BIからデータ解析ツールまで大量データを手軽に解析するツールとテクニック

株式会社INTHEFOREST

自己紹介

冨田　和孝

肩書き：　株式会社INTHEFOREST 代表取締役社長
Cassandraコンサルティング

Cassandra Conference in Tokyo / Cassandra Summit JPN Organizer
Cassandra勉強会主宰
２か月に一度程度開催。第37回まで開催しています。

職種：DB・インフラ屋
以前、某レストランサーチのDBA
高負荷・大容量・大規模のOracleRACとPostgreSQLと
MySQLに苦しめられ続けた経験あり。

Agenda

- ビッグデータを取り巻く環境
- データ解析
- 解析ツール
- まとめ

ビッグデータを取り巻く環境

- 既存のBIとビッグデータ
  - データは扱う人が把握できる量を超えるととたんにコントロールが行えなくなります。
  - 今までのBI　→　個別のPCでコントロールできるデータ量でレポートを作成
  - 現在のBI　→　人が把握できないデータ量を把握できる用に加工してレポートを作成

このことを把握できない限りビックデータはバズワードのまま

ビッグデータを取り巻く環境

- ビッグデータの歴史
  - Webとインデクサー
  - Googleとインデックス
  - Hadoop登場
  - 構造化データと非構造化データ
  - IoTとセンサーデータ

ビッグデータを取り巻く環境

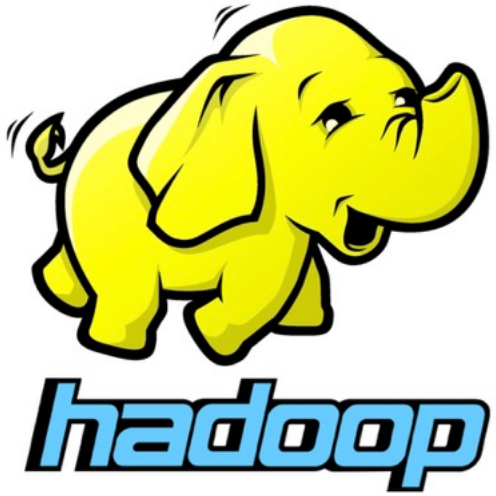
- ビッグデータの歴史
  - Webとインデクサー
    - インターネット黎明期、インターネット上でWebページとして日々データが増加していきました。増加するのすべての Web サイトを索引付けすることで、利用者が検索語を入力して関連 Web ページを一覧表示できるようにインデックスサービスが誕生していきます。
    - この頃のビッグデータは、Web ページの内容を保存してそれを分解し、検索可能な形に変換するという形で用いられました。基本的に、検索エンジンはデータを様々な構造化されていないフォーマットで入手し、より構造化された形で解釈する必要がありました。

ビッグデータを取り巻く環境

- ビッグデータの歴史
  - Googleとインデックス
    - Googleは様々な方法を用いて効率的な索引を作成する方法を構築していきます。
    - 世界最大のアプリケーション、データセット、データセンター並の規模と量を扱い、何百、何千ものサーバを利用する極端なパラレル処理を行っていました。

ビッグデータを取り巻く環境

- ビッグデータの歴史
  - Hadoop登場
    - Googleが考え出したデータ解析用のツール群を汎用性を持たせて再実装
    - Googleが発表した論文を元に汎用的に人々が利用できるようにHadoopツール群が作成されます。色々な種類のツールを組み合わせてHadoopエコシステムと呼ばれます。



ビッグデータを取り巻く環境

- ビッグデータの歴史
  - 構造化データと非構造化データ
    - 企業のデータセンターや外部 Web リソースのデータも増していました。
    - 電子メール、文書、画像および動画の中の非構造化データが、構造化データを凌ぐ勢いで増大していきました。

ビッグデータを取り巻く環境

- ビッグデータの歴史
  - IoTとセンサーデータ

消費財から計器、自動車まで、企業はあらゆるものにセンサーを設置。
ペタバイトのデータが普通に生成される環境にあります。

従来型の業界におけるビッグデータの革新的利用の例が現れるにつれ、企業はその増大するデータセットや外部データを分析することで得られる運営および競争における利点を認識し始めました。

データ解析

- データ解析とは
  - データ解析とは何らかの目的を持って表現された文字や符号、数値などを収集し、分類、整理、成型、取捨選択したうえで解釈して、価値のある意味を見出すこと
  - BIは、経営・会計・情報処理などの用語で、企業などの組織のデータを、収集・蓄積・分析・報告することで、経営上などの意思決定に役立てる手法や技術のこと。(wikipedia)

データ解析

データ解析方法の変化

- 小さいデータ
- 固定化したデータ
- サンプリング

- 巨大なデータ
- 連続したデータ
- 繰り返し

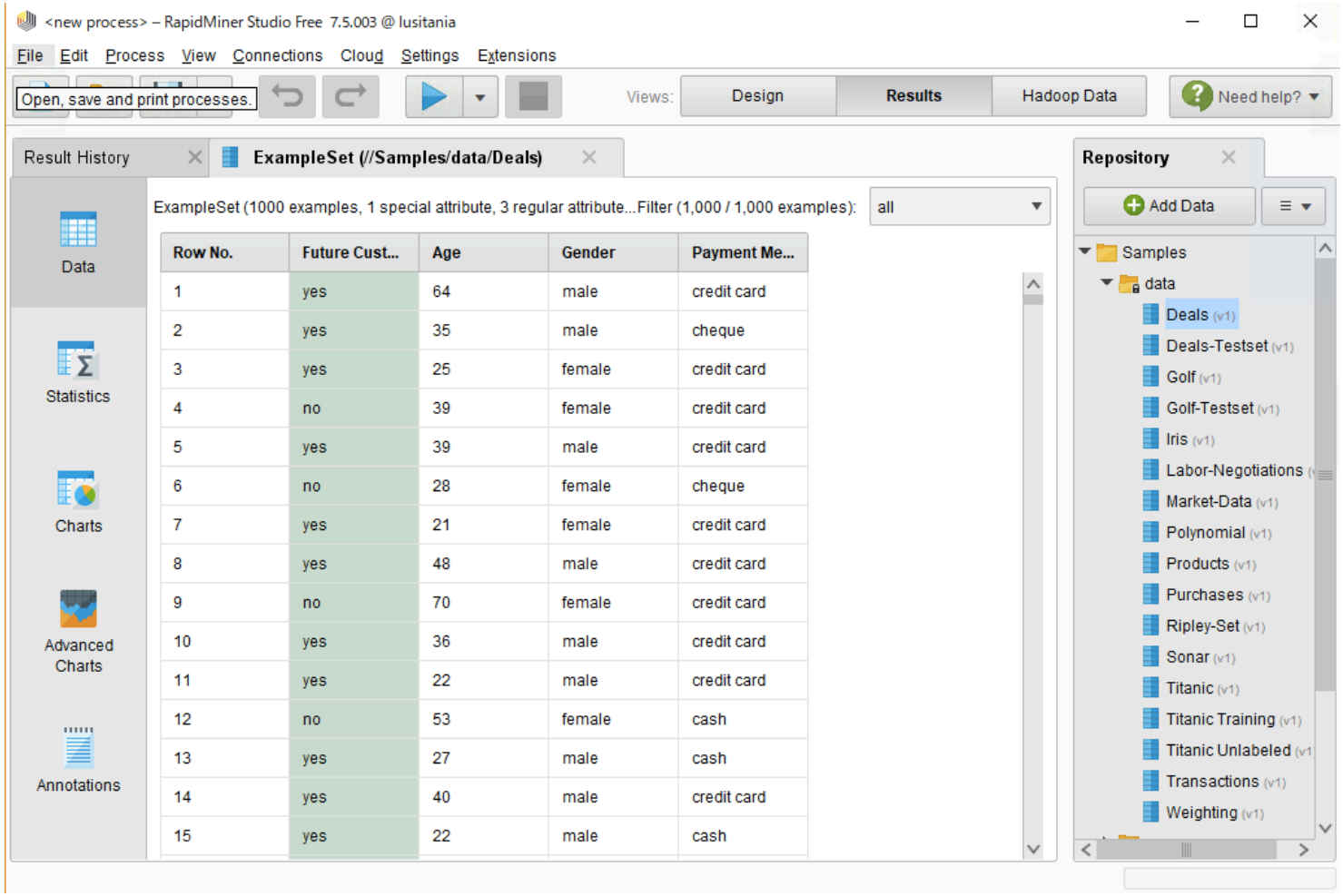
データ解析

データ解析方法の変化

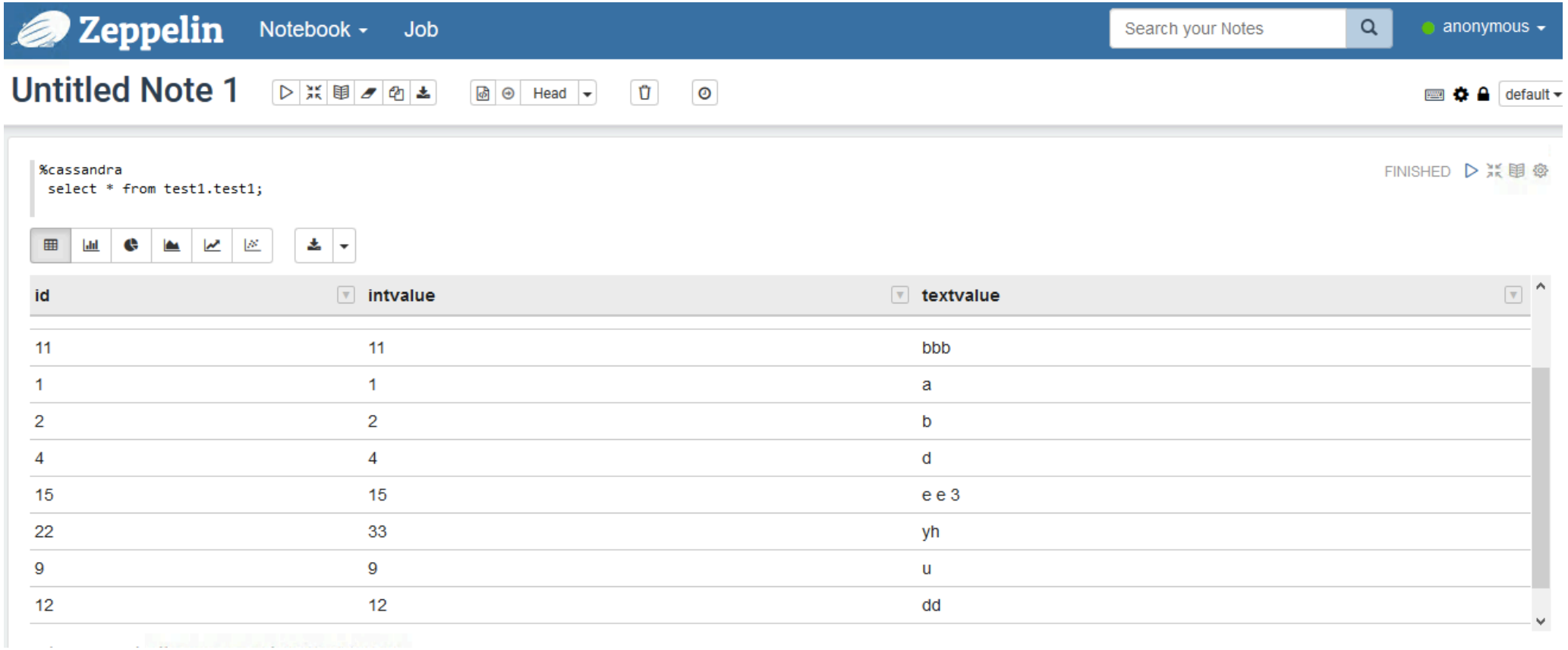
データの解析方法は膨大に貯められたデータをインタラクティブに繰り返し解析することにより傾向や変異などを抽出する方法に変化。

ツールにも抽出から解析まで繰り返しトライ ＆ エラーを繰り返すことのできることが求められるようになった。

RapidMinor


<new process> – RapidMiner Studio Free 7.5.003 @ lusitania
File Edit Process View Connections Cloud Settings Extensions
Open, save and print processes.
Views: Design Results Hadoop Data
Need help?
Result History
ExampleSet (//Samples/data/Deals)
ExampleSet (1000 examples, 1 special attribute, 3 regular attribute...Filter (1,000 / 1,000 examples): all
Data
Statistics
Charts
Advanced Charts
Annotations
Row No. Future Cust... Age Gender Payment Me...
1 yes 64 male credit card
2 yes 35 male cheque
3 yes 25 female credit card
4 no 39 female credit card
5 yes 39 male credit card
6 no 28 female cheque
7 yes 21 female credit card
8 yes 48 male credit card
9 no 70 female credit card
10 yes 36 male credit card
11 yes 22 male credit card
12 no 53 female cash
13 yes 27 male cash
14 yes 40 male credit card
15 yes 22 male cash
Repository
Add Data
Samples
data
Deals (v1)
Deals-Testset (v1)
Golf (v1)
Golf-Testset (v1)
Iris (v1)
Labor-Negotiations
Market-Data (v1)
Polynomial (v1)
Products (v1)
Purchases (v1)
Ripley-Set (v1)
Sonar (v1)
Titanic (v1)
Titanic Training (v1)
Titanic Unlabeled
Transactions (v1)
Weighting (v1)

Apache Zeppelin



id	intvalue	textvalue
11	11	bbb
1	1	a
2	2	b
4	4	d
15	15	e e 3
22	33	yh
9	9	u
12	12	dd

Jupiter Notebook