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 traditional data processing 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

SCALE
To Big Data systems like Hadoop

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

MATLAB EXPO 2017
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

Process out-of-memory data on your Desktop to explore, analyze, gain insights and to develop analytics.

Use Parallel Computing Toolbox for increased performance.

Run on Compute Clusters or Spark + Hadoop (HDFS), for large scale analysis.

MATLAB Distributed Computing Server, Spark+Hadoop
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)

**SCALE**
- Scale to compute clusters and Hadoop/Spark for data stored in 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

**Datastores**
- Images
- Spreadsheets
- Tabular Text
- Custom Files
- SQL
- Hadoop (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

Find a partner

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

FEATURED SOLUTIONS

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 (SV)

<table>
<thead>
<tr>
<th>Clouds Version</th>
<th>Partner Product Name</th>
<th>Partner Product Version</th>
<th>Interface Components</th>
<th>Supports Kerberos</th>
<th>Supports Apache Sentry</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDHv7</td>
<td>MATLAB R2018b</td>
<td>R2018b</td>
<td>HDP, MapReduce, Spark</td>
<td>Yes</td>
<td>Not Applicable</td>
</tr>
<tr>
<td>CDHv7</td>
<td>Statistics and ML: ToolBox: R2018b</td>
<td>R2018b</td>
<td>HDP, MapReduce, Spark</td>
<td>Yes</td>
<td>Not Applicable</td>
</tr>
<tr>
<td>CDHv7</td>
<td>MATLAB Compiler: R2018b</td>
<td>R2018b</td>
<td>HDP, MapReduce, Spark</td>
<td>Yes</td>
<td>Not Applicable</td>
</tr>
<tr>
<td>CDHv7</td>
<td>MATLAB Distributed Computing Server: R2018b</td>
<td>R2018b</td>
<td>HDP, MapReduce, Spark</td>
<td>Yes</td>
<td>Not Applicable</td>
</tr>
<tr>
<td>CDHv6</td>
<td>MATLAB Compiler: R2017b</td>
<td>R2017b</td>
<td>HDP, MapReduce</td>
<td>Yes</td>
<td>Not Applicable</td>
</tr>
<tr>
<td>CDHv6</td>
<td>MATLAB R2017b</td>
<td>R2017b</td>
<td>HDP, MapReduce</td>
<td>Yes</td>
<td>Not Applicable</td>
</tr>
<tr>
<td>CDHv6</td>
<td>MATLAB Distributed Computing Server: R2017b</td>
<td>R2017b</td>
<td>HDP, MapReduce</td>
<td>Yes</td>
<td>Not Applicable</td>
</tr>
</tbody>
</table>
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:

- Web search for:
  “Big Data MATLAB”