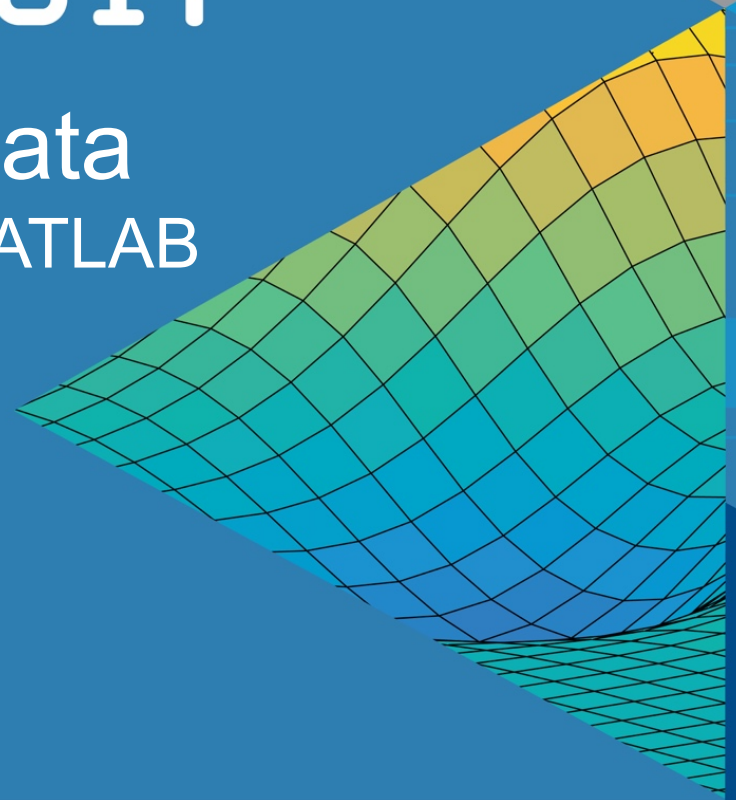
MATLAB EXPO 2017

Machine Learning auf Big Data

praktische Programmierkonzepte in MATLAB

Dmytro Martynenko
Applikationsingenieur, MathWorks



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 tra[MATLAB EXPO 2017]cessing applications are inadequate to deal with them.”

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

ACCESS

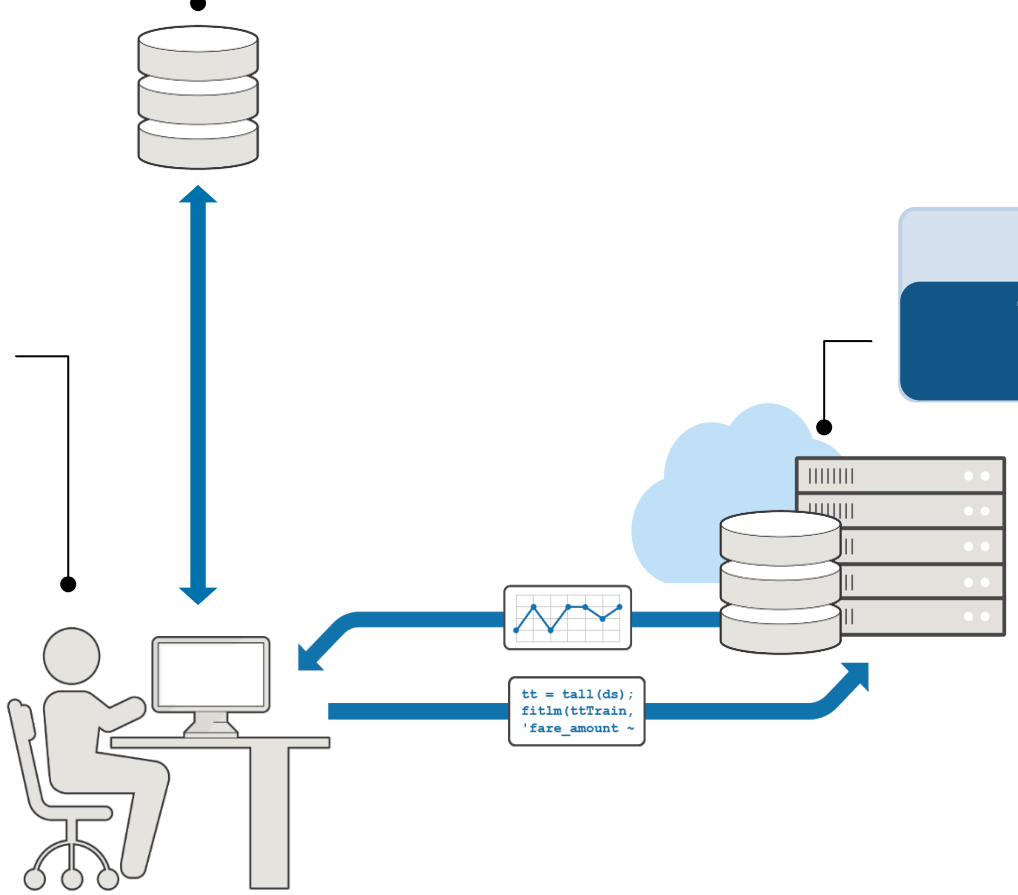
More data and collections of files than fit in memory

PROCESS AND ANALYZE

Adapt traditional processing tools or learn new tools to work with Big Data

SCALE

To Big Data systems like Hadoop

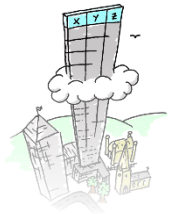


Big solutions

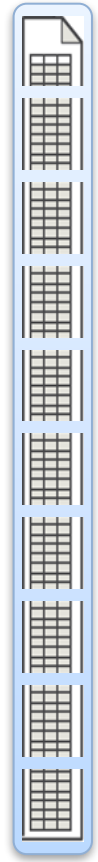
Wouldn't it be nice if you could:

- 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**

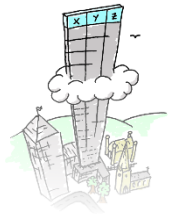
`tall` arrays R2016b



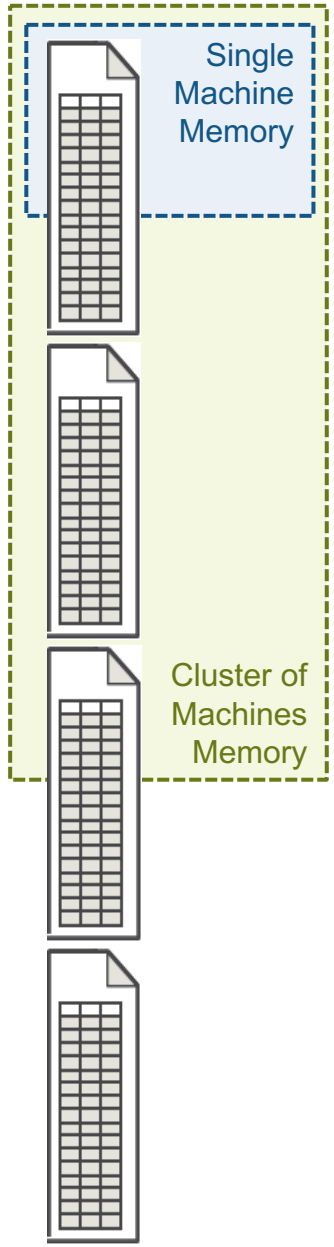
- For data that doesn't fit into memory
- Lots of observations (hence "tall")
- Looks like a normal MATLAB array
  - Supports numeric types, tables, datetimes, strings, etc…
  - Supports basic math, stats, indexing, etc.
  - **Statistics and Machine Learning Toolbox** support (clustering, classification, etc.)



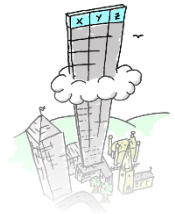
`tall` arrays R2016b



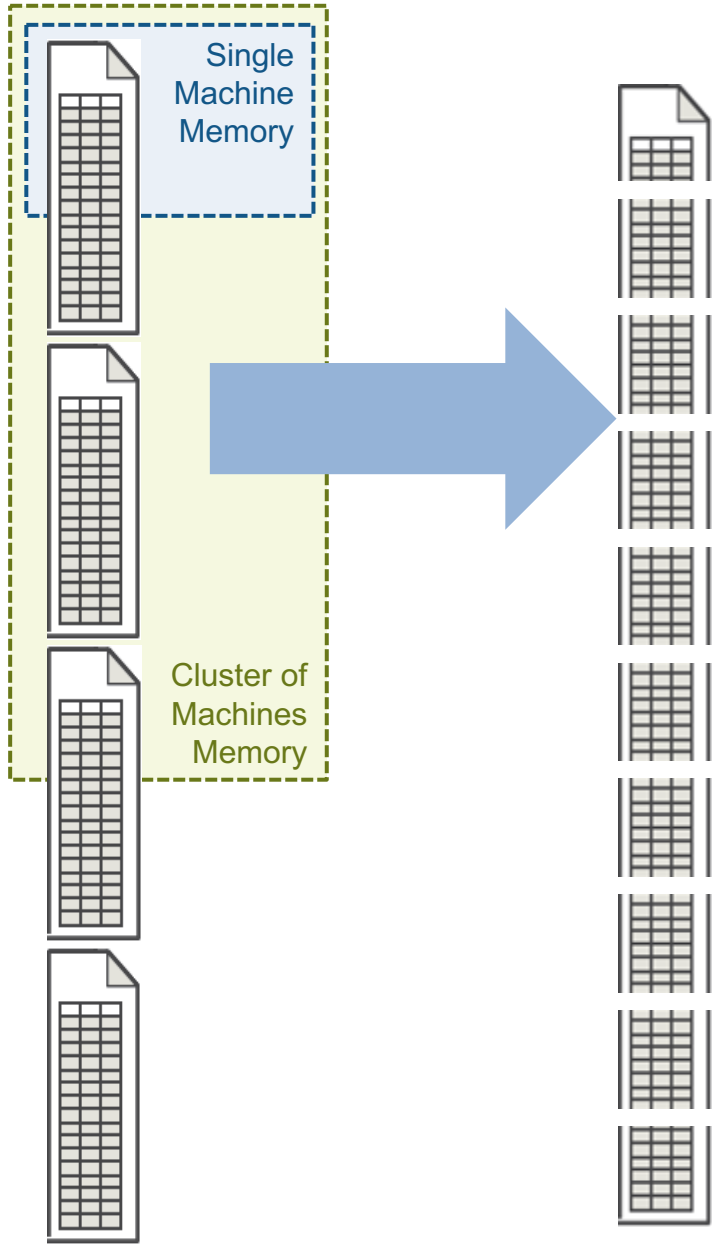
- Data is in one or more files
- Typically tabular data
- Files stacked vertically
- Data doesn't fit into memory
  (even cluster memory)



`tall` arrays R2016b

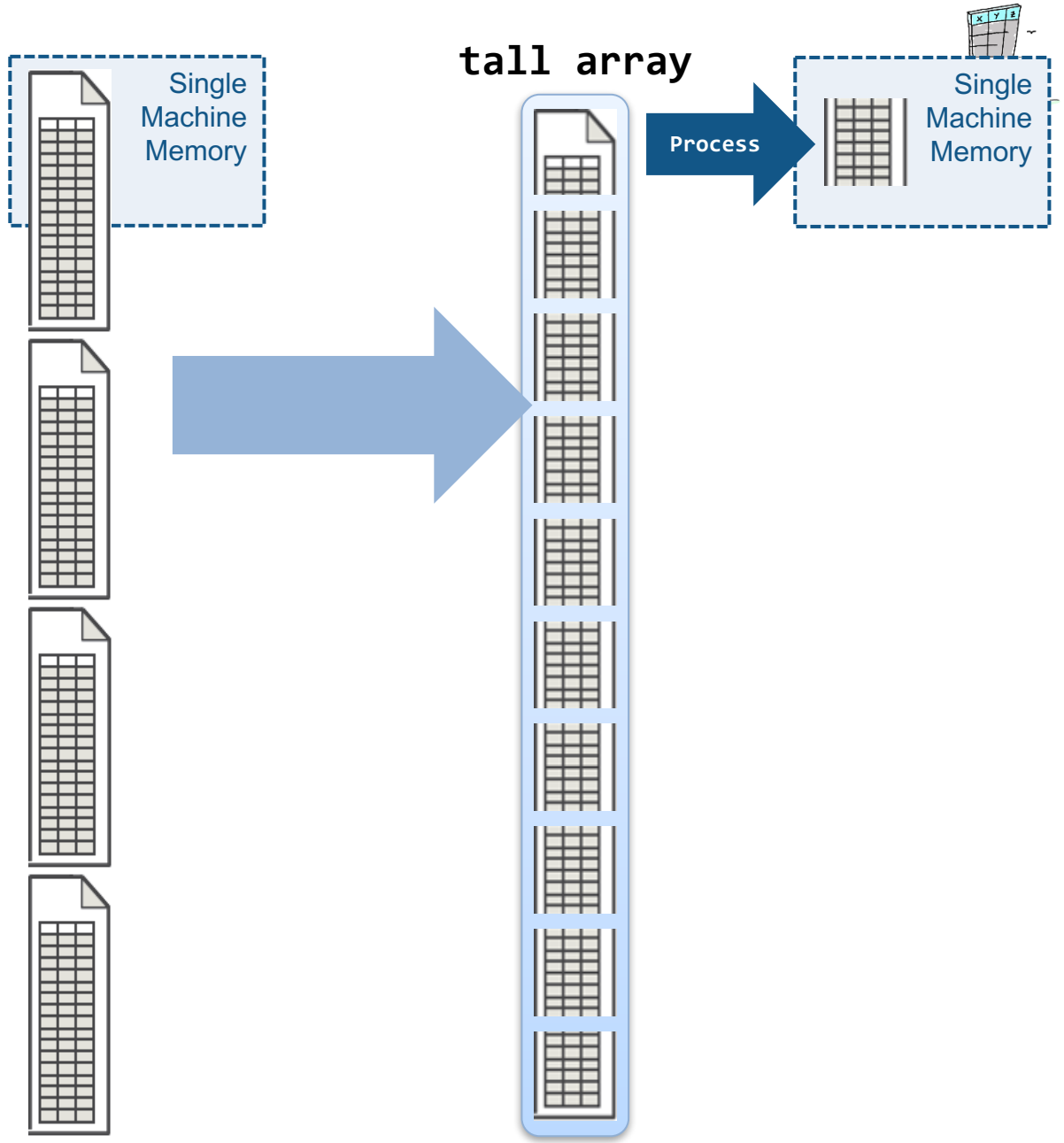


- Automatically breaks data up into small “chunks” that fit in memory



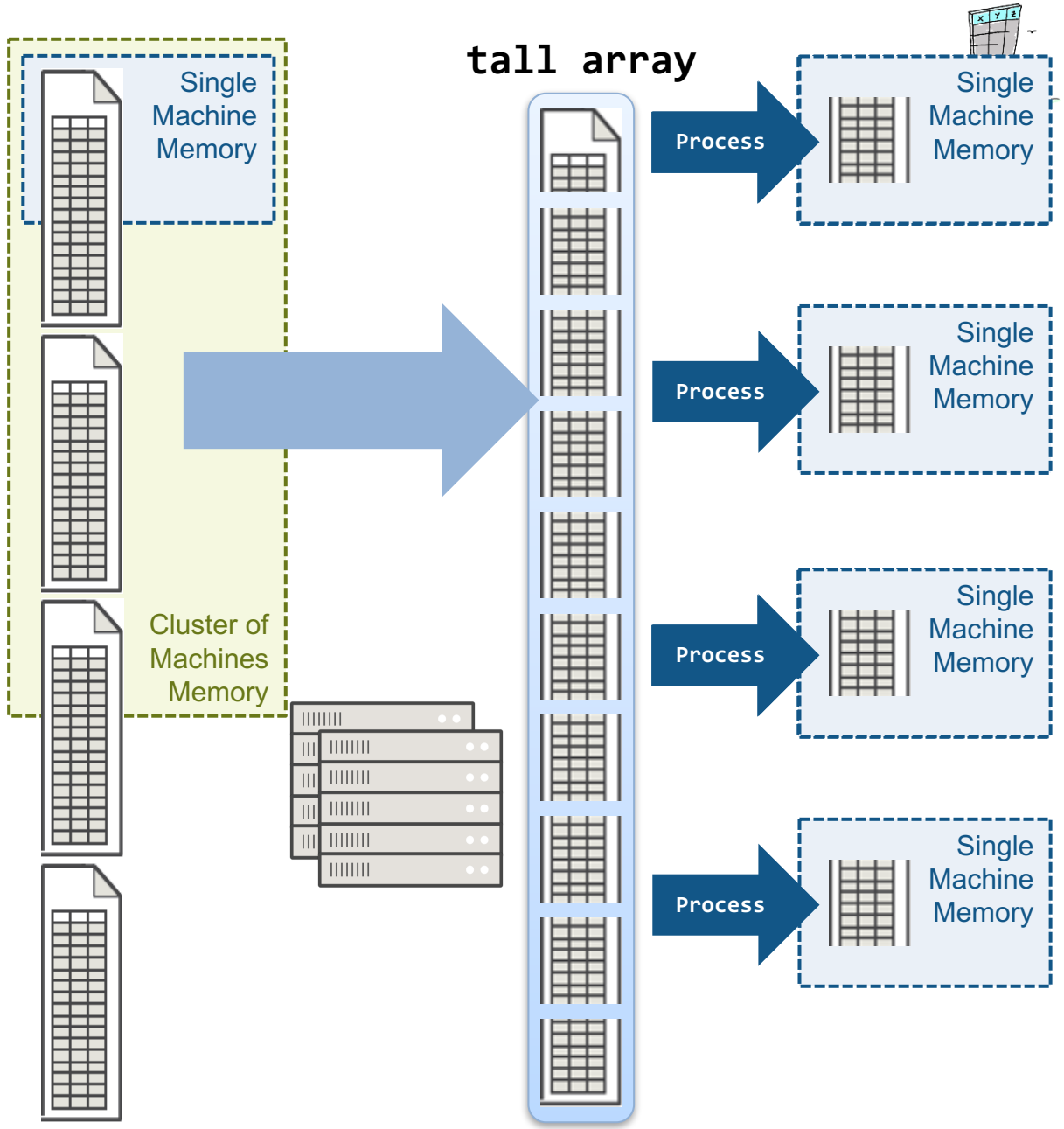
`tall` arrays R2016b

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

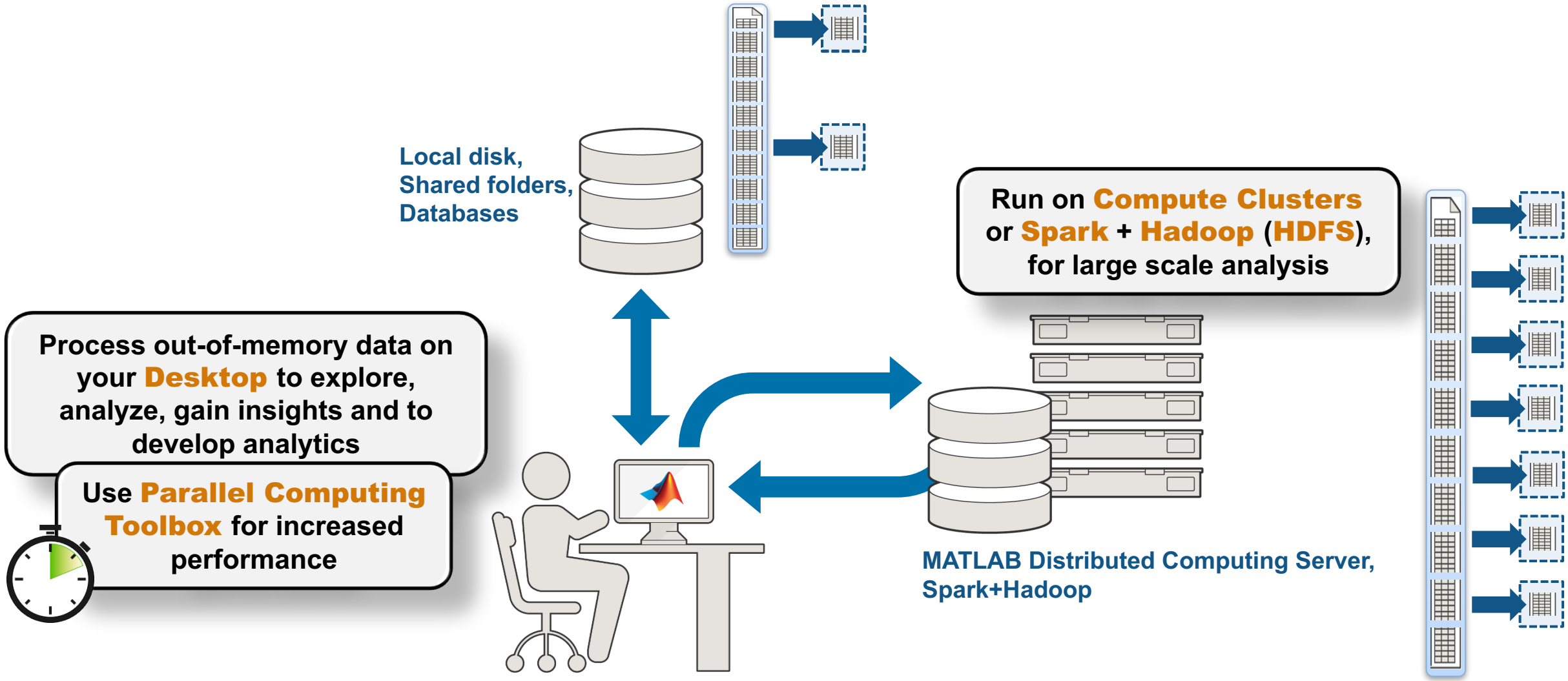


`tall` arrays R2016b

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



Summary for `tall` arrays



Big Data Workflow With Tall Data Types

Access Data

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

Datastores 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

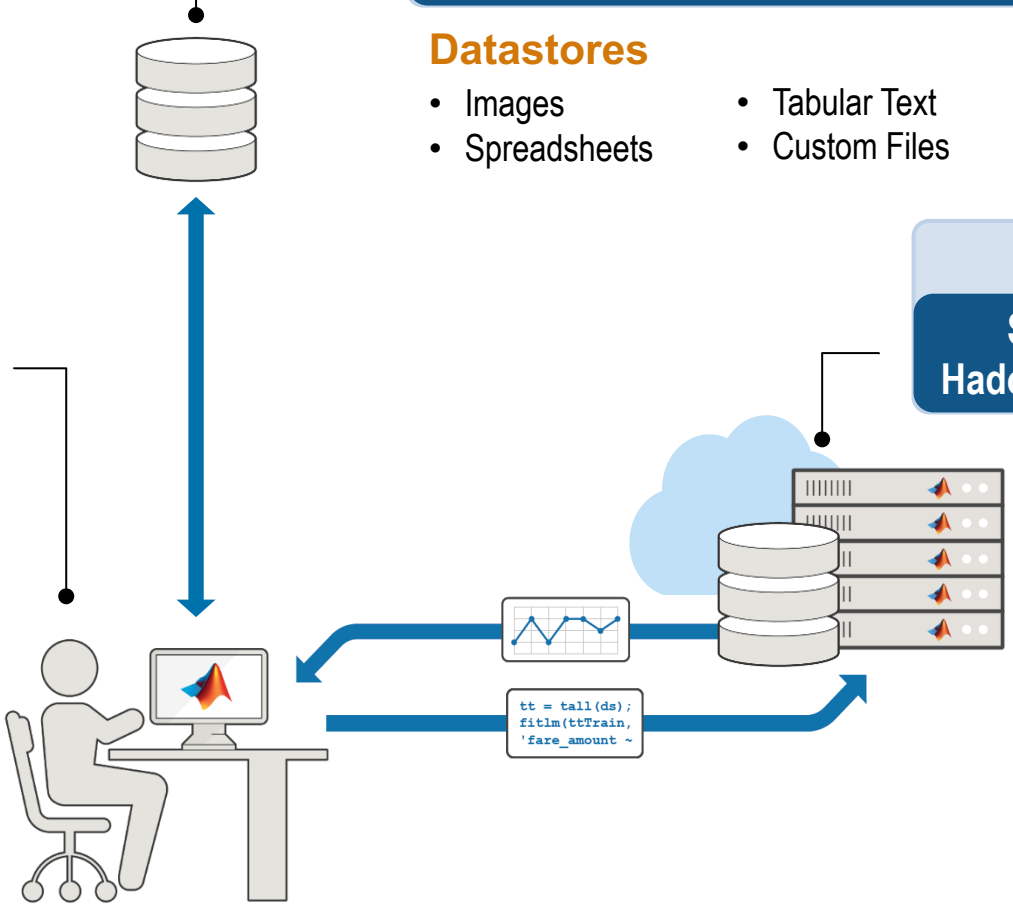
Big Data capabilities in MATLAB

ACCESS

Access data and collections of files that do not fit in memory

Datastores

- Images
- Spreadsheets
- Tabular Text
- Custom Files
- SQL
- Hadoop (HDFS)



PROCESS AND ANALYZE

Purpose-built capabilities for domain experts to work with big data locally

Tall Arrays

- Math
- Statistics
- Visualization
- Machine Learning

GPU Arrays

- Matrix Math
- Image Processing

Deep Learning

- Image Classification

SCALE

Scale to compute clusters and Hadoop/Spark for data stored in HDFS

Tall Arrays

- Math, Stats, Machine Learning on Spark

Distributed Arrays

- Matrix Math on Compute Clusters

MDCS for EC2

- Cloud-based Compute Cluster

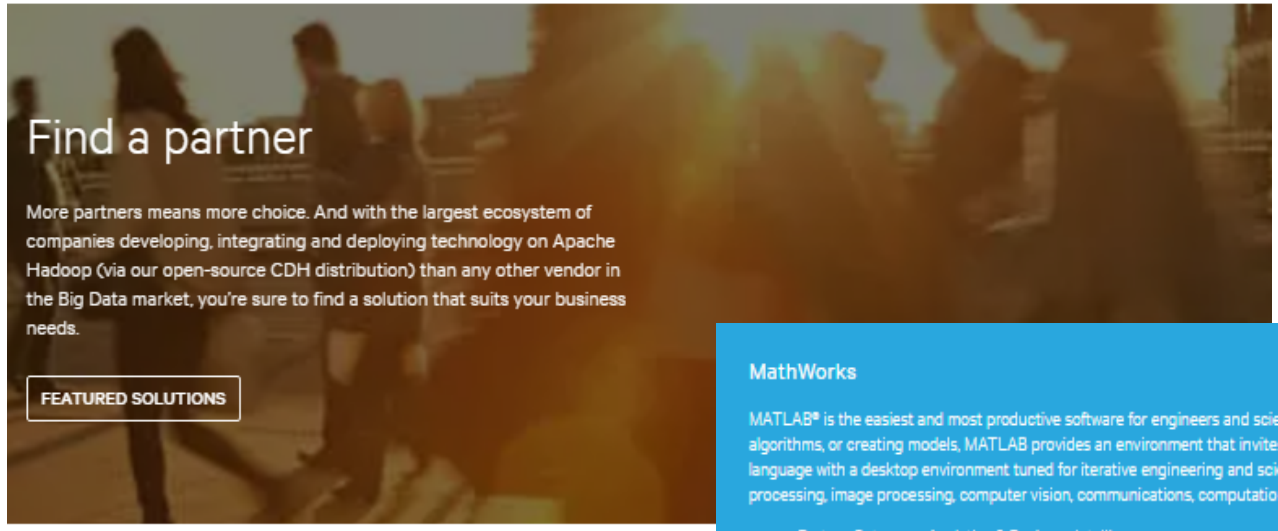
MapReduce

MATLAB API for Spark

MathWorks: certified partner by Cloudera

cloudera

Why Cloudera Products Services & Support Solutions Get Started



CERTIFICATION PARTNER TYPE ISV/IHV CATEGORY APPL

Certified

MathWorks

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 (ISV)

Cloudera Versions	Partner Product Name	Partner Product Version	Interface Components	Supports Kerberos	Supports Apache Sentry
CDH5.7	MATLAB R2016B	R2016B	HDFS, MapReduce, Spark	Yes	Not applicable
CDH5.7	Statistics and ML Toolbox R2016B	R2016B	HDFS, MapReduce, Spark	Yes	Not applicable
CDH5.7	MATLAB Compiler R2016B	R2016B	HDFS, MapReduce, Spark	Yes	Not applicable
CDH5.7	MATLAB Distributed Computing Server R2016B	R2016B	HDFS, MapReduce, Spark	Yes	Not applicable
CDH5.4	MATLAB Compiler R2015b	R2015b	HDFS, MapReduce	Yes	Not applicable
CDH5.4	MATLAB R2015b	R2015b	HDFS, MapReduce	Yes	Not applicable
CDH5.4	MATLAB Distributed Computing Server R2015b	R2015b	HDFS, MapReduce	Yes	Not applicable

Summary

- MATLAB makes it easy, convenient, and scalable to apply machine learning on 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.

For more information

- Website:

  https://www.mathworks.com/solutions/big-data-matlab

- Web search for:

  “**Big Data MATLAB**”

© 2017 The MathWorks, Inc.