Tamrのご紹介

機械学習を原動力とした企業データの統合

2018 年 10 月

tamr
data united

Tamr の DNA

会社概要

Tamr のソリューションは、革新的な分析法と運用上の成果を強化するために、機械学習と人間の専門知識を組み合わせることにより企業データを統一するものです。

本部: マサチューセッツ州ケンブリッジ

追加オフィス: サンフランシスコ、ニューヨーク、ロンドン

設立: 2013 年

従業員 : 100 名

インベスター

NEA
Google ventures
SBI Holdings
GE VENTURES
intage
THOMSON REUTERS
SAMSUNG
MassMutual FINANCIAL GROUP
Hewlett Packard Enterprise

中心的な創立者

Dr. Michael Stonebraker
共同創立者＆ CTO
前：Ingres/Postgres、HP Vertica
創立者

Andy Palmer
共同創立者＆ CEO
前：HP Vertica 創立者兼 CEO

VERTICA An HP Company
tamr data united
PostgreSQL

Tamr: Global 2000 においてデータエンジニアリングの「腕力」を築く

「データネイティブ」と競合する必要のある企業を支援

精選されたデータは、変革的分析と運用上の成果の原動力となる

10倍もの削減
新しいデータセットの統合
6 ヵ月から 2 週間

ケーススタディ動画

$500M+の節約
ソーシング分析から
サイロ化されたビジネス全体まで

ケーススタディ動画

(Europe)
カスタマー・インサイト
30 点超の地域にあるサイロ化された販売代理店システム全体を網羅する統一された購買者プロフィール

ケーススタディ動画

5000 件超のケーススタディ
研究者に活力を与える統一された臨床研究データ

ケーススタディ

エンジニアリング及びオペレーションのための全てのパーツの統一的視点

支出分類
パイロットモデルから
Google Cloud Platform でのライブまで 6 週間

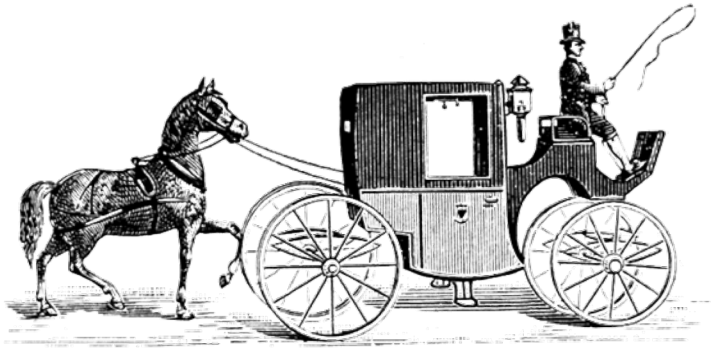
在庫の最適化
5 船隊全体でのパーツの平準化により在庫 $100M の削減

※カリブ海を中心に展開するクルーズ会社

マーケットの可視性
新しい分析を可能にするための販売代理店からの商品販売データを統一

企業 AI 採用における大きなリスク：

格言のごとく、馬車である AI を馬であるデータの前に繋ぐという行為



「アルゴリズムはコピーが簡単である。
データは守れる障壁である。」

Andrew Ng, スタンフォード大学
「百度」チーフサイエンティスト

ハーバード・ビジネス・レビュー：**AI** が今できることとできないこと

念願：実際に必要になる前にどのパーツが必要となるか予測する

現実：実際のどのパーツを在庫しているか？

240 0.312 丸棒材
FORTACERO, S.A. DE C.V.
在庫：20

在庫：3

丸棒材
FORTACERO

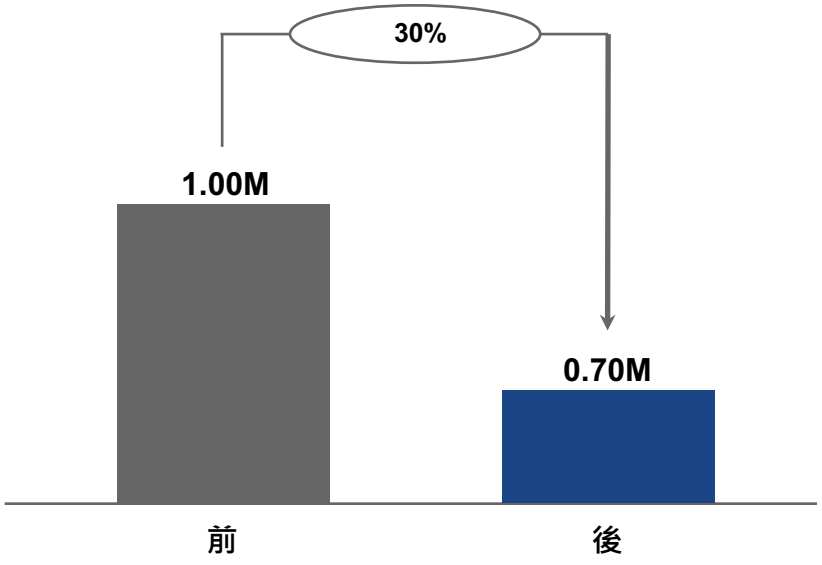
丸棒材 0.312x240
ERIE CONCRETE & STEEL CO
在庫：5

丸棒材 0.312 X 240.HRS
ERIE STEEL SUPPLY CO
在庫：7

基本的な質問に正しく答えることから得られる大きな成果

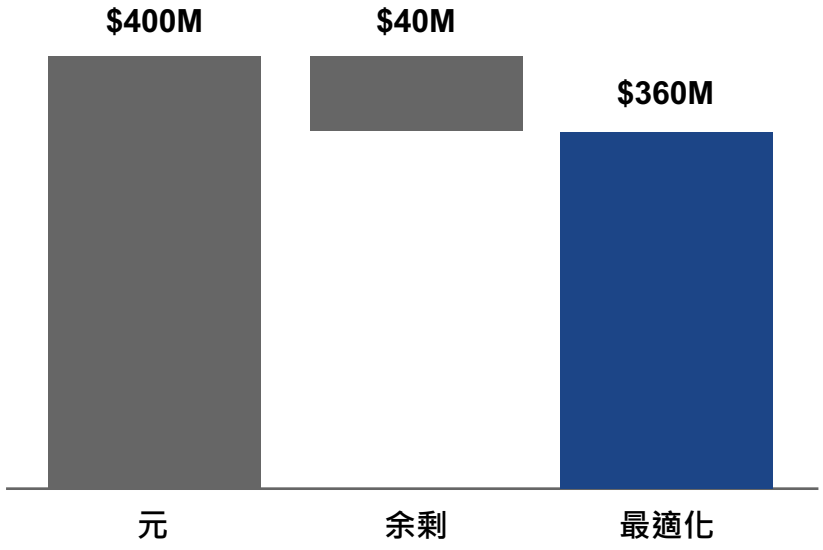
質問実際にどれだけスペアパーツを在庫しているのか？

ユニークなスペアパーツ総数

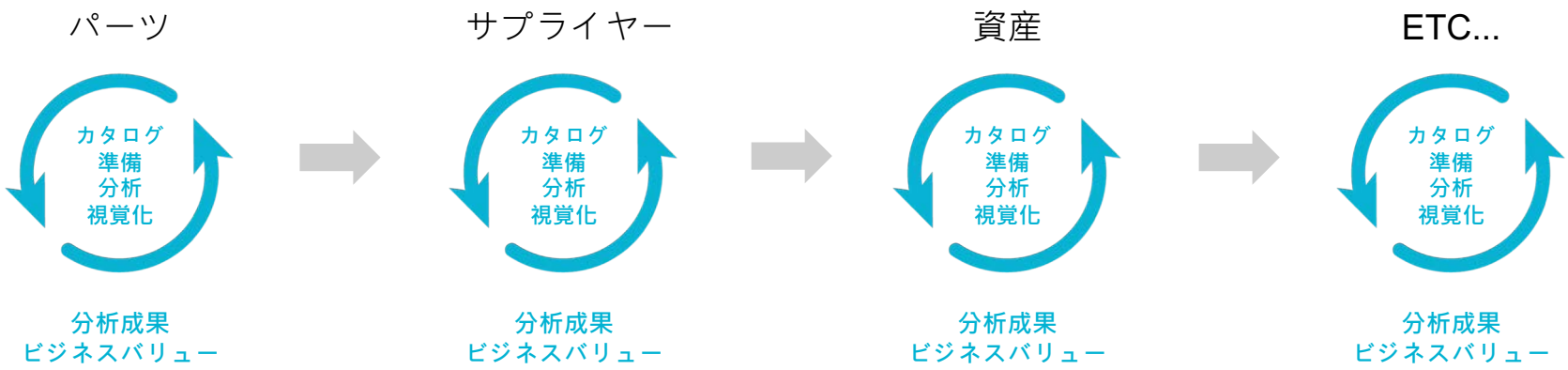


インパクト：在庫を $40M 削減

在庫の価値



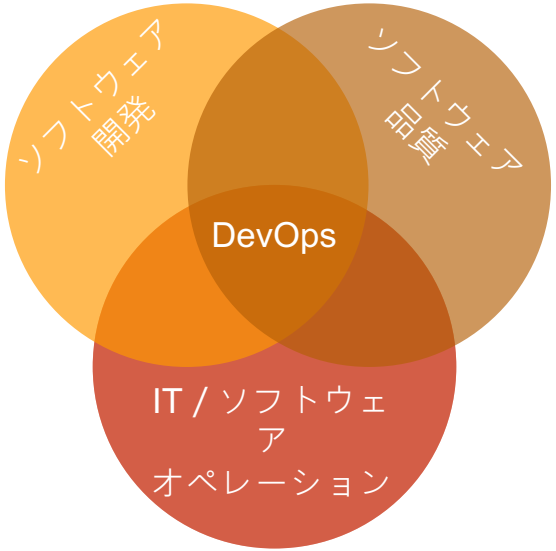
アジャイルなデータマスタリング：行動力を高める迅速な成果



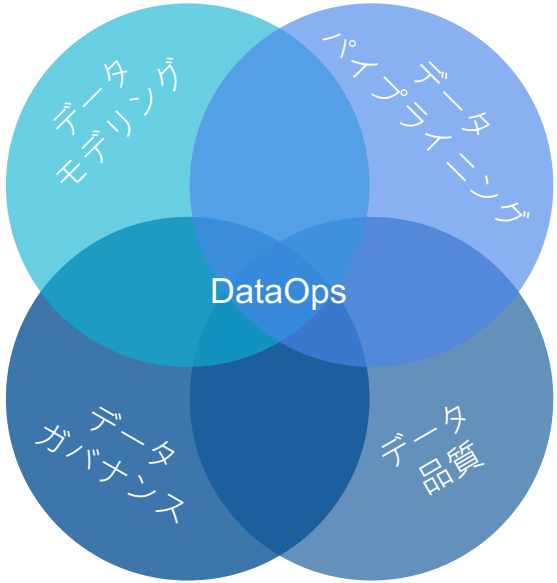
実現される成果

- 在庫の削減
- 資産設計判断の改善
- より良い価格条件の交渉
- 最適なサプライヤーとのパートナーシップ
- 資産寿命の改善
- オーナーシップの総コストの削減
- ...

DevOps >> DataOps



大インターネット企業においてフィーチャーベロシティを高める



大企業において分析ベロシティを高める

オープンな DataOps エコシステム：キーとなる技術的原則

ソース

企業データ

社外データ

人、プロセス、ツール、サービス、データ

- **自動化**
- アジャイル / **連続的** (Kafka/Lambda)
- キーエンティティのための論理データモデル
- **オープン / ベスト・オブ・ブリード** (1 つのプラットフォーム / ベンダーでない)
- 双方向 (幅広い**フィードバック**)
- コラボラティブ (**人間** / マシンがコアにある)
- SOA/レストフル・インターフェース/テーブル・In/Out/RESTful
- スケールアウト / ディストリビューティッド
- **ハイブリッド** (クラウドとオンプレムのミックス)

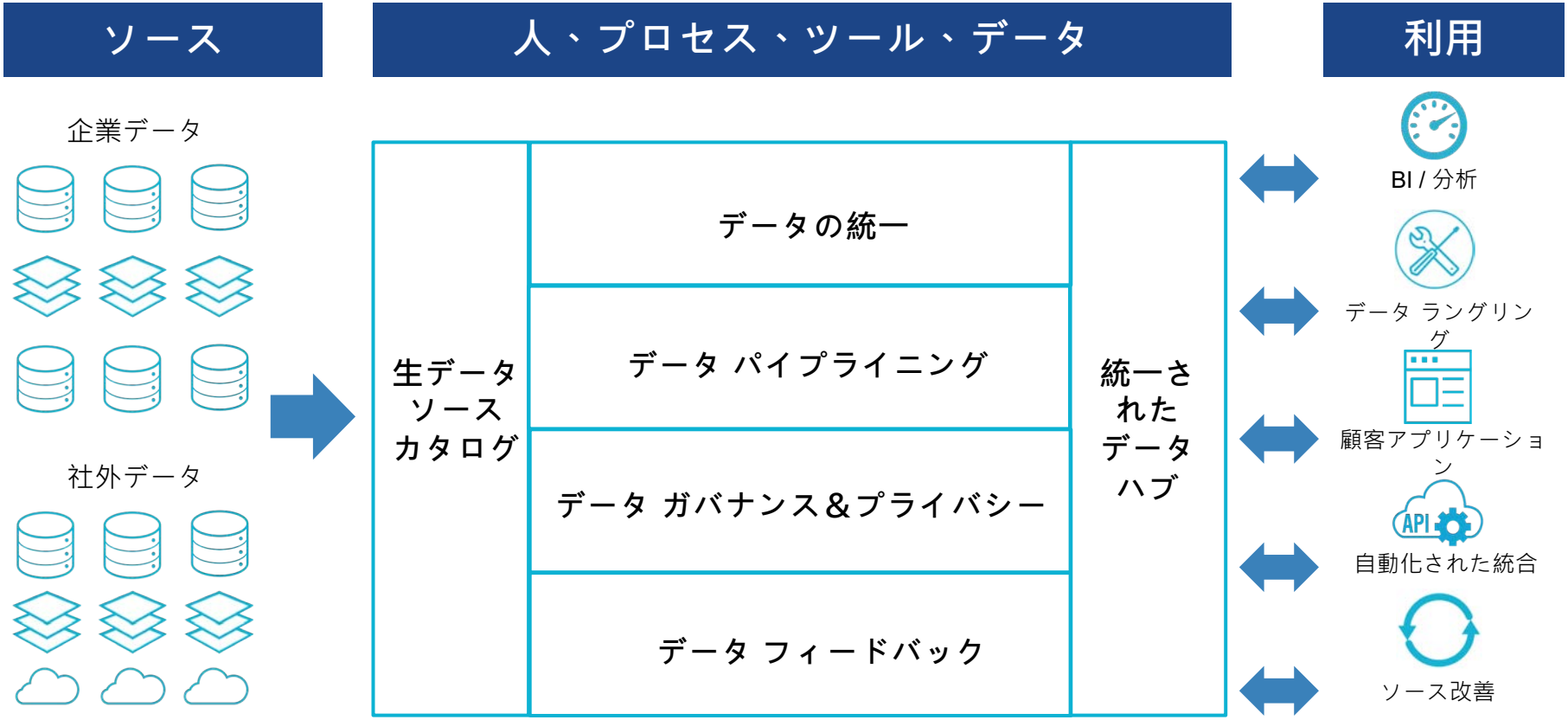
使用

BI / データ視覚化 / 分析

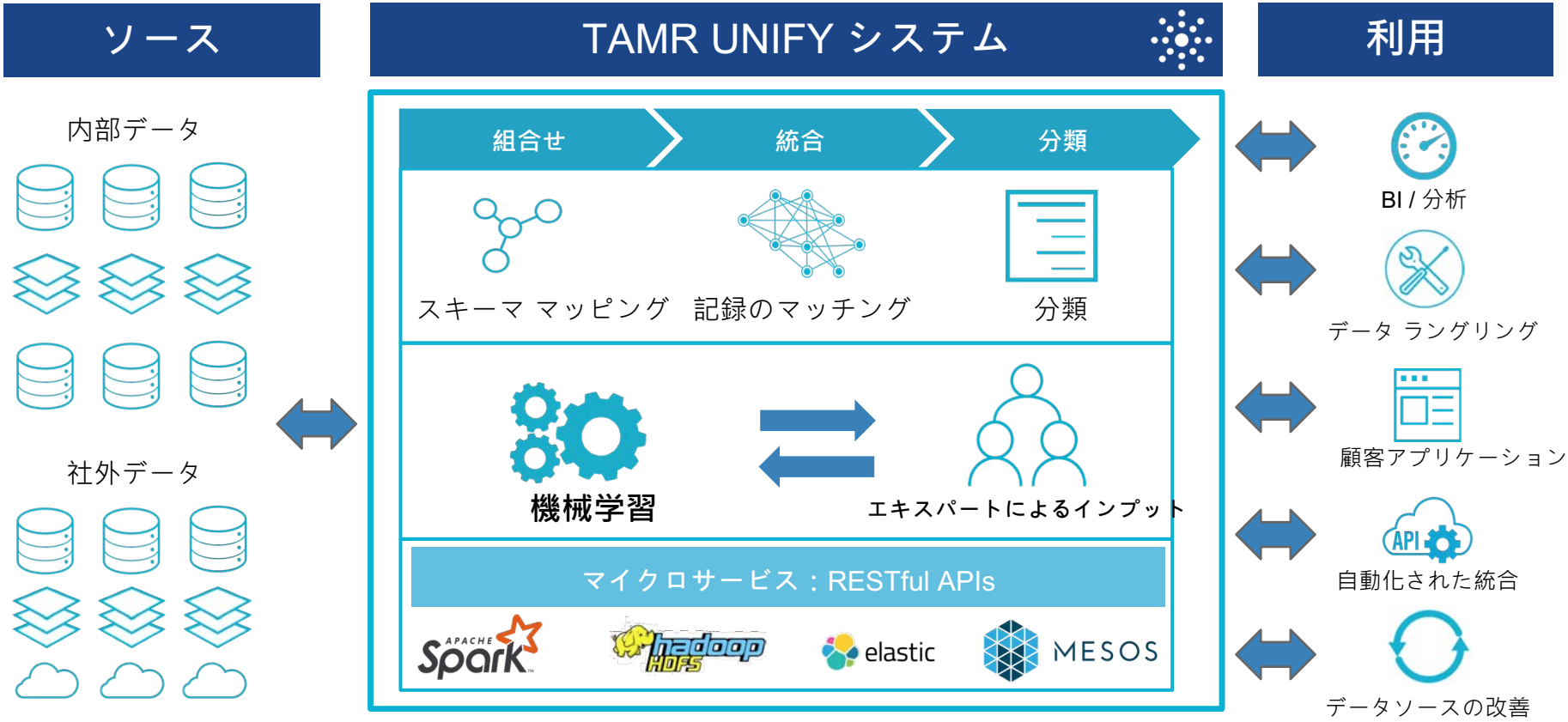
データ ラングリング

顧客アプリケーション

DataOps ソリューション

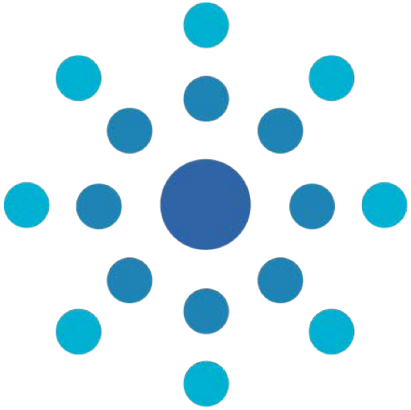


Tamr Unify


ソース
TAMR UNIFY システム
利用
内部データ
社外データ
組合せ
統合
分類
スキーマ マッピング
記録のマッチング
分類
機械学習
エキスパートによるインプット
マイクロサービス：RESTful APIs
APACHE Spark
hadoop HDFS
elastic
MESOS
BI / 分析
データ ラングリング
顧客アプリケーション
API
自動化された統合
データソースの改善

重要な結論

- データの多様性を理解し受け入れる
- 少量でも高品質のデータの方が多くのユーザーには意味がある
- 大企業は、データインフラを迅速に変えつつある
- 大海を煮詰めるようなことはしない。アジャイルなデータマスタリングの方が優れている
- データに関しては人間の知識の方がテクノロジーより同程度かそれ以上に重要



tamr

data united

andy.palmer@tamr.com

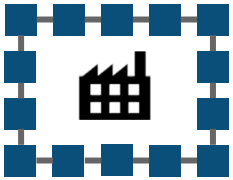
活用ケース＆
カスタマーの成功例

活用ケース例

カスタマー

目標とされるインタラクション：ある特定のカスタマーとのインタラクションを全て把握し、目標とされる関与プロセスを促進する

サプライヤー
/ 契約業者

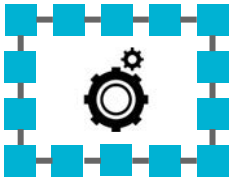


サプライヤー関係マネジメント：選別判定や継続的管理を改善するため、サプライヤーや契約業者とのアクティビティの統一視点を得る

トランザクショ

支出の最適化購入コスト削減のために何を購入しているかへの粒度の細かい洞察

資産



フリートマネジメントフリートの不調や問題点をより良く予想できるよう、フリート関連データの組織化 (例、作業指示書、パーツ、etc)

在庫の最適化ダブりのパーツを特定し、在庫を削減し、スペアパーツ購入ではより良い価格を交渉する

ケーススタディ：Toyota Motor Europe

高度に断片化したデータ集合体から顧客情報を全体的に統一

“顧客データに対して一貫性のある視点が欠けていたので、イノベーションを実現し顧客の期待に応えることが制限されていました。このような問題に立ち向かうために、企業データの統一アプローチとTamr に出会うことができました。MDM ツールのような従来からの商業製品は、そのトップダウンアプローチが単一のデータモデルを必要とするので、我々には受け入れられなかったのです。”

Matt Stevens
情報システム ディレクター

技術的なチャレンジ

- 30 カ国各地で管理されていた顧客データを効果的に統合
- 国レベルでのデータの収集と管理の柔軟性を維持
- 新しいソースを迅速に統合
- 各地のマイグレーションプロジェクトをサポート

技術的な結果

- 現在 125 点のソースを統合、プロジェクト完了時には 500 点
- 現地のシステムを中断することなく、新しいマスターデータが利用可能になる
- 新たなデータ統合に平均 1 ~ 2 週間
- TME France CRM のマイグレーションは 6 週間で完成

ビジネスチャレンジ

- 国々を移動する顧客に対してより良いサービスを提供
- 顧客のタッチポイント全体にわたり、一貫性のあるエクスペリエンスを提供
- 期待を超える能力を強化することにより顧客のニーズを理解・予測する

ビジネス成果

- CSRs は Tamr マスターの顧客データを検索するのに単一 UI のみを持つ
- POS (ポイント・オブ・セール) / サービスにおける、より良い顧客の知識
- 顧客に対する統一された視点により新しい分析・運用利用が可能

ケーススタディ：Société Générale

グローバルな Soc Gen の支出を常に現在時点で反映できる、初めて実現された統一的視点

SOCIETE GENERALE

「30 時間の作業で、120 億ユーロの支出の 75%を正確に分類したが、これは 600 万件に相当する」

- ソーシングメソッド&情報システム責任者 Jean Baptiste

総コストを低減させると同時に社内期日とセキュリティ要件を満たすため、Google Cloud Platform を利用

技術的なチャレンジ

- レガシーソリューション (ルールベースのオンプレミス ERP アドオン) は全ての支出をカバーするようスケール調整ができない
- 維持とサポートにコストがかかる
- 厳しいセキュリティとインフラストラクチャ要件のため、ソリューションの変更が困難である。

技術的な結果

- 新しいソースの追加に 60 日かかっていなのが 5 日に短縮
- 手作業によるサポート作業が 90%削減 (IT&調達)
- プロジェクト全体が 2 ヵ月で完成
- GCP で実行

ビジネスチャレンジ

- 300 名を超える従業員のユーザーグループ (主なソーシング) は、会社の支出に対して不完全で不正確な全体像しか把握していない。経営陣の指示は 3 ヵ月以内に素早くこの問題を解決すること。

ビジネス成果

- Société Générale では、グローバルなの支出を常に現在時点で反映できる、初めて実現された統一的視点を手に入れた。
- 分類精度が 90%超向上したことにより、分析と信頼が大幅に向上

ケーススタディ：General Electric

迅速なマルチドメイン企業マスタリングは、$500M以上の価値をもたらす

“サプライヤーのデータ統合は大きな勝利。”
Bill Ruh
CEO GE Digital 兼
GE CDO

“我々は、Tamr のテクノロジーが GE の規模の企業でも変革的な結果をもたらすことを直接体験している。企業データセットを統合する際のコストと複雑さが大きく低減されるので、分析打開策の結果として、これまでは利用できなかったチャンスを生み出している。”
Lisa Coca
GE Ventures マネージングディレクター

技術的なチャレンジ

- **サプライヤー**：75+ ERP システム及び 2M サプライヤー記録から、統合された視点を構築
- **パーツ**：BU 8 点全体にわたる調達システムにおける 25M 点のユニークでないパーツ
- M&A：既存のマスタービューで取得したエンティティからのデータを統合

技術的な結果

- パイロット版からグローバルに配備されたデータパイプラインまで 6 ヵ月以下； 記録 2M 点を 700k に統合
- 25M 点をユニークなパーツ 6.4M 点に削減； 5 層の分析可能分類
- GE のデータと取得 3 点からのデータ統合が 2 週間以内

ビジネスチャレンジ

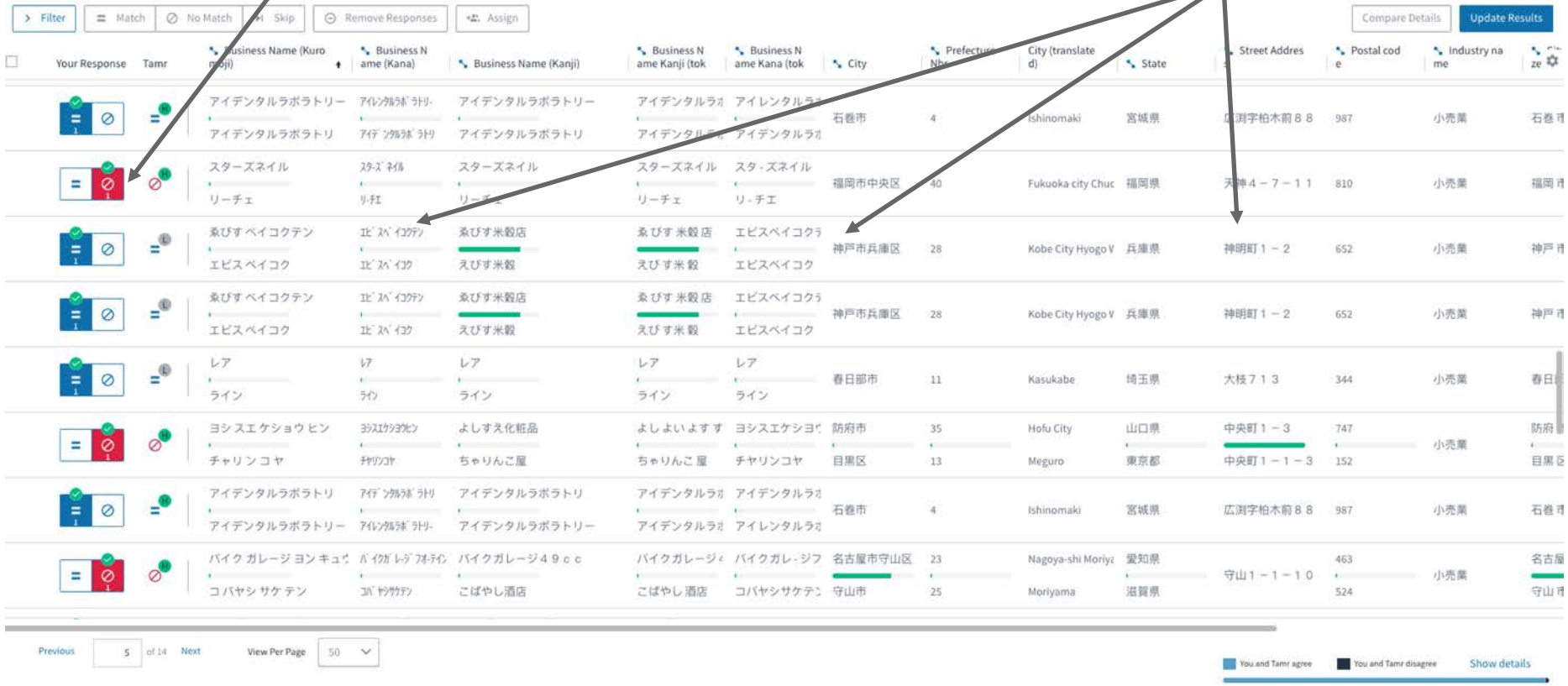
- **サプライヤー**：すべての交渉においてサプライヤーは GE の最適条件を得ることができる
- **パーツ**：最もコスト効果の高いサプライヤーにソーシング戦略を最適化
- M&A：取得後のシナジーの実現速度を高める

ビジネス成果

- Tamr によってマスター統合されたサプライヤービューにより最初の年で **$80M 節約**
- **年間の節約** $300M (直接支出の 0.5% 削減)
- サプライヤー、調達、カスタマーベースのチャンスを迅速に識別

Tamr は、SME レスポンスを正確に予測している

Tamr は、推薦を行えるよう全てのカラムの類似性を監視している



tamr

Tamr は、SME の知識の再現するため、非常に高度な精度、正確さ、リコールを実現できる

商品 / マーケットの適合性の確立

Tamr のプラットフォームは、複数のコア使用ケースで繰り返し展開されており、データ統一のさまざまなニーズに対応できる

カスタマー 360度ビュー

カスタマーエクスペリエンスの改善、アップセル / X セルの駆動、規制コンプライアンスへの合致のための B2B / B2C データ統合

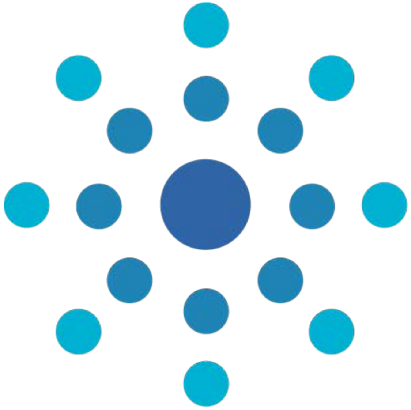
調達最適化

サプライチェーン効率化のためのサプライヤーマスタリング、パーツマスタリング、支出分類

臨床研究

基礎研究から応用分野におよぶ研究のブレイクスルーを可能にするために従来の臨床研究データを新しいデータ基準 (SDTM) に揃える

さらに90 件を超えるUse Caseの一覧で、繰り返しのパターンが明らかになるにつれて市場開拓ソリューションの基盤を提供する
(例、商品マスタリング、M&A データ統合、在庫統合)



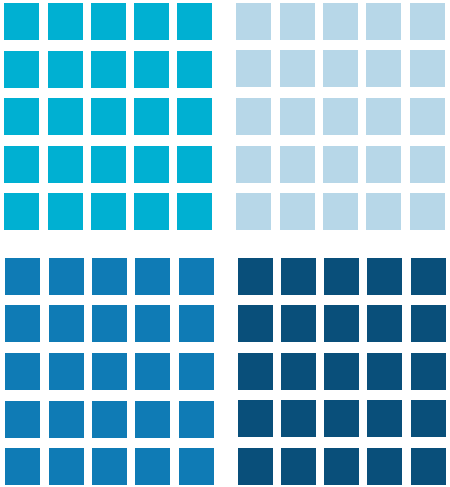
tamr

data united

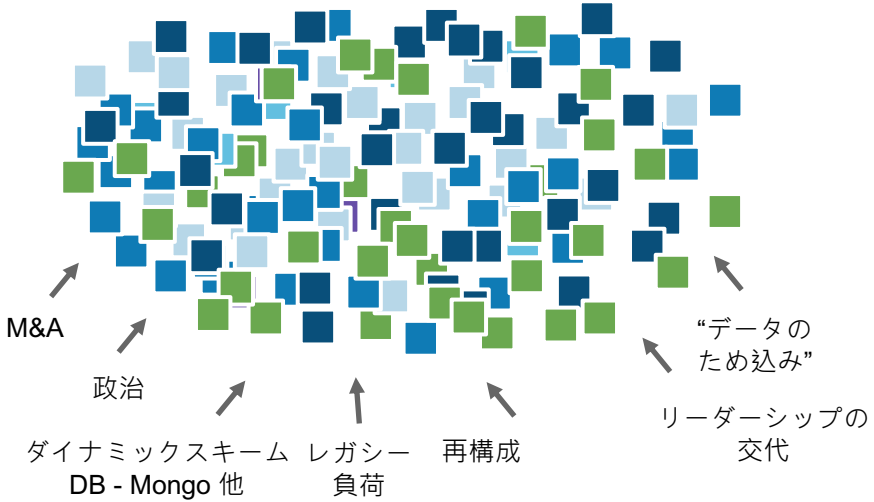
andy.palmer@tamr.com

企業内のデータの状態大規模なデータ負債

認識：企業データは、一般にこのようなものであると考えられている



現実：“ランダムなデータのサラダ”
継続的な変更やエントロピーによるデータ負債

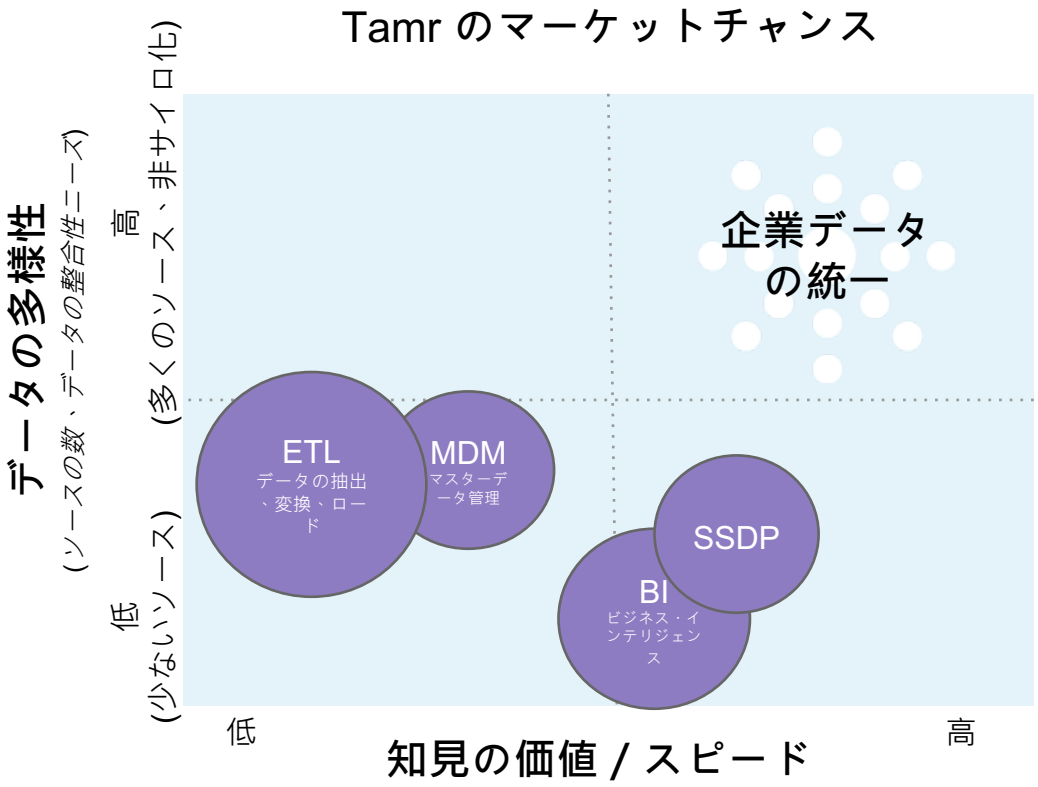


結果：

1.分析・利用に対してデータ準備に時間がかかり過ぎる
2.BI / 分析プロジェクトの高い失敗率
3.抜本的改革イニシアチブが「不可能」で始められない

企業はデータ統合のスケール化を行い始めたばかりである

データが多様になり分析コンテキストが拡張するので、ルールベースのアプローチは、スケーラブルでない



- Tamr は、既存のルールベースのETL /MDMを補完するものであり置き換えるものではない
  - 確率論的モデルは、加速度の大きさの桁が違う
  - コラボラティブ (エキスパートのインプット) アプローチは、高品質の結果を実現
- 対象となりうるチャンスには、伝統的なデータマーケットセグメントの $5.4B分が含まれる
  - データ統合：$2.7B, 6.7% CAGR
  - MDM：$1.3B, 3% CAGR
  - データ品質：$1.4B, 16.7% CAGR
- 新しい種類の問題でもTamrの確率論的アプローチなら解決が可能
- 幅広いデータからの知見の価値は、相対的に変革的である (つまり、単なるレポーティングではない)

参考：Gartner Inc. によるData Preparationの発展的見解

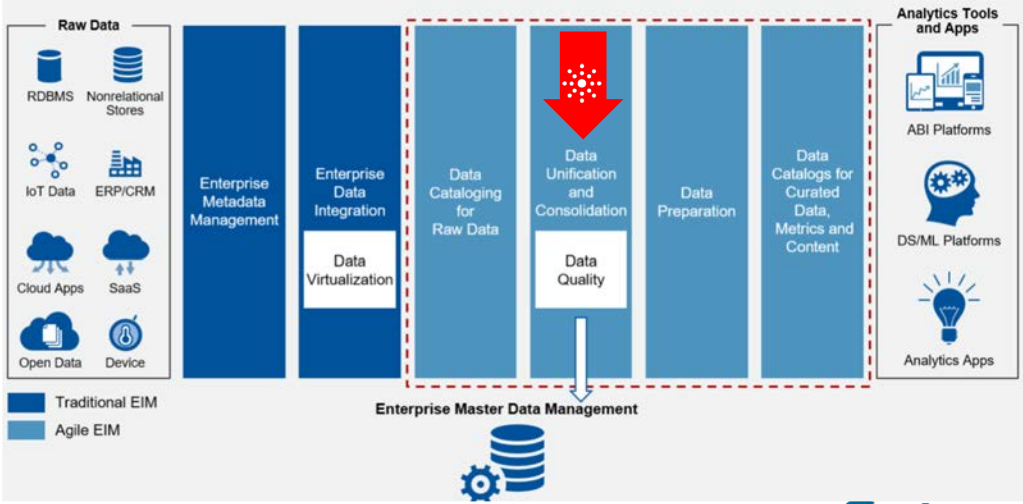
“2019 年までに、俊敏で精選された社内外のデータセットを幅広いコンテンツ制作者に提供しているデータ分析機関は、そうしていない機関よりも 2 倍のビジネス利益を実現できるであろう。”

Gartner Market Guide for Data Preparation, 2017 年 12 月

Gartner の重要な所見

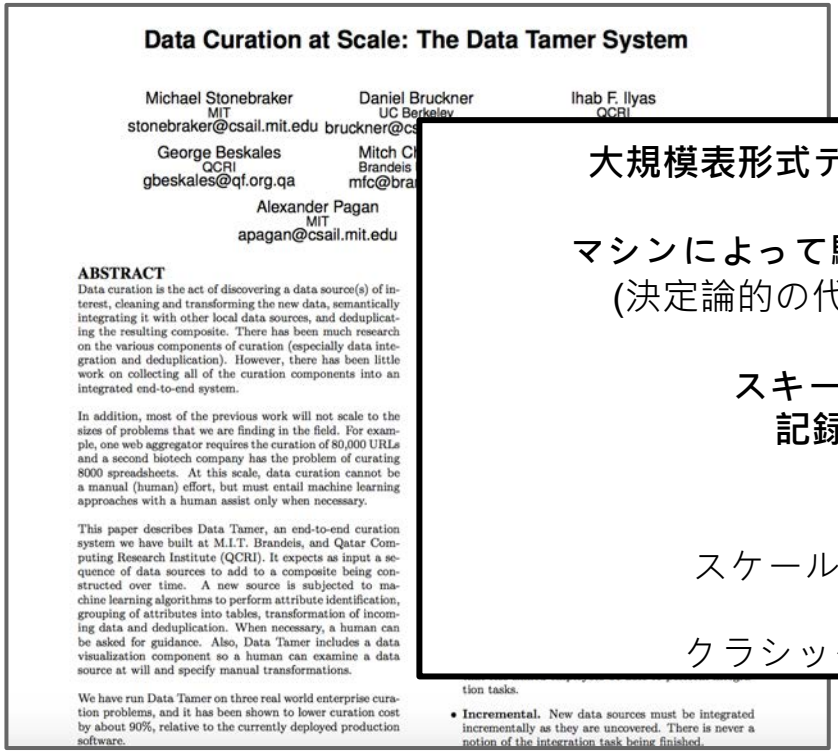
- データ作成マーケットは大きく拡大している
  - $750M → 2021 年には $1.5B = 年成長率 18.5%
- 問題点は大きい
  - 時間の 60% はData Prepatationに使われる
- データ準備ツールがデータ統合の新しい基準になりつつある
  - 2020 年までには、新規プロジェクトの 50% がData Prepatationツールを利用する
- 機械学習は必須
  - “データ準備のパーソナル化と自動化の重要な機能”
- データの統一がカギ！
  - Gartner のアジャイル企業情報管理の一部

伝統的 + 新しい企業情報管理の集合



Gartner

MIT 学術プロジェクト発行特許



Data Curation at Scale: The Data Tamer System

Michael Stonebraker
MIT
stonebraker@csail.mit.edu

Daniel Bruckner
UC Berkeley
bruckner@cs

Ihab F. Ilyas
QCRI

George Beskales
QCRI
gbeskales@qf.org.qa

Mitch Cherniack
Brandeis
mfc@bra

Alexander Pagan
MIT
apagan@csail.mit.edu

ABSTRACT

Data curation is the act of discovering a data source(s) of interest, cleaning and transforming the new data, semantically integrating it with other local data sources, and deduplicating the resulting composite. There has been much research on the various components of curation (especially data integration and deduplication). However, there has been little work on collecting all of the curation components into an integrated end-to-end system.

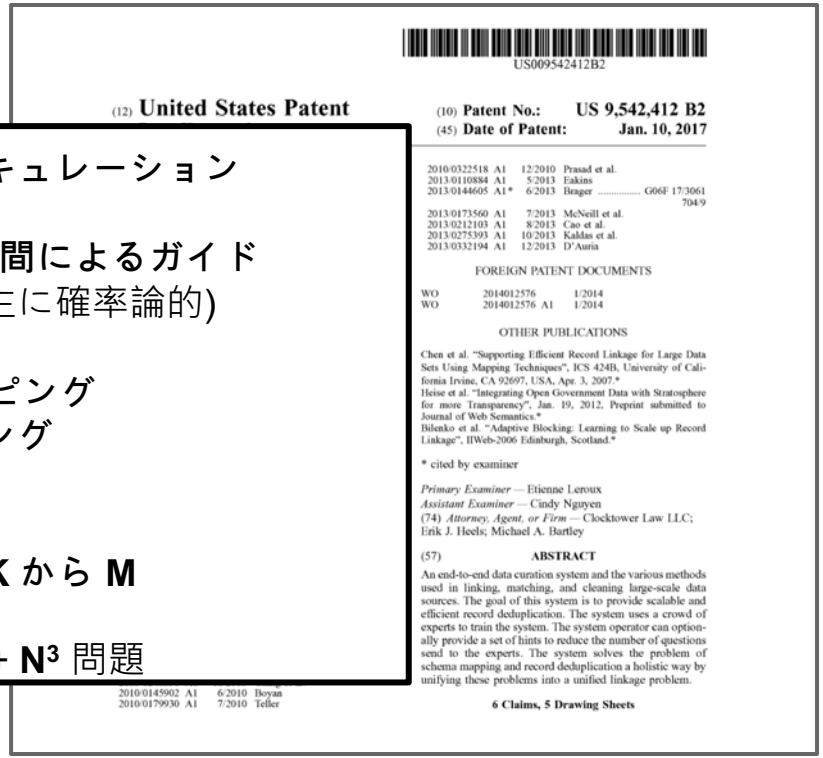
In addition, most of the previous work will not scale to the sizes of problems that we are finding in the field. For example, one web aggregator requires the curation of 80,000 URLs and a second biotech company has the problem of curating 8000 spreadsheets. At this scale, data curation cannot be a manual (human) effort, but must entail machine learning approaches with a human assist only when necessary.

This paper describes Data Tamer, an end-to-end curation system we have built at M.I.T. Brandeis, and Qatar Computing Research Institute (QCRI). It expects as input a sequence of data sources to add to a composite being constructed over time. A new source is subjected to machine learning algorithms to perform attribute identification, grouping of attributes into tables, transformation of incoming data and deduplication. When necessary, a human can be asked for guidance. Also, Data Tamer includes a data visualization component so a human can examine a data source at will and specify manual transformations.

We have run Data Tamer on three real world enterprise curation problems, and it has been shown to lower curation cost by about 90%, relative to the currently deployed production software.

tion tasks.

- **Incremental.** New data sources must be integrated incrementally as they are uncovered. There is never a notion of the integration task being finished.



US009542412B2

(12) United States Patent

(10) Patent No.: US 9,542,412 B2
(45) Date of Patent: Jan. 10, 2017

2010/0322518 A1 12/2010 Prasad et al.
2013/0110884 A1 5/2013 Eakins
2013/0144605 A1* 6/2013 Beager G06F 17/3061 704/9
2013/0173560 A1 7/2013 McNeill et al.
2013/0212103 A1 8/2013 Cao et al.
2013/0275393 A1 10/2013 Kaldas et al.
2013/0332194 A1 12/2013 D'Auria

FOREIGN PATENT DOCUMENTS

WO 2014012576 1/2014
WO 2014012576 A1 1/2014

OTHER PUBLICATIONS

Chen et al. "Supporting Efficient Record Linkage for Large Data Sets Using Mapping Techniques", ICS 424B, University of California Irvine, CA 92697, USA, Apr. 3, 2007.*
Heise et al. "Integrating Open Government Data with Stratosphere for more Transparency", Jan. 19, 2012, Preprint submitted to Journal of Web Semantics.*
Bilenko et al. "Adaptive Blocking: Learning to Scale up Record Linkage", IIWeb-2006 Edinburgh, Scotland.*

* cited by examiner

Primary Examiner — Etienne Leroux
Assistant Examiner — Cindy Nguyen
(74) *Attorney, Agent, or Firm* — Clocktower Law LLC; Erik J. Heels; Michael A. Bartley

(57) ABSTRACT

An end-to-end data curation system and the various methods used in linking, matching, and cleaning large-scale data sources. The goal of this system is to provide scalable and efficient record deduplication. The system uses a crowd of experts to train the system. The system operator can optionally provide a set of hints to reduce the number of questions send to the experts. The system solves the problem of schema mapping and record deduplication a holistic way by unifying these problems into a unified linkage problem.

6 Claims, 5 Drawing Sheets

2010/0145902 A1 6/2010 Boyan
2010/0179930 A1 7/2010 Teller

大規模表形式データのキュレーション

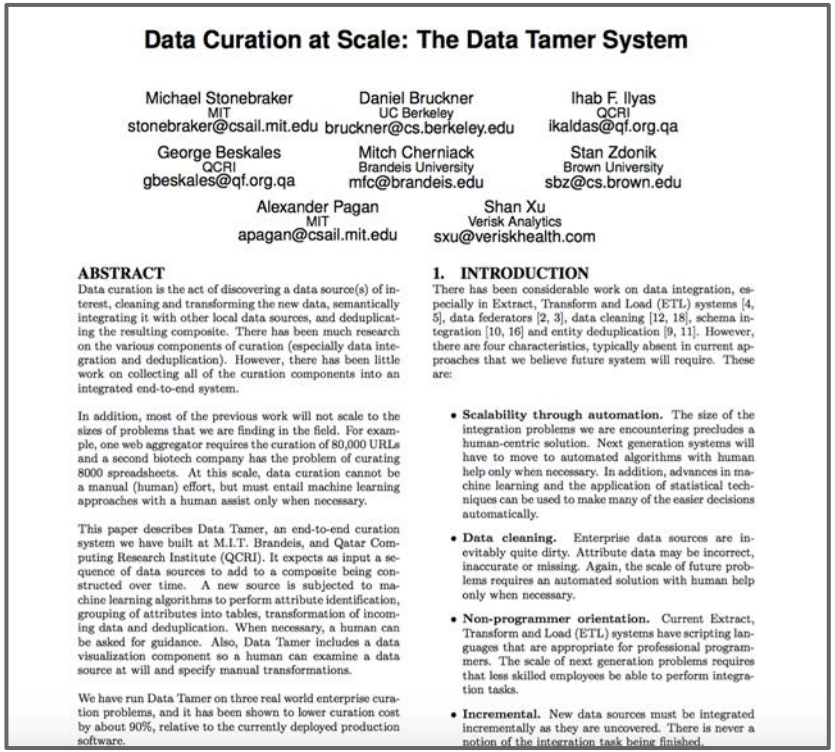
マシンによって駆動 + 人間によるガイド
(決定論的の代わりに主に確率論的)

スキーマ マッピング
記録マッチング
分類

スケール：**表が K から M**

クラシックな **N^2 + N^3** 問題

学術的プロジェクトから特許取得まで 4 年以下



Data Curation at Scale: The Data Tamer System

Michael Stonebraker
MIT
stonebraker@csail.mit.edu

Daniel Bruckner
UC Berkeley
bruckner@cs.berkeley.edu

Ihab F. Ilyas
QCRI
ikaldas@qf.org.qa

George Beskales
QCRI
gbeskales@qf.org.qa

Mitch Cherniack
Brandeis University
mfc@brandeis.edu

Stan Zdonik
Brown University
sbz@cs.brown.edu

Alexander Pagan
MIT
apagan@csail.mit.edu

Shan Xu
Verisk Analytics
sxu@veriskhealth.com

ABSTRACT

Data curation is the act of discovering a data source(s) of interest, cleaning and transforming the new data, semantically integrating it with other local data sources, and deduplicating the resulting composite. There has been much research on the various components of curation (especially data integration and deduplication). However, there has been little work on collecting all of the curation components into an integrated end-to-end system.

In addition, most of the previous work will not scale to the sizes of problems that we are finding in the field. For example, one web aggregator requires the curation of 80,000 URLs and a second biotech company has the problem of curating 8000 spreadsheets. At this scale, data curation cannot be a manual (human) effort, but must entail machine learning approaches with a human assist only when necessary.

This paper describes Data Tamer, an end-to-end curation system we have built at M.I.T. Brandeis, and Qatar Computing Research Institute (QCRI). It expects as input a sequence of data sources to add to a composite being constructed over time. A new source is subjected to machine learning algorithms to perform attribute identification, grouping of attributes into tables, transformation of incoming data and deduplication. When necessary, a human can be asked for guidance. Also, Data Tamer includes a data visualization component so a human can examine a data source at will and specify manual transformations.

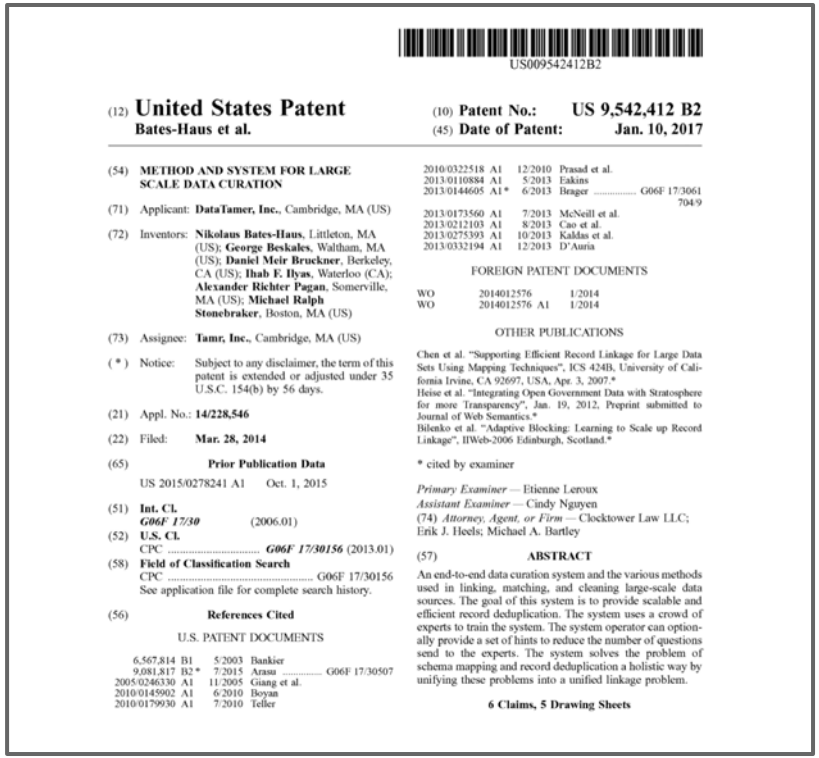
We have run Data Tamer on three real world enterprise curation problems, and it has been shown to lower curation cost by about 90%, relative to the currently deployed production software.

1. INTRODUCTION

There has been considerable work on data integration, especially in Extract, Transform and Load (ETL) systems [4, 5], data federators [2, 3], data cleaning [12, 18], schema integration [10, 16] and entity deduplication [9, 11]. However, there are four characteristics, typically absent in current approaches that we believe future system will require. These are:

- **Scalability through automation.** The size of the integration problems we are encountering precludes a human-centric solution. Next generation systems will have to move to automated algorithms with human help only when necessary. In addition, advances in machine learning and the application of statistical techniques can be used to make many of the easier decisions automatically.
- **Data cleaning.** Enterprise data sources are inevitably quite dirty. Attribute data may be incorrect, inaccurate or missing. Again, the scale of future problems requires an automated solution with human help only when necessary.
- **Non-programmer orientation.** Current Extract, Transform and Load (ETL) systems have scripting languages that are appropriate for professional programmers. The scale of next generation problems requires that less skilled employees be able to perform integration tasks.
- **Incremental.** New data sources must be integrated incrementally as they are uncovered. There is never a notion of the integration task being finished.

2013 年 CIDR 研究論文



US009542412B2

(12) **United States Patent**
Bates-Haus et al.

(10) **Patent No.: US 9,542,412 B2**
(45) **Date of Patent: Jan. 10, 2017**

(54) **METHOD AND SYSTEM FOR LARGE SCALE DATA CURATION**

(71) Applicant: **DataTamer, Inc.**, Cambridge, MA (US)

(72) Inventors: **Nikolaus Bates-Haus**, Littleton, MA (US); **George Beskales**, Waltham, MA (US); **Daniel Meir Bruckner**, Berkeley, CA (US); **Ihab F. Ilyas**, Waterloo (CA); **Alexander Richter Pagan**, Somerville, MA (US); **Michael Ralph Stonebraker**, Boston, MA (US)

(73) Assignee: **Tamr, Inc.**, Cambridge, MA (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 56 days.

(21) Appl. No.: **14/228,546**

(22) Filed: **Mar. 28, 2014**

(65) **Prior Publication Data**

US 2015/0278241 A1 Oct. 1, 2015

(51) **Int. Cl.**
G06F 17/30 (2006.01)

(52) **U.S. Cl.**
CPC ***G06F 17/30156*** (2013.01)

(58) **Field of Classification Search**
CPC G06F 17/30156
See application file for complete search history.

(56) **References Cited**

U.S. PATENT DOCUMENTS

6,567,814	B1	5/2003	Bankier	
9,081,817	B2 *	7/2015	Arasu	G06F 17/30507
2005/0246330	A1	11/2005	Giang et al.	
2010/0145902	A1	6/2010	Boyan	
2010/0179930	A1	7/2010	Teller	
2010/0322518	A1	12/2010	Prasad et al.	
2013/0110884	A1	5/2013	Eakins	
2013/0144605	A1 *	6/2013	Beager	G06F 17/3061 704/9
2013/0173560	A1	7/2013	McNeill et al.	
2013/0212103	A1	8/2013	Cao et al.	
2013/0275393	A1	10/2013	Kaldas et al.	
2013/0332194	A1	12/2013	D'Auria	

FOREIGN PATENT DOCUMENTS

WO	2014012576	1/2014
WO	2014012576 A1	1/2014

OTHER PUBLICATIONS

Chen et al. "Supporting Efficient Record Linkage for Large Data Sets Using Mapping Techniques", ICS 424B, University of California Irvine, CA 92697, USA, Apr. 3, 2007.*
Heise et al. "Integrating Open Government Data with Stratosphere for more Transparency", Jan. 19, 2012, Preprint submitted to Journal of Web Semantics.*
Bilenko et al. "Adaptive Blocking: Learning to Scale up Record Linkage", IIWeb-2006 Edinburgh, Scotland.*

* cited by examiner

Primary Examiner — Etienne Leroux
Assistant Examiner — Cindy Nguyen
(74) *Attorney, Agent, or Firm* — Clocktower Law LLC; Erik J. Heels; Michael A. Bartley

(57) **ABSTRACT**

An end-to-end data curation system and the various methods used in linking, matching, and cleaning large-scale data sources. The goal of this system is to provide scalable and efficient record deduplication. The system uses a crowd of experts to train the system. The system operator can optionally provide a set of hints to reduce the number of questions send to the experts. The system solves the problem of schema mapping and record deduplication a holistic way by unifying these problems into a unified linkage problem.

6 Claims, 5 Drawing Sheets

2017 年特許
大規模データ精選のための方法及びシステム