Oracle Direct Seminar

ORACLE

Oracle DBでできる！簡単データマイニング

日本オラクル株式会社

Oracle Direct

Oracle Direct Seminar

OTN×ダイセミ　でスキルアップ!!

・一般的な技術問題解決方法などを知りたい！
・セミナ資料など技術コンテンツがほしい！

Oracle Technology Network(OTN)を御活用下さい。

http://otn.oracle.co.jp/forum/index.jspa?categoryID=2

一般的技術問題解決にはOTN掲示版の
「データベース一般」をご活用ください

※OTN掲示版は、基本的にOracleユーザー有志からの回答となるため100%回答があるとは限りません。
ただ、過去の履歴を見ると、質問の大多数に関してなんらかの回答が書き込まれております。

http://www.oracle.com/technology/global/jp/ondemand/otn-seminar/index.html

過去のセミナ資料、動画コンテンツはOTNの
「OTNセミナー　オンデマンド コンテンツ」へ

※ダイセミ事務局にダイセミ資料を請求頂いても、お受けできない可能性がございますので予めご了承ください。
ダイセミ資料はOTNコンテンツ　オン　デマンドか、セミナ実施時間内にダウンロード頂くようお願い致します。

データマイニングとは

データマイニングとは

- データ・マイニングとは、大量に保管されているデータを自動的に検索して、単純な分析では得られないパターンや傾向を見つける手続きです。

- データ・マイニングでは高度な数学的アルゴリズムを使用して、データを分割し、将来のイベントの発生確率を判断します。

- データ・マイニングは、データからの知識発見（KDD）としても知られています。
  - KDD：knowledge-discovery in databases

データマイニングで何ができる？（一例）

大量のデータの中から「有意なデータ」を発掘

- 大量の顧客データの中から、「有意な顧客」を発掘
  - アップセルできそうな顧客／解約しそうな顧客
  - 顧客のセグメンテーション
- 大量の取引データの中から、「有意な取引」を発掘
  - 特定の併売パターン（一緒に売れる傾向の強い商品の組み合わせ）
  - 不正な取引
- その他大量のアイテムデータの中から、「有意なアイテム」を発掘
  - 製造部品データの中から、壊れる可能性の高い部品/不良品
  - 新薬検証データの中から、効果の高い新薬

データマイニングで何ができる？（イメージ）

顧客データ

顧客ID	性別	年齢	職業
101	男性	31	会社員
102	女性	28	主婦
103	女性	36	主婦
104	男性	43	会社員

マイニング
エンジン

解約顧客

顧客ID	解約予想	確率
101	Y	70%
102	Y	85%
103	N	58%
104	N	92%

販売明細データ

取引ID	商品	顧客ID	日時	数量
5020	CD-R	103	1/10/2010	1
5021	CD-R	110	1/10/2010	2
5022	CD-R	121	1/11/2010	1
5023	マウスパッド	103	1/10/2010	1

マイニング
エンジン

併売傾向（ルール）

購入（A）	併売（B）	信頼度	支持度
CD-R	CDケース	90%	7%
マウス	マウスパッド	88%	3%
CD-R	マウスパッド	51%	2%

マイニングエンジンとは？

- 関数のようなもの
  - Oracleの場合、PL/SQLまたはJava APIとして提供
  - データセットをインプットすると、有意な結果をアウトプットしてくれる
  - 有意な結果：データ、ルール、予測モデル（新たな関数のようなもの）

- ただし、インプットするデータによっては、有意な結果が出てこない場合もあるので、インプットデータは仮説を元に色々試してみる必要がある

Oracleのデータマイニング

- データマイニング機能
- Oracleの強み

Oracle Data Mining機能

機能名	内容
Classification 分類	履歴データを元に、未知のデータ（質的）を予測 ※異常値の検出としても利用可能
Regression 回帰	履歴データを元に、未知のデータ（量的）を予測
Clustering クラスタリング	データ全体の中で自然なグループを識別
Association 相関	AかつBなどの関係がデータ全体で発生する確率を特定 （ex.バスケット分析）
Attribute_Importance 属性評価	あるデータの予測に対して、使用する各データ項目の相対的な重要度を識別
Feature_extraction 特徴抽出	各データ項目を結合して、より上質な項目に変換 （データを説明する項目の削減・品質向上）

Oracle Data Miningアルゴリズム

ファンクション	アルゴリズム
分類(Classification)	**Decision Tree** **Naïve Bayes** **Support Vector Machine** **Generalized Liner Models**
回帰(Regression)	**Support Vector Machine** **Generalized Liner Models**
クラスタリング(Clustering)	**拡張*k*-Means** **OCluster**
異常検出(Anomaly Detection)	**Support Vector Machine**
相関(Association)	**Apriori**
属性評価(AI:Attribute Importance)	**MDL（最小記述長）**
特徴抽出(Feature Extraction)	**NMF**

【参考】PL/SQLによるマイニング処理例

- DBMS_DATA_MINING.CREATE_MODELプロシージャ

```
BEGIN
  DBMS_DATA_MINING.CREATE_MODEL(
    model_name          => ‘svmC_model',
    mining_function     => dbms_data_mining.classification,
    data_table_name     => ‘sample_data',
    case_id_column_name => 'cust_id',
    target_column_name  => ‘Affinity_card‘);
END;
/
```

サンプルの顧客データセットから、カード会員になりそうな
顧客を予測するモデルを作成

【参考】SQLの活用例

年収2000万円以上と思われる顧客の名前と仕事を抽出

```
SELECT ENAME, JOB from customers
where PREDICTION (svmR_model USING *) > 2000
AND COMM IS NOT NULL ;
```

PRIDICTION()・・・顧客の年収を予測

顧客をセグメント化し、重要度の高いセグメントごとに解約する危険性を確率で表示

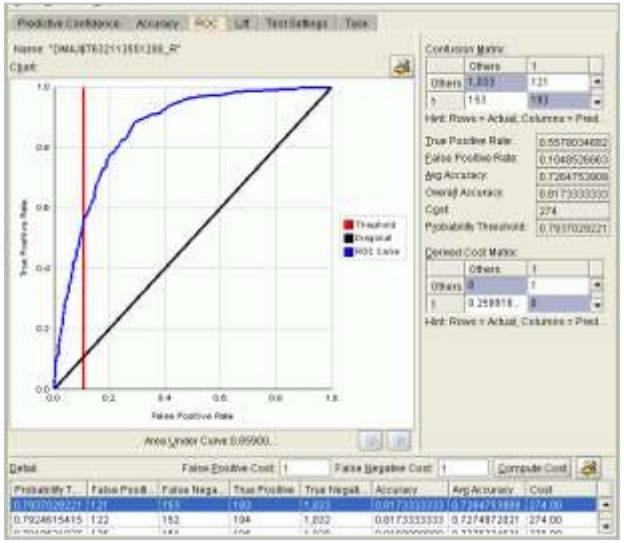
```
SELECT count(*) as cnt,
 AVG(PREDICTION_PROBABILITY(svmC_model, 'attrite' USING *)) as
  avg_attrite,
 AVG(cust_value_score)
FROM customers
GROUP BY CLUSTER_ID(clus_model USING *) ORDER BY avg_attrite
  DESC;
```

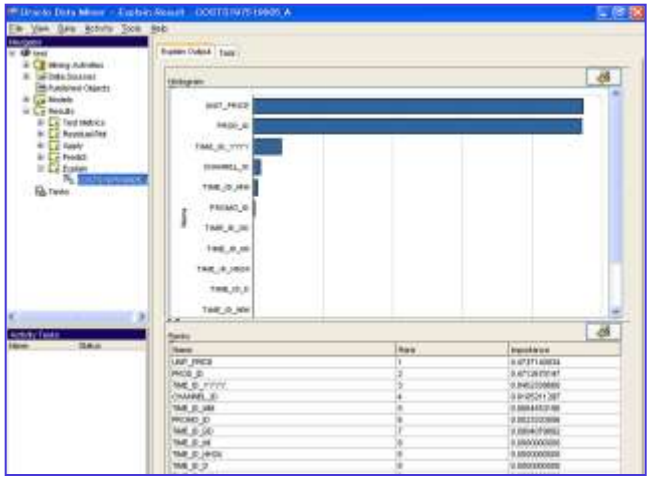
PRIDICTION_PROBABILITY()・・・顧客の解約確率を算出

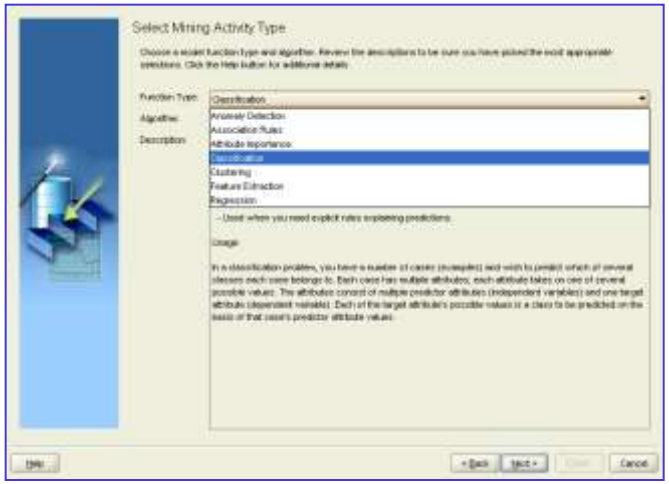
CLUSTER_ID()・・・顧客をクラスタに分類し、そのクラスタの識別IDを返す

【参考】Oracle Data Miner

- データマイニング用GUIツール（無償）
- マイニング作業の簡易化
  - ウィザードに沿って作業を実行
  - 結果の可視化
    - 残差プロット、ROC、リフトチャートなど





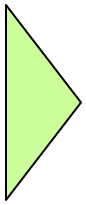


http://www.oracle.com/technology/products/bi/odm/odminer.html

Demonstration　バスケット分析

- シナリオ
  - パソコンショップの売上明細データから、よく併売されている商品を抽出します

商品	顧客ID	日時	数量
CD-R	103	1/10/2010	1
CD-R	110	1/10/2010	2
CD-R	121	1/11/2010	1
マウスパッド	103	1/10/2010	1
マウスパッド	115	1/11/2010	3



購入（A）	併売（B）	信頼度	支持度
CD-R	CDケース	90%	7%
マウス	マウスパッド	88%	3%
CD-R	マウスパッド	51%	2%

信頼度・・・A全体のうち、AかつBの割合

支持度・・・全ケースのうち、AかつBの割合

Oracleのデータマイニング

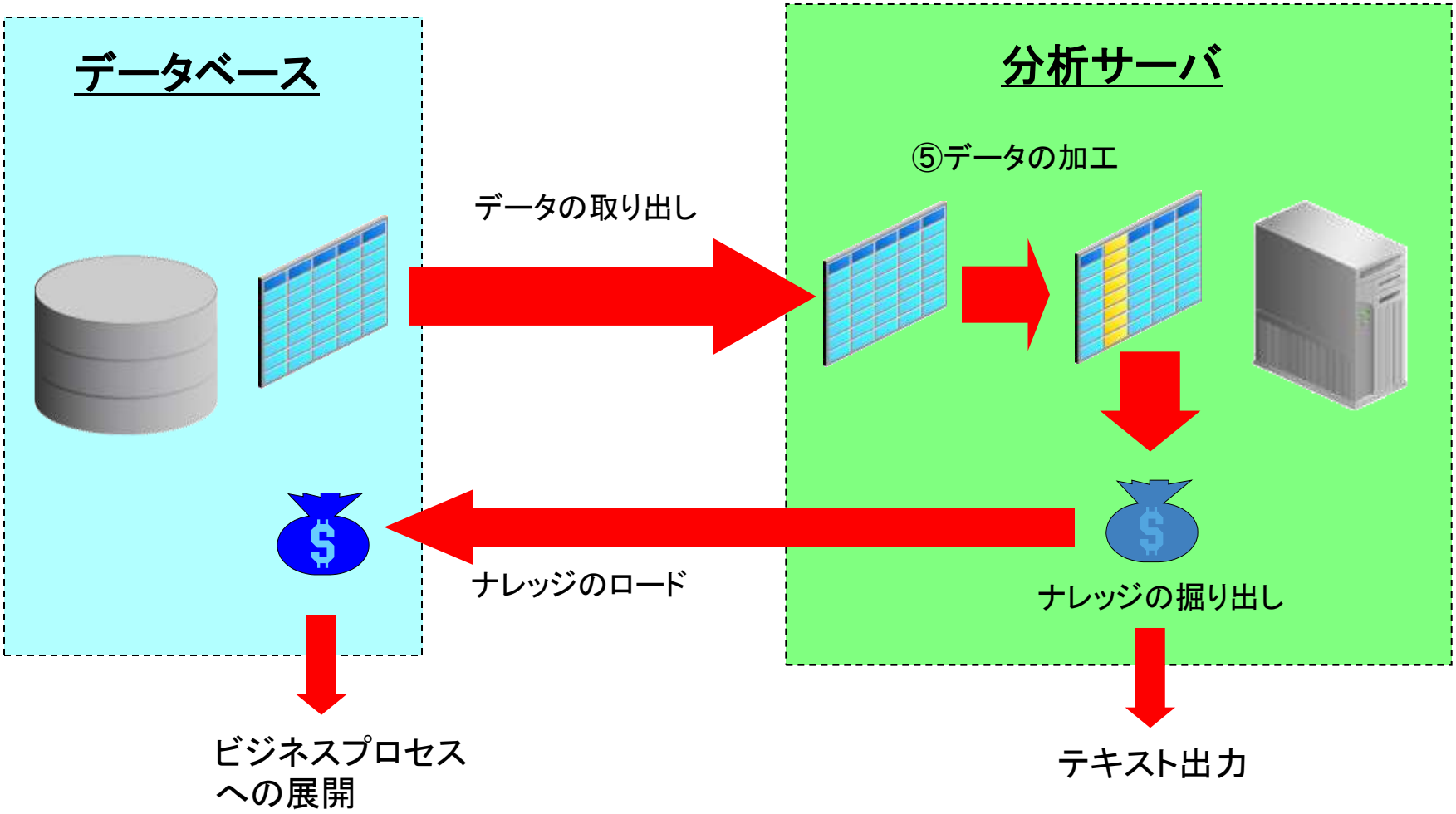
- データマイニング機能
- Oracleの強み

分析ニーズの拡大と大量データ時代

- 分析ニーズの拡大
  - 経営者・専門家から現場へ
  - 人による判断から自動化へ
  - 傾向の可視化から詳細の分析へ
  - 過去データから直近のデータへ（分析データの鮮度）
- 分析目的・背景
  - 売上拡大（マーケティング分析、顧客分析、レコメンデーション）
  - リスク縮小（例外・異常分析、チャーン分析）
- 大量データ
  - 業務の明細データ
  - ログ（通信ログ、ライフログ、センサーデータ）

ORACLE

従来の分析システム



従来の分析システムの課題

データベースと分析サーバ間のデータのやり取りが発生

- データベースからのアンロード
- データベースから取り出したデータを分析サーバへ転送
- 分析サーバでモデリングしたデータをDBへ転送
- モデリングデータをDBへローディング

分析作業を高速化するため、高スペックな分析サーバが必要

分析サーバ上でのデータセキュリティ対策が別途必要

データソースの管理、分析処理、セキュリティの管理ごとにI/Fや処理方式が異なり、作業の標準化や他システムとの連携が困難

データ移動の工数が増大

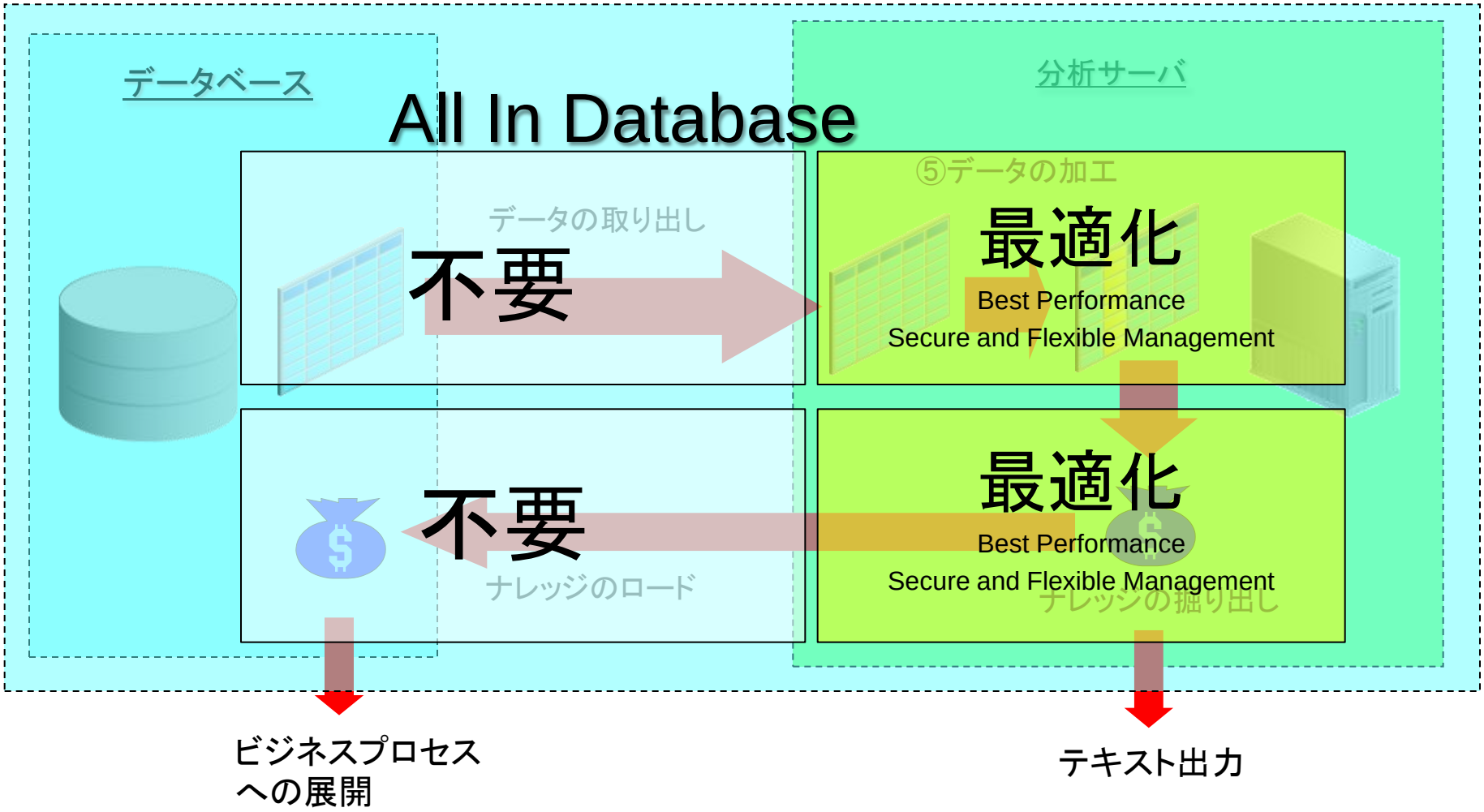
特にデータ量が増えるほど大きな影響

分析サーバの高コスト化

H/W、運用管理面でコストが増加

Oracle Architecture

In database Analytics system



In Databaseにおけるデータ分析最適化

- Best Performance
  - データの加工・集計・フィルタリング・モデリング・スコアリング
    - In Parallel
    - No CPU Overflow
    - Minimum I/O

- Secure and Flexible Management
  - Secure Data Management
    - アクセス制御、監査ロギング
  - Low cost of System Management
    - 自動モニタリング、自動診断
    - 他のアプリケーション（BI、CRMなど）との連携
      - ODBC、JDBC、SQL、アダプタ・・・

Oracle Architecture

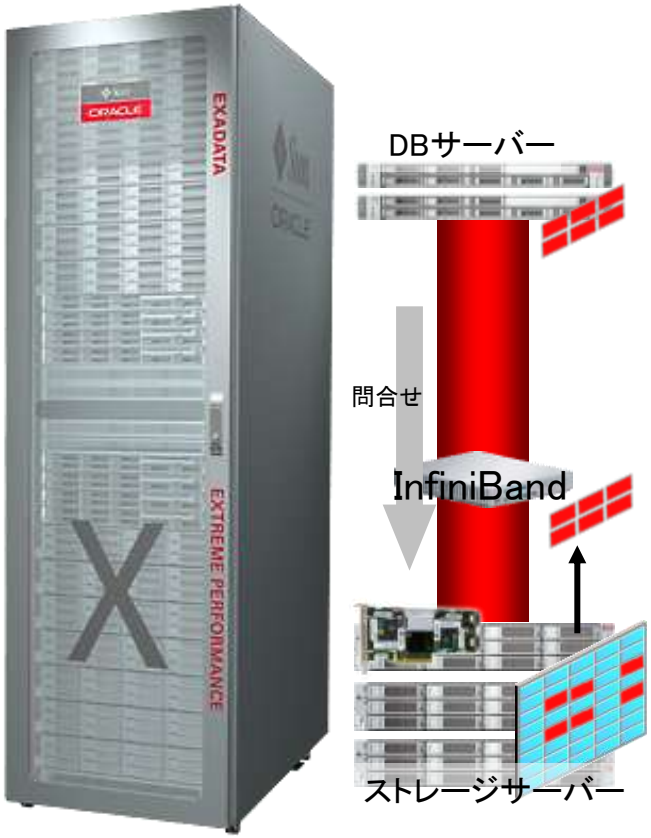
In database Analytics system

データベースと分析サーバ間のデータのやり取りが発生 ・データベースからのアンロード ・データベースから取り出したデータを分析サーバへ転送 ・分析サーバでモデリングしたデータをDBへ転送 ・モデリングデータをDBへローディング	データベースと分析サーバが統合 データのアンロードの必要はなし データの転送の必要もなし データの再ロードの必要もなし
分析作業を高速化するため、高スペックな分析サーバが必要	そもそも高価な分析サーバは不要 DBサーバの高速エンジンでデータクレンジングやモデリングを最適化
分析サーバ上でのデータセキュリティ対策が別途必要	データの操作はDB内で完了するため、DBのセキュリティ機能で対策は万全
データソースの管理、分析処理、セキュリティの管理ごとにI/Fや処理方式が異なり、作業の標準化や他システムとの連携が困難	全ての作業をPL/SQLやSQLベースで　標準化し、他システムとの連携も容易

Software. Hardware. Complete.

Oracleのソフトウェアと Sunのハードウェアを組み合わせて最適化された
データベースマシン「Exadata」

Oracle Exadata



Minimize I/O

列単位でデータを大幅に圧縮

ストレージ上のインデックスにより、不要なI/Oを削減

ストレージ上で、必要なデータだけを抽出・加工して
DBサーバへ転送

Express Highway Architecture

Fibre Channelの8倍のデータ転送帯域(InfiniBand 40Gb/s)

利用頻度の高いデータはフラッシュストレージ上に格納

データは自動的に Hot Spot のないように分散配置され、
ストレージサーバーを並列稼動する仕組み

顧客事例 - Xerox

THE DOCUMENT COMPANY XEROX UNITED STATES

“データマイニングアルゴリズムとデータがOracle Databaseに一緒に格納されているため、膨大なデータをアルゴリズムにかけ、分析、予測するために外部プログラムに移動する必要がありません。 このことは75%以上のコスト削減につながっています。”

Tracy E. Thieret Ph.D Principal Scientist
Xerox Innovation Group Imaging and Solution Technology Center

- コピー機、複合機からのデータを収集し、以下の分析を実施
  - お客様の使用状況の分類
    - ジョブの長さ、使用量、
  - リプレースの典型的な要因
  - CRUライフタイム
    - 使用量やジョブの長さなどとの相関関係
  - 消耗品の生産
    - お客様の使用状況との相関関係

ORACLE

Data Mining をはじめるには

- Oracle Databaseソフトウェアのインストール
  - インストール時に「Data Mining RDBMS Files」にチェック
  - Data MinerはOTNより別途インストール（解凍するだけ）

- ユーザの作成
  - データマイニングで使用するユーザを作成します
  - ユーザに必要な権限を設定します
    - CREATE MINIG MODELなど

- サンプルスキーマのインストール
  - チュートリアルを試してみる場合は、サンプルスキーマのデータを使用しますので、インストールが必要です。
  - サンプルスキーマはDBCAにてDBを作成（カスタム以外）時、または別途CompanionCDに含まれるスクリプトから手動にてインストールできます。

Data Miningをはじめてみよう

- Data Mining概要
  - http://otndnld.oracle.co.jp/document/products/oracle11g/111/doc_dvd/datamine.111/E05704-02/toc.htm
- Data Mining管理者ガイド
  - http://otndnld.oracle.co.jp/document/products/oracle11g/111/doc_dvd/datamine.111/E05705-02/toc.htm
- Data Miningアプリケーション開発者ガイド
  - http://otndnld.oracle.co.jp/document/products/oracle11g/111/doc_dvd/datamine.111/E05706-02/toc.htm
- Oracle Data Miner & Tutorial
  - http://www.oracle.com/technetwork/database/options/odm/downloads/index.html

※マニュアルのリンク先は11gR1のものです

あなたにいちばん近いオラクル

Oracle Direct

まずはお問合せください

Oracle Direct　検索

システムの検討・構築から運用まで、ITプロジェクト全般の相談窓口としてご支援いたします。

システム構成やライセンス/購入方法などお気軽にお問い合わせ下さい。

Web問い合わせフォーム

専用お問い合わせフォームにてご相談内容を承ります。
http://www.oracle.co.jp/inq_pl/INQUIRY/quest?rid=28

※フォームの入力には、Oracle Direct Seminar申込時と同じ
　ログインが必要となります。
※こちらから詳細確認のお電話を差し上げる場合がありますので、ご登録されている連絡先が最新のものになっているか、ご確認下さい。

フリーダイヤル

0120－155－096

※月曜~金曜　9:00~12:00、13:00~18:00
（祝日および年末年始除く）

ORACLE

以上の事項は、弊社の一般的な製品の方向性に関する概要を説明するものです。また、情報提供を唯一の目的とするものであり、いかなる契約にも組み込むことはできません。以下の事項は、マテリアルやコード、機能を提供することをコミットメント（確約）するものではないため、購買決定を行う際の判断材料になさらないで下さい。オラクル製品に関して記載されている機能の開発、リリースおよび時期については、弊社の裁量により決定されます。