方圆并济：基于 Spark on Angel 的高性能分布式机器学习

Tencent——数据平台部

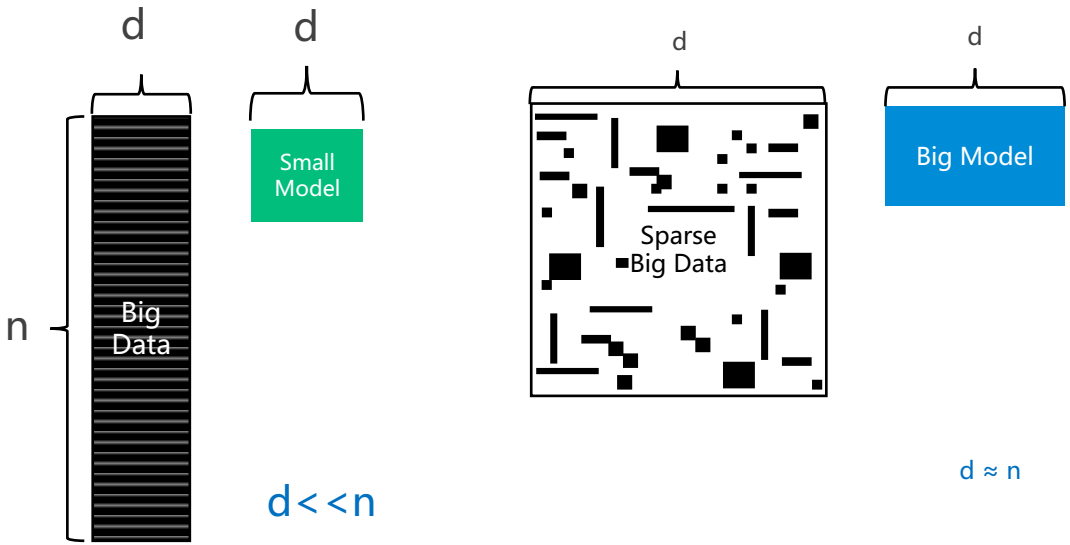
Andymhuang（黄明）

目录

- 源起
- Spark on Angel
- Spark on Angel的开发
- Spark on Angel的算法
- 性能和比较
- 后续规划

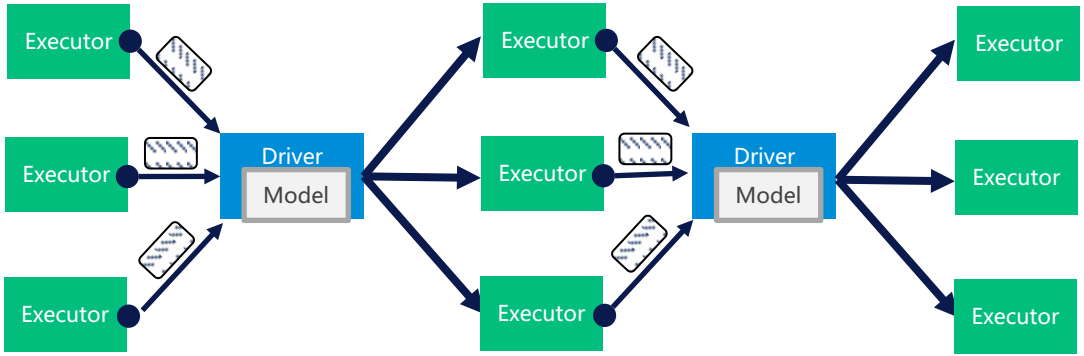
源起

腾讯的产品需求



寻找满足十亿级维度的工业级的分布式机器学习平台

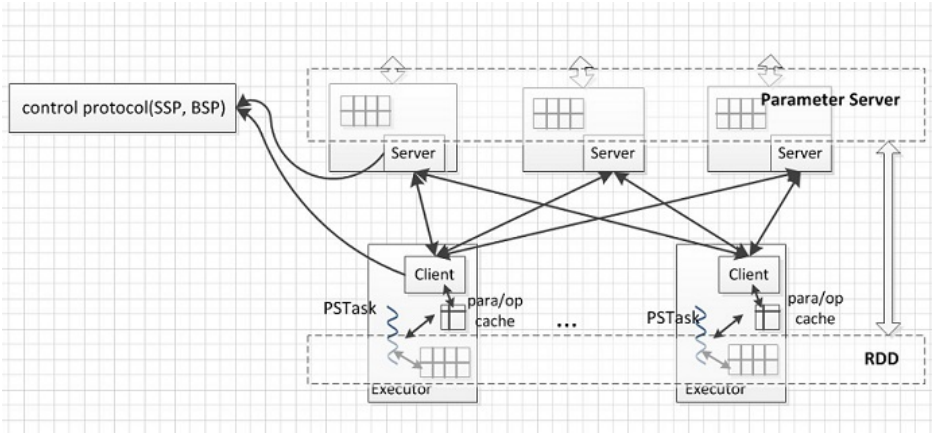
Spark机器学习的瓶颈



- Driver成为参数汇总的单点瓶颈，难以支撑大规模模型及数据
- 满足不了十亿级维度的模型训练，实际应用中需要进行降维处理
- Executor之间相互等待，整体效率不高

One Issue

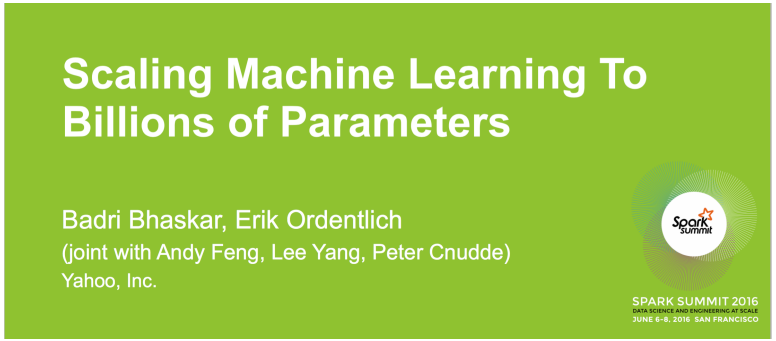
A Prototype of Parameter Server



https://issues.apache.org/jira/browse/SPARK-6932

2015

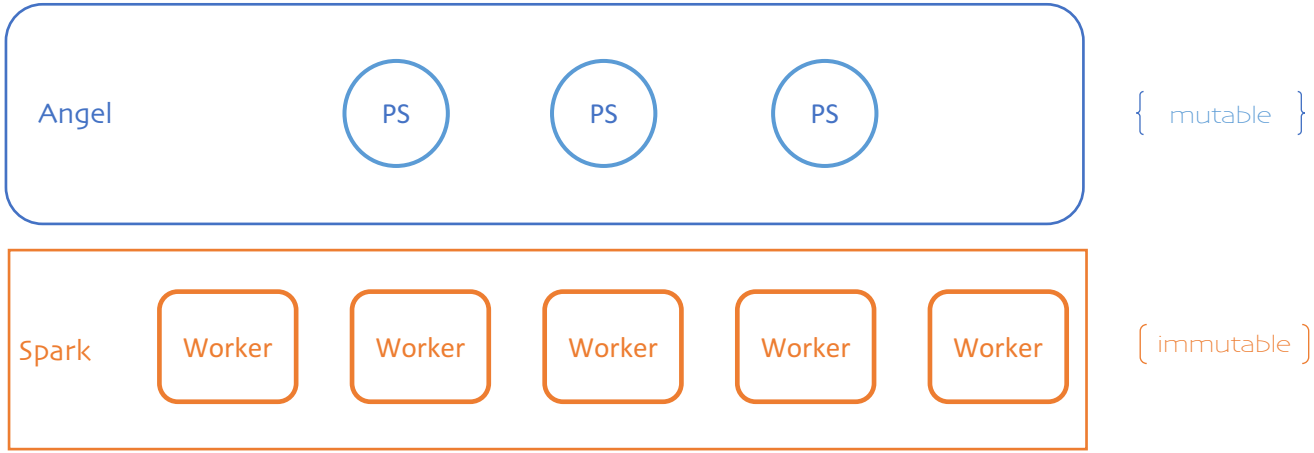
Glint & Yahoo



2016

理念 —— 方圆并济

Angel



1. 分离系统中的变和不变
2. 以少博多
3. 对Spark Core的侵入性越少越好

Spark on Angel

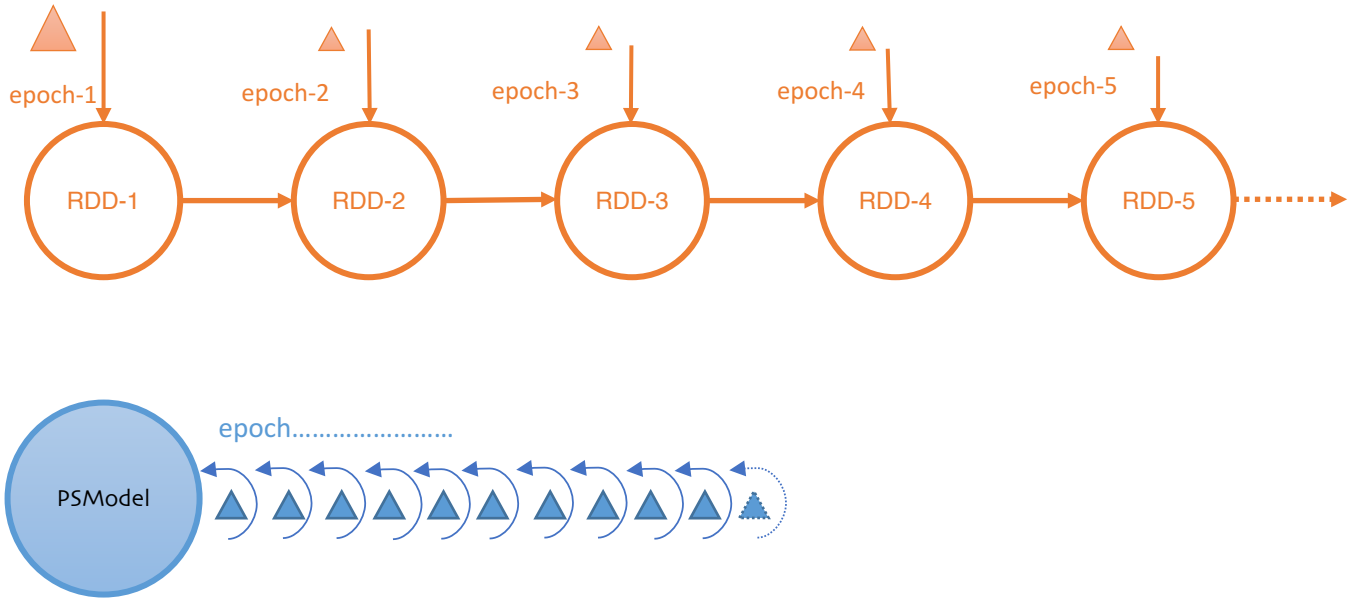
核心抽象

Mapper
Reducer

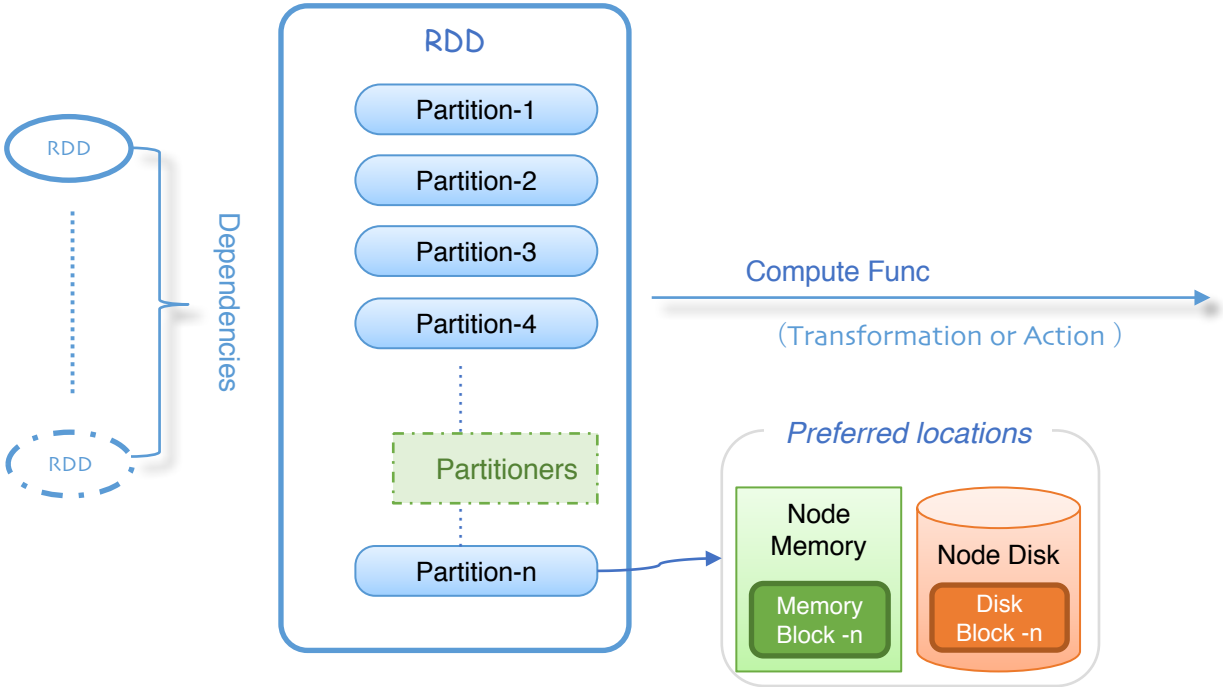
RDD

PSModel

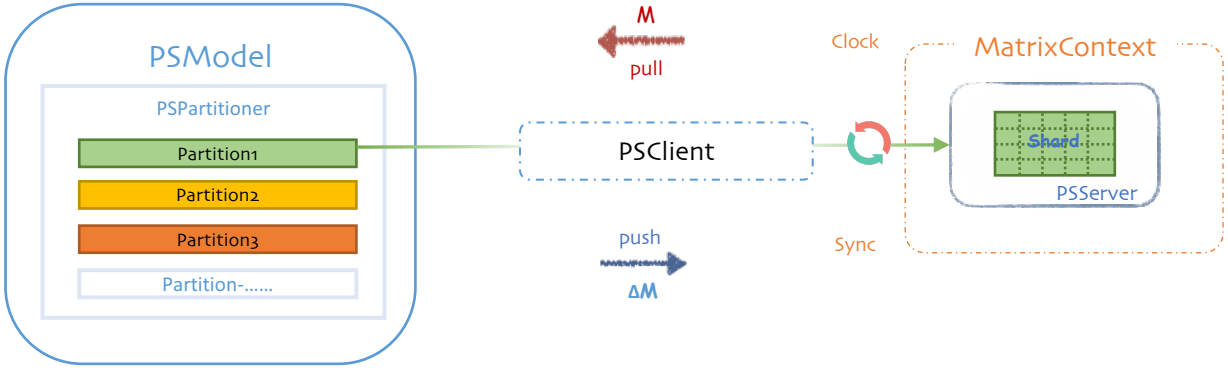
RDD vs PSModel



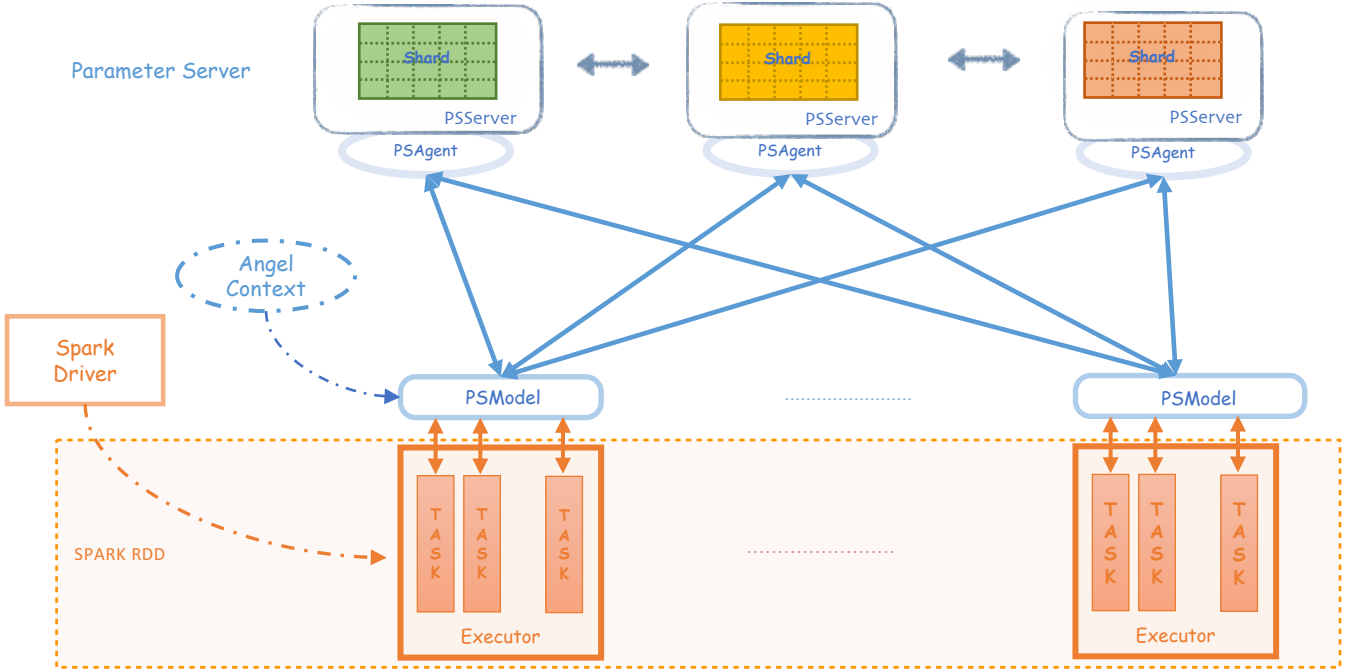
RDD的核心抽象


RDD
RDD
Dependencies
RDD
Partition-1
Partition-2
Partition-3
Partition-4
Partitioners
Partition-n
Compute Func
(Transformation or Action)
Preferred locations
Node Memory
Memory Block -n
Node Disk
Disk Block -n

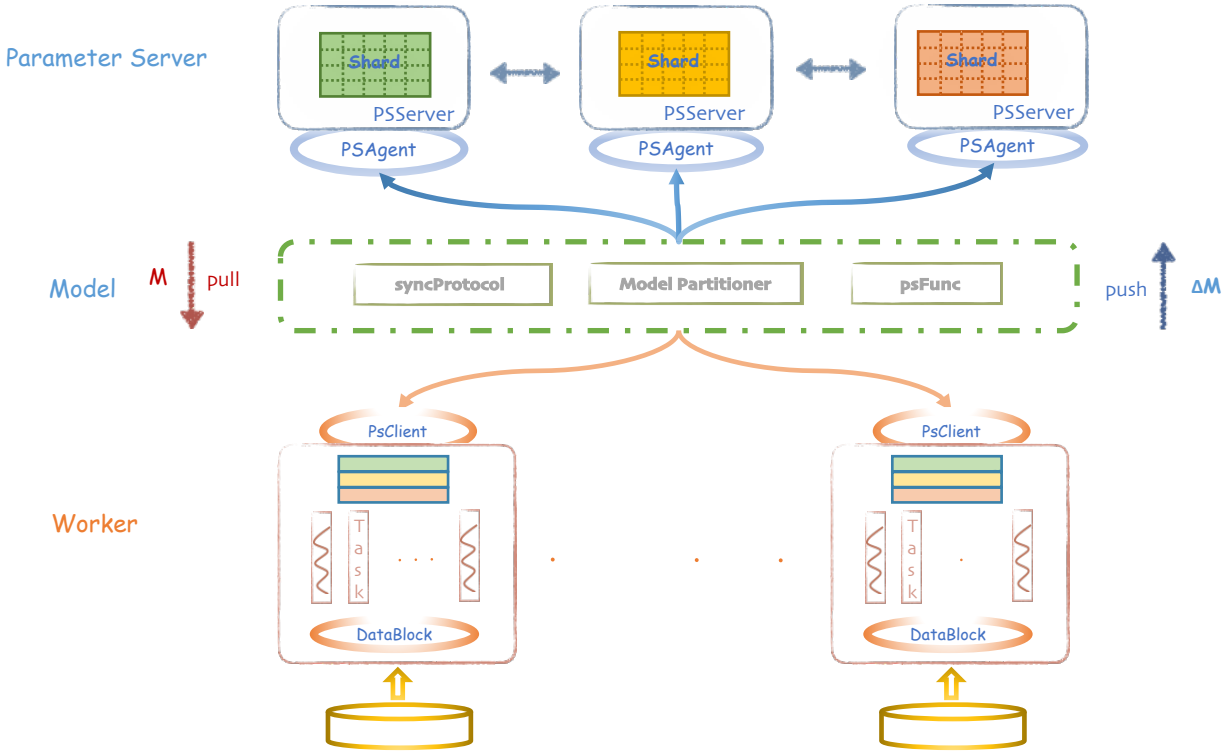
PSModel的核心抽象



Spark on Angel的架构



Angel的系统框架

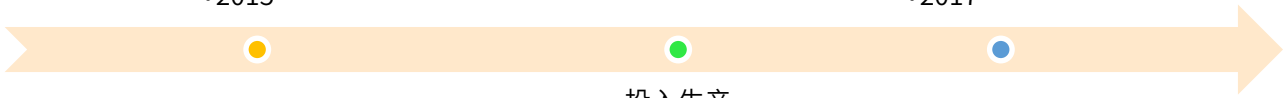

Parameter Server
Shard
PSServer
PSAgent
Shard
PSServer
PSAgent
Shard
PSServer
PSAgent
Model
M
pull
syncProtocol
Model Partitioner
psFunc
push
ΔM
Worker
PsClient
Task
DataBlock
PsClient
Task
DataBlock

Angel

开始研发
•2015

投入生产
•2016

正式开源 V1.0.0
•2017



- 能支持十亿级别维度的模型训练
- 基于Matrix/Vector的模型自动切分和管理，兼顾稀疏和稠密两种格式
- 提供多种同步控制机制（BSP/SSP/ASP）

工业级别可用的参数服务器

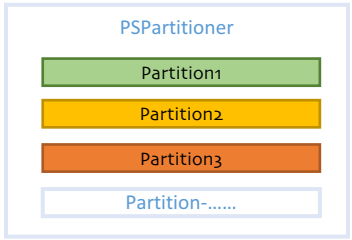
丰富的机器学习及数学计算库

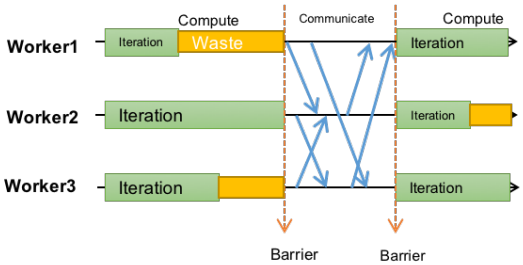
- 集成LR（ADMM-LR），SVM，KMeans，LDA，MF，GBDT等机器学习算法
- 多种优化方法，包括ADMM，OWLQN， LBFGS和GD
- 支持多种损失函数、评估指标，包含L1、L2正则项算法

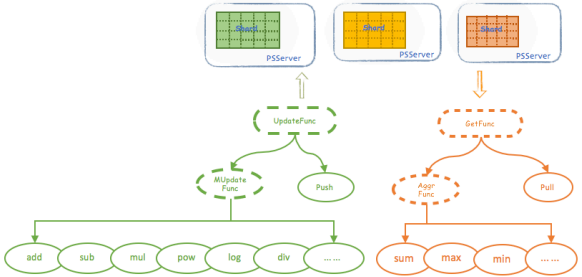
- 基于PSModel的机器学习友好接口，以Model为核心编程
- 支持Spark on Angel，Spark代码小量改动就可以运行Angel之上
- 灵活的psFunc，便于复杂算法的开发，实现模型并行

友好的用户编程接口

Angel和Glint的比较







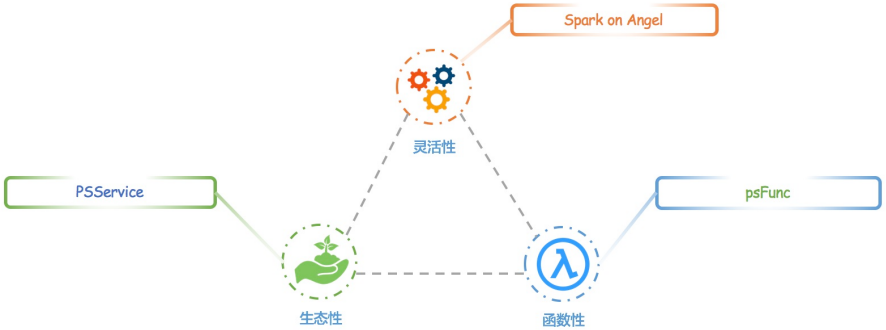
更丰富的模型切分

更灵活的异步模式

更强大的psFunc

Angel的定位

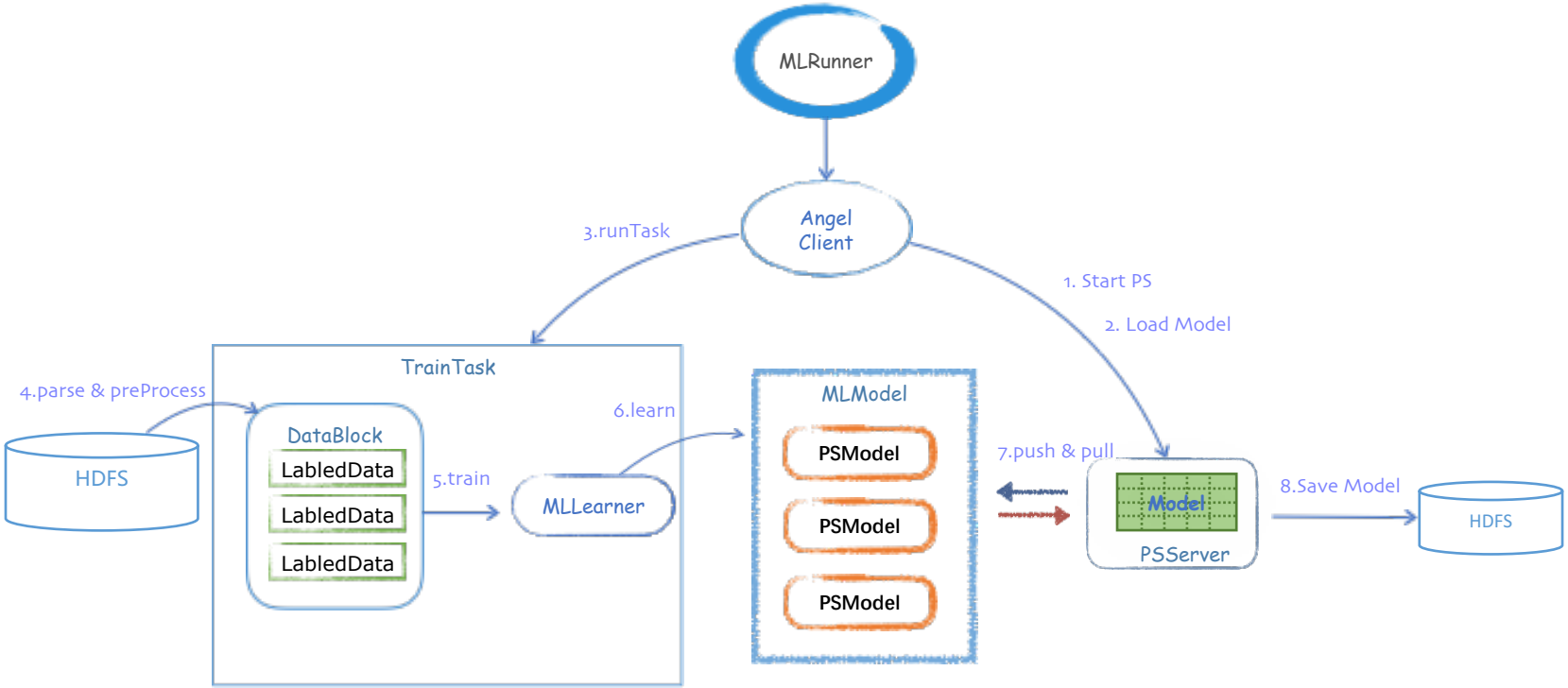
一个源于Parameter Server理念的高性能分布式机器学习平台



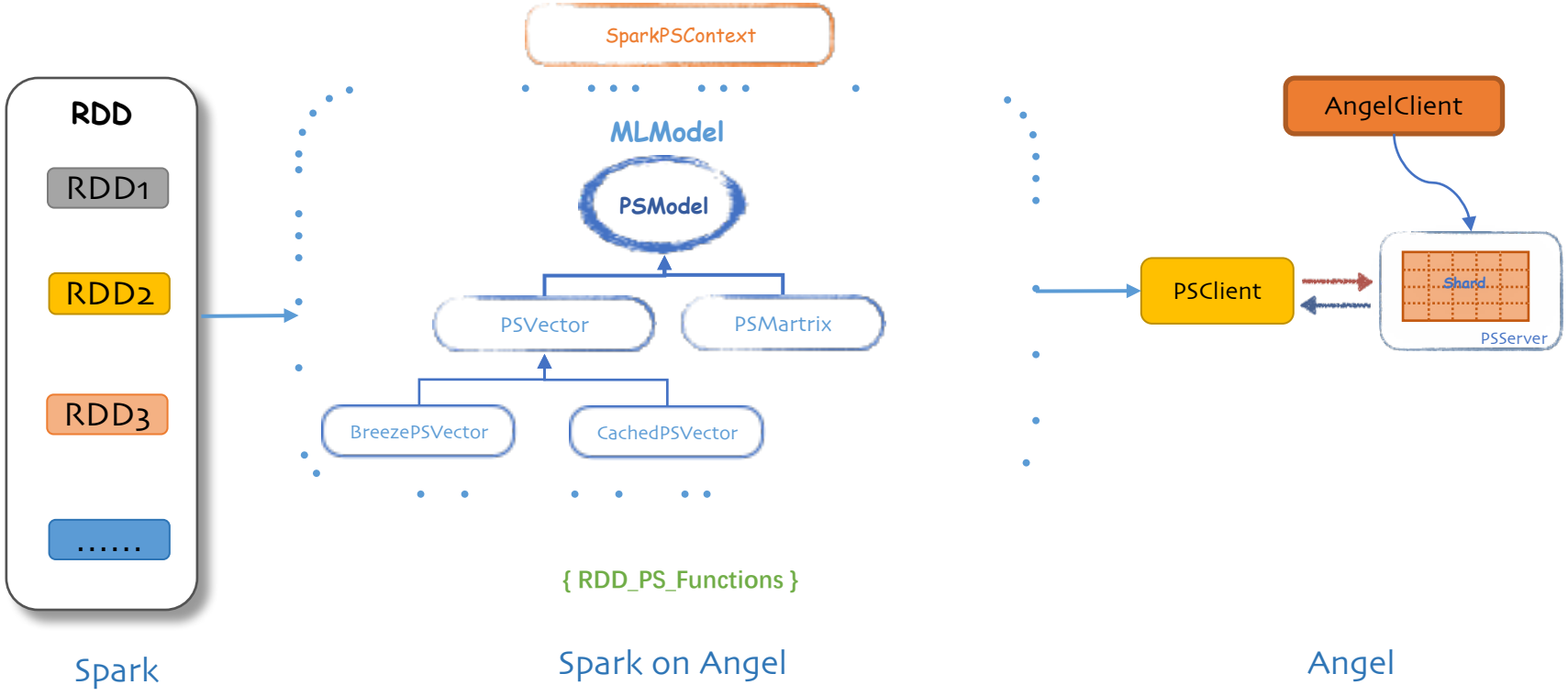
https://github.com/tencent/angel

Spark on Angel的开发

Angel的API设计



Spark on Angel的API设计

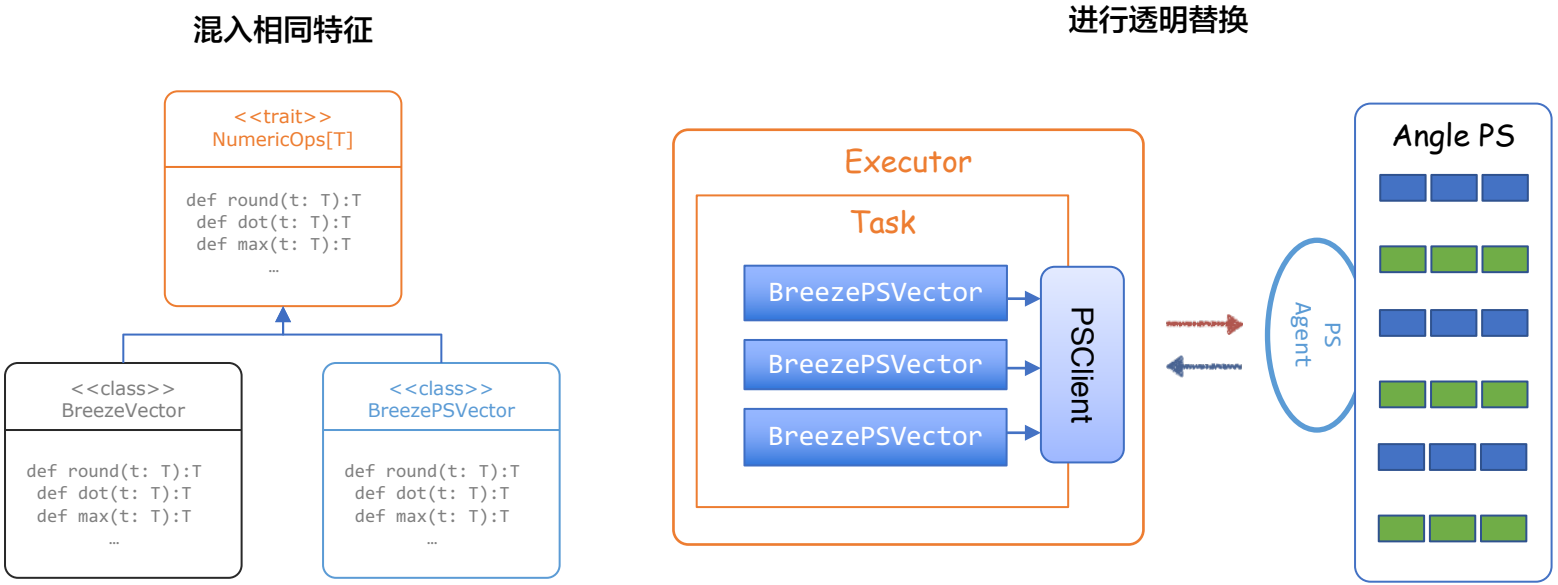


Spark on Angel的基础写法

```
val psContext = PSContext.getOrCreate(spark.sparkContext)
val psVector = PSVector.dense(0.0)
rdd.map { case (label , feature) =>
      psVector.increment(feature)
      ...
}
println("Feature sum size:" + psVector.dimension)
```

- 启动SparkSession
  - 初始化PSContext，启动Angel的PSServer
  - 通过PSContext, 创建PSVector
  - 在RDD的运算中，直接调用PSVector，进行模型更新
  - 终止PSContext
- 停止SparkSession

Vector的透明替换



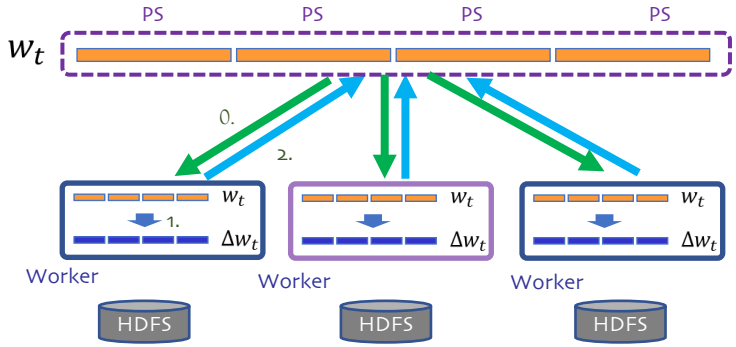
- 将BreezeVector透明替换为BreezePSVector
- 适用于MLLib大部分算法
- 替代成本非常低

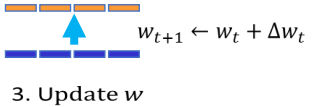
Spark on Angel的算法

Angel的算法

Spark on Angel Available

LR on Angel



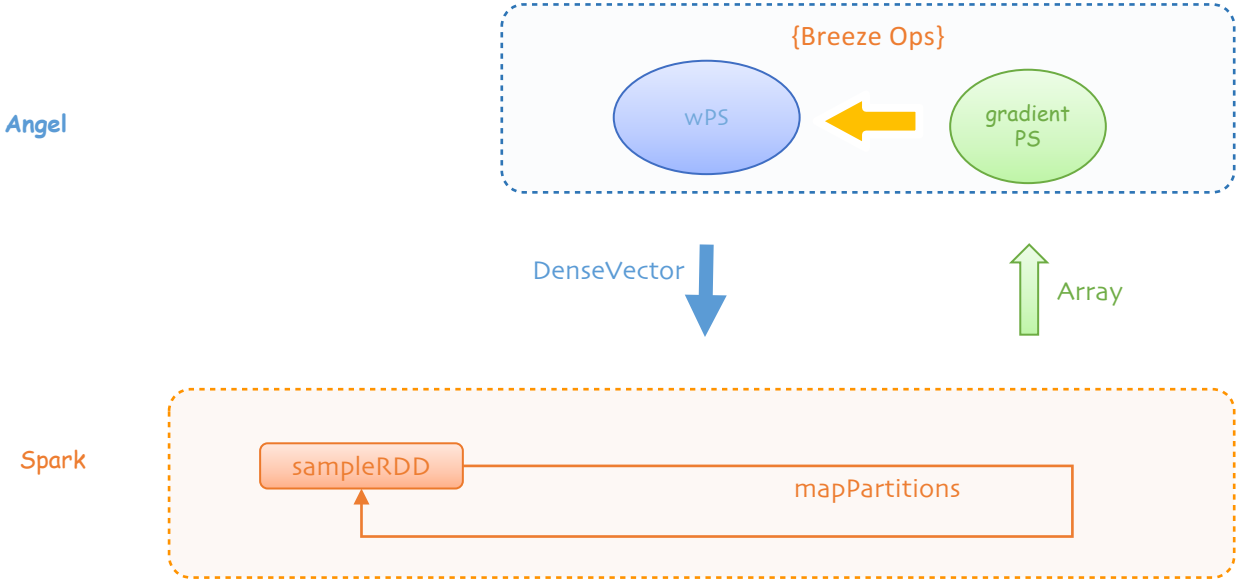


Pull parameters from PS

Push update value to PS

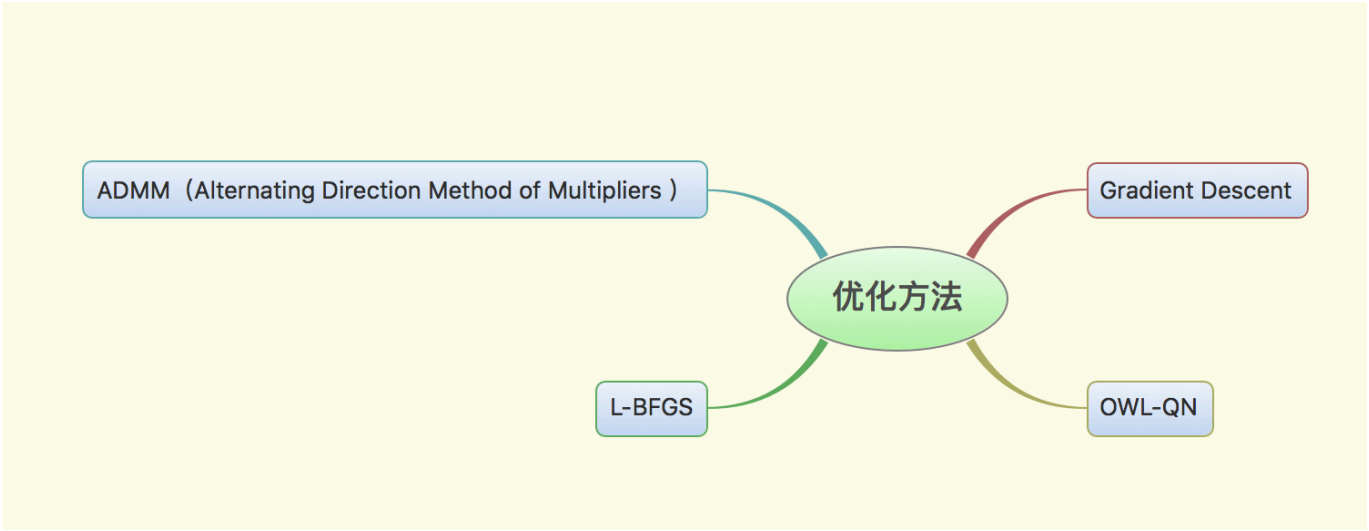
Step 0: Worker从PS获得参数W_t
Step 1: Worker计算参数的更新值ΔW_t
Step 2: Worker把ΔW_t推送给PS
Step 3: PS更新参数（$W_{t+1} \leftarrow W_t + \Delta W_t$）

[Spark on Angel] LR

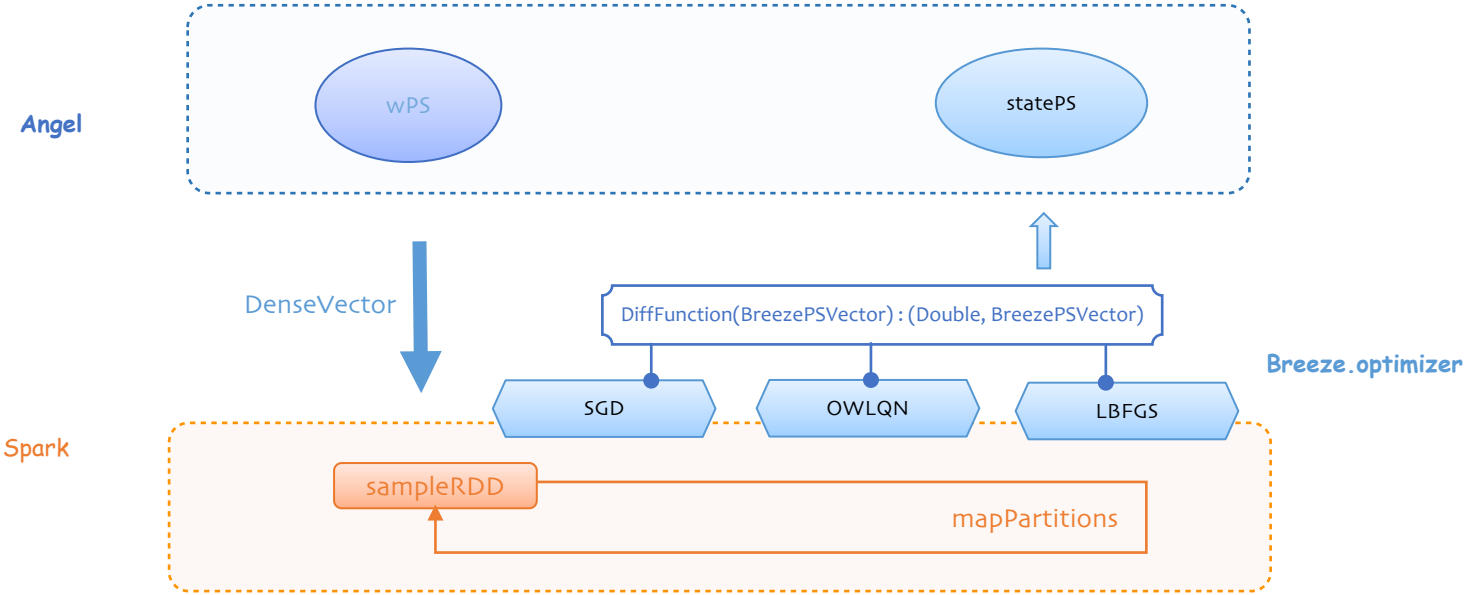


[spark_on_angel_quick_start.md]

优化方法

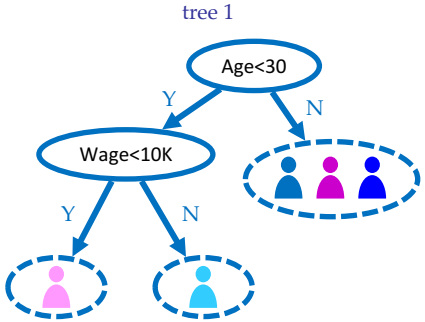

ADMM（Alternating Direction Method of Multipliers）
Gradient Descent
优化方法
L-BFGS
OWL-QN

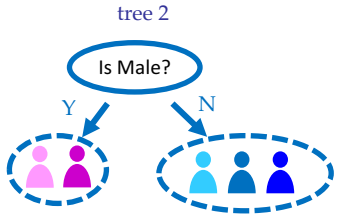
[Spark on Angel] LR with Optimizer



[spark_on_angel_optimizer.md]

GBDT：树模型+Boosting





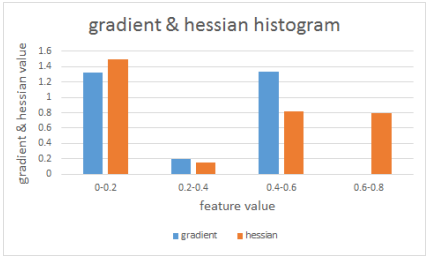
A predict( ) 5+0.5=5.5

B predict( ) 10+1.5=11.5

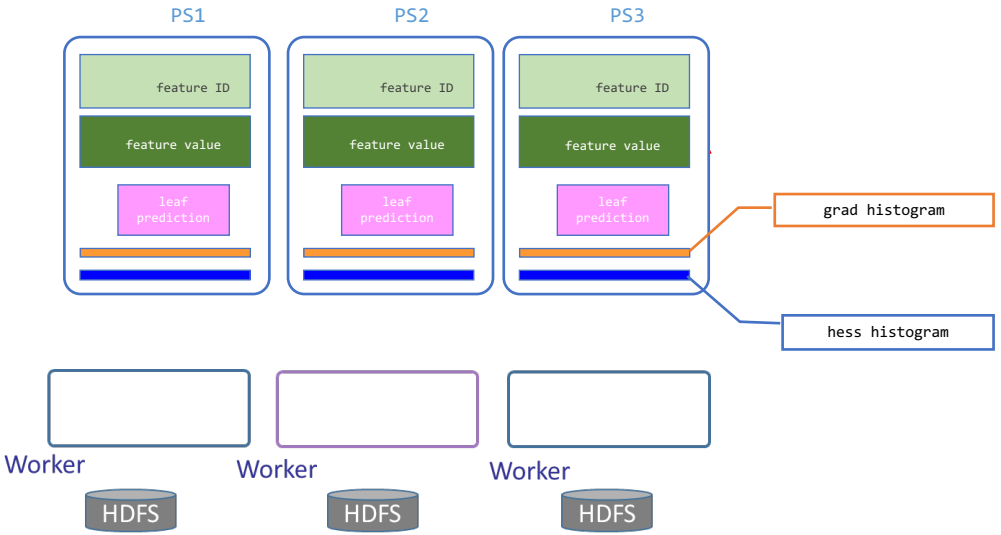
C predict( ) 1+1.5=2.5

D predict( ) 1+0.5=1.5

E predict( ) 1+1.5=2.5

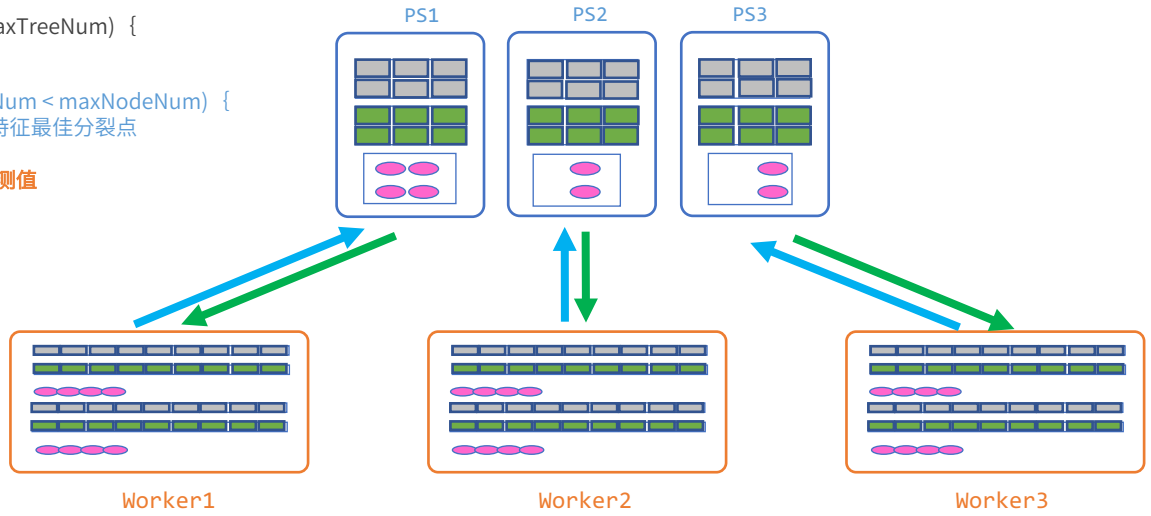


GBDT on Angel: 模型存储



GBDT on Angel（1）：构建森林

```
while (treeNum < maxTreeNum) {
    创建一棵新树
        while (nodeNum < maxNodeNum) {
                寻找特征最佳分裂点
        }
    计算叶子节点的预测值
    完成一棵决策树
}
```

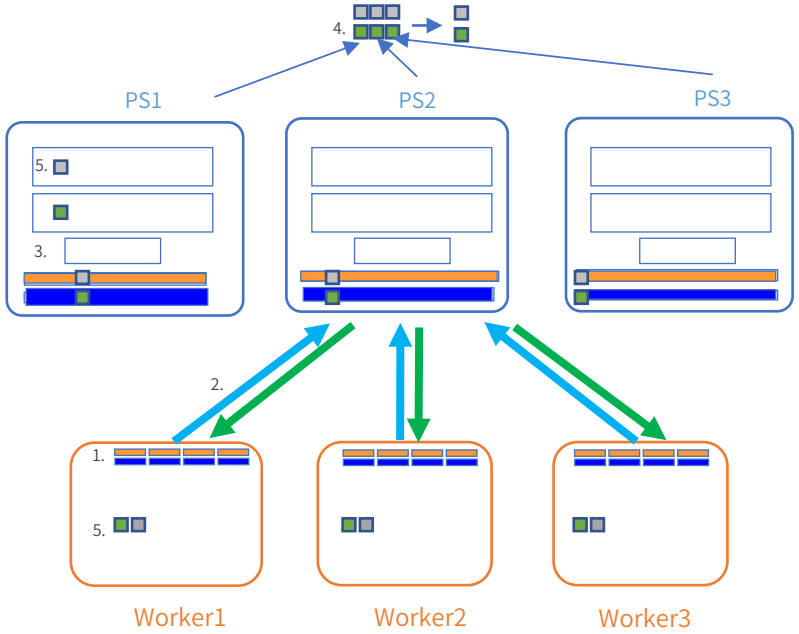


GBDT on Angel（2）：分裂树节点

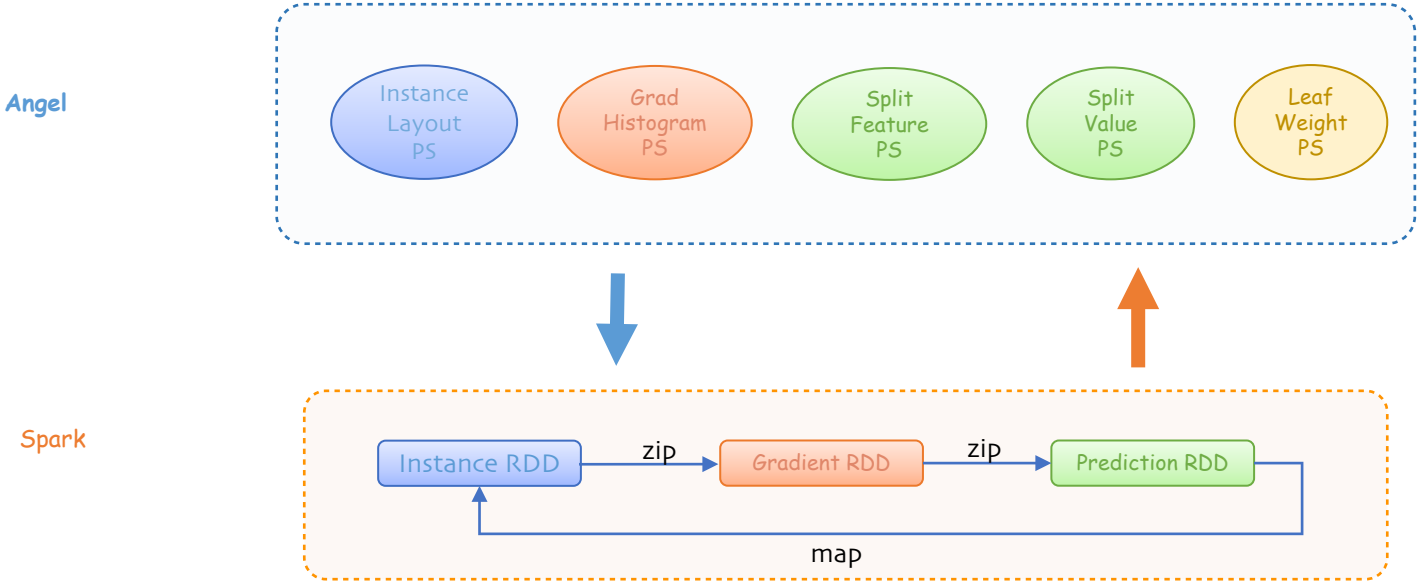
find split feature & value

1. worker计算梯度直方图（一阶&二阶）
2. worker推送梯度直方图到PS
3.每个 PS计算局部最佳分裂点
4.PS之间计算出全局最佳分裂点
5.创建分裂点，Worker从Ps拉取最佳分裂点

[gbdt_on_angel.md]



[Spark on Angel] GBDT



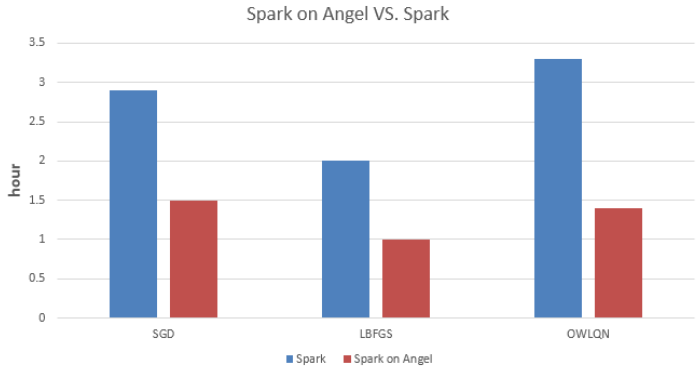
[spark_on_angel_gbdt.md]

性能比对

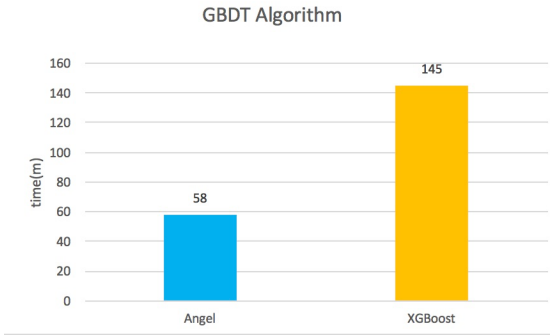
——生产数据，现网环境，尽量公平

（Spark on Angel） vs Spark —— LR

	Spark	Spark on Angel	加速比例
SGD LR (stepSize=0.05,maxIter=100)	2.9 hour	1.5 hour	48.3%
L-BFGS LR (m=10, maxIter=50)	2 hour	1 hour	50.0%
OWL-QN LR (m=10, maxIter=50)	3.3 hour	1.4 hour	57.6%



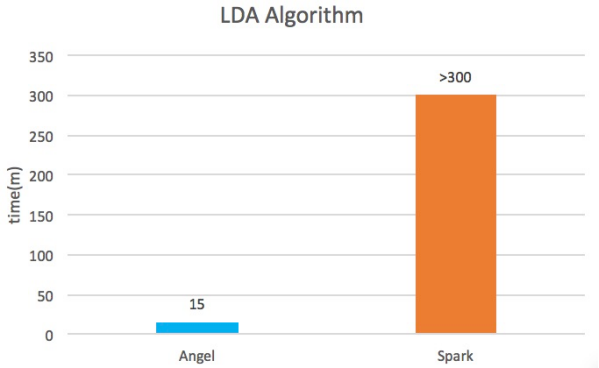
Angel vs XGBoost —— GBDT



框架	Worker	PS	建立20棵树时间
Angel	50 个(内存：10G / Worker)	10个 (内存：10G / PS)	58 min
XGBoost	50个 (内存：10G / Worker)	N/A	2h 25 min

数据：腾讯内部某性别预测数据集，3.3×10^5 特征，1.2×10^8 样本

Angel vs Spark —— LDA



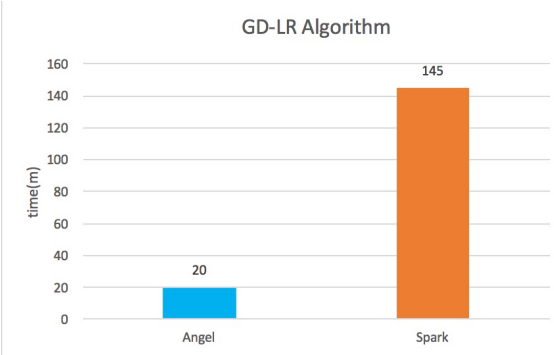
框架	Worker	PS	时间
Angel	20个(内存：8G/Worker)	20个(内存：4G/PS)	15min
Spark	20个(内存：20G/Worker)	N/A	>300min

数据：PubMED

框架	Worker	PS	时间
Angel	50个(内存：10G/Worker)	50个(内存：4G/PS)	1h 7min

DataSet：40G Token： 2 billion
Word：52w Topic： 1000

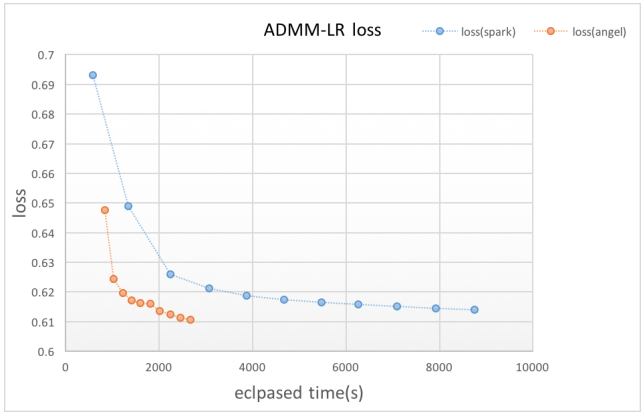
Angel vs Spark —— GD-LR



框架	Worker	PS	迭代100次时间
Angel	50个(内存:10G/Worker)	20个(内存: 5G/PS)	20min
Spark	50个(内存:14G/Worker)	N/A	145min

数据：腾讯内部某推荐数据，5×10^7 特征，8×10^7 样本

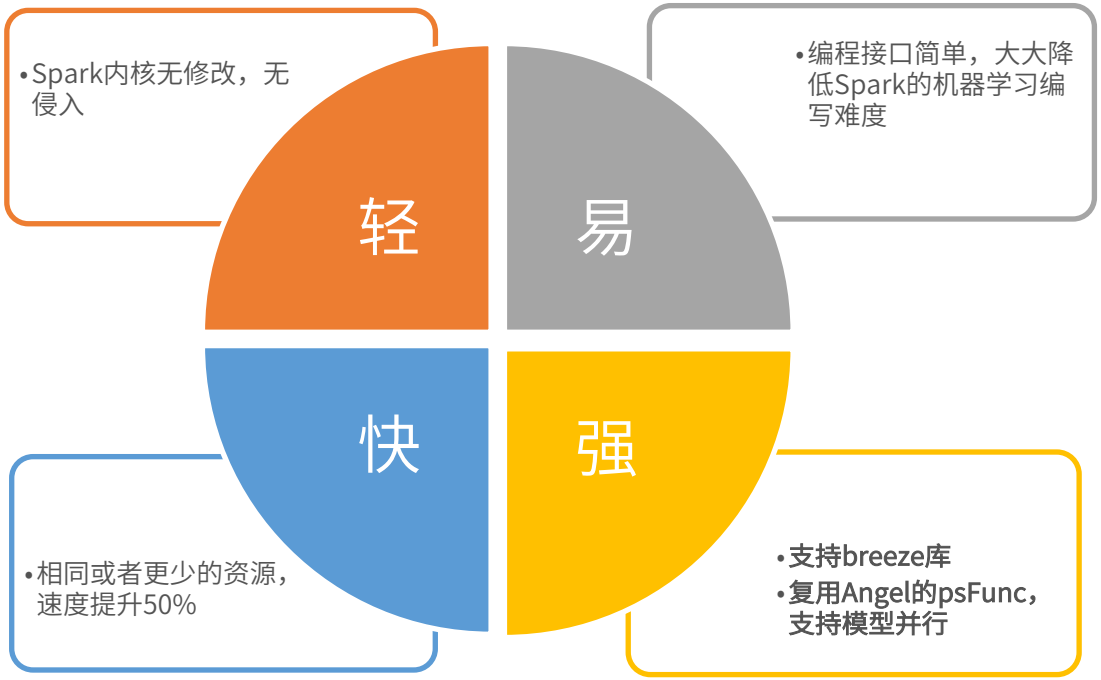
Angel vs Spark —— ADMM-LR



框架	Worker	PS	收敛退出
Angel	100个(内存:10G/Worker)	50个(内存：5G/PS)	27 min
Spark	200个(内存:20G/Worker)	N/A	145 min

数据：腾讯内部某推荐数据，5千万特征，1亿样本

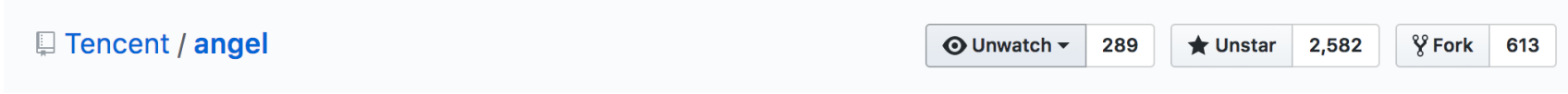
Spark on Angel的特点



开源和展望

OpenSource & Perspective

Angel开源



github:issues

(PR 60)

- LightBGM作者：[GBDT] The purposes of using parameter server in GBDT #7
- 海外华人：English translation of documents #95
- 华为工程师：[WIP]Upgrade the netty version of RPC to 4.x #94
- 新浪微博：增强LR算法，加入y截距因子
- ……

学术创新

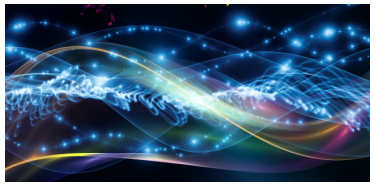
- **国际顶级会议Paper（CCF A类）**

  - LDA*: A Robust and Large-scale Topic Modeling System VLDB, 2017
  - Heterogeneity-aware Distributed Parameter Servers. SIGMOD, 2017
  - Angel: a new large-scale machine learning system. National Science Review (NSR), 2017
  - TencentBoost: A Gradient Boosting Tree System with Parameter Server. ICDE, 2017
  - ……

版本展望（What is Next）

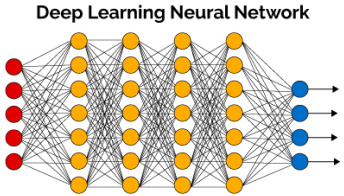
Python API

V1.3



Spark Streaming on Angel

V1.5



Deep Learning Framework Support

V2.0

Q & A

微博：@明风

喜欢记得给个Star噢

andymhuang@tenent.com

机器学习系统 & 算法工程师

We are Hiring