Real-World Example: Sports Analytics

- Too much data to handle and capture it
- Difficult to predict
- Real-Time dependence
Big data workflow: from desktop to production

ACCESS DATA

PROCESS ON DESKTOP
- Visualization
- Preprocessing
- Machine Learning

SCALE PROBLEM SIZE

MATLAB EXPO 2018
So, what’s the big (data) challenges?

- Standard tools won’t work
- Time-consuming
- Need to learn new tools & rewrite algorithms
Solution!

- Standard tools won’t work
- Time-consuming
- Need to learn new tools & rewrite algorithms

Prototype algorithms quickly
Run directly from MATLAB with **tall arrays**
Use the **same MATLAB code**
Datastore & tall arrays

1. Use datastore to define file-list
   >> ds = datastore('*.csv')

2. Create tall table from datastore
   >> tt = tall(ds)

3. Act like ordinary table in parallel
   >> model = fitlm(tt.Temp=...)

4. Request on local machine
   >> result = gather(tt.result)
Tall arrays: very small changes

1 file

Access Data
measured = readtable('PumpData.csv');
measured = table2timetable(measured);

Preprocess Data
Select data of interest
measured = measured(timerange(seconds(1),seconds(2)),:);

Work with missing data
measured = fillmissing(measured,'linear');

Calculate statistics
m = mean(measured.Speed);
s = std(measured.Speed);

---

1000+ files

Access Data
measured = datastore('PumpData.csv');
measured = tall(measured);
measured = table2timetable(measured);

Preprocess Data
Select data of interest
measured = measured(timerange(seconds(1),seconds(2)),:);

Work with missing data
measured = fillmissing(measured,'linear');

Calculate statistics
m = mean(measured.Speed);
s = std(measured.Speed);
[m,s] = gather(m,s);
Workflow Pattern

Access out of memory data  
datastore & tall

Work with subsets of your data  
findgroups, splitapply

Develop functions for event detection and calculation  
Normal MATLAB code

Apply functions to all of your data  
cellfun

Aggregate, summarize, & visualize  
table, histogram, heatmap, boxplot, binScatterPlot
MATLAB Distributed Computing Server (MDCS)

Desktop System

Local Parallel Computing

Parallel Computing Toolbox

Local Workers

Simulink, Blocksets, and Other Toolboxes

MATLAB

Computer Cluster

MATLAB Distributed Computing Server

Workers

Scheduler

Deployed Parallel Computing

MATLAB EXPO 2018
What is Hadoop/Spark?

- Machine Learning
- Fault tolerance
- Cluster-computing framework
- Data parallelism
Scaling with Spark: Very small changes too!

**Desktop Code**

```
mapreduce(gcp);
```

**Define the Execution Environment**

**Access Data**

```
measured = datastore('PumpData*.csv');
measured = tall(measured);
```

**Spark + Hadoop Code**

```
setenv('HADOOP_HOME', '/path/to/hadoop/install')
setenv('SPARK_HOME', '/path/to/spark/install');
cluster = parallel.cluster.Hadoop;
mapreduce(cluster);
```

**Define the Execution Environment**

**Access Data**

```
measured = datastore('PumpData*.csv');
measured = tall(measured);
```
Big Data with MATLAB & Spark

ACCESS DATA

datastore

PROCESS ON DESKTOP

Tall arrays
Subset of your data
Local parallel computing

SCALE PROBLEM SIZE

MATLAB EXPO 2018
The MathWorks Fleet Data

- 21 unique vehicles
- Data collected over 1.5 years
- 1300 trips
- log files
- 39 unique channels

Challenges

Big Data

Needle in the Haystack

Objectives

Testing Ideas

Event Detection

CO₂
Example Setup at MathWorks

- **Data Warehouse**
- **Server**
- **4G LTE**
- **Bluetooth**
- **Engineers**
Analyze fleet data with MATLAB
Access & Explore Data: *MATLAB & Spark*  
MathWorks Vehicle Fleet

**Challenge**  
Develop and deploy Data Analytics to run on Spark against vehicle fleet data stored on Hadoop

**Solution**  
Use MATLAB `tall` arrays to develop analytics on the desktop and then scale out to the Spark cluster

**Results**  
Developed insight and understanding of over 1300 vehicle trips  
Fuel efficiency performance under real-world driving conditions
Analysis Domains

**Statistics**
- Summary Statistics
- Regression, ANOVA, Machine Learning

**Signal Processing**
- Sound quality analysis
- LIDAR analysis

**Image Processing**
- Active Safety

**Location/Mapping**
- Analyzing GPS Data
- Custom Visualizations
Key Takeaways

- Use the **same MATLAB code**
- Use new MATLAB data types **datastore & tall arrays** for **out of memory** data sets
- **Scale** your work up with **Parallel Computing Toolbox** on the desktop or the **MATLAB Distributed Computing Server (MDCS)** on **Spark**