

MATLABによる大規模フリートデータ解析

Part 2: クラスター編

MathWorks Japan

アプリケーションエンジニアリング部

齊藤 甲次郎

アジェンダ

- フリートデータ解析実践
 - クラスターへのスケールアウト

ビッグデータの扱い フリートデータ解析 サマリー

ステップ1



デスクトップPCでの解析

ステップ2



Hadoop® / Spark™

Hadoopクラスターでの解析

フリートデータ解析実践

クラスターへのスケールアウト

- Hadoop
 - ビッグデータのためのプラットフォーム
 - **HDFS** : 分散ファイルシステム
 - **YARN** : リソース管理やスケジューリング
 - **Hadoop MapReduce**: MapReduceフレームワークの実装
- Spark
 - クラスター計算のフレームワーク
 - フォールトトレラントな分散処理
 - **機械学習**のような繰り返し演算に適性あり
 - **Hadoop YARN**と連携可能

フリートデータ解析実践 クラスターへのスケールアウト

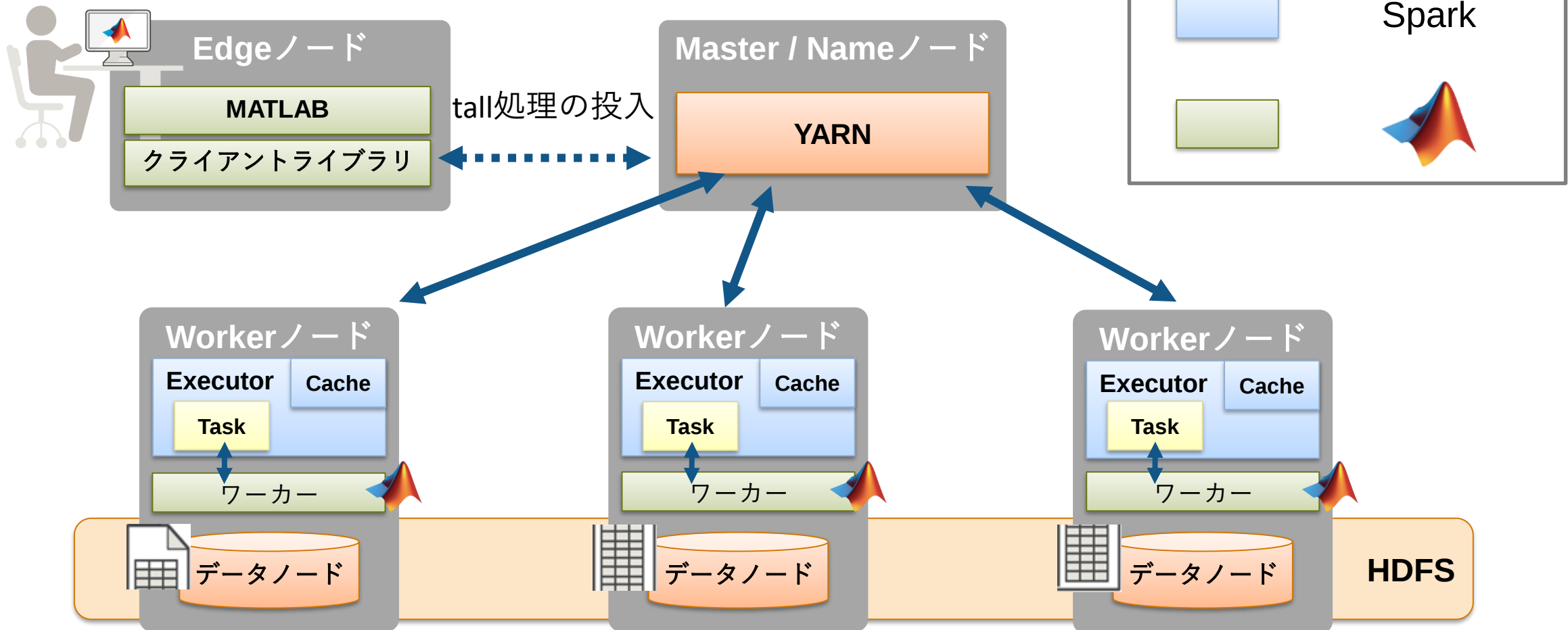
■ MATLABとHadoop/Sparkの2つの連携方法

#	用途	使用する製品	対象ユーザー
1	MATLABからtall配列の処理をSparkで実行	MATLAB Parallel Server™ 	試行錯誤でアルゴリズムを検討するユーザー
2	- tall配列の処理をSpark上でスタンドアロンアプリで実行 - Spark APIを使ったスタンドアロンアプリの実行	MATLAB Compiler™  	定形処理を利用するユーザー - SparkのAPIを細かく制御した処理を作成したいユーザー

※MATLAB Distributed Computing Server™は、R2019aからMATLAB Parallel Server™に名称が変わりました。

フリートデータ解析実践 クラスターへのスケールアウト

- MATLAB Parallel Serverを使ったHadoop/Spark連携



フリートデータ解析実践 クラスターへのスケールアウト

MATLABデスクトップでの処理

個々のトリップファイルの読み取り

```
dataDir = fullfile('..', 'LogFiles', '*.mat');
```

fileDatastoreを使ったカスタム読み込み関数での読み取り

```
readFcn = @(filename) readTrip(filename, tripTable);  
fds = fileDatastore(dataDir, 'ReadFcn', readFcn, 'UniformRead', true);
```

tall配列の作成

```
tt = tall(fds);
```

クラスターの設定とファイルパスの変更だけで
MATLABの処理は書き直さずにスケールアウト

Hadoop/Sparkクラスターでの処理

Hadoopクラスターの設定

```
setenv('HADOOP_HOME', '/usr/local/hadoop/hadoop-2.9.1');  
setenv('HADOOP_PREFIX', getenv('HADOOP_HOME'));  
setenv('SPARK_HOME', '/usr/local/spark/spark-2.3.2-bin-hadoop2.7');  
cluster = parallel_cluster.Hadoop;
```

MDCSインストールディレクトリ及びクラスターに渡すスクリプトの指定

```
cluster.ClusterMatlabRoot = '/usr/local/MATLAB/MDCS/R2018b';  
cluster.AttachedFiles = {'readTrip.m'};
```

Hadoopクラスターをtall配列の実行環境として設定

```
mapreducer(cluster);
```

個々のトリップファイルの読み取り

```
dataDir = 'hdfs://hadoopmaster:54310/Fleet/*.mat';
```

fileDatastoreを使ったカスタム読み込み関数での読み取り

```
readFcn = @(filename) readTrip(filename, tripTable);  
fds = fileDatastore(dataDir, 'ReadFcn', readFcn, 'UniformRead', true);
```

tall配列の作成

```
tt = tall(fds);
```

フリートデータ解析実践

クラスターへのスケールアウト

Browsing HDFS - Mozilla Firefox

hadoopmaster:50070/explorer.html#/
All Applications Namenode information

Hadoop Overview Datanodes Datanode Volume Failures Snapshot Startup Progress Utilities

Browse Directory

/ Go! [Icons]

Show 25 entries Search: [Input]

☐	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	
☐	drwxr-xr-x	matlab	supergroup	0 B	Oct 29 05:08	0	0 B	Fleet	[Icon]
☐	drwxr-xr-x	matlab	supergroup	0 B	Oct 29 05:14	0	0 B	tmp	[Icon]
☐	drwxr-xr-x	matlab	supergroup	0 B	Oct 29 05:14	0	0 B	user	[Icon]

Showing 1 to 3 of 3 entries Previous 1 Next

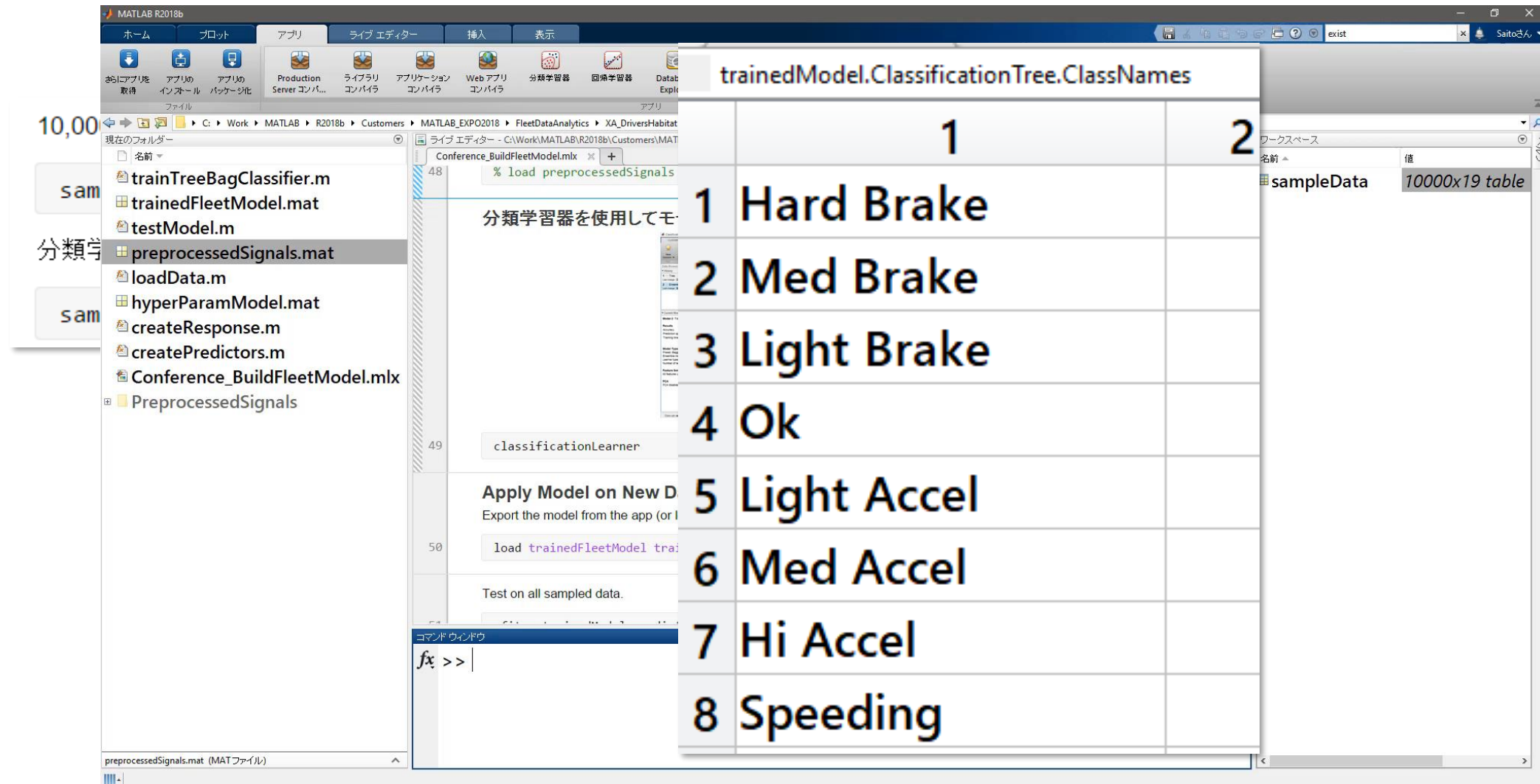
Hadoop, 2018.

フリーデータ解析実践

走行データ解析

分類学

分類学習器でドライブの8クラスのモデルを作成



フリートデータ解析実践

走行データ解析

- Hadoop/Sparkクラスターで分類

学習済みモデルの読み込み

```
trainedModel = load('trainedFleetModel');  
trainedModel = trainedModel.trainedModel;
```

分類学習器で作成した
モデルを読み込み

データの読み込み

```
dataDir = 'hdfs://hadoopmaster:54310/PreprocessedSignals/*.mat';  
ds = fileDatastore(dataDir, 'ReadFcn', @loadData, ...  
    'UniformRead', true)  
tt = tall(ds);
```

時刻の表示を英語表記に変更

```
tt.time.Format = 'default';
```

モデルに適用するためにtableに変換

```
tt_all = timetable2table(tt);
```

分類モデルにtall配列を
適用

モデルの適用

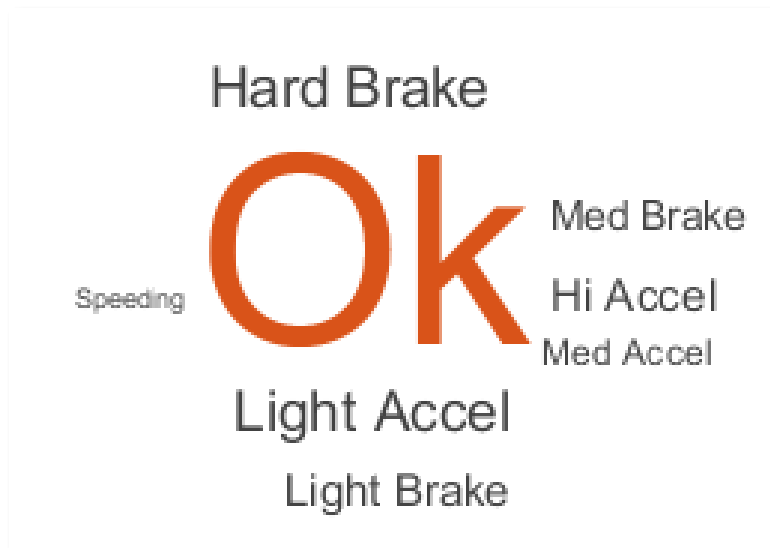
```
yfit = trainedModel.predictFcn(tt_all);  
yfitGathered = gather(yfit);
```

フリートデータ解析実践

走行データ解析

分類結果をwordcloudでプロット

ドライバー1



ドライバー2



ドライバー3



ドライバー5



ドライバー4

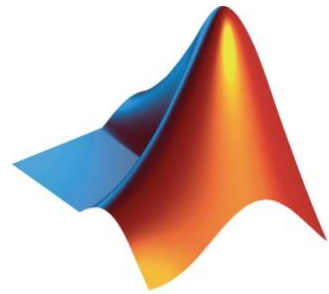


まとめ

Key Takeaways

1. 大規模なフリートデータを **datastore** や **tall** を用いて簡単に処理
2. デスクトップPCからHadoop/Sparkクラスターへ、 **コードを書き直さずにスケールアウトが容易**

眠っているビッグデータを
MATLABで解析してみませんか？



MathWorks®

Accelerating the pace of engineering and science

© 2019 The MathWorks, Inc. MATLAB and Simulink are registered trademarks of The MathWorks, Inc. See www.mathworks.com/trademarks for a list of additional trademarks. Other product or brand names may be trademarks or registered trademarks of their respective holders.