

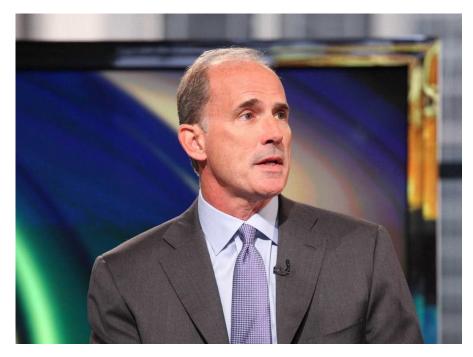
David Willingham Senior Application Engineer, MathWorks david.willingham@mathworks.com.au



© 2015 The MathWorks, Inc.



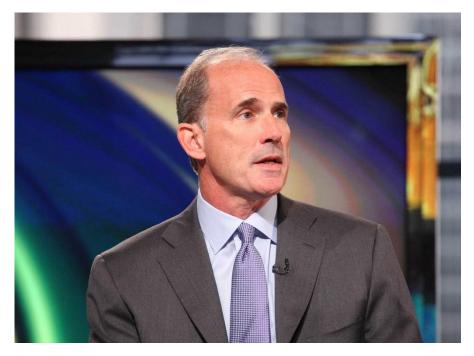
## "Data is the sword of the 21st century, those who wield it the samurai."



Google's Former SVP - Jonathan Rosenberg



## "Data is the sword of the 21st century, those who wield it the samurai."



Google's Former SVP - Jonathan Rosenberg

- Big data how to create it, manipulate it, and put it to good use.
- "If you want to work at Google, make sure you can use MATLAB."





• 2012: 2.5 billion GB (  $2.5 \times 10^{18}$  ) of data each day.



• 2012: 2.5 billion GB (  $2.5 \times 10^{18}$  ) of data each day.





• 2012: 2.5 billion GB (  $2.5 \times 10^{18}$  ) of data each day.



How can I work with large data sets?



• 2012: 2.5 billion GB ( 2.5 x 10<sup>18</sup> ) of data each day.



- How can I work with large data sets?
- How can I get business information from the data?



• 2012: 2.5 billion GB ( 2.5 x 10<sup>18</sup> ) of data each day.



- How can I work with large data sets?
- How can I get business information from the data?
- Do I need significant technical & theoretical knowledge?





Wikipedia

"Any collection of data sets so large and complex that it becomes difficult to process using ... traditional data processing applications."



Wikipedia

"Any collection of data sets so large and complex that it becomes difficult to process using ... traditional data processing applications."

Described with the 3 V's

Volume : amount of data

Velocity : speed at which data is generated or needs to be analysed

*Variety* : range of data types/data sources



Wikipedia

"Any collection of data sets so large and complex that it becomes difficult to process using ... traditional data processing applications."

Described with the 3 V's

Volume : amount of data

Velocity : speed at which data is generated or needs to be analysed

Variety : range of data types/data sources

But now there is a 4<sup>th</sup> V, Value: What Business Information can be obtained from Big Data.





- Software Developers
  - Programmers using languages such as Java / .NET
  - They may not be domain experts



- Software Developers
  - Programmers using languages such as Java / .NET
  - They may not be domain experts
- End Users (non technical)
  - Non programmers
  - They may or may not be domain experts



- Software Developers
  - Programmers using languages such as Java / .NET
  - They may not be domain experts
- End Users (non technical)
  - Non programmers
  - They may or may not be domain experts
- Domain Users
  - Scientists, Engineers, Analysts, etc..
  - Have some programming skills
  - Might use MATLAB to prototype ideas, algorithms models
  - They want their ideas to scale with the size of data easily



## **Key Industries**

- Aerospace and defense
- Automotive
- Biotech and pharmaceutical
- Communications
- Education
- Electronics and semiconductors
- Energy production
- Financial services
- Industrial automation and machinery
- Medical devices





#### Astrium Creates World's First Two-Way Laser Optical Link Between an Aircraft and a Communication Satellite

#### Challenge

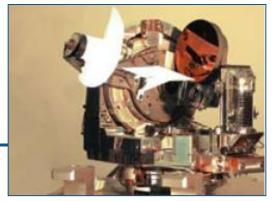
Develop controls to ensure the precision of a laser optical link between an aircraft and a communication satellite

#### **Solution**

Use MathWorks tools to model control algorithms and pointing hardware, conduct hardware-in-theloop tests, and deploy a real-time system for flight tests

#### **Results**

- First of its kind optical link demonstrated
- Design iterations reduced from days to hours
- Overall development time reduced by six months



LOLA telescope assembly, as fitted to aircraft in Artemis laser link trials.

"Using MathWorks tools for Model-Based Design, we simulated not only our control algorithms but also the physical hardware. By automatically generating code for the control software and the test bench, we reduced development time and implemented changes quickly. We visualized simulation and test results, which gave us confidence in the design we ultimately deployed."

> David Gendre Astrium



#### **Research Engineers Advance Design** of the International Linear Collider



Queen Mary high-throughput cluster.

#### Challenge

Design a control system for ensuring the precise alignment of particle beams in the International Linear Collider

#### Solution

Use MATLAB, Simulink, Parallel Computing Toolbox, and Instrument Control Toolbox software to design, model, and simulate the accelerator and alignment control system

#### Results

- Simulation time reduced by an order of magnitude
- Development integrated
- Existing work leveraged

"Using Parallel Computing Toolbox, we simply deployed our simulation on a large group cluster. We saw a linear improvement in speed, and we could run 100 simulations at once. MathWorks tools have enabled us to accomplish work that was once impossible."

> Dr. Glen White Queen Mary, University of London

Link to user story



#### BuildingIQ Develops Proactive Algorithms for HVAC Energy Optimization in Large-Scale Buildings

#### Challenge

Develop a real-time system to minimize HVAC energy costs in large-scale commercial buildings via proactive, predictive optimization

#### Solution

Use MATLAB to analyze and visualize big data sets, implement advanced optimization algorithms, and run the algorithms in a production cloud environment

#### **Results**

- Gigabytes of data analyzed and visualized
- Algorithm development speed increased tenfold
- Best algorithmic approaches quickly identified



Large-scale commercial buildings can reduce energy costs by 10–25% with BuildinglQ's energy optimization system.

"MATLAB has helped accelerate our R&D and deployment with its robust numerical algorithms, extensive visualization and analytics tools, reliable optimization routines, support for object-oriented programming, and ability to run in the cloud with our production Java applications."

Borislav Savkovic BuildinglQ







 4 years of sales data held in a Teradata data warehouse.





- 4 years of sales data held in a Teradata data warehouse.
- MATLAB is used to forecast the optimum stock levels.





- 4 years of sales data held in a Teradata data warehouse.
- MATLAB is used to forecast the optimum stock levels.



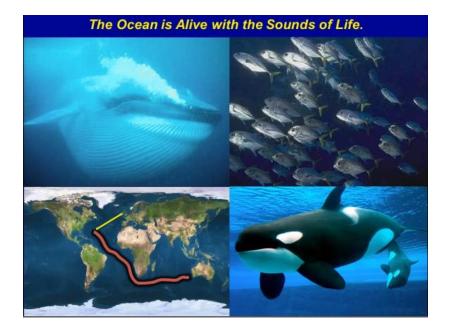
 "We can run 100 stores for 100 days in about half an hour. We can figure out quickly whether what we are doing is right and we can optimise that."







 Cornell University collected terabytes of ocean acoustic data.





- Cornell University collected terabytes of ocean acoustic data.
- Crowdsourced algorithms to detect and classify animal signals in the presence of noise.





- Cornell University collected terabytes of ocean acoustic data.
- Crowdsourced algorithms to detect and classify animal signals in the presence of noise.



"A data set that would have taken months to process can now be processed multiple times in just a few days using different detection algorithms."



## **Challenges of Big Data**

"Any collection of data sets so large and complex that it becomes difficult to process using ... traditional data processing applications." (Wikipedia)

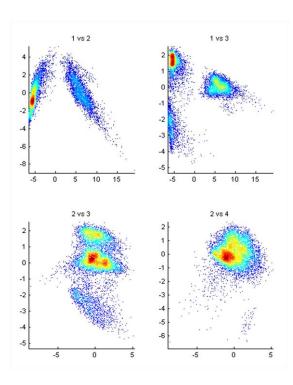
- Getting started
- Rapid data exploration
- Development of scalable algorithms
- Ease of deployment



## **Big Data Capabilities in MATLAB**

#### **Memory and Data Access**

- 64-bit processors
- Memory Mapped Variables
- Disk Variables
- Databases
- Datastores R2014b



#### **Programming Constructs**

- Streaming
- Block Processing
- Parallel-for loops
- GPU Arrays
- SPMD and Distributed Arrays
- MapReduce R2014b

#### **Platforms**

- Desktop (Multicore, GPU)
- Clusters
- Cloud Computing (MDCS on EC2)
- Hadoop R2014b



out-of-memory

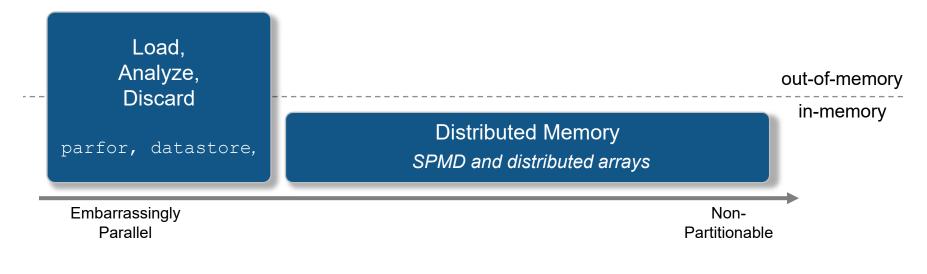
in-memory

Embarrassingly Parallel Non-Partitionable

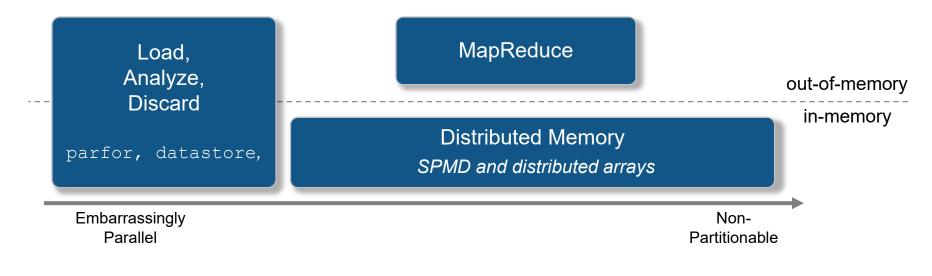














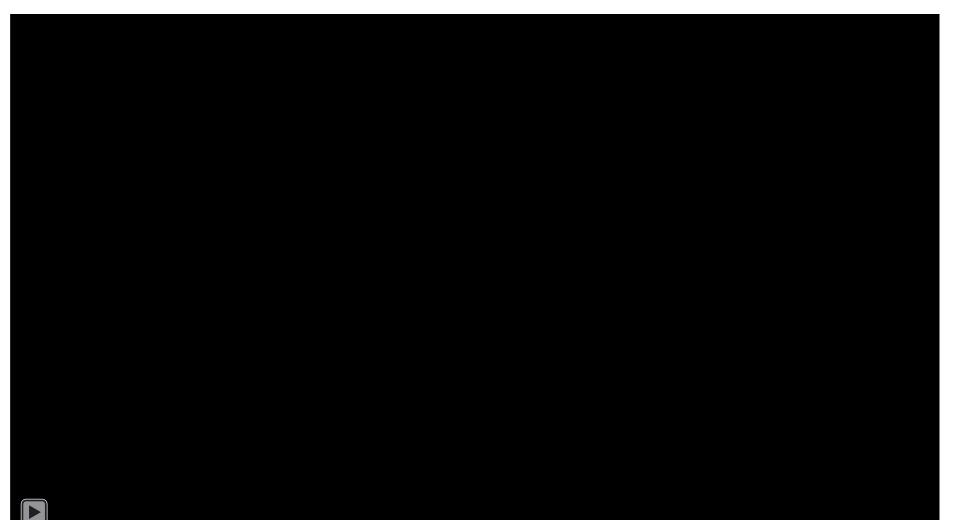
## **MATLAB Example**

### Analyze Airline Flight Data using MapReduce

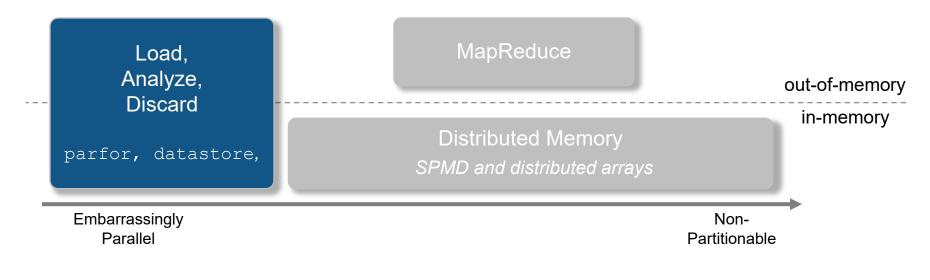
📣 MATU	AB R2014	4b		-		-		_				_ 0 X	
Пом	AC	PLOTS	APP3	SIICRTCUTS	EDITOR	PUDLISII	VIEW			1 <b>1</b> 9 c <b>2</b> 0	Search Documentatio	on 🔎 🗖	
	Dpen Sa	Find Files	Gr Tr V Gr Tr V Gr Find V	Insert 🛃 fx ( Comment % ‰ 3 Indent F 📲 (	Breakpoints	<ul> <li>Advar</li> </ul>	nd 🛃 Advance	Run and Time					
🔏 Editor				es\airlineMapReduce m	and the second		KUN					⊙ x	
airlin	airlineMapReduce.m × +												
1	- <b>%</b>	% Flight De	elay Dis	tributions									
2	8	% Get the flight data from a CSV file											
3 —	d	<pre>ds = datastore('airlinesmall.csv', 'TreatAsMissing', 'NA','SelectedVariableNames','ArrDelay');</pre>											
4													
5	응	% Define the Map and Reduce functions											
6 —	e	edges = -60:599;											
7 -	m.	<pre>mapFunction = @(data, into, intermKVstore) visualizationMapper(data, into, intermKVstore, edges);</pre>											
8 —	r	reduceFunction = @visualizationReducer;											
9													
10	응	% Call the "mapreduce" function I											
11 -	d	<pre>disp(sprintf('\nMapReduce job for calculating Flight Delay Distributions\n'));</pre>											
12 -	d	<pre>delayData = mapreduce(ds, mapFunction, reduceFunction);</pre>											
13													
14	8	% Plot the distribution of Flight Delays											
15 -	<sup>_</sup> p	plotDelays(delayData,edges);											
16													
17	- 8	% Flights ]	per Day	- Moving Av	erage								
18	용	% Reset the current datastore object and choose the variables you need											
	) r	eset (ds) ;											
		a Coloatodi	Tamiable	Name a fly	and Man	+ 6 1 175-	fManth!	L TI SS S	mic Commissille			*	
•									2	cript	Ln	1 Col 30	



#### **Demo: Vehicle Registry Analysis** *Running on Hadoop*







#### Complexity



## 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	
light Recent Places		🔍 1990.csv	8/13/2014 3:45 PM	WinZip File	50,822 KB	
🗂 Libraries		🔍 1991.csv	8/13/2014 3:43 PM	WinZip File	48,709 KB	
Documents	=	🔍 1992.csv	8/13/2014 3:46 PM	WinZip File	48,869 KB	
Music		🌒 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	
Tomegroup		🔍 1998.csv	8/13/2014 4:06 PM	WinZip File	74,887 KB	
Computer	-	•			,	

>> preview(ds)										
ans =										
Year	Month	DayofMonth	DayOfWeek							
1987	10	21	3							
1987	10	26	1							
1987	10	23	5							
1987	10	23	5							

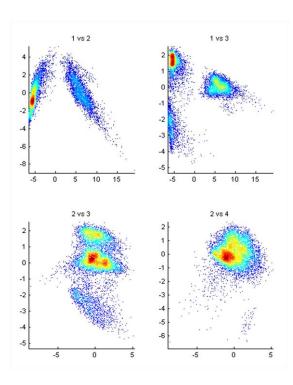
```
airdata = datastore('*.csv');
airdata.SelectedVariables = {'Distance', 'ArrDelay`};
data = read(airdata);
```



## **Big Data Capabilities in MATLAB**

#### **Memory and Data Access**

- 64-bit processors
- Memory Mapped Variables
- Disk Variables
- Databases
- Datastores R2014b



#### **Programming Constructs**

- Streaming
- Block Processing
- Parallel-for loops
- GPU Arrays
- SPMD and Distributed Arrays
- MapReduce R2014b

#### **Platforms**

- Desktop (Multicore, GPU)
- Clusters
- Cloud Computing (MDCS on EC2)
- Hadoop R2014b



out-of-memory

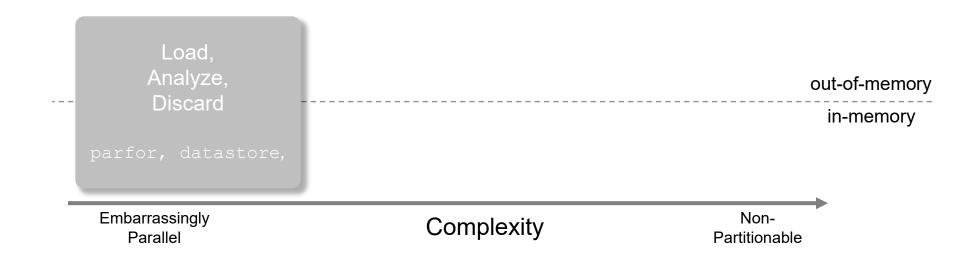
in-memory

Embarrassingly Parallel

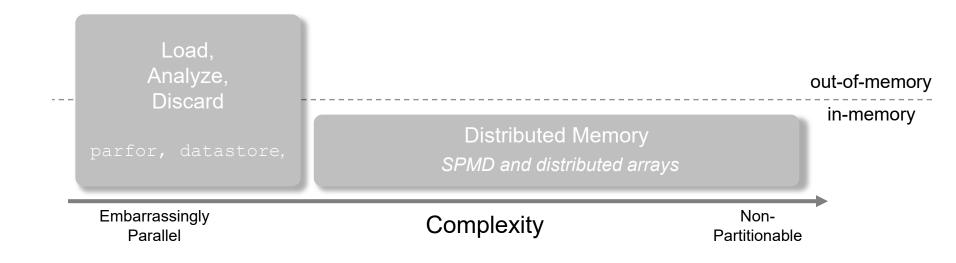
Complexity

Non-Partitionable

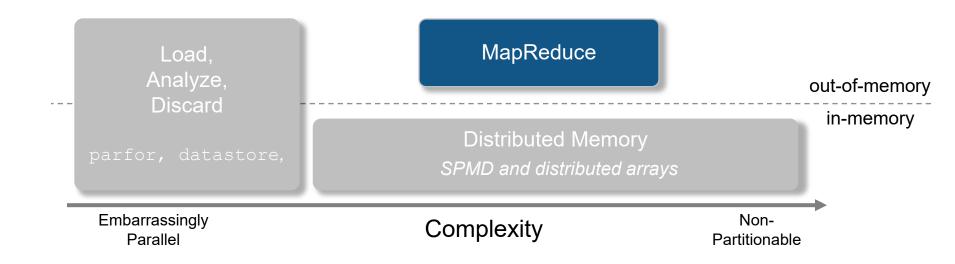














# Analyze Big Data

- Use the powerful MapReduce programming technique to analyze big data
  - mapreduce uses a datastore to process data in small chunks that individually fit into memory
  - Useful for processing multiple keys, or when Intermediate results do not fit in memory
- mapreduce on the desktop
  - Increase compute capacity (Parallel Computing Toolbox)
  - Access data on HDFS to develop algorithms for use on Hadoop
- mapreduce with Hadoop
  - Run on Hadoop using MATLAB Distributed Computing Server
  - Deploy applications and libraries for Hadoop using MATLAB Compiler

*	MAPH	REDUCE PRO	GRESS *
* * * :	* * * * * * * *	*******	*********
Мар	0%	Reduce	0%
Мар	20%	Reduce	0%
Мар	40%	Reduce	0%
Мар	60%	Reduce	0%
Мар	80%	Reduce	0%
Мар	100%	Reduce	25%
Мар	100%	Reduce	50%
Мар	100%	Reduce	75%
Мар	100%	Reduce	100%



#### Data Store







# mapreduce Data Store Map Veh\_typ Q3\_08 Q4\_08 Q1\_09 Hybrid



# Data Store





Veh_typ	Q3_08	Q4_08	Q1_09	Hybrid	
Car	1	1	1	0	
SUV	0	1	1	1	
Car	1	1	1	1	
Car	0	0	1	1	
Car	0	1	1	1	
Car	1	1	1	1	



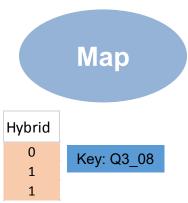
	Dat	ta Sto	ore	
Veh_typ	Q3_08	Q4_08	Q1_09	Hybrid
Car	1	1	1	0
SUV	0	1	1	1

Car

Car

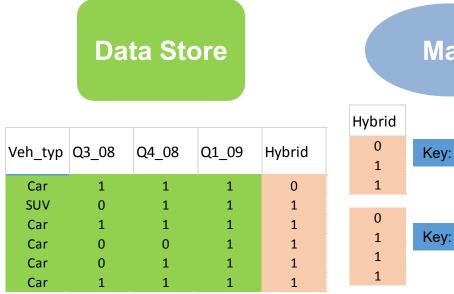
Car

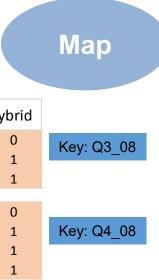
Car





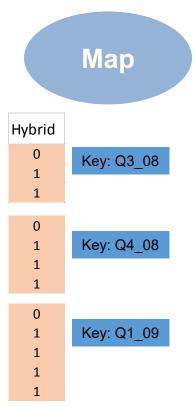






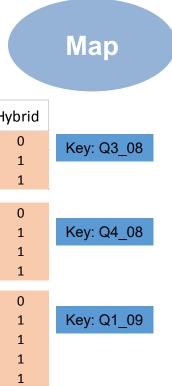


Data Store										
Veh_typ	Q3_08	Q4_08	Q1_09	Hybrid						
Car	1	1	1	0						
SUV	0	1	1	1						
Car	1	1	1	1						
Car	0	0	1	1						
Car	0	1	1	1						
Car	1	1	1	1						



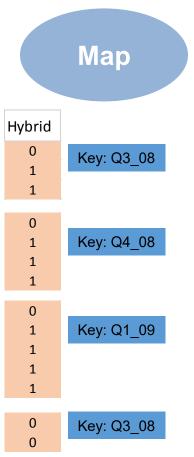


	Data Store										
[						F					
1	Veh_typ	Q3_08	Q4_08	Q1_09	Hybrid						
	Car	1	1	1	0						
	SUV	0	1	1	1						
	Car	1	1	1	1						
	Car	0	0	1	1						
	Car	0	1	1	1						
	Car	1	1	1	1						
	Car	0	0	1	1						
	SUV	0	1	1	0						
	Car	1	1	1	0						
	SUV	1	1	1	1						
	Car	0	1	1	1						
	Car	1	0	0	0						



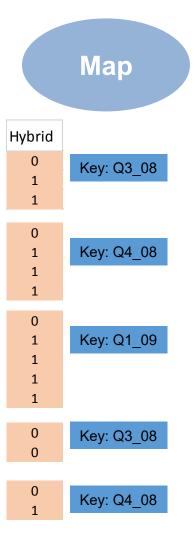


Data Store									
Veh_typ	Q3_08	Q4_08	Q1_09	Hybrid					
Car	1	1	1	0					
SUV	0	1	1	1					
Car	1	1	1	1					
Car	0	0	1	1					
Car	0	1	1	1					
Car	1	1	1	1					
Car	0	0	1	1					
SUV	0	1	1	0					
Car	1	1	1	0					
SUV	1	1	1	1					
Car	0	1	1	1					
Car	1	0	0	0					



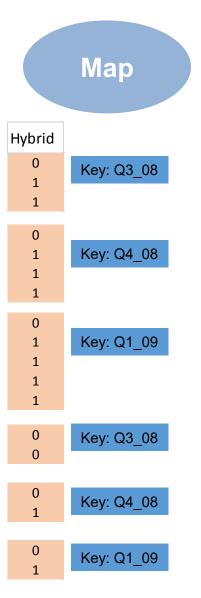


Data Store										
Veh_typ	Q3_08	Q4_08	Q1_09	Hybrid						
Car	1	1	1	0						
SUV	0	1	1	1						
Car	1	1	1	1						
Car	0	0	1	1						
Car	0	1	1	1						
Car	1	1	1	1						
Car	0	0	1	1						
SUV	0	1	1	0						
Car	1	1	1	0						
SUV	1	1	1	1						
Car	0	1	1	1						
Car	1	0	0	0						

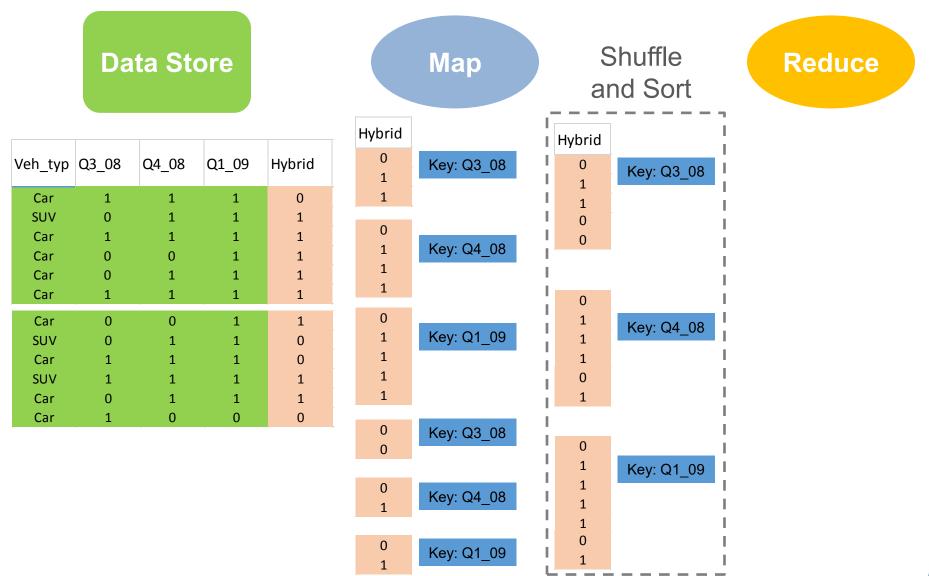




Data Store										
Veh_typ	Q3_08	Q4_08	Q1_09	Hybrid						
Car	1	1	1	0						
SUV	0	1	1	1						
Car	1	1	1	1						
Car	0	0	1	1						
Car	0	1	1	1						
Car	1	1	1	1						
Car	0	0	1	1						
SUV	0	1	1	0						
Car	1	1	1	0						
SUV	1	1	1	1						
Car	0	1	1	1						
Car	1	0	0	0						









Data Store				Мар		Shuffle nd Sort		Reduce		
Veh_typ	Q3_08	Q4_08	Q1_09	Hybrid	Hybrid 0 1	Key: Q3_08	Hybrid 0 1	Key: Q3_08	Кеу	% Hybrid (Value)
Car SUV Car Car Car	1 0 1 0 0	1 1 1 0 1	1 1 1 1 1	0 1 1 1 1	1 0 1 1	Key: Q4_08			1     	
Car Car SUV Car SUV	1 0 0 1 1	1 0 1 1 1	1 1 1 1 1 1	1 1 0 0 1	1 0 1 1 1	Key: Q1_09	0 1 1 1 0	Key: Q4_08	1       	
Car Car	0	1 0	1	1 0	1 0 0	Key: Q3_08	1 0 1	Key: Q1_09	     	
					0 1 0 1	Key: Q4_08 Key: Q1_09	1 1 0 1		     	57



Data Store				Мар		Shuffle nd Sort	R	Reduce		
					Hybrid	·	I I Hybrid			
Veh_typ	Q3_08	Q4_08	Q1_09	Hybrid	0 1	Key: Q3_08	0	Key: Q3_08	Кеу	% Hybrid (Value)
Car SUV	1 0	1 1	1 1	0 1	1		1		Q3_08	0.4
Car	1	1	1	1	0 1	Key: Q4_08	0			
Car Car	0 0	0 1	1 1	1 1	1	<u>_</u>				
Car Car	1	1	1	1	0		0 1			
SUV	0	1	1	0	1	Key: Q1_09	1	Key: Q4_08		
Car SUV	1 1	1 1	1 1	0 1	1					
Car Car	0 1	1 0	1 0	1 0	1		1 			
					0 0	Key: Q3_08	0			
					0 1	Key: Q4_08	1 1 1 1	Key: Q1_09	1   	
					0 1	Key: Q1_09	0 1		   ,	58



	Da	ita St	ore			Мар		Shuffle nd Sort	R	educe
					Hybrid		Hybrid			
Veh_typ	Q3_08	Q4_08	Q1_09	Hybrid	0 1	Key: Q3_08	0 1 1	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 1	0		0		Q4_08	0.67
Car	0	0	1	1	1	Key: Q4_08	0			
Car	0	1	1	1	1		1			
Car	1	1	1	1	1		0		I	
Car	0	0	1	1	0		1	Key: Q4_08	I	
SUV	0	1	1	0	1	Key: Q1_09	1	1.09. Q1_00		
Car	1	1	1	0	1		1			
SUV	1	1	1	1	1		0		1	
Car	0	1	1	1	1		1			
Car	1	0	0	0	0	Key: Q3_08		, 		
					0	, <u>,</u>	0		I	
							1	Key: Q1_09		
					0	Key: Q4_08	1			
					1		1			
					0					
					0 1	Key: Q1_09	1			
					-		<u> </u>			59



	Da	ita St	ore			Мар		Shuffle nd Sort	R	educe
					Hybrid		l Hybrid			
Veh_typ	Q3_08	Q4_08	Q1_09	Hybrid	0	Key: Q3_08	0	Key: Q3_08	Кеу	% Hybrid (Value)
Car	1	1	1	0	1				Q3 08	0.4
SUV	0	1	1	1	0		0	I	Q4 08	0.67
Car Car	1 0	1 0	1 1	1 1	1	Key: Q4_08	0		Q1_09	0.75
Car	0	1	1	1	1		i i			
Car	1	1	1	1	1		0	i		
Car	0	0	1	1	0		1	Key: Q4_08		
SUV	0	1	1	0	1	Key: Q1_09	1			
Car	1	1	1	0	1		1			
SUV	1	1	1	1	1		0 1	i		
Car Car	0 1	1 0	1 0	1 0				I		
	-			Ŭ	0	Key: Q3_08				
					0		0 1	Kaun 04,00		
					0		1	Key: Q1_09		
					1	Key: Q4_08	1	l		
							1	I		
					0	Key: Q1_09	0 1			
					1		L			60

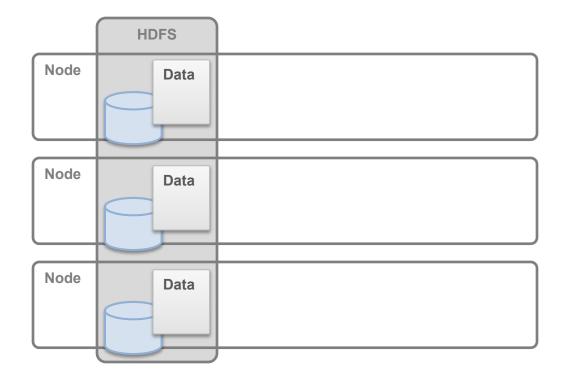






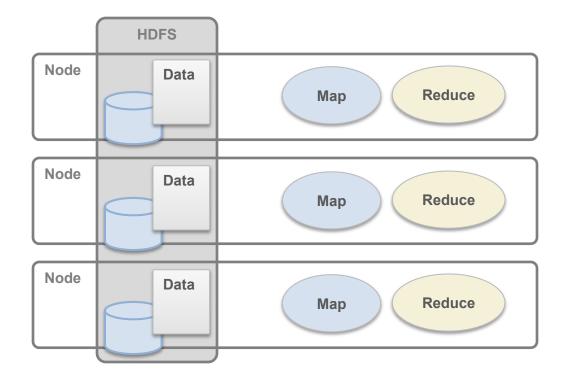






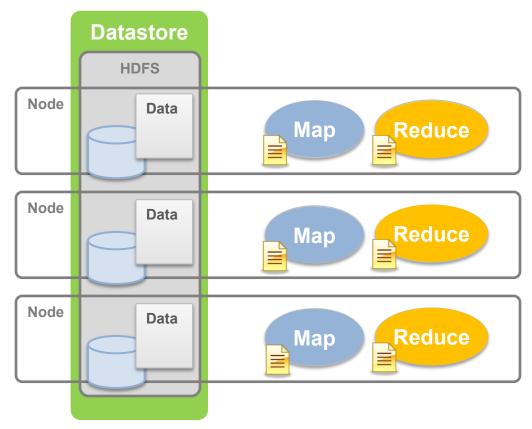






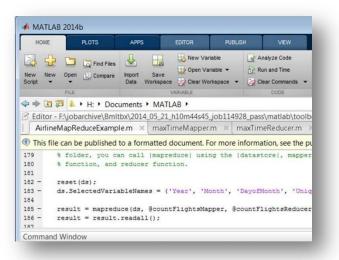


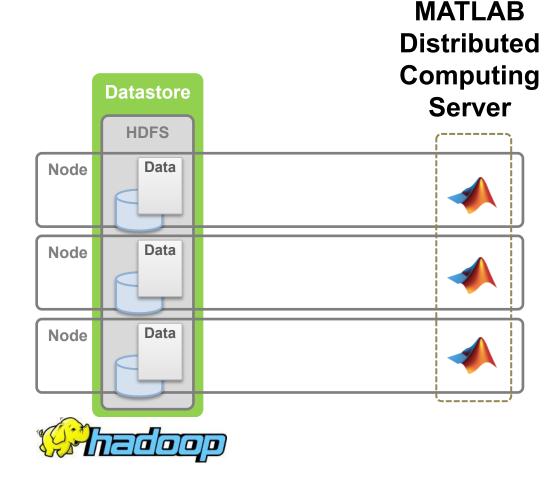






## **Explore and Analyze Data on Hadoop**

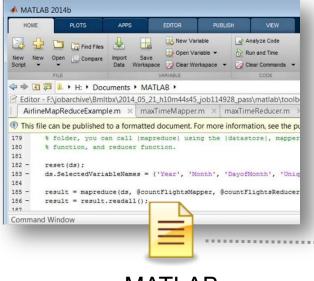




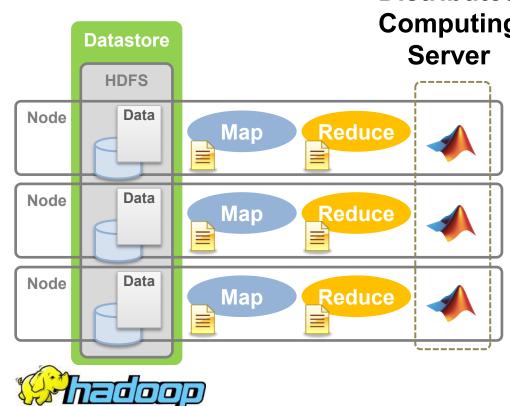


## **Explore and Analyze Data on Hadoop**

MATLAB Distributed Computing

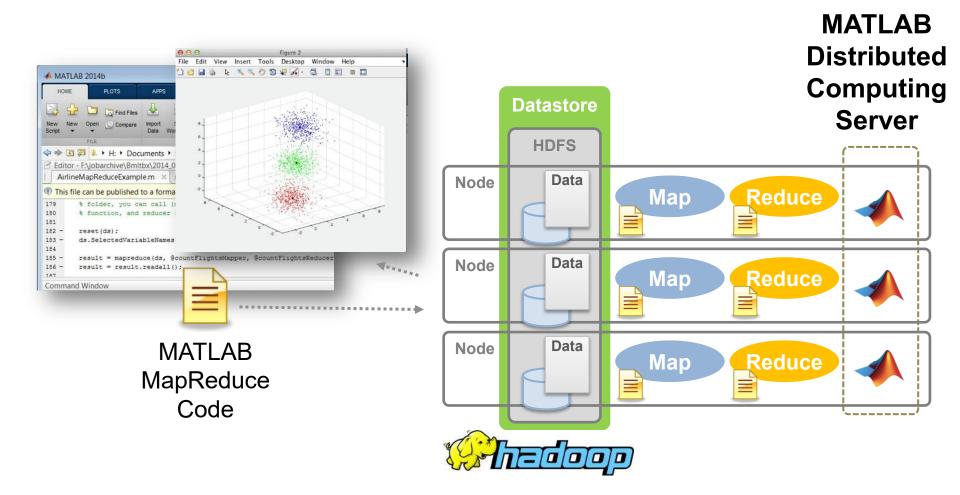


MATLAB MapReduce Code

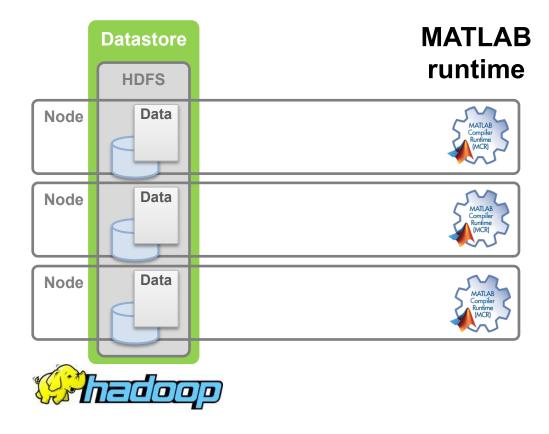




## **Explore and Analyze Data on Hadoop**







HOME	PLOTS	APPS	EDITOR	PUBLE	SH	VIEW
New New C	pen	Import Save Data Workspace	New Va	iriable 👻	& Ru	alyze Code n and Time ar Commands 👻
	ILE		VARIABLE			CODE
	L + H: + Doc	uments • MATI	LAB +			
Editor - E1	inharchive Rmlt	by) 2014 05 21	h10m/4c45	ioh11/0	28 0355	matiah) tool
		tbx\2014_05_21				
		tbx\2014_05_21 le.m × maxT				
AirlineMa	pReduceExamp		imeMapper	.m × n	naxTime	Reducer.m
AirlineMa This file ca	pReduceExamp n be published	le.m × maxT	imeMapper document. F	or more i	naxTime nformat	eReducer.m tion, see the p
AirlineMa This file ca	pReduceExamp n be published folder, you ca	le.m 🙁 maxT to a formatted	imeMapper document. F educe   usi	or more i	naxTime nformat	eReducer.m tion, see the p
AirlineMa This file ca	pReduceExamp n be published folder, you ca	le.m × maxT to a formatted an call (mapre	imeMapper document. F educe   usi	or more i	naxTime nformat	eReducer.m tion, see the p
AirlineMa This file ca This file ca 179 180 181	pReduceExamp n be published folder, you ca	le.m × maxT to a formatted an call (mapre	imeMapper document. F educe   usi	or more i	naxTime nformat	eReducer.m tion, see the p
AirlineMa This file ca 179 180 181 182 - res	pReduceExamp n be published older, you ca function, and set (ds);	le.m × maxT to a formatted an call (mapre	imeMapper document. F educe   usi tion.	or more i	naxTime nformat datast	eReducer.m tion, see the p ore], mappe
AirlineMa This file ca 179 180 181 182 - res	pReduceExamp n be published older, you ca function, and set (ds);	le.m × maxT to a formatted an call [mapre reducer funct	imeMapper document. F educe   usi tion.	or more i	naxTime nformat datast	eReducer.m tion, see the p ore], mappe
AirlineMa This file ca 179 180 181 182 - rei 183 - ds 184	pReduceExamp n be published folder, you co function, and set(ds); SelectedVaria	le.m × maxT to a formatted an call [mapre reducer funct	imeMapper document.F educe  usi tion. 'Year', 'M	or more ing the I	naxTime nformal datast DayofM	eReducer.m tion, see the p ore[, mappe orth', 'Uni



HOME

179

180

181 182 -

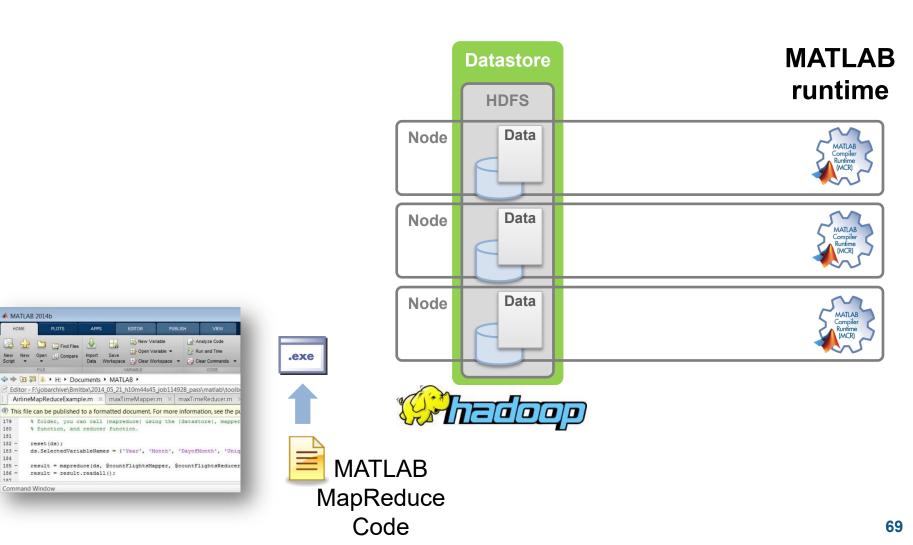
183 -

185 -

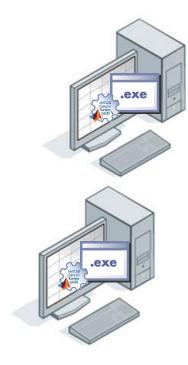
186 -

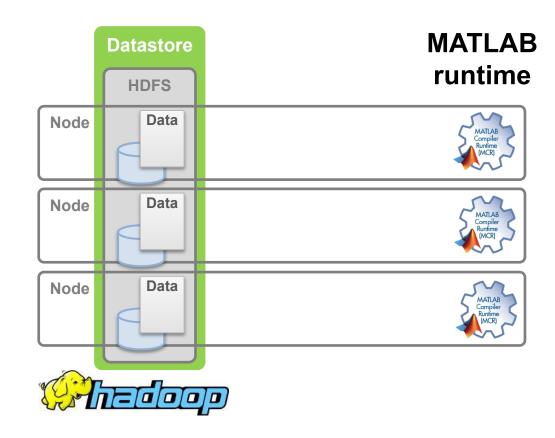
187

184

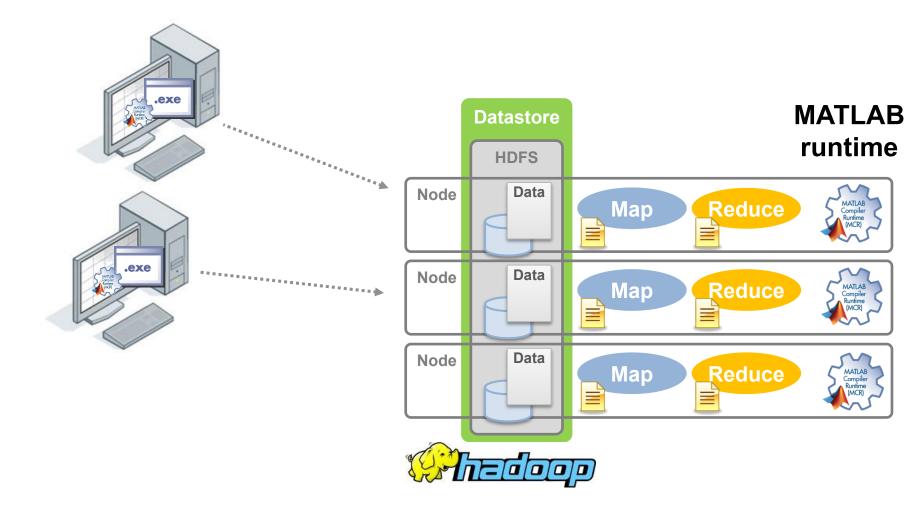




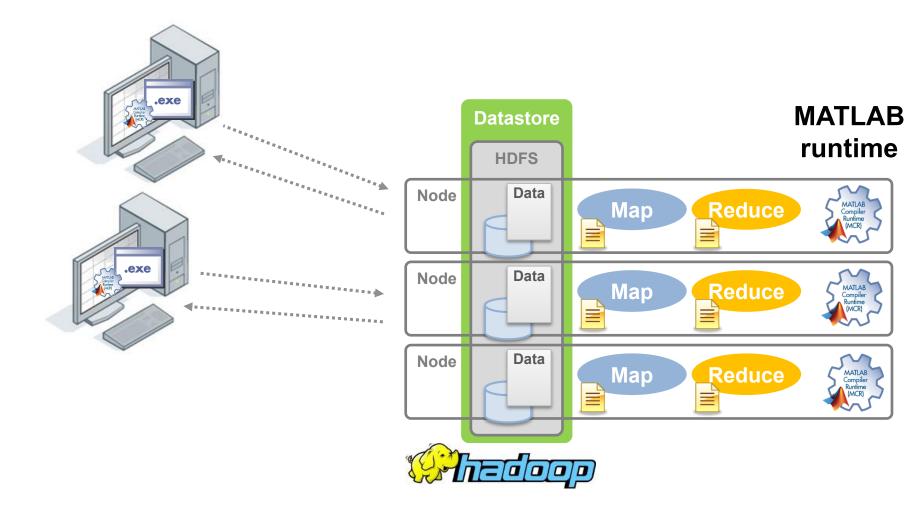














## **Big Data on the Desktop**



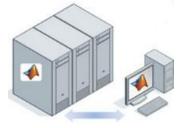
- Expand workspace
  - 64 bit processor support increased in-memory data set handling
- Access portions of data too big to fit into memory
  - Memory mapped variables huge binary file
  - Datastore huge text file or collections of text files R2014b
  - Database query portion of a big database table
- Variety of programming constructs
  - System Objects analyze streaming data
  - MapReduce process text files that won't fit into memory R2014b
- Increase analysis speed
  - Parallel for-loops with multicore/multi-process machines
  - GPU Arrays

## **Further Scaling Big Data Capacity**

MATLAB supports a number of programming constructs for use with clusters

- General compute clusters
  - Parallel for-loops embarrassingly parallel algorithms
  - SPMD and distributed arrays distributed memory
- Hadoop clusters
  - MapReduce analyze data stored in the Hadoop Distributed File System

Use these constructs on the desktop to develop your algorithms Migrate to a cluster without algorithm changes









#### 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
  - www.mathworks.com/discovery/matlab-mapreduce-hadoop.html

#### 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.



Create applications and libraries based upon MATLAB MapReduce for deployment within production instances of Hadoop, using MATLAB Compiler.

» Deploy MATLAB MapReduce applications to Hadoop