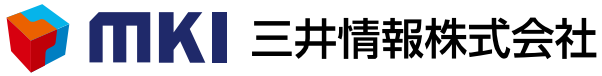
2018/03/09

企業内データ分析におけるGraphデータベース活用

三井情報株式会社
デジタルトランスフォーメーションセンター R&D部 研究開発室

荒木一馬



本日の内容

1. MKIのソリューションと研究開発の取り組み紹介
2. Graph分析機能の範疇と期待効果
3. 業務データ分析における課題
4. 取り扱いデータとユースケースごとの課題
5. 直近の研究開発の結果
6. まとめ

©2018 MITSUI KNOWLEDGE INDUSTRY CO., LTD. All rights reserved.

事業概要

デジタル時代の「攻めのIT」へ。

MKIはITのスペシャリストとして、
お客様の企業戦略に沿ったIT活用だけではなく、
中長期的な企業価値の向上や競争力強化に結びつく
ビジネスモデルの変革を推進するＩＴ活用 に向けて
パートナー企業と共にお客様をご支援していきます。

コンサルティングから、アプリケーション開発、クラウド等のサービスやネットワークインフラ、運用・保守までトータルサービスをご提供

©2018 MITSUI KNOWLEDGE INDUSTRY CO., LTD. All rights reserved.

三井情報株式会社について

事業概要（事業領域）①

コンサルティング

【主なソリューション】

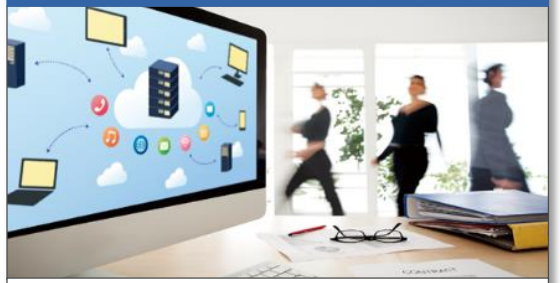
- ITマネジメントサービス
- IT活用コンサルティング

アプリケーション開発

【主なソリューション】

- 基幹システム
  SAP ERP、SAP S/4 HANA
- 商社･卸売業･メーカ販社向けテンプレート
  「MKI-Trade Suite」
- 金融業向けシステム
  「財務分析システム：CASTER」

サービスインテグレーション



【主なソリューション】

- データ分析プラットフォーム
  「Black Swan」
- 需要予測/自動発注ソリューション
- クラウド型省エネルギーマネジメントサービス「GeM2」
- サブスクリプションビジネス支援プラットフォーム「Zuora」
- 電子署名/デジタル･トランザクション・マネジメント「DocuSign」

©2018 MITSUI KNOWLEDGE INDUSTRY CO., LTD. All rights reserved.

三井情報株式会社について

事業概要（事業領域）②

インフラソリューション

【主なソリューション】

- クラウド型無線LANサービス「MKI マネージド Wi-Fi」
- コミュニケーションツール（Cisco、Microsoft）
- コンタクトセンターソリューション

運用・保守サービス

【主なソリューション】

- 統合監視
- マネージドサービス
- アウトソーシングサービス

その他

【主なソリューション】

- がん遺伝子解析サービス「OncoPrime」
- 園とご家庭をつなぐ総合連絡ツール「きっずノート」

さまざまな業界に、最適なソリューションをご提供します

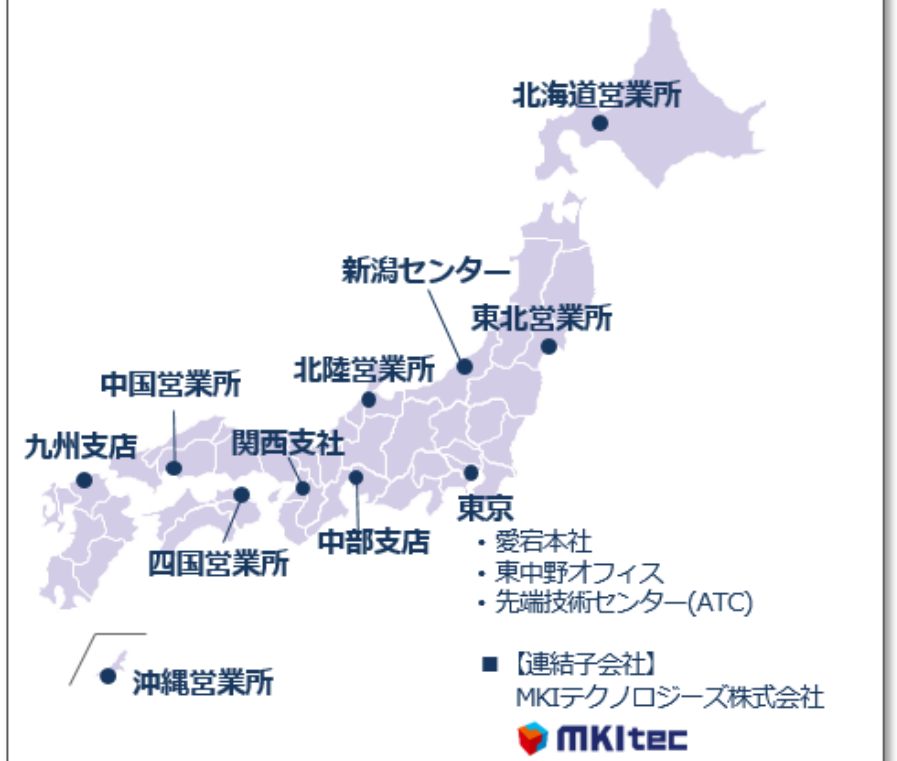
商社　流通　不動産　金融　通信　公共　教育　ヘルスケア

©2018 MITSUI KNOWLEDGE INDUSTRY CO., LTD. All rights reserved.

三井情報株式会社について

MKIグループのネットワーク

国内11拠点



・愛宕本社
・東中野オフィス
・先端技術センター(ATC)

■【連結子会社】
MKIテクノロジーズ株式会社
MKItec

海外6拠点

MKIのグローバル拠点は、世界に視野を広げたICT技術調査と情報発信を行うアンテナ拠点として、革新的なICTソリューションをお客様へご提供しています。

また、米国三井物産、欧州三井物産のシステム運用も行っています。

©2018 MITSUI KNOWLEDGE INDUSTRY CO., LTD. All rights reserved.

三井情報株式会社について

MKIの分析分野への取り組み

データ分析プラットフォーム

英国Black Swan Data Ltd社が開発した「Nest」©は、SNS等の社外データや社内の売上や広告費データを取り込み最適な分析アルゴリズムを用いて将来の売上、来場者の予測、新規商品の開発、広告予算の影響度分析による効果的な広告費投下の支援等、ビジネスを幅広くサポートする分析プラットフォームです。

▸ データ分析プラットフォーム

Black Swan

分析予測ソリューション

データ分析に関わる専門知識がない・・・。分析に時間がかかる・・・。
そんなお悩みはございませんか？
三井情報の分析予測ソリューションなら、専門知識やノウハウの有無に関わらず、高度な時系列データを分析・予測し、新たな将来をサポートします。

▸ 分析予測ソリューション

データ分析ソリューション

個別に持っている重要な取引先情報をつないでリスク管理をしたい
社内にある知財を新たなビジネスの創出に活用し競争力を強化したい

三井情報のデータ分析ソリューション～Graph Engine活用サービス～は、「会社」「人」「情報」などのあらゆるつながりを可視化。新たな発見をサポートします。

▸ データ分析ソリューション

需要予測ソリューション

▸ 需要予測ソリューション

http://www.mki.co.jp/biz/solution/analytics/index.html

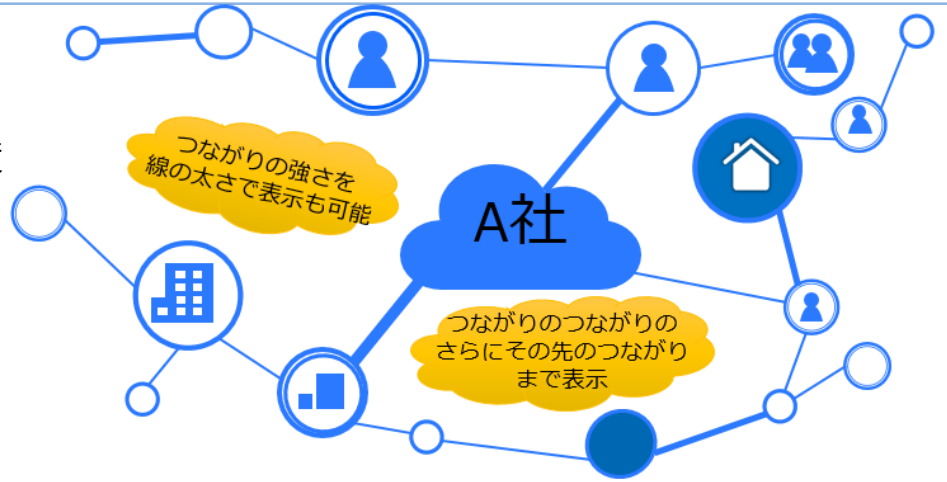
http://dtc.mki.co.jp/solution/

©2018 MITSUI KNOWLEDGE INDUSTRY CO., LTD. All rights reserved.

Graphへの取り組み紹介（１）

Graph分析エンジンの活用により下のような問題解決を支援

- グループ会社の債権管理
- サプライチェーンの可視化と把握
- 会社間関係の可視化と倒産による影響の可視化
- 人・会社・製品などの関係性による高度なドキュメント検索



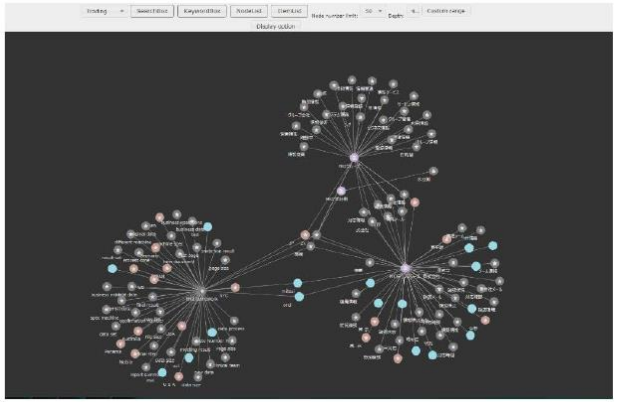
どんなデータを分析できるのか

- メール、予定表（の宛先関係およびテキスト）
- SNS
- 取引トランザクション
- ドキュメント・Webページ（が含むテキスト）

グループ経営を行う企業
↓
グループ会社の債権管理

製造業／流通業
↓
サプライチェーンの把握による突発的なトラブルや急激な需要変動に対応

金融業
↓
企業の倒産予知や連鎖倒産を把握



システムを横断した検索が可能

各システムのデータを集約し、共通の情報で繋げる

情報の繋がりを可視化することで、新たな気付きをもたらすことができる

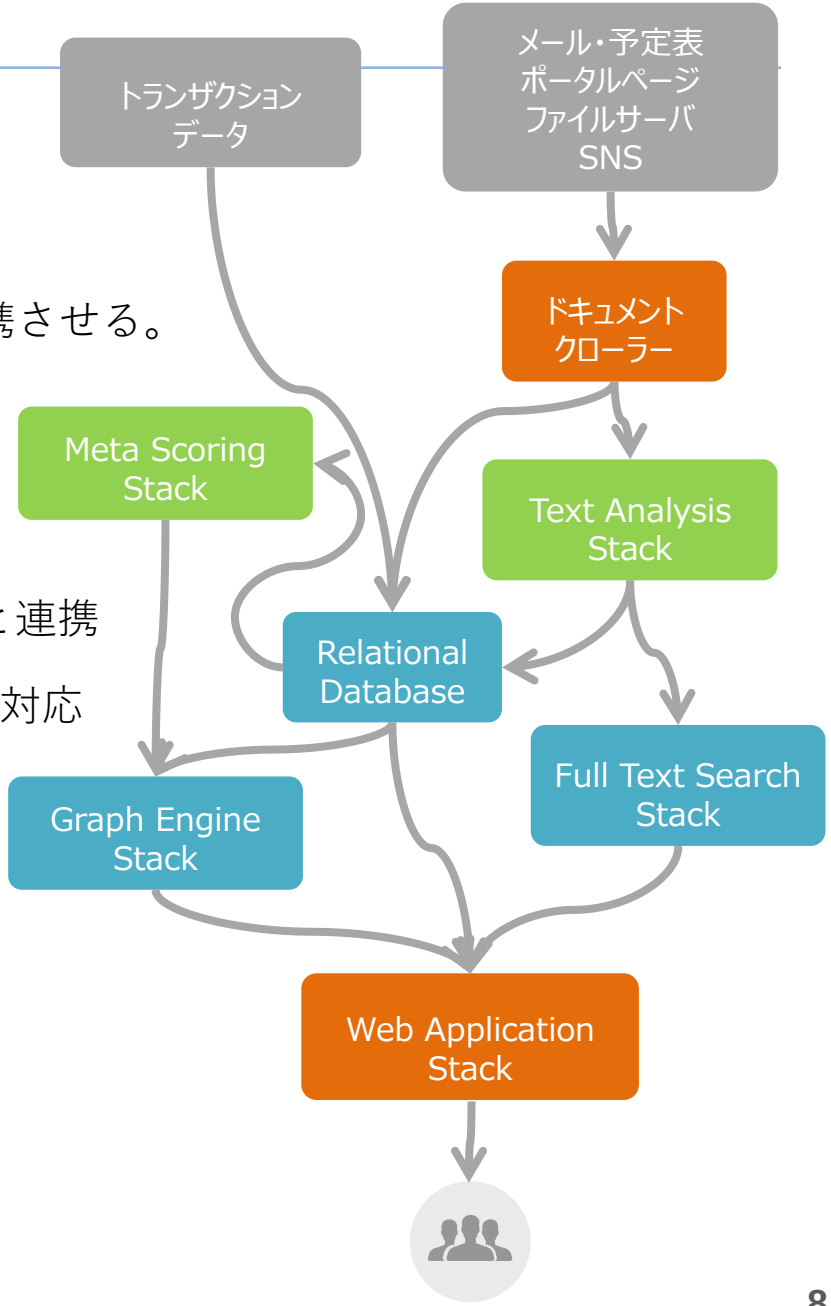
http://www.mki.co.jp/biz/solution/analytics/graph.html
https://www.youtube.com/watch?v=k3diFLtkaz0

©2018 MITSUI KNOWLEDGE INDUSTRY CO., LTD. All rights reserved.

Graphへの取り組み紹介（２）

- どのような構成なのか
  - 右の図の通り
  - データストアとしてはRDBMS、Graph Engine、Full-text Searchの3つを連携させる。

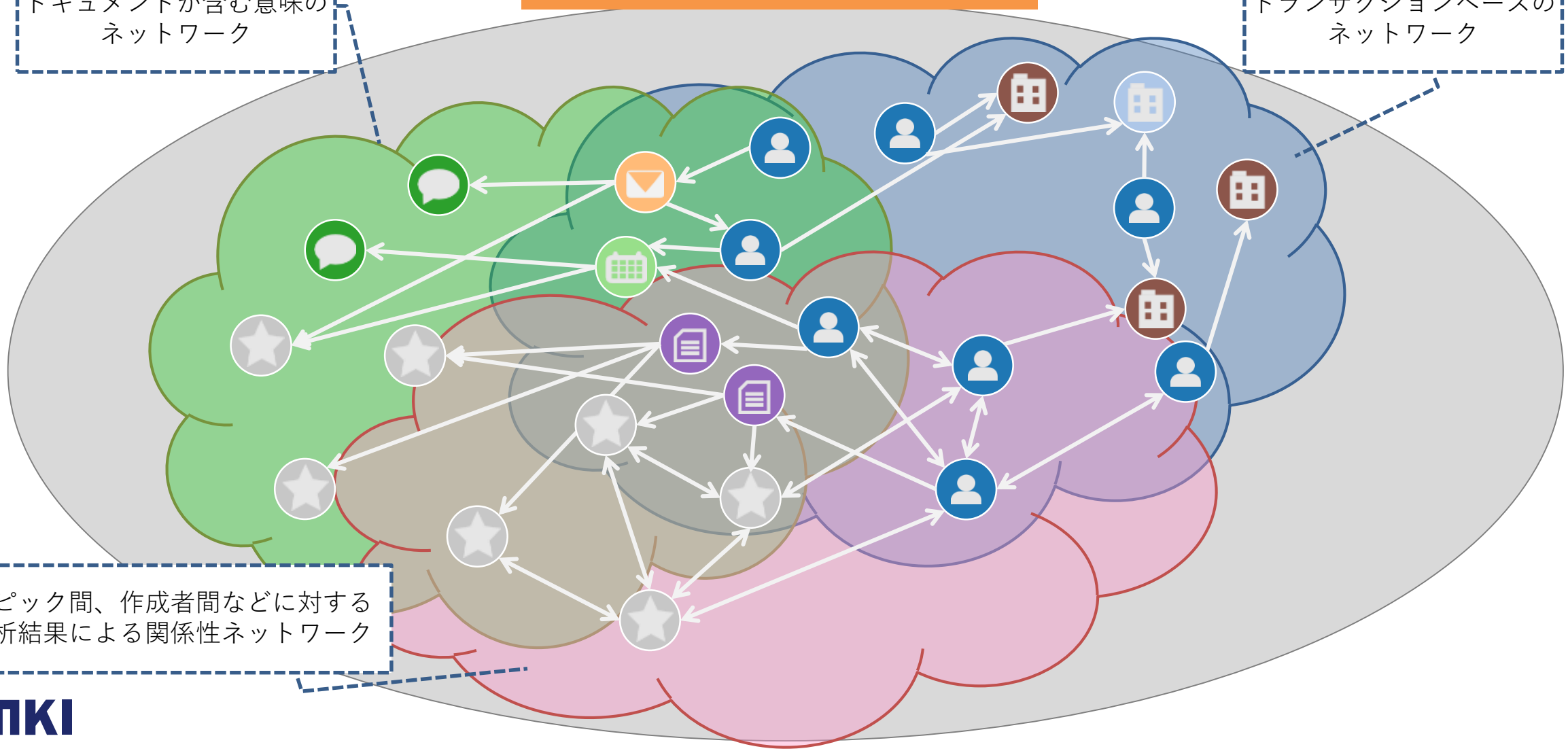
- どこが特徴的なのか
  - 複数のデータソース、データ種別を定期的・自動的に収集
  - テキスト分析と連携することで、ドキュメント内の単語間の関係性をGraphと連携
  - 主データストアをRDBMSとして持つことで、Graph分析外の処理にも高速に対応
  - Graph分析へは、トランザクションや宛先・作成者など直接的な関係と、
    別の機械学習の分析結果による間接的なスコアの両方を投入
  - Graph分析が生かせる処理のみをGraph Engineに任せることで、
    高速かつ柔軟に分析目的に対応



©2018 MITSUI KNOWLEDGE INDUSTRY CO., LTD. All rights reserved.

Graphへの取り組み紹介（３）

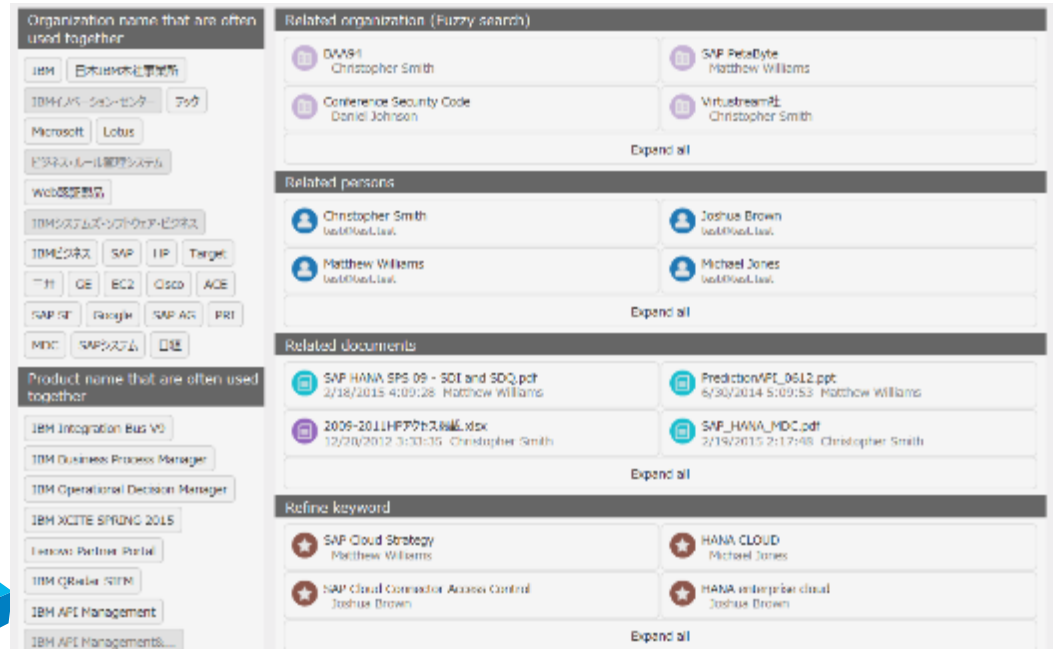
Graphネットワーク内スキーマ
（イメージ図）

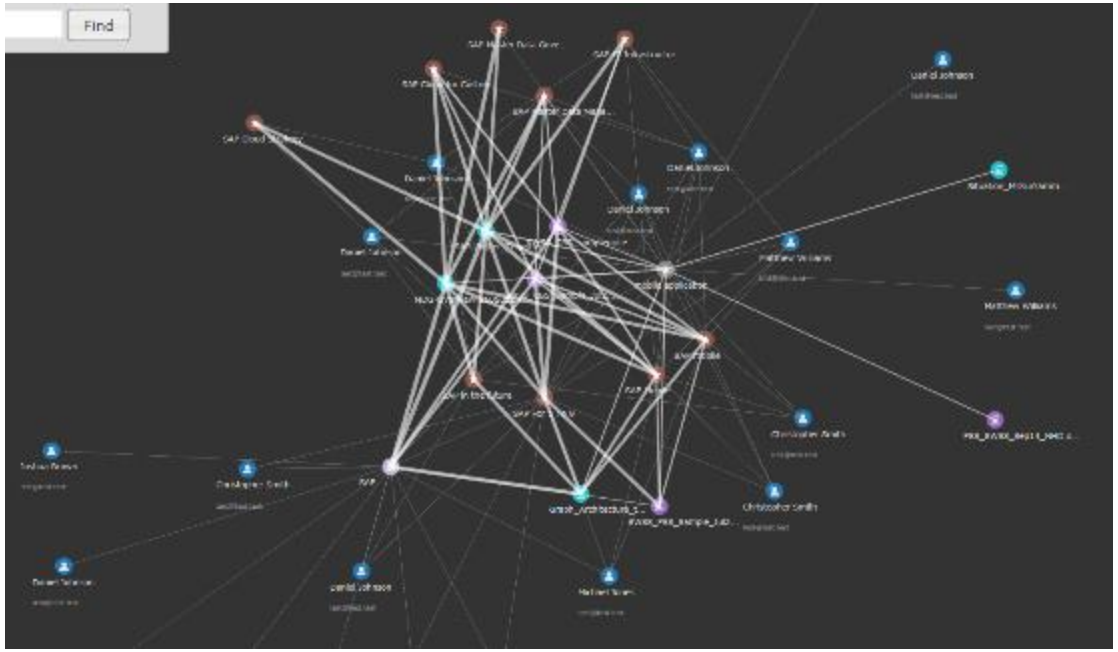


©2018 MITSUI KNOWLEDGE INDUSTRY CO., LTD. All rights reserved.

Graphへの取り組み紹介（４）

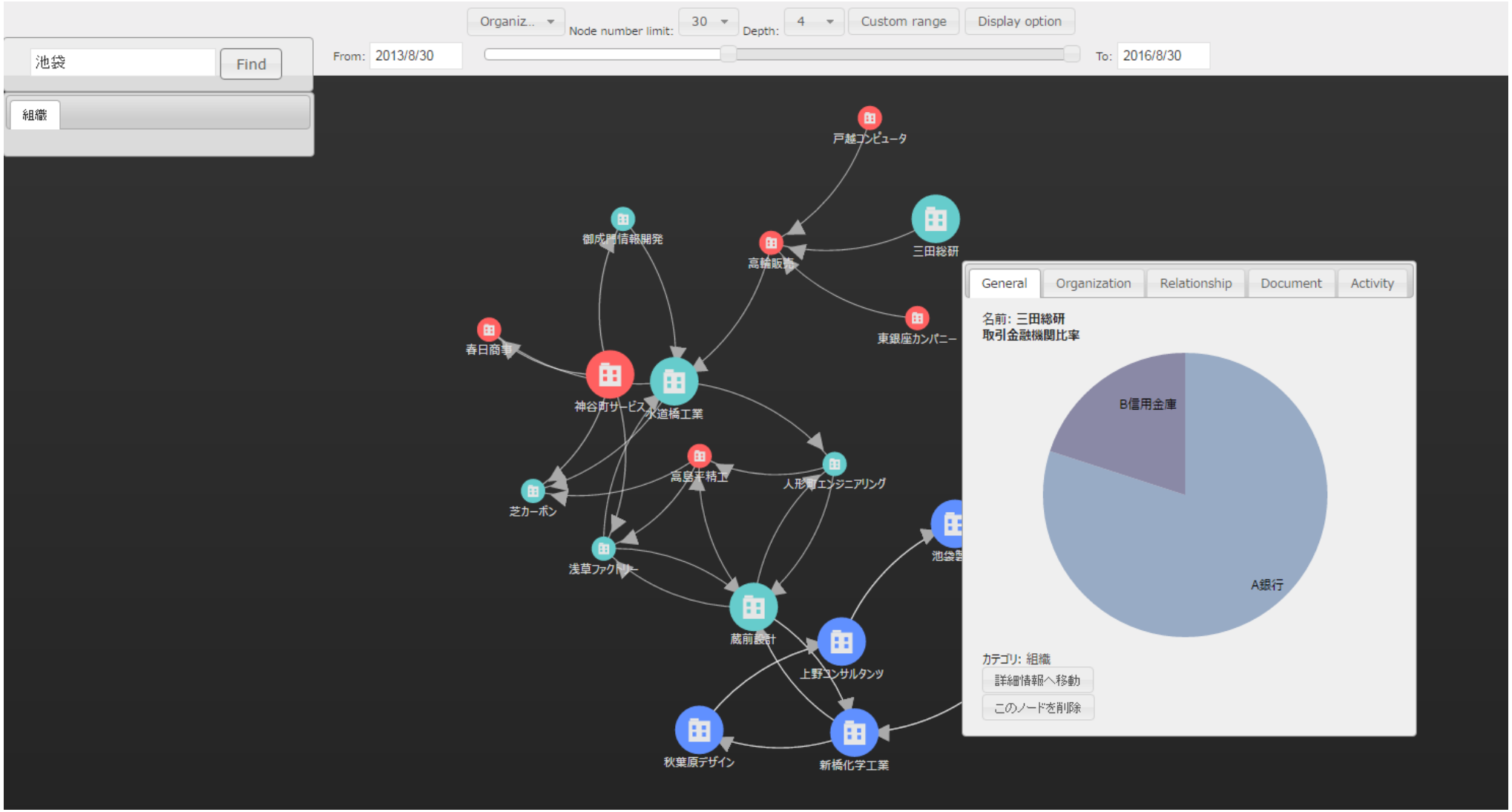
- 何ができるのか
  - 取引関係の可視化
  - 関係性の高い会社の検出
  - 会社グループが共通して持つ特徴を検出
  - 循環取引の検出
  - 人間関係の可視化
  - 暗黙的なグループの検出
  - 会社や技術に詳しい人間を検出
  - 個性のサマリ化
  - ファジー検索、類似文書の検出　など





©2018 MITSUI KNOWLEDGE INDUSTRY CO., LTD. All rights reserved.

Graphへの取り組み紹介（５）



©2018 MITSUI KNOWLEDGE INDUSTRY CO., LTD. All rights reserved.

Graph分析機能の範疇と期待効果（１）

- Graph分析、およびGraphデータベースと呼ばれるものの特徴

#	機能概要	機能詳細
1	Graph独特のデータ保持	- Distributed Graphs（分散グラフ） - Property Graph（属性グラフ） - Nested Properties（階層属性グラフ）
2	Graph的クエリ	- Graph Query language（Gremlin / SPARQL / GRAPHQL / Cypher） - Complex graph traversals（グラフ横断検索） - Full blown Traversals
3	Graph特有の分析アルゴリズム	Centrality： PageRank Betweenness Centrality Closeness Centrality Partitioning: Label Propagation (Weakly) Connected Components Strongly Connected Components Union-Find Path-Finding: Shortest Path Minimum Weight Spanning Tree All Pairs – and Single Source – Shortest Path Multi-Source, Breadth-First Search
4	NoSQL的特徴	分散処理、柔軟なデータ型、etc…

©2018 MITSUI KNOWLEDGE INDUSTRY CO., LTD. All rights reserved.

Graph分析機能の範疇と期待効果（２）

- なにを期待するのか

#	機能概要	機能詳細
1	Graph独特のデータ保持	- Distributed Graphs（分散グラフ） - Property Graph（属性グラフ） - Nested Properties（階層属性グラフ）
2	Graph的クエリ	- Graph Query language (Gremlin, Cypher, PGQL, GraphQL) - Complex graph traversals（グラフ横断検索） Mind blown Traversals
3	Graph特有の分析アルゴリズム	Centrality： PageRank Betweenness Centrality Closeness Centrality Partitioning: Label Propagation (Weakly) Connected Components Strongly Connected Components Union-Find Path-Finding: Shortest Path Minimum Weight Spanning Tree All Pairs – and Single Source – Shortest Path Multi-Source – Breadth-First Search
4	NoSQL的特徴	分散処理、柔軟なスキーマ、etc.

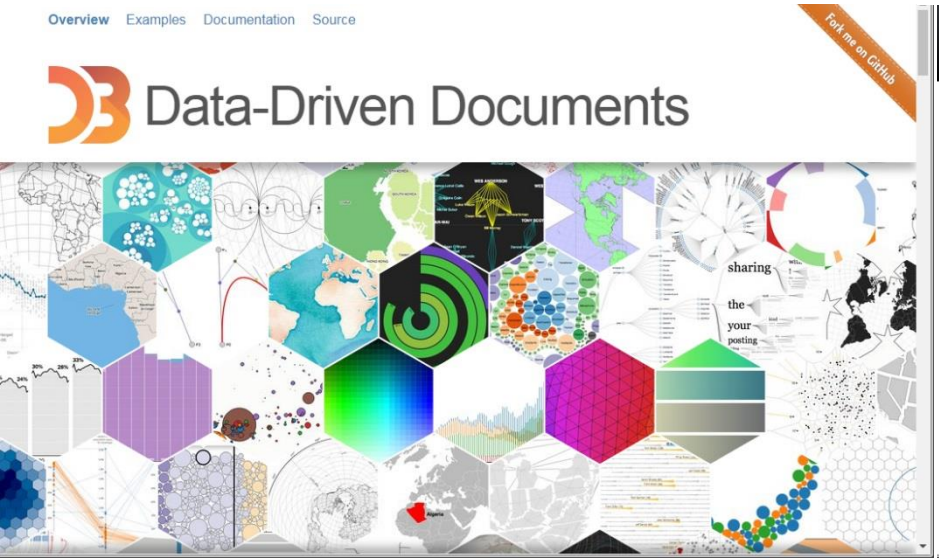
- （SQLだと）JOINの連鎖になるようなクエリ
  - ➢ ループ処理の記述が容易であり、再帰的な構造の存在する関係データに有効
- 直感的なクエリ
  - ➢ クエリ記法がデータに対し直感的であり、開発効率・分析効率が高い
- 各ノードの重みの算出
- ループ構造の検出
- ノード間の最短ルートの検出
- 指定条件によるパーティション化（クラスタリング）

©2018 MITSUI KNOWLEDGE INDUSTRY CO., LTD. All rights reserved.

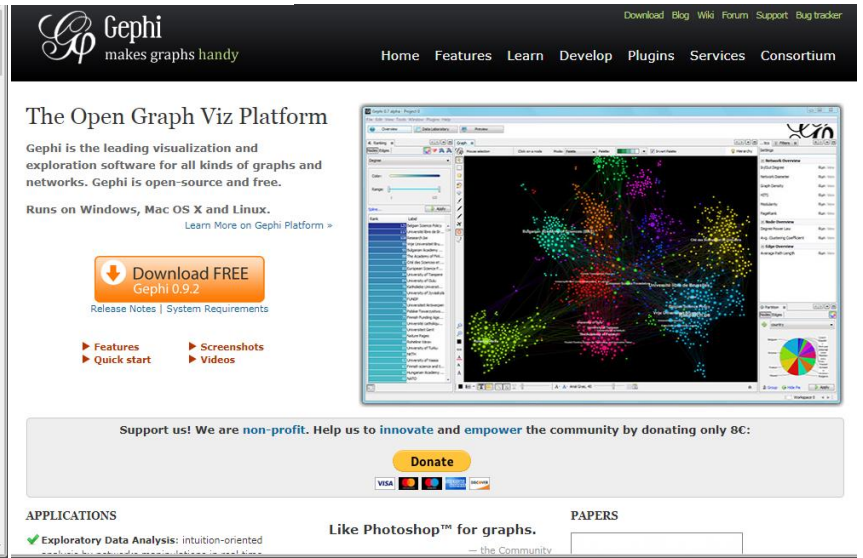
Graph分析機能の範疇と期待効果（３）

- Graphから連想される、ネットワーク構造の図
  - ➢ 分析のためのGraph処理とは直接関係しない。
  - – 経路を含む結果をすべて取得し、UI側で再度関連付けて描画する
- 分析結果ではなく人間が分析する材料（＝可視化）がほしい場合、こちらの方が重要な場合も
  - – 分析結果と根拠が必要な場合　→　最も○○なパス経路を取得して描画
  - – ファジィに考える材料がほしい場合　→　○○の周辺のノード/エッジを良しなに取得して全描画

https://d3js.org/



https://gephi.org/



Alchemy.js
http://graphalchemist.github.io/Alchemy/

Linkurious
https://linkurio.us/

KeyLines
https://cambridge-intelligence.com/keylines/

Cytoscape
http://www.cytoscape.org/

etc

パブリッククラウドPaaSによるGraph Databaseと想定シナリオ

Azure CosmosDB Graph API

- ソーシャルネットワーク
  - 行動予測、嗜好の追跡
- レコメンデーションエンジン
  - 小売向け商品提案
- GeoSpatial
  - 通信、物流、旅行分野での最短ルート算出
- IoT
  - デバイス間の接続関係、ネットワーク構造の把握、影響範囲の特定

https://docs.microsoft.com/ja-jp/azure/cosmos-db/graph-introduction

Amazon Neptune

- ソーシャルネットワーク
  - 個人に提供する情報の優先順位を最適化
- レコメンデーションエンジン
  - 小売向け商品提案
- 不正検出
  - 購入取引の不正パターン検出
- 知識グラフ
  - 索引・辞書型の情報のモデル化
- ライフサイエンス
  - 分子構造パターンの探索、研究出版物の整理
- ネットワーク/IT運用
  - 稼働状況の確認、影響範囲の特定

https://aws.amazon.com/jp/neptune/

IBM Graph

- ソーシャルネットワーク
- レコメンデーションエンジン
- 不正検出
- 大スケールの関連呼び出しクエリ

https://www.ibm.com/analytics/jp/ja/technology/cloud-data-services/graph/

©2018 MITSUI KNOWLEDGE INDUSTRY CO., LTD. All rights reserved.

Apache Tinkerpop & Gremlin

- Graph処理フレームワークであるTinkerPopと、Graphトラバーサル言語であるGremlin
- Tinkerpop対応のGraphデータベース / Graph分析基盤であれば共通してGremlin言語でクエリが書ける
  - OLTP（リアルタイム処理）
    - JanusGraph / Titan
    - Neo4j
    - OrientDB
    - 等々
  - OLAP（バッチ処理、実行時間が長く分散処理したいもの）
    - Apache Giraph
    - Apahce Spark
- ただし、データベース管理系のコマンドや、スキーマ定義・インデックス周りは各ソフトウェア独自である
- 特にインデックスはパフォーマンスにダイレクトに影響することに注意

http://tinkerpop.apache.org/providers.html
http://tinkerpop.apache.org/gremlin.html

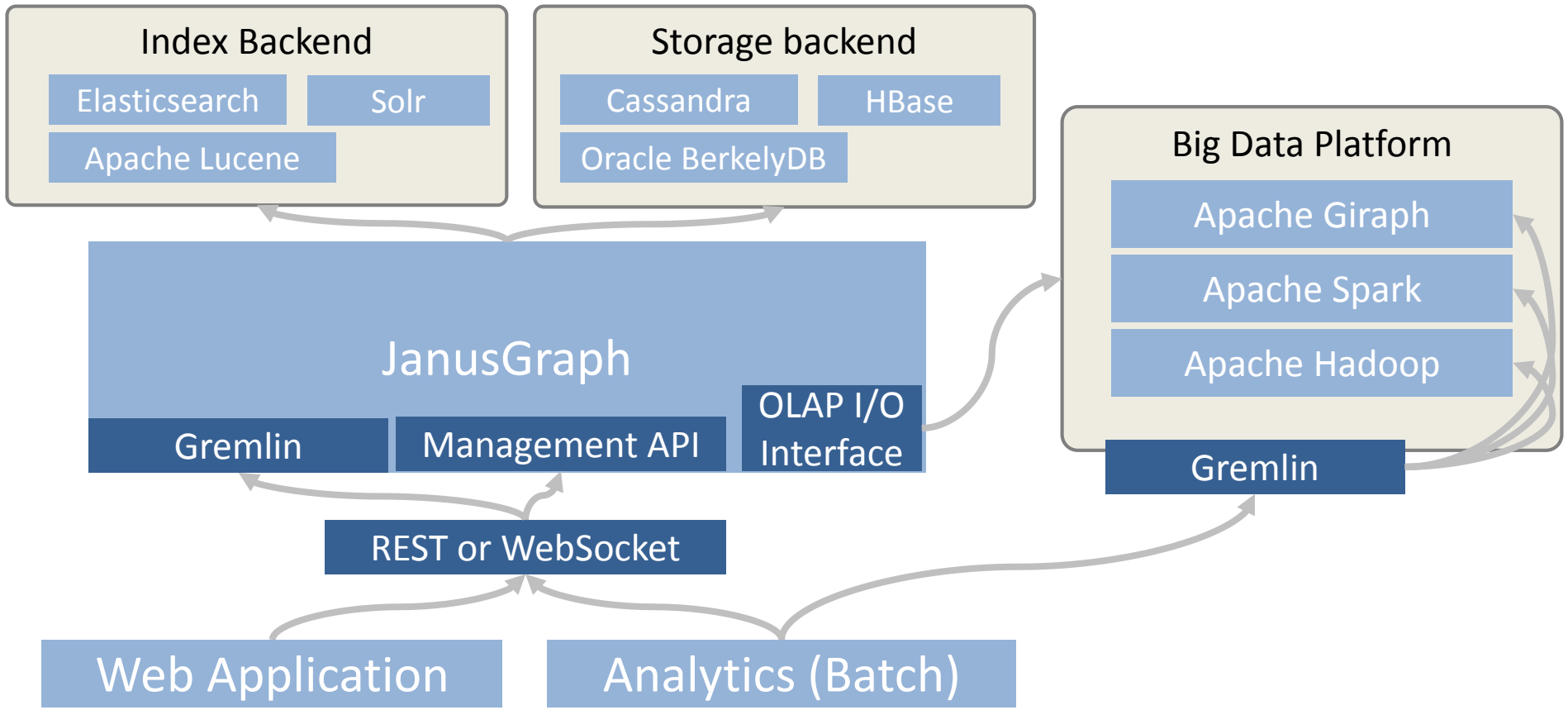
©2018 MITSUI KNOWLEDGE INDUSTRY CO., LTD. All rights reserved.

モダンなGraph分析環境の構成（JanusGraphでの例）

JanusGraphとは　…Linux Foundationによるグラフデータベースプロジェクト。Titan DBのfork。

- http://docs.janusgraph.org/latest/arch-overview.html

- プロパティグラフ。Apache TinkerPop対応
- ストレージエンジン、インデックスエンジンとの連携
- 分散処理、可用性
- Apache 2 License



©2018 MITSUI KNOWLEDGE INDUSTRY CO., LTD. All rights reserved.

業務データ分析への適用における課題

1. データ格納先としての位置付け

✓ グラフ基盤を：分析結果算出のための環境（メモリ上のみ）として使うか／永続的格納先として使うか

- Graphへの初期データロードに大量に時間をかけたくない／ソース側に負荷をかけたくない
- Graph基盤のみではシングルソースとしての能力が足りない

（トランザクション、レスポンス速度、インデックス機能、権限管理、LIKE/正規表現　等々）

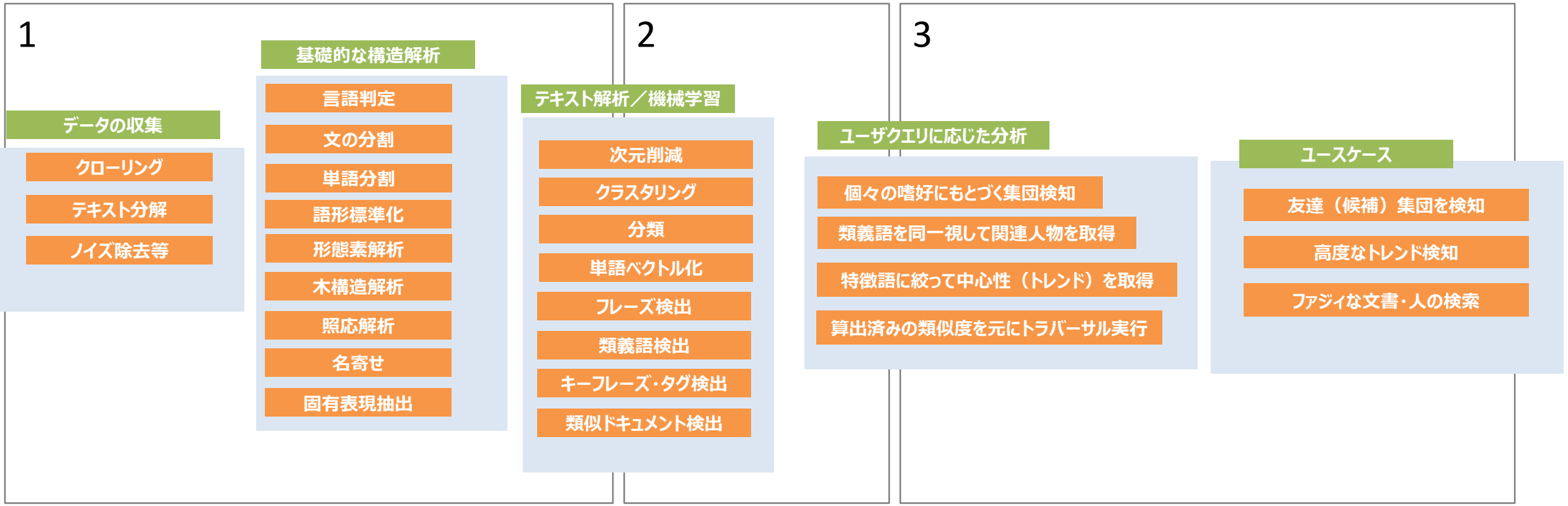
✓ データの格納先を：　グラフ基盤に統一するか／他の格納先（RDBMS等）も並行して使うか

- データの二重管理は、整合性の観点からも物理的な領域（コスト）の観点からも避けたい

©2018 MITSUI KNOWLEDGE INDUSTRY CO., LTD. All rights reserved.

業務データ分析への適用における課題

2. Graph内で分析すべきこと、Graph外で分析すべきことの区分け
   - 例えばテキストの分析結果から個々人の嗜好を分析する場合、以下のような工程がある。
   - このうち　①Graphでできない部分　②Graphでもできる部分　③Graphでしかできない部分があるうち、②をどちらに任せるか



©2018 MITSUI KNOWLEDGE INDUSTRY CO., LTD. All rights reserved.

業務データ分析への適用における課題

3. 分析結果がほしいのか　／　考える材料がほしいのか

- 分析結果のみ出力すれば良いパターン
  - 出力がグラフ構造である必要はなく、結果の一覧があればよい　（ノード一覧、最も一致するパス　等）
  - 固定されたシナリオ、固定されたクエリ

  ⇒ グラフエンジン側の処理の比重が大きい
- Graph構造を意思決定の補助材料として使いたいパターン
  - アイデアの根拠としたい　⇒　なにが取り出すべきNodeなのかわかっていない　⇒　クエリが定められない
  - 関連しそうなノードをファジィに一括取得し、UI側でグラフ構造を再現する。
  - 動的・かつ複合的に呼び出されるクエリ

  ⇒UI側の処理が効果に大きく影響し、グラフエンジン側の性能引き出しも、UI側で動的に構築するクエリ次第

©2018 MITSUI KNOWLEDGE INDUSTRY CO., LTD. All rights reserved.

取り扱いデータとユースケースごとの課題（1）

- Graphクエリにおけるテキストデータの取り扱い
  - LIKE検索、正規表現、（高速な取得のための）テキストインデックス化　：　Gremlin仕様標準では機能が不足
- 空間情報（位置情報）によるフィルタリング
  - Polygonによる範囲限定によるフィルタリング　：　Gremlin仕様標準では機能が不足
- レコメンデーションエンジンとしての活用方法
  - Graphの頻出ユースケースとして現れるレコメンデーション方法は協調フィルタリング方式である。
    コンテンツベース方式でレコメンドしたい場合、前述のようなNLPにもとづき関連性算出が必要である。
    - ドキュメントごとにベクトル化しコサイン類似等で関連性を算出
    - トピック検出と類似度の算出により関連性を算出
    - タグ推定により個々のドキュメントとタグ間の関連性を推定

©2018 MITSUI KNOWLEDGE INDUSTRY CO., LTD. All rights reserved.

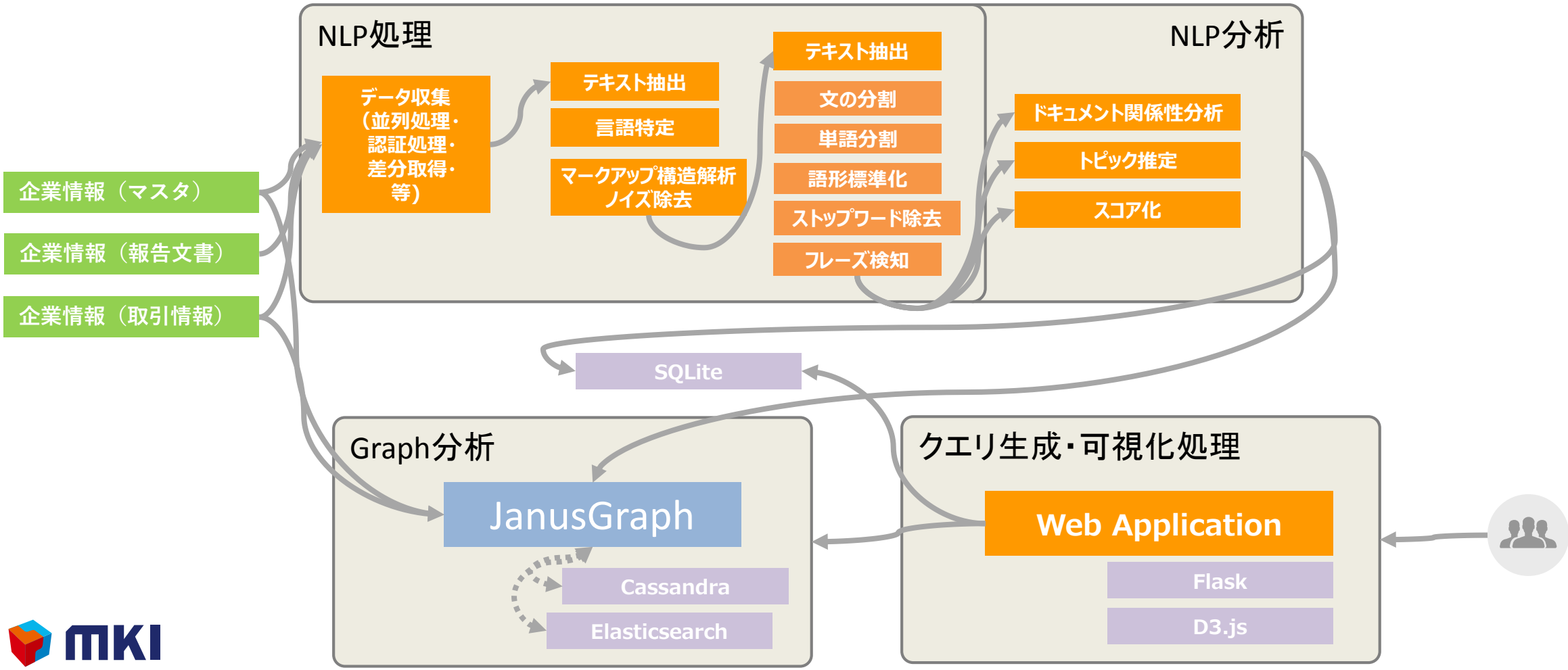
取り扱いデータとユースケースごとの課題（２）

- 他の機械学習結果との連携
  - Graph (Gremlin, Graph Database等）そのものには、PythonやRのような汎用的・豊富な機械学習環境があるわけではない。
  - 他の分析結果をNode/Edgeの根拠としたい場合、またはGraph分析の結果を他の機械学習の根拠にする場合、それらの区分けを明確に定める必要がある。
- 時系列データの格納方法
  - 同じノードの持つ値が時系列で変化する場合、その値をどう持つか
  - トランザクションを記録するか、各時点での情報を記録するか
  - 後者の場合Graphでは工夫が必要になる。
- 権限情報を各データに付与したい場合
  - 特定の分析者による分析用途ではなく、個別にアクセスが制限されるべきユーザが分析用クエリを呼び出す場合

©2018 MITSUI KNOWLEDGE INDUSTRY CO., LTD. All rights reserved.

直近の研究開発の結果

- Graph処理基盤にJanusGraphを使い、分析手法の拡張を検討中。
  - 現行バージョンのMKI Graphソリューションでよりも効率的にクエリを処理する構成・シナリオ



©2018 MITSUI KNOWLEDGE INDUSTRY CO., LTD. All rights reserved.

動作イメージ

DEMO

©2018 MITSUI KNOWLEDGE INDUSTRY CO., LTD. All rights reserved.

まとめ

- Graph分析の導入は非常に容易である
- Graphは万能の分析基盤ではない
- Graphは万能のストレージ基盤ではなく、またRDBMSや他のNoSQL代替ではない

導入を考える場合、下の項目を検討すべき

- ✓ なにを入れるのか
- ✓ どの範囲までGraphで分析するのか
- ✓ 誰が使うのか（分析者、一般ユーザ）

  分析者が分析結果を得るために使うのか、それとも（ファジィ検索など）サービスの一部として使うのか
- ✓ 結果がほしいのか、考える材料がほしいのか
- ✓ Graph単体での分析で要件は満たせているのか。なにと組み合わせれば効果が出るのか

©2018 MITSUI KNOWLEDGE INDUSTRY CO., LTD. All rights reserved.