



## 分散データシステムを活かすETLとBIソ リューションPentaho

株式会社INTHEFOREST 富田 和孝

# 自己紹介

---



## 富田 和孝

肩書き: 株式会社INTHEFOREST 代表取締役社長  
Cassandra商用サポート、Cassandraコンサルティング他

職種: 元々はDB・インフラ系エンジニア  
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

---

- ▶ 解析を行うこと
- ▶ Pentaho Spoon
- ▶ Pentaho Report Designer



# ◇解析を行うこと ビッグデータはバズワード？



## ビッグデータの定義とは？

情報技術分野の用語としては、通常のデータベース管理ツールなどで取り扱う事が困難なほど巨大な大きさのデータの集まりのこと。(wikipedia)

## 巨大なデータとは？

人によって定義が異なる

10GB?  
100GB?  
1TB?  
10TB?  
100TB?  
1PB?  
1EB?

100万行?  
1000万行?  
1億行?  
10億行?  
100億行?

言葉の定義は曖昧

# ◇解析を行うこと データを貯めるということ



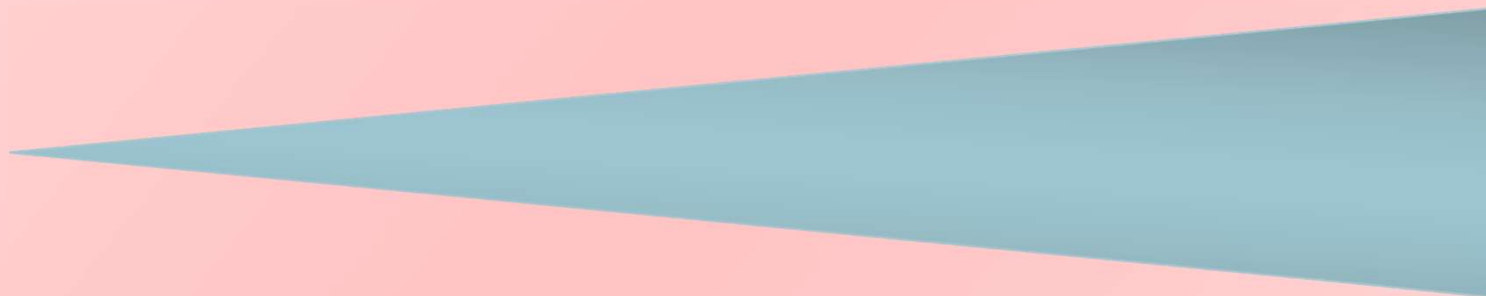
データ種別

データのタイプによって生成量が異なる

マスター系

更新系

アーカイブ系



## ◇解析を行うこと アーカイブ系データ



ビッグデータ = 巨大データ

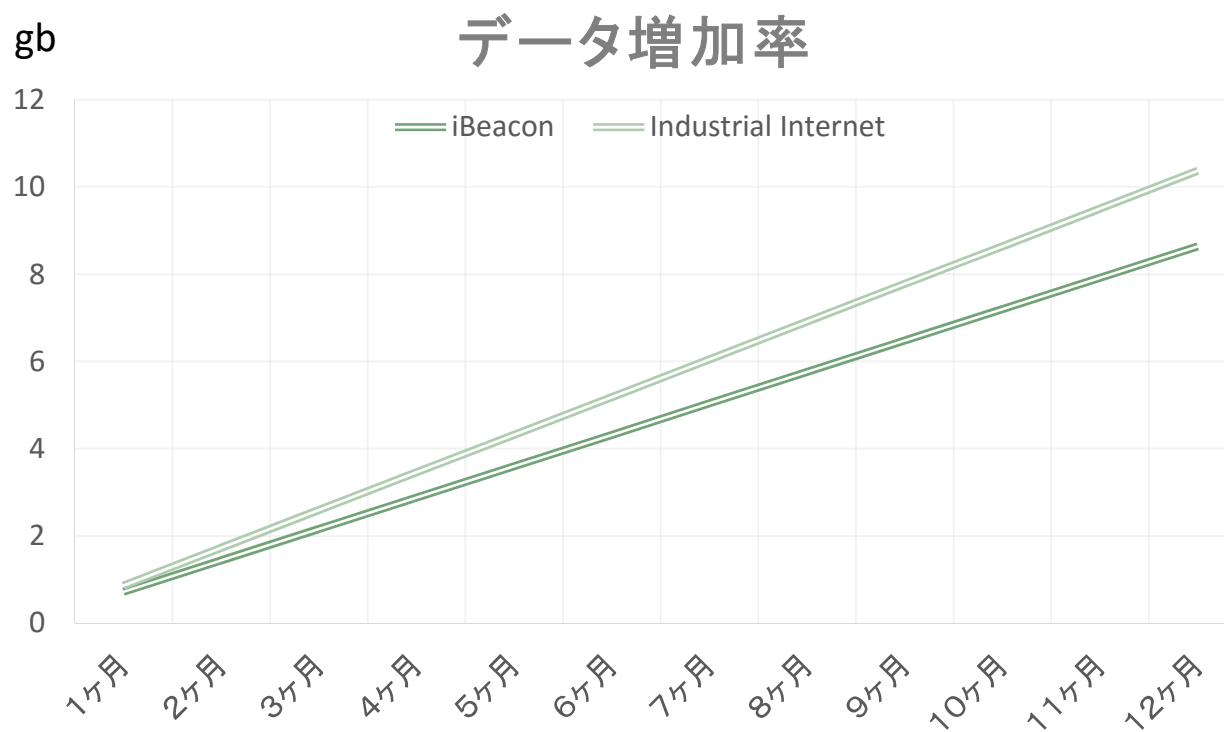
そんなに大きいデータなんかあるの？

- 各種ログ
  - 行動履歴
  - イベントログ
  - サーバーログ
- センサーデータ



データの廃棄を辞めると言う選択肢

# 解析を行うこと アーカイブ系データとは



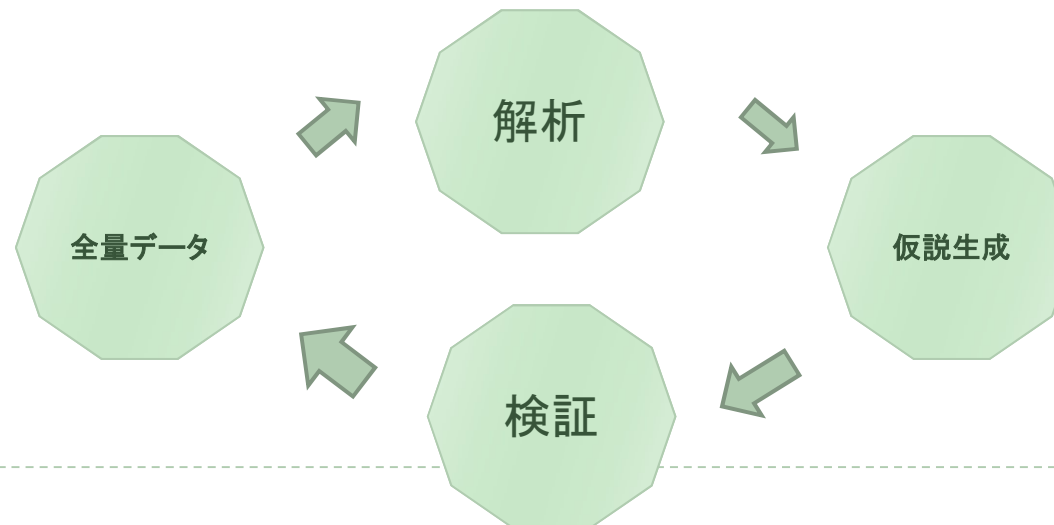
# ◇解析を行うこと 解析方法の変遷



## 従来のデータ解析

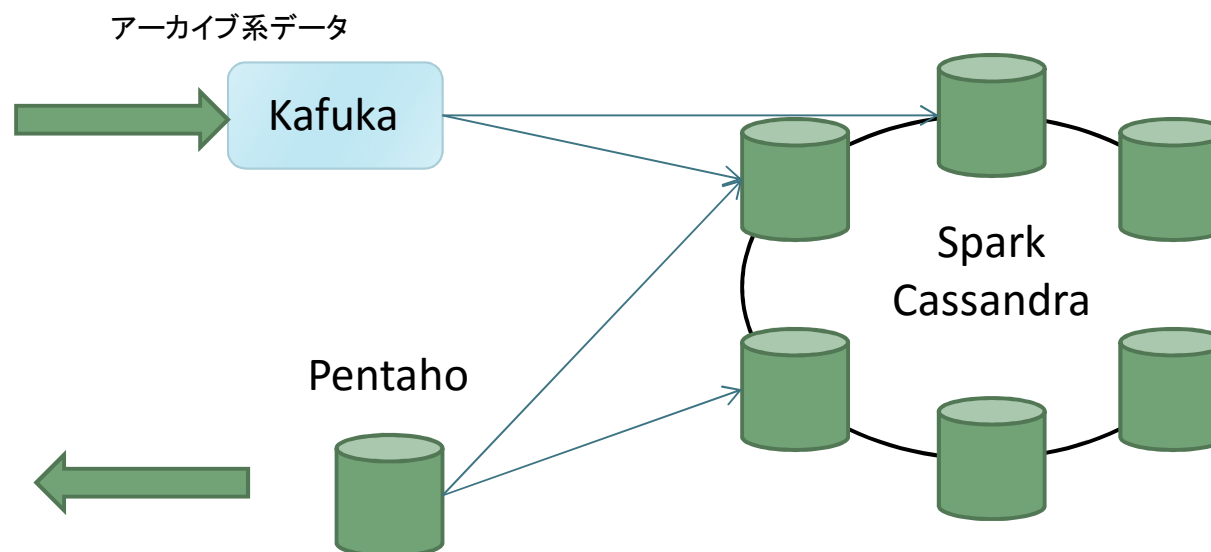


## 最近のデータ解析





# ◇解析を行うこと 巨大データを解析する仕組み



Kafuka + Spark + Cassandra + Pentaho を用いたテキストマイニングソリューション

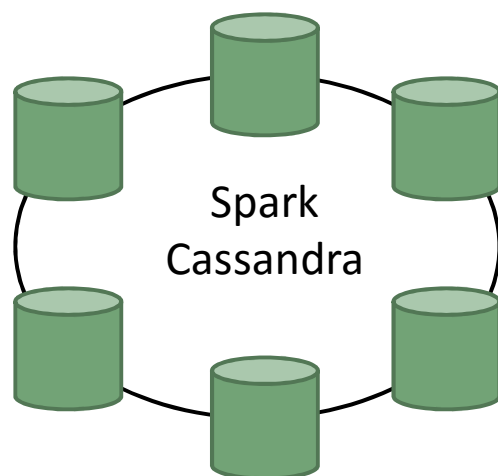
語群の意味を用いたクラスタリングなど



# ◇解析を行うこと 多段解析



一次解析



機械学習  
Map Reduce

MySQL等



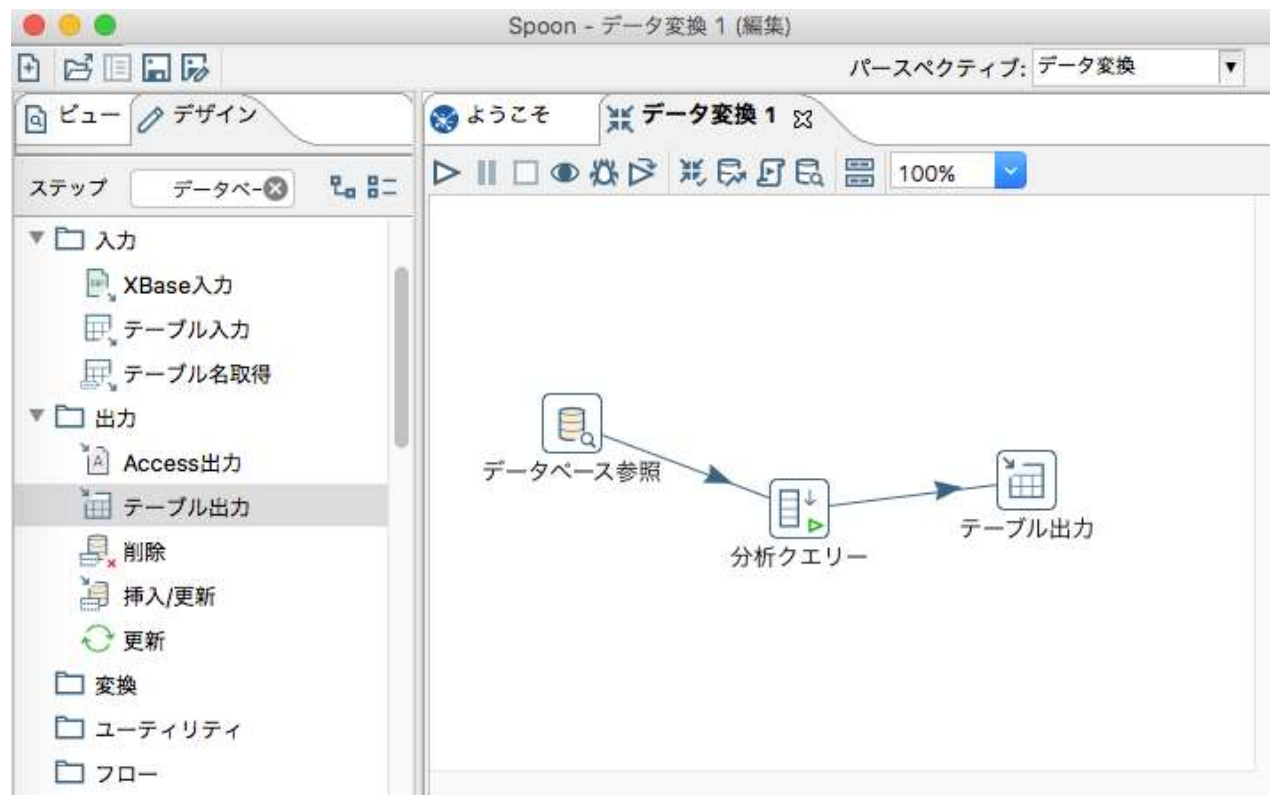
一次マージデータ保存

Pentaho



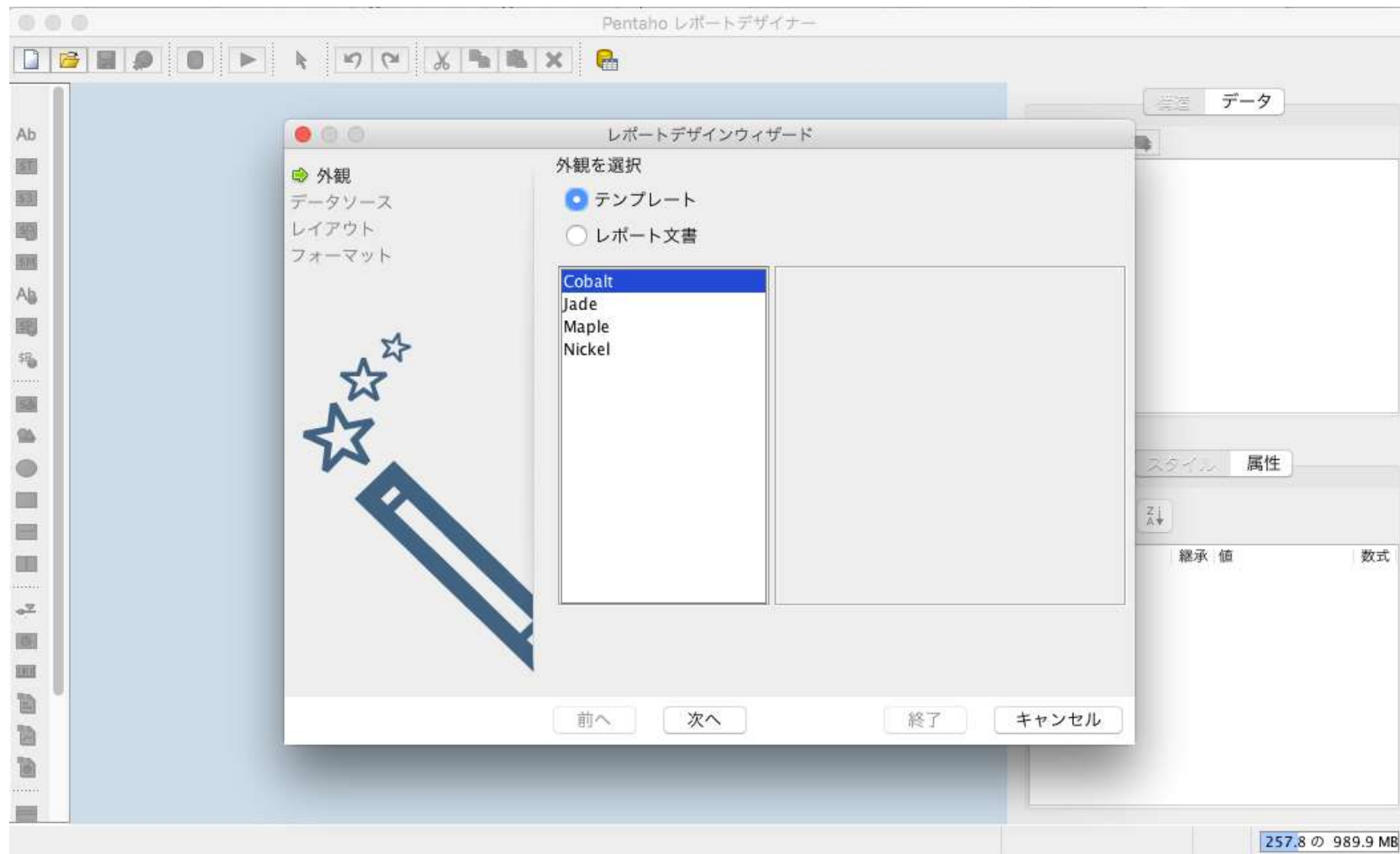
一次解析  
レポート作成

# Pentaho Spoon ETLパッケージ



# Pentaho Report Designer

## BIレポート作成ツール



# まとめ

---



- ▶ 日々増大するデータを一次解析のみで対応するのは難しい
- ▶ 解析の軸は日々変わる
- ▶ 都度都度に会わせてデータ解析を手軽に行えることの重要性

