MathWorks

MATLAB EXPO 2018

MATLAB与Spark/Hadoop相集成：实现大数据的处理和价值挖

马文辉

内容

- **大数据及其带来的挑战**
- **MATLAB大数据处理**
  - tall数组
  - 并行与分布式计算
- **MATLAB与Spark/Hadoop集成**
  - MATLAB访问HDFS(Hadoop分布式文件系统）
  - 在Spark/Hadoop集群上运行MATLAB代码
- **应用演示 – 汽车传感器数据分析**

大数据概述

大数据的”4V”特征：

- Volumes - 数据规模，数据规模巨大

  互联网、社交网络的普及，全社会的数字化转型，数据规模向PB级发展

- Variety - 数据种类 ，数据种类繁多

  结构化数据，半结构化数据，非结构化数据

- Value - 数据价值，数据价值密度低

  价值密度的高低与数据总量的大小成反比

- Velocity - 数据处理速度，数据处理速度需要快速

  数据处理速度是决定大数据应用的关键

大数据带来的挑战

- 传统的工具和方法不能有效工作
  - 访问和处理数据变得困难；
  - 需要学习使用新的工具和新的编程方式；
  - 不得不重写算法以应对数据规模的增大；
- 现有处理或计算方法下的结果质量受到影响
  - 被迫只能处理一部分数据（数据子集）；
  - 采用新的工具或重写算法会对现有生产力产生影响；
- 数据处理与分析所需时间增长
  - 数据规模增大、数据复杂度增加，增加处理难度和所需时间；

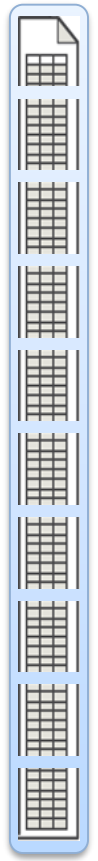
MATLAB的大数据处理

- 内存与数据访问
  - 64-bit processors
  - Memory Mapped Variables
  - Disk Variables
  - Databases
  - Datastore R2014b
  - ImageDatastore R2016a
- 编程
  - Streaming
  - Block Processing
  - Parallel-for loops
  - GPU Arrays
  - SPMD and Distributed Arrays
  - MapReduce R2014b
  - MapReduce (MDCS/PCT) R2014b
  - MATLAB API for Spark API R2016b
  - Tall Arrays R2016b
- 计算
  - Desktop (Multicore, GPU)
  - Clusters
  - Cloud Computing (MDCS on EC2)
  - Hadoop R2014b
  - Spark R2016b

tall arrays R2016b

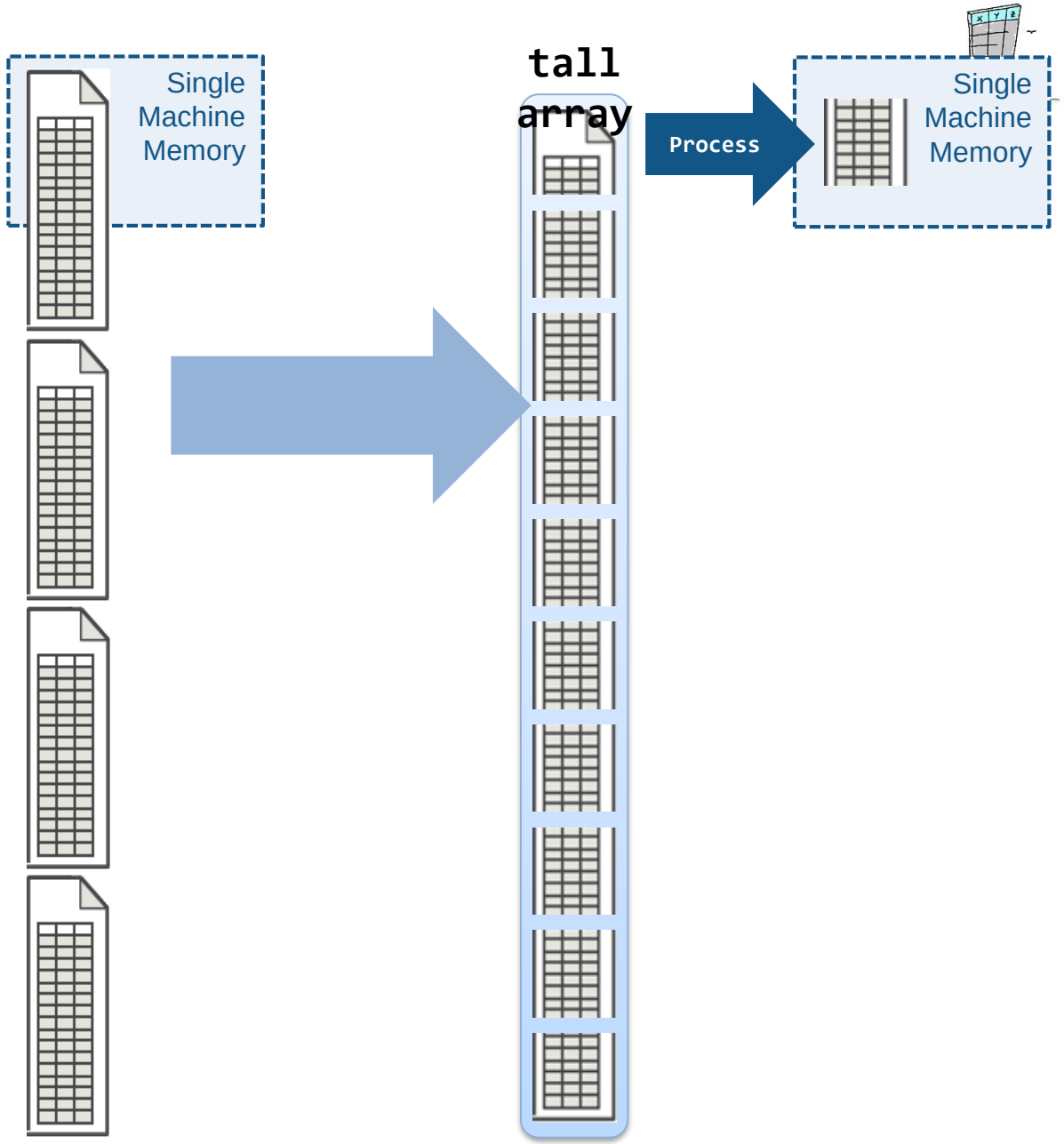
- tall array
  - 一种新的数据类型，专门用于处理大数据.
  - 用于处理数据规模超过单个机器或群集的内存承载能力的数据集合

- 使用方式等同于MATLAB 数组(array)
  - 支持数据类型包括数值型、字符串、时间类型、表等…
  - 支持众多基本的数学函数、统计函数、索引函数等.
  - 支持机器学习算法包括分类、聚类和回归

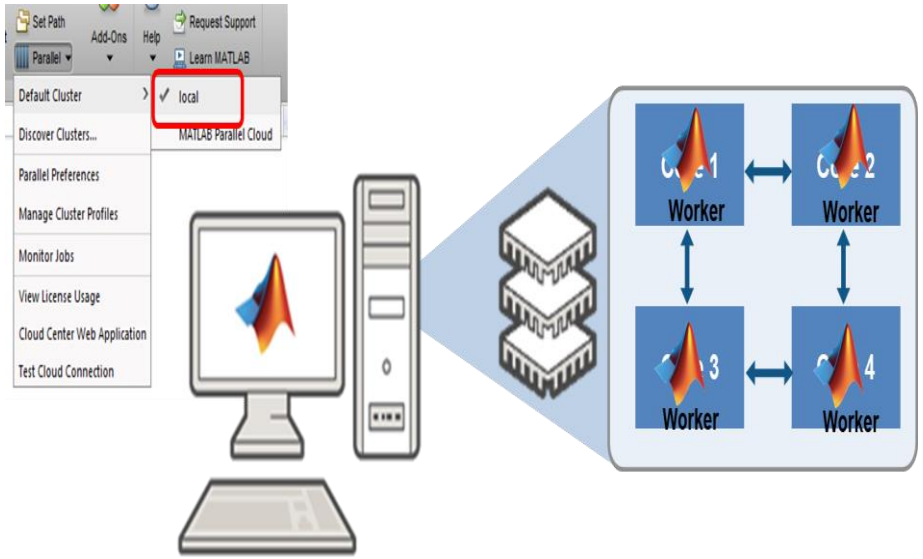


tall arrays R2016b

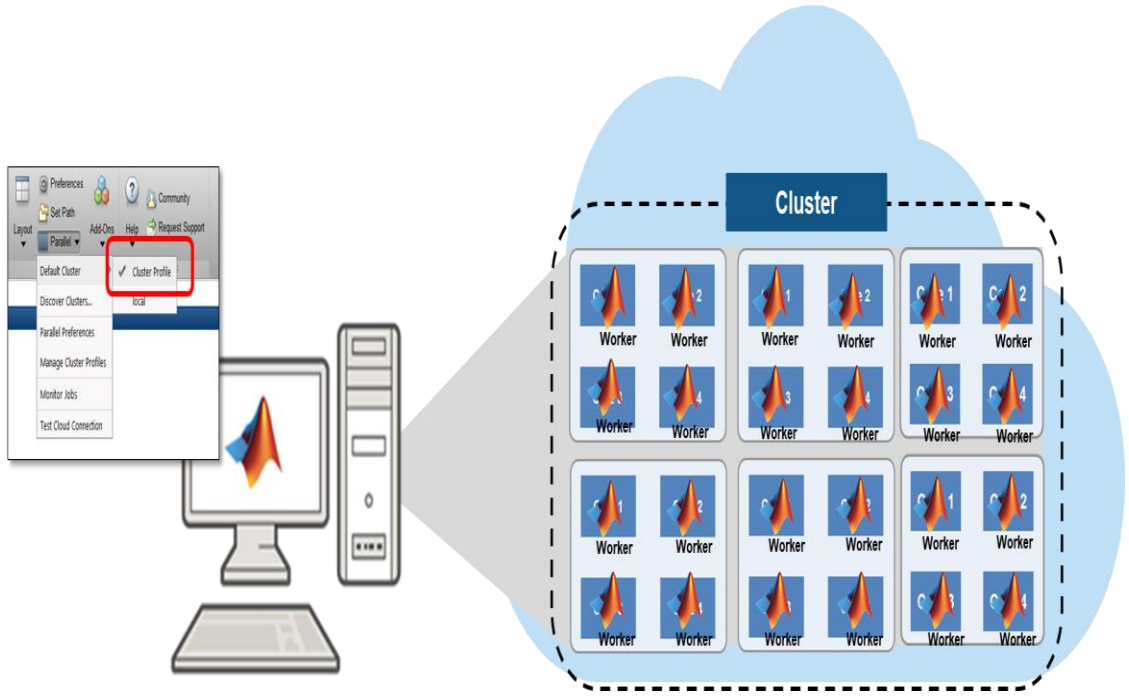
- 自动将数据分解成适合内存的小“块”(chunk)
- 计算过程中，一次处理一个“块”(chunk)的数据
- 对tall数组(tall array)的编程方式与MATLAB标准数组 编程方式一致



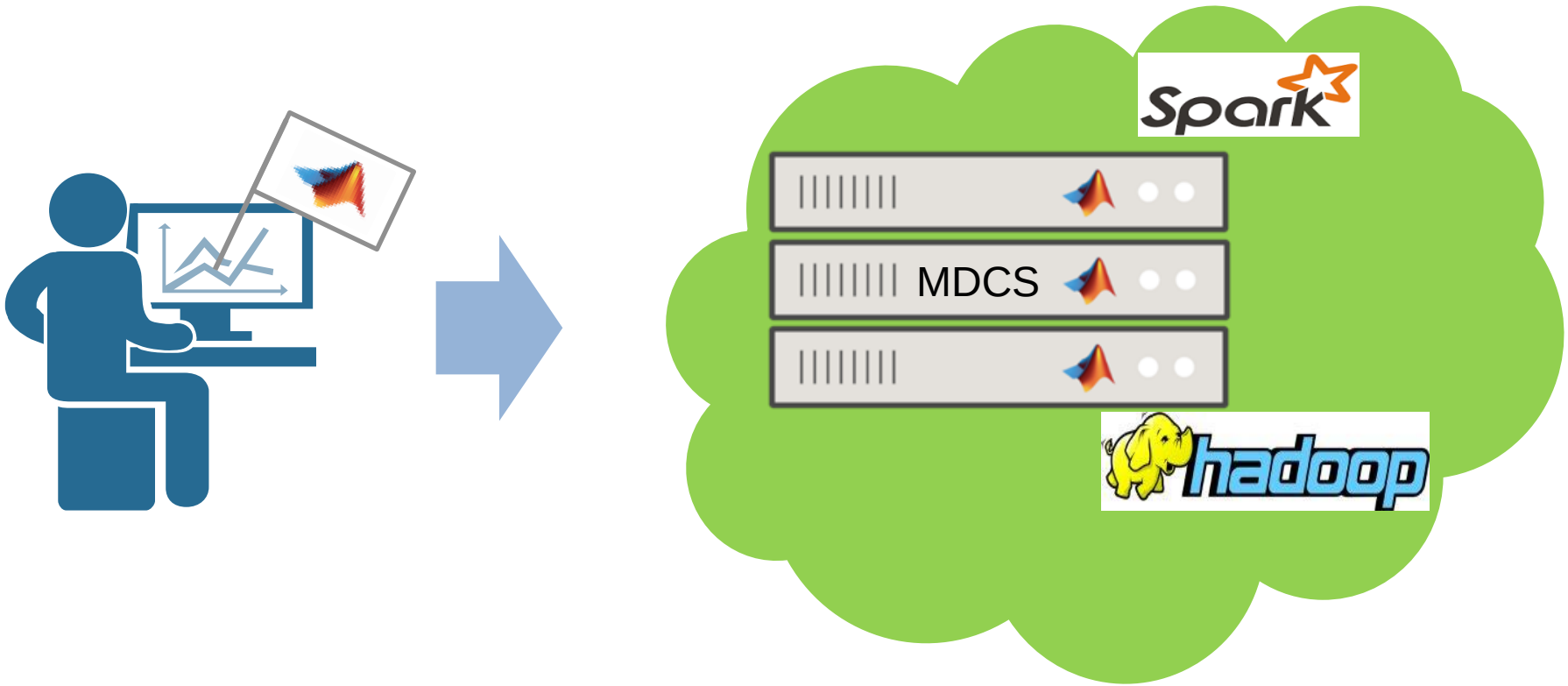
- MATLAB本地多核并行计算计
  (PCT, Parallel Computing Toolbox)



- MATLAB集群之上的分布式计算
  (MDCS, MATLAB Distributed Computing Server)



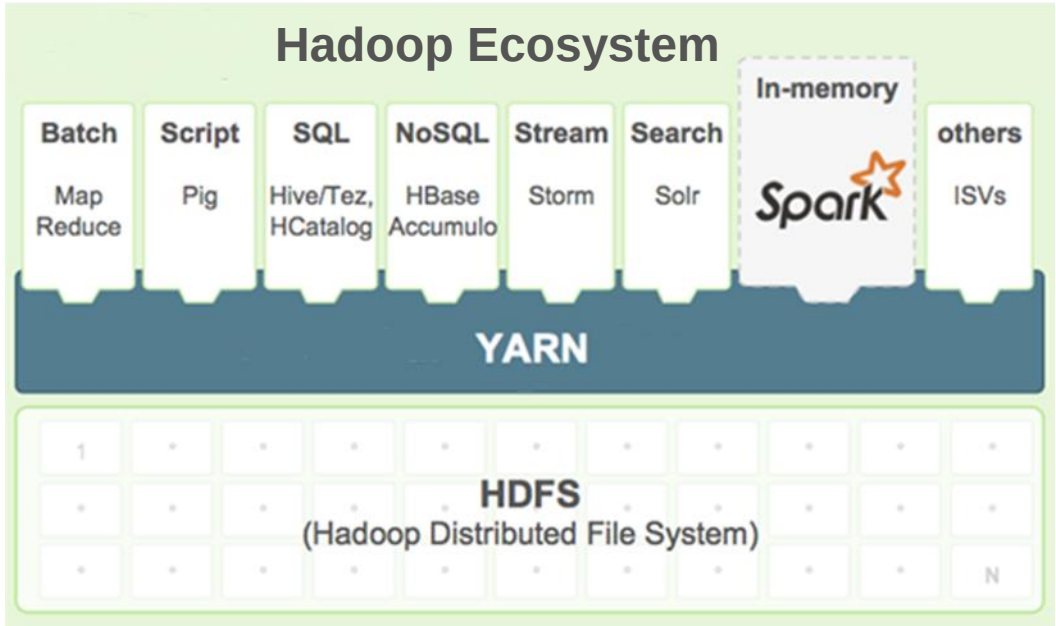
MATLAB与Spark/Hadoop集成



Hadoop

Hadoop是跨计算机集群的分布式大数据处理平台，由两部分组成：

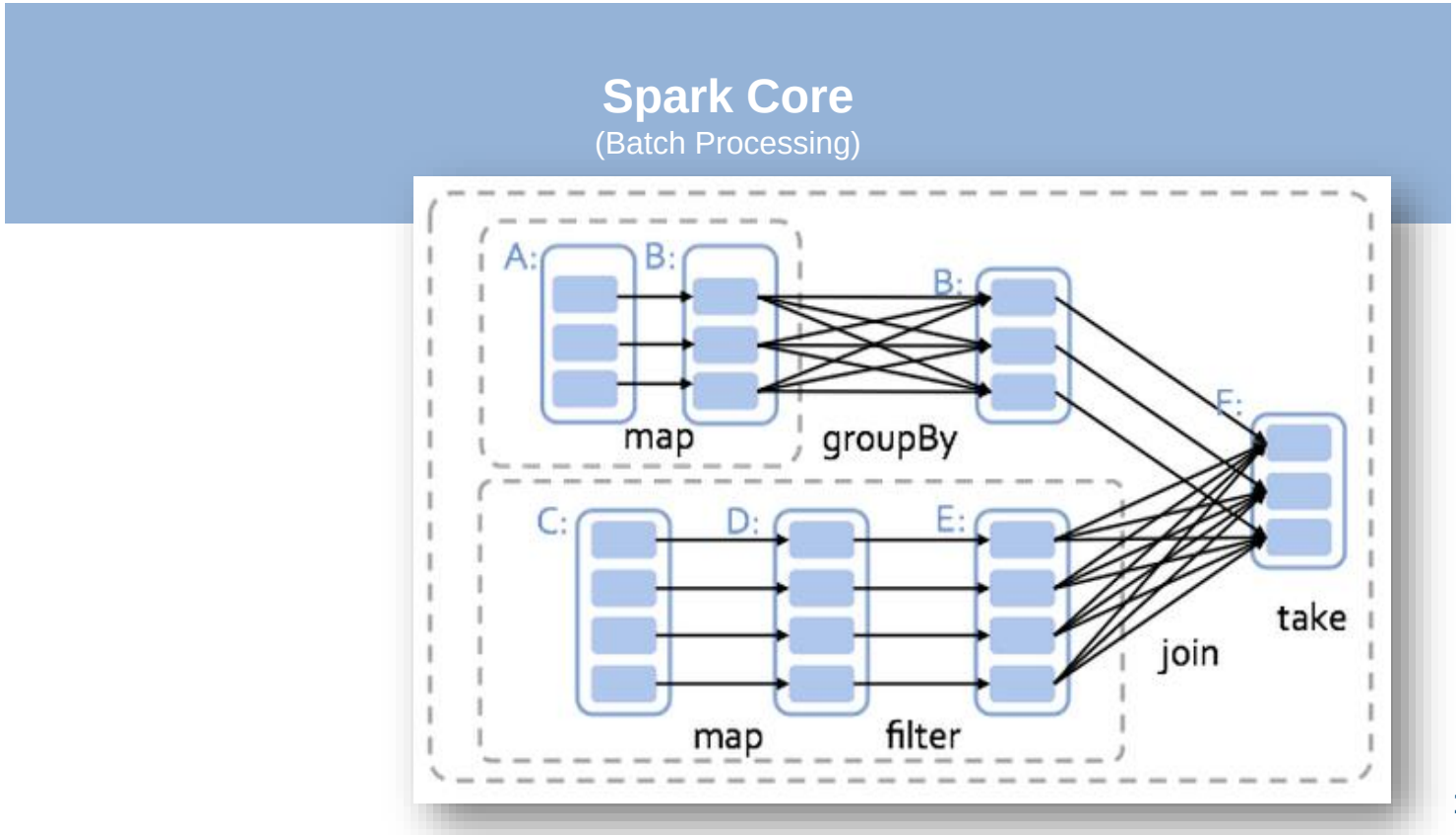
- YARN (Yet Another Resource Negotiator) – 资源调度模型，实现数据跨节点的最小移动
- Map/Reduce – 跨节点分布式计算模型
- HDFS (Hadoop Distributed File System) - 跨节点的分布式文件系统



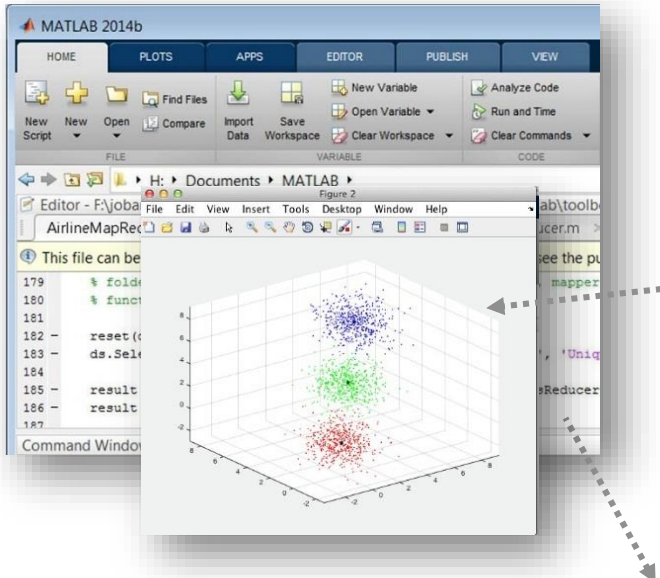
Spark

Spark是一个流行的开源集群计算框架

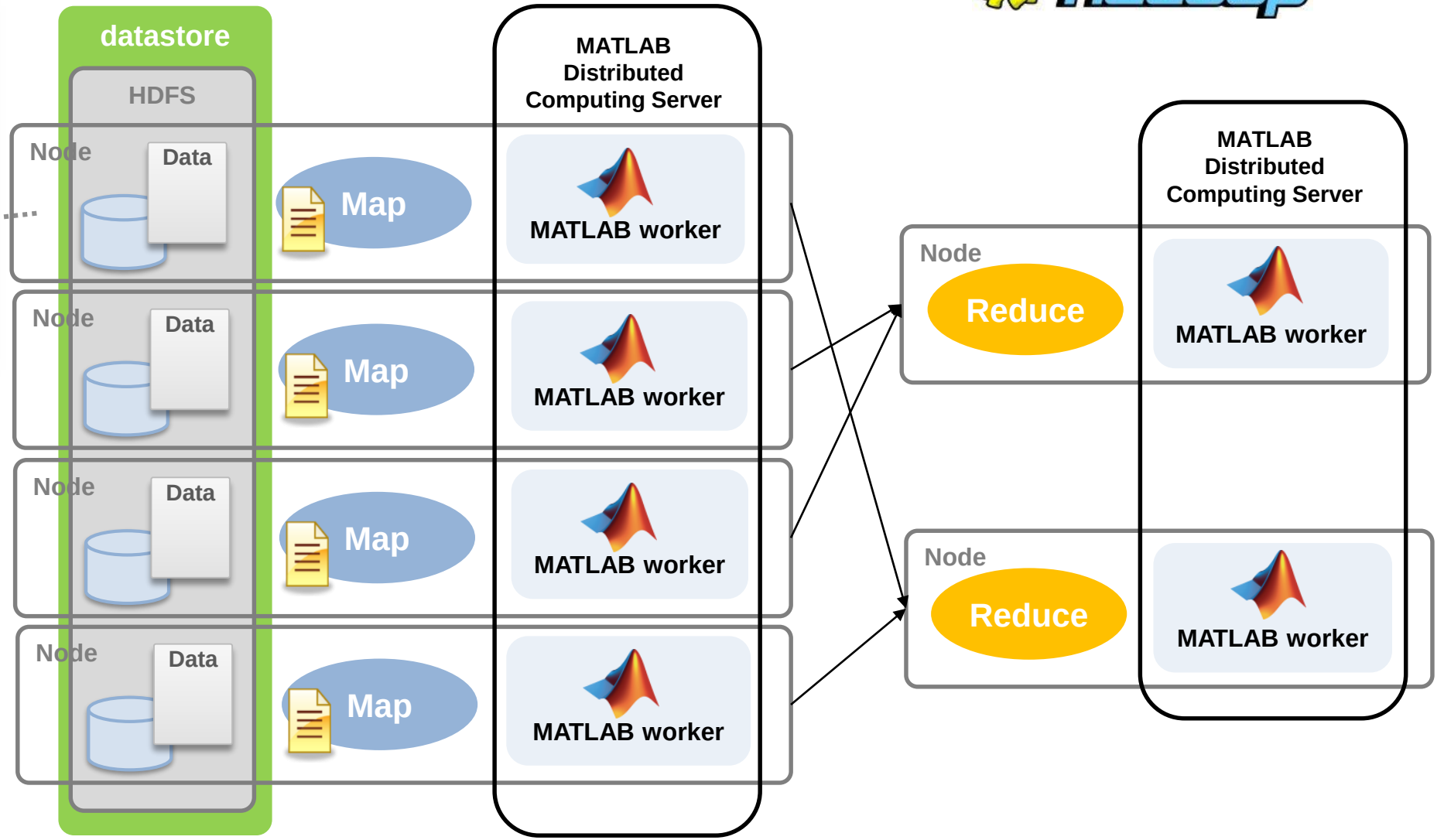
- 并行计算引擎
- 使用广义的计算模型
- 基于内存进行计算（内存计算）



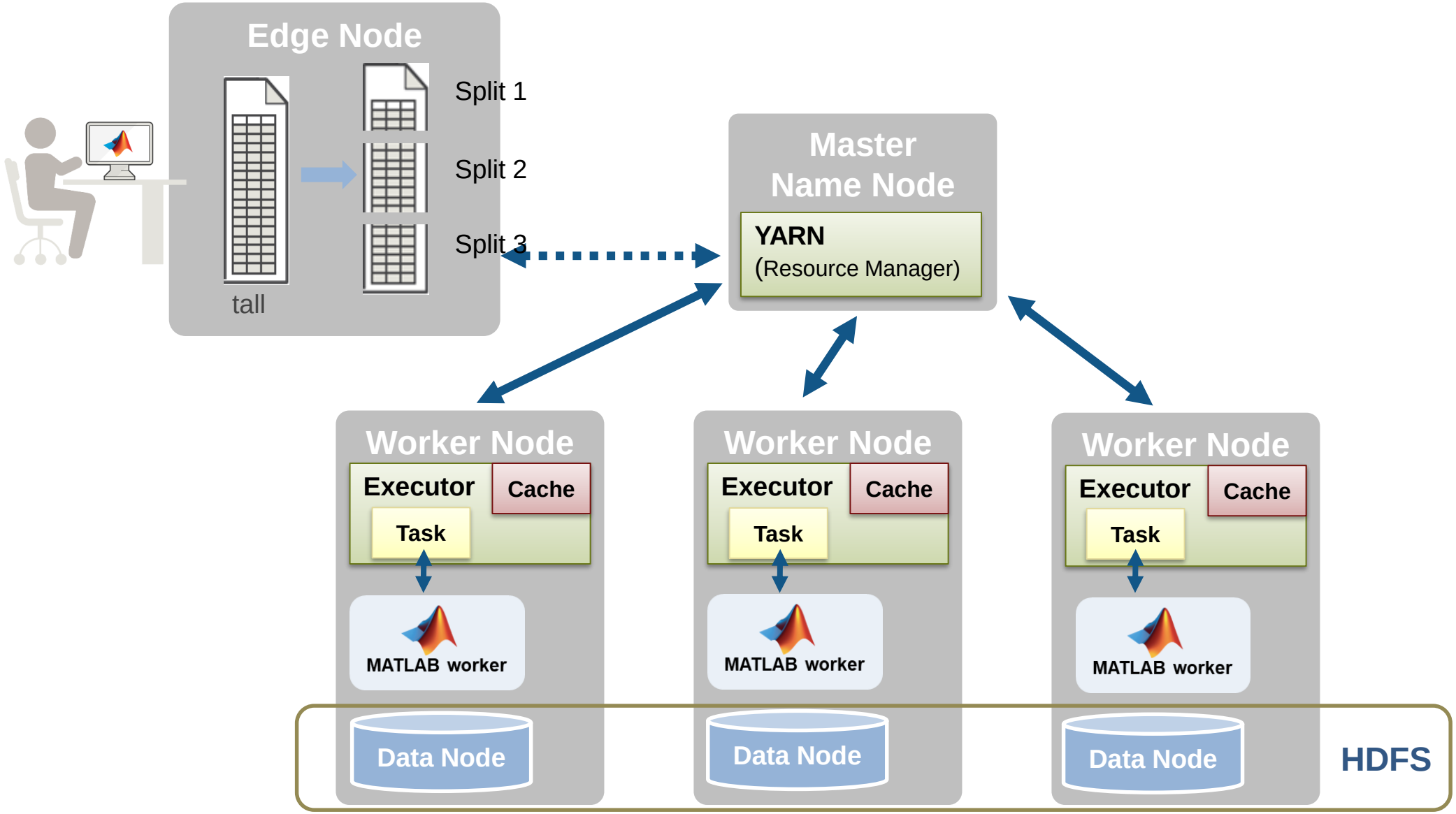
MATLAB与Hadoop



map.m
reduce.m



MATLAB tall 与Spark



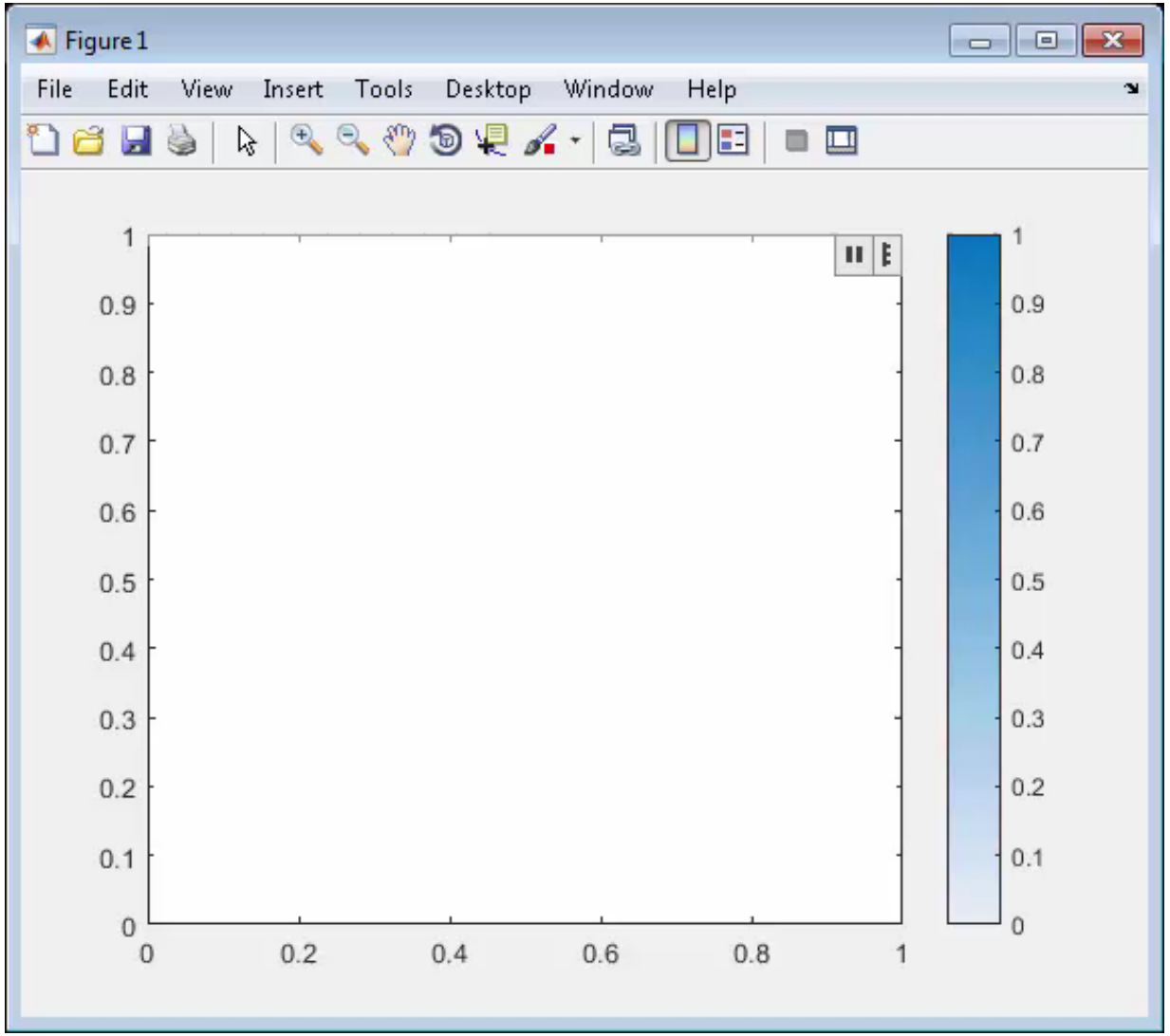
Tall支持的大数据可视化

R2016b

- histogram
- histogram2
- ksdensity

R2017b

- plot
- scatter
- binscatter



tall 支持的大数据机器学习算法

- K-means Clustering (`kmeans`)
- Linear Regression (`fitlm`)
- Logistic & Generalized Linear Regression (`fitglm`)
- Discriminant Analysis Classification (`fitcdiscr`)
- Partition for Cross Validation (`cvpartition`)

R2016b

- Linear Support Vector Machine (SVM) Classification (`fitclinear`)
- Naïve Bayes Classification (`fitcnb`)
- Random Forest Ensemble Classification (`TreeBagger`)
- Lasso Linear Regression (`lasso`)

R2017a

- Linear Support Vector Machine (SVM) Regression (`fitrlinear`)
- Single Classification Decision Tree (`fitctree`)
- Linear Classification with Random Kernel Expansion (`fitckernel`)

R2017b

应用演示 – 汽车传感器数据分析

- 1300 trip log files
- 21 unique vehicles
- Approx 39 unique channels
- Data collected over 1.5 years