INTHEFOREST会社紹介

～超大規模データ・並列分散処理に強みのあるデータエンジニアリングカンパニー～

株式会社INTHEFOREST

Mail: sales@intheforest.co.jp / Tel:03-5848-2424
〒176-0023 東京都練馬区中村北1-13-13 OHD練馬ビル502

INTHEFOREST社紹介

～超大規模データ・並列分散処理に
強みのあるデータエンジニアリングカンパニー～

会社紹介

会　社　名	株式会社 INTHEFOREST (インザフォレスト)
設　　立	平成23年1月
所　在　地	東京都練馬区中村北1-13-13 OHD練馬ビル 5F
T　E　L	03-5848-2424
U　R　L	https://www.intheforest.co.jp/
代表取締役	冨田 和孝

主要お取引先様

国立研究開発法人 理化学研究所様、コネクト株式会社様、株式会社Two Five様、伊藤忠テクノソリューションズ株式会社様、ヤフー株式会社様、株式会社ワークスアプリケーションズ様、楽天株式会社様、カルチュア・コンビニエンス・クラブ様、KSKアナリティクス様、ユニアデックス様

事業内容

- Webサービス・データベースなど分散処理OSSを中心に超大規模データ処理基盤の構築・運用・保守コンサルティング
- 分散DB Apache Cassandra商用サポート
- 機械学習・自然言語処理技術を用いたデータ分析コンサルティング

経営陣紹介（ITアーキテクト・エンジニアリング担当）

冨田 和孝 Tomita Kazutaka

・株式会社INTHEFOREST代表取締役

・日本Cassandraコミュニティーメンバー

・DBエンジニア・システムアーキテクト

ぐるなび、シンプレックス（FX）、ISPなどでDB中心としたシステム構築・運用を担当。
ユーザ数千万人規模のデータ処理基盤作りに強み。並列分散処理・リソース管理OSSなども
ソースコードレベルで理解し、サポートや修正パッチ作成も可能。
分散DB Apache Cassandraでは業界の第一人者。

エンジニアチーム紹介

有本 絵麻

フロントサイドエンジニア

企業内研究機関にて研究員向けデータ分析環境の開発及び分析業務支援に従事。IFでは日本語解析システム、行動履歴取得システム、など多数システムの構築に従事

村岡 志保

基盤エンジニア
データベースエンジニア

分散DB Cassandraスペシャリスト、OSSを組み合わせたデータ処理環境構築に強み。
IFではレコメンデーション基盤など、ビッグデータ保有の企業向けコンサルティングに多数従事

川浪拓也Timothy

基盤エンジニア
データベースエンジニア

データスペシャリスト。
キャリア向けシステムにおけるデータ解析処理等。
Python-Djangoを用いたWebシステムにも精通

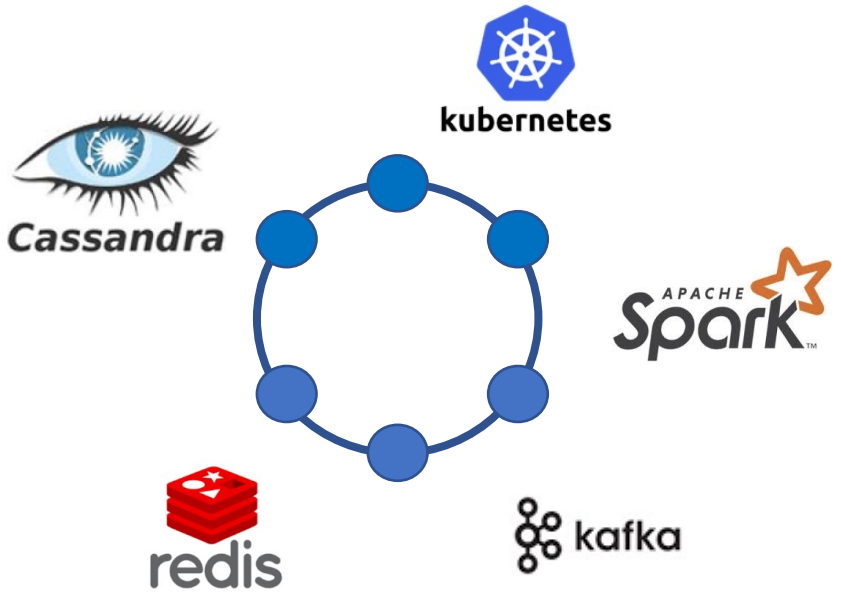
ミハイ・シュテウ

アナリティクスアドバイザー

インペリアルカレッジ・ロンドン博士課程在籍※
研究領域：Machine Learning/Deep Learning on Time Series

※ The Times Higher Education World University Rankings 世界総合８位 2018年

スキル・バックグラウンド

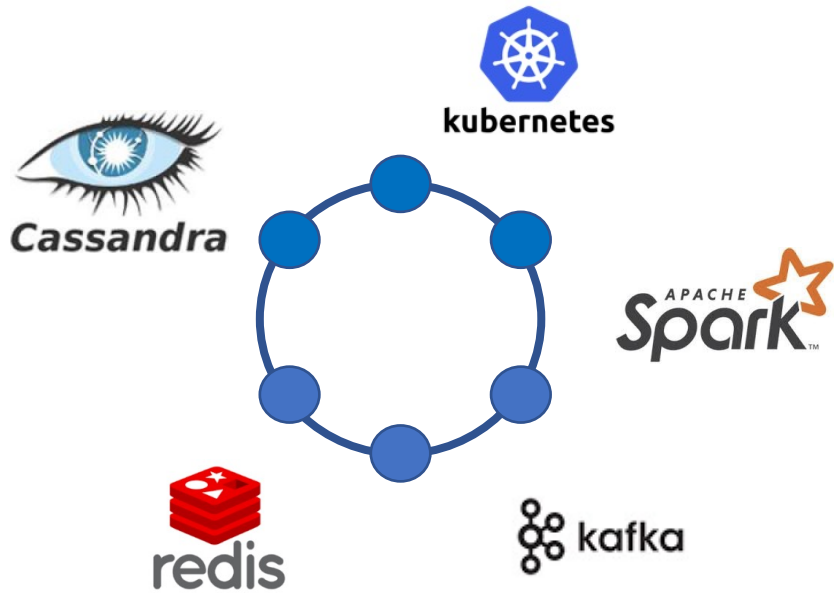


- 超大規模Web・基幹システムの構築経験
- 上場企業・金融監査に耐えうるシステム内部統制対応

- Cassandraはじめとした分散処理・リソース管理OSSの世界レベルのエンジニアリング能力
- ITアーキテクトとしてインフラ～ミドルウェア～フロントエンドまで一気通貫した総合的知見

弊社訴求ポイント：超大規模システム構築実績

- 数千万人以上ユーザを抱える**超大規模システムの構築経験・運用経験**を豊富に有しています。
- 複数の**並列分散処理・リソース管理OSS**を組み合わせた超大規模システム構築が可能です。
- 各OSSをソースコードレベルで理解しており、**OSSサポート**も承っております。

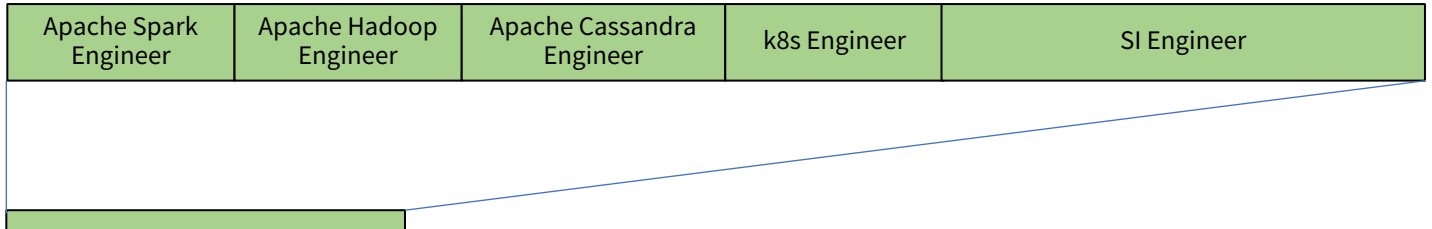


Apache Zeppelin / Jupyter Notebook
Apache Mesos / Apache Hadoop(Yarn)
Presto/drill /impala Spark/Hadoop MR MongoDB/Redis/Cassandra Oracle/MySQL/PostgreSQL
Kubernetes/Docker Swarm
Docker
OpenStack / DC/OS
Xen/KVM/VMWare/VirtualBox
Debian (Linux) /CentOS(Linux)/FreeBSD
Hardware

弊社訴求ポイント：他社コスト比較

- 日本最高ランクのデータエンジニアが短期間・少人数でシステム設計・構築を行う為、トータルコストは抑えられ・システム品質は非常に高いものとなります。

他社:



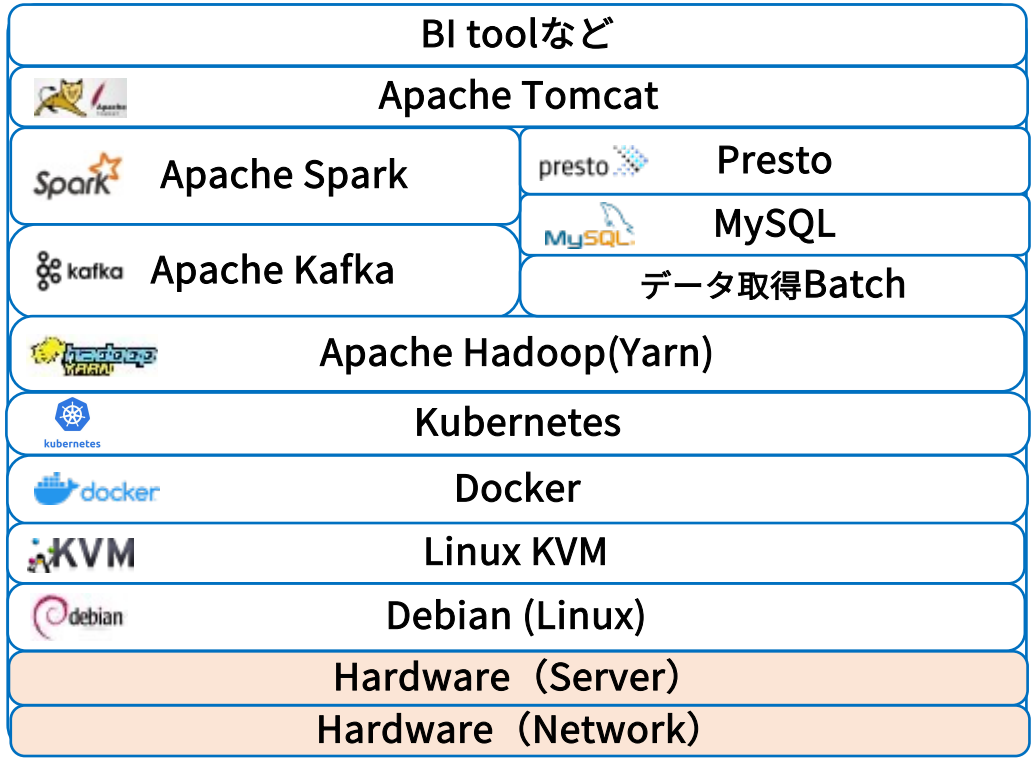
INTHEFOERST:

・他社：各分野のエンジニアを集めて大人数で設計・構築（大人数＋マネジメントオーバーヘッド発生）
・弊社IF：スペシャリストが少人数・短期間で設計・構築

データレイク・
ビッグデータ処理基盤

データ処理基盤（アプリケーション構成例）

- エンタープライズソフトウェア構成、OSS構成それぞれに弊社は対応しております。



- BIなど
- Apache Tomcat
- Presto
- Apache Spark
- MySQL
- Apache Kafka
- データ取得Batch
- Apache Hadoop（Yarn）
- Kubernetes
- Docker
- Linux KVM
- Debian（Linux）

※上記はサンプル構成です。詳細構成などはプロジェクト開始後に詳細検討・設計とします。

システム構成の特徴・メリット

- Hadoop・Spark・Levyx社製品など並列分散処理によるデータ処理高速化
（同時並行でのクエリ処理・予測計算にも対応）

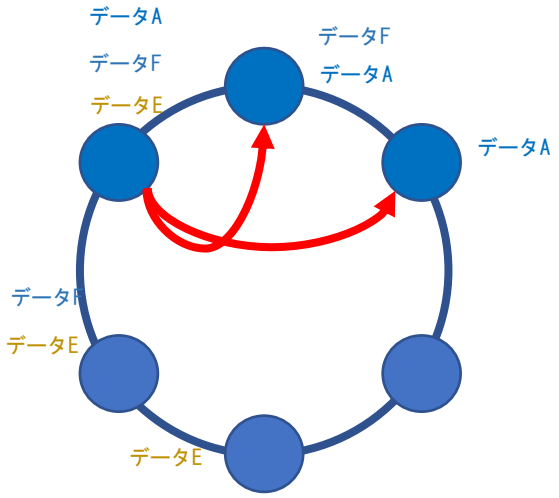
- ペタバイト以上の大容量データ・ユーザ数百万人以上の高負荷トランザクションにも対応

- データサイズ・処理規模に合わせて拡張が非常に容易、将来的なデータ・ユーザ増にも対応
（追加データ増にも対応、ノード数を増やすことで調整可能）

- クラウドに比較して圧倒的なコストパフォーマンス
（クラウド利用料は非常に高額）

（参考）分散DB Apache Cassandra

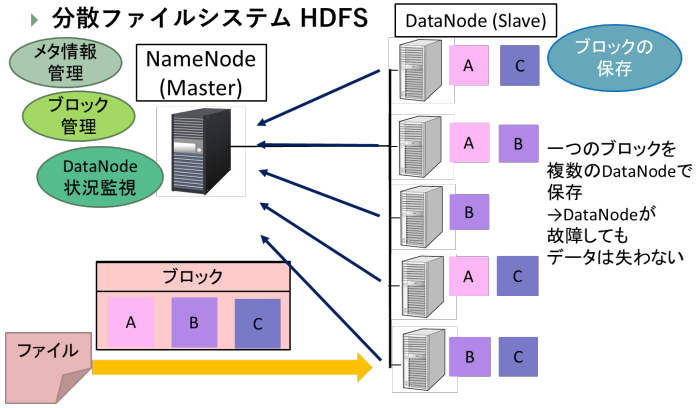
既存のデータベースに代わる次世代のPtoP型分散DB

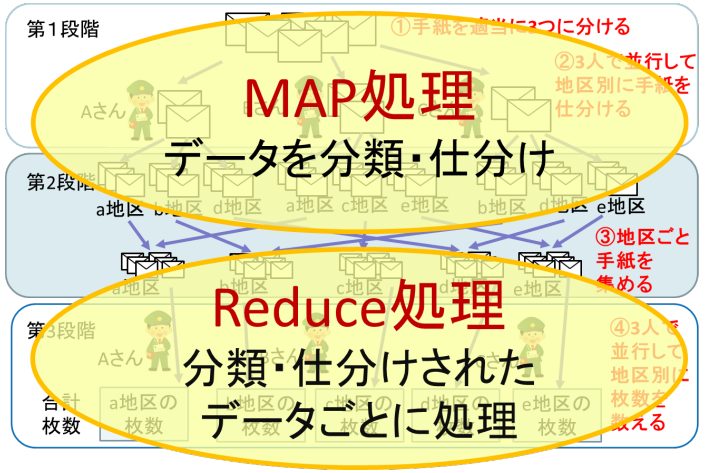


- iTunes・Instagram・NETFLIXなど億人ユーザ、ペタバイト・エクサバイト級システムで広く活用
  ※1ペタバイト=1024テラバイト、1エクサバイト=1024ペタバイト
- 超大規模・高可用性・高パフォーマンス
  （1000万クエリ/秒以上も対応可能）
- データセンター間連携が可能
  （マルチクラウド：AWS-Azure-オンプレ間のデータ連携も可能）
- 非常に高い可用性
  （複数ノードが落ちても問題発生せず、障害に強い）
- リニアなスケールパフォーマンス
  （トランザクション・データ量増大に対しても
  Cassandraノード数を増やすことで対応可能）

（参考）分散処理フレームワーク Apache Hadoop

分散ファイルシステム：Hadoop Distributed File System、分散処理管理：MapReduce Framework

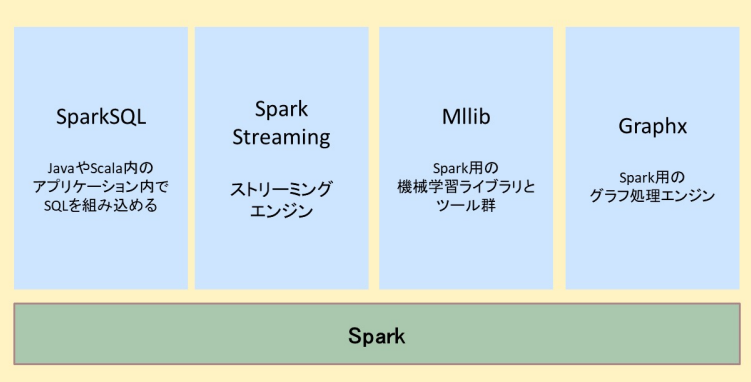




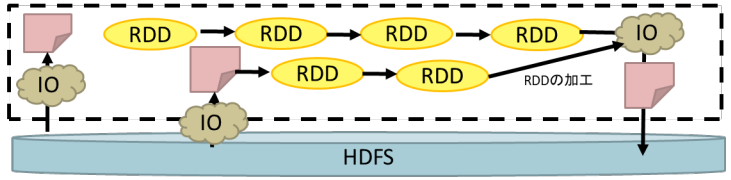
- 複数のIAサーバを束ねて、
  一つの大きな処理システムとして利用
  （特に大量データの格納・処理に最適化）
- Map処理、Reduce処理のみを指定すれば
  あとはフレームワークが並列分散処理を実現
- ノード数を増やせば、基本スケールする
- サーバが故障しても、ジョブは実行される

（参考）分散処理エンジン Apache Spark

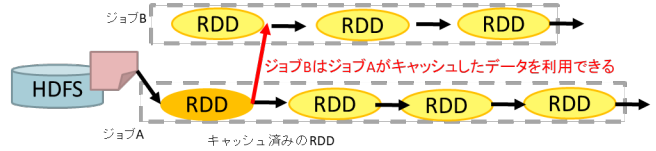
MapReduceに限らず、DAG(有向非循環グラフ)型で柔軟・高速に実行できる並列分散処理エンジン



①ジョブが多段に構成される場合、複雑な処理を少ないジョブ数で実現できる



②複数のジョブで何度も同じデータを利用する場合
何度も利用するRDDは複数のサーバのメモリに分割してキャッシュできる



- 様々な分散処理のライブラリ群
- Hadoopが普及にするにつれて、MapReduceフレームワークの処理効率が課題となっていた
- データ管理には向かないが、高速なデータ処理に向く
  (Hadoopの逆)

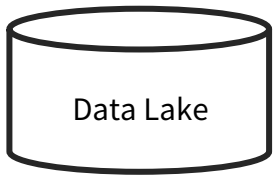
機械学習・
データ解析コンサルティング

データ処理全体の流れと機械学習予測

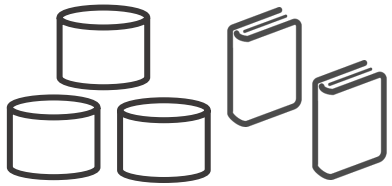
Data Prep

-Data Type Conversion
-Table Join/Sort/Count
-Unification of Units/ Names/DateTime

Loading Data from the Data Lake



Data Linkage from Multiple Data Sources



Data Sources

Feature Engineering

-Normalize
-Binning
-One Hot Encoding
-Replace Missing Values
-Up Sampling
-Dimensionality Reduction
-Feature Generation from Exploratory Data Analysis

Feature Selection

-Statistical Test (Uncorrelated Data for Labels)
-High Correlation between Variables(Multicollinearity)
-Detect Outlier
-Stepwise

Predictive Modelling

-Generalized Linear Model using Elastic Net
-Extra Trees Regression
-Light GBM Regression

Ensemble Modeling

-Voting
-Bagging
-Boosting
-Stacking

Predict By ML

Convert to Data for Business Intelligence

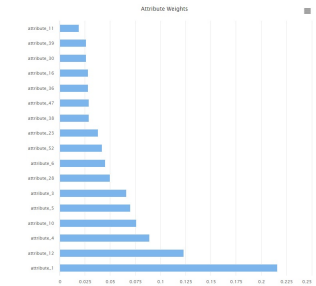
Provide to Web Service API

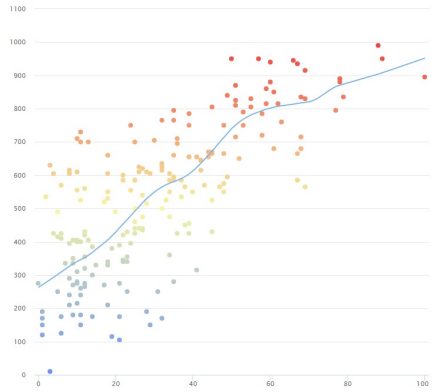
探索的データ分析によるデータ理解

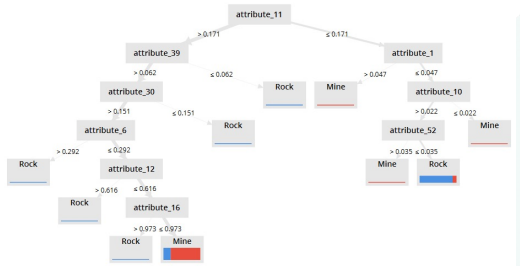
探索的データ分析(Exploratory Data Analysis)

・可視化・統計・機械学習などの手法を用いて、
多岐にわたり探索的にデータを分析・理解する行為。
・正しく予測モデルを作成することや予測精度向上のためには、
深くデータそのものを理解することが非常に重要。

Attribute	Coefficient	Std. Error	Std. Coeffici...	Tolerance	t-Stat	p-Value	Code
attribute_1	2.053	1.970	0.094	0.904	1.042	0.299	
attribute_2	2.007	1.800	0.132	0.930	1.115	0.267	
attribute_3	-5.631	1.679	-0.433	0.886	-3.353	0.001	***
attribute_4	3.253	1.035	0.303	0.953	3.142	0.002	***
attribute_8	-1.191	0.631	-0.203	0.910	-1.888	0.061	*
attribute_9	0.943	0.753	0.223	0.850	1.251	0.213	
attribute_10	-0.381	0.689	-0.102	0.802	-0.552	0.582	

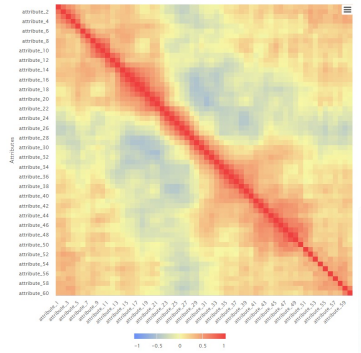


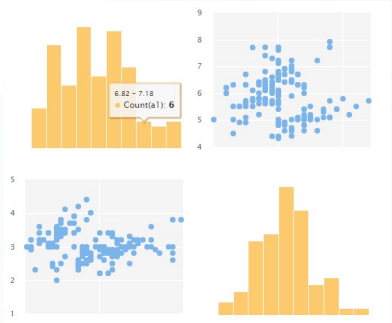




EDAによるデータ確認・検討（例）

・主要変数のクロス集計と可視化
・基本統計量・データの分布の確認
・各変数間の相関性・多重共線性の確認
・データ欠損値の有無・外れ値/異常値の有無
・Tree系アルゴリズムの変数重要度
・リーク情報は含まれていないか
・各変数のビジネス的意味合いとデータ内容の整合性に違和感はないか
・仮説検証





（参考）異常検知アプローチ

- 外れ値検知(Outliner detection)
  ー静的なデータを対象、データ分布から大きく外れているか？
  　Ex. One class SVM、k近傍法、Local Outliner Factor(LOF)

- 変化点検出(Change point detection)
  -動的なデータを対象(ex時系列)、データに変化が起きたポイントを検出
  　Ex. AR(自己回帰予測モデル)乖離スコア

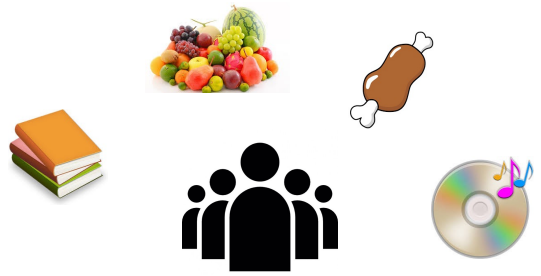
- 異常状態検出(Anomaly detection)
  ー動的なデータを対象、状態の正常・異常判定
  　Ex.部分時系列の近傍スコア

弊社システム構築事例

システム構築・機械学習事例（大手ポイントメーカ様）

- 競合大手SI：費用2億円以上、約1年以上の設計構築期間
- 弊社IF：総額4000万円、期間3ヶ月、人員3名（データ基盤2名+予測モデル1名）
  非常に厳しい個人情報保護・セキュリティポリシーにも対応

目的：商品レコメンデーション基盤構築

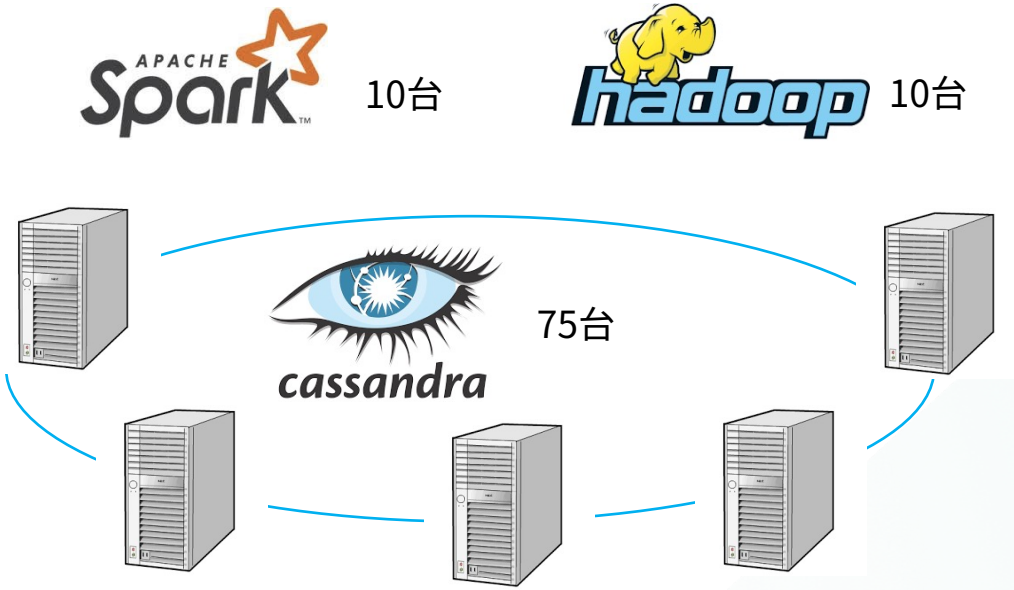


予測対象商品数
10万アイテム

予測対象者数
7200万人

データサイズ
250TB

サーバ100台規模の並列分散処理環境
（インフラAWS、全てOSSで構築）



システム構築事例（公益財団法人 高輝度光科学研究センター様）

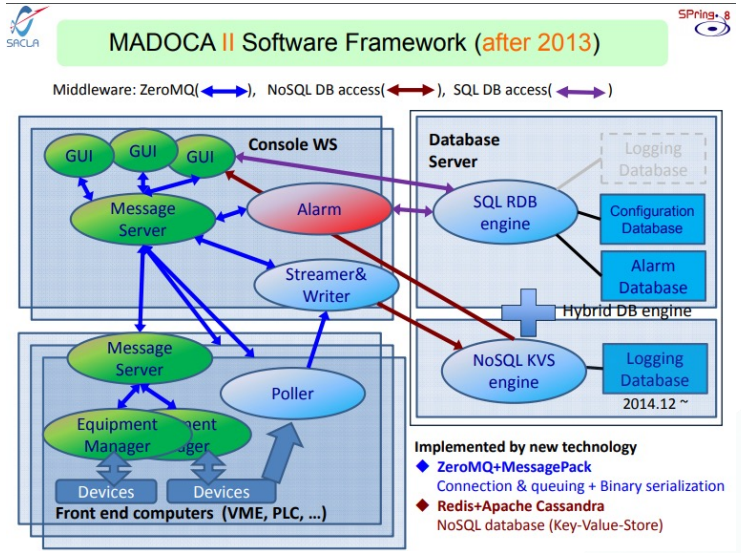
- 大型放射光施設SPring-8において、加速器とビームライン制御システム MADOCA II のログ保管DB Cassandraの設計支援・サポートを弊社が担当

目的：放射光（センサーデータ）の保管

24時間切れ目なく50,000クエリ/秒を要求する高負荷処理

MADOCA II Software Framework

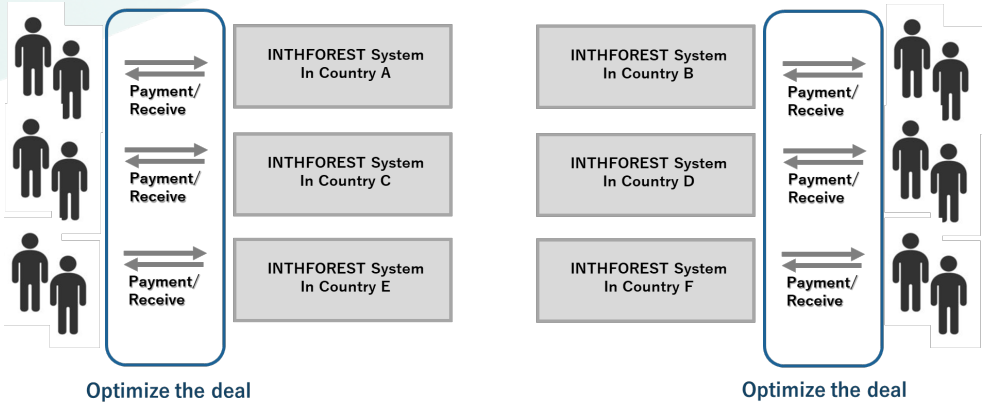
（MADOKA : Message And Database Oriented Control Architecture）

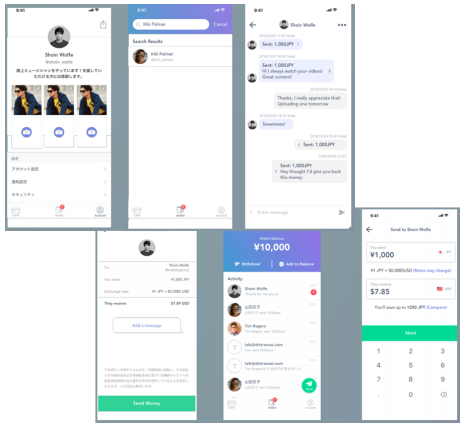


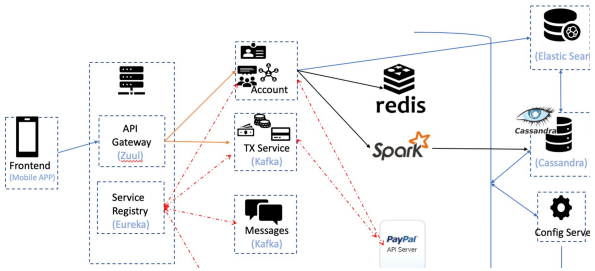
Ryotaro Tanaka ” Conceptual Design of the Control System for SPring-8-II” 10th International Workshop on Personal Computers and Particle Accelerator Controls 14th - 17th October, 2014,Karlsruhe, Germany

システム構築事例（CtoC国際送金システム）

目的：個人間の国際送金プラットフォーム







- 各国間リアルタイムのデータ処理連携
- ユーザ数万～数千万～数億人規模に対応した分散処理システム
  （シームレスなスケール）
- 為替価格ロジックを組み込んだシステム設計
  （対抗可能な企業が日本で数社）
- 圧倒的なコストパフォーマンス
  （構築コスト 50億円～100億円 が1/10以下へ）
- 処理規模に合わせてシステム運用コストが可変
  （ノード数増減で調整）
- 金融庁監査に耐えうるシステム設計・監査ログ取得
  （法令・監査対応）

データ処理基盤・データ分析など
お気軽にお問合せください

株式会社INTHEFOREST
お問い合わせ sales@intheforest.co.jp