PyConChina 2019

PySpark在大数据/机器学习方面的应用

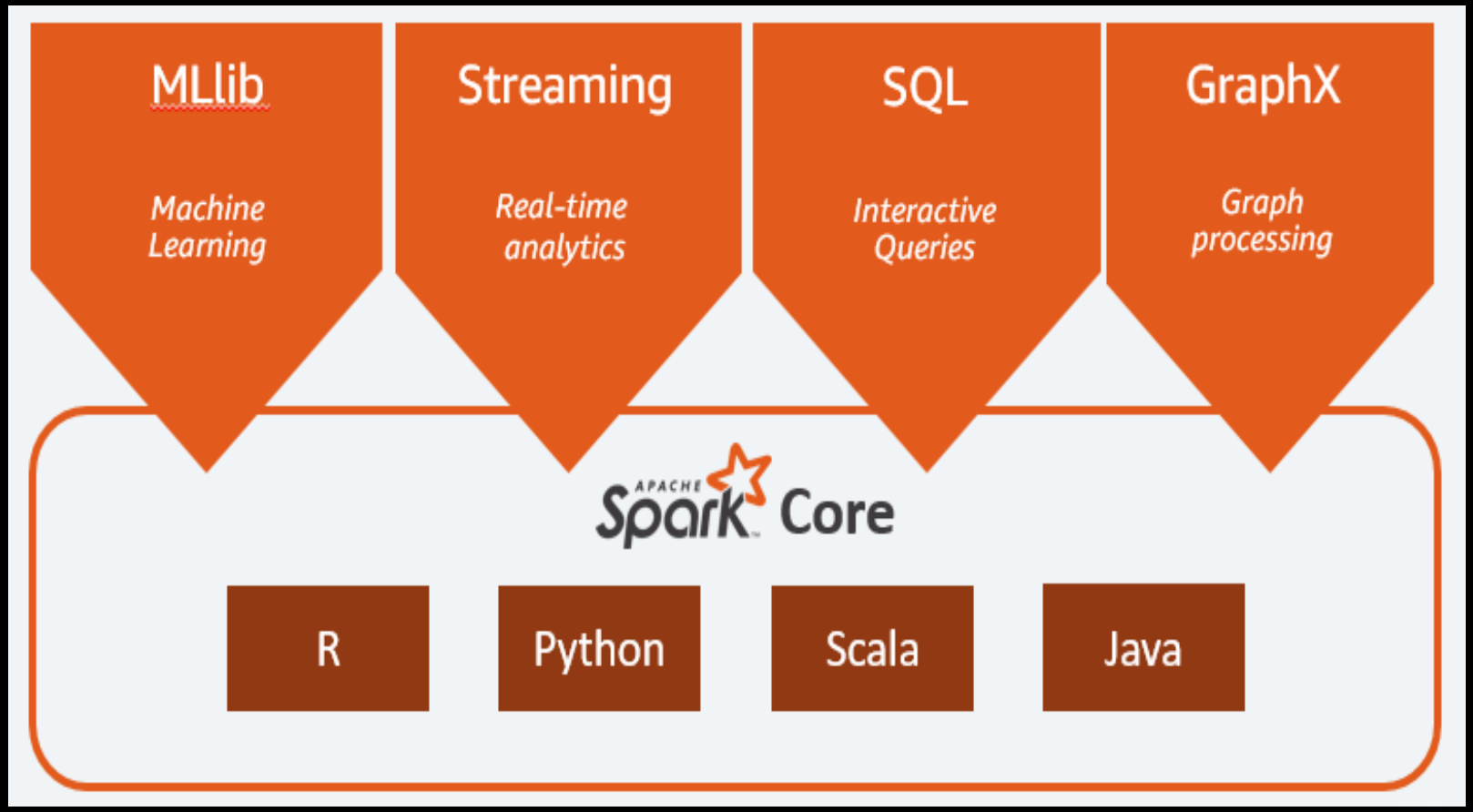
张宏柯

目录 CONTENTS

一、什么是PySpark?

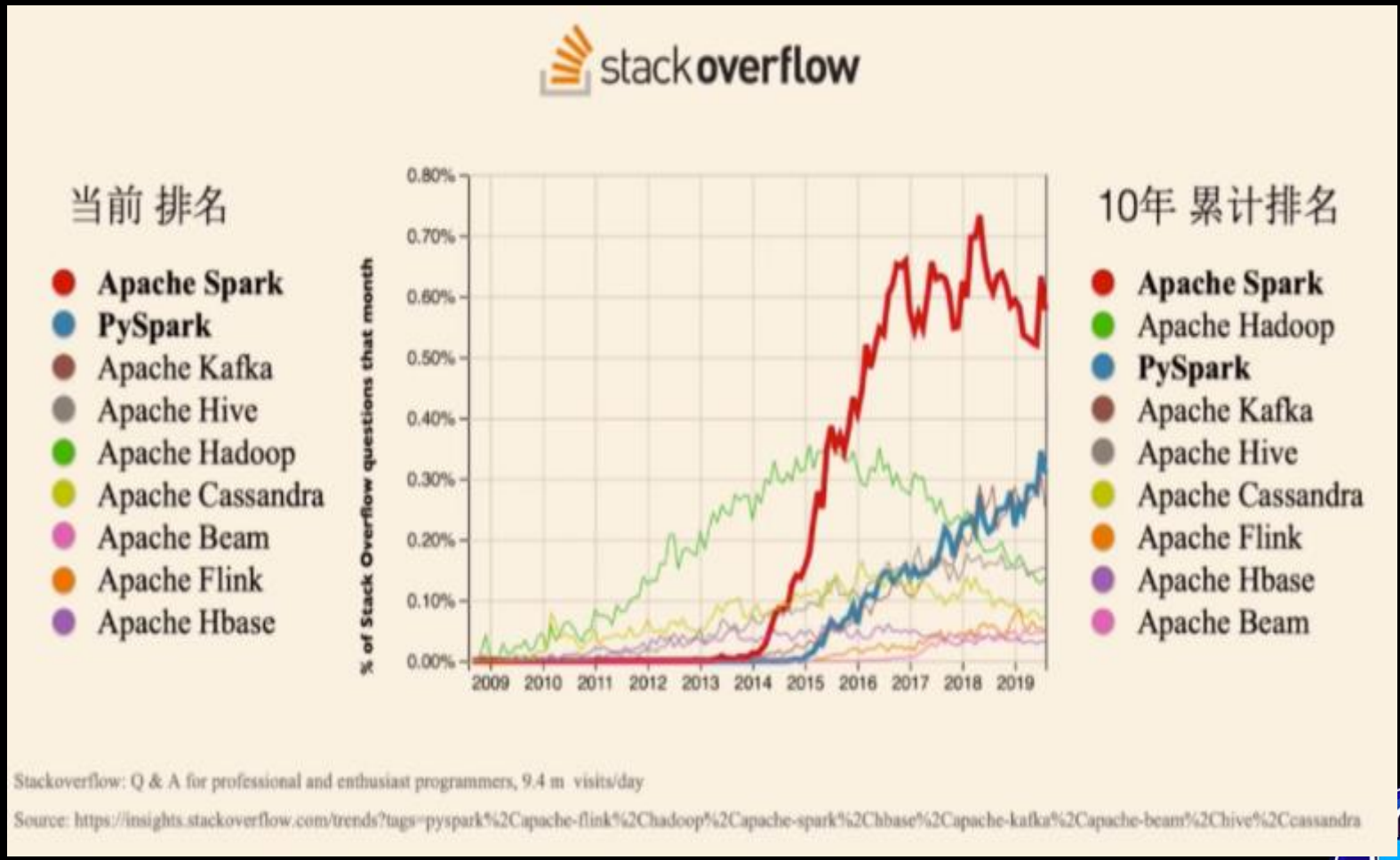
PySpark是针对Spark的Python API

spark


MLlib
Machine Learning
Streaming
Real-time analytics
SQL
Interactive Queries
GraphX
Graph processing
APACHE Spark Core
R
Python
Scala
Java

二、为什么选择PySpark?

•为什么选择PySpark?



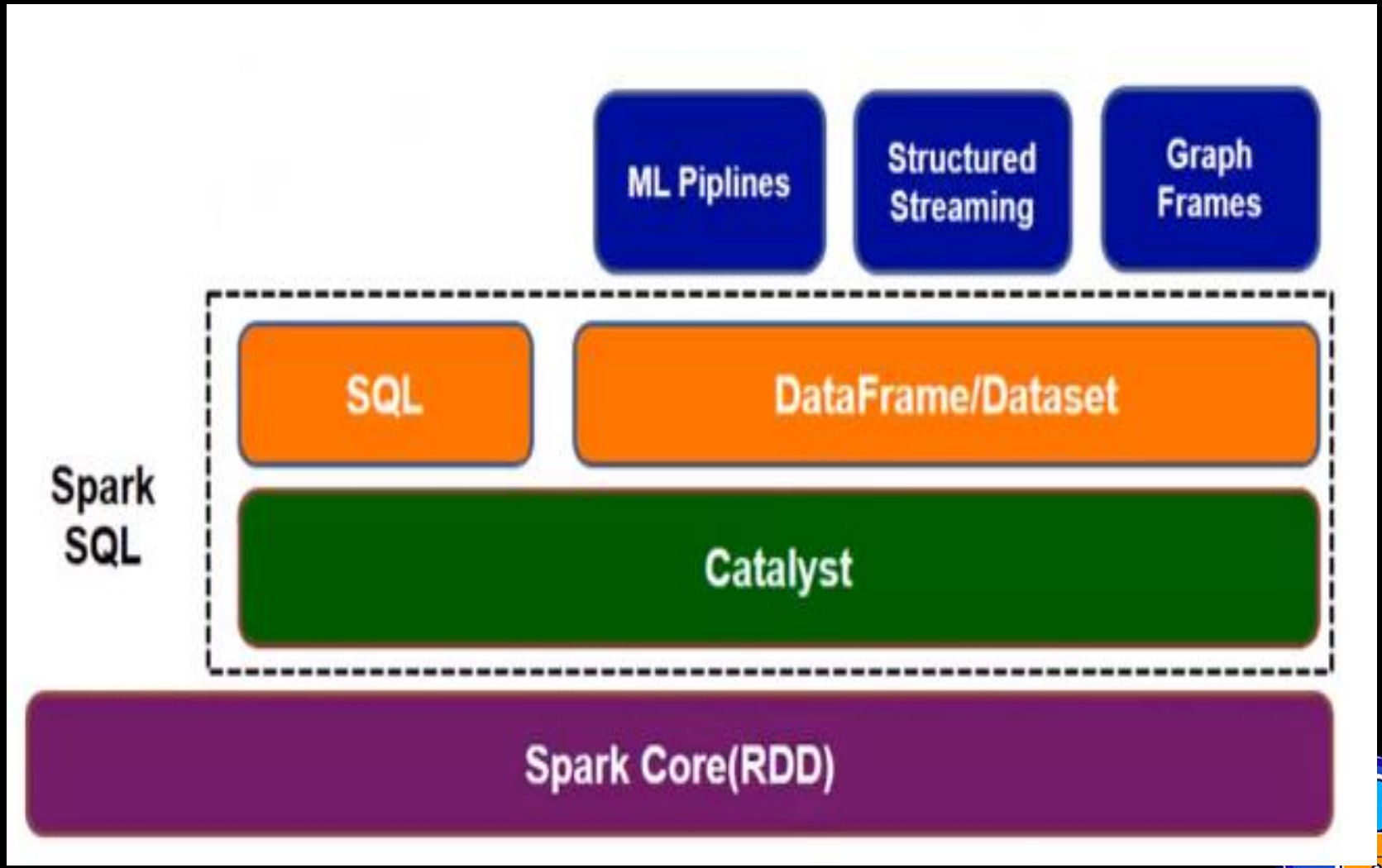
·为什么选择PySpark?

- 一、**易于学习：**Python因其语法和标准库相对容易学习
- 二、**大量的库：** Scala没有足够的数据科学工具，Scala缺乏良好的可视化和本地数据转换
- 三、**巨大的社区支持：** Python拥有一个全球社区，拥有数百万开发人员，可在数千个虚拟和物理位置进行在线和离线交互
- 四、无需编译打包，直接运行

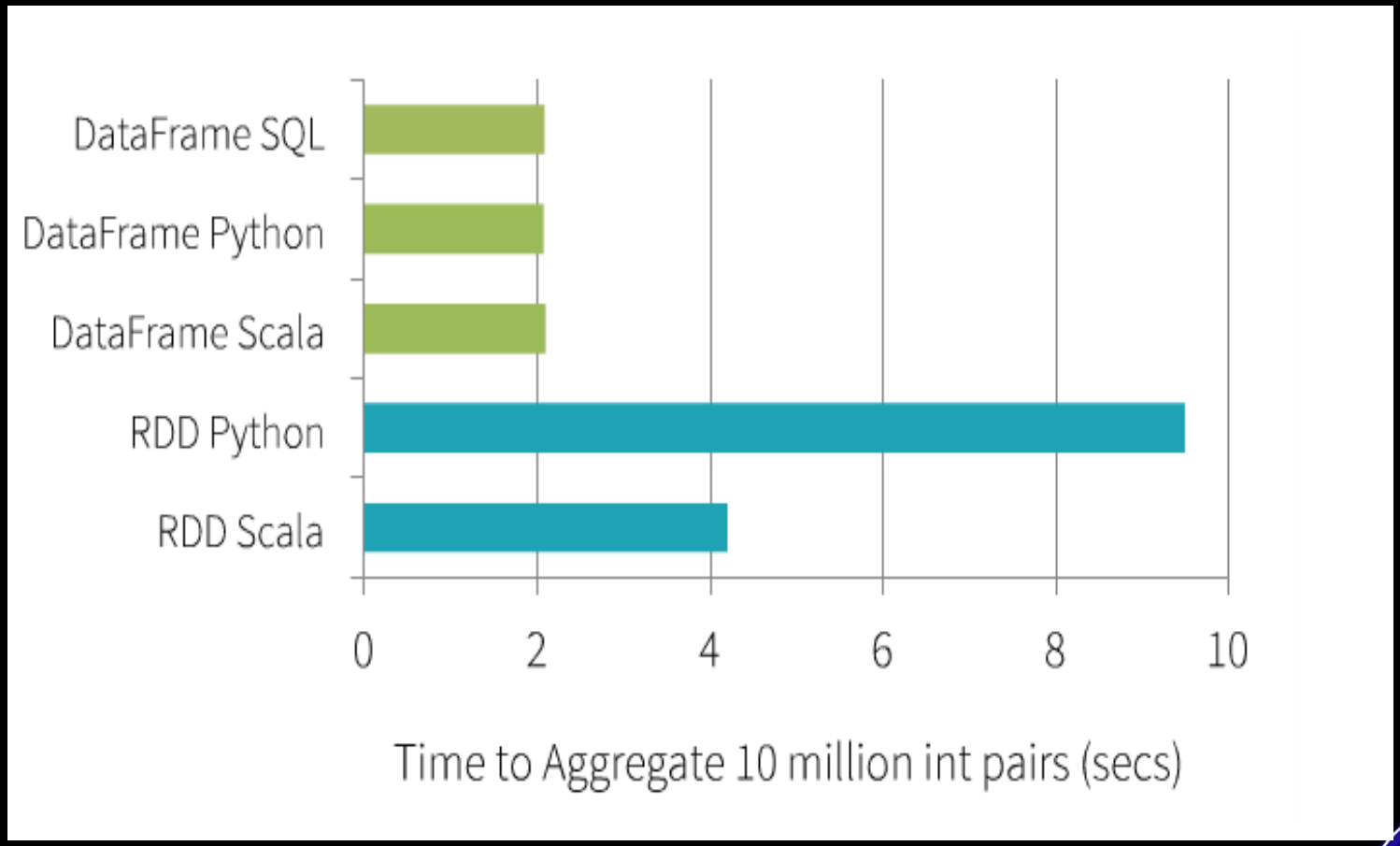
三、使用PySpark进行大数据处理

SparkSQL & SparkStreaming

•使用PySpark进行大数据处理



•使用PySpark进行大数据处理



•使用PySpark进行大数据处理

- 离线数据处理(数据仓库)：sparksql
- 实时数据处理(ETL)：sparkstreaming

- 优点：简单上手，易于结合python库，如：spark dataframe可与pandas dataframe互相转换

- 有待改进：py版本sparkstreaming不是很稳定

•Pandas vs PySpark DataFrame

Pandas和Pyspark DataFrame一些区别如下：

1、在Pyspark DataFrame上的操作在群集中的不同节点上并行运行， Pandas单机的

2、PySpark DataFrame中的操作本质上是懒惰的， Pandas ，只要应用任何操作，我们都会立即得到结果

3、在PySpark DataFrame中，由于其不可变的属性，我们无法更改该DataFrame，我们需要对其进行转换。但是Pandas却并非如此。

4、Pandas API比PySpark DataFrame支持更多的操作。

四、使用PySpark进行机器学习

Sparkmllib & Sparkml

Spark 机器学习算法

1、两个算法包

- Spark.mllib:包含原始API,构建在RDD之上
- Spark.ml:基于dataframe构建的高级API

2、如何选择

- Spark.ml具备更优的性能和更好的拓展性，建议优先使用
- 从Spark 2.0开始，spark.mllib软件包中基于RDD的API已进入维护模式。 Spark的主要机器学习API现在是spark.ml软件包中基于DataFrame的API

Spark 机器学习算法

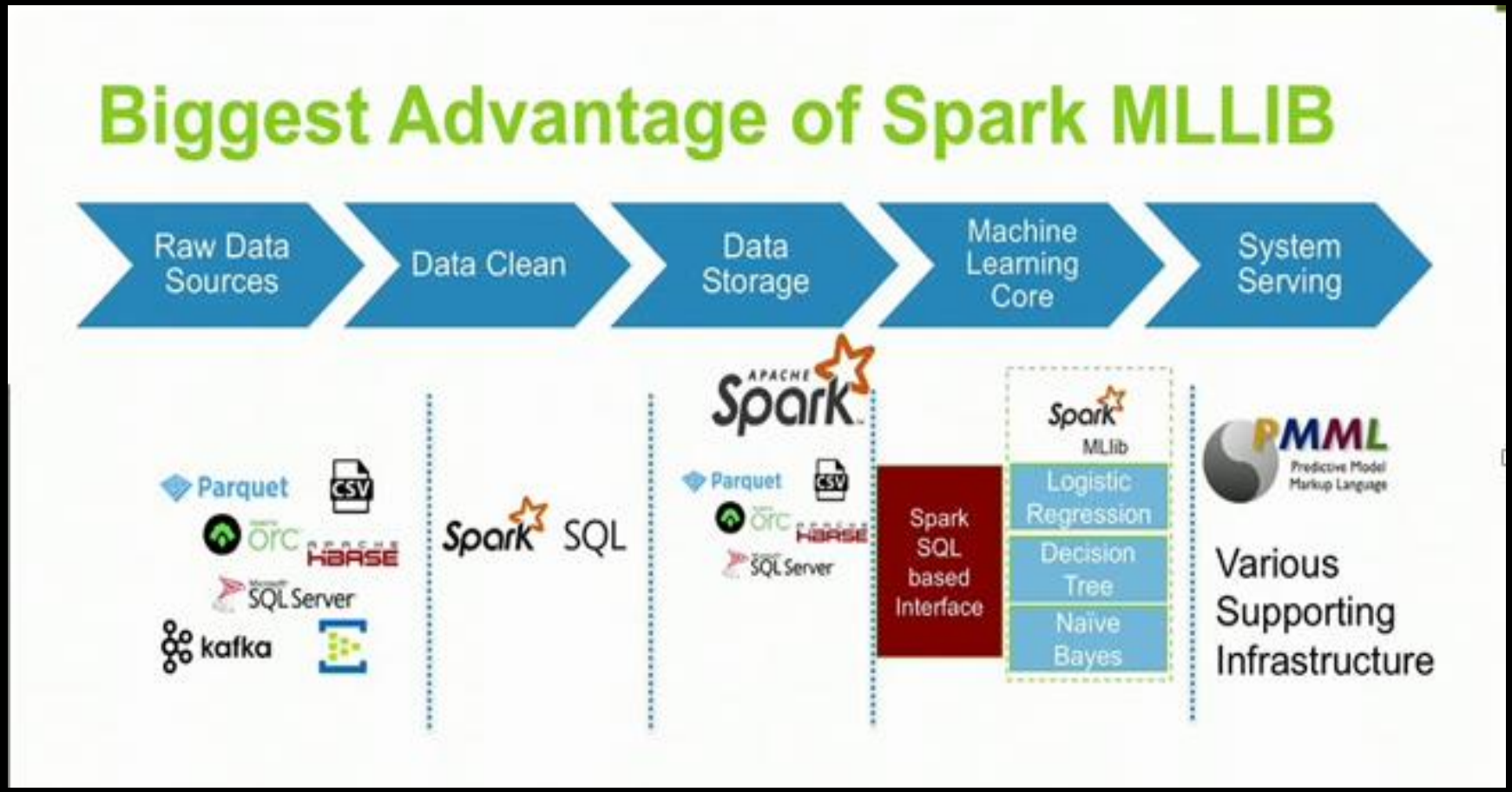
- 一、分类算法
- 二、回归算法
- 三、聚类算法
- 四、推荐算法

Spark 机器学习算法

- Classification
  - Logistic regression
    - Binomial logistic regression
    - Multinomial logistic regression
  - Decision tree classifier
  - Random forest classifier
  - Gradient-boosted tree classifier
  - Multilayer perceptron classifier
  - One-vs-Rest classifier (a.k.a. One-vs-All)
  - Naive Bayes
- Regression
  - Linear regression
  - Generalized linear regression
    - Available families
  - Decision tree regression
  - Random forest regression
  - Gradient-boosted tree regression
  - Survival regression
  - Isotonic regression
    - Examples

- K-means
  - Input Columns
  - Output Columns
  - Example
- Latent Dirichlet allocation (LDA)
- Bisecting k-means
  - Example
- Gaussian Mixture Model (GMM)
  - Input Columns
  - Output Columns
  - Example

Spark 机器学习算法



Spark 机器学习算法

- 优点：能够并行训练
- 缺点：支持库不够全面 比如kmeans 有 dbscan 没有
- ==========一个Demo==========

THANK YOU

zxczhkzxc