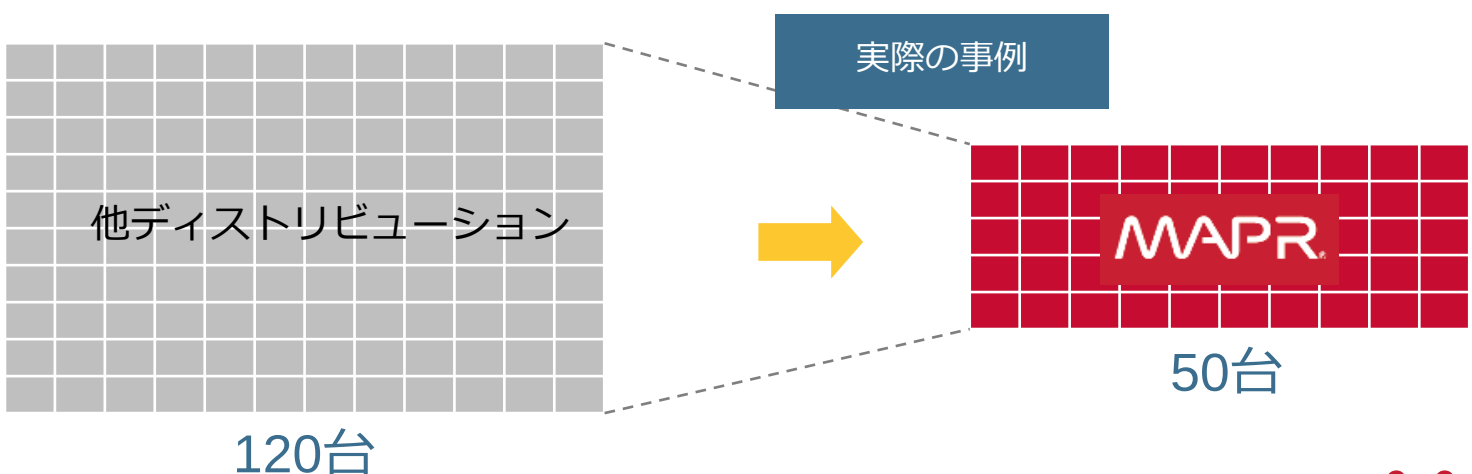
ハードウェアの限界性能を引き出すための
アーキテクチャ設計・再実装

- ロックの排除による並列処理の最適化
- ビルトイン圧縮によるI/O削減
- 分散NameNode
- RPC経由のShuffle転送
- Java GCの影響の排除

高パフォーマンス = TCO 改善



ソース: Flux7 Labs Study, 2013年10月



バイアスのないオープンソース (Linux を例に)

- オープンソースディストリビューションは選択肢の提供が鍵
  - Linux は MySQL, PostgreSQL, SQLite の**すべて**を含む
  - Linux は Apache httpd, nginx, Lighttpd の**すべて**を含む
- MapR はバイアスのない選択肢を提供する唯一の Hadoop

	MapR Distribution for Hadoop	**ディストリビューションA**	**ディストリビューションB**
Spark	Spark および Shark （全スタックをサポート）	Spark のみ	なし
インタラクティブ SQL	Shark, Impala, Drill, Hive/Tez	単一の選択肢 (Impala)	単一の選択肢 (Hive/Tez)
バージョン	Hive 0.11, 0.12, 0.13, 0.14 Pig 0.11, 012 HBase 0.94, 0.98	単一のバージョン	単一のバージョン

まとめ

- Hadoopをよりフレンドリーにするツールが存在
  - Asakusa：基幹系のも活用が可能
- Hadoop自身もエンタープライズでの利用を考慮したものに進化
  - MapR
- さらにユーザはリアルタイム性とセルフサービス化を求める
  - Drill：SQL on Hadoopを使って既存BIやSQLを活用

MAPR

ご静聴ありがとうございました

MAPR