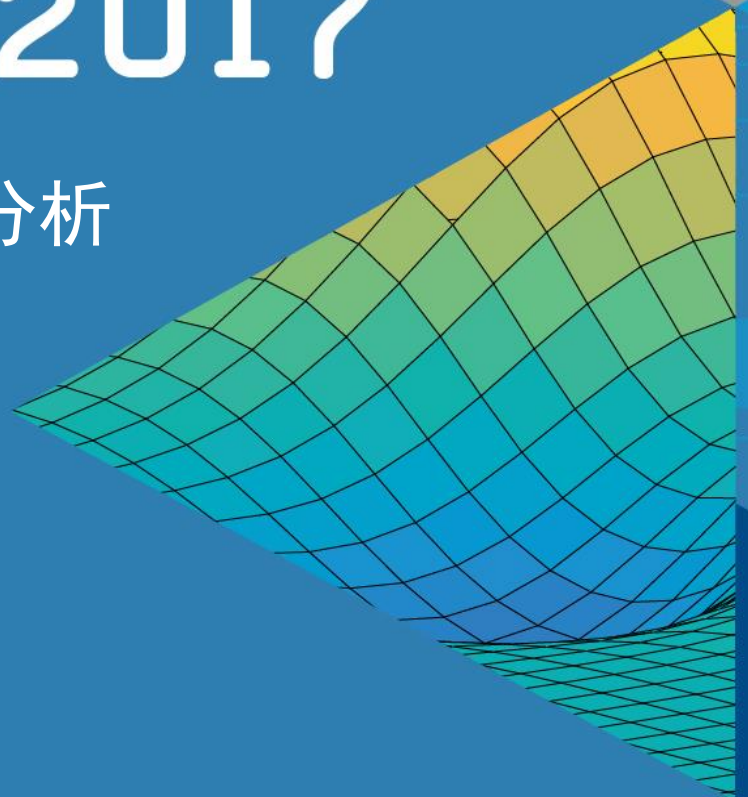
MathWorks

MATLAB EXPO 2017

使用MATLAB实现大数据的处理与分析



马文辉　高级应用工程师，MathWorks 中国

什么是大数据?

“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.”

Wikipedia

大数据会带来哪些挑战?

- 传统的工具和方法不能有效工作
  - 数据获取和处理变得越来越困难;
  - 需要学习使用新的工具和编程方法;
  - 需要重写算法和更改代码，以应对数据的规模和复杂程度的增加;
- 计算结果的质量会受到影响
  - 例如，很多情况下不得不使用子集进行计算;

大数据工作流程

数据访问

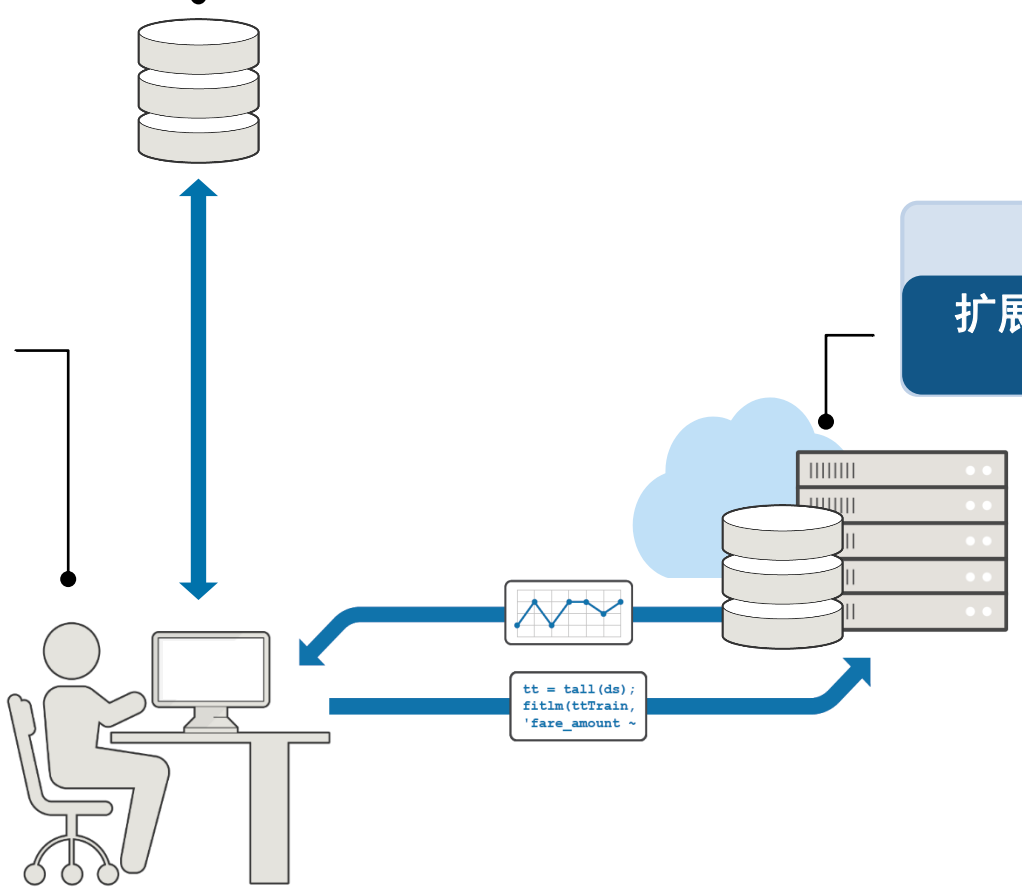
数据规模超过内存限制

数据处理与分析

使用传统工具或新的大数据工具

计算扩展

扩展到大数据系统，例如 Hadoop



大数据所需解决方案

- 轻松快速的访问数据，无论它们存储在哪里；
- 小数据集上实现原型算法开发；
- 数据处理和分析的大数据集扩展；
- **使用与数据集规模大小无关的MATLAB语法；**

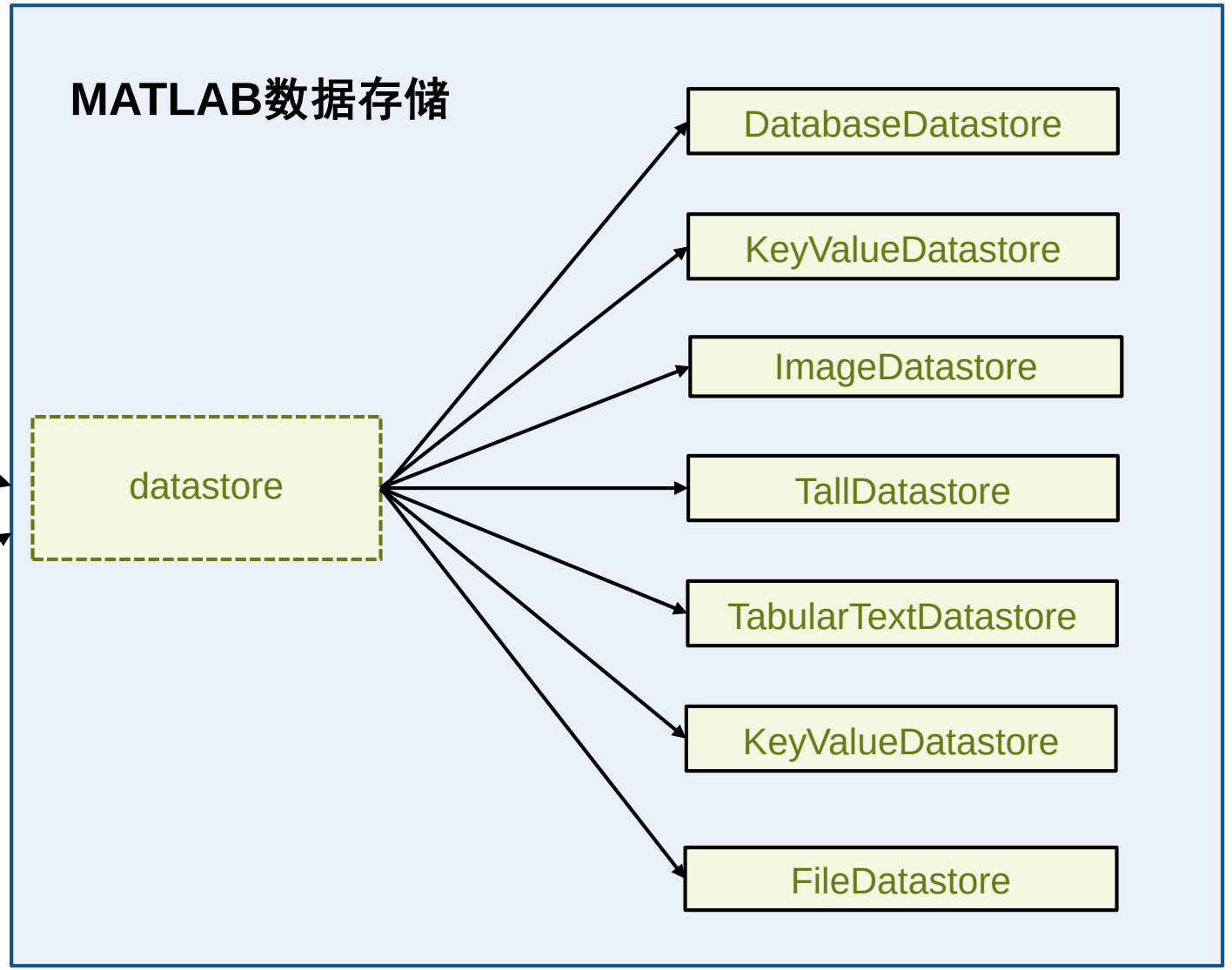
多数据源

- 业务数据与工程数据

Repositories
- Databases (SQL)
- NoSQL
- Hadoop

File I/O
- Text
- Spreadsheet
- Image

Tall Array

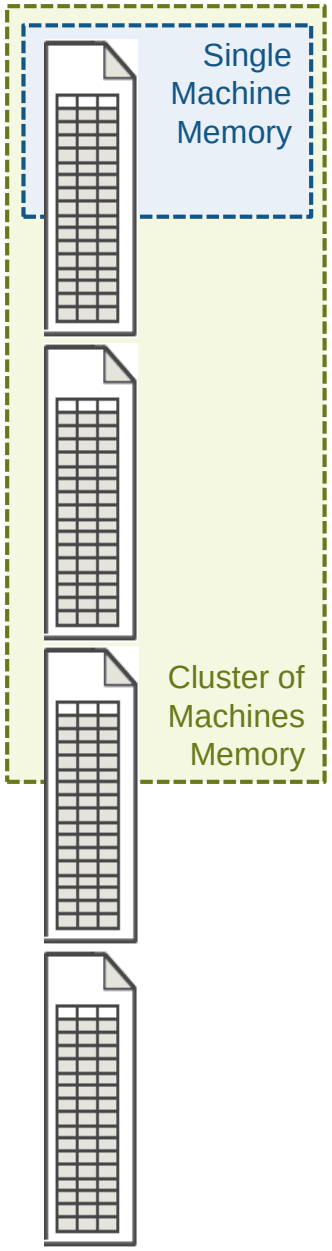


tall arrays R2016b

- 数据规模超过内存的限制
- 数据记录条数巨大 (“tall”)
- 与MATLAB array相似
  - 支持数据类型：numeric types, tables, datetimes, strings, 等等；
  - 支持算术运算、统计运算、索引等；
  - **支持统计与机器学习工具箱的众多算法** (regression, clustering, classification, etc.)；

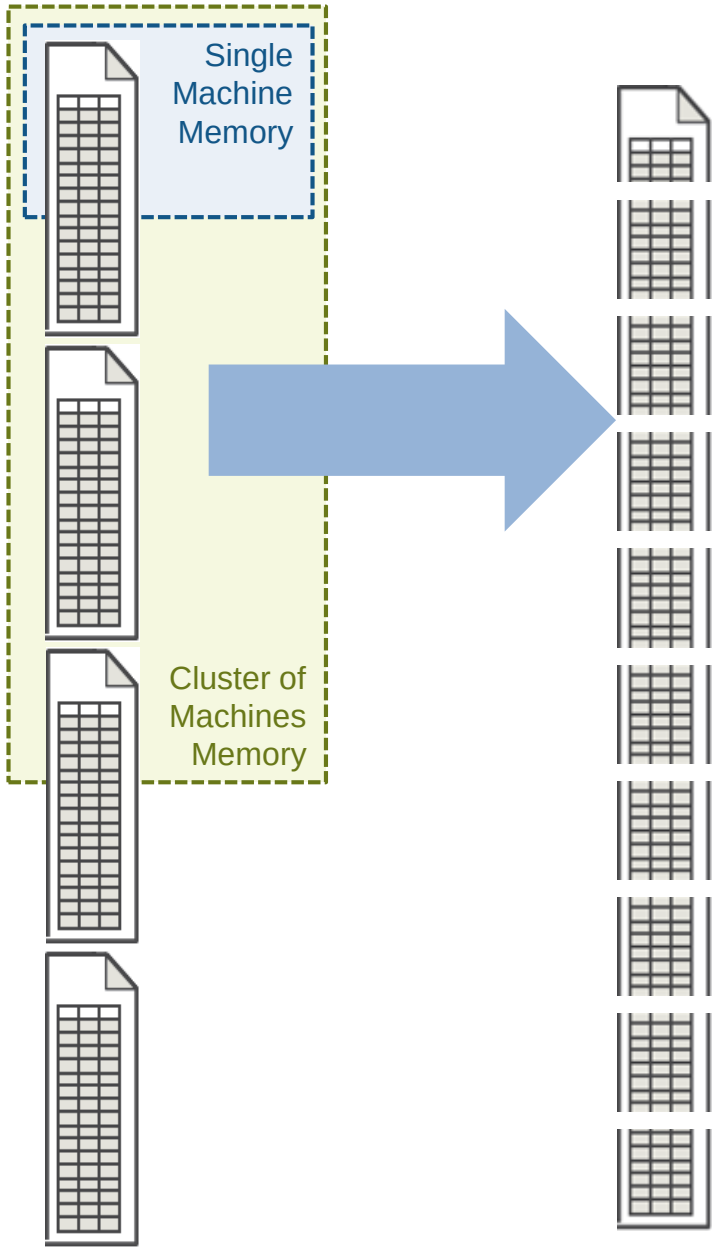
tall arrays R2016b

- 数据存储在多个文件中；
- 通常是表格数据；
- 垂直堆叠的文件；
- 数据规模超过内存容量，甚至是集群内存容量



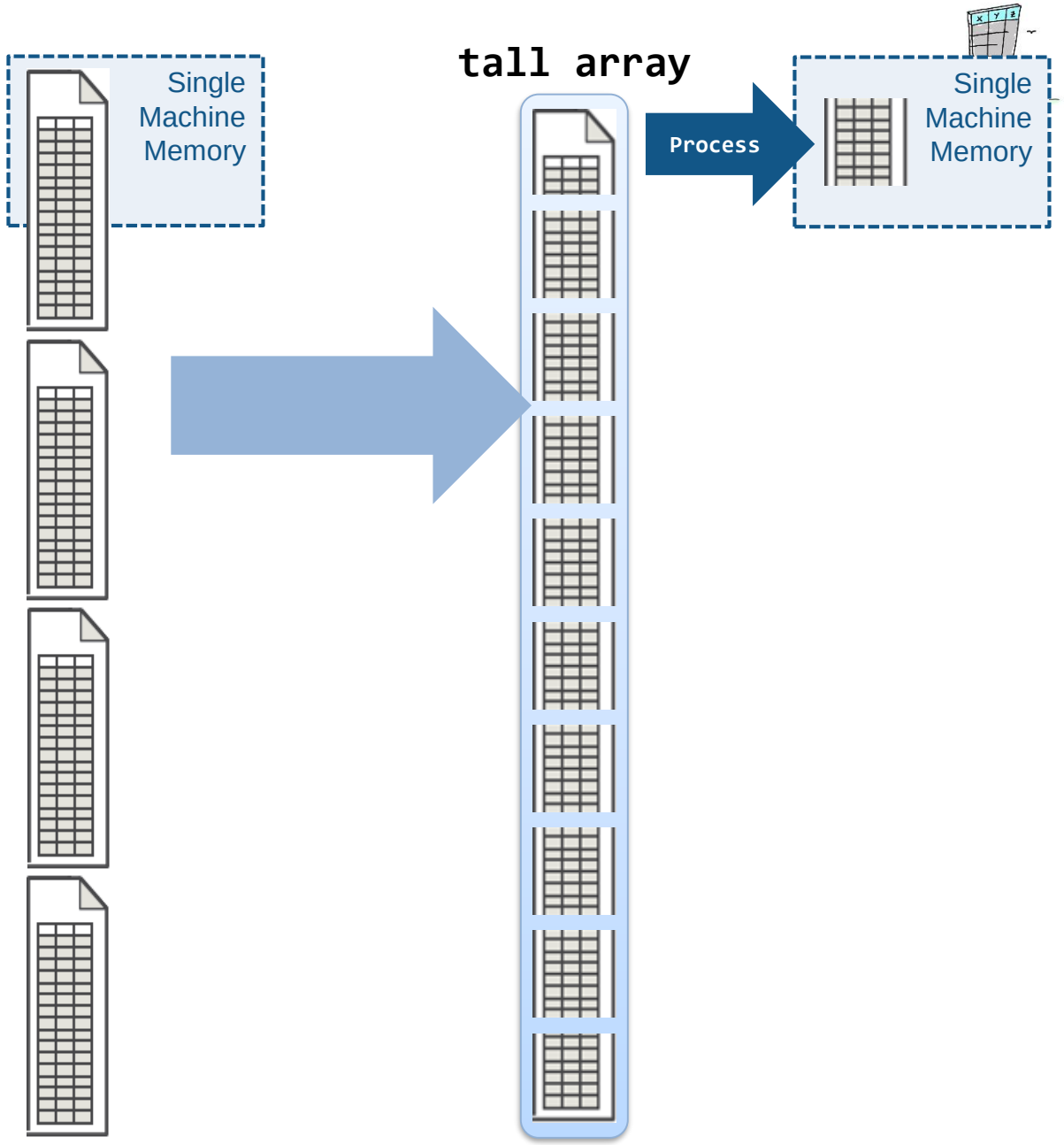
tall arrays R2016b

- 自动将数据分解成适合内存的小块（“chunks”）



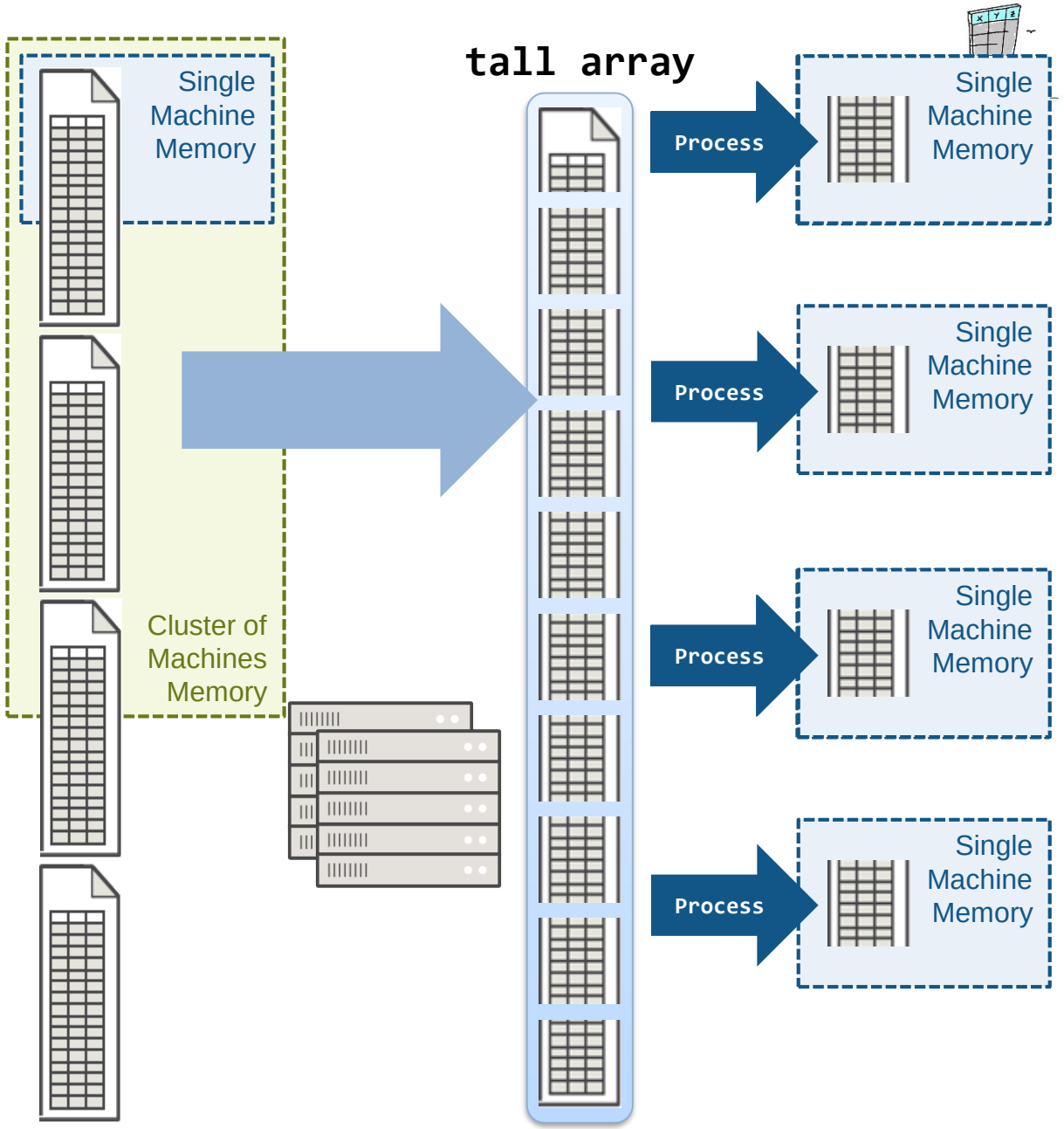
tall arrays R2016b

- 数据块（“Chunk”）的处理自动进行；
- 对 tall arrays 数据类型的处理与普通数组相同；



tall arrays R2016b

- 利用Parallel Computing Toolbox（并行计算工具箱），可以并行处理数据块；
- 可以通过MATLAB Distributed Computing Server（MATLAB分布式计算）将计算扩展到整个集群；



基于tall array的大数据处理流程

Access Data

- Text
- Spreadsheet (Excel)
- Database (SQL)
- Custom Reader

Datastores for common types of data

Tall Data Types

- `table`
- `cell`
- `double`
- `numeric`
- `cellstr`
- `datetime`
- `categorical`

Tall versions of commonly used MATLAB data types

Exploration & Pre-processing

- Numeric functions
- Summary 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
- Validation techniques

Key statistics and machine learning algorithms

"Tall" data types and functions for use with out-of-memory data

示例: 利用MATLAB处理大数据

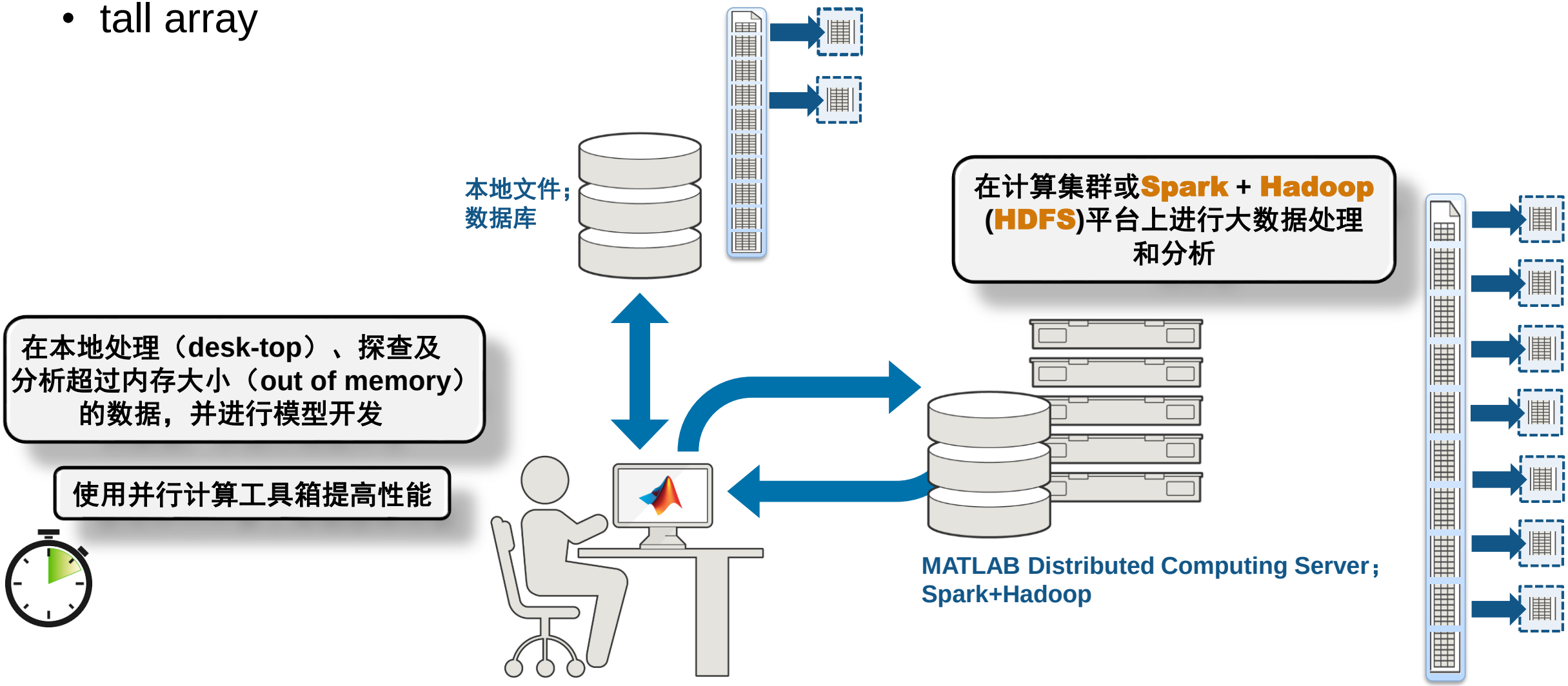
- **目标: 创建一个模型，用于预测纽约出租车乘坐费用**

- **输入:**
  - 每月出租车日志文件
  - 本地小规模数据集 (~2 MB)
  - 全量数据集 (~25 GB)

- **方法:**
  - 数据预处理和探查
  - 开发并验证预测模型
    - 使用小的数据子集做模型原型开发
    - 将模型扩展的HDFS上全量数据集

总结

- tall array



总结

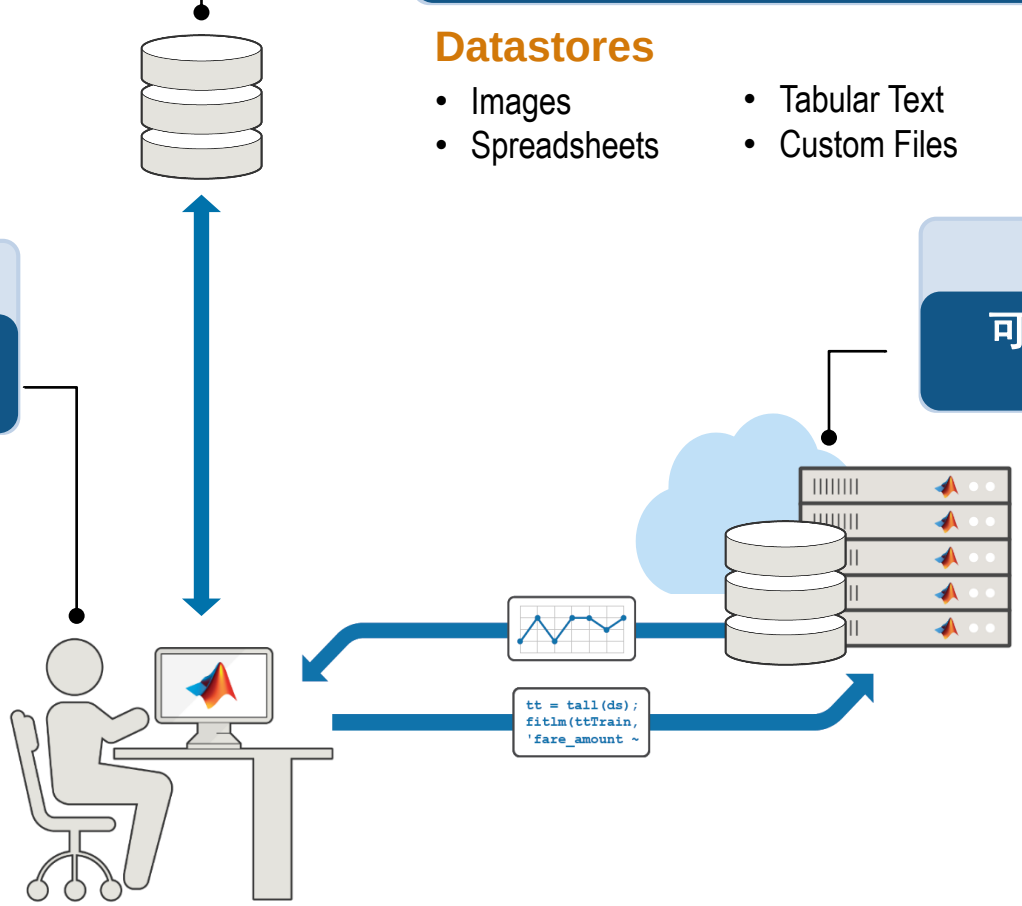
- MATLAB大数据处理

数据访问

可以访问超过内存大小的数据源

Datastores

- Images
- Spreadsheets
- Tabular Text
- Custom Files
- SQL
- Hadoop (HDFS)



数据处理与分析

适合于领域专家使用的内置数据处理与分析能力，并且可以实现本地的处理大数据

Tall Arrays

- Math
- Statistics
- Visualization
- Machine Learning

GPU Arrays

- Matrix Math
- Image Processing

Deep Learning

- Image Classification

扩展性

可以平滑扩展的计算集群或者Spark+Hadoop集群

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

获取更多信息

- 网站:

  https://www.mathworks.com/solutions/big-data-matlab

- Web search for:

  “**Big Data MATLAB**”

谢 谢