# 大数据基础平台算法调研

|  |  |
| --- | --- |
| 作者： | 刘炜 |
| 版本： | 0.2 |
| 时间： | 2016-07-18 |

## 调研简介

项目“大数据基础平台搭建及关键技术研究”中，大数据应用层将重点面向智慧城市的特定行业需求，集成并应用大数据素材，提供舆情分析（信息安全）、商业智能（电子商务）等服务。

对此需求，调研需要的算法以及开源库的选择，并规划设计解决问题方案。

## 调研结果

结论：

1. 算法目标合适，满足项目需求。[参见](#_算法目标确认)
2. 算法采用MLlib，完全没问题。[参见](#_MLlib)
3. 平台采用Spark集群，Spark + Hadoop（HDFS）

## 设计方案

1. 通过Spark集群，运用MLlib可实现大数据的分析，可对1-100TB的数据进行高效的运算。
2. 先运行MLlib算法测试用例，再结合具体需求包装及改进算法。
3. 对于超大数据量，文件数据的存储与分析通过Spark结合Hadoop的HDFS实现，对于数据库的存储与分析通过分布式数据库实现。

## 风险

暂时没有

## 算法目标确认

通过项目需求知道本平台主要应用于智慧城市特定行业需求

例如：

舆情分析（信息安全）：通过现象分析本质，趋势预测等。

商业智能（电子商务）：数据仓库、线上分析处理、决策系统、推荐系统等。

以上需求需要算法：

决策树、KNN、SVM等常见分类算法

K均值等常见聚类算法

Apriori等常见关联分析算法

统计与回归分析算法

由于目前大数据应用非常广泛，以上需求的实现已经有相关企业应用在实际工作中。而且对于某个问题的解决往往是基于多的算法，而进行综合评价。所以所用的算法为以上所列（经典算法）但也不局限于此。

目前很多开源包含了很多算法，几乎面面俱到，在实际应用中也屡见不鲜。在Spark集群中内置的MLlib库，而Spark的应用非常广泛，可以从2015 Spark 技术峰会中了解到。

**综上：**需求的解决可以由以上的经典算法解决，而且由于机器学习的白热化，任何一个机器学习开源库都包含这些算法，对于非常活跃的开源库也将不断有测试稳定的新的算法入驻，所以用任何一个开源库都可以满足。

## 开源库的比较

|  |  |  |  |
| --- | --- | --- | --- |
| 开源库 | 语言 | 优缺点 | 项目地址 |
| TensorFlow | C++ | **优点：**谷歌开源深度学习系统。  **缺点：**太大，东西太多 | <https://github.com/tensorflow/tensorflow> |
| Scikit-learn | Python | **优点：**单机学习库，机器学习  分类、回归、聚类、数据降维、模型选择、数据预处理  **缺点：**单机 | <https://github.com/scikit-learn/scikit-learn> |
| PredictionIO | Scala | **优点：**PredictionIO 机器学习服务器，MongoDB数据库，Hadoop系统架构。  **缺点：**不是算法库，只是方便搭建推荐系统的一个包装。 | <https://github.com/PredictionIO/PredictionIO> |
| Oryx 2 | java | **优点：**回归、分类、集群和协同过滤，也基于Spark内存计算  **缺点：**暂无 | Oryx2  <https://github.com/OryxProject/oryx> |
| MLlib | JAVA  Scala | **优点：**用的非常多，含在Spark里面，包含各种算法。Spark为开源的  **缺点：**无 | <http://spark.apache.org/mllib/>  <https://github.com/apache/spark> |

经过比较分析，由于MLlib用的非常多，并且在Spark集群的内置库，所以对大规模的数据有相当的应用范例，比如百度的运行着Spark集群1300个节点，包含数万核心和上百TB内存。

由于Spark的广泛和大规模应用以及开源泉，MLlib作为其中的机器学习库，势必将越来越完善。

## Spark

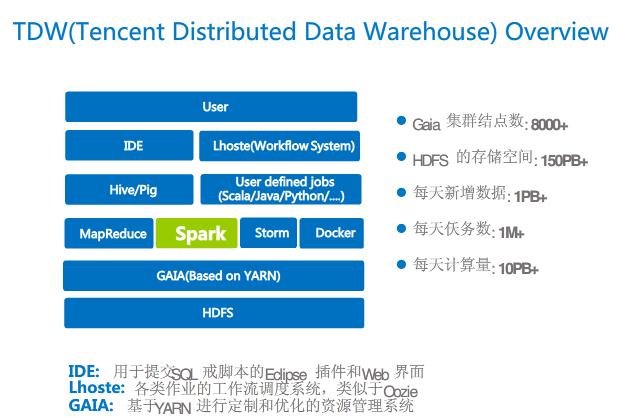
Spark是一个基于内存计算的开源的集群计算系统，大数据领域最活跃的开源项目之一，在BAT都用的比较广泛。

支持数据处理1－100TB的快速处理，完成过1PB数据的排序。

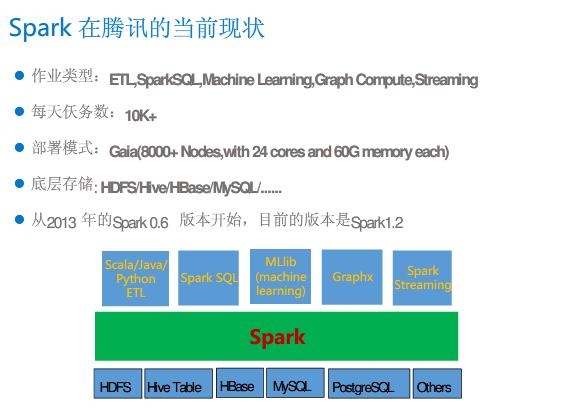
1. 轻量级快速处理
2. 易于使用，支持多语言
3. 支持复杂查询
4. 实时的流处理Sparks Streaming
5. 可以与Hadoop整合

## Spark在腾讯的应用（[参见附录](#_附录)）

MLlib在腾讯Spark集群机器学习中处于主要地位，用于预测用户点击率、计算两个人之间的共同好友、用于ETL的SparkSQL和DAG任务等。



（图：腾讯分布式数据仓库）



（图：腾讯Spark当前现状）

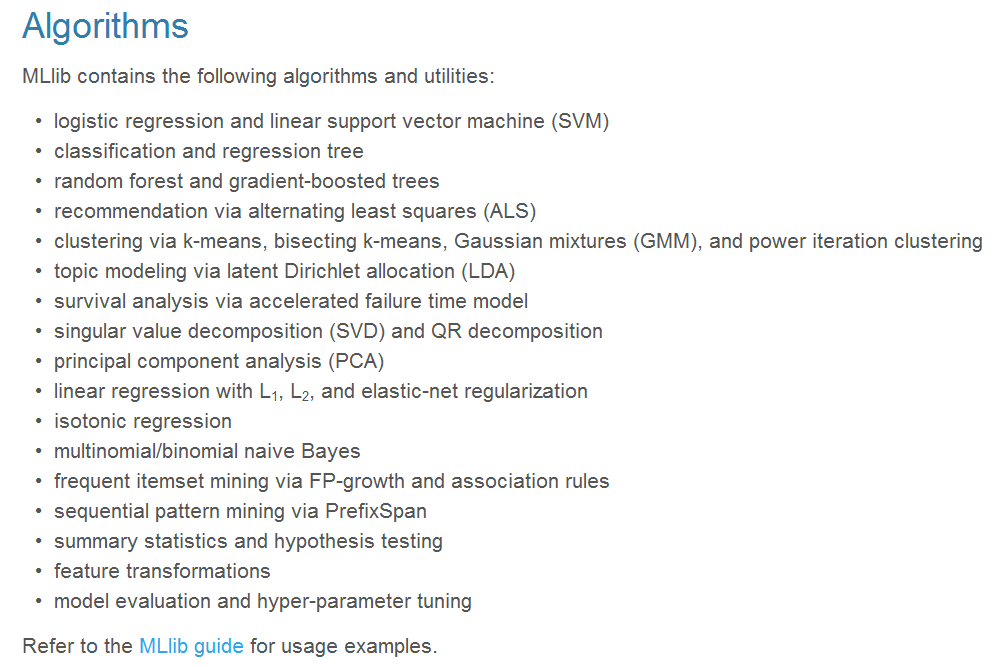
## Spark在百度的应用

在实际生产环境，百度运行着1300台规模的单集群（包含数万核心和上百TB内存），公司内部同时还运行着大量的小型Spark集群。

## MLlib

MLlib支持常见的机器学习问题：分类、聚类、回归、协同过滤、梯度下降基础算法。完全满足项目需求。

MLlib是高效的算法，比MapReduce快一百倍左右。由其对于需要频繁迭代的计算。



（图：MLlb包含算法）

## 附录

1、2015 Spark技术峰会-腾讯在Spark上的应用与实践优化-王联辉

<http://download.csdn.net/detail/happytofly/8637461>

2、2015 Spark技术峰会资料分享

<http://download.csdn.net/album/detail/1611>

3、机器学习库（MLlib）指南

<http://ifeve.com/spark-mllib/>

4、MLlib介绍

<http://spark.apache.org/mllib/>

201522220203\_liuwei