Strataで盛り上がっていた データ処理系のOSS

株式会社　NTTドコモ
サービスイノベーション部 ビッグデータ担当
原野　将大

もくじ

- Strata Data Conferenceって？
  - 会議の紹介
- 全体の中で盛り上がっていたセッションの紹介(一部)
  - KSQL：kafka対応ストリーミングSQLエンジン
  - BigDL：Spark用分散DLライブラリ
- まとめ

自己紹介

- 所属/名前
  - NTTドコモ　サービスイノベーション部　ビッグデータ担当
  - 原野 将大(はらの　まさひろ)
- 業務
  - ドコモR&D部門におけるビッグデータ基盤の開発と運用(1年ぐらい)
  - 内製でDevOps
- 好きなAWSサービス
  - Cloudformation

- 趣味

どんなシステム？

- ビッグデータ統合分析システム
- オンプレ ＋ AWSクラウド
- 社内分析者 数百名
- データサイズ　数PB
- 毎日　数十TB
- 内製（ピザ2枚人程度？）

Strata Data Conferenceって？

Strata Data Conference概要

- Strata Data Conferenceとは
  - O'Reillyが主催するビッグデータ処理に関するカンファレンス(主にOSS)
    - 機械学習
    - データ管理
    - ストリーム処理
    - 分散処理
    - など

- 年に3回開催(2018年度)
  - San Jose
  - London
  - New York

- カンファレンス構成
  - 約20のチュートリアル(1日)
  - 約170のセッション (2日間の合計)
  - 展示ブース

Strata Data Conference@San Jose

- 参加日
  - 2018/03/06～2018/03/08
- 場所
  - Convention Center
- 参加者数
  - 1000以上
  - 日本人は50人？

- 全体所感
  - SparkやStreamingの他に**DL**(TensorFlow、Kerasなど)のセッションが多かった(聴講者の人気は**リアルタイム処理**が圧倒的)

スポンサー

NTT docomo

Presented by

cloudera

O'REILLY

Elite Sponsors

MAPR DATA TECHNOLOGIES

Microsoft

Strategic Sponsors

Google Cloud

IBM

MEMSQL

Zettabyte Sponsor

DATASCIENCE.COM

Contributing Sponsors

HITACHI Inspire the Next

Kyligence

Exabyte Sponsors

aws

intel

MicroStrategy

NetApp

sas

snowflake

talend

TalkingData

Impact Sponsors

confluent

DOMINO

IMPETUS

kinetica

kyvos insights

redislabs home of redis

striim

syncsort

vmware

参加動機

- ビッグデータ処理関係OSSツールの最新動向を調査するため

- 社内ビッグデータ分析基盤では、オンプレミスとAWSのハイブリッド構成で、統合管理の観点でOSSも注目している。
  - 弊社でも活用できそうなOSSがあるか事例や生の声を聴く
  - 今後、自社プライベートクラウドやマルチクラウド化も見据え、目利きをする

セッション選定理由

- 社内で力を入れているキーワード
  “Realtime処理” “AI/deep learning”

- 取り組んでいる/みたい領域(社内で/個人で)
  ＋ 自チームの課題領域（データ処理/ETL）

- 人の流れ≒人気のセッション

セッションの一部をピックアップ

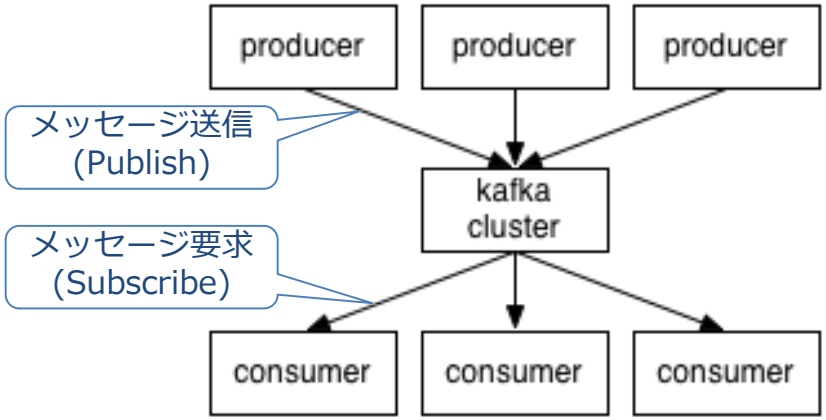
① KSQL：kafka対応ストリーミングSQLエンジン

② BigDL：分散DLライブラリ

参考：Kafkaとは

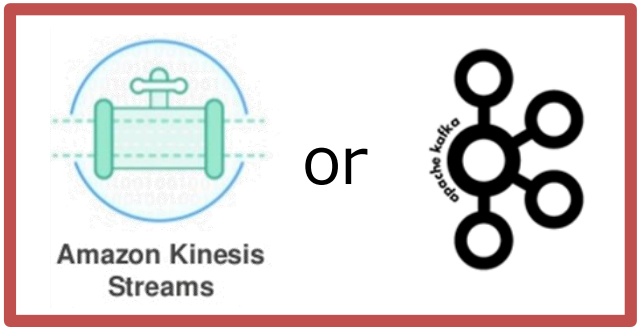
- **分散型メッセージング/キューイングシステムのOSS**
  - 2011年にLinkedInが開発

- 特徴
  - 大量のメッセージをリアルタイムに処理することが可能
  - クラスタ構成でスケールアウトが可能



- 導入企業
  - Yahoo、twitter、Netflix、Uber、他
- 類似サービス
  - **AWS Kinesis Data Streams**
  - Google Cloud PubSub
  - Activ MQ
  - Rabbit MQ

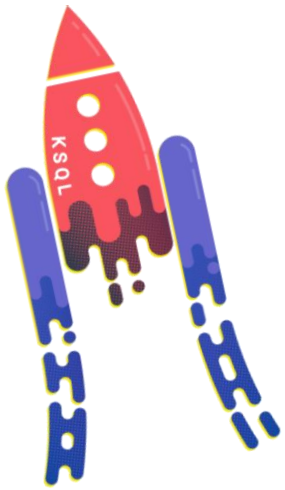
機能的には
ほぼ同等



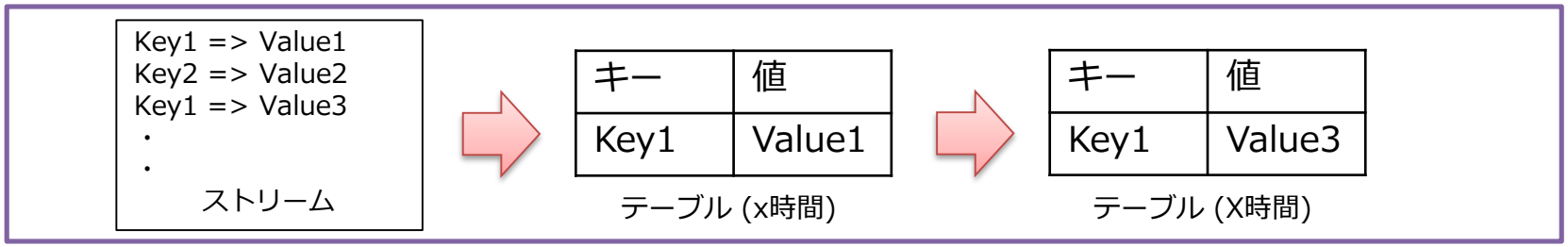
- バージョン
  - v1.1.0 (2018年3月28日)

KSQLとは

- **Kafka用ストリーミングSQLエンジン**
  - **Conflent（元Linkedinのkafka開発元）が開発**
  - **2017年8月に公開 (現在最新 v0.5.0)**
  - **Conflent社のConfluent Platformで 2018/4 GA予定**



- **特徴**
  - **ストリーミングデータに対し、SQLライクな構文でクエリ実行可能**



- **類似サービス**
  - **Kinesis Data Analytics**

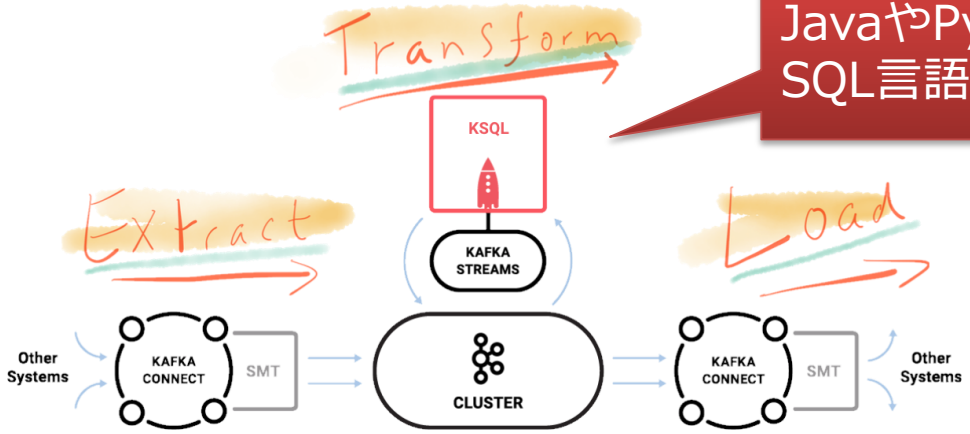
The future of ETL isn't what it used to be

- **発表者**
  - **Gwen Shapira (Confluent社)**

KafkaとSqoop
のコミッタ

- **内容**
  - **Kafkaに対応したOSSのSQLライクなエンジン(KSQL)の紹介と、それを活用することで、従来のバッチ処理の他に、ストリームデータに対し直接処理（例：ETL）できるという話**

JavaやPythonではなく、
SQL言語で処理を書ける



☺ **KSQLは、次々流れてくるストリームデータに対してクエリを実行可能で、データの加工処理を容易に実装できる**

適応例：

- **リアルタイムETL**
- **異常検知**
- **モニタリング(データの遅延状況などの監視)**

比較

KSQLとSQLの違い

- KSQLは基本的にANSI SQL準拠
- ただし、Streamデータに対する処理としてWindowの概念が考慮されている

Streaming ETL

```
CREATE STREAM vip_actions AS
  SELECT userid, page, action
  FROM clickstream c
  LEFT JOIN users u ON c.userid = u.user_id
  WHERE u.level = 'Platinum';
```

Anomaly Detection

```
CREATE TABLE possible_fraud AS
  SELECT card_number, count(*)
  FROM authorization_attempts
  WINDOW TUMBLING (SIZE 5 SECONDS)
  GROUP BY card_number
  HAVING count(*) > 3;
```

Kafka KSQLとKinesis Data Analyticsとの違い

- 基本的に機能は同じ
- 当然だが、OSSか否か
  - オンプレで頑張れるならKSQL、分析処理等のみに集中したいのであればマネージドなKinesis Data Analytics (ただし、東京リージョン未対応)

セッションの一部をピックアップ

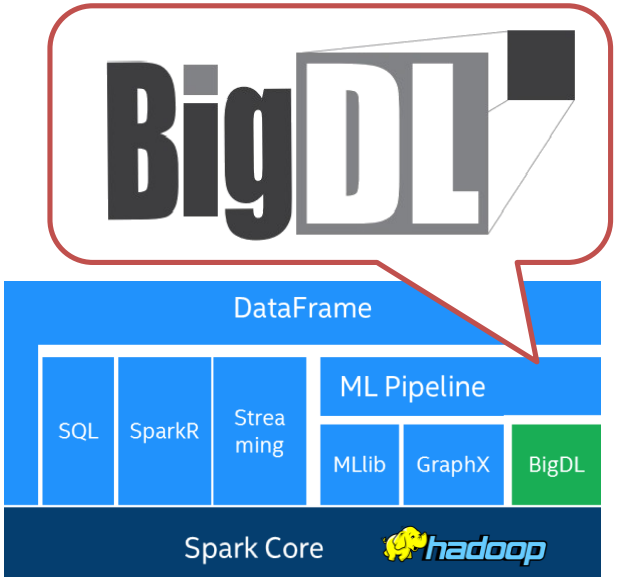
① KSQL：kafka対応ストリーミングSQLエンジン

② **BigDL：分散DLライブラリ**

BigDLとは

- **Spark上で動く分散DLライブラリのOSS**
  - 2017年にIntelが開発　（現在最新 v0.5）

- **特徴**
  - Hadoop/Sparkクラスタの流用可能
  - スケールアウトによる高速処理
  - CPU(Xeon)上で設計されている
    ※既存リソースの有効活用
  - Caffe/Torchの学習済みモデルが利用可能

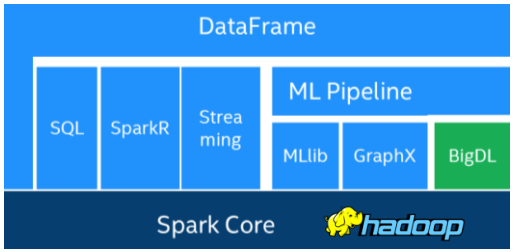


[背景]
世の中でDL基盤としてGPU利用に対する障壁（コスト・知見）の課題があり、既存リソース・知見を有効活用する手段としてIntelが提供

- **その他**
  - AWS Marketplaceで「BigDL with Apache Spark」のAMIを提供中
  - Githubで公開　(https://github.com/intel-analytics/BigDL)

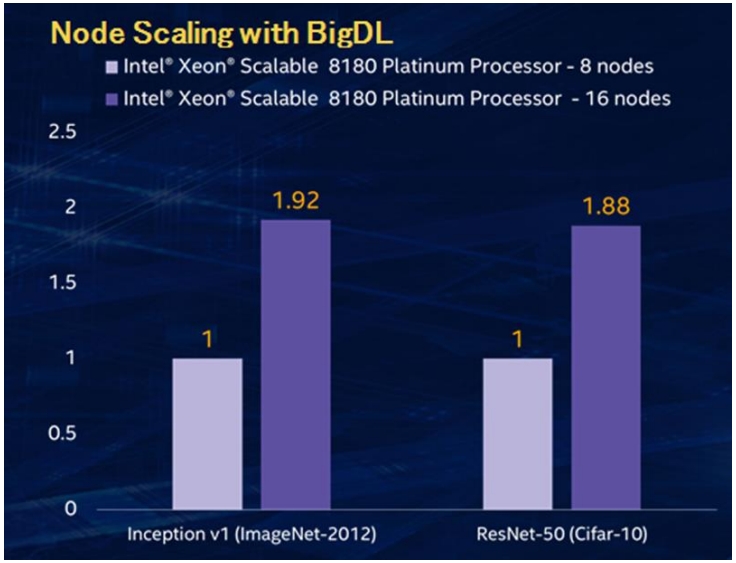
Accelerating deep learning on Apache Spark using BigDL with coarse-grained scheduling

- **発表者**
  - **Sergey Ermolin (Intel)**



- **内容**
  - **Hadoop/Sparkクラスタがあれば、DLの専門家でなくても学習済みモデルを直接Loadして実行可能なDLライブラリの紹介(GPUの対抗？)**

- **ベンチマーク**
  - **Node数を 8 → 16 にスケールアウトした結果**
    約 x1.9 性能向上を実現



☺ **BigDLは、 Hadoop/Sparkクラスタ上のデータに対して新規言語を習得することなくDLを容易に実装ができる(CaffeやTorchの学習モデルを利用可能)**

適用例：
- **既存のクラスタを流用して、クラスタ上のデータの分析(GPUサーバの購入は不要)**

全体まとめ・所感

Strata Data Conferenceの個人的に盛り上がっていたと感じたセッションの紹介

- **KSQL：kafkaに対応するSQLライクな構文を発行可能なOSS**
  - **Kafka KSQLとKinesis Data Analyticsで機能的差異を感じなかった**
  - **Kafkaに対してJavaやPythonではなく、SQLライクな言語で処理を記述できる点が良い**
  - **オンプレや非AWS環境ではKafka+KSQLは選択しとなりうる。**

- **BigDL：分散DLライブラリOSS**
  - **Hadoop/Sparkクラスタ上でDLが実行でき、CPUマシン上でCaffeやTorchの学習済みモデルを利用できる点が良い（マイグレーション）**
  - **ただし、Hadoop/Sparkの既存資産がなく、すでにGPUマシンでがっつりDLをやっているなら魅力はあまり感じないかもしれない**
  - **★おもしろい内容なので、今後の動きを見る**

KSQLを利用している方でここが
良いなどあれば教えてください