

IT技術センター 2020年度技術発表会



全集中！！ 大石がビッグデータを斬る
～無限データ編～

AI製品技術部

チーム名：大石率いるAI軍団

遠藤、大石、星、青島、田澤、長谷川 with 金シ部門

SCSK

夢ある未来を、共に創る。

1. テーマ選定の背景
2. H2O Sparkling Waterについて
3. システム構成
4. 活動内容
5. 結果
6. まとめ、所感

1. テーマ選定の背景

1. テーマ選定の背景

AIを使うには？

→ ビッグデータの活用が有効

ビッグデータはあるが…

→ ビッグデータの加工、分析などが難しい

→ 次世代のビッグデータ機械学習基盤が必要

○今年度のAI製品技術部の取り組み

1. Cloudera + H2O + Apache Sparkによる

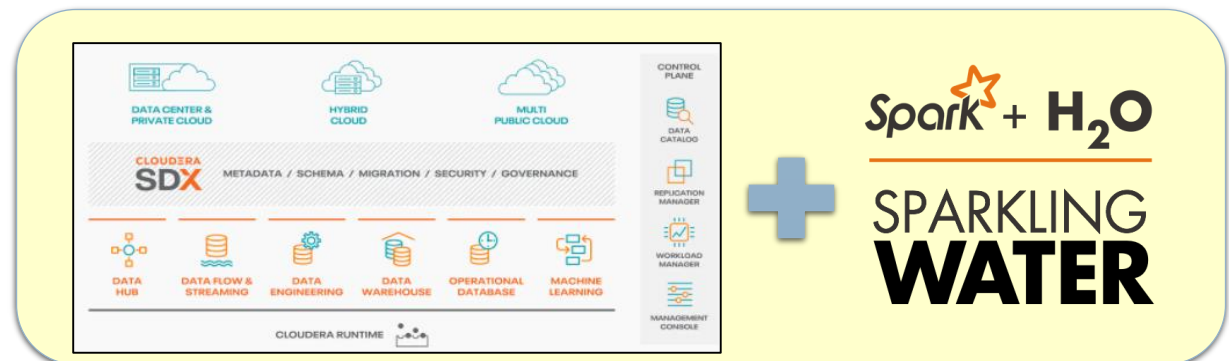
「次世代ビッグデータ機械学習基盤」を構築

2. 「機械学習ツール」と「次世代ビッグデータ機械学習基盤」の
速度検証・評価を実施



機械学習ツール

VS



次世代ビッグデータ機械学習基盤

1. テーマ選定の背景

○メンバー紹介

Clouderaについてはノウハウが無かったため、金融部門へ協力を依頼し、共同で本テーマに取り組んだ。



○AI製品技術部

長谷川 峰子 課長
遠藤 宏
星 雅人
青島 健太
大石 隆之
田澤 圭祐



○金融部門

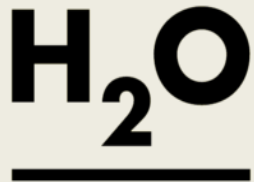
長濱 啓文 課長
武藤 賢司
高岸 大路

2. H2O Sparkling Waterについて

2. H2O SparklingWaterとは

H2O.ai社の製品群

オープンソース

The logo for H2O, featuring the chemical formula H₂O in a bold, black, sans-serif font, with a horizontal line underneath the O.

インメモリ型の
機械学習アルゴリズム

The logo for Spark + H2O Sparkling Water. It features the word "Spark" in a black, sans-serif font, followed by a small orange star icon and the text "+ H₂O". Below this, the words "SPARKLING" and "WATER" are stacked in a bold, black, sans-serif font.

Sparkを統合したH2O AI

- 100%オープンソース
- データサイエンティスト向け

The logo for H2O.ai Driverless AI. It features the text "H₂O.ai" in a small, yellow, sans-serif font, followed by the words "DRIVERLESSAI" in a large, bold, yellow, sans-serif font.

特徴量抽出を内部で自動処理,
機械学習と解釈可能性

- エンタープライズソフトウェア
- データ取込みから、モデル展開まで完全自動化
- GUIベースであらゆるレベルの分析者、データサイエンティストが使用可能

2. H2O SparklingWaterとは

Spark + H2O = Sparkling Water (*nice* 命名センス！！)

Sparkling Water は、“Apache Spark” と “H2O” を統合したものです。



- 大規模データを処理するための分析エンジン
- SparkのMLlib(機械学習エンジン)
 - ✓ 強力なデータ変換機能
 - ✓ NLP(自然言語処理)に適したアルゴリズム



- 分散型インメモリ機械学習プラットフォーム
- 高度なアルゴリズム
 - ✓ 計算速度、精度
 - ✓ 高度なパラメータ設定
- 分散および並列処理
- R & Python のインタフェース

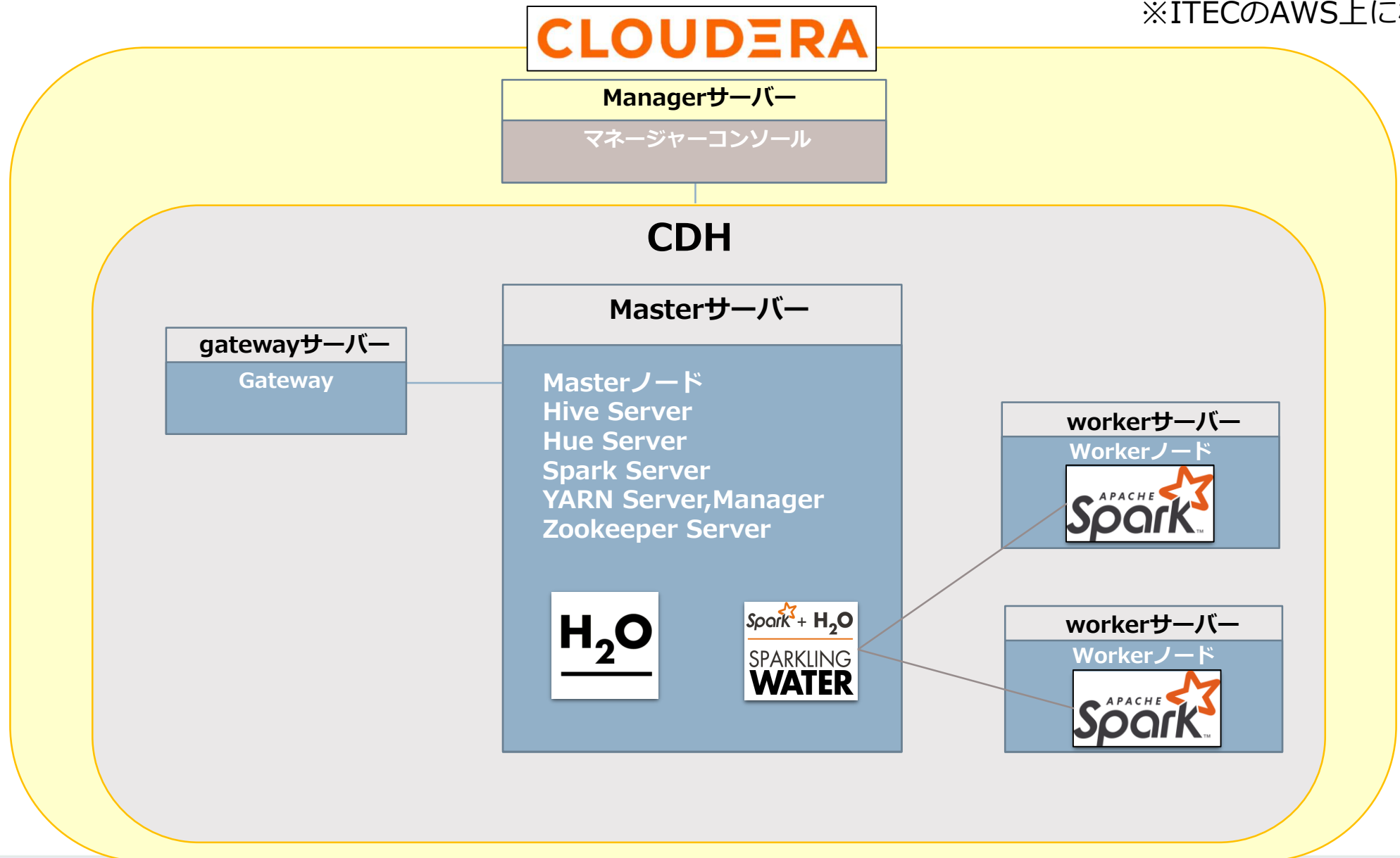
両プロダクトのメリットを活用し、「大規模データ」を「早く」「精度高く」分析可能

3. システム構成

3. システム構成

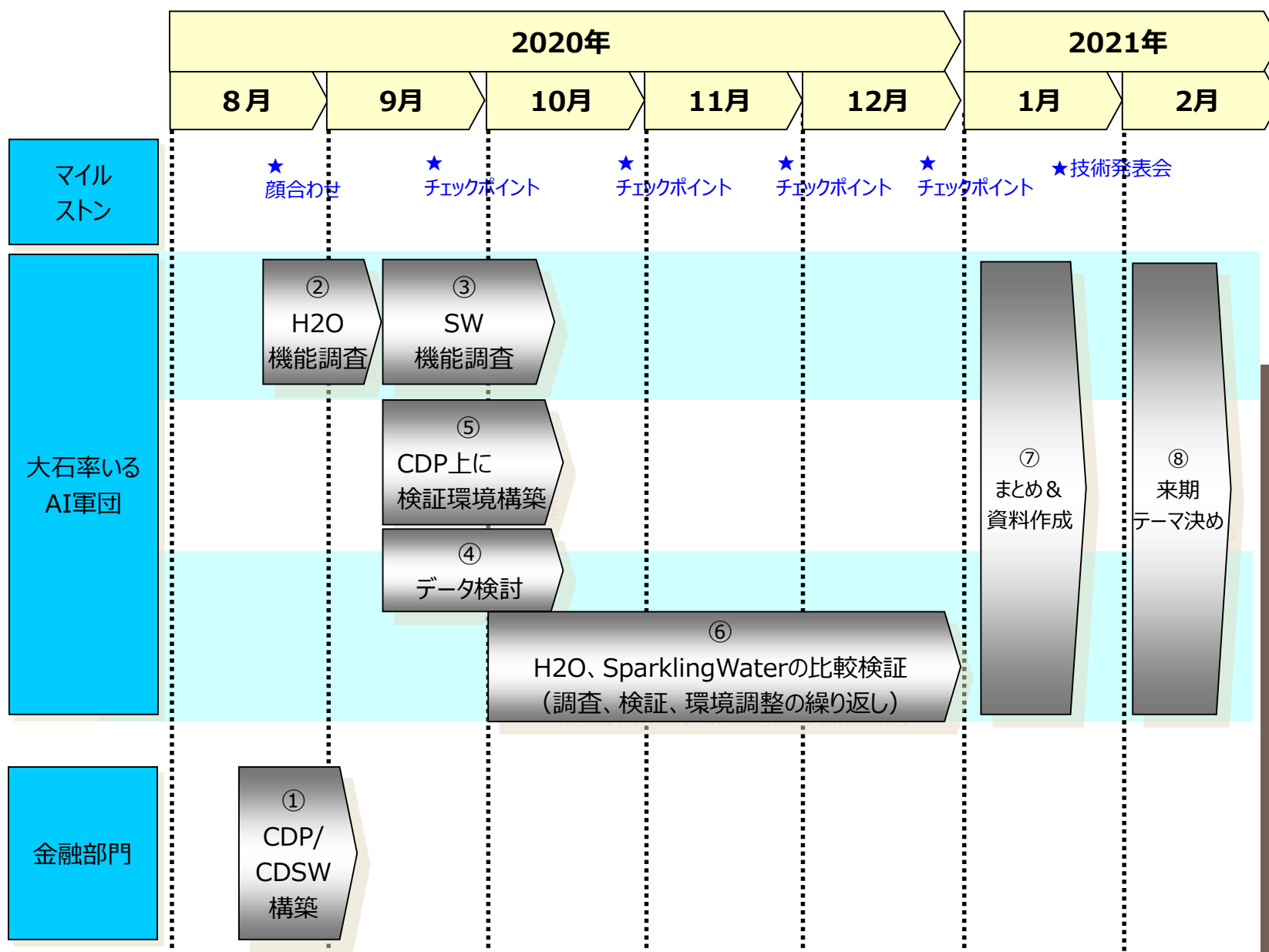
今回構築したシステムは下図のようにサーバー5台で構成となっている。

※ITECのAWS上に構築



4. 活動内容

4. 活動内容 ～活動履歴～



○実施・検証項目

- ① CDP/CDSWの構築と検証
- ② H2O(OSS)検証@ローカル環境
- ③ SparklingWater(OSS)検証 @ローカル環境
- ④ 検証に使用するデータ検索 対象：大容量データ
- ⑤ H2O/SparklingWater 検証環境構築@CDP
- ⑥ 検証活動 H2O単体/SparklingWater
- ⑦ まとめ&資料作成
- ⑧ 今後の活動について

4. 活動内容 ～コミュニケーションに活用したツール～

Teamsの投稿を活用して連絡事項共有、Wiki機能を活用した技術情報共有



The screenshot shows a Microsoft Teams interface. On the left, a chat window for 'cloudera' is visible, showing a message from '大石 隆之' dated 2020/11/11 13:10. The main part of the screen displays a Wiki page titled 'Sparkling Water' by '高岸 大路', dated 2020/12/10. The page content includes a list of requirements for running Sparkling Water on Linux/OS X/Windows, such as Java 1.8+, Python 2.7+, and Spark 3.0. It also lists data sets like 'Fannie Mae' and 'HIGGS' with their respective URLs.

cloudera 投稿 ファイル Wiki +

大石 隆之 2020/11/11 13:10
高岸 大路
masterのサーバーですが検証中に、大容量データをアップロードした際、一度落ちてしまったらしく、コンソール上からのステータスチェックも「！」マークが出ている状態です。復旧方法について教えていただけますでしょうか。

11 件の返信、送信者: 大石 および 隆之

2020年12月9日

高岸 大路 2020/12/09 15:47

h2oai 投稿 ファイル Wiki +

12 件の返信

Sparkling Water
最終編集: 2020/12/10

要件

- Linux/OS X/Windows
- Java 1.8+
- Python 2.7+ For Python version of Sparkling Water (PySparkling)
- Spark 3.0 and `SPARK_HOME` shell variable must point to your local Spark installation

Fannie Mae の一戸建て住宅のローン運用実績データ セット

<https://docs.rapids.ai/datasets/mortgage-data>

航空路線のデータ セット

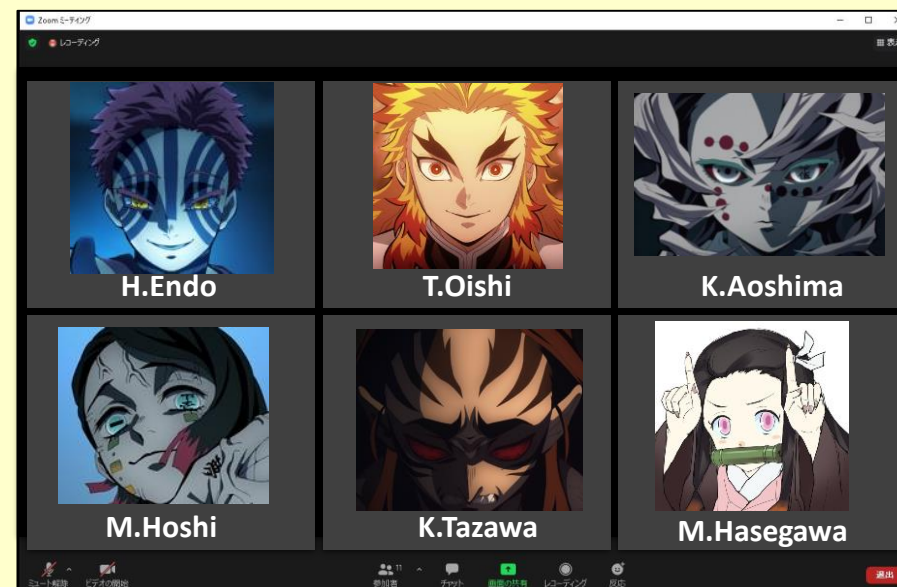
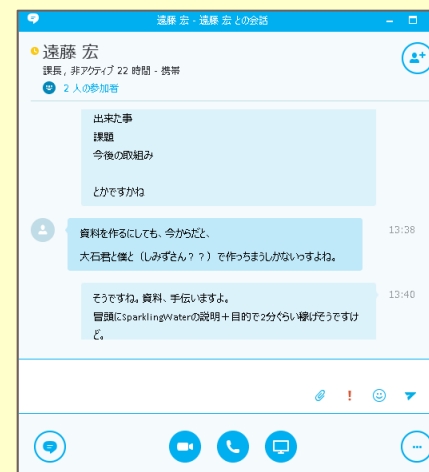
<https://github.com/h2oai/h2o-2/wiki/Hacking-Airline-DataSet-with-H2O>

HIGGS データ セット

<https://archive.ics.uci.edu/ml/datasets/HIGGS>

+

Zoom/Skypeを活用したコミュニケーション



5、 结果

H2O.aiとSparklingWaterの速度比較を実施

水柱

H₂O

インメモリ型の
機械学習アルゴリズム

VS

雷柱

Spark[☆] + H₂O

SPARKLING
WATER

Sparkを統合したH2O AI

第一回戦



航空会社のデータセット(2,000行 4.5MB)

データは次のとおりです。航空会社のデータセット データは元々RITAからのものであり、詳細に説明されています。
データは次のようになります。

	名前	説明
1	年	1987-2008
2	月	1~12
3	DayofMonth	1-31
4	曜日	1(月曜日)-7(日曜日)
5	DepTime	実際の出発時間(ローカル、hhmm)
6	CRSDepTime	出発予定時刻(現地時間、hhmm)
7	ArrTime	実際の到着時間(ローカル、hhmm)
8	CRSArrTime	到着予定時刻(現地時間、hhmm)
9	UniqueCarrier	<u>固有のキャリアコード</u>
10	FlightNum	フライトナンバー
11	TailNum	飛行機の尾翼番号
12	ActualElapsedTime	分で
13	CRSElapsedTime	分で
14	AirTime	分で
15	ArrDelay	到着遅延(分単位)

	名前	説明
16	DepDelay	出発遅延(分単位)
17	原点	出発地 <u>IATA空港コード</u>
18	目的地	目的地 <u>IATA空港コード</u>
19	距離	マイル単位
20	TaxiIn	時間内のタクシー、分単位
21	TaxiOut	分単位のタクシー乗車時間
22	キャンセル	フライトはキャンセルされましたか?
23	CancellationCode	キャンセルの理由(A=運送業者、B=天気、C=NAS、D=セキュリティ)
24	流用	1=はい、0=いいえ
25	CarrierDelay	分で
26	WeatherDelay	分で
27	NASDelay	分で
28	SecurityDelay	分で
29	LateAircraftDelay	分で

第一回戦

4. 5MB程度の
スモールデータでは、
当然SparklingWaterの
真価は発揮できない！！

引用元 https://www.google.com/imgres?imgurl=https%3A%2F%2Fup.gc-img.net%2Fpost_img%2F2020%2F03%2F874NdANuesWXEkQ_KIPSG_16.gif&imgrefurl=https%3A%2F%2Fgirlschannel.net%2Ftopics%2F2625557%2F&tbid=ndnLzOz06d5SPM&vet=12ahUKEwj9m6uvzafuAhVF9pQKHSwOCm0QMygOegUIARDxAQ..i&docid=ygQRGYyCzUCb3M&w=400&h=225&q=%E9%9C%B9%E9%9D%82%E4%B8%80%E9%96%83%20GIF&ved=2ahUKEwj9m6uvzafuAhVF9pQKHSwOCm0QMygOegUIARDxAQ

第二回戦

HIGGSデータセット(11,000,000行 3.6GB)

ソース:

ダニエル ホワイトソングダニエル'@'uci.edu、カリフォルニア大学物理学・天文学助教授 カリフォルニア大学アーバイン校

データセット情報:

データは、モンテカルロシミュレーションを使用して作成されています。最初の21個の特徴(列2~22)は、加速器の粒子検出器によって測定された運動学的特性です。最後の7つの機能は、最初の21の機能の機能です。

これらは、2つのクラスを区別するために物理学者によって導出された高レベルの機能です。

物理学者がそのような機能を手動で開発する必要をなくすために、ディープラーニング手法を使用することに関心があります。

標準の物理パッケージのベイジアンディシジョンツリーと5層ニューラルネットワークを使用したベンチマーク結果は、元の論文に示されています。

最後の500,000の例は、テストセットとして使用されます。

属性情報:

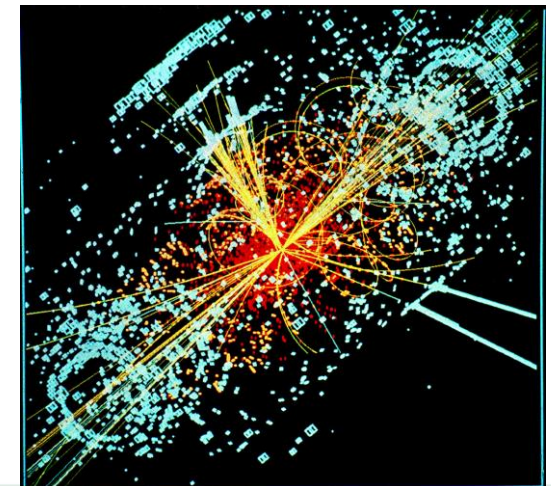
最初の列はクラスラベル(信号の場合は1、バックグラウンドの場合は0)で、その後28の機能(21の低レベル機能、次に7つの高レベル機能)が続きます:

レプトン p_T 、レプトン η 、レプトン ϕ 、エネルギーの大きさが欠落、欠落エネルギーファイ、ジェット1 p_T 、ジェット1 η 、ジェット1 ϕ 、ジェット1 b タグ、

ジェット2 p_T 、ジェット2 η 、ジェット2 ϕ 、ジェット2 b タグ、ジェット3 p_T 、ジェット3 η 、ジェット3 ϕ 、jet 3 b -tag、jet 4 p_T 、jet 4 η 、jet 4 ϕ 、jet 4 b -tag、

m_{jj} 、 m_{jjj} 、 m_{lv} 、 m_{jlv} 、 m_{bb} 、 m_{wbb} 、 m_{wwbb} 。各機能の詳細については、元の論文を参照してください。

質量を与える神の粒子



第二回戦

3. 6GBのデータでも、
SparklingWaterの真価は
発揮できない！？

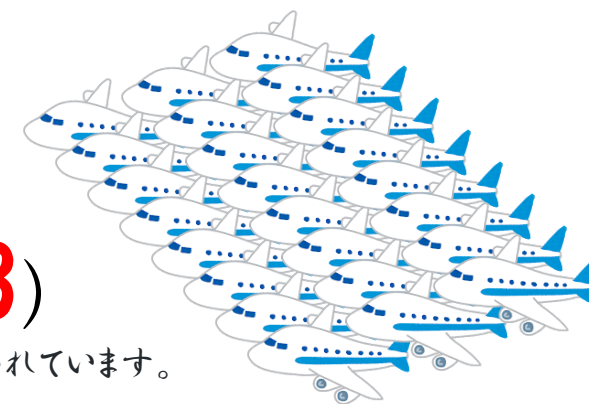
引用元

<https://www.google.com/url?sa=i&url=https%3A%2F%2Fsteamcommunity.com%2Fsharedfiles%2Ffiledetails%2F%3Fid%3D1834788871&psig=AOvVaw199twlZ6e9im81ILMK3vK0&ust=1611132221663000&source=images&cd=vfe&ved=0CAIQjRxqFwoTCLjCkrXNp-4CFQAAAAAdAAAAABAP>

第参回戦

航空会社のデータセット(113,308,324行 12GB)

データは次のとおりです。航空会社のデータセット データは元々RITAからのものであり、詳細に説明されています。
データは次のようになります。



	名前	説明
1	年	1987-2008
2	月	1~12
3	DayofMonth	1-31
4	曜日	1(月曜日)-7(日曜日)
5	DepTime	実際の出発時間(ローカル、hhmm)
6	CRSDepTime	出発予定時刻(現地時間、hhmm)
7	ArrTime	実際の到着時間(ローカル、hhmm)
8	CRSArrTime	到着予定時刻(現地時間、hhmm)
9	UniqueCarrier	<u>固有のキャリアコード</u>
10	FlightNum	フライトナンバー
11	TailNum	飛行機の尾翼番号
12	ActualElapsedTime	分で
13	CRSElapsedTime	分で
14	AirTime	分で
15	ArrDelay	到着遅延(分単位)

	名前	説明
16	DepDelay	出発遅延(分単位)
17	原点	出発地 <u>IATA空港コード</u>
18	目的地	目的地 <u>IATA空港コード</u>
19	距離	マイル単位
20	TaxiIn	時間内のタクシー、分単位
21	TaxiOut	分単位のタクシー乗車時間
22	キャンセル	フライトはキャンセルされましたか?
23	CancellationCode	キャンセルの理由(A=運送業者、B=天気、C=NAS、D=セキュリティ)
24	流用	1=はい、0=いいえ
25	CarrierDelay	分で
26	WeatherDelay	分で
27	NASDelay	分で
28	SecurityDelay	分で
29	LateAircraftDelay	分で

第参回戦

エラー

! ?

12GBだとサーバの
物理メモリが足りない!!

allocate

mem

6. まとめ、所感

6. まとめ、所感

①H2O.ai

強点：スモールデータならとことん早い。

弱点：データサイズ分のメモリを1サーバに搭載する必要がある。
(スケールアップしかできないため1サーバの搭載メモリが処理量の限界)

②SparklingWater

強点：どんなにビッグデータであろうが、複数サーバにまたがってメモリを分散できる
(スケールアウトができる)

弱点：スモールデータだと遅い

③どっちを使うべきか

今回の検証では、学習データが少なくとも3.6GB以下の場合はH2O.aiの方が早い。

今回の反省点と次回への取り組み

①今回は欲張り過ぎた

- ・全てが初めて触れるソフトウェアだった。
- ・Cloudera製品をフルセットで環境を構築したため、システム構成が複雑過ぎた。
- ・エラーが発生してもどこが問題なのかの特定が困難だった。
- ・サーバーを5台以上で構成する必要があり、リソースを逼迫していた。
- ・正常に動作させるだけで精いっぱい、結果検証までできなかった。

②次回はシンプルな環境で検証したい

- ・Spark+Yarnのみで環境を構築し、適切なリソース割り当てを行いたい。
- ・分散学習可能なアルゴリズムや分析結果の考察までやりたい。
- ・H2O Driverless AIで作成したAIモデルをSparklingWaterで動かしたい。



こんな事で俺の情熱は
なくなったりしない！！
心の炎は消える事はない！！
決して俺は挫けない！！
来年に向けてチャレンジだ！！