

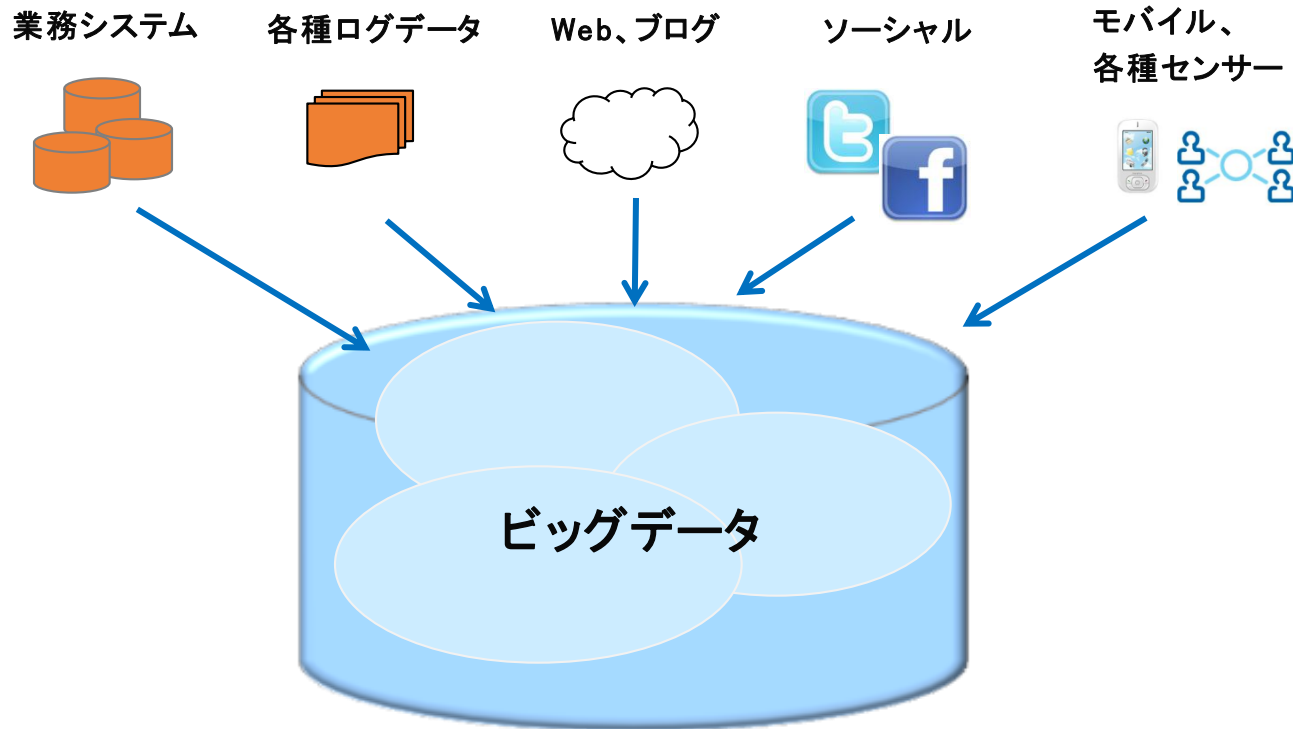
BI × ビッグデータ

PentahoのHadoopに対する取組みのご紹介

株式会社KSKソリューションズ

ビッグデータ？

- TバイトやPバイトクラスのデータ、日次で発生



ユースケース

- トランザクション
 - 不正検知
 - 金融サービス、株取引(アルゴリズム取引)

- サブトランザクション
 - Webログ
 - ソーシャル／オンラインメディア
 - 通信イベント

ユースケース

- ノントランザクション
 - Webページやブログ
 - ドキュメント
 - アプリケーションイベント
 - マシーンイベント(M2M)

データレイク(データの池)

- シングルソース
- 大きいボリューム
- 蒸留されていない



データレイクの4つの要件

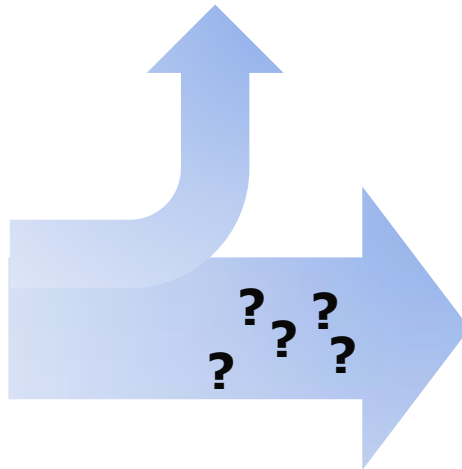
- すべてのデータをストアする
- 日常の定型レポートや分析を行う
- アドホック(非定型)のレポートや分析を行う
- パフォーマンスとコストのバランス

いままでのBI

データマート



データソース



テープor削除



ビッグデータ・アーキテクチャ



データマート



アドホック



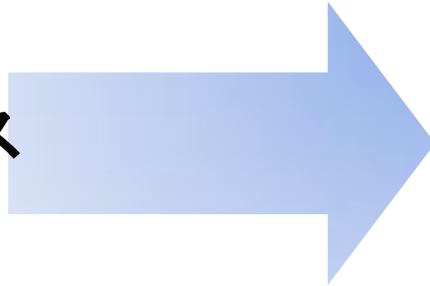
データウェアハウス



データレイク



データソース



ビッグデータはデータマートに代わるものではない

- 大きいレイテンシー
- 大量データのバッチ処理に最適化されている
- データベースとして成熟していない
- no-SQLデータベースである

ビッグデータ・アーキテクチャー

データマート



アドホック



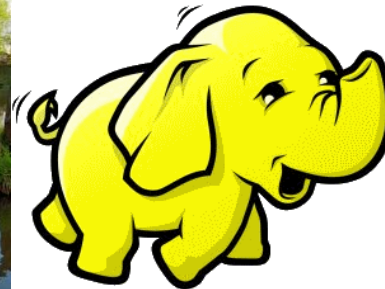
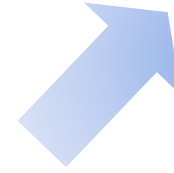
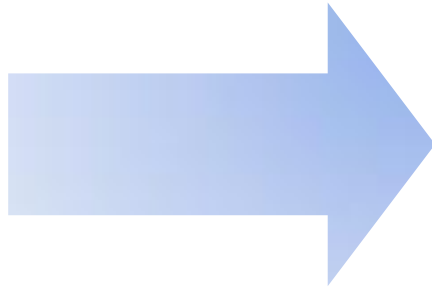
データウェアハウス



データレイク



データソース



Hadoopとは？

データ分散処理のためのJavaソフトウェアフレームワーク

- Apacheプロジェクト
- Yahoo!, Googleのアイデアによって作成
- 分散ファイルシステム+MapReduceエンジン
- コモディティのハードウェアを使用
- スケールアウト

Hadoopの特徴とBI

- 分散処理システム(バッチ処理の高速化)
- 分散ファイルシステム(耐障害性)
- コモディティのハードウェアを使用(低コスト)
- プラットフォームの独立性
- RDBの限界を超えて、スケールアウトが可能

多くのケースにおいて、それは唯一のソリューション

Googleのユースケース

- インターネットをインデキシングする必要性
- 大量の非構造化データ
- 事前に定義されたインプット
- 事前に定義されたアウトプット
- 事前に定義された質問
- シングル・ユーザーコミュニティ
- 並列実行とストアの必要性








彼らの出した答えは、MapReduce

Yahoo!のユースケース

- インターネットをインデキシングする必要性
- 大量の非構造化データ
- 事前に定義されたインプット
- 事前に定義されたアウトプット
- 事前に定義された質問
- シングル・ユーザーコミュニティ
- 並列実行とストアの必要性

彼らの出した答えは、MapReduce

BIユーザーは？

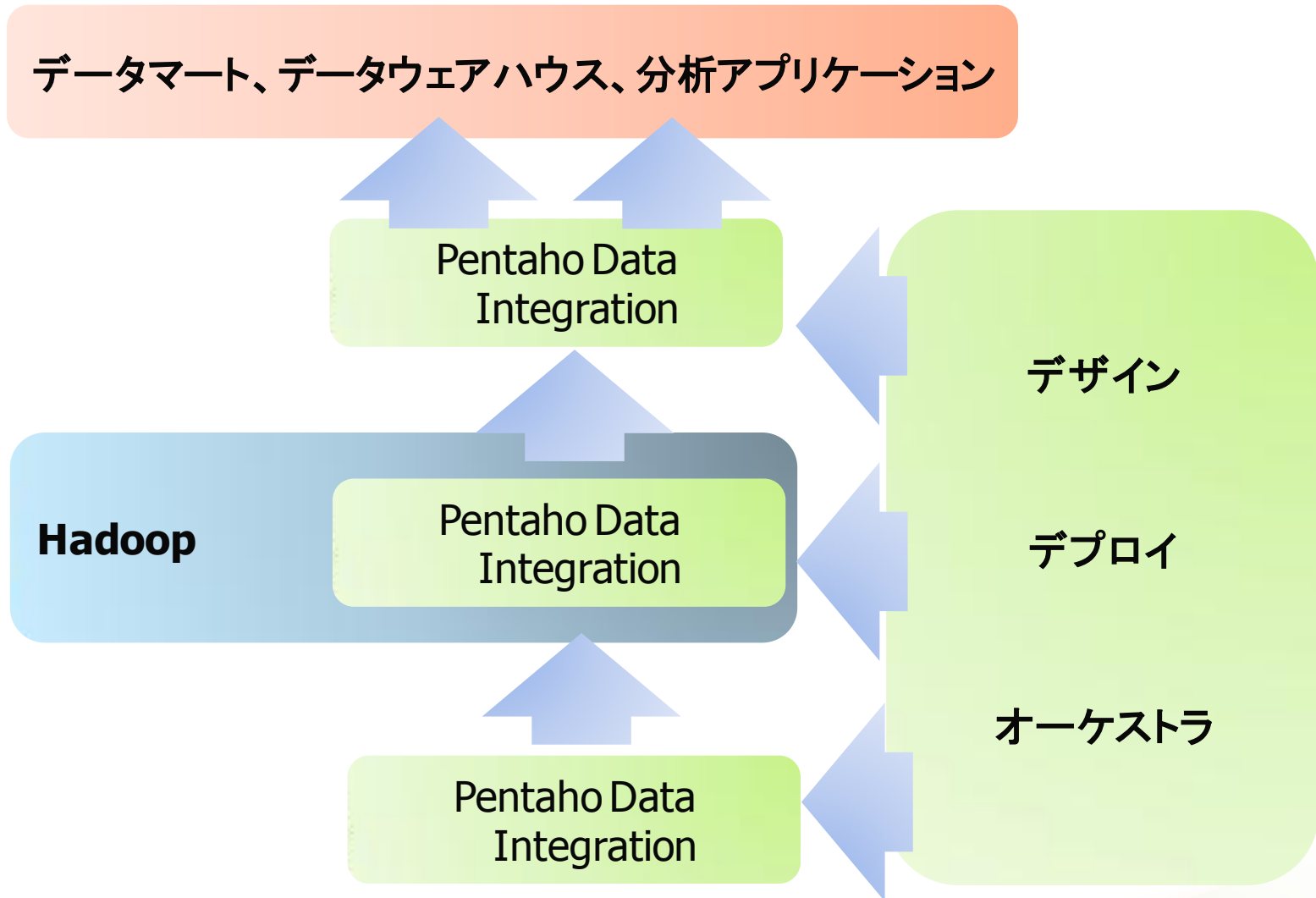
-  ● インターネットをインデキシングしない
-  ● 大量の構造化されたデータ
-  ● 異なるインプットソースとフォーマット
-  ● 異なるアウトプット
-  ● 異なる質問
-  ● 複数のユーザーコミュニティ
-  ● 並列実行とストアの必要性

HadoopをBIをためには使わない

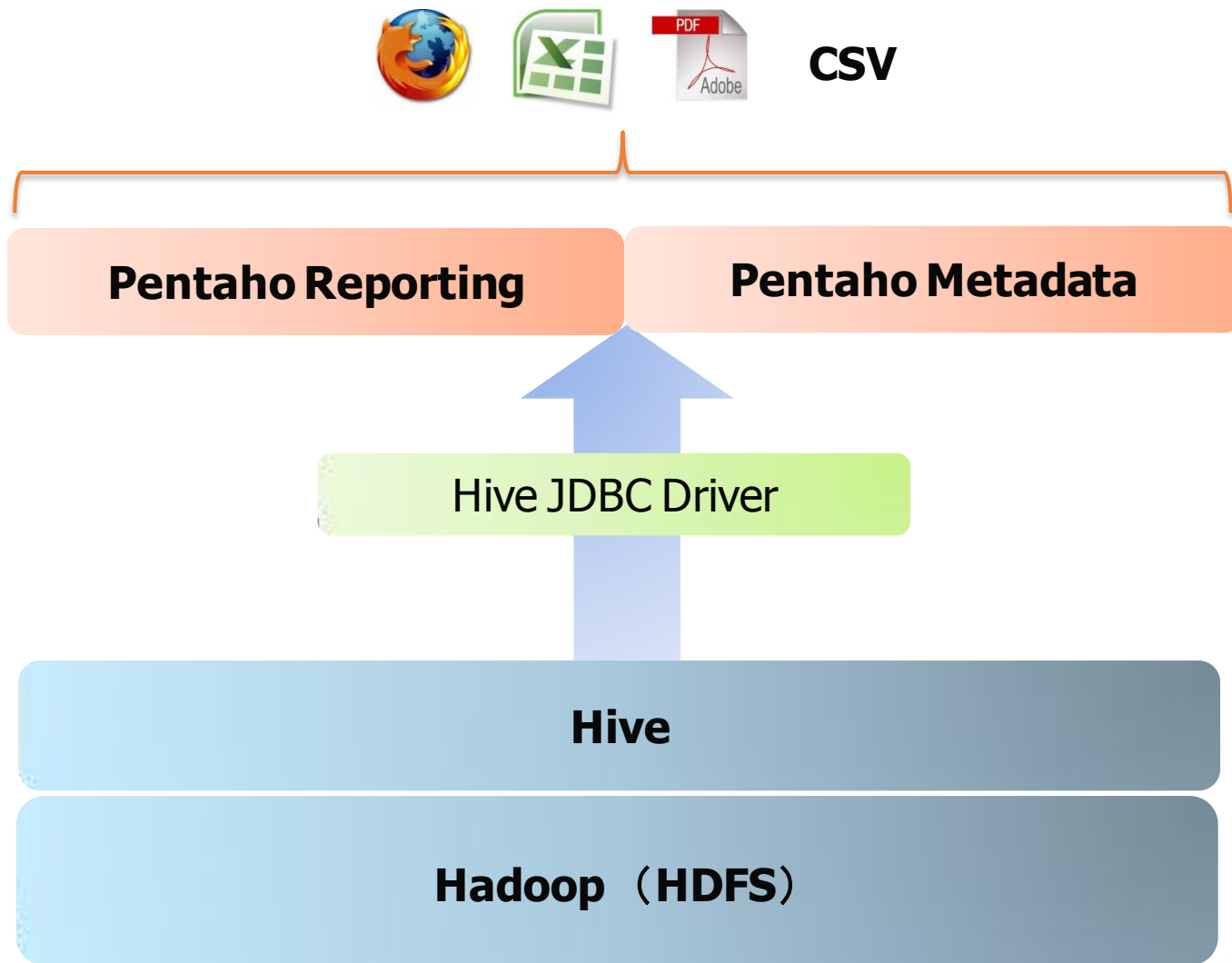
なぜなら、それはBI用途には向いていないから

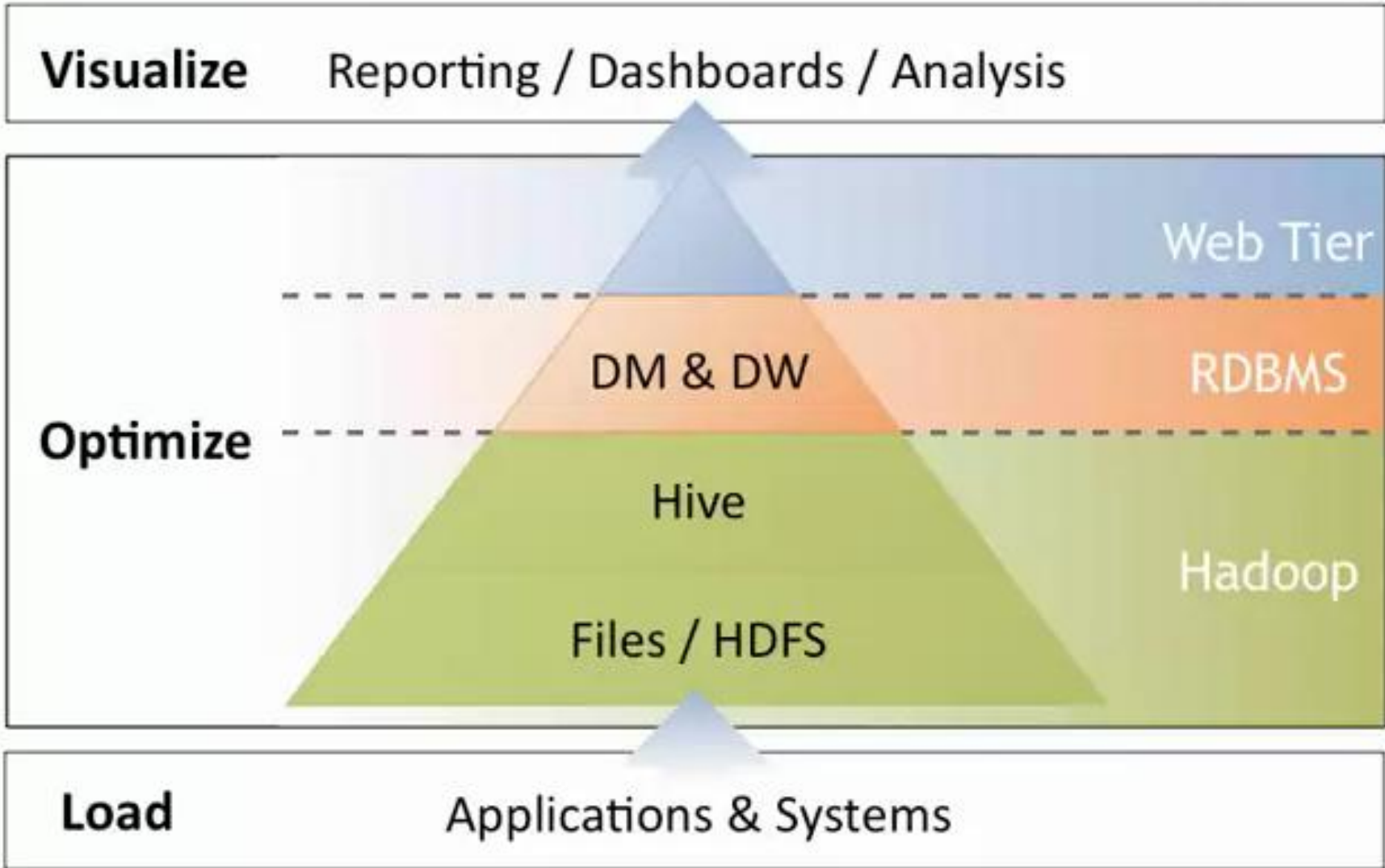
でも、Hadoopに足りないところを補う方法があったら？

Pentaho Data Integration (Pentahoデータ統合)



Pentaho Reporting (Pentahoレポートイング)





ビッグデータ・アーキテクチャー

データマート



アドホック



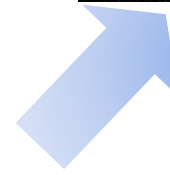
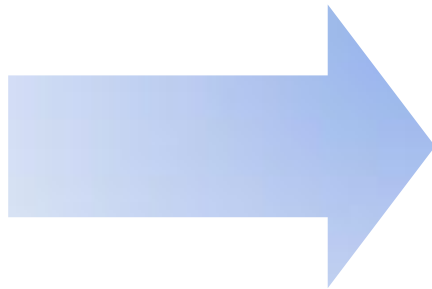
データウェアハウス



データレイク



データソース

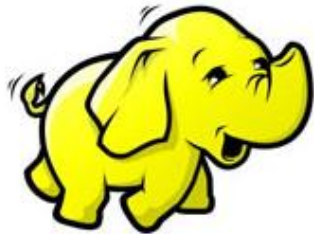


Reporting / Dashboards / Analysis



Applications & Systems

Pentaho BI Suite 4.1



Hadoop



NoSQL



分析データベース

- ビジネスユーザー

- Hadoopデータソースに対するインタラクティブな分析
- Hiveを経由したHadoopデータに対するレポートのスケジューリング

- テクニカルユーザー

- MapReduceジョブ、Pigスクリプト、Hive、Hbaseクエリー作成のためのグラフィカルデザイナー
- Hadoopジョブやその他のジョブのスケジューリングやモニタリング
- Hadoopデータと他のデータソースの統合

- NoSQLデータとリレーショナルデータを容易に統合

- ネイティブで高パフォーマンスなアクセス

- サポートしているNoSQL

- Apache Cassandra
- DataStax
- Hbase
- Hive
- MongoDB
- HPC systems
- Elasticsearch
- XML Streaming

- ネイティブなSQLをサポート

- ネイティブなバルクローダーをサポート

- サポートしている分析データベース

- Infobright
- EMC Greenplum
- HP NonStop SQL/MX
- HP Vertica
- IBM Netezza
- Actian Vectorwise
- LucidDB
- MonetDB
- Teradata

まとめ

ビッグデータをデータウェアハウスとして使える？

たぶん可能

ビッグデータをデータウェアハウスとして使うべき？

あまりお勧めしません

データウェアハウス／データマートとは

データマート

- クエリーやレポートのための構造化されたデータ

データウェアハウス

- すべてのシステムに対するデータマートを作成したら、それを統合しないとならない、データマートが整理・統合されたもの

データウェアハウスの条件

- 複数のデータソース
- クレンジング処理されている
- 整理されている
- 集計されている



データウェアハウスに対する、ビッグデータ

- 大きいレイテンシー
- 大量データのバッチ処理に最適化されている
- データベースとして成熟していない
- no-SQLデータベースである

でも、データウェアハウスとして、使いたくて使いたくて、たまらない場合は？

データ・ウォーターガーデン

- 複数の大きなプールと池
- 整理されている
- クレンジングされている
- 関連性がある

