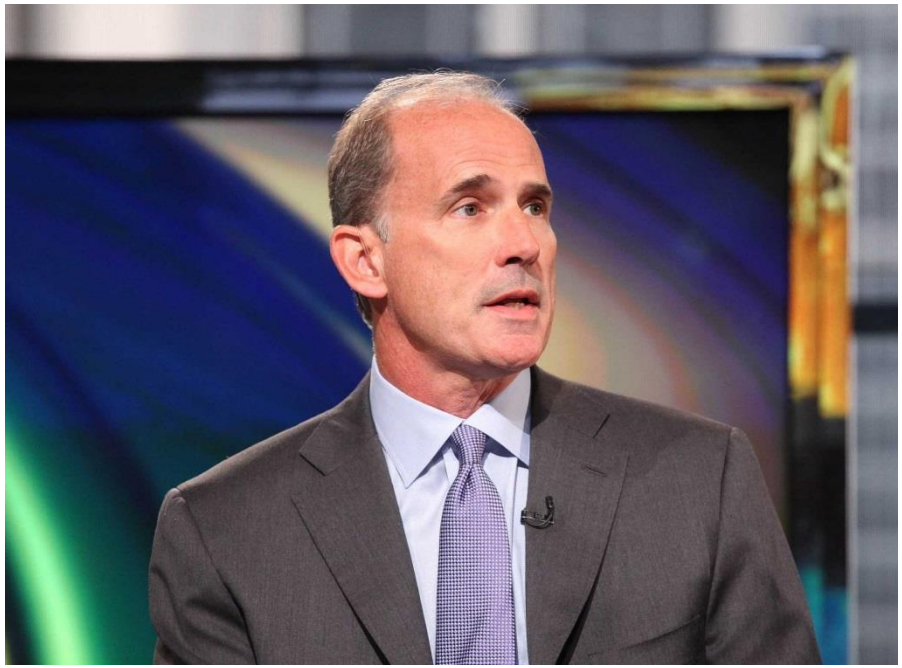
Big Data & Predictive Analytics

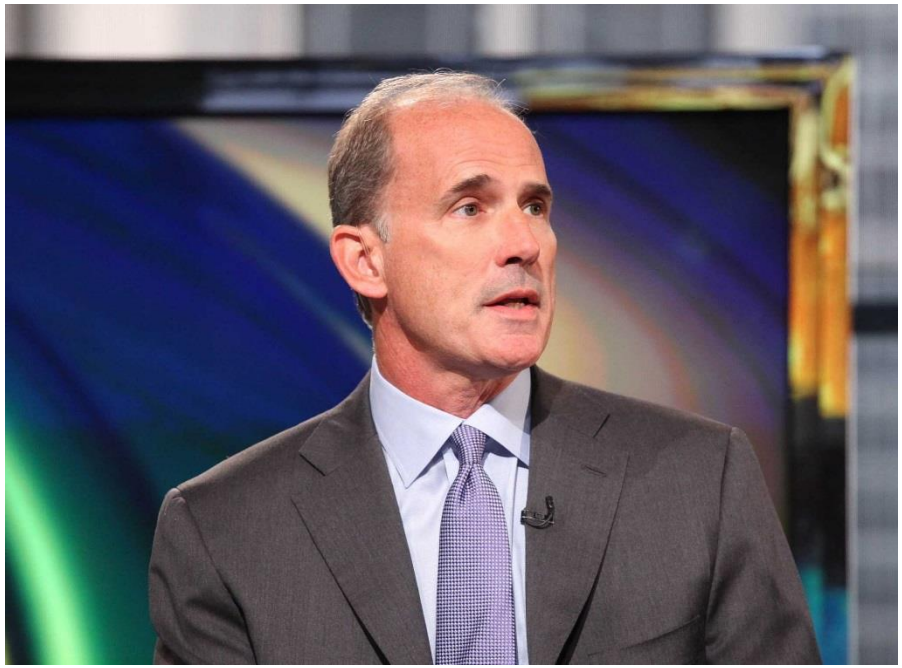
David Willingham
Senior Application Engineer, MathWorks
david.willingham@mathworks.com.au

"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.”

Big Data & Predictive Analytics

Big Data & Predictive Analytics

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

Big Data & Predictive Analytics

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






Big Data & Predictive Analytics

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






- How can I work with large data sets?

Big Data & Predictive Analytics

- 2012: 2.5 billion GB ($2.5 \; x \; 10^{18}$) of data each day.






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

Big Data & Predictive Analytics

- 2012: 2.5 billion GB ($2.5\ x\ 10^{18}$) 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?

What is BIG in Big Data?

What is BIG in Big Data?

Wikipedia

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

What is BIG in Big Data?

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*

What is BIG in Big Data?

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 4th V,
***Value**: What Business Information can be obtained from Big Data.*

Who is looking to do Big Data Analytics now?

Who is looking to do Big Data Analytics now?

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

Who is looking to do Big Data Analytics now?

- 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

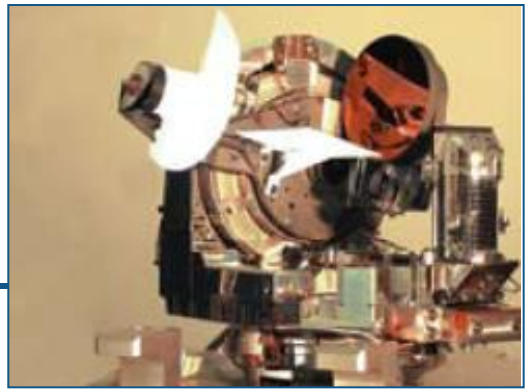
Who is looking to do Big Data Analytics now?

- 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



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

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-the-loop 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

"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

Link to user story

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

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

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

"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
BuildingIQ

Link to user story

Tesco uses supply chain analytics to save £100m a year.

Tesco uses supply chain analytics to save £100m a year.

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

Tesco uses supply chain analytics to save £100m a year.

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

Tesco uses supply chain analytics to save £100m a year.

- 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.”

Preventing lethal ship collisions with whales

Preventing lethal ship collisions with whales

- Cornell University collected terabytes of ocean acoustic data.

Preventing lethal ship collisions with whales

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

Preventing lethal ship collisions with whales

- 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

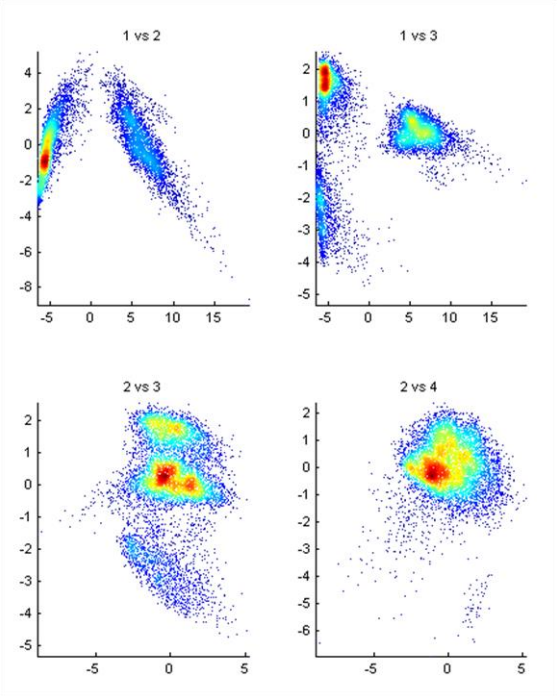
"Any collection of data sets so large and complex that it becomes difficult to process using … traditional data processing applications."
(Wikipedia)

- 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



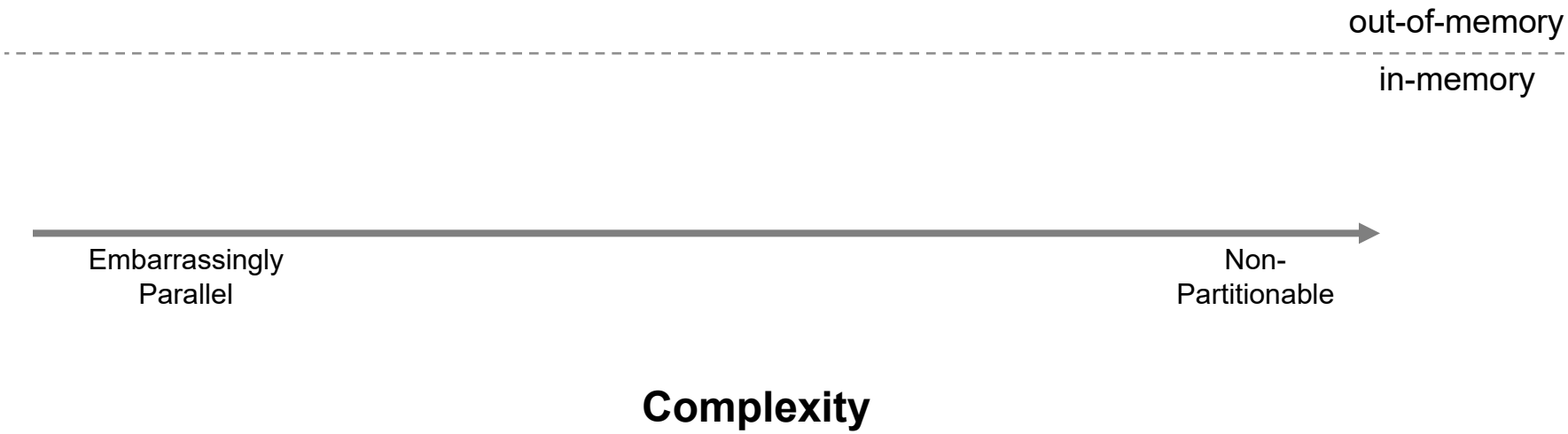
Programming Constructs

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

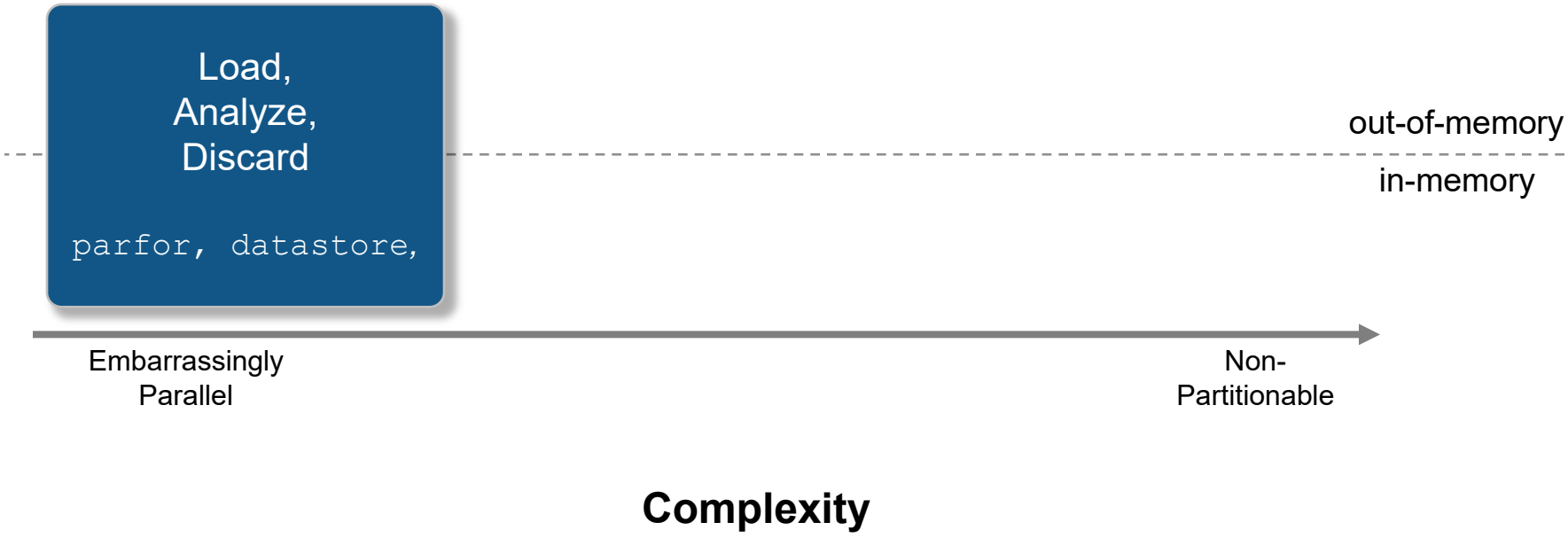
Platforms

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

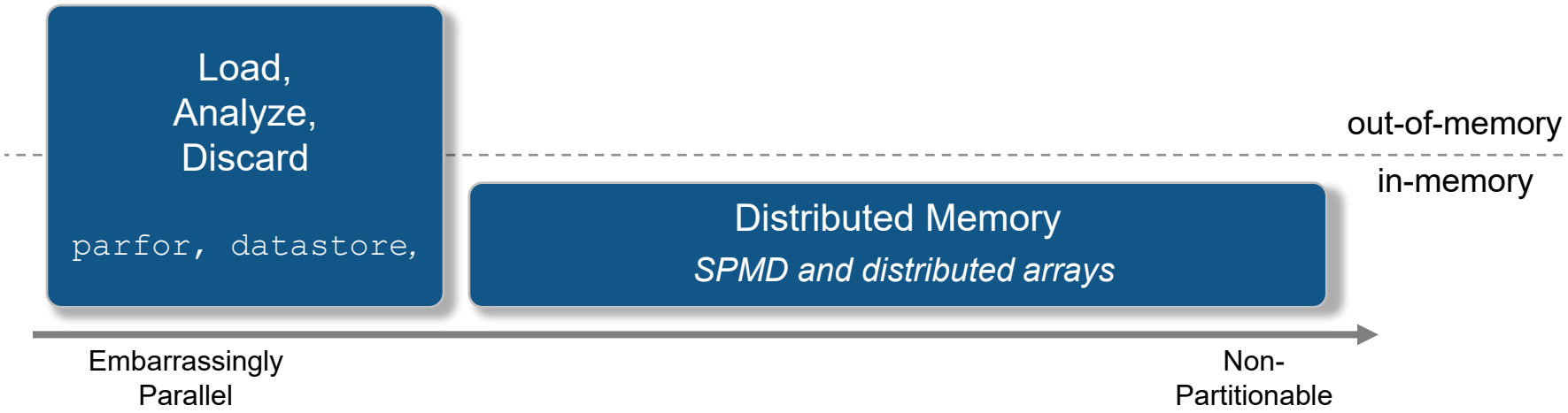
Techniques for Big Data in MATLAB



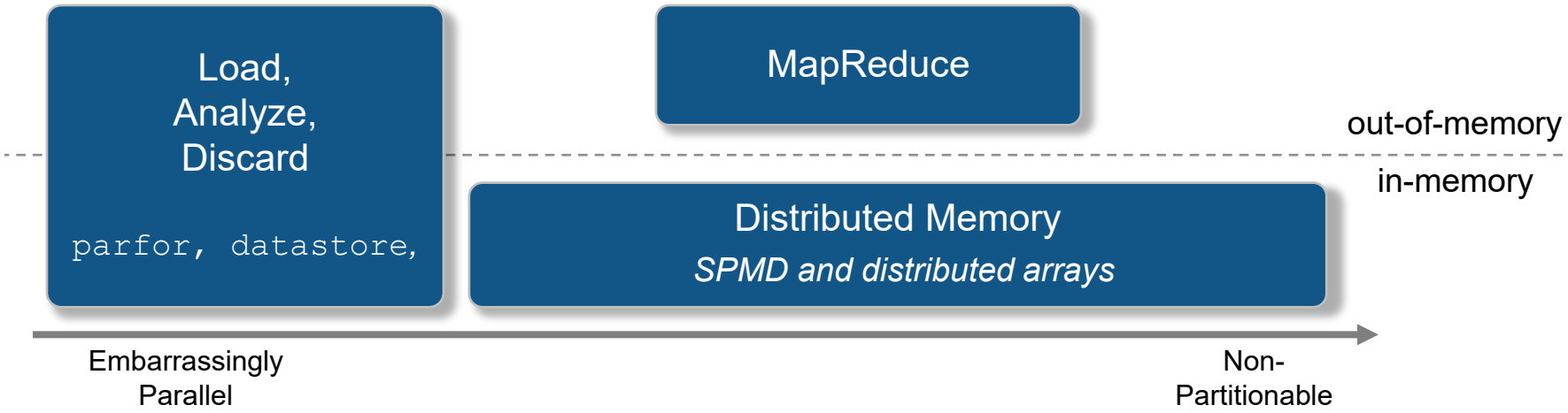
Techniques for Big Data in MATLAB



Techniques for Big Data in MATLAB

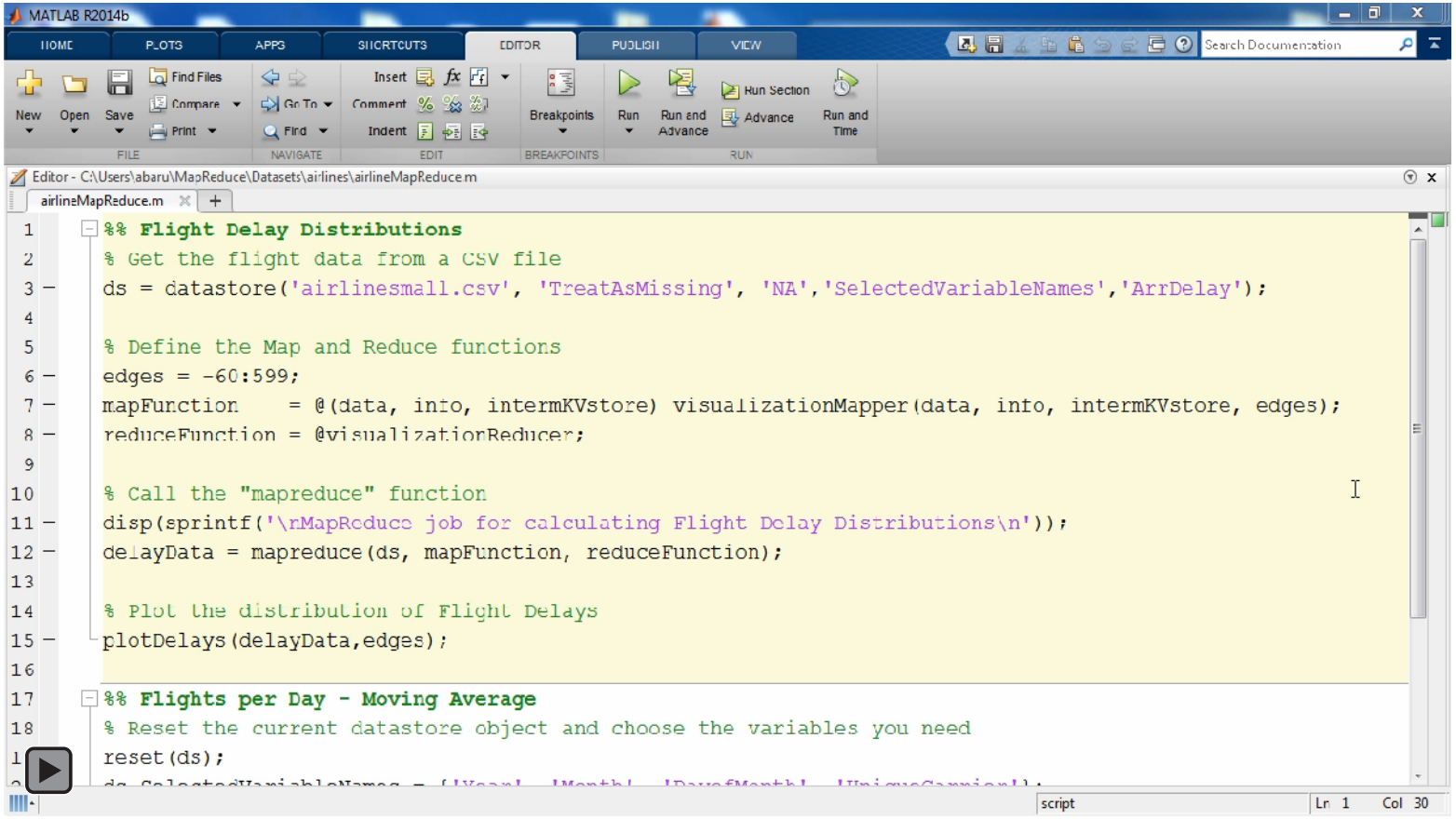


Techniques for Big Data in MATLAB



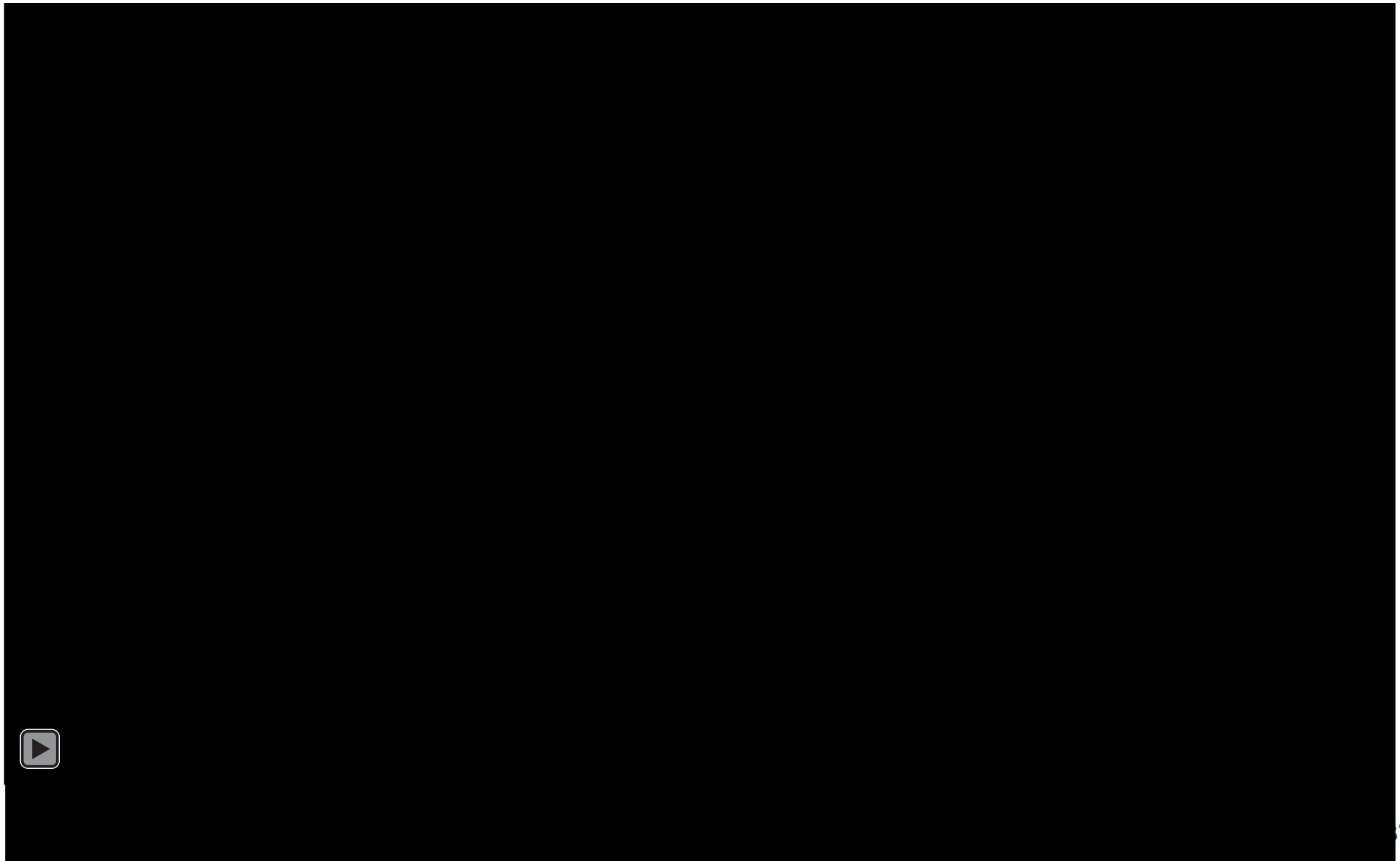
MATLAB Example

Analyze Airline Flight Data using MapReduce

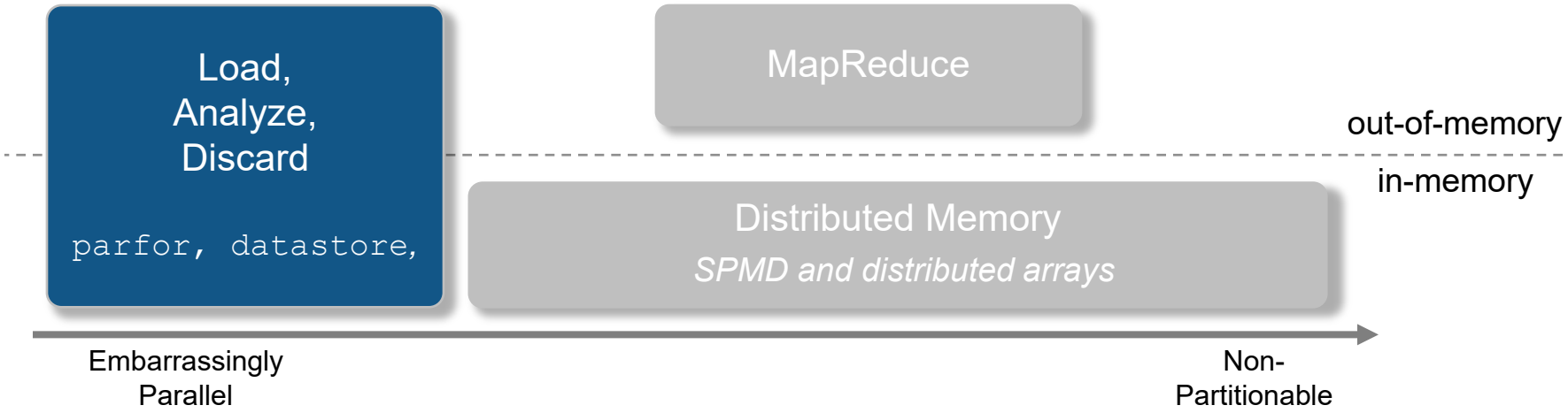


Demo: Vehicle Registry Analysis

Running on Hadoop



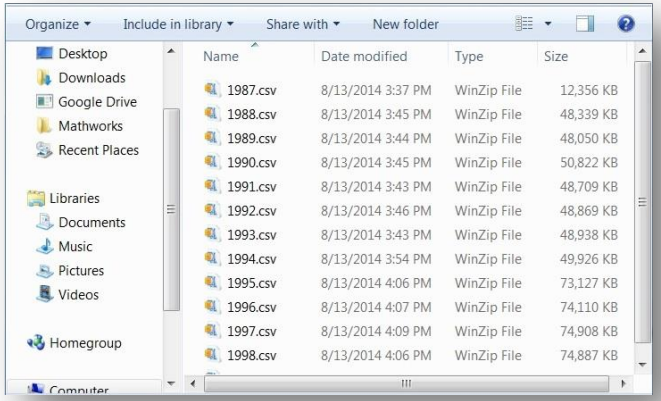
Techniques for Big Data in MATLAB



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




```
>> preview(ds)
ans =
    Year    Month    DayofMonth    DayOfWeek
    ____    _____    __________    _________
    1987    10       21            3
    1987    10       26            1
    1987    10       23            5
    1987    10       23            5
```

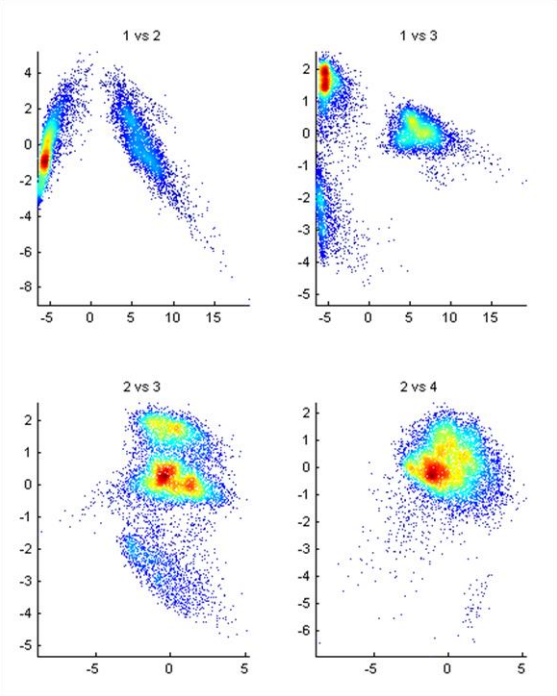
```
airdata = datastore('*.csv');
airdata.SelectedVariables = {'Distance', 'ArrDelay'};

data = read(airdata);
```

Big Data Capabilities in MATLAB

Memory and Data Access

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



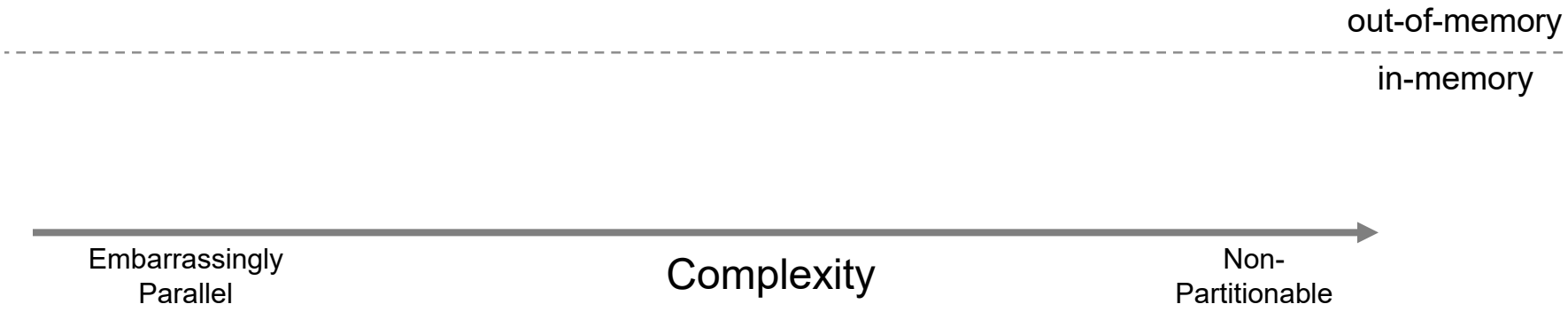
Programming Constructs

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

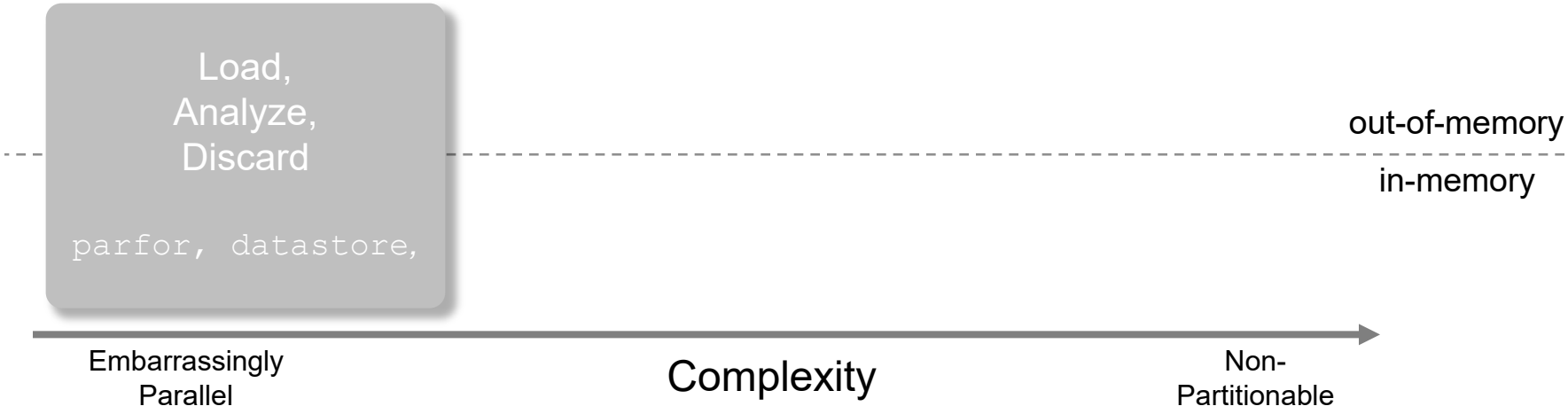
Platforms

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

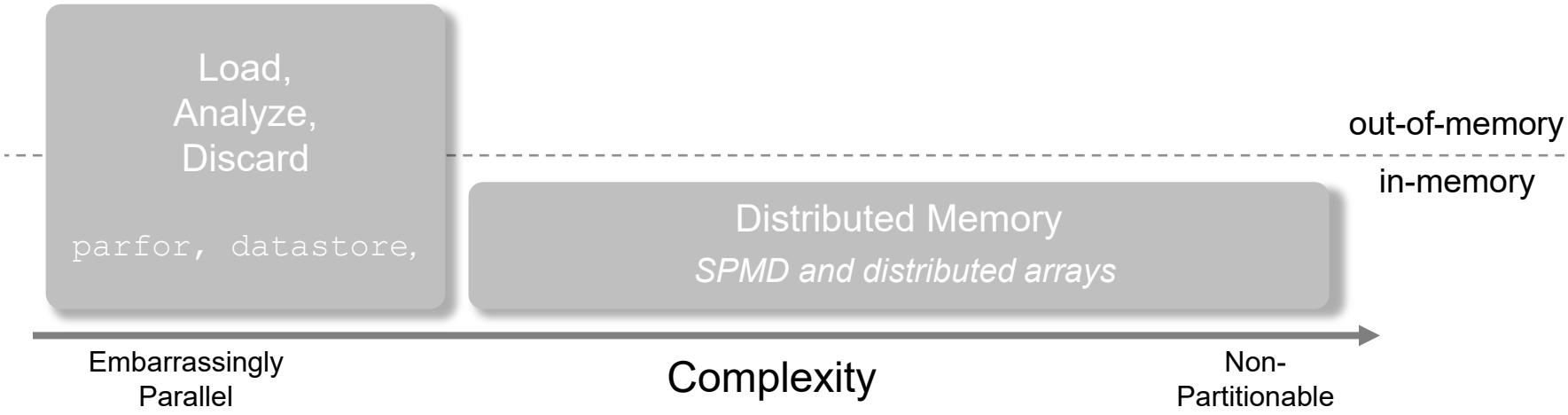
Techniques for Big Data in MATLAB



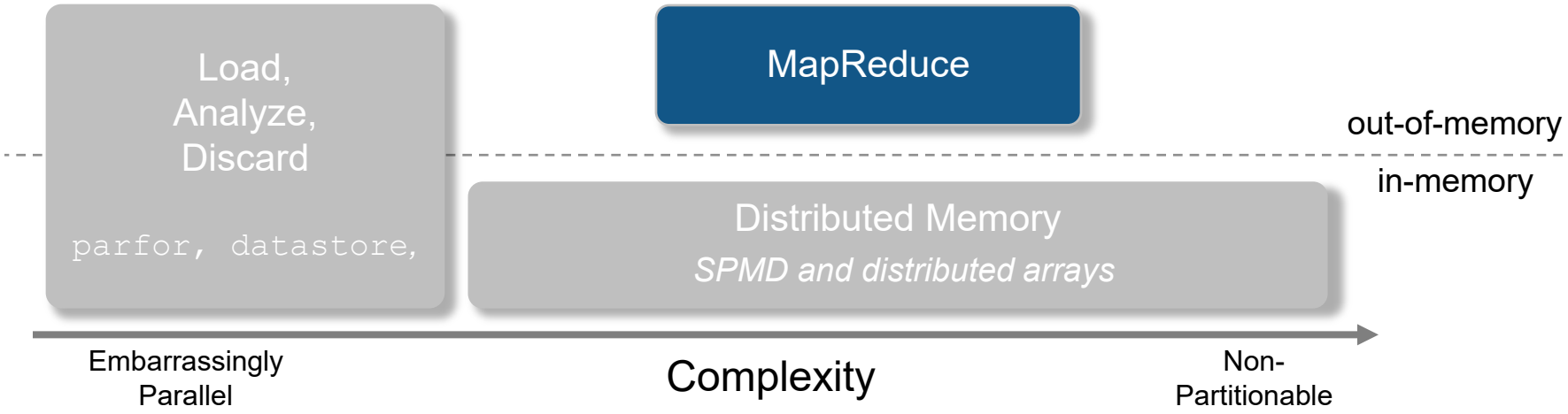
Techniques for Big Data in MATLAB



Techniques for Big Data in MATLAB



Techniques for Big Data in MATLAB



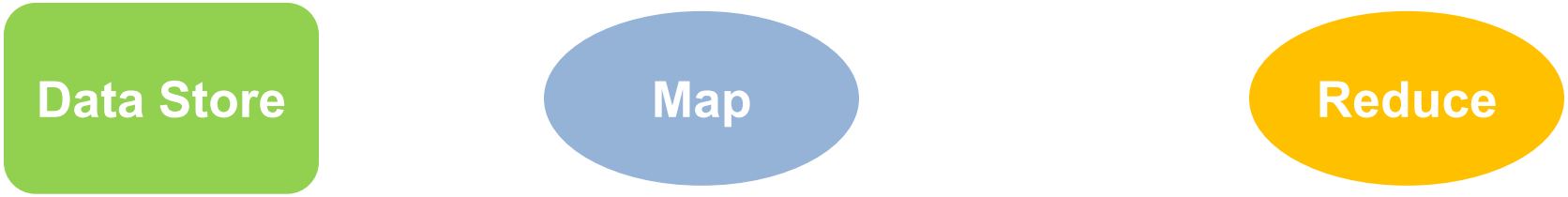
Analyze Big Data

`mapreduce`

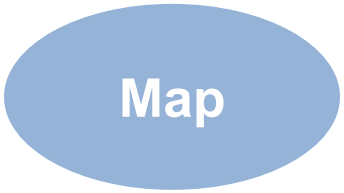
```
********************************
*      MAPREDUCE PROGRESS      *
********************************
Map 0%        Reduce 0%
Map 20%       Reduce 0%
Map 40%       Reduce 0%
Map 60%       Reduce 0%
Map 80%       Reduce 0%
Map 100%      Reduce 25%
Map 100%      Reduce 50%
Map 100%      Reduce 75%
Map 100%      Reduce 100%
```

- 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

`mapreduce`



mapreduce



Veh_typ	Q3_08	Q4_08	Q1_09	Hybrid

mapreduce

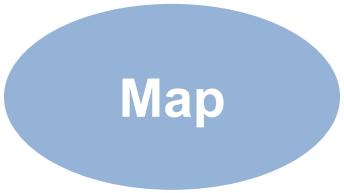
Data Store

Map

Reduce

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

mapreduce

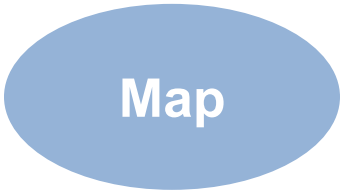


Veh_typ	Q3_08	Q4_08	Q1_09	Hybrid
Car	1	1	1	0
SUV	0	1	1	1
Car	1	1	1	1
Car	0	0	1	1
Car	0	1	1	1
Car	1	1	1	1

Hybrid
0
1
1

Key: Q3_08

mapreduce



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

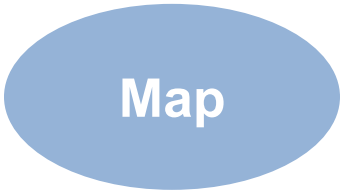
Hybrid
0
1
1

Key: Q3_08

Hybrid
0
1
1
1

Key: Q4_08

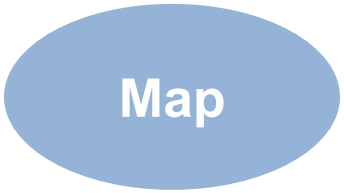
mapreduce



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

Hybrid	
0	Key: Q3_08
1	
1	
0	Key: Q4_08
1	
1	
1	
0	Key: Q1_09
1	
1	
1	
1	

mapreduce



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

Hybrid	
0	Key: Q3_08
1	
1	
0	Key: Q4_08
1	
1	
1	
0	Key: Q1_09
1	
1	
1	
1	

mapreduce

Data Store

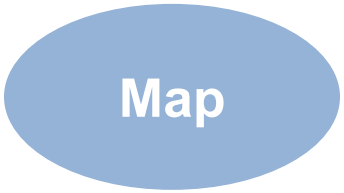
Veh_typ	Q3_08	Q4_08	Q1_09	Hybrid
Car	1	1	1	0
SUV	0	1	1	1
Car	1	1	1	1
Car	0	0	1	1
Car	0	1	1	1
Car	1	1	1	1
Car	0	0	1	1
SUV	0	1	1	0
Car	1	1	1	0
SUV	1	1	1	1
Car	0	1	1	1
Car	1	0	0	0

Map

Hybrid	
0	Key: Q3_08
1	
1	
0	Key: Q4_08
1	
1	
1	
0	Key: Q1_09
1	
1	
1	
1	
0	Key: Q3_08
0	

Reduce

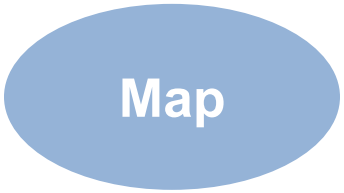
mapreduce



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

Hybrid	
0 1 1	Key: Q3_08
0 1 1 1	Key: Q4_08
0 1 1 1 1	Key: Q1_09
0 0	Key: Q3_08
0 1	Key: Q4_08

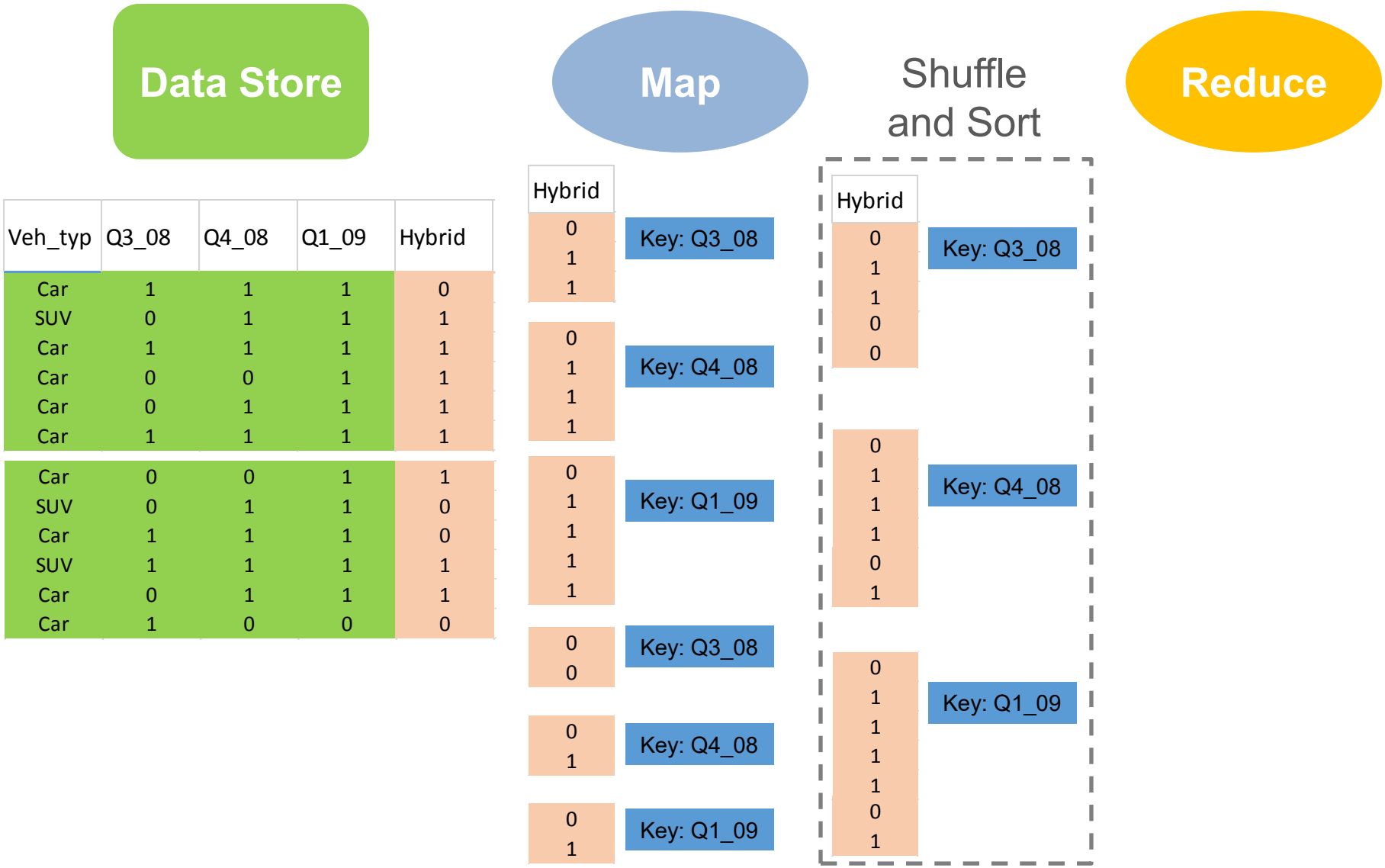
mapreduce



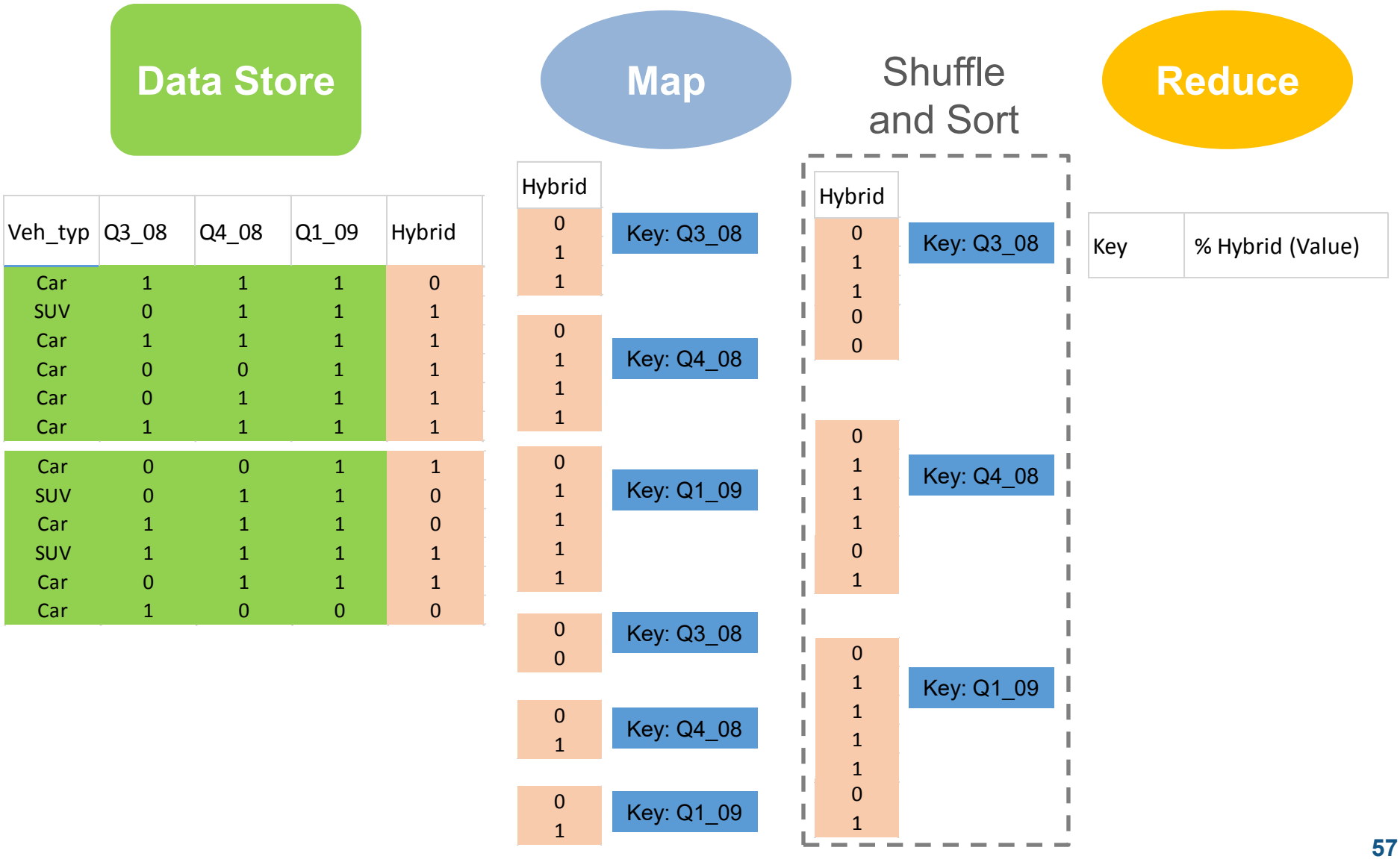
Veh_typ	Q3_08	Q4_08	Q1_09	Hybrid
Car	1	1	1	0
SUV	0	1	1	1
Car	1	1	1	1
Car	0	0	1	1
Car	0	1	1	1
Car	1	1	1	1
Car	0	0	1	1
SUV	0	1	1	0
Car	1	1	1	0
SUV	1	1	1	1
Car	0	1	1	1
Car	1	0	0	0

Hybrid	
0 1 1	Key: Q3_08
0 1 1 1	Key: Q4_08
0 1 1 1 1	Key: Q1_09
0 0	Key: Q3_08
0 1	Key: Q4_08
0 1	Key: Q1_09

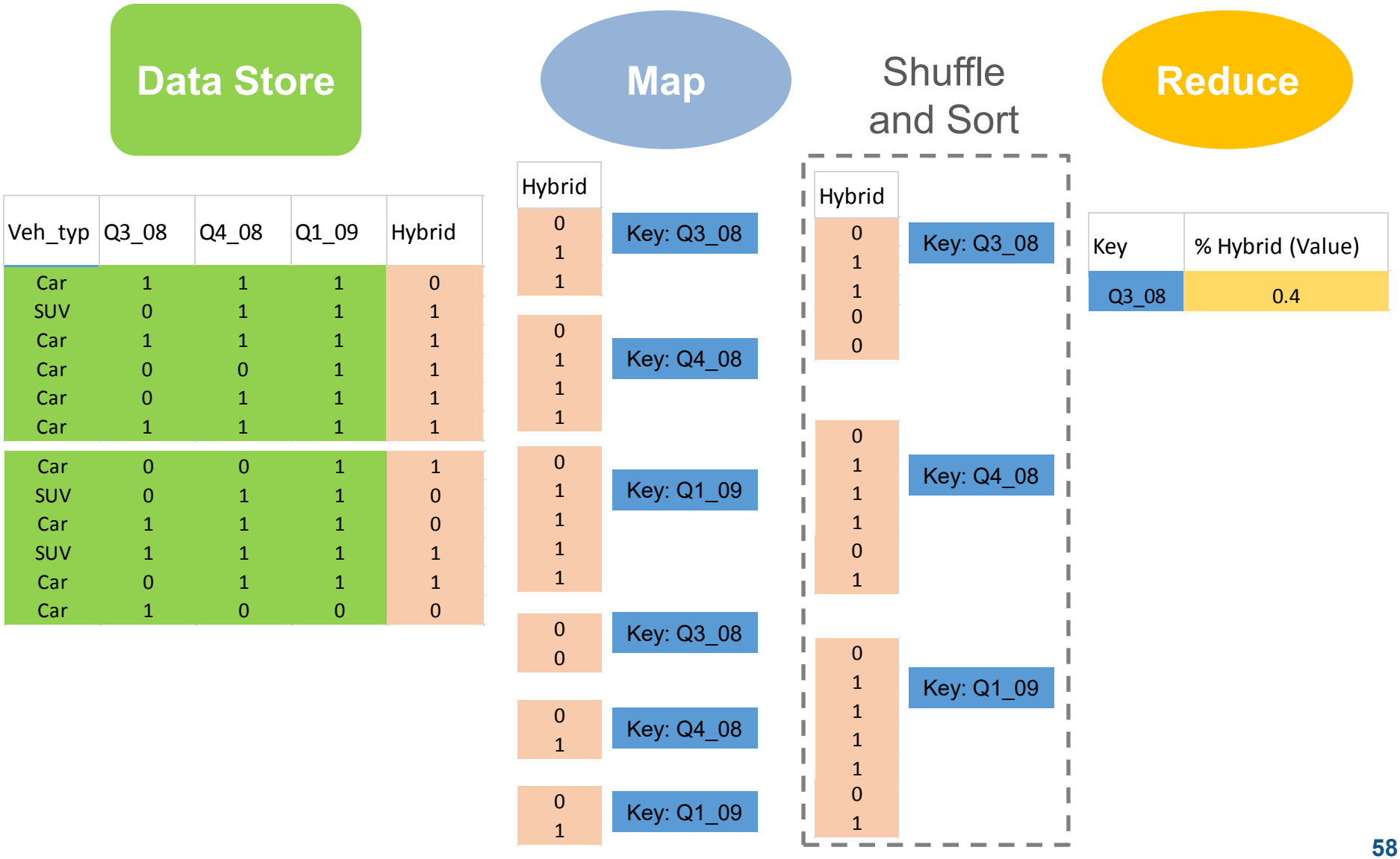
mapreduce


Data Store
Map
Shuffle and Sort
Reduce
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
Hybrid
0 1 1
Key: Q3_08
0 1 1 1
Key: Q4_08
0 1 1 1 1
Key: Q1_09
0 0
Key: Q3_08
0 1
Key: Q4_08
0 1
Key: Q1_09
Hybrid
0 1 1 0 0
Key: Q3_08
0 1 1 1 0 1
Key: Q4_08
0 1 1 1 1 0 1
Key: Q1_09

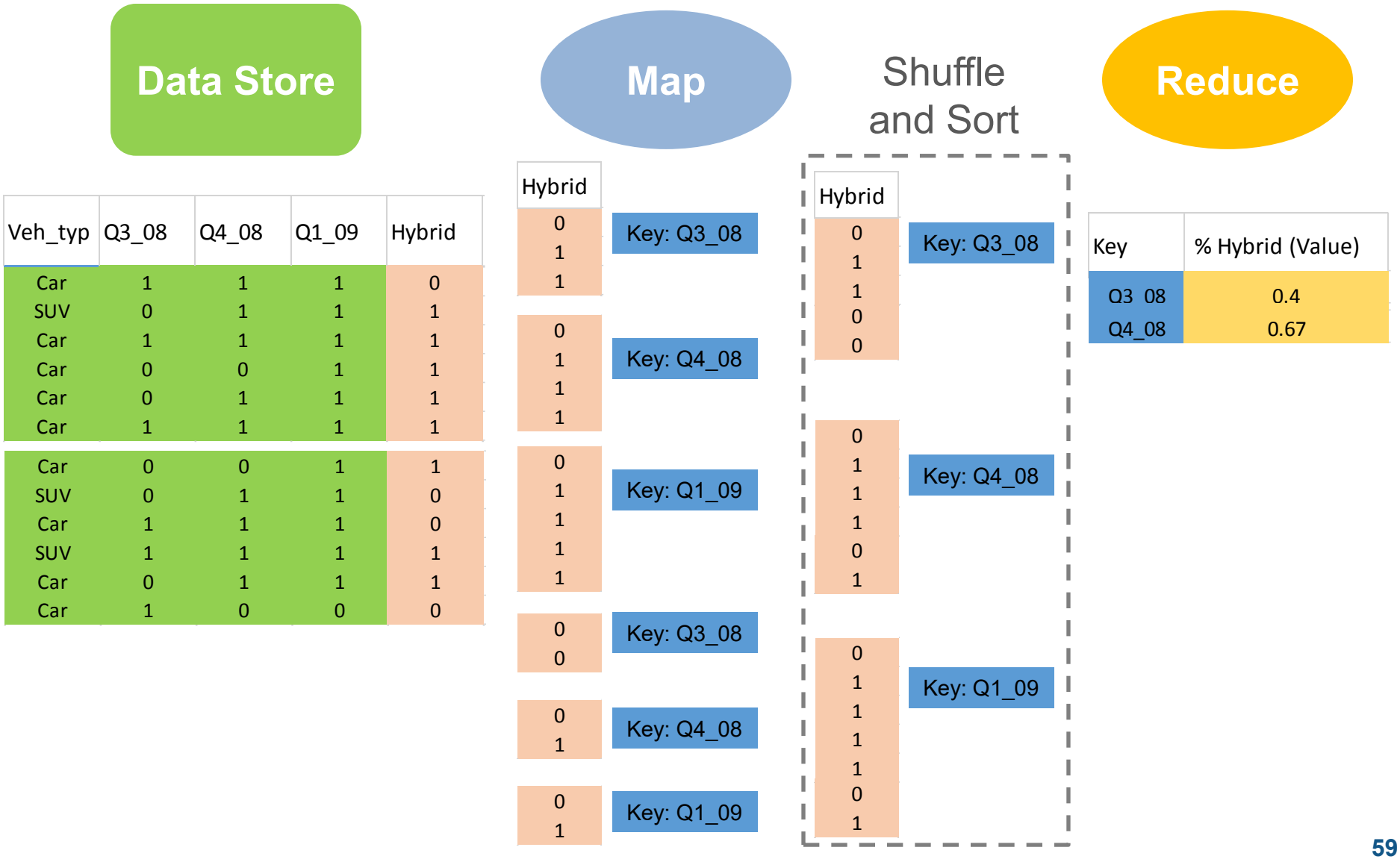
mapreduce


Data Store
Map
Shuffle and Sort
Reduce
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
Hybrid
0 1 1
Key: Q3_08
0 1 1 1
Key: Q4_08
0 1 1 1 1
Key: Q1_09
0 0
Key: Q3_08
0 1
Key: Q4_08
0 1
Key: Q1_09
Hybrid
0 1 1 0 0
Key: Q3_08
0 1 1 1 0 1
Key: Q4_08
0 1 1 1 1 0 1
Key: Q1_09
Key
% Hybrid (Value)

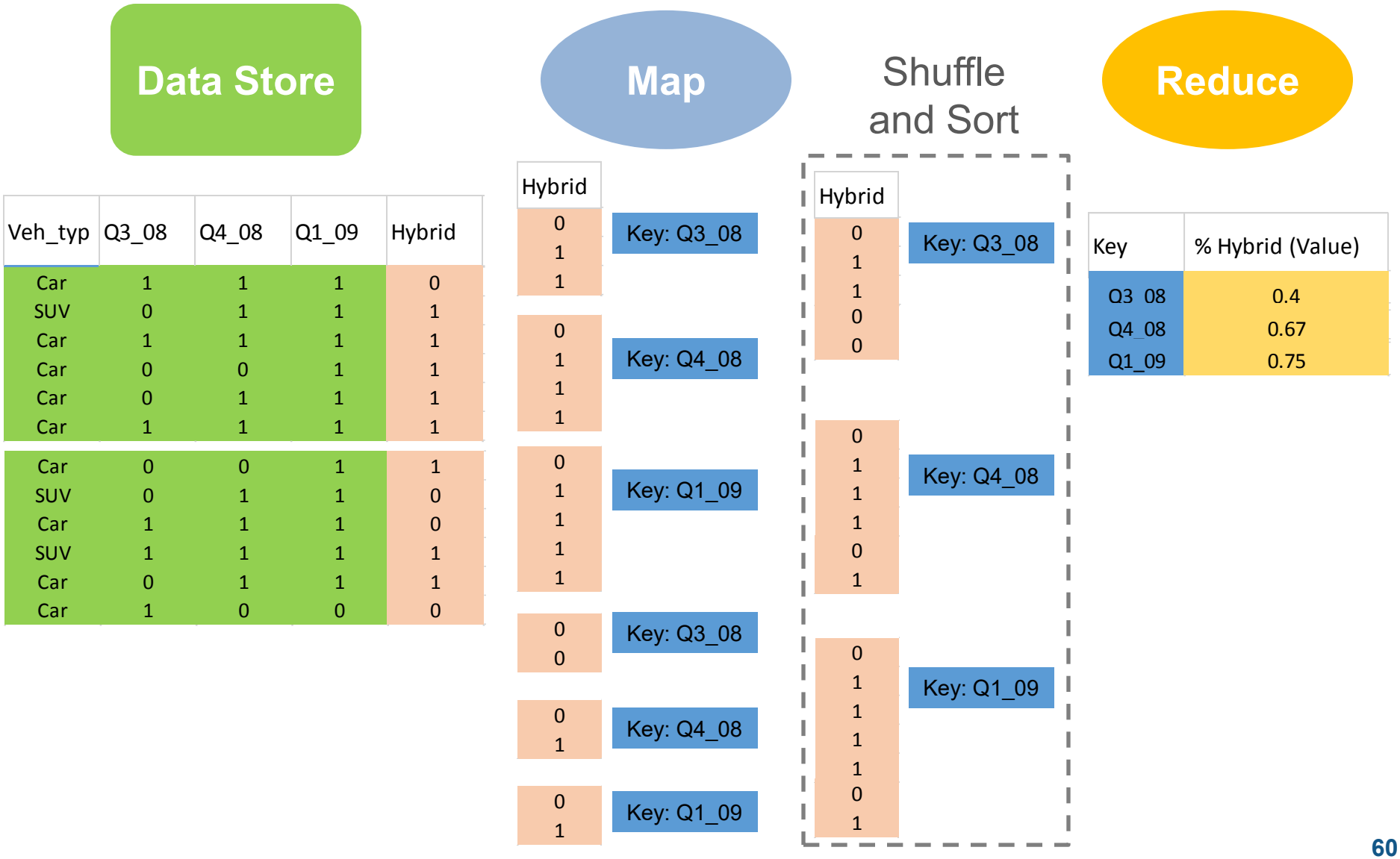
mapreduce


Data Store
Map
Shuffle and Sort
Reduce
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
Hybrid
0 1 1
Key: Q3_08
0 1 1 1
Key: Q4_08
0 1 1 1 1
Key: Q1_09
0 0
Key: Q3_08
0 1
Key: Q4_08
0 1
Key: Q1_09
Hybrid
0 1 1 0 0
Key: Q3_08
0 1 1 1 0 1
Key: Q4_08
0 1 1 1 1 0 1
Key: Q1_09
Key % Hybrid (Value)
Q3_08 0.4

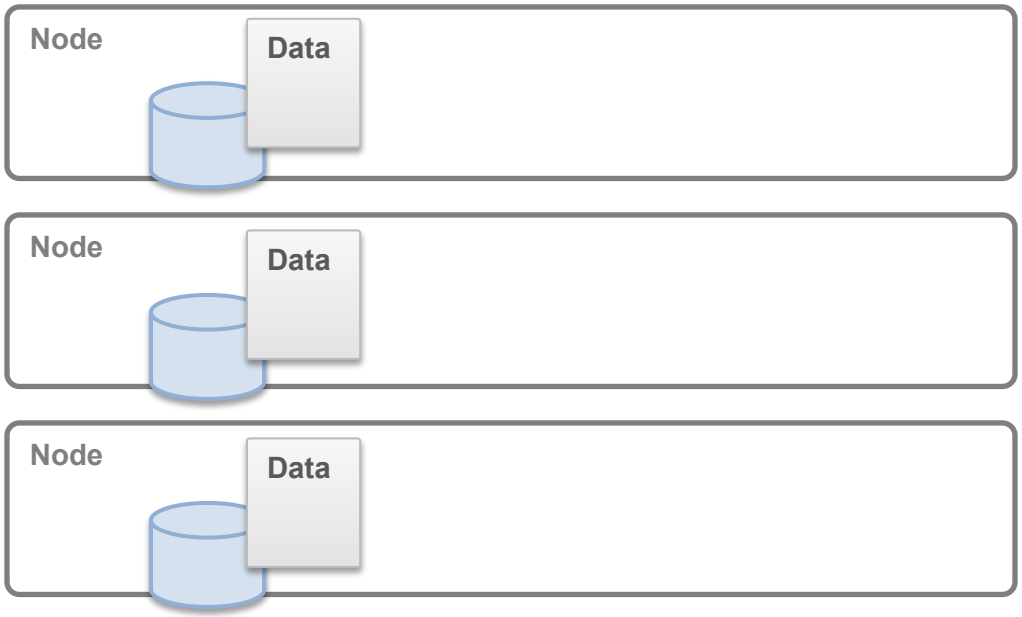
mapreduce


Data Store
Map
Shuffle and Sort
Reduce
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
Hybrid
0 1 1
Key: Q3_08
0 1 1 1
Key: Q4_08
0 1 1 1 1
Key: Q1_09
0 0
Key: Q3_08
0 1
Key: Q4_08
0 1
Key: Q1_09
Hybrid
0 1 1 0 0
Key: Q3_08
0 1 1 1 0 1
Key: Q4_08
0 1 1 1 1 0 1
Key: Q1_09
Key % Hybrid (Value)
Q3_08 0.4
Q4_08 0.67

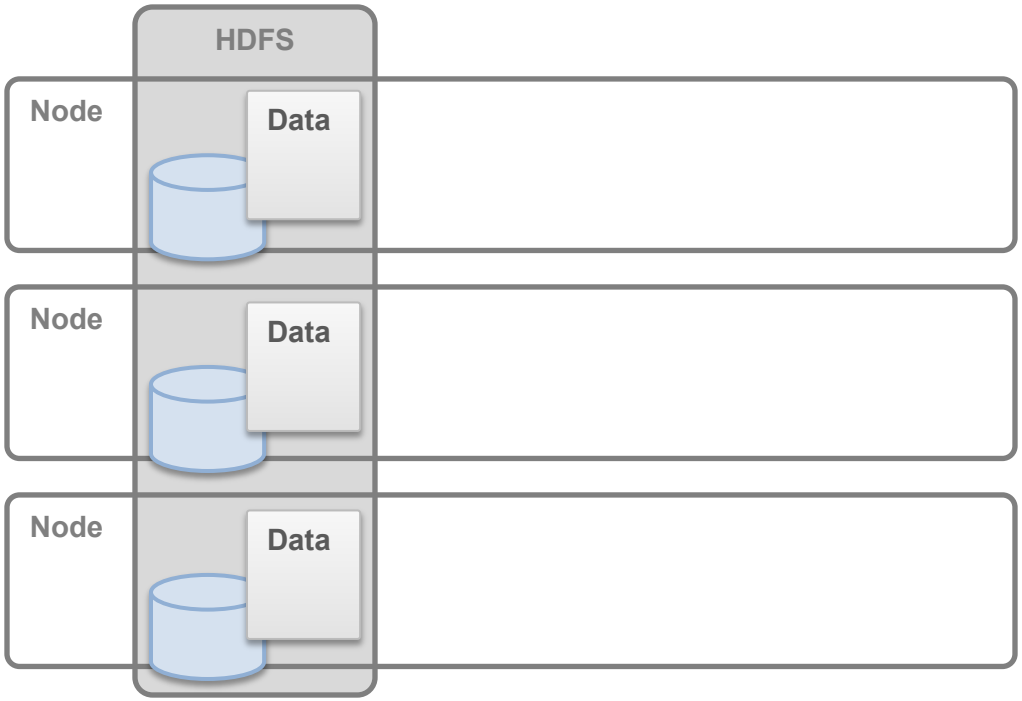
mapreduce


Data Store
Map
Shuffle and Sort
Reduce
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
Hybrid
0 1 1
Key: Q3_08
0 1 1 1
Key: Q4_08
0 1 1 1 1
Key: Q1_09
0 0
Key: Q3_08
0 1
Key: Q4_08
0 1
Key: Q1_09
Hybrid
0 1 1 0 0
Key: Q3_08
0 1 1 1 0 1
Key: Q4_08
0 1 1 1 1 0 1
Key: Q1_09
Key % Hybrid (Value)
Q3_08 0.4
Q4_08 0.67
Q1_09 0.75

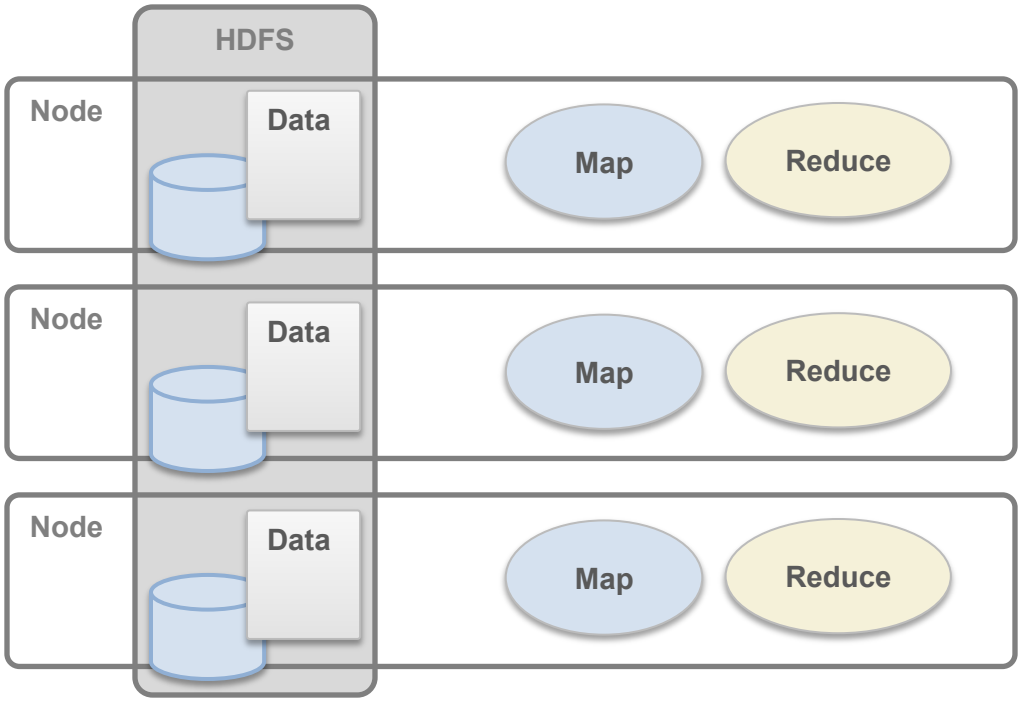
The Big Data Platform



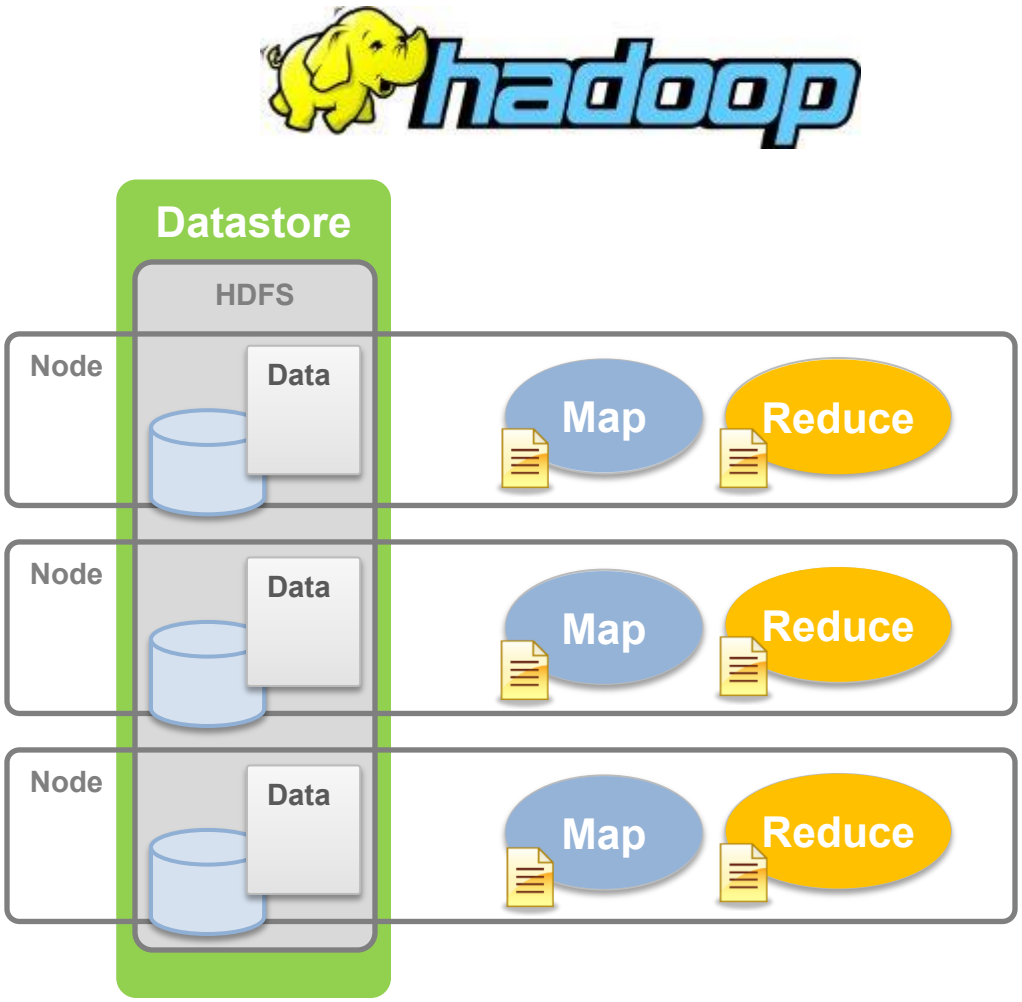
The Big Data Platform



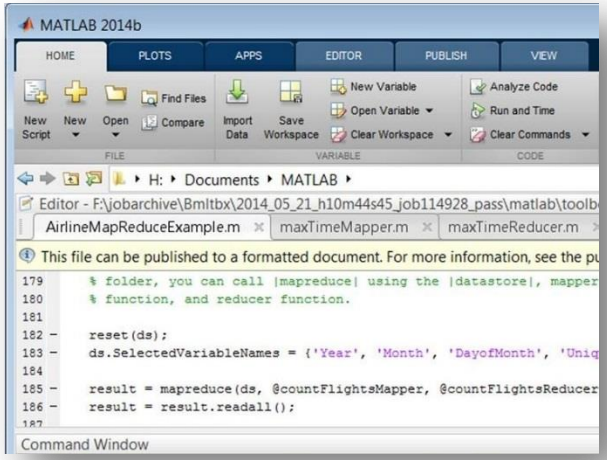
The Big Data Platform

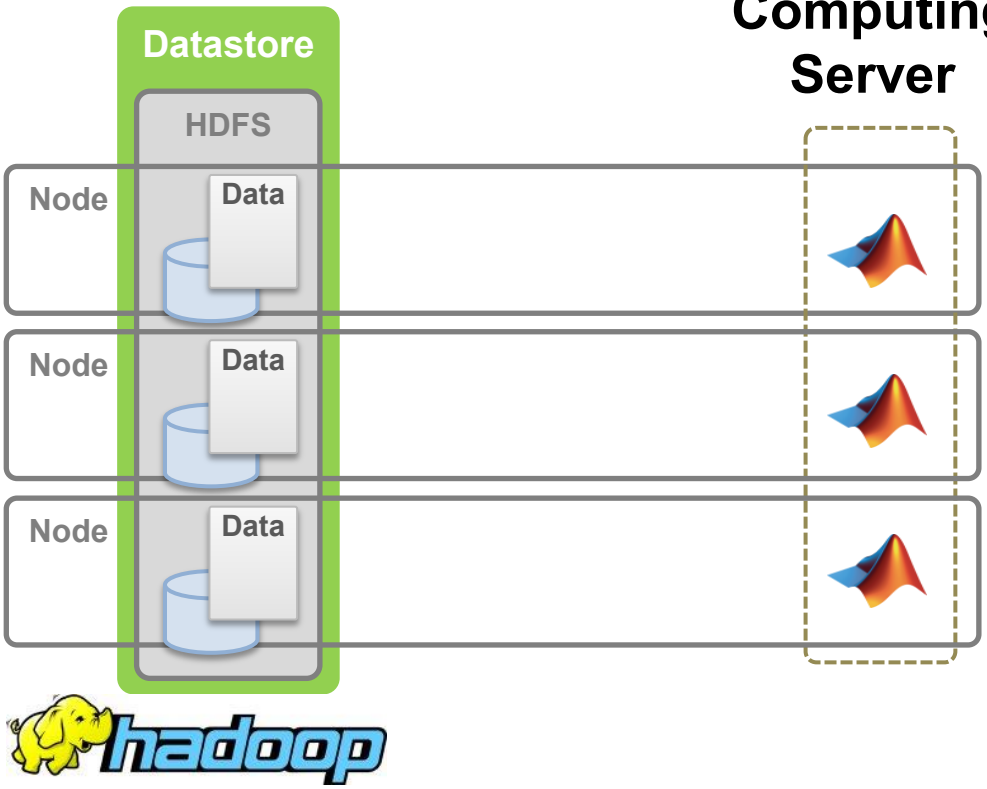


The Big Data Platform

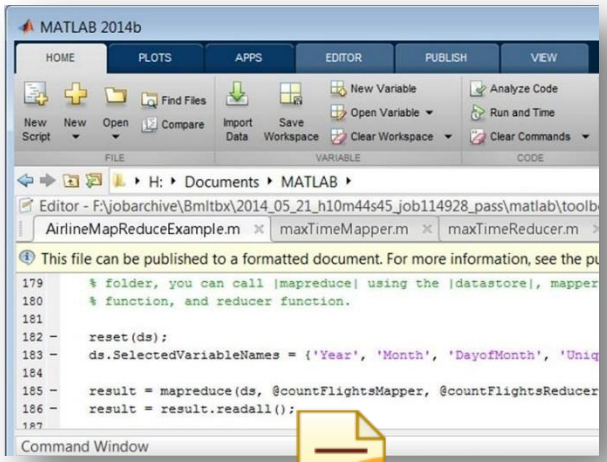


Explore and Analyze Data on Hadoop



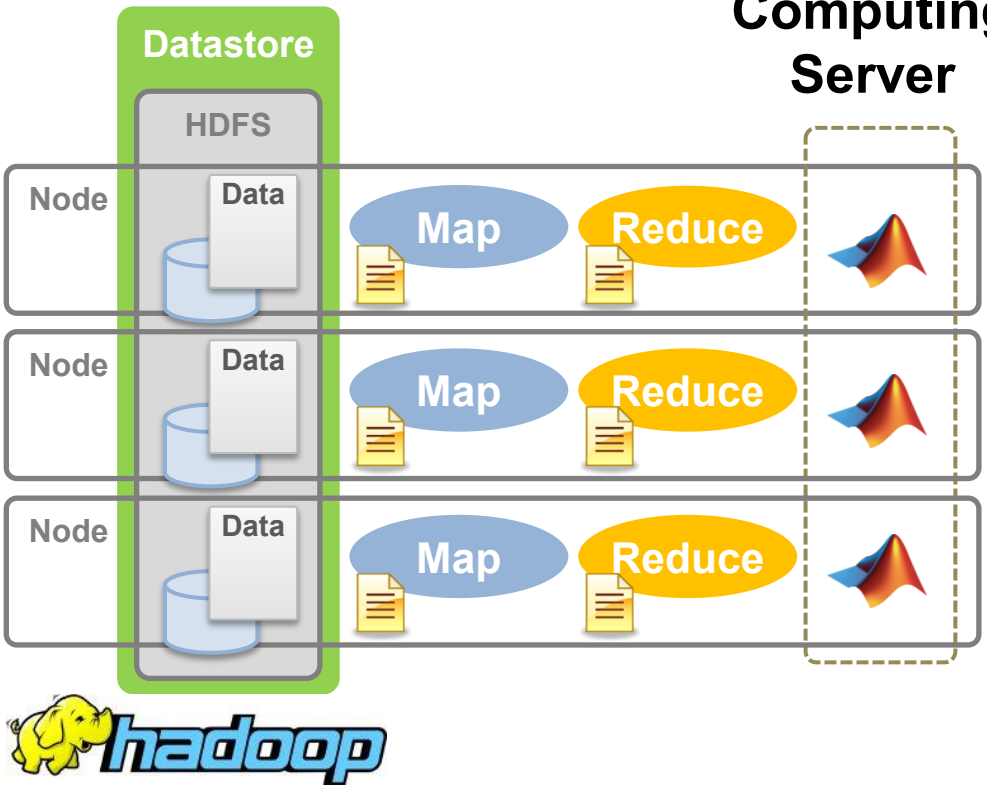


Explore and Analyze Data on Hadoop

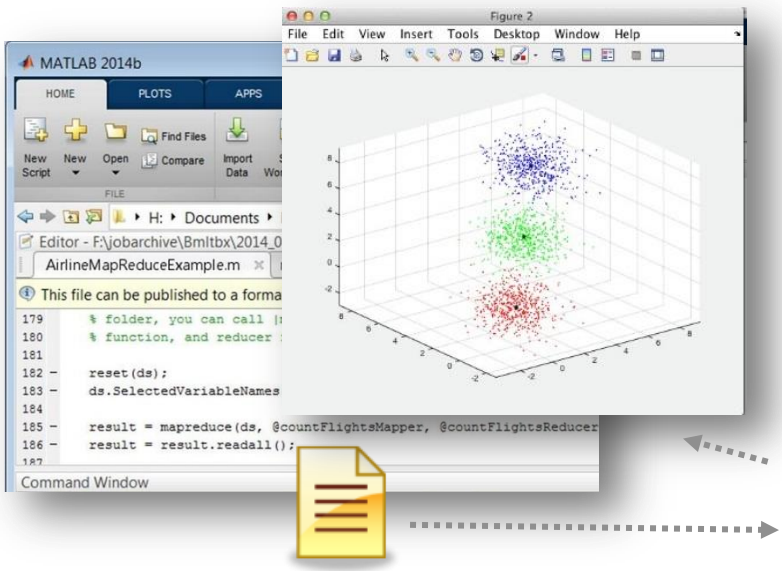


MATLAB MapReduce Code

MATLAB Distributed Computing Server

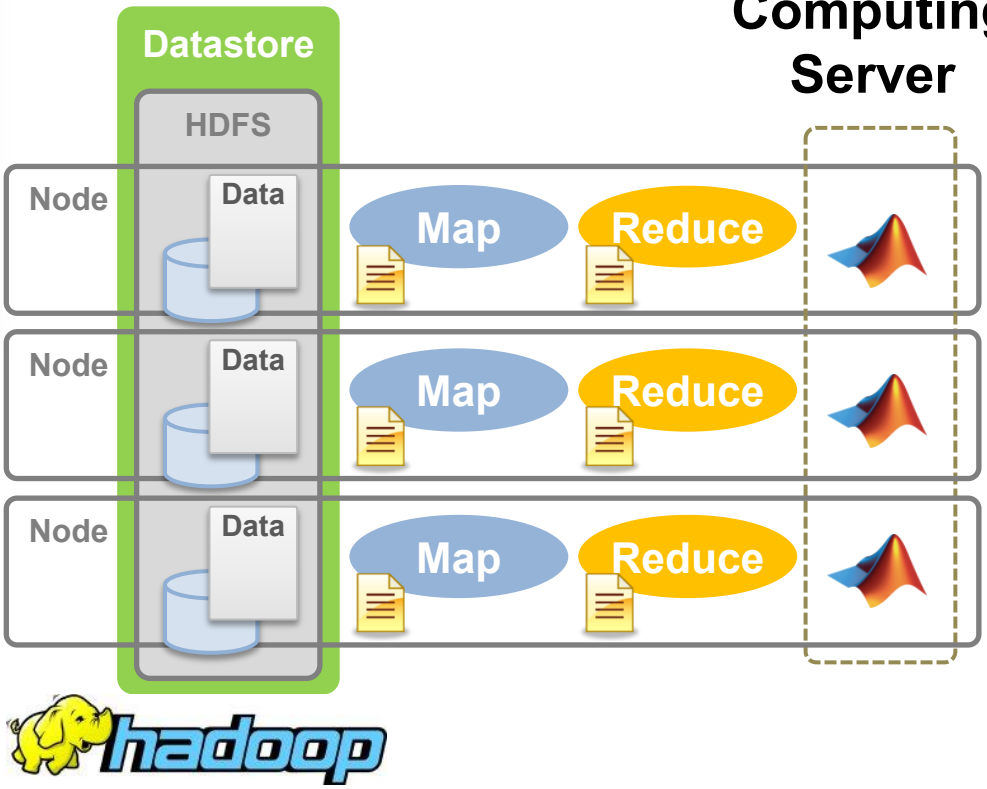


Explore and Analyze Data on Hadoop

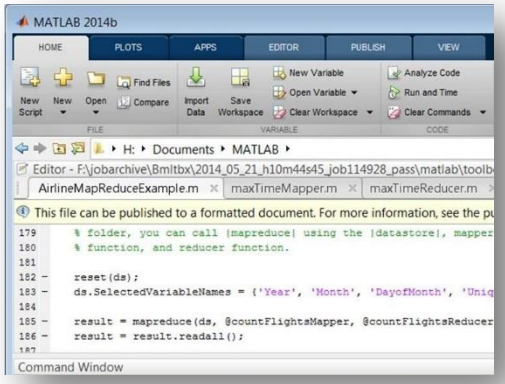


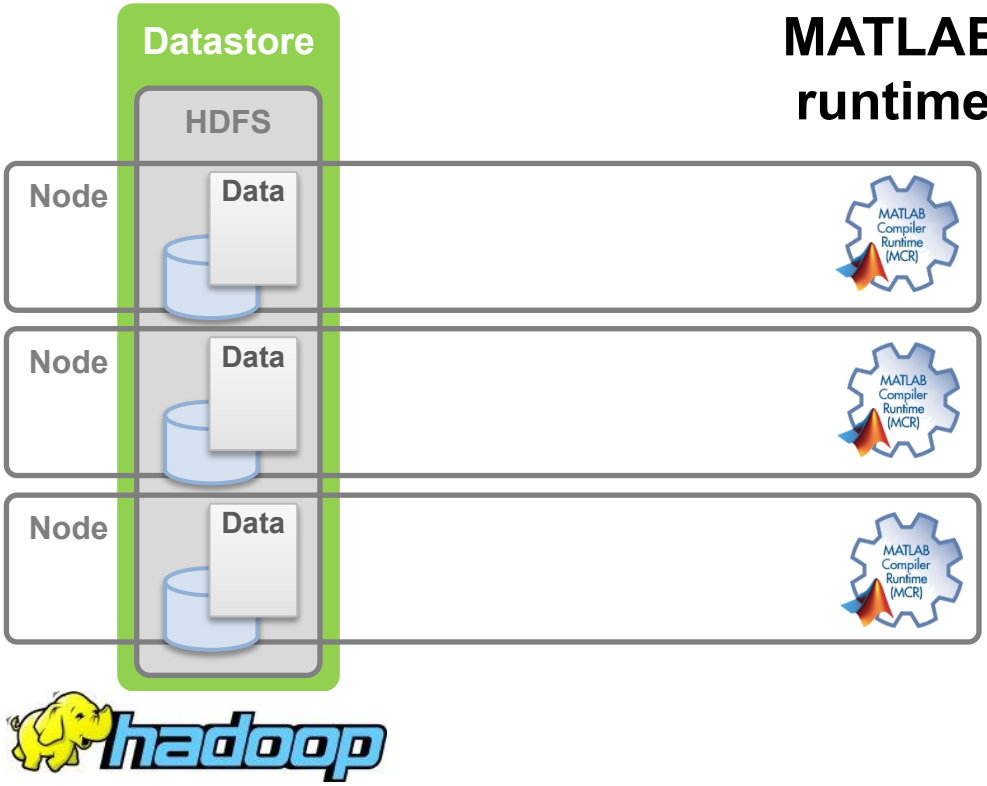
MATLAB
MapReduce
Code

MATLAB Distributed Computing Server

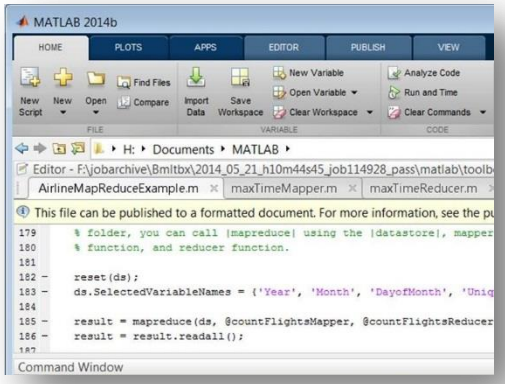


Deployed Applications with Hadoop

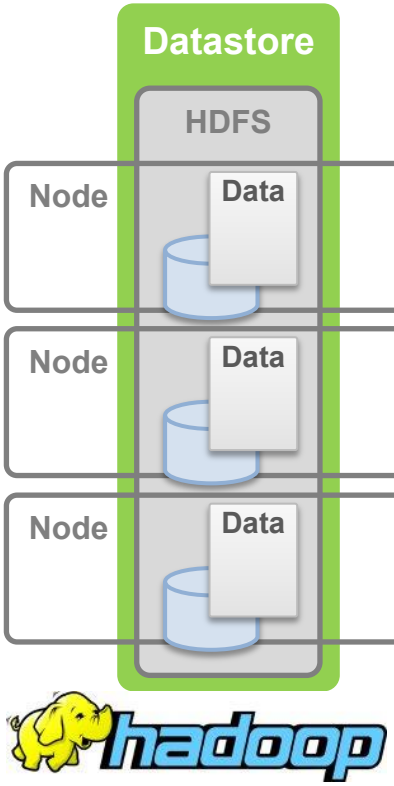




Deployed Applications with Hadoop

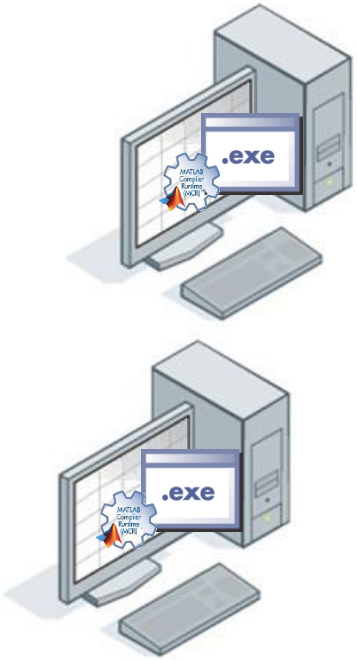


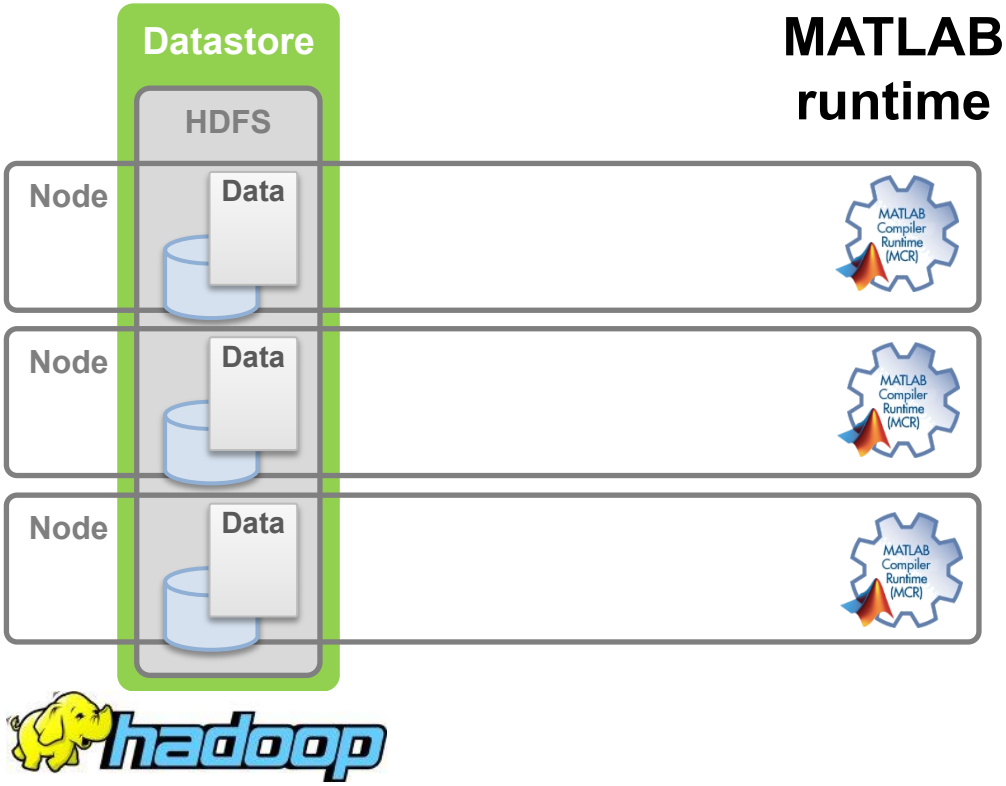
MATLAB MapReduce Code



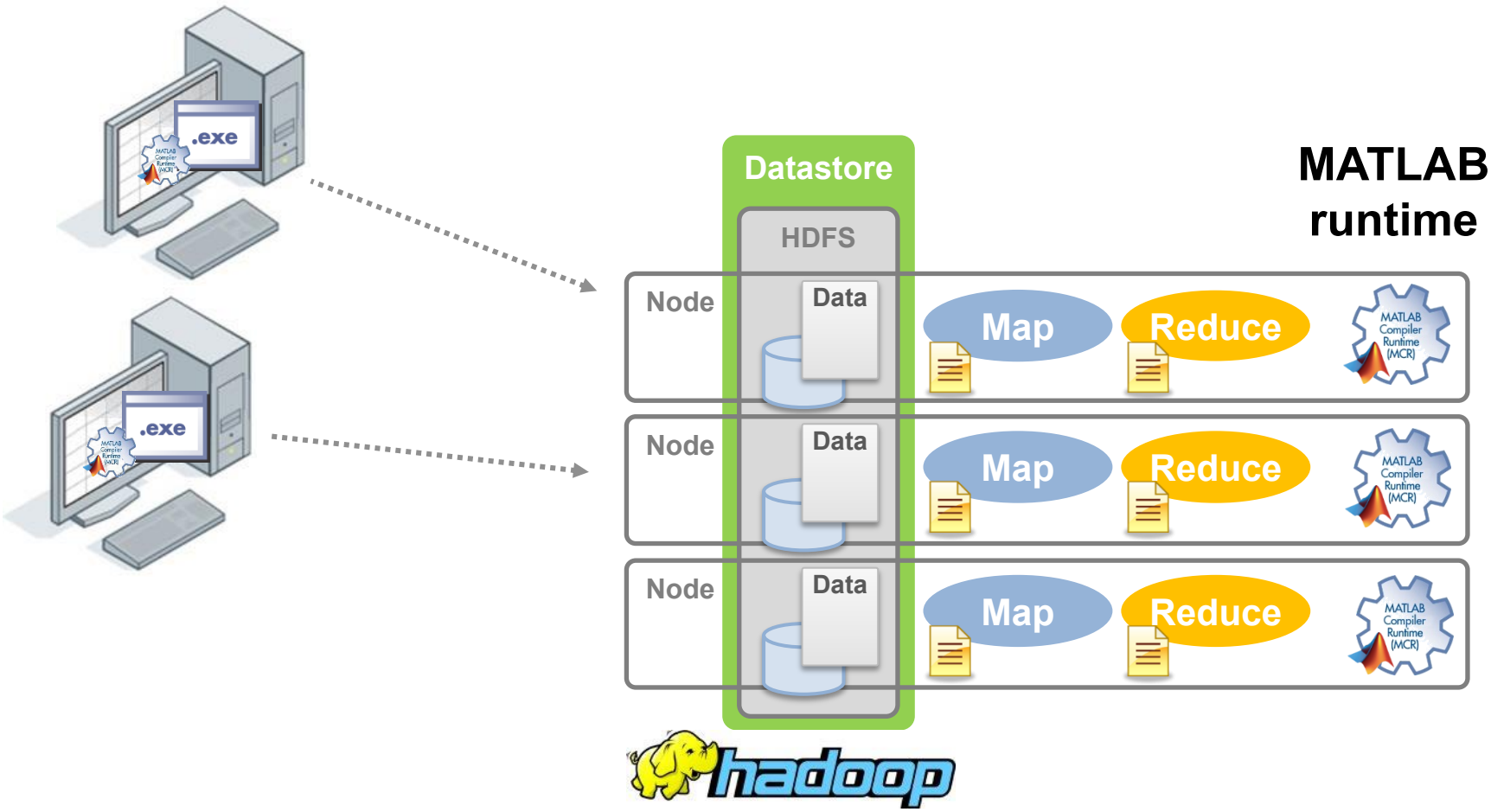
MATLAB runtime

Deployed Applications with Hadoop

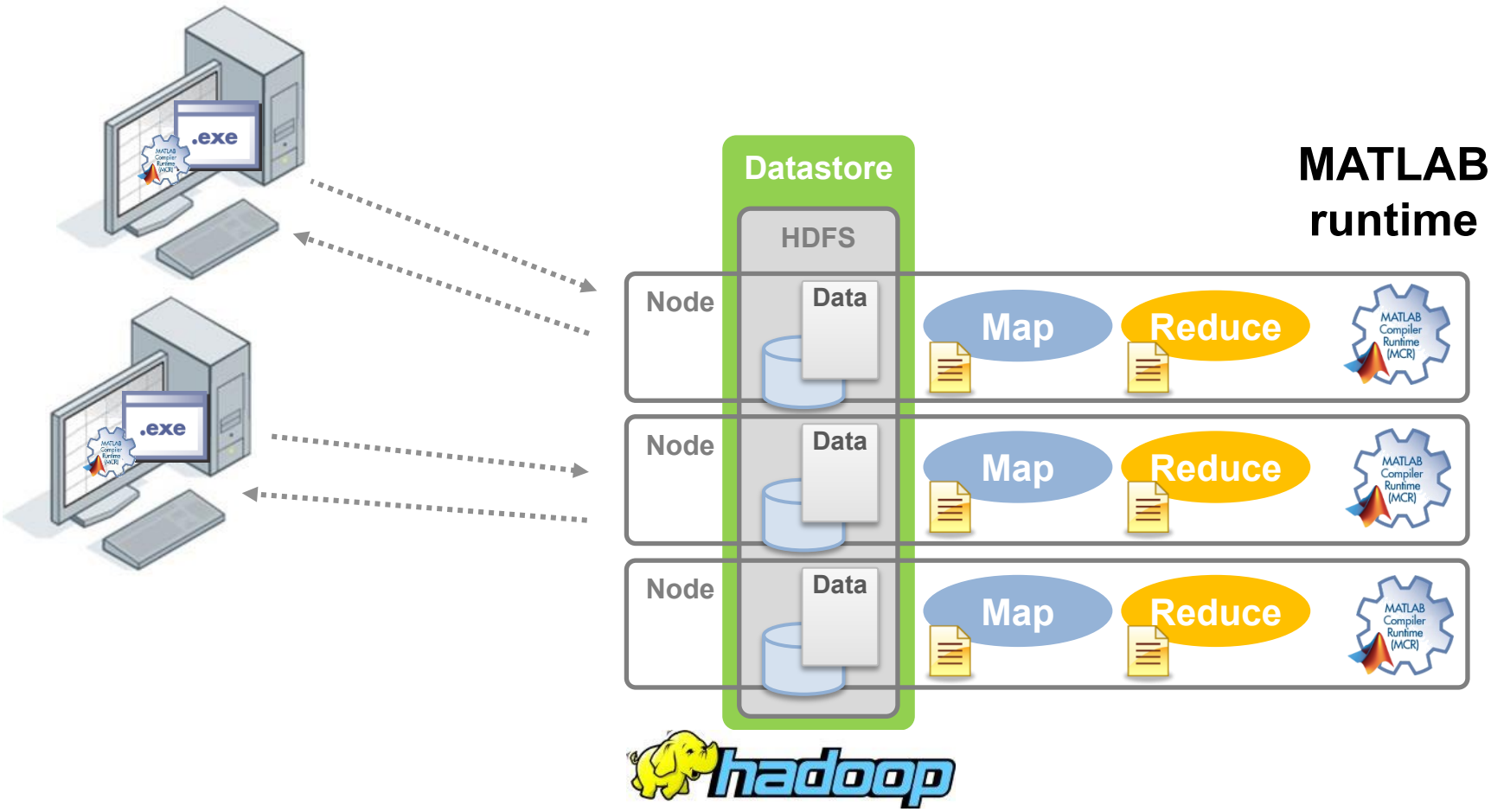




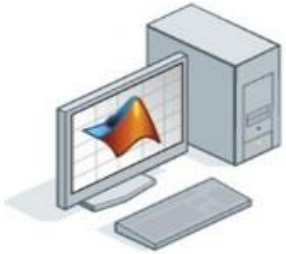
Deployed Applications with Hadoop



Deployed Applications with Hadoop



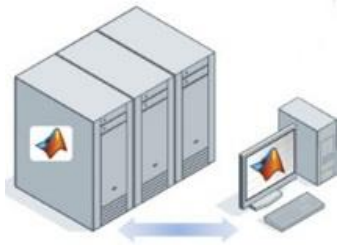
Big Data on the Desktop



- 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

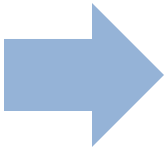
MATLAB supports a number of programming constructs for use with clusters



- 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

R2014b

Use these constructs on the desktop to develop your algorithms

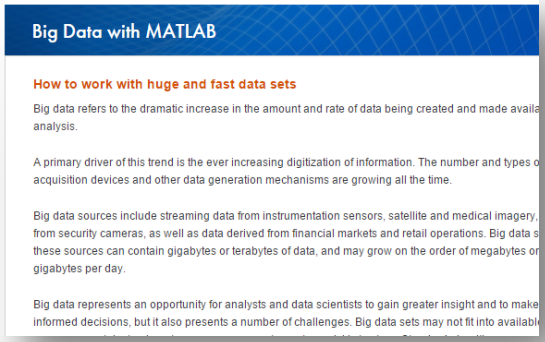


Migrate to a cluster without algorithm changes

Learn More

- MATLAB Documentation
  - Strategies for Efficient Use of Memory
  - Resolving "Out of Memory" Errors

- Big Data with MATLAB
  - www.mathworks.com/discovery/big-data-matlab.html



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