

Building a Big Engineering Data Analytics System using MATLAB

International Data Science Conference
12 June 2017



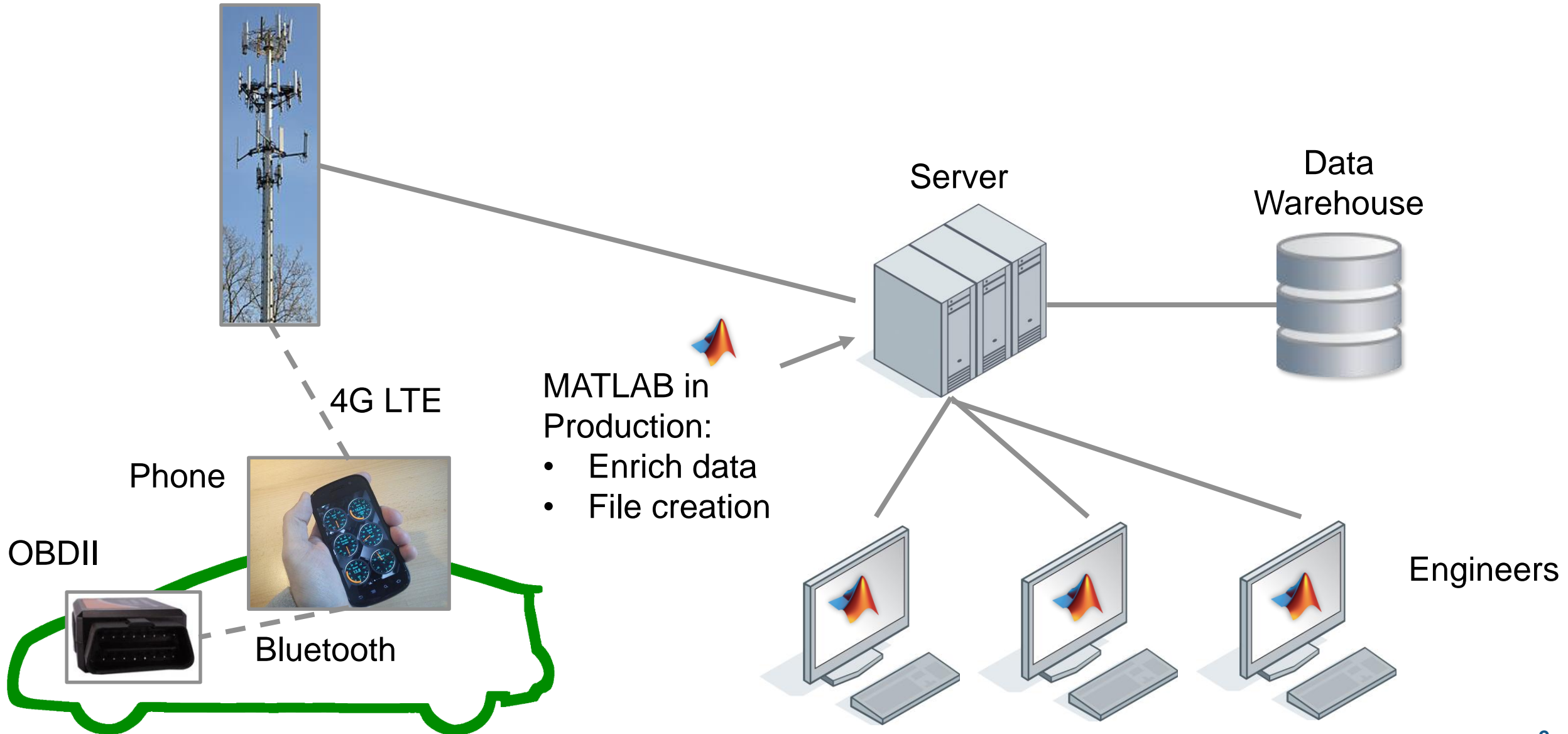
International
Data Science Conference
June 12th – 13th, 2017
Salzburg, Austria



Dmytro Martynenko – Application Engineering, The MathWorks GmbH

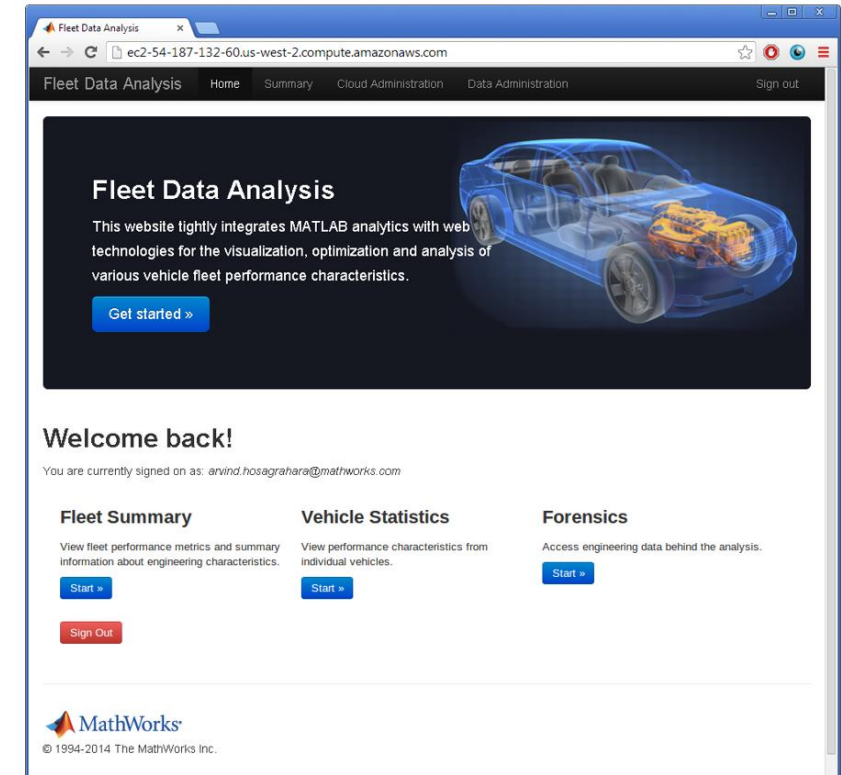
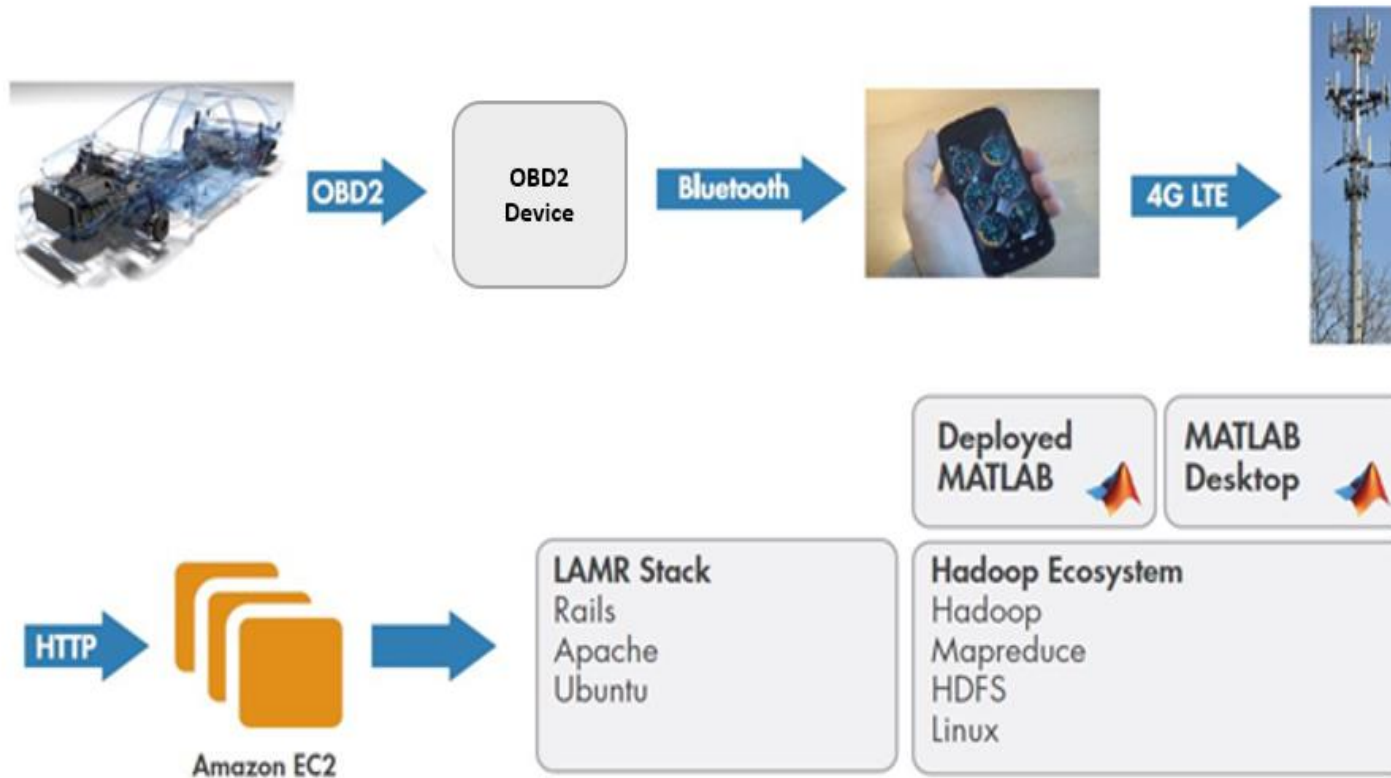
Dmitrij.Martynenko@mathworks.de

Example Setup at MathWorks



Fleet Data Analytics MathWorks Paper

[Engine Vehicle Design White Paper on mathworks.com](http://mathworks.com)



Customer Example: Scania

Automatic Emergency Braking

Opportunity

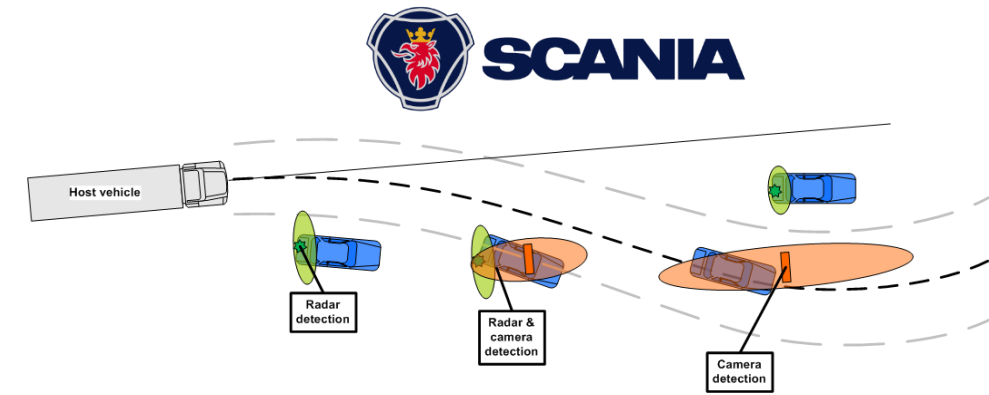
Real-time crash avoidance by detecting imminent collisions and automatically taking action

Analytics Use

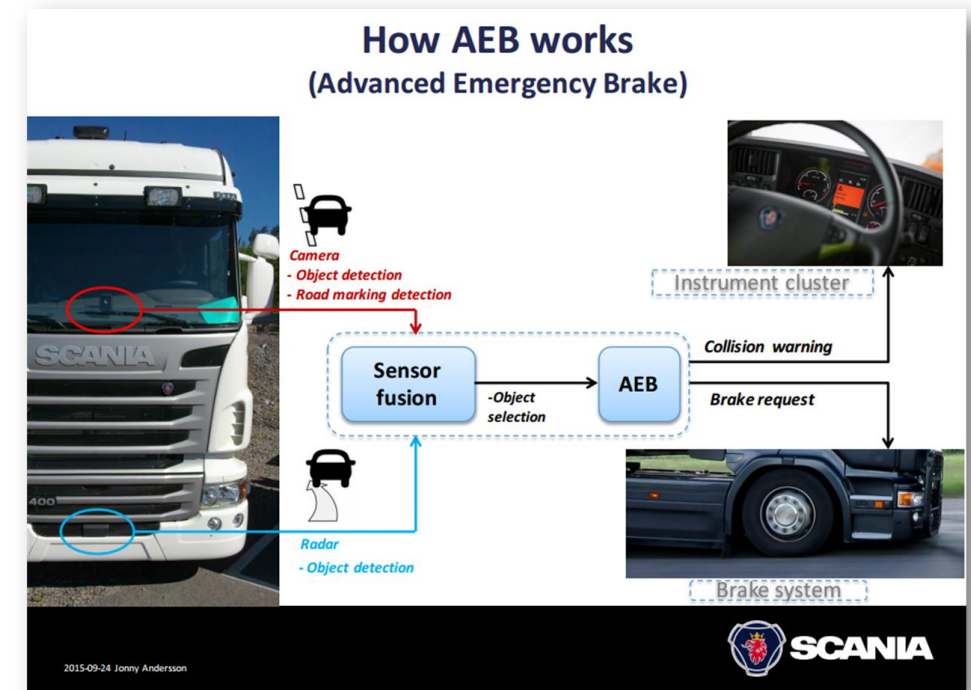
- **Data:** 80 TB – 1.5 million km of driving
- **Machine Learning:** Object detection
- **Control Systems:** Brake application
- **Test and Verify:** System model with simulated, recorded, and live data.

Benefit

- Reduced accidents
- Meet EU Regulations



Radar and camera for object-detection and real-time collision warning and braking.



Perception

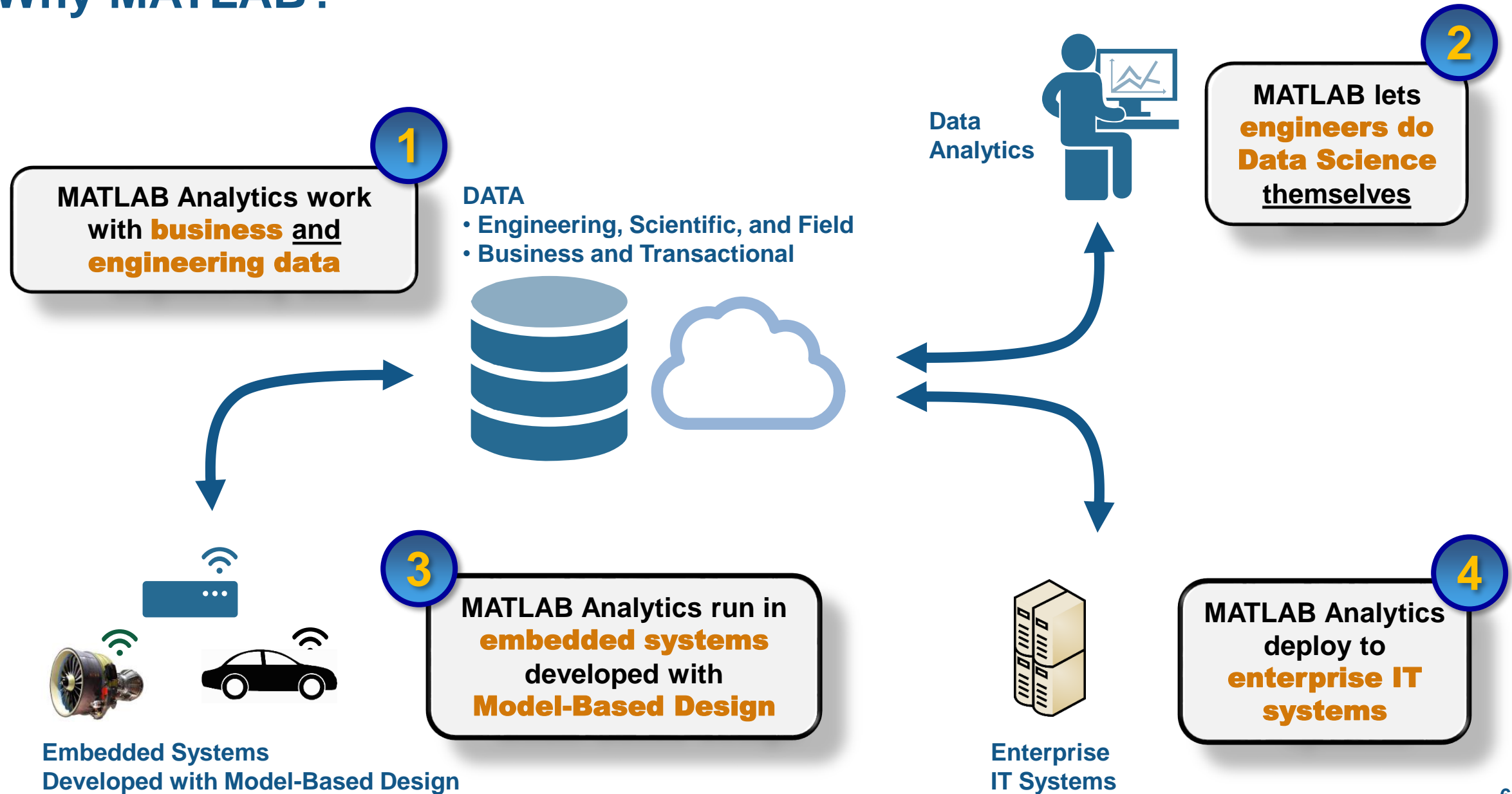
Prediction

Action

50 km/h - sudden brake



Why MATLAB?



Machine Learning Workflow



Files

Databases

Sensors

Working with Messy Data

Data Reduction/Transformation

Feature Extraction

Model Creation e.g. Machine Learning

Parameter Optimization

Model Validation

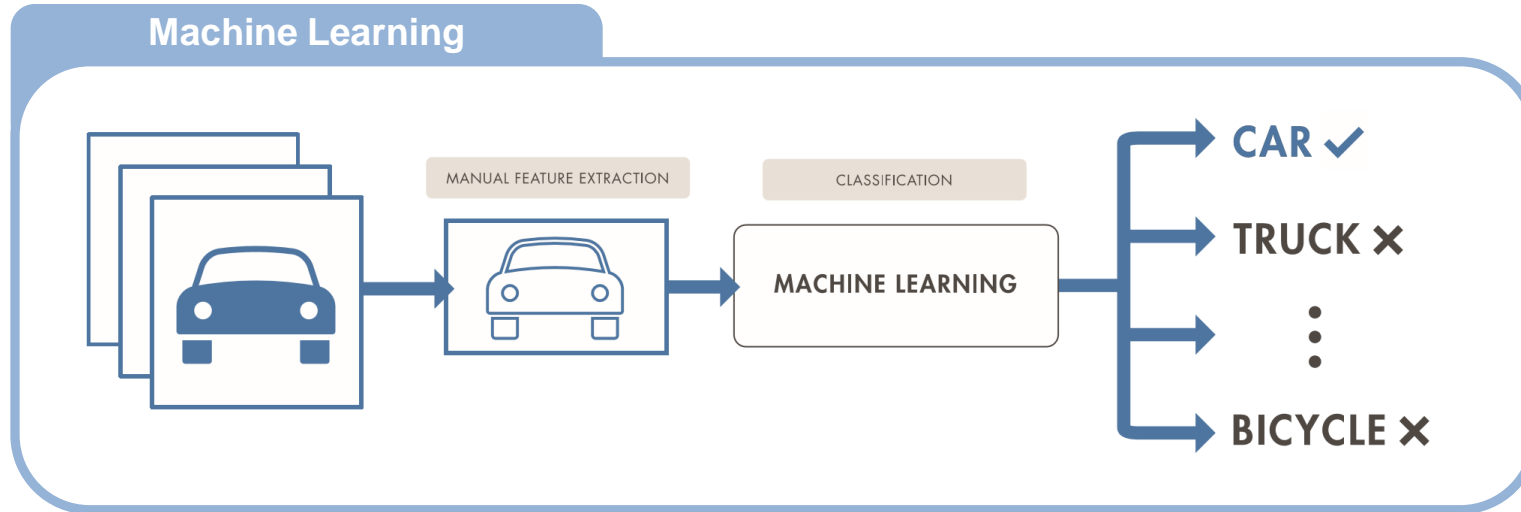
Desktop Apps

Enterprise Scale Systems

MATLAB Excel
 .NET C/C++
 .exe Java .dll

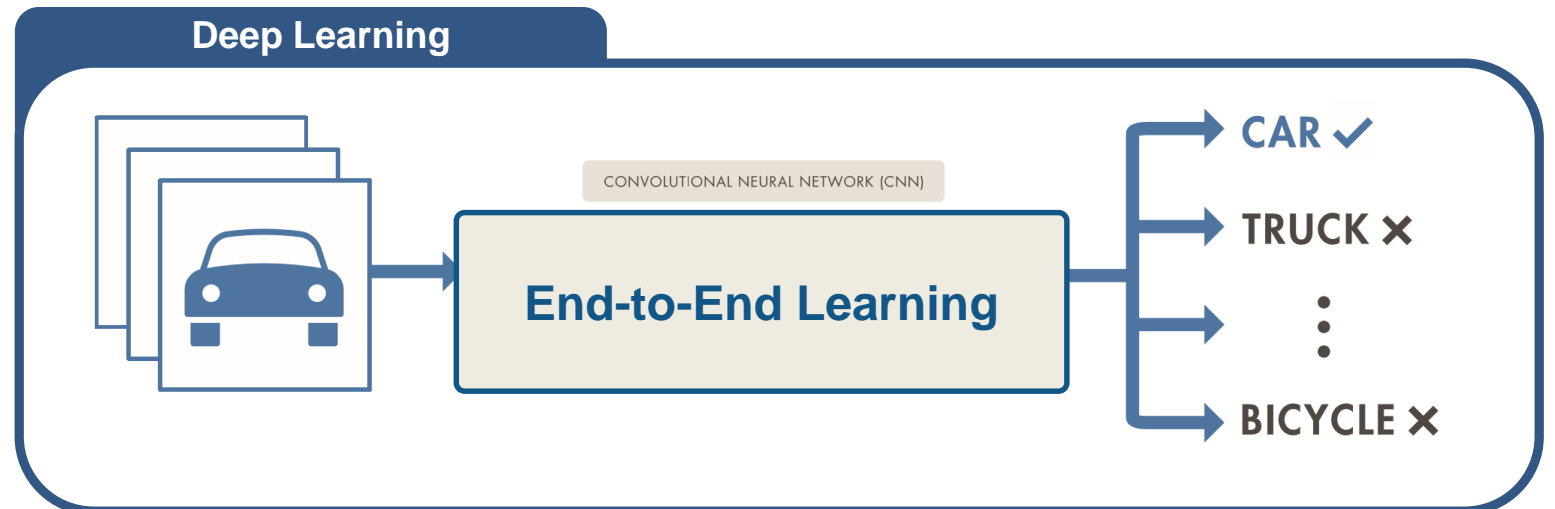
Embedded Devices and Hardware

What is Deep Learning?

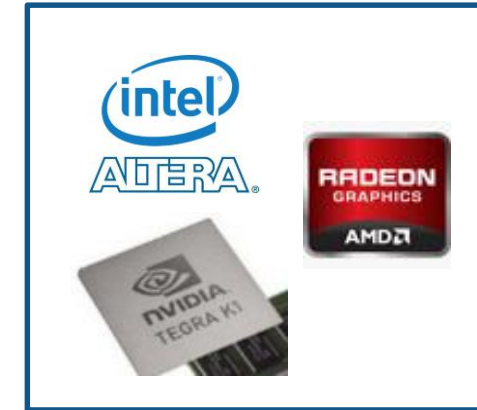


Machine Learning learns tasks using features extracted manually from data

Deep Learning learns both features and tasks directly from data



Deep Learning Workflow



Access & Explore

Preprocess Data

Develop Predictive Models

Integrate Analytics to Systems

Training Data

- Large open datasets
- Recorded labeled data (e.g., images, video)

Pre-Trained Models

- Access state-of-art networks already trained on large datasets

Data Augmentation

- Crop, resize and rotate images – create more training data

Label Training Data

- For image data: draw ROI's, label individual pixels with labels

Train from Scratch

- Configure and train a deep network with massive amount of training data

Transfer Learning

- Tune pre-trained model for a different task with smaller datasets

Share Models

- Publish models for others to use

Embedded Deployment

- Embedded processors
- FPGA

HPC

- Servers (multi-GPU)
- Clusters

How big is big?

What does “Big Data” even mean?

“Big data is a term for data sets that are so large or complex that traditional data processing applications are inadequate to deal with them.”

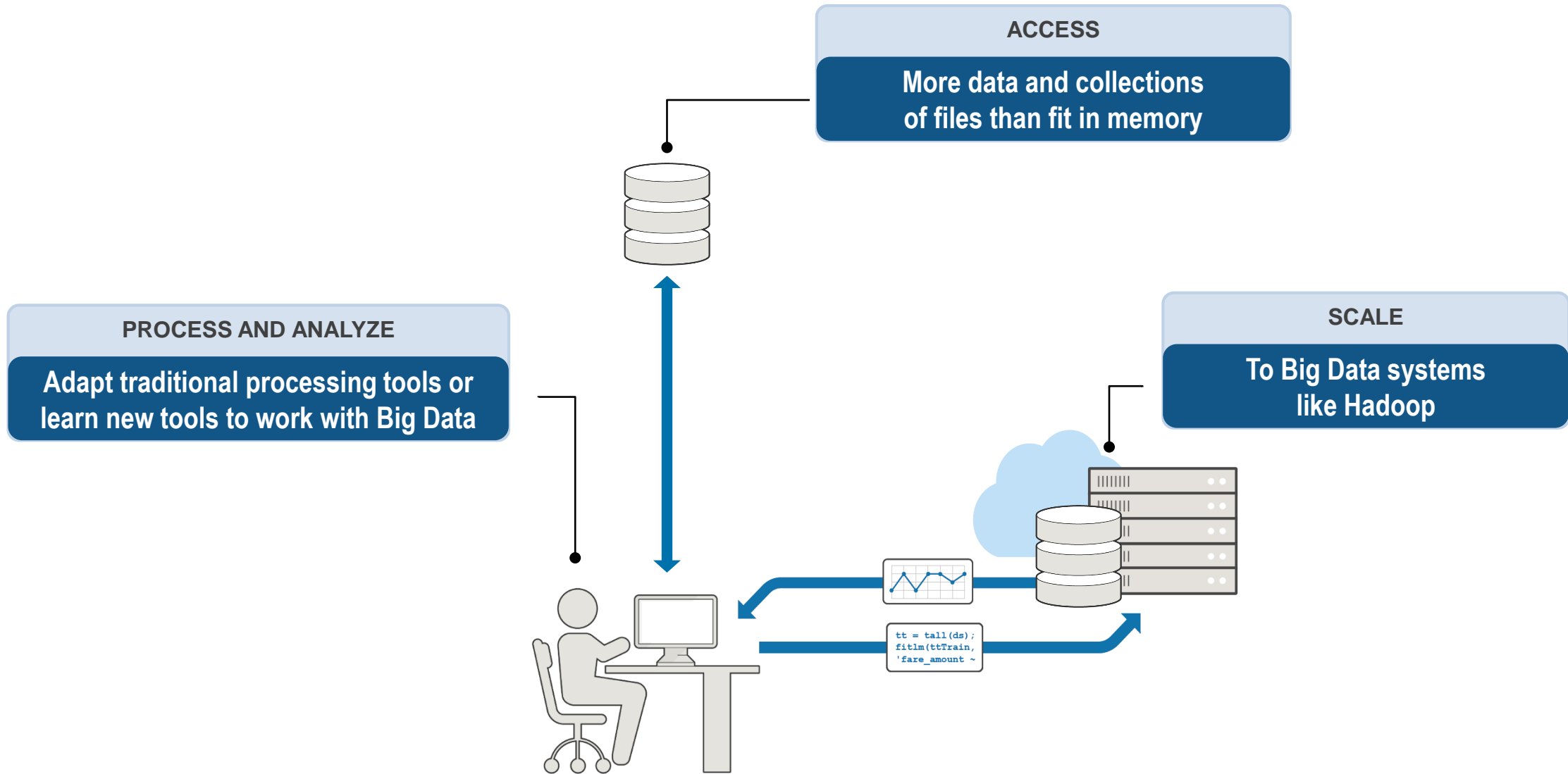
Wikipedia

So, what's the (big) problem?

- Traditional tools and approaches won't work
 - **Getting** the data is hard; **processing** it is even harder
 - Need to learn **new tools** and **new coding styles**
 - Have to rewrite algorithms, often at a lower level of abstraction
- Quality of your results can be impacted
 - e.g., by being forced to work on a subset of your data



Big Data workflow



Big solutions

Wouldn't it be nice if you could:

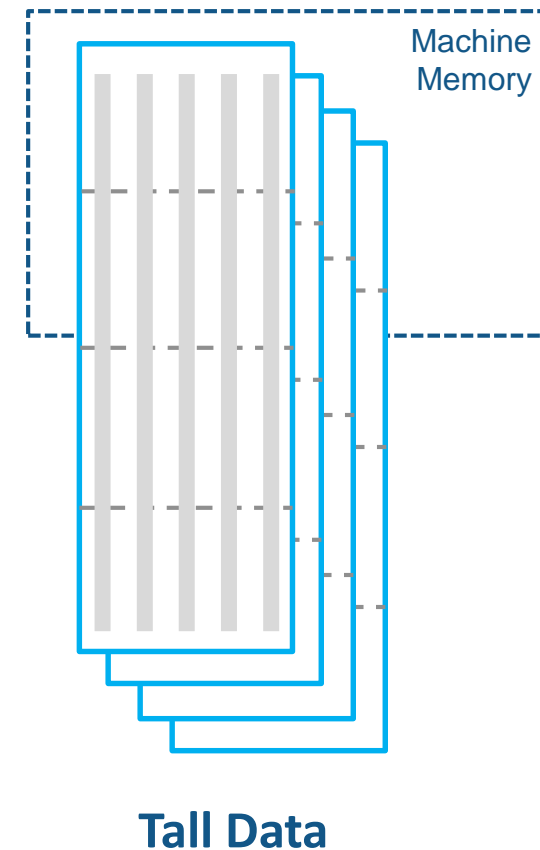
- Easily access data however it is stored
- Prototype algorithms quickly using small data sets
- Scale up to big data sets running on large clusters
- **Using the same intuitive MATLAB syntax you are used to**



Tall Arrays

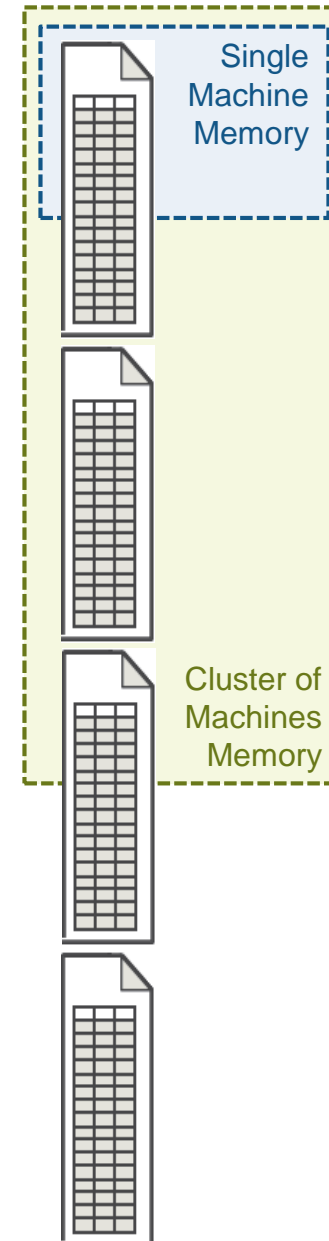
Scaling your code to big data

- Applicable when:
 - Data is **columnar** – with **many** rows
 - Overall data size is **too big to fit into memory**
 - Operations are mathematical/statistical in nature
- Statistical and machine learning applications
 - Hundreds of functions supported in MATLAB and Statistics and Machine Learning Toolbox



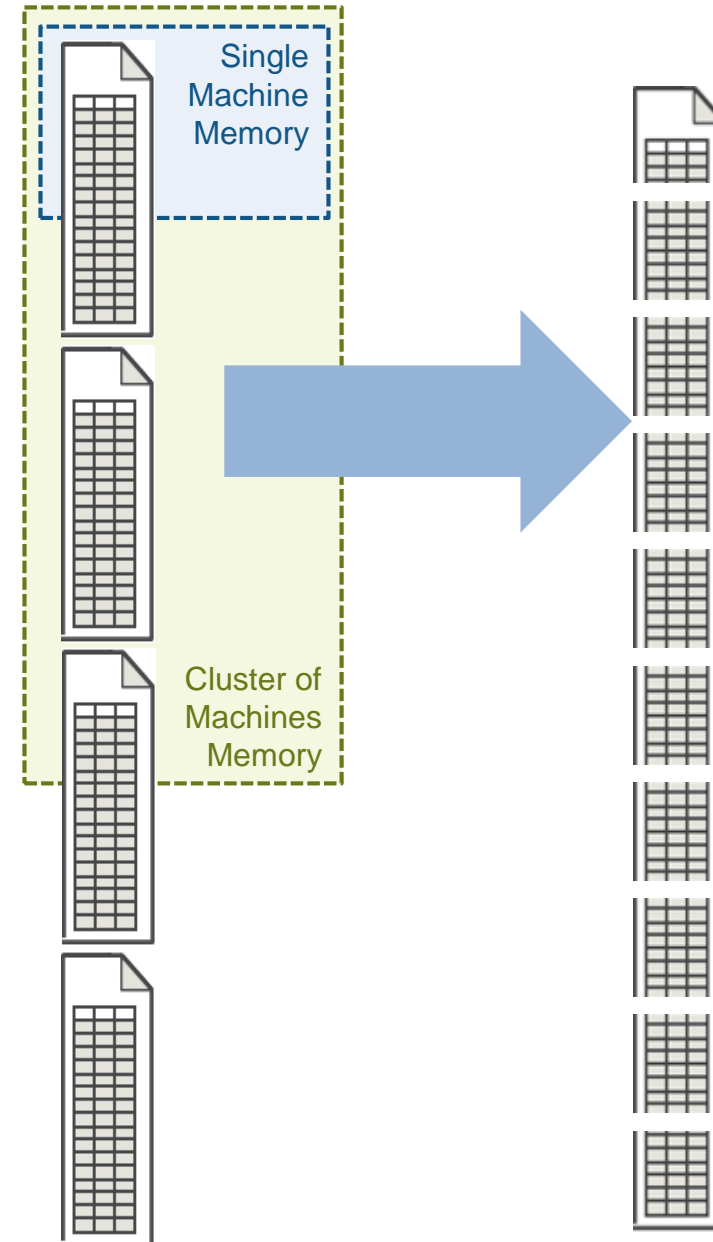
ta11 arrays R2016b

- Data is in one or more files
- Typically tabular data
- Files stacked vertically
- Data doesn't fit into memory (even cluster memory)



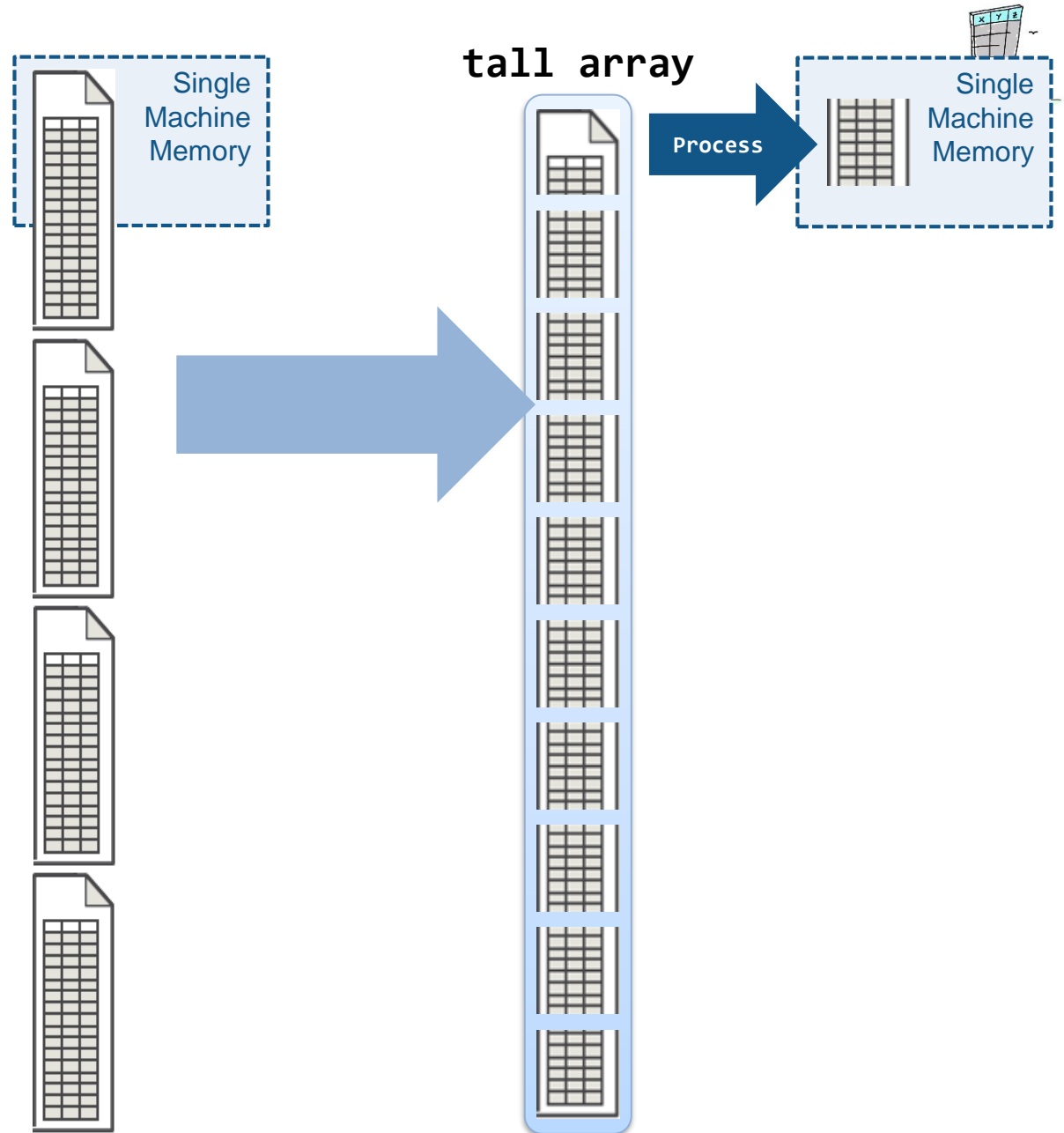
ta11 arrays R2016b

- Automatically breaks data up into small “chunks” that fit in memory



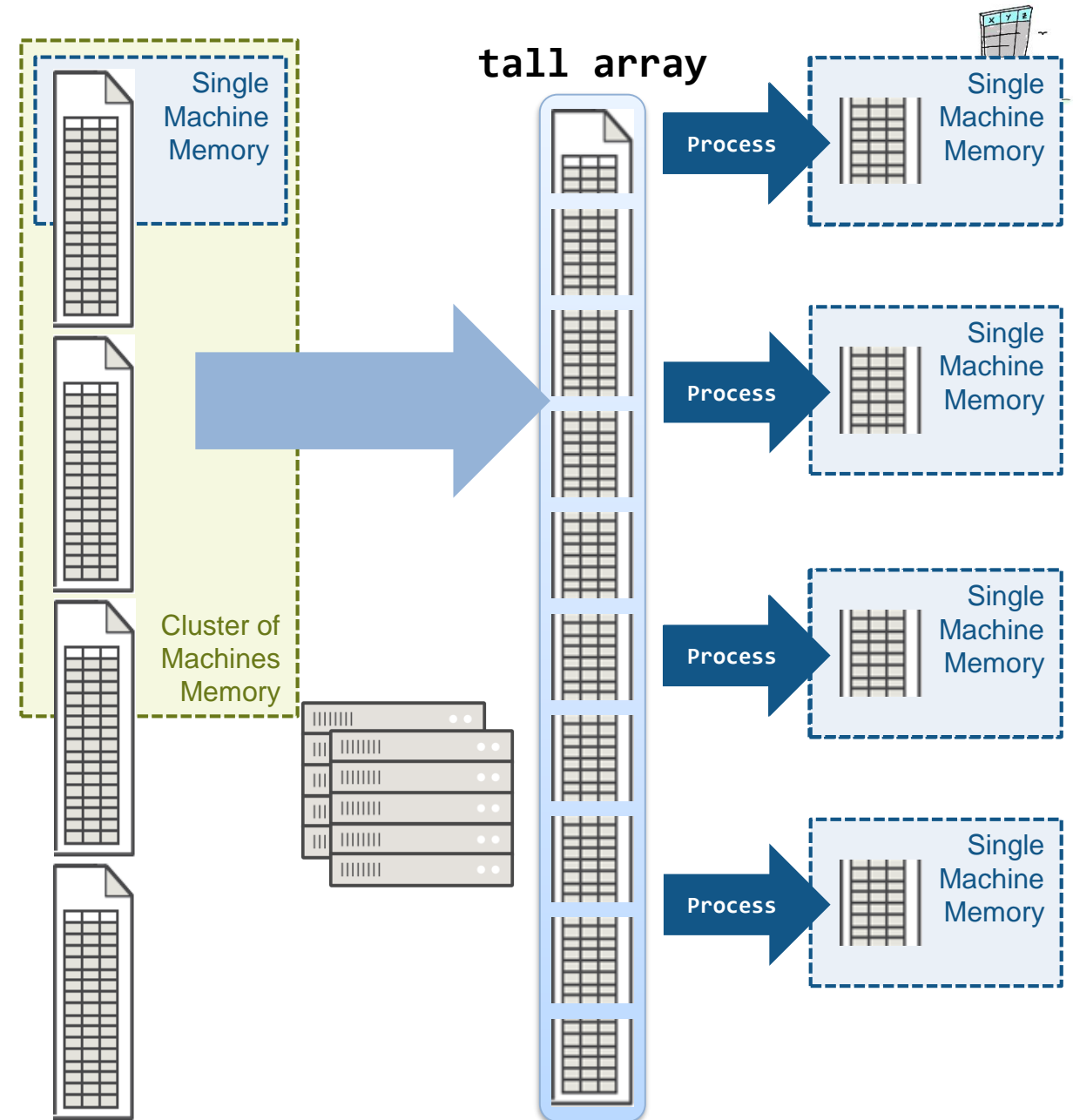
tall arrays R2016b

- “Chunk” processing is handled automatically
- Processing code for tall arrays is the same as ordinary arrays



ta11 arrays R2016b

- With Parallel Computing Toolbox, process several “chunks” at once
- Can scale up to clusters with MATLAB Distributed Computing Server



Example: Running on Spark + Hadoop

% Hadoop/Spark Cluster

```
numWorkers = 16;
```

```
setenv('HADOOP_HOME', '/dev_env/cluster/hadoop');  
setenv('SPARK_HOME', '/dev_env/cluster/spark');
```

```
cluster = parallel_cluster.Hadoop;  
cluster.SparkProperties('spark.executor.instances') = num2str(numWorkers);  
mr = mapreducer(cluster);
```

% Access the data

```
ds = datastore('hdfs://hadoop01:54310/datasets/taxiData/*.csv');  
tt = tall(ds);
```

Big Data Workflow With Tall Data Types

Access Data

- Text
- Spreadsheet (Excel)
- Database (SQL)
- Custom Reader

**Datstores for
common types of
structured data**

Tall Data Types

- table
- timetable (R2017a)
- cell
- double
- numeric
- cellstr
- datetime
- categorical

**Tall versions of
commonly used
MATLAB data types**

Exploration & Pre-processing

- Numeric functions
- Basic stats reductions
- Date/Time capabilities
- Categorical
- String processing
- Table wrangling
- Missing Data handling
- Summary visualizations:
 - Histogram/histogram2
 - Kernel density plot
 - Bin-scatter

**Hundreds of pre-built
functions**

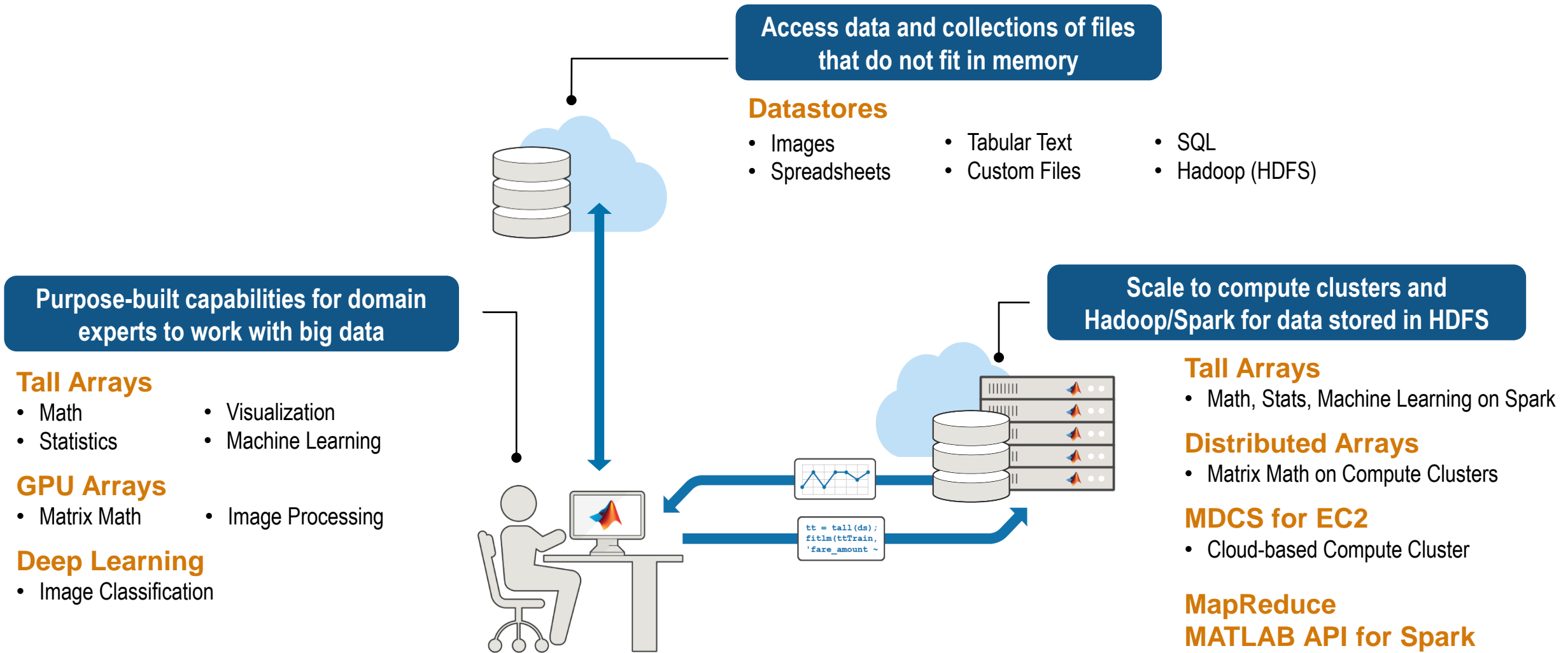
Machine Learning

- Linear Model
- Logistic Regression
- Discriminant analysis
- K-means
- PCA
- Random data sampling
- Summary statistics
- Decision trees (R2017a)

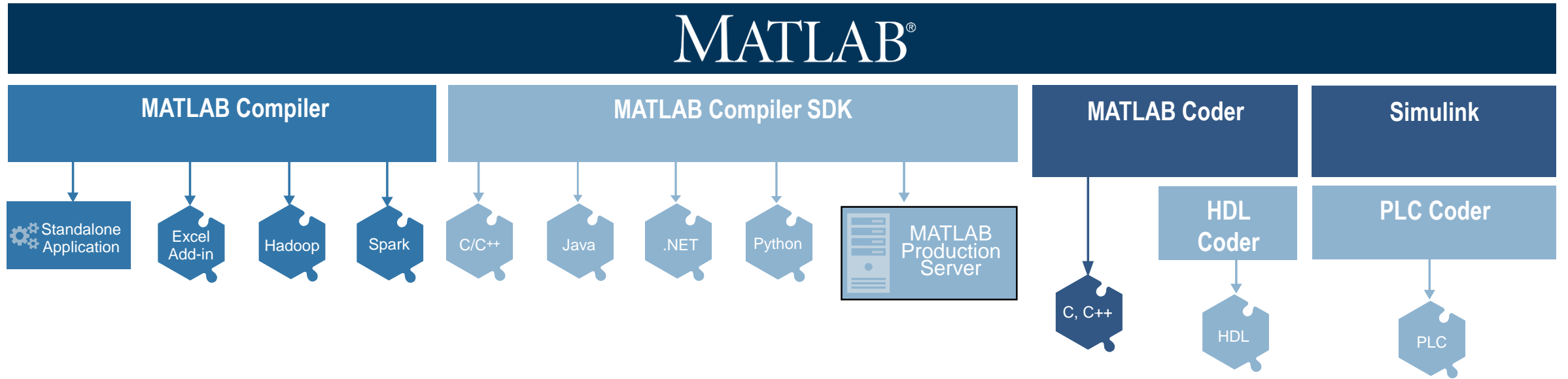
**Key statistics and
machine learning
algorithms**

MATLAB programming for data that does not fit into memory

Big Data Capabilities in MATLAB



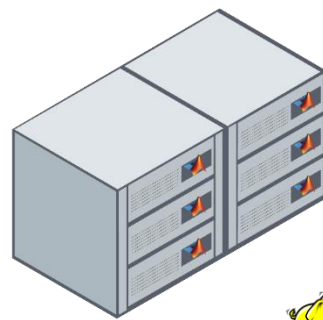
MathWorks Products for Operationalizing Analytics



Desktop Users



Enterprise IT Systems

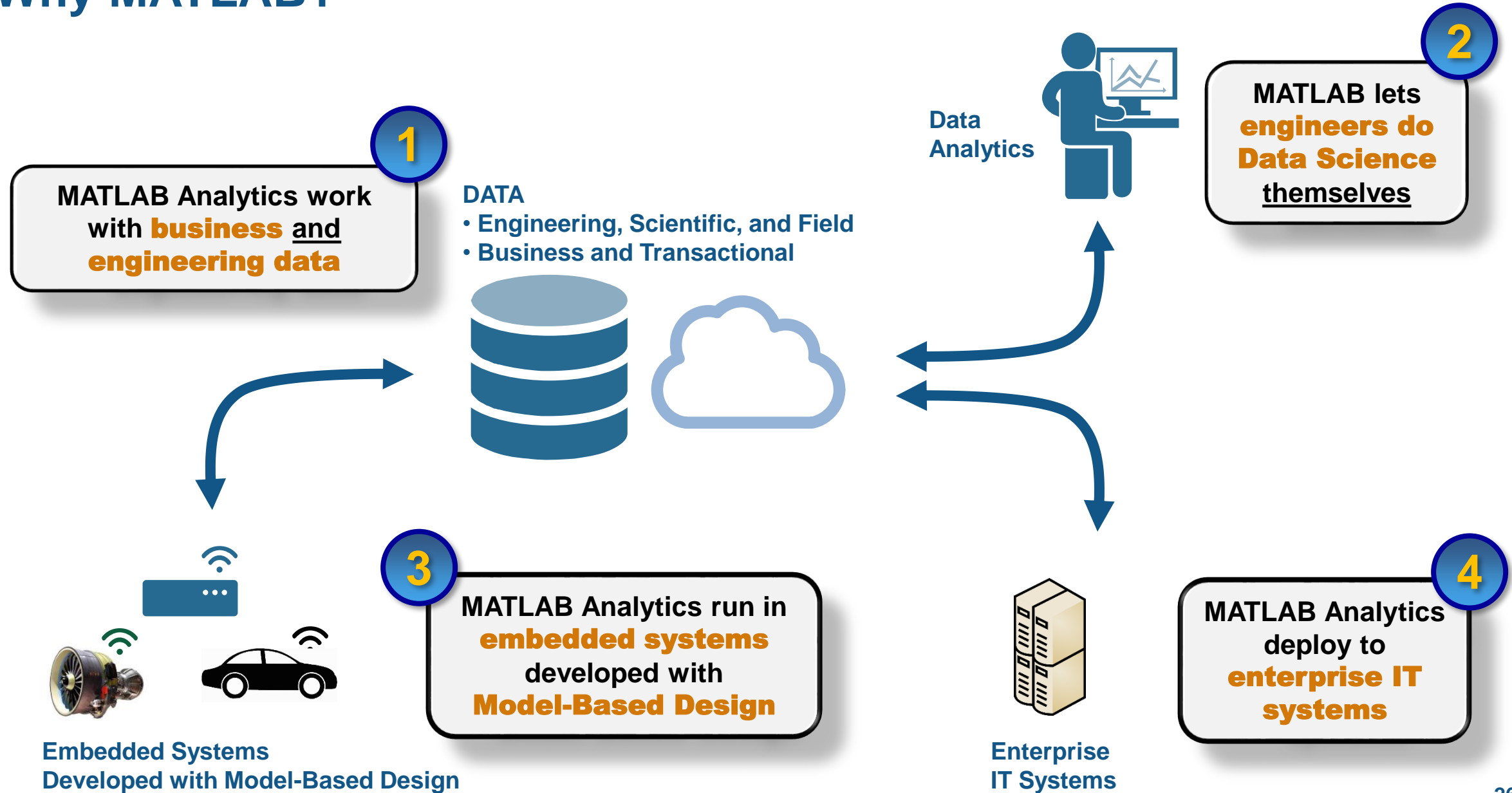


Embedded Systems (Including Edge Devices)



- Microcontrollers
- DSP chips
- FPGAs
- ARM-based
- Low-cost:
 - Arduino
 - Raspberry Pi
 - BeagleBone

Why MATLAB?



Summary

- MATLAB makes it easy, convenient, and scalable to work with big data
 - **Access** any kind of big data from any file system
 - Use tall arrays to **process and analyze** that data on your desktop, clusters, or on Hadoop/Spark

There's no need to learn big data programming or out-of-memory techniques -- simply use the same code and syntax you're already used to.

Resources to learn and get started

mathworks.com/machine-learning

Machine Learning with MATLAB
Search MathWorks.com

Overview | Examples
Trial software | Contact sales

Machine Learning with MATLAB Webinar
Learn how to get started using machine learning tools to detect patterns and build predictive models from your data sets.

[Watch video](#)

Choosing the Best Classification Model and Avoiding Overfitting

[» Download white paper](#)

Explore Products for Machine Learning

- Statistics and Machine Learning Toolbox™
- Neural Network Toolbox™
- Computer Vision System Toolbox™
- Fuzzy Logic Toolbox™

Engineers and data scientists work with large amounts of data in a variety of formats such as sensor, image, video, telemetry, databases, and more. They use machine learning to find patterns in data and to build models that predict future outcomes based on historical data. With MATLAB®, you have immediate access to prebuilt functions, extensive toolboxes, and specialized apps for [classification](#), [regression](#), and [clustering](#). You can:

- Compare approaches such as logistic regression, classification trees, support vector machines, ensemble methods, and deep learning.
- Use model refinement and reduction techniques to create an accurate model that best captures the predictive power of your data.
- Integrate machine learning models into enterprise systems, clusters, and clouds, and target models to real-time embedded hardware.

Machine Learning with MATLAB

Machine Learning with MATLAB Overview

Deep Learning in 11 Lines of MATLAB Code

Deep Learning with MATLAB

[Contact sales](#)

[Get a trial](#)

[Read: The Netflix Prize and Production Machine Learning](#)

eBook

Machine Learning Challenges: Choosing the Best Model and Avoiding Overfitting

Modeling with machine learning is a challenging but valuable skill for any user working with data. No matter what you use machine learning for, chances are you have encountered a classification or overfitting concern along the way. This paper shows you how to mitigate the effects of these challenges using MATLAB.

WHITE PAPER

Using Analytics and Machine Learning to Build Intelligent Products and Services

WHITE PAPER

Learn More

Big Data with MATLAB

<http://www.mathworks.com/discovery/big-data-matlab.html>

MapReduce and Hadoop

[mathworks.com/discovery/matlab-mapreduce-hadoop](http://www.mathworks.com/discovery/matlab-mapreduce-hadoop)

cloudera Why Cloudera Products Services & Support Solutions Get Started

Find a partner

More partners means more choice. And with the largest ecosystem of companies developing, integrating and deploying technology on Apache Hadoop (via our open-source CDH distribution) than any other vendor in the Big Data market, you're sure to find a solution that suits your business needs.

FEATURED SOLUTIONS

MathWorks PARTNER WEBSITE

MATLAB® is the easiest and most productive software for engineers and scientists. Whether you're analyzing data, developing algorithms, or creating models, MATLAB provides an environment that invites exploration and discovery. It combines a high-level language with a desktop environment tuned for iterative engineering and scientific workflows. It is used for machine learning, signal processing, image processing, computer vision, communications, computational finance, control design, robotics, and much more.

- Partner Category: Analytics & Business Intelligence
- Partner Type: Software Vendor (S/V)

Cloudera Versions	Partner Product Name	Partner Product Version	Interface Components	Supports Kerberos	Supports Apache Sentry
CDH57	MATLAB R2016B	R2016B	HDFS, MapReduce, Spark	Yes	Not applicable
CDH57	Statistics and ML Toolbox R2016B	R2016B	HDFS, MapReduce, Spark	Yes	Not applicable
CDH57	MATLAB Compiler R2016B	R2016B	HDFS, MapReduce, Spark	Yes	Not applicable
CDH57	MATLAB Distributed Computing Server R2016B	R2016B	HDFS, MapReduce, Spark	Yes	Not applicable
CDH54	MATLAB Compiler R2015b	R2015b	HDFS, MapReduce	Yes	Not applicable
CDH54	MATLAB R2015b	R2015b	HDFS, MapReduce	Yes	Not applicable
CDH54	MATLAB Distributed Computing Server R2015b	R2015b	HDFS, MapReduce	Yes	Not applicable

Big Data with MATLAB Contact sales Trial Software

MATLAB MapReduce and Hadoop Contact sales Trial Software

How to work with huge data

Big data refers to the dramatic increase in data acquisition devices and other data sources.

A primary driver of this trend is the proliferation of acquisition devices and other data sources. Big data sources include streaming video from security cameras, as well as data from sensors that can contain gigabytes per day.

Big data represents an opportunity for informed decisions, but it also presents challenges. Large data sets, which may take too long to process, may not be designed to process big data. Therefore, MATLAB provides a powerful and established programming technique for applying filtering, statistics and other general analysis methods to big data.

The MapReduce functionality built into MATLAB lets you analyze data that does not fit into memory. By running your MapReduce based algorithms in parallel (using Parallel Computing Toolbox™), you can better utilize the processing resources on your desktop without changing your algorithms.

To analyze data in MATLAB using MapReduce:

- Specify the data you want to analyze using `datastore`.
- Create your map and reduce functions in MATLAB.
- Execute your map and reduce functions using `mapreduce`.

While MATLAB MapReduce is optimized for array-based analysis, it is fully compatible with Hadoop MapReduce, so you can run your MapReduce based algorithms within the Hadoop MapReduce framework:

- Execute MapReduce based algorithms on Hadoop directly from the MATLAB desktop, using MATLAB Distributed Computing Server™.
- Package MapReduce based algorithms for deploying to production Hadoop systems, using MATLAB Compiler™.

Working with Big Data in MATLAB

- 64-bit Computing.** The amount of data that can be held in memory – typically up to 2 GB of the OS. For Windows 8, you need 64-bit MATLAB.
- Memory Mapped Variables.** The amount of data that can be held in memory is limited by the amount of RAM available on the system.
- Disk Variables.** The amount of data that can be held in memory is limited by the amount of RAM available on the system.
- Datastore.** Use the `datastore` function to read data from files, collections of files, or databases.

MapReduce on the Desktop

Explore and analyze big data sets on your desktop with the MapReduce programming technique built into MATLAB.

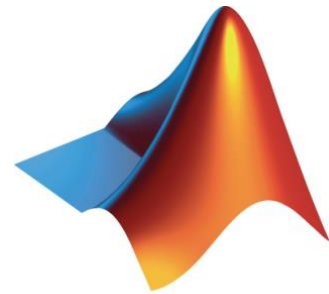
Creating algorithms using MapReduce: `max`, `mean`, `mean by group`, `histograms`, `covariance` and related quantities, `summary statistics by group`, `logistic regression`, `tall skinny QR`.

- Get started with MATLAB MapReduce
- MapReduce design patterns
- Use MATLAB MapReduce with relational databases

MapReduce on Hadoop

Execute MATLAB MapReduce based algorithms within Hadoop MapReduce to explore and analyze data that is stored and managed on Hadoop, using MATLAB Distributed Computing Server.

- Run MATLAB MapReduce on Hadoop
- Create applications and libraries based upon MATLAB MapReduce for deployment within production instances of Hadoop, using MATLAB Compiler.
- Deploy MATLAB MapReduce applications to Hadoop



MathWorks®

Accelerating the pace of engineering and science

© 2017 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.