

Tackling Big Data with MATLAB



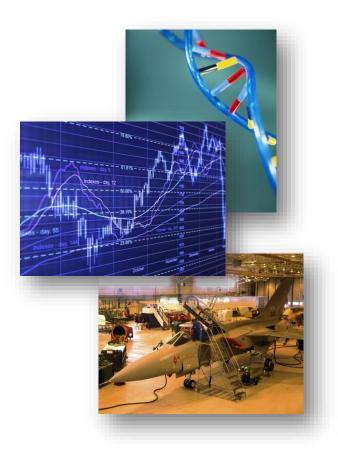
Francesca Perino Application Engineering Team - MathWorks

© 2014 The MathWorks, Inc.



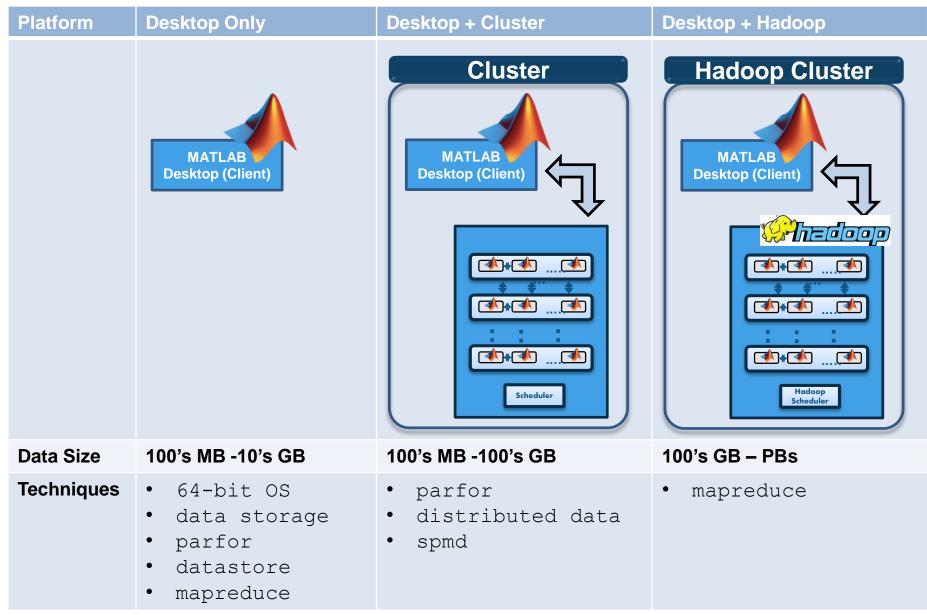
Running into "Big Data" Issues?

- "Out of memory"
 - Running out of address space
- Performance
 - Takes too long to process all of your data
- Slow processing (swapping)
 - Data too large to be efficiently managed between RAM and virtual memory



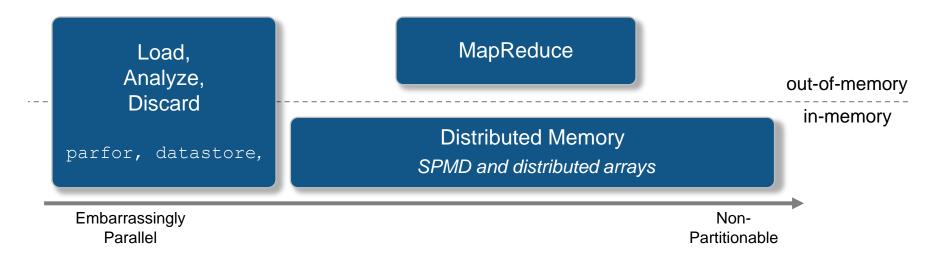


Options for Handling Large Data





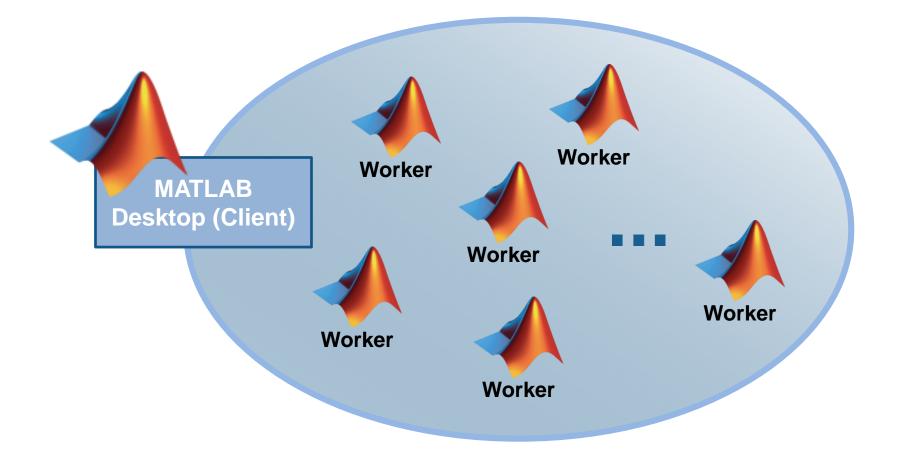
Techniques for Big Data in MATLAB



Complexity



Parallel Computing with MATLAB





Example: Determining Land Use Using Parallel for-loops (parfor)

Data

- Arial images of agriculture land
- 24 TIF files
- Analysis
 - Find and measure irrigation fields
 - Determine which irrigation circles are in use (by color)
 - Calculate area under irrigation

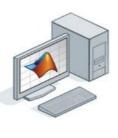


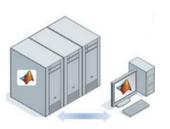


When to Use parfor

Data Characteristics

- Can be of any format (i.e. text, images) as long as it can be broken into pieces
- The data for each iteration must fit in memory
- Compute Platform
 - Desktop (Parallel Computing Toolbox)
 - Cluster (MATLAB Distributed Computing Server)
- Analysis Characteristics
 - Each iteration of your loop must be independent







Access Big Data

datastore

- Easily specify data set
 - Single text file (or collection of text files)
- Preview data structure and format
- Select data to import using column names
- Incrementally read subsets of the data

| E Desktop | * | Name | Date modified | Туре | Size | |
|-----------------|---|------------|-------------------|-------------|-----------|---|
|) Downloads | | 1987.csv | 8/13/2014 3:37 PM | WinZip File | 12,356 KB | |
| Google Drive | | 1988.csv | 8/13/2014 3:45 PM | WinZip File | 48,339 KB | |
| Mathworks | | 🍕 1989.csv | 8/13/2014 3:44 PM | WinZip File | 48,050 KB | |
| S Recent Places | | 🔍 1990.csv | 8/13/2014 3:45 PM | WinZip File | 50,822 KB | |
| ARC 1.1 | | 🔍 1991.csv | 8/13/2014 3:43 PM | WinZip File | 48,709 KB | |
| Libraries | Ξ | 🔍 1992.csv | 8/13/2014 3:46 PM | WinZip File | 48,869 KB | = |
| Jocuments | | 🍕 1993.csv | 8/13/2014 3:43 PM | WinZip File | 48,938 KB | |
| Pictures | | 🍕 1994.csv | 8/13/2014 3:54 PM | WinZip File | 49,926 KB | |
| Videos | | 🔍 1995.csv | 8/13/2014 4:06 PM | WinZip File | 73,127 KB | |
| S videos | | 🔍 1996.csv | 8/13/2014 4:07 PM | WinZip File | 74,110 KB | |
| 🝓 Homegroup | | 🍕 1997.csv | 8/13/2014 4:09 PM | WinZip File | 74,908 KB | |
| | | 🔍 1998.csv | 8/13/2014 4:06 PM | WinZip File | 74,887 KB | |

| ans = | | | |
|-------|-------|------------|-----------|
| Year | Month | DayofMonth | DayOfWeek |
| 1987 | 10 | 21 | 3 |
| 1987 | 10 | 26 | 1 |
| 1987 | 10 | 23 | 5 |
| 1987 | 10 | 23 | 5 |

```
airdata = datastore('*.csv');
airdata.SelectedVariables = {'Distance', 'ArrDelay`};
data = read(airdata);
```



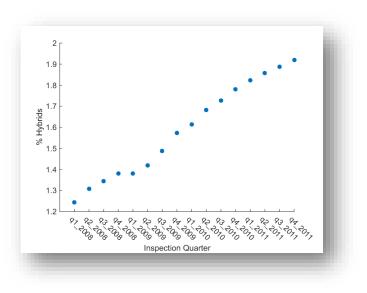
Example: Vehicle Registry Analysis Using a DataStore

Data

- Massachusetts Vehicle Registration
 Data from 2008-2011
- 16M records, 45 fields

| muni_id | veh_zip | insp_year | model_year | make |
|---------|---------|-----------|------------|------------|
| | | | | |
| 325 | 1089 | 2011 | 2008 | 'Hyundai' |
| 325 | 1089 | 2009 | 2008 | 'Hyundai' |
| 288 | 1776 | 2011 | 2008 | 'Acura' |
| 288 | 1776 | 2008 | 2008 | 'Acura' |
| 145 | 2364 | 2011 | 2005 | 'Chevrolet |
| 325 | 1089 | 2010 | 2008 | 'Hyundai' |
| 325 | 1089 | 2011 | 2008 | 'Hyundai' |
| 288 | 1776 | 2009 | 2008 | 'Acura' |

- Analysis
 - Examine hybrid adoptions
 - Calculate % of hybrids registered by quarter
 - Fit growth to predict further adoption



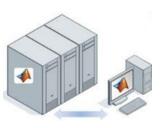


When to Use datastore

Data Characteristics

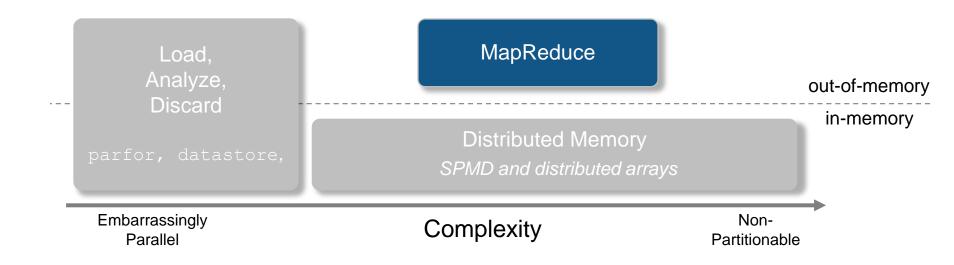
- Text data in files, databases or stored in the Hadoop Distributed File System (HDFS)
- Compute Platform
 - Desktop
- Analysis Characteristics
 - Supports Load, Analyze, Discard workflows
 - Incrementally read chunks of data, process within a while loop







Techniques for Big Data in MATLAB





mapreduce

| | Da | ita St | ore | | | Мар | | Shuffle nd Sort | R | educe |
|------------|--------|--------|--------|--------|-------------|------------|----------|--------------------|-------|------------------|
| Veh_typ | 03.08 | Q4_08 | Q1_09 | Hybrid | Hybrid 0 | Key: Q3_08 | l Hybrid | | | |
| | | | | | 1 | Ney. Q5_00 | | Key: Q3_08 | Кеу | % Hybrid (Value) |
| Car | 1 | 1 | 1 | 0 | 1 | | 1 | | Q3 08 | 0.4 |
| SUV Car | 0 1 | 1 1 | 1 1 | 1 | 0 | | 0 | | Q4_08 | 0.67 |
| Car | 0 | 0 | 1 | 1 | 1 | Key: Q4_08 | 0 | | Q1_09 | 0.75 |
| Car | 0 | 1 | 1 | 1 | 1 | | i i | I | | |
| Car | 1 | 1 | 1 | 1 | 1 | | 0 | | | |
| Car | 0 | 0 | 1 | 1 | 0 | | 1 | Key: Q4_08 | | |
| SUV | 0 | 1 | 1 | 0 | 1 | Key: Q1_09 | 1 | Ney. Q4_00 | 1 | |
| Car | 1 | 1 | 1 | 0 | 1 | | 1 | | | |
| SUV | 1 | 1 | 1 | 1 | 1 | | 0 | | | |
| Car | 0 | 1 | 1 | 1 | 1 | | 1 | | I | |
| Car | 1 | 0 | 0 | 0 | 0 | Key: Q3_08 | I | | | |
| | | | | | 0 | | 0 | | | |
| | | | | | | | 1 | Key: Q1_09 | | |
| | | | | | 0 | Key: Q4_08 | 1 | | | |
| | | | | | 1 | | | | - | |
| | | | | | 0 | | | | I | |
| | | | | | 0 1 | Key: Q1_09 | 1 | | | |
| | | | | | | | \sim | | I | 12 |

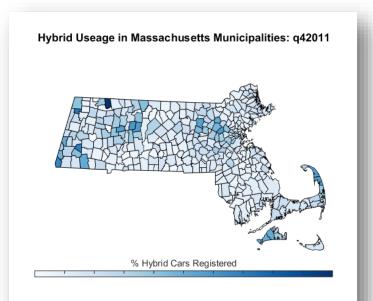


Example: Vehicle Registry Analysis Using MapReduce

Data

- Massachusetts Vehicle Registration
 Data from 2008-2011
- 16M records, 45 fields
- Analysis
 - Examine hybrid adoptions
 - Calculate % of hybrids registered
 - By Quarter
 - By Regional Area
 - Create map of results

| muni_id | veh_zip | insp_year | model_year | make |
|---------|---------|-----------|------------|-----------|
| 325 | 1089 | 2011 | 2008 | 'Hvundai' |
| 325 | 1089 | 2009 | 2008 | 'Hyundai' |
| 288 | 1776 | 2011 | 2008 | 'Acura' |
| 288 | 1776 | 2008 | 2008 | 'Acura' |
| 145 | 2364 | 2011 | 2005 | 'Chevrole |
| 325 | 1089 | 2010 | 2008 | 'Hyundai' |
| 325 | 1089 | 2011 | 2008 | 'Hyundai' |
| 288 | 1776 | 2009 | 2008 | 'Acura' |

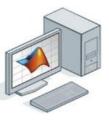


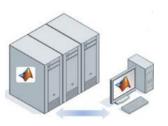


When to Use mapreduce

Data Characteristics

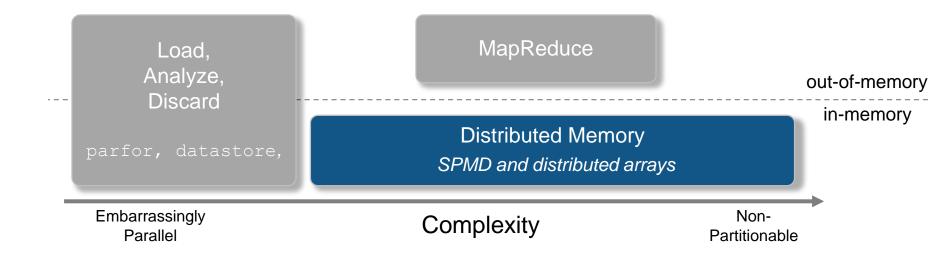
- Text data in files, databases or stored in the Hadoop Distributed File System (HDFS)
- Dataset will not fit into memory
- Compute Platform
 - Desktop
 - Scales to run within Hadoop MapReduce on data in HDFS
- Analysis Characteristics
 - Must be able to be Partitioned into two phases
 - 1. Map: filter or process sub-segments of data
 - 2. Reduce: aggregate interim results and calculate final answer







Techniques for Big Data in MATLAB





spmd blocks

spmd

% single program across workers
end

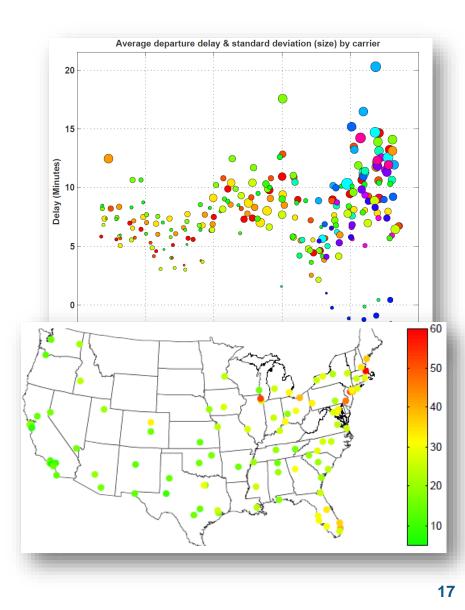
- Mix parallel and serial code in the same function
- Run on a pool of MATLAB resources
- Single Program runs simultaneously across workers
- Multiple Data spread across multiple workers



Example: Airline Delay Analysis

Data

- BTS/RITA Airline
 On-Time Statistics
- 123.5M records, 29 fields
- Analysis
 - Calculate delay patterns
 - Visualize summaries
 - Estimate & evaluate predictive models

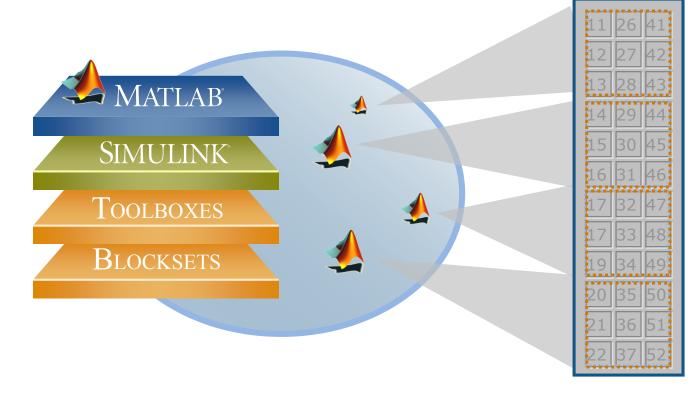




Distributed Arrays

Available from

- Parallel Computing Toolbox
- MATLAB Distributed Computing Server



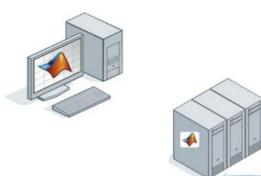
Remotely Manipulate Array from Desktop

Distributed Array Lives on the Cluster



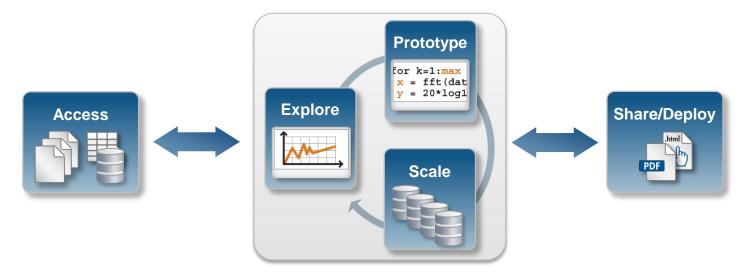
When to Use Distributed Memory

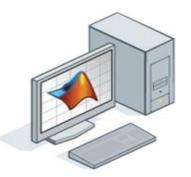
- Data Characteristics
 - Data must be fit in collective memory across machines
- Compute Platform
 - Prototype (subset of data) on desktop
 - Run on a cluster or cloud
- Analysis Characteristics
 - Consists of:
 - Parts that can be run on data in memory (spmd)
 - Supported functions for distributed arrays



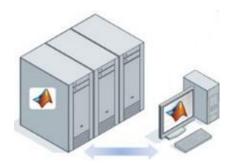


Big Data Analysis with MATLAB





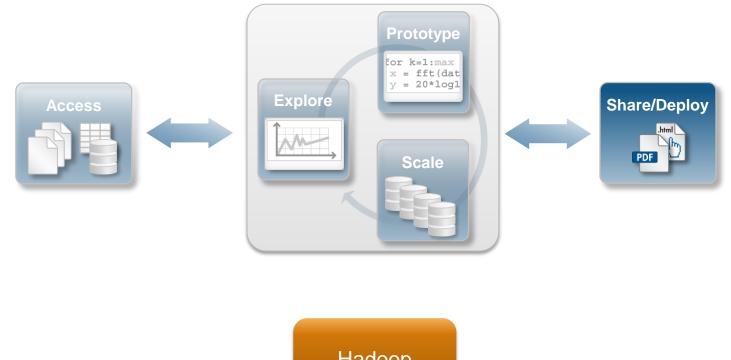
Work on the desktop

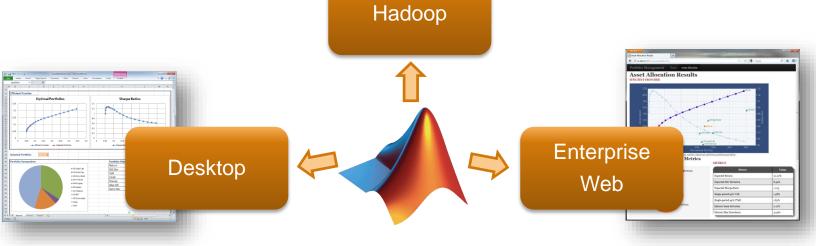


Scale capacity as needed



Deploy Big Data Algorithms







Learn More

- MATLAB Documentation
 - Strategies for Efficient Use of Memory
 - Resolving "Out of Memory" Errors
- Big Data with MATLAB
 - <u>www.mathworks.com/discovery/big-data-matlab.html</u>



How to work with huge and fast data sets

Big data refers to the dramatic increase in the amount and rate of data being created and made availa analysis.

A primary driver of this trend is the ever increasing digitization of information. The number and types o acquisition devices and other data generation mechanisms are growing all the time.

Big data sources include streaming data from instrumentation sensors, satellite and medical imagery, from security cameras, as well as data derived from financial markets and retail operations. Big data s these sources can contain gigabytes or terabytes of data, and may grow on the order of megabytes or gigabytes per day.

Big data represents an opportunity for analysts and data scientists to gain greater insight and to make informed decisions, but it also presents a number of challenges. Big data sets may not fit into available

- MATLAB MapReduce and Hadoop
 - <u>www.mathworks.com/discovery/matlab-mapreduce-hadoop.html</u>

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