大規模データも手軽に分析！
Big Data Module 紹介

数理システムユーザーコンファレンス 2014
（株）NTTデータ数理システム データマイニング部
五十嵐 健太

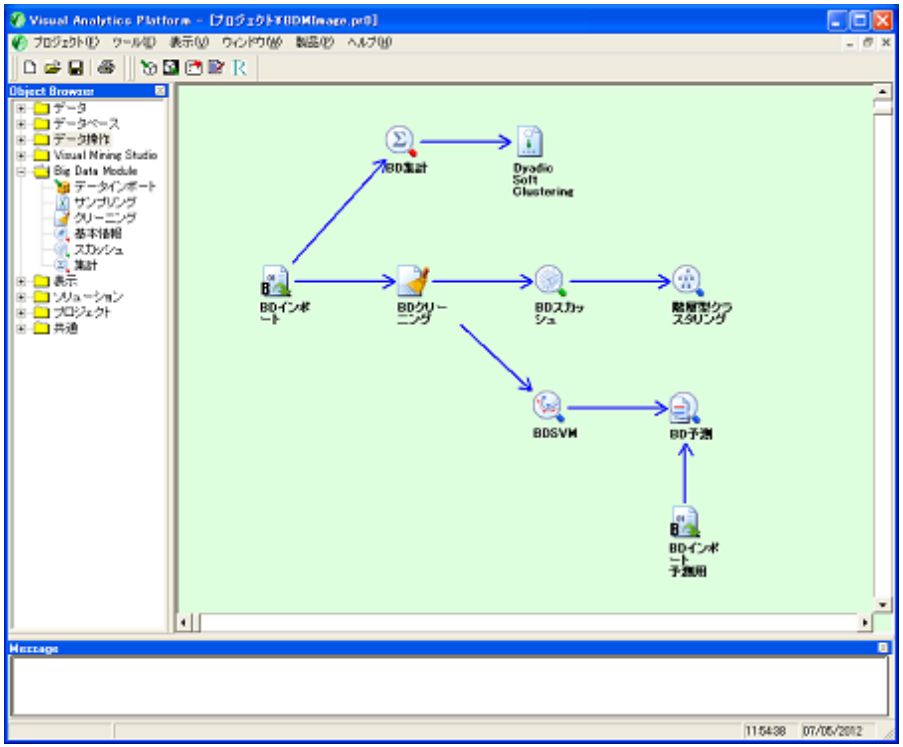
Global IT Innovator NTT DATA Group

NTT DATA MATHEMATICAL SYSTEMS INCARNATION

Big Data Module とは

- マウス操作で簡単に
  大規模データのデータマイニングを実現

- 大規模データのための
  高速分析アルゴリズムを搭載
  - オンラインアルゴリズム
  - 並列処理

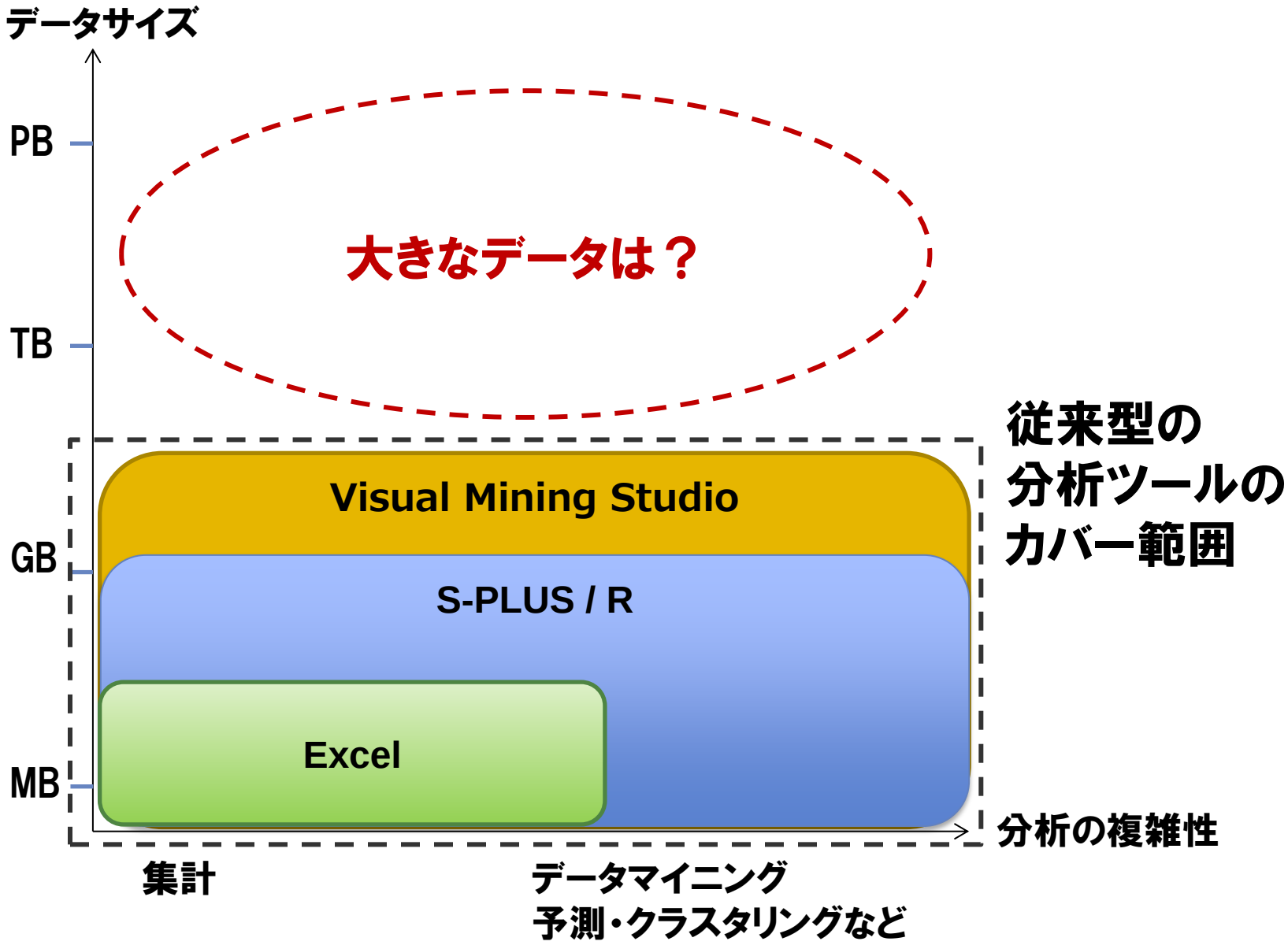


- 特殊な分析専用マシンは不要
  市販のマシンを1台用意すれば
  それだけで分析が実行可能

- Visual Mining Studio を
  はじめとした、数理システム製品と連係

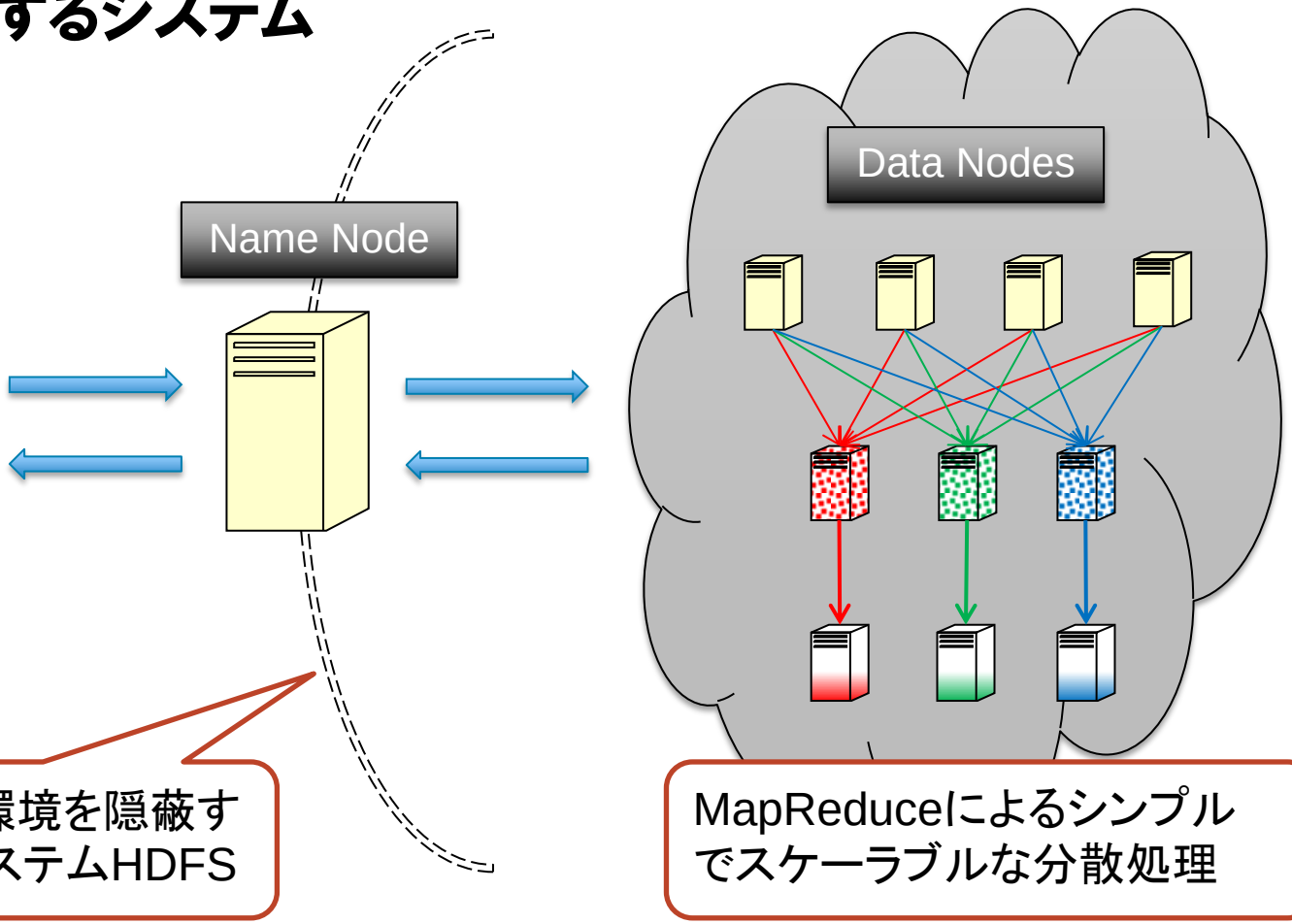
ー 対象データサイズ ー

分析ツールとデータサイズ



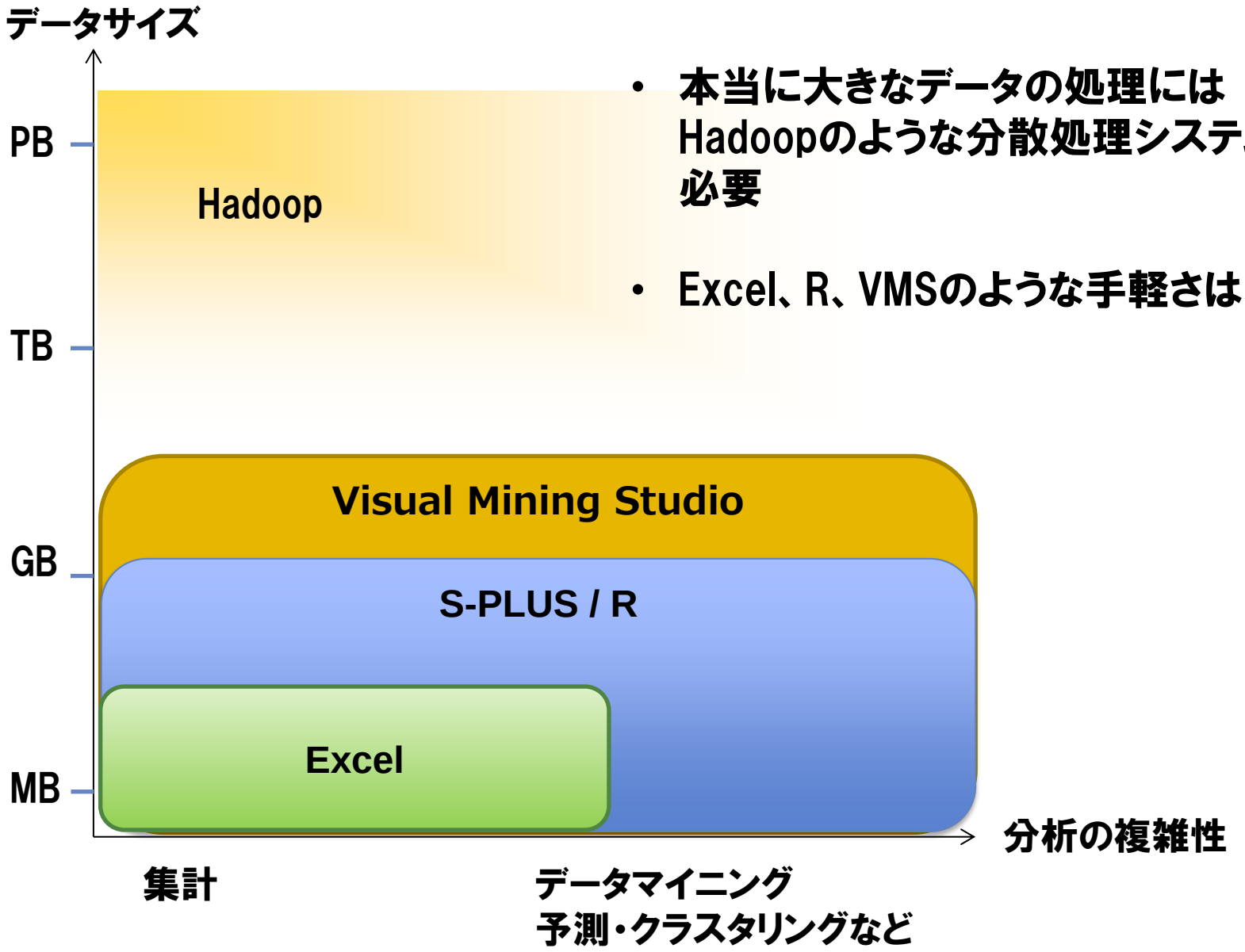
Hadoop とは

従来は面倒だった複数マシンでの分散並列処理を簡単に実現するシステム



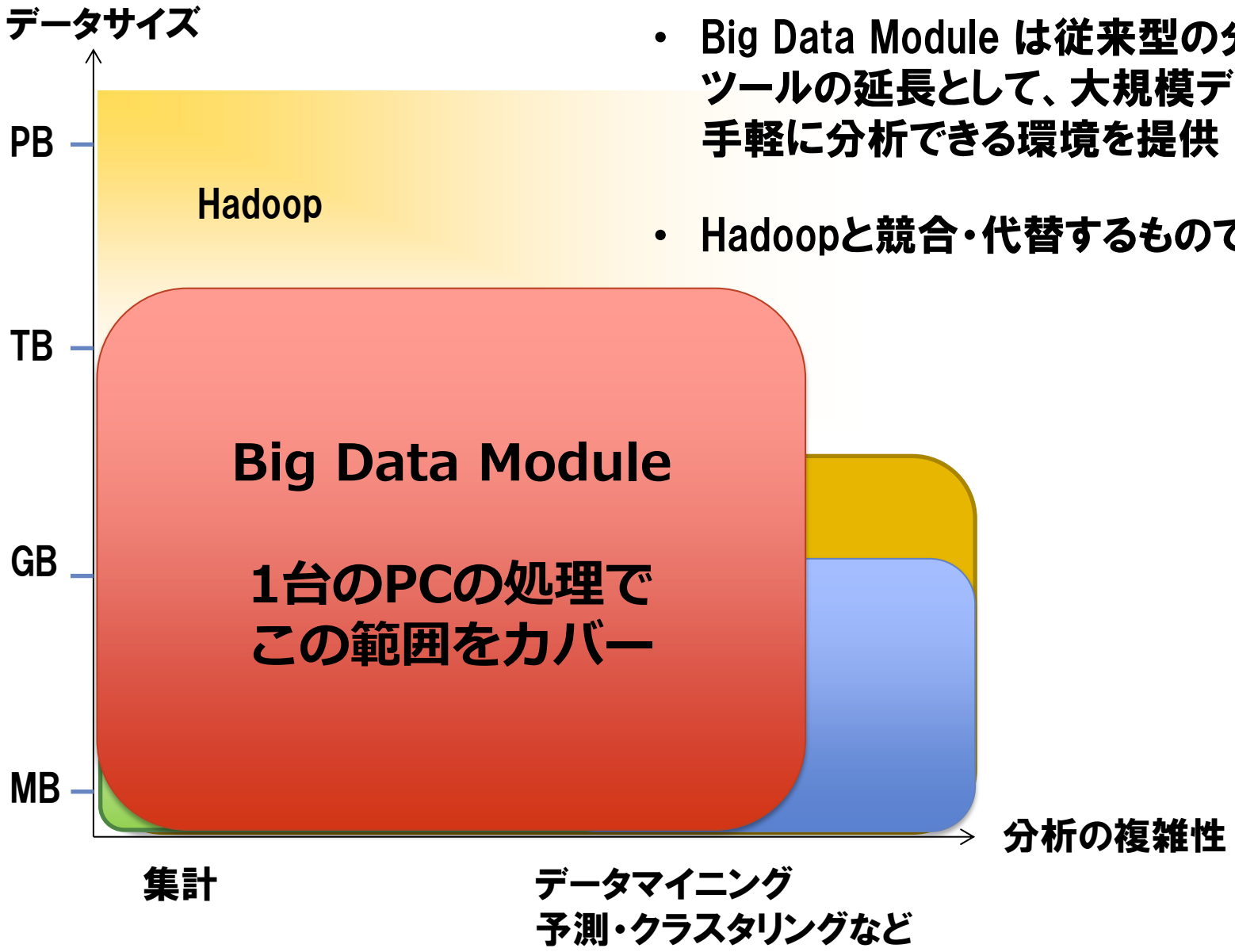
多数のマシンで処理を分散することで、テラバイト、ペタバイトスケールのデータの処理が可能に

Hadoop の位置づけ



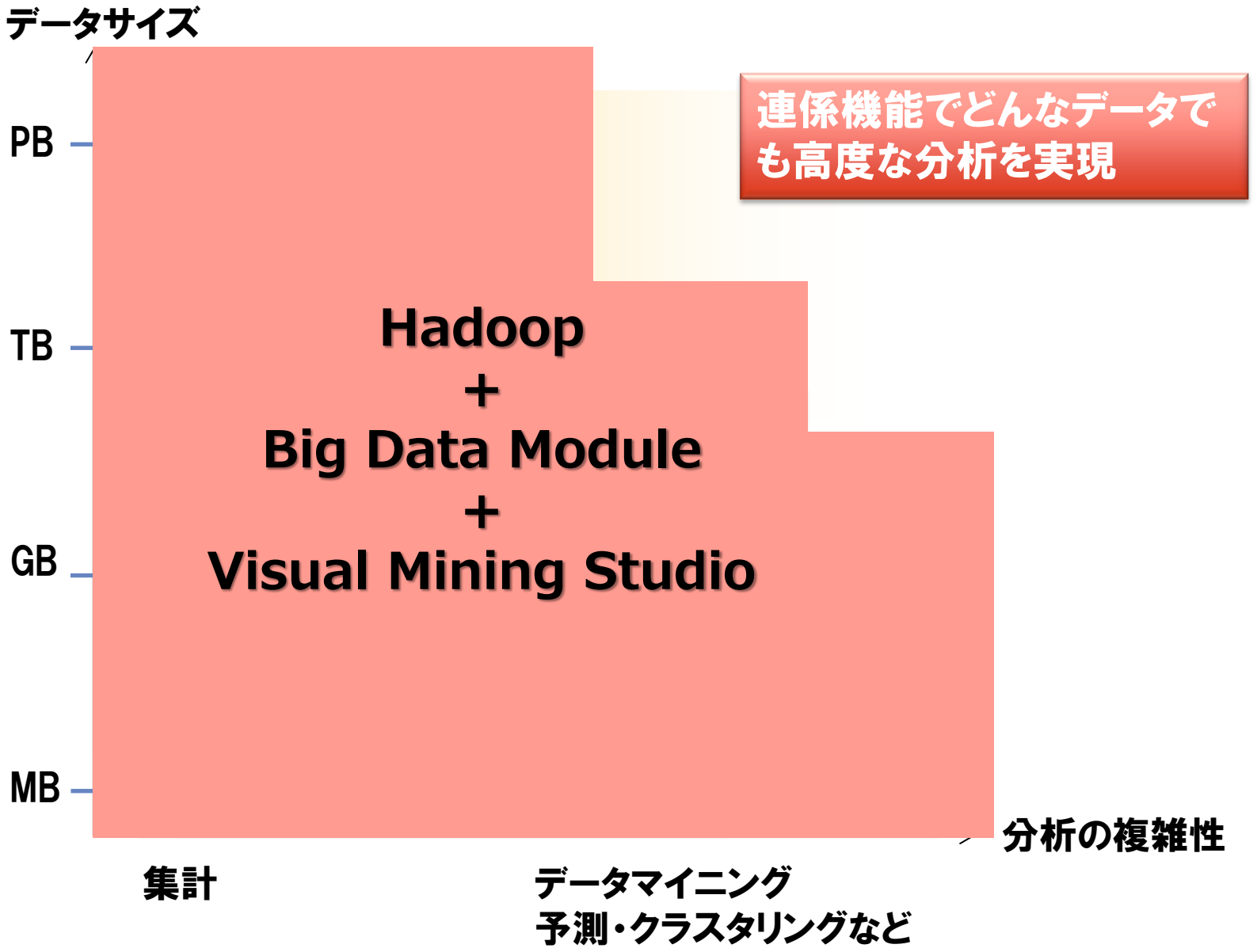
- 本当に大きなデータの処理には Hadoopのような分散処理システムが必要
- Excel、R、VMSのような手軽さはない

Big Data Module の位置づけ



- Big Data Module は従来型の分析ツールの延長として、大規模データを手軽に分析できる環境を提供
- Hadoopと競合・代替するものではない

Hadoop 連係、Visual Mining Studio 連係



ー 使用イメージ ー

数値予測

売上予測

株価予測

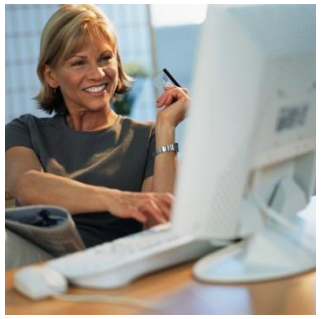
電力需要予測

予測精度を高めるには…
- データ数を増やす
- 説明変数を増やす

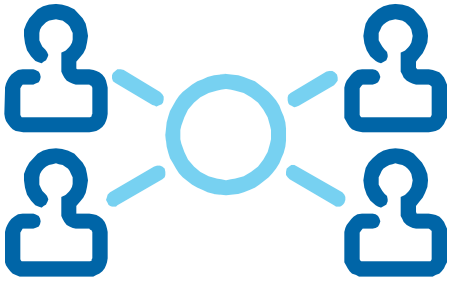
計算時間の
爆発的な増加で
計算不可能

Big Data Moduleなら…
オンライン線形回帰でビッグデータでも予測可能
- データ数の線形オーダーの計算時間→**超高速**
- データ数に依存しないメモリ使用量→**超省メモリ**

レコメンデーション



- ECサイト
ユーザー×アイテムの
マッチング



- SNS
ユーザー×ユーザー
のマッチング

- セールス
営業マン×営業先
のマッチング

膨大な組み合わせの中から
ベストなパターンを発見

- 時間をかけて終わるかどうか
わからない計算？
- 一部の情報だけを使って分析？

Big Data Moduleなら…
オンライン行列分解で協調フィルタリングによる
レコメンデーション
高速かつ**高精度**なレコメンデーション

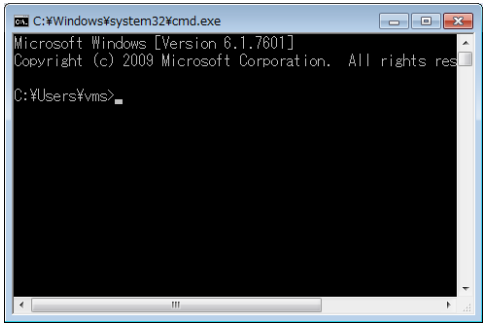
システムへの組み込みも

バッチ処理

- 今の分析ツールだと、処理時間がかかりすぎる
- (WEBアクセス、システム)ログデータの分析がしたい

→GUIツールとして
Big Data Module を使用

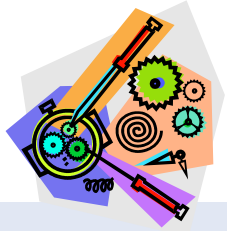
→定型化した分析はバッチ化して
コマンドプロンプトから実行



システム組み込み

- リアルタイムで処理したいのに、分析処理が追いつかない

→Big Data Module のアルゴリズム部分を取り出して
システムへ組み込み



ー 今後の開発について ー

今年度の開発予定

- データベースへの接続機能の追加
  - テーブル選択
  - SQL 発行

- データインポート機能の高速化

- 線形回帰での自動変数選択
  正則化方法として
  - Lasso
  - elastic net（Lasso と Ridge のハイブリッド）

  を選択可能に

来年度以降の開発課題

- データ加工アイコンの追加
  スクリプトの記述で一通りのデータ操作はできるが…
  アイコン化してGUIで簡単に使えるようにしたい

- 分析機能の追加
  - 最近流行のアルゴリズムを検討
    Deep Learning？
    non-parametric bayes？

- 実務的に使いやすいツールに
  - 外部ツール(python)からの実行
  - VAP関連製品との連係の強化

テスト利用制度もございます
お気軽にご相談ください

＜お問合わせ先＞
(株) NTTデータ数理システム Big Data Module 担当
bigdata-info@msi.co.jp
http://www.msi.co.jp
Tel :03-3358-6681 ［ 営業部直通 ］
Fax :03-3358-1727

変える力を、ともに生み出す。
NTT DATA グループ