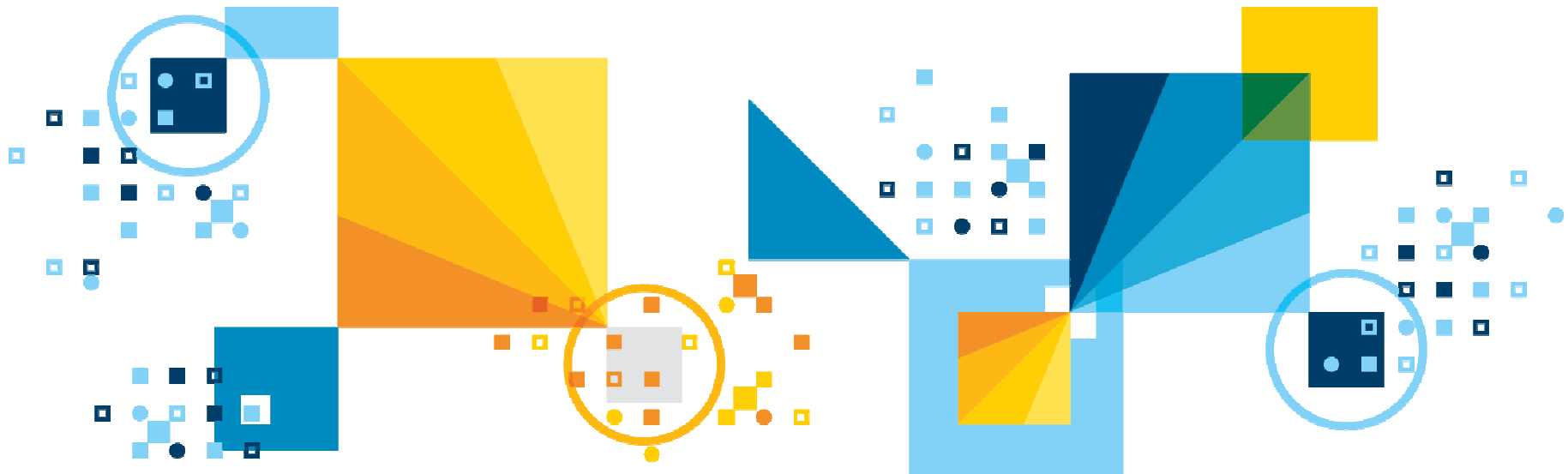


Apache SparkとDeepLearningについて

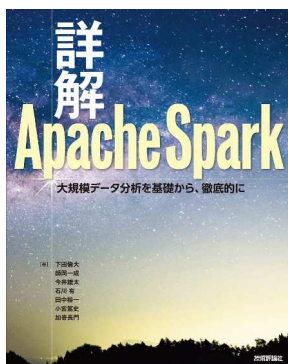


自己紹介

田中裕一 (yuichi tanaka)



主にアーキテクチャとサーバーサイドプログラムを担当することが多い。Hadoop/Spark周りをよく触ります。Node.js、Python、最近はSpark周りの仕事でScalaを書くことが多い気がします。休日はOSS周りで遊んだり。



詳解 Apache Spark



自己紹介



大規模データを、リアルタイムに分析！

DMM の Spark への取り組み



艦これ bot 解析

Spark Streaming



SparkStreaming によるリアルタイムレコメンド

Spark Streaming



Spark MLLib



Kafka+Spark による大規模ログ収集基盤

Spark Streaming



Apache Kafka



アジェンダ

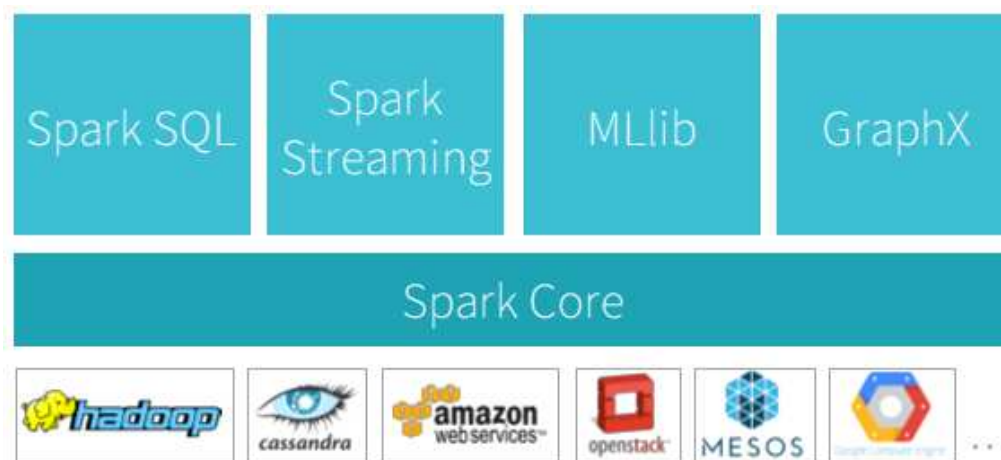
- Apache Sparkのおさらい
 - 改めてSparkとはどんなフレームワークなのか？
 - Sparkはどのように動くのか？
 - SparkでMachiLearning
 - 従来の課題は何か、Sparkがなぜマッチするのか
- DeepLearningのおさらい
- SparkでDeepLearningはどうか？
- 既存のDeepLearningFrameworkの問題はどこにあったのか
- DeepLearning4J on Sparkはどう動いているのか
- TensorflowOnSparkはどう動いているのか

Sparkとは

従来Hadoopでは難しかったBigDataにおける
アドホック分析やニアリアルタイム処理を実現するための
InMemory分散並列処理フレームワーク。



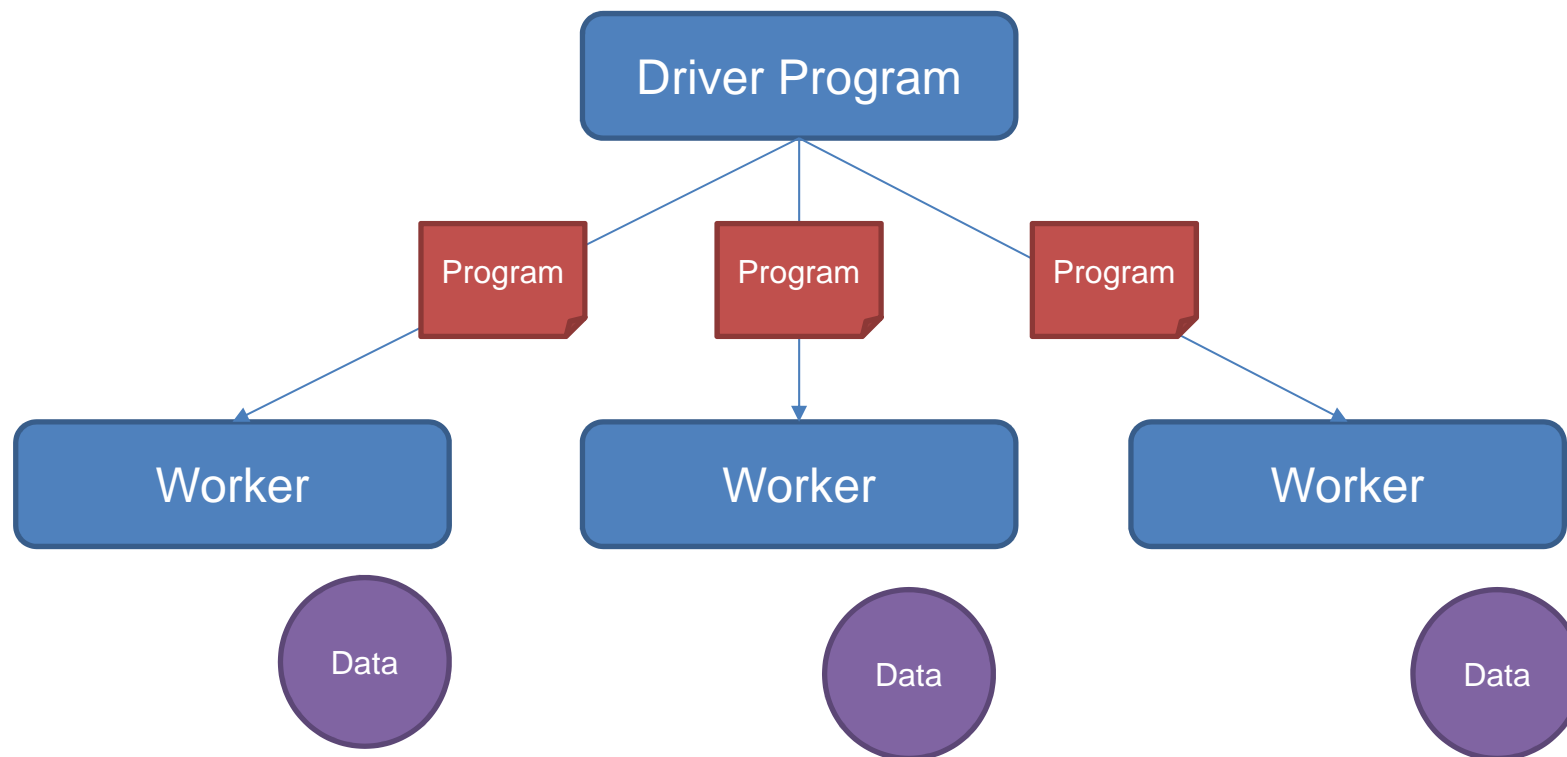
- HDFSを筆頭にCassandraなど分散ストレージのデータと相性が良い
- YARN,Mesos,Standaloneの3種類の分散処理基盤の上で動作
- SparkSQL,Streaming,MLlib,GraphXといった処理の拡張を持つ



Sparkはどう動くのか？

Sparkとは

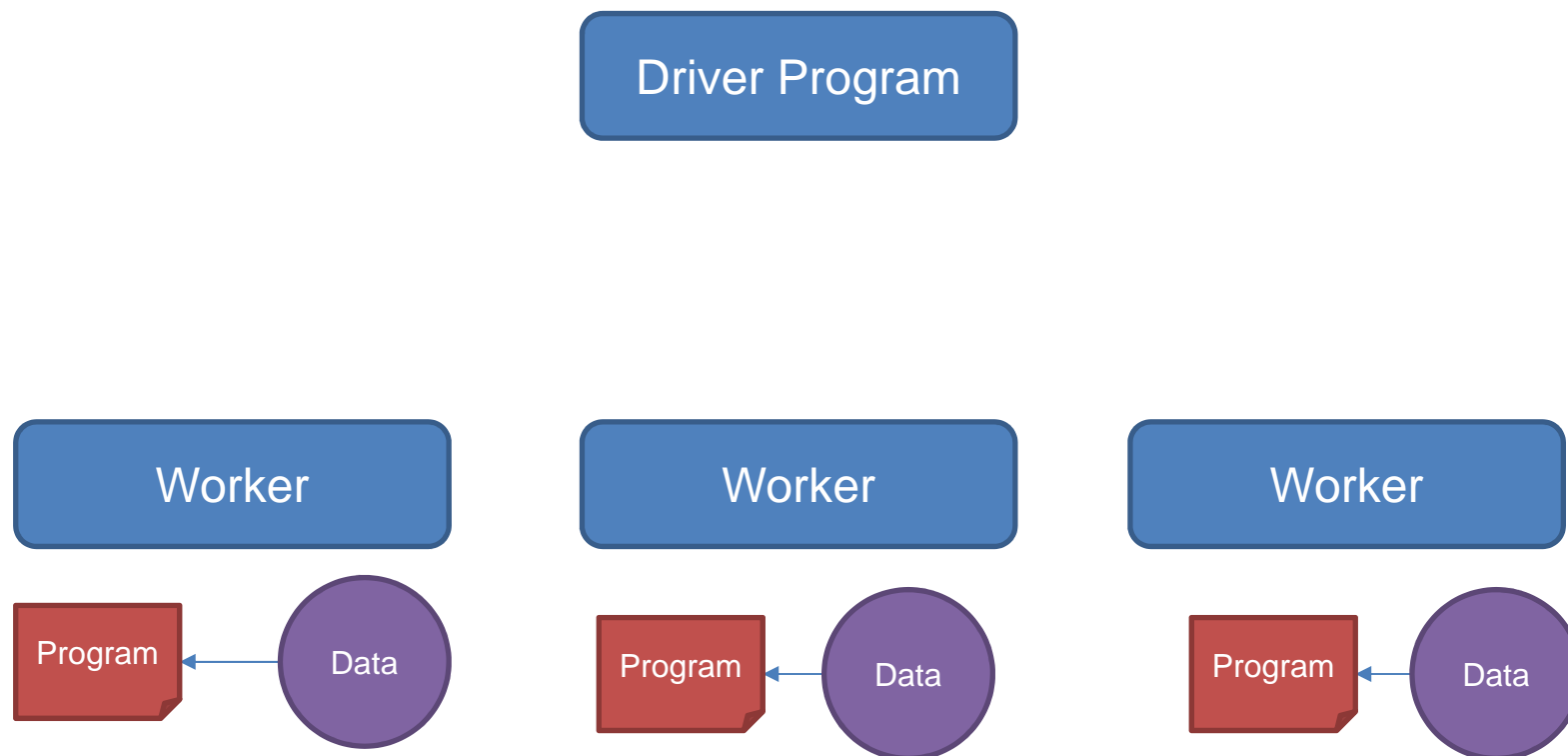
RDDとDAGをコアコンセプトとして設計された分散並列処理フレームワーク



Sparkはどう動くのか？

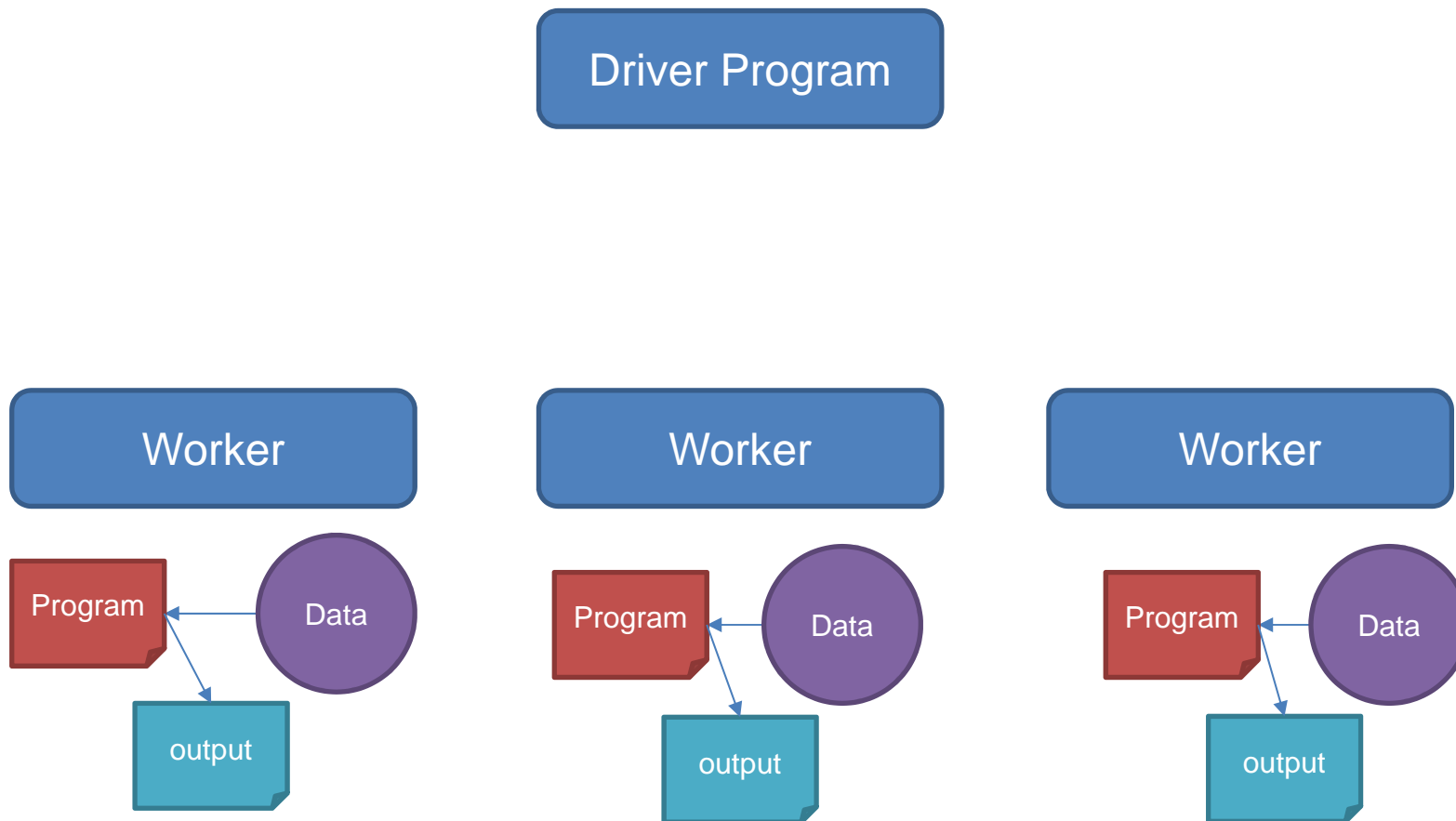
Sparkとは

RDDとDAGをコアコンセプトとして設計された分散並列処理フレームワーク

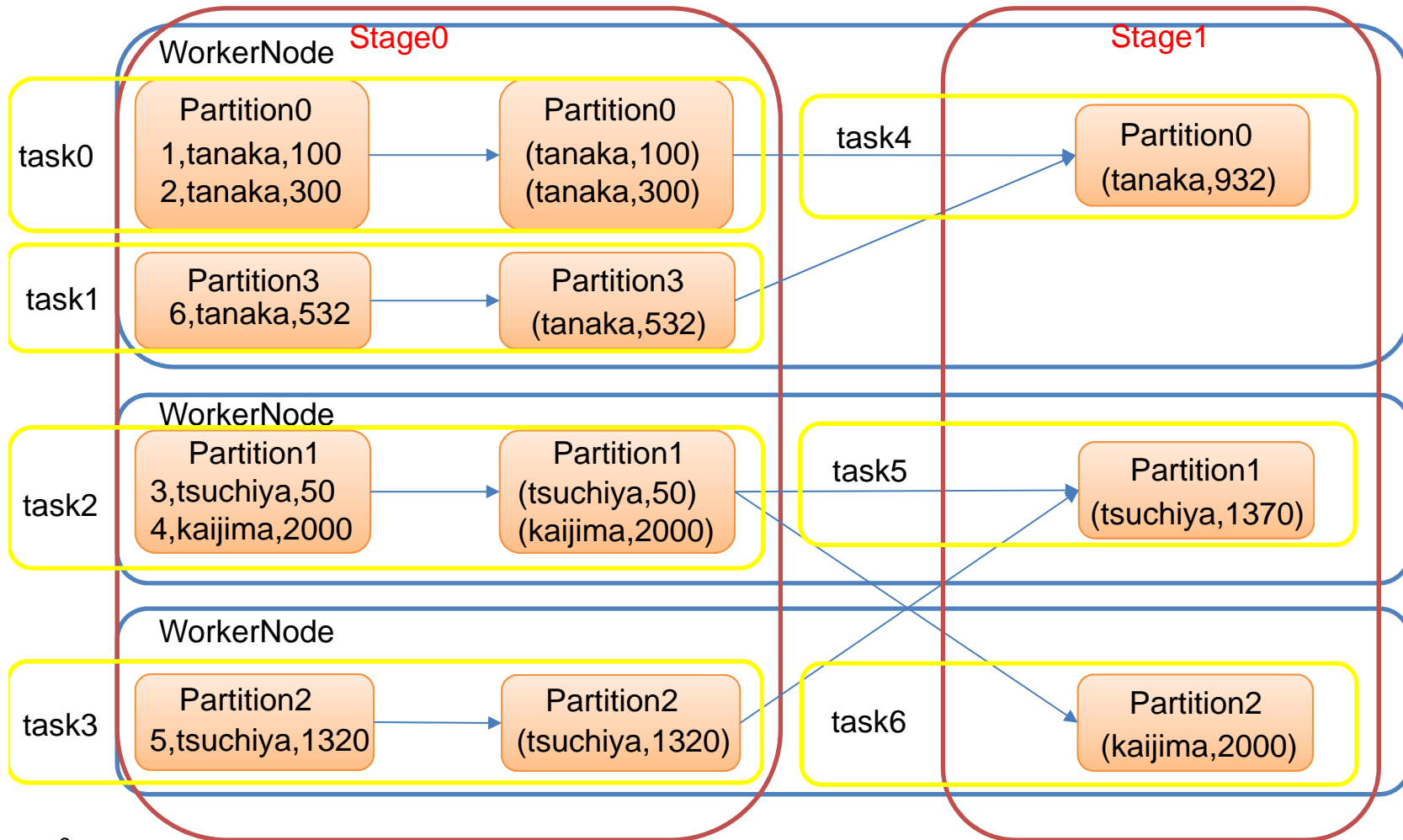


Sparkはどう動くのか？

Sparkとは
RDDとDAGをコアコンセプトとして設計された分散並列処理フレームワーク



Sparkはどう動くのか？



SparkでMachineLearning

- MLlibは大きく2つの実装に分かれる
 - spark.mllib
 - spark.ml
- どちらを使えば良いか？
 - spark.mlを使うこと！
 - spark.mllibは2系でメンテモードになりました。
 - 今後新しい機能は追加されません。タブン
 - (予定では3.0でremoveされる)

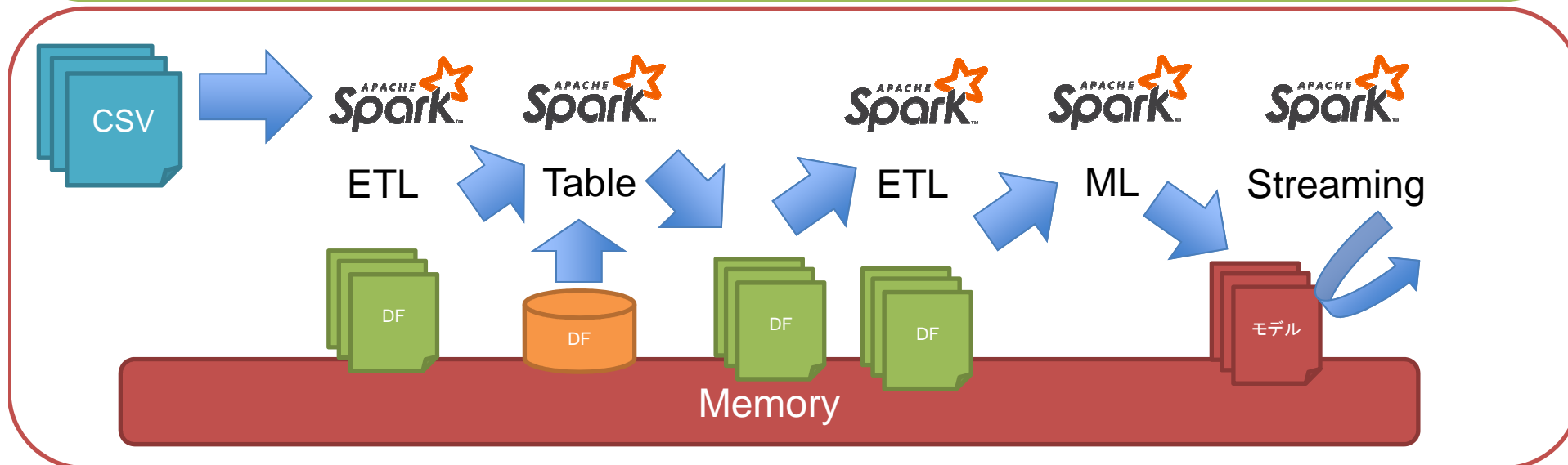
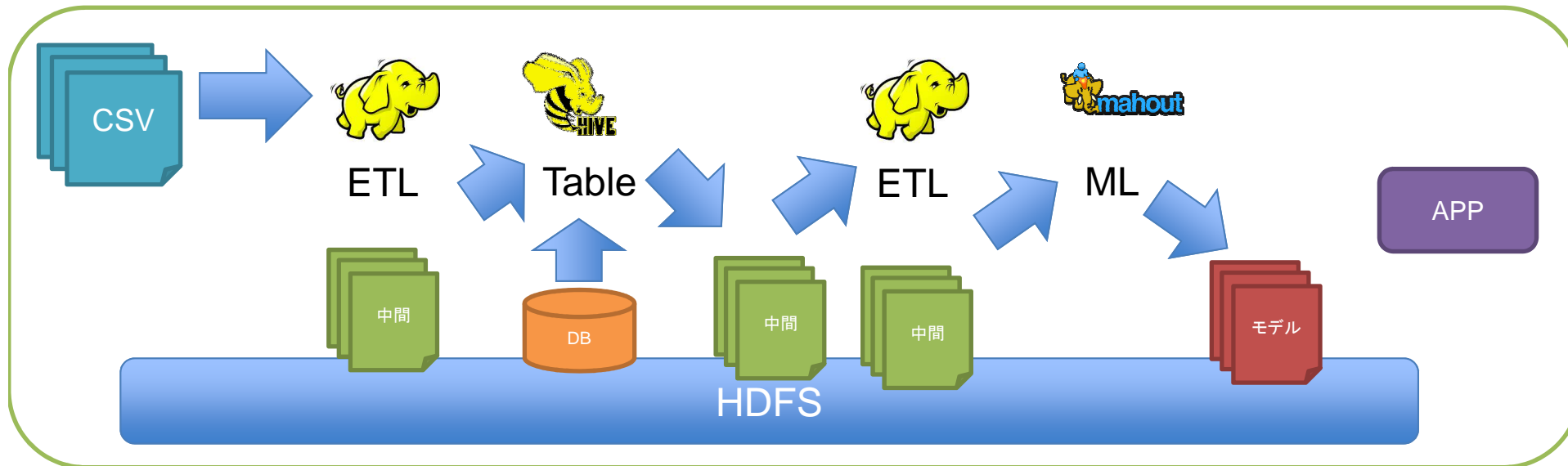


SparkでMachineLearning

- そもそもMachineLearningとは
 - データから新たな価値を見出すための手法の一つ
- MachineLearningでは前処理が重要
 - 前処理8割といわれる程、データサイエンスの占める作業のうち、大部分がこの作業
- 前処理の例
 - フィールドのコード化(性別変換・カテゴリ変換)
 - ID変換(Join, CookieSync, ジオコーディング)
 - 集計(GroupBy, Sum, Max, Min)
 - 形態素解析系処理(構文解析、分かち書き、ストップワード)
 - 名寄せ(表記ゆれ)
 - クレンジング(無効値処理、欠損値補完、外れ値補正)
 - 画像系処理(特徴抽出、スライス、サイズ変換、グレースケール)
- 背景
 - データは分析されることを前提としていない。こう言った配慮は事業において負荷となる

従来の課題は何か、Sparkがなぜマッチするのか

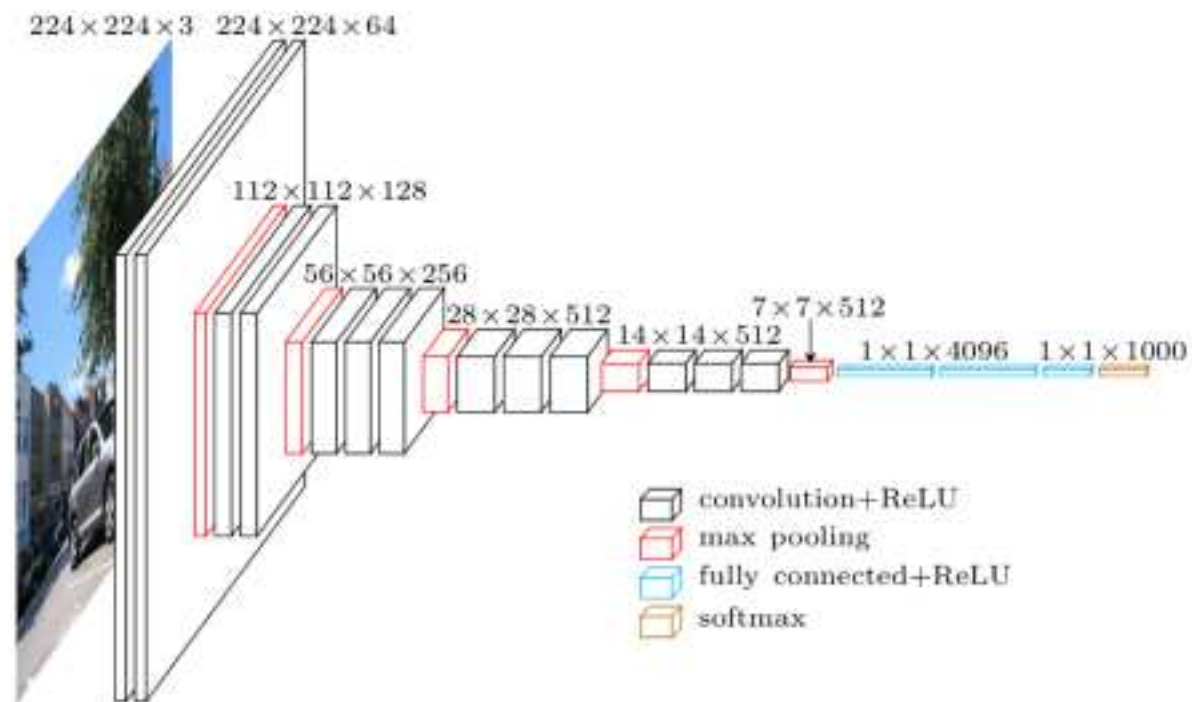
- こうした多様な前処理は従来のHadoop Ecosystemでの実現は難しかった



DeepLearningとは

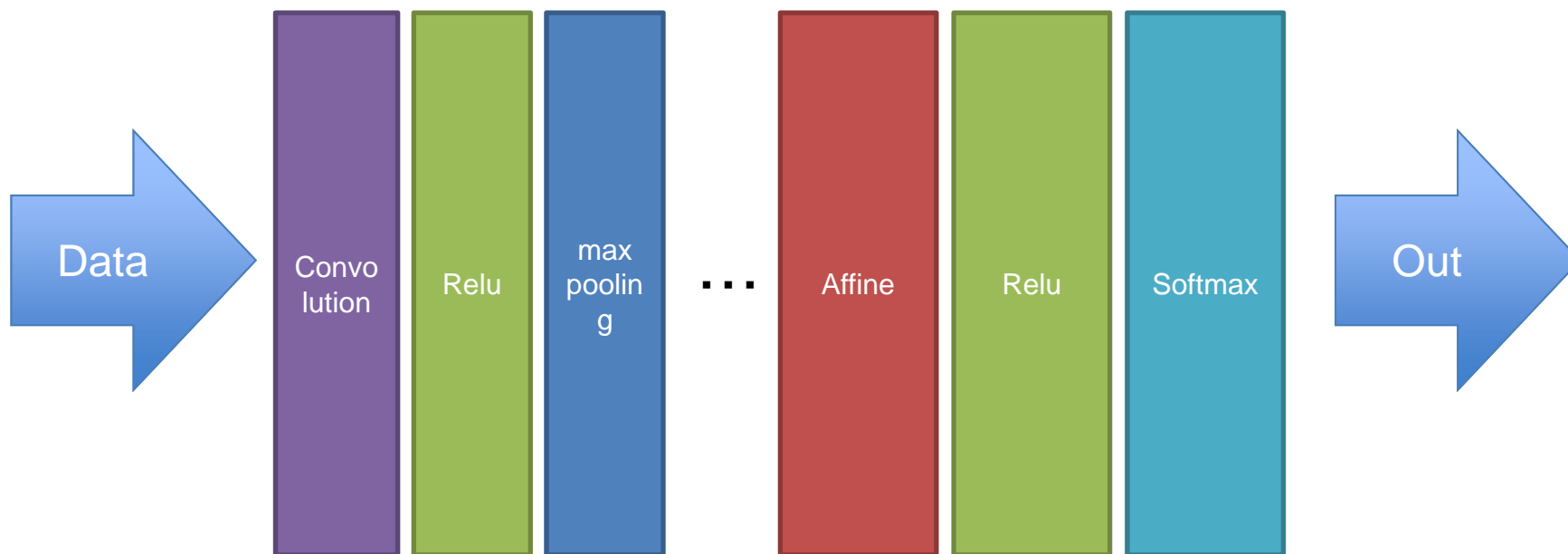
- 機会学習の一つであるNNの層を重ね高精度な予測・分類処理を行うための手法
 - 色々な関数の層を重ねて精度を向上させる
 - 画像識別の分野で精度が高くて有名 (ResNet/GoogLeNet等)

例) VGG16



さっきのを簡略に書き直すと

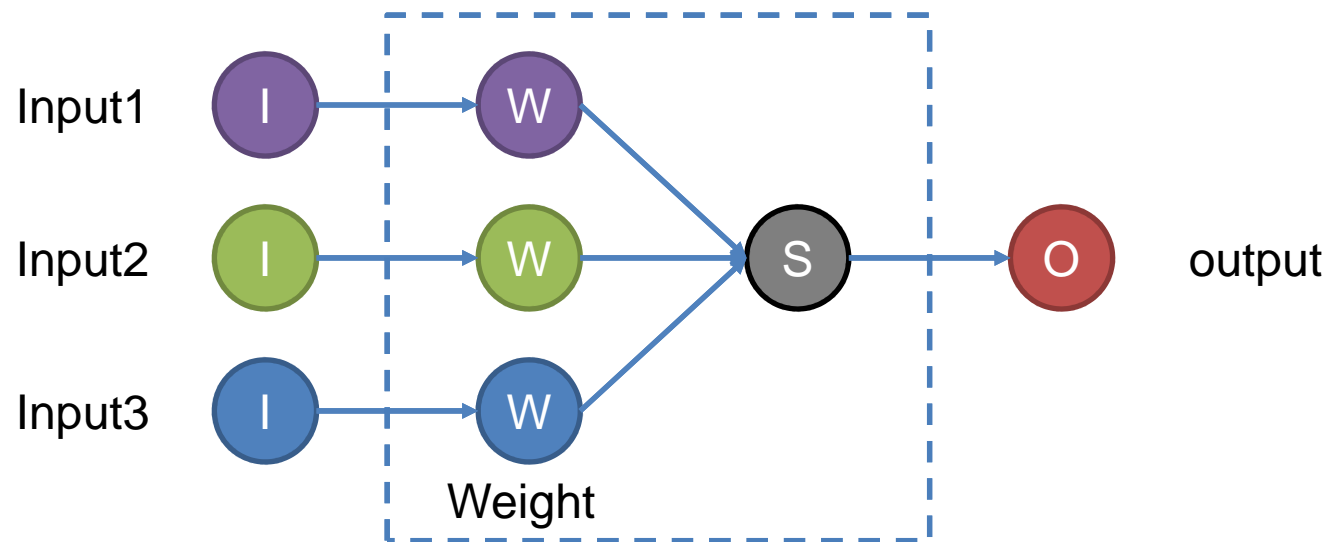
VGG16の簡略図



背景

- 人間の脳の構造を模してやれば上手く識別できるのは？

例) 単純パーセプトロン



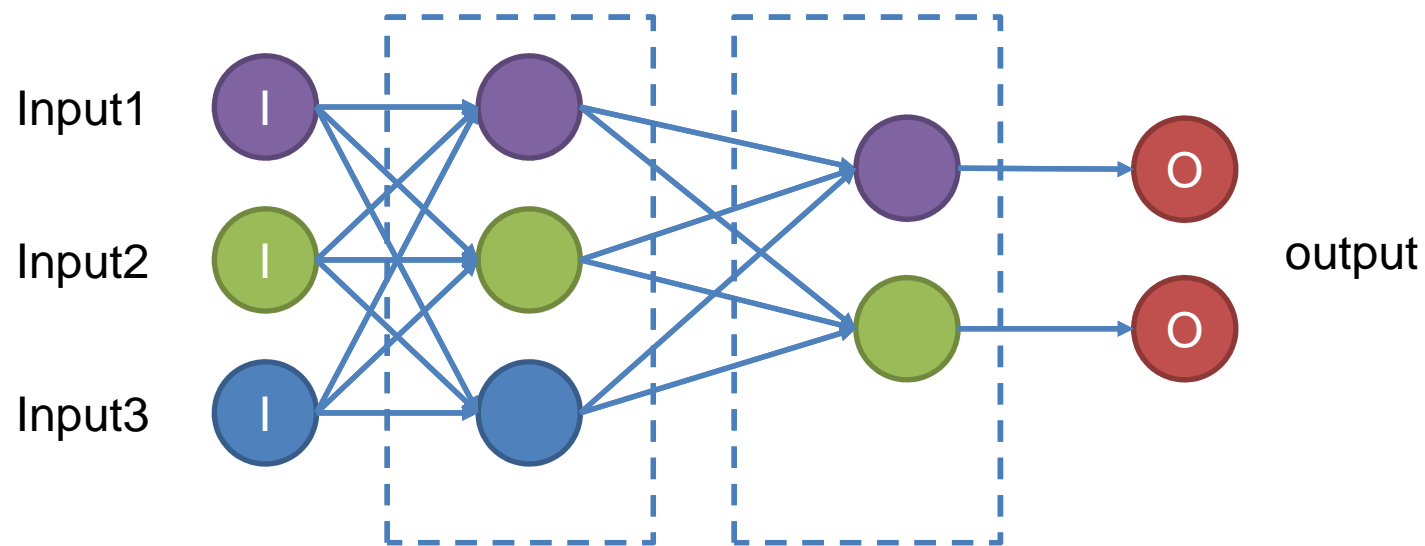
ニューロン情報処理モデル

参考)
[単純パーセプトロンの基本のき](#)

背景

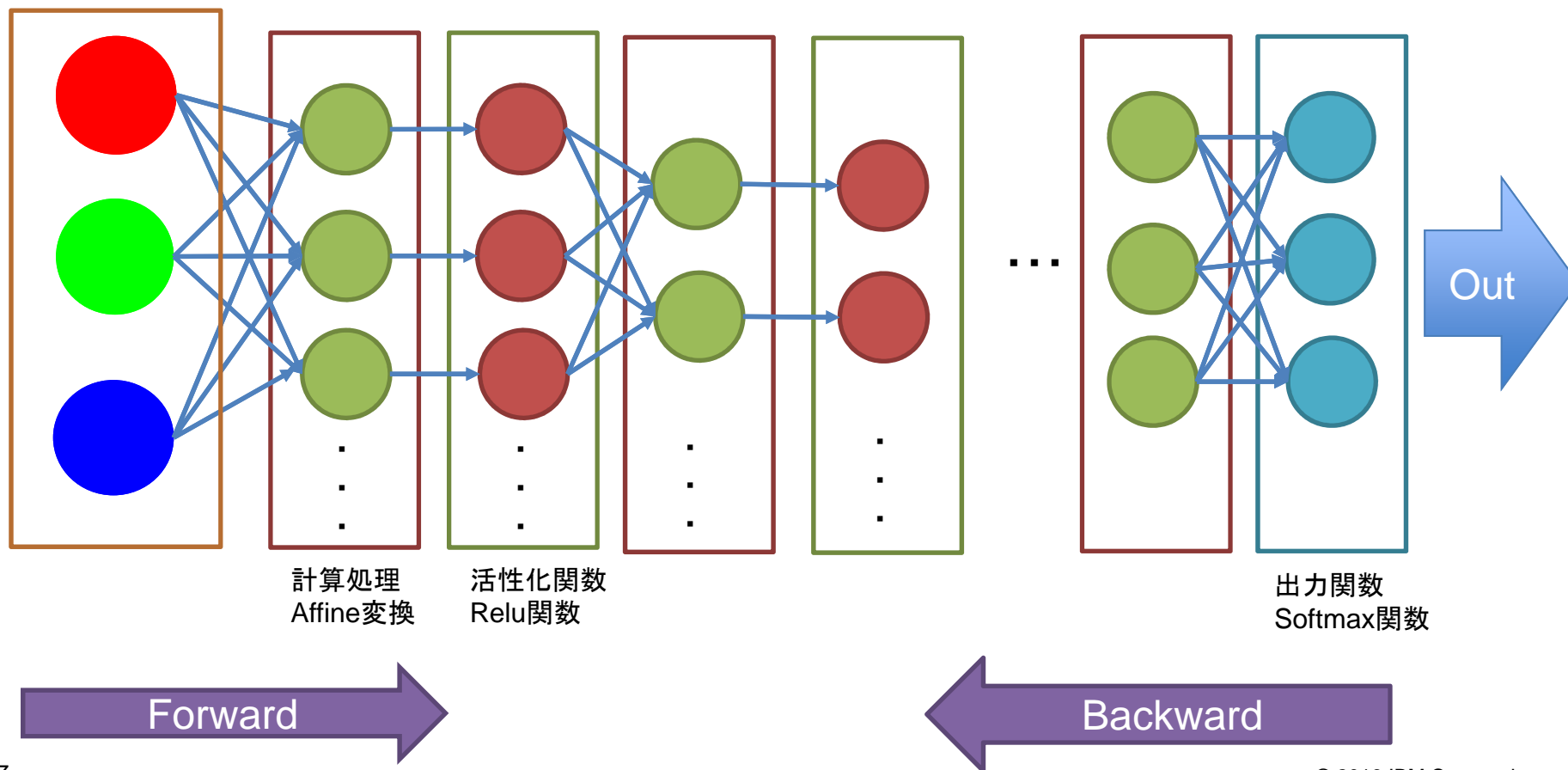
- 層をふやしてやれば上手く識別できるのは？

例) FF-NN



DeepLearning

- 階層をより深くし、ForwardとBackwardによりパラメータ設定まで自動化



SparkでDeepLearningはどうか？

- MLlibだけではDeepLearningは難しい(多層パーセプトロンはある)
- Spark v2.1でも難しい
- 16年に各DLフレームワークがSparkに対応
 - DL4J on Spark : <https://deeplearning4j.org/>
 - Caffe on Spark : <https://github.com/yahoo/CaffeOnSpark>
 - TensorSpark(Tensorflow on Spark) : <https://github.com/adatao/tensorspark>
 - Distributed Keras : <https://github.com/cerndb/dist-keras>
 - Sparkling Water : <https://github.com/h2oai/sparkling-water/tree/rel-2.0>
 - TensorFlowonSpark : <https://github.com/yahoo/TensorFlowOnSpark>

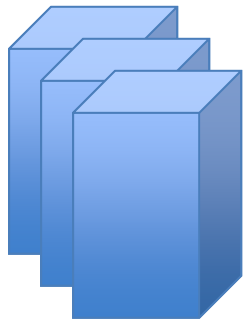
その他にも雨後の筍のように。

※Web上でも様々な記事が上がっていますが、すでにメンテされていないリポジトリもあるので選定する際は注意が必要

既存のDeepLearningFrameworkの課題

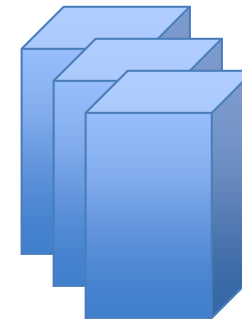
- 機械学習の基盤とDeepLearningの基盤を別に作る必要がある
- pipelineを構築するために別のプログラムを書く必要がある

Hadoop/Spark Cluster



1.前処理
3.モデルの適用

DeepLearning Cluster



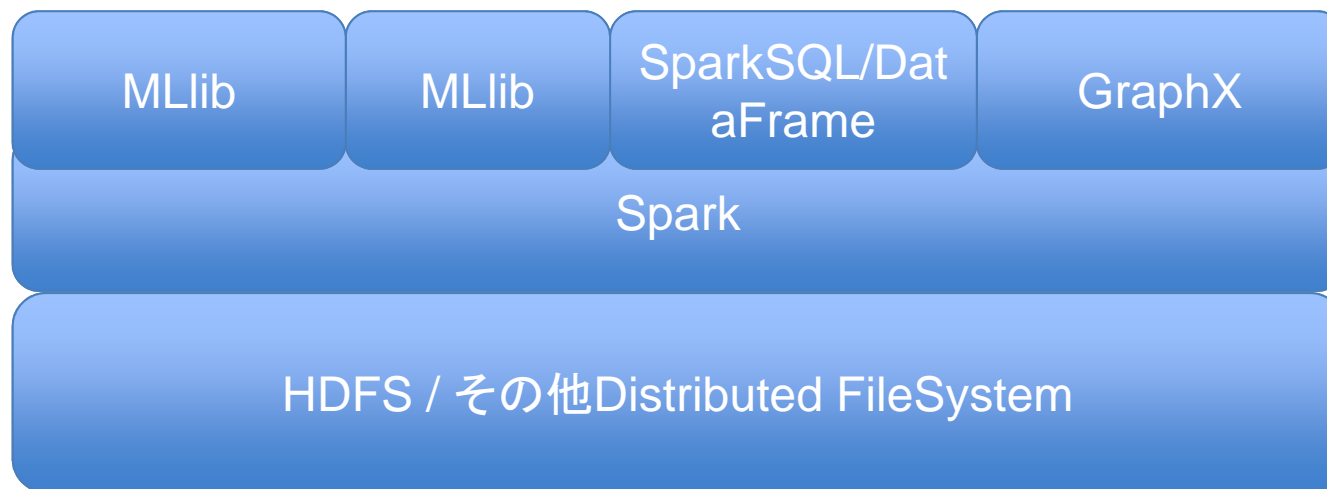
2.DL学習・テスト



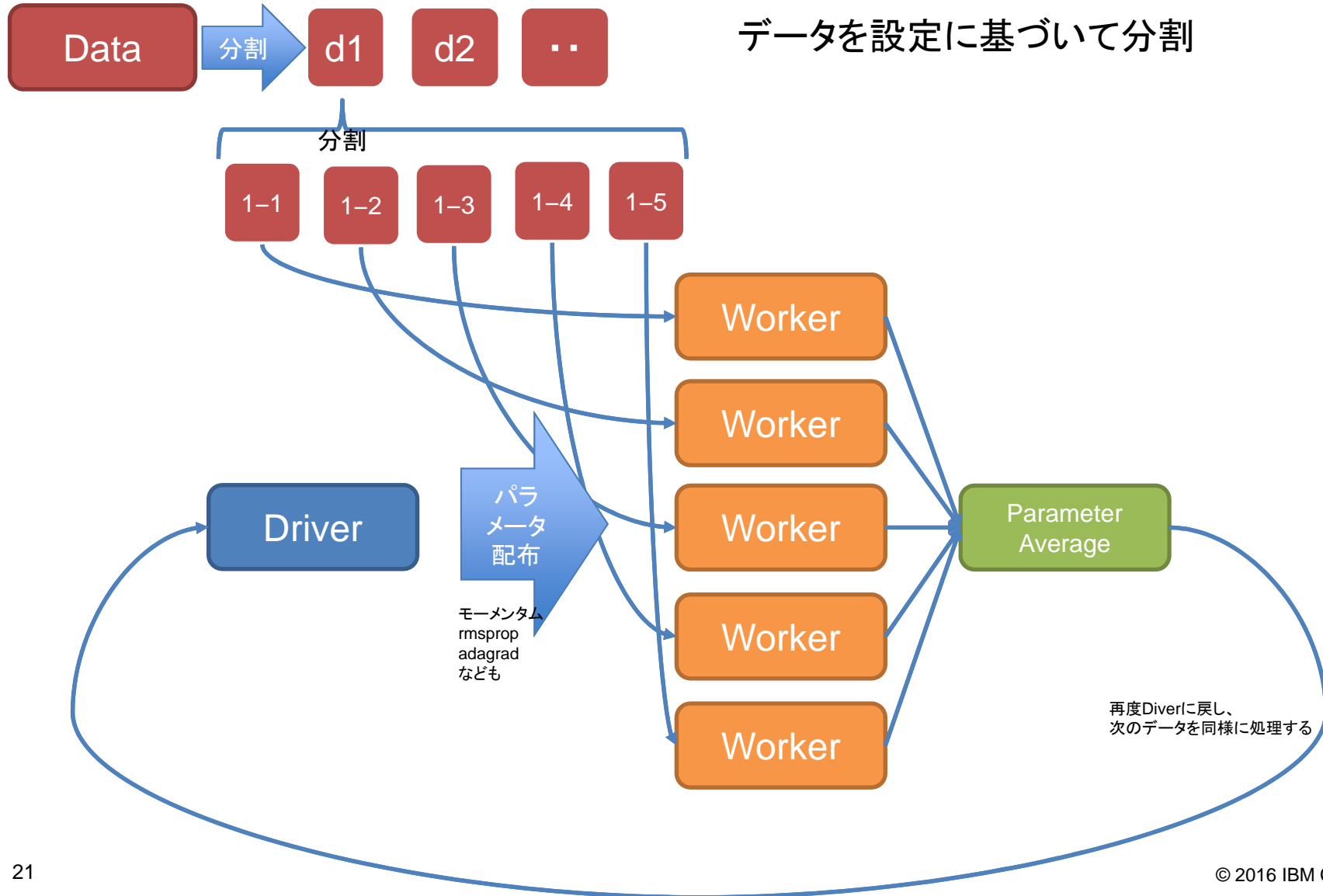
参考:

<https://github.com/yahoo/TensorFlowOnSpark>

昨今のDeepLearningFrameworkの動き



DeepLearning4Jを例にSpark上でのDLを見てみる



TensorflowOnSpark(TFoS)

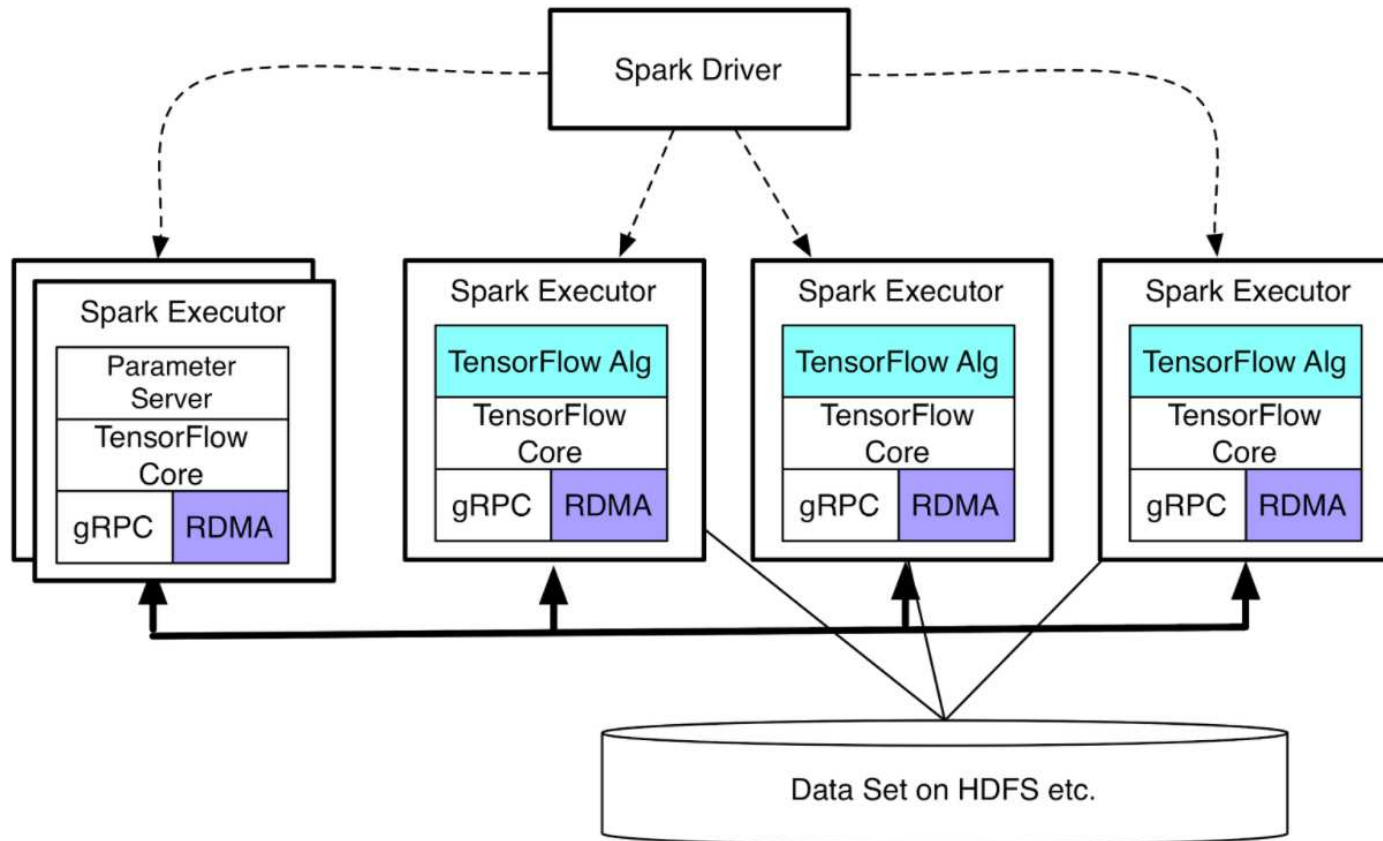


Figure 3: TensorFlowOnSpark system architecture

参考:
<https://github.com/yahoo/TensorFlowOnSpark>

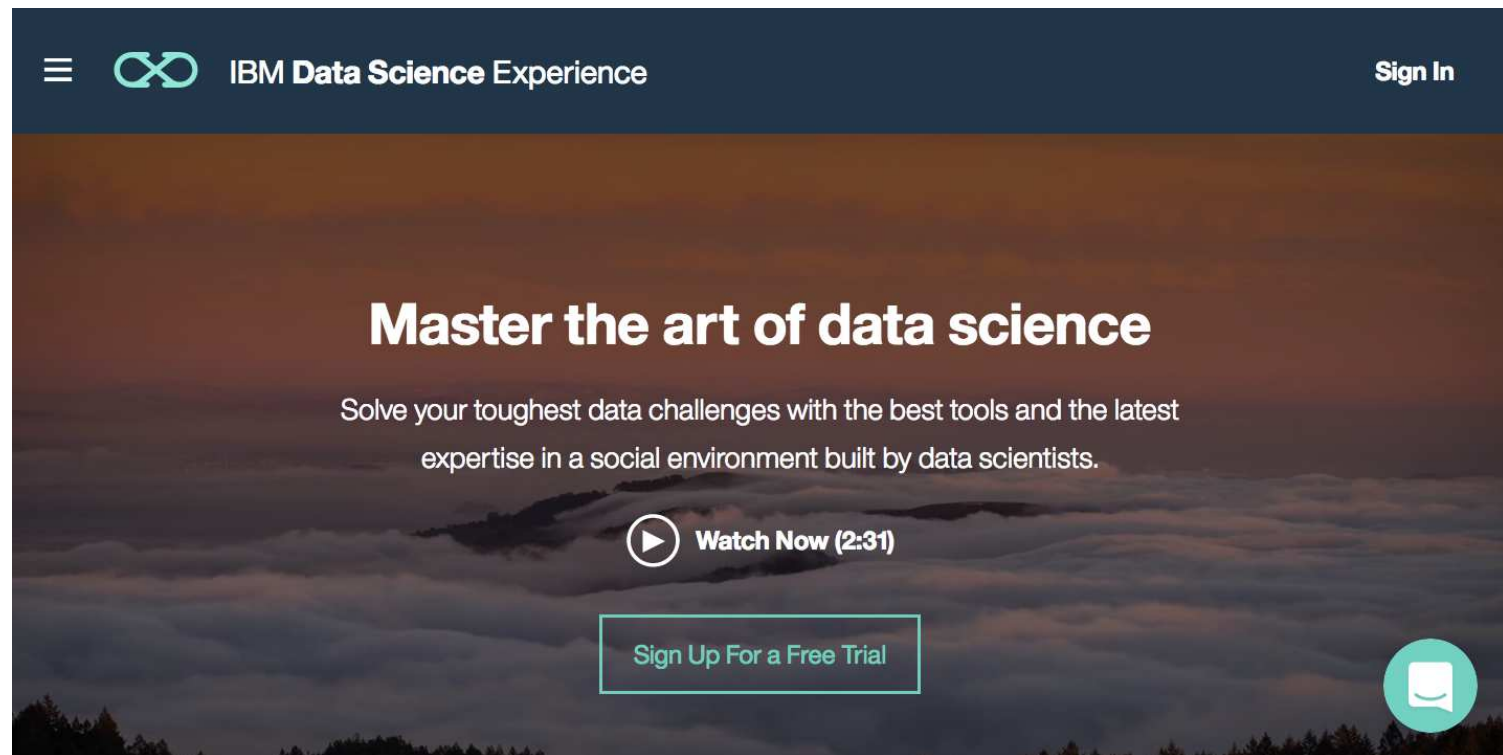
Spark + DeepLearningのまとめ


- MLlibだけでなく前処理やデータのつながりの部分でもSparkは重要
- DeepLearningを分散処理するための基盤としてSparkへの対応が進んでる
 - 深層学習だけでなく、「データサイエンス基盤」としてもSparkは重要
 - その他の機械学習の機能もSparkの対応が始められている
- 高速にDeepLearningを分散処理させるためにRDMAなどの対応もされ始めている

一家に一台のSparkの時代が来るか？

Data Science Experience


Data Science ExperienceはIBMが提供するCloud上の
Data Scientist, Data Engineerに向けたPlatform




☰  IBM Data Science Experience Sign In

Master the art of data science

Solve your toughest data challenges with the best tools and the latest expertise in a social environment built by data scientists.

 Watch Now (2:31)

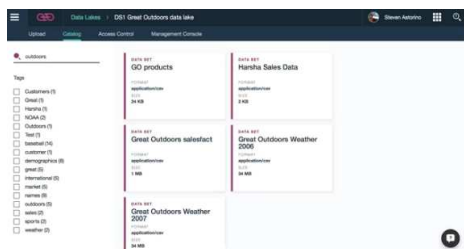
[Sign Up For a Free Trial](#)



Watson Data Platform の要素技術



データ・エンジニア

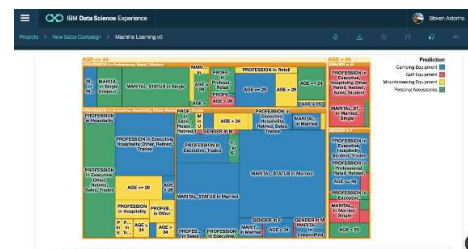


Data Connect

カタログ機能によりキーワードを入れて必要なデータを検索



データ・サイエンティスト

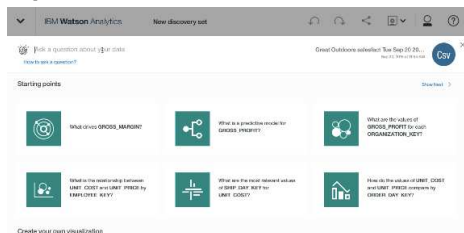


Data Science Experience (DSX) IBM Machine Learning

データ・サイエンティストが分析モデルを作成、顧客ベースをセグメンテーション



ビジネス・アナリスト

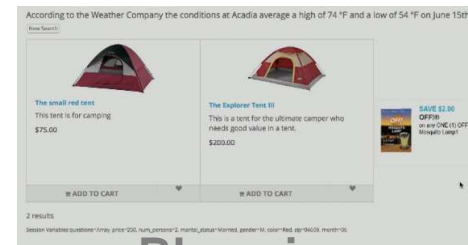


Watson Analytics

同じデータアセットをビジネス・アナリストが自然言語の入力で探索・分析



アプリケーション開発者



Bluemix

アプリケーション開発者がBluemix上に構築した、特定顧客セグメント向けのクーポンを表示するWebアプリ

プロジェクト・コミュニティ機能

コントロール

コメントセル 正しいデータ (教師データ) をグラフ化します

コードセル

```
In [4]: codeauth_p = codeauth_df.toPandas()  
%brunel data('codeauth_p') x(VOLUME) y(WEIGHT) color(AUTH) :: width=800, height=400
```

グラフ

Out[4]:

データベースとオブジェクト・ストレージ