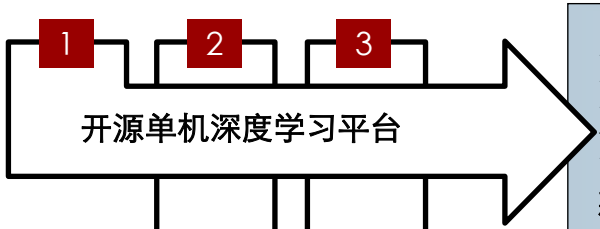
分布式数据分析与人工智能平台及算法实践

涂丹丹 金鑫
tudandan@huawei.com
jinxin11@huawei.com

业界分析：深度学习平台演进趋势



◆Caffe: 模块化，支持单机多卡，不支持异构平台，主要用于图像分类、识别。
◆Torch: 可读性好，其中网络的基本组成为module，方便实现任意网络的构建
◆Theano: 实现符号微分，DAG表达深度网络、符号计算的开拓者，效率较低，主要用于NLP等任务，不适合大规模的深度学习模型训练。
◆百度PADDLE：基于MPI实现分布式深度学习，已经开源了OCR、人脸识别API
主要是学术界深度学习科研人员开发，用于新算法探索研究。大部分还不支持模型并行机制、不支持异构平台。

企业闭源深度学习平台

◆腾讯Marina:目前公开的是单机多卡的平台，主要用于语音识别，采用线性拓扑结构通讯，支持模型并行和数据并行机制，还没开放API。
◆Google DistBelief:支持数据并行模型并行机制，支持上千节点，主要用于Google内部Gmail，搜索等业务。
闭源不利于很好使用生态圈，Google策略有闭源、开源，或二者结合。

开源分布式深度学习平台

◆MXNet: 混合符号编程和命令式编程，支持数据并行，不支持模型并行；
◆CNTK: 支持数据并行和模型并行，对SGD算法做了优化，并行机制并未开源；
◆TensorFlow: 支持数据并行、模型并行，开源的分布式能力逐步增强
◆Spark+Caffe: 在深度学习使用Caffe，Spark在大规模参数计算没有优势；
◆Poseidon：结合Petuum SSP能力，但深度学习方面仍依赖于Caffe。
主要是Google,微软等大公司推动，目前业界开源较多，在平台并行机制、集群管理、调度、支持算法丰富度等方面还很不成熟。

深度学习平台的演进趋势：从单机到多机，计算能力提升；平台化，可视化，易用性提升；开源+闭研结合，依赖生态圈，构建关键竞争力。

业界分析：大数据分析与人工智能算法技术发展趋势

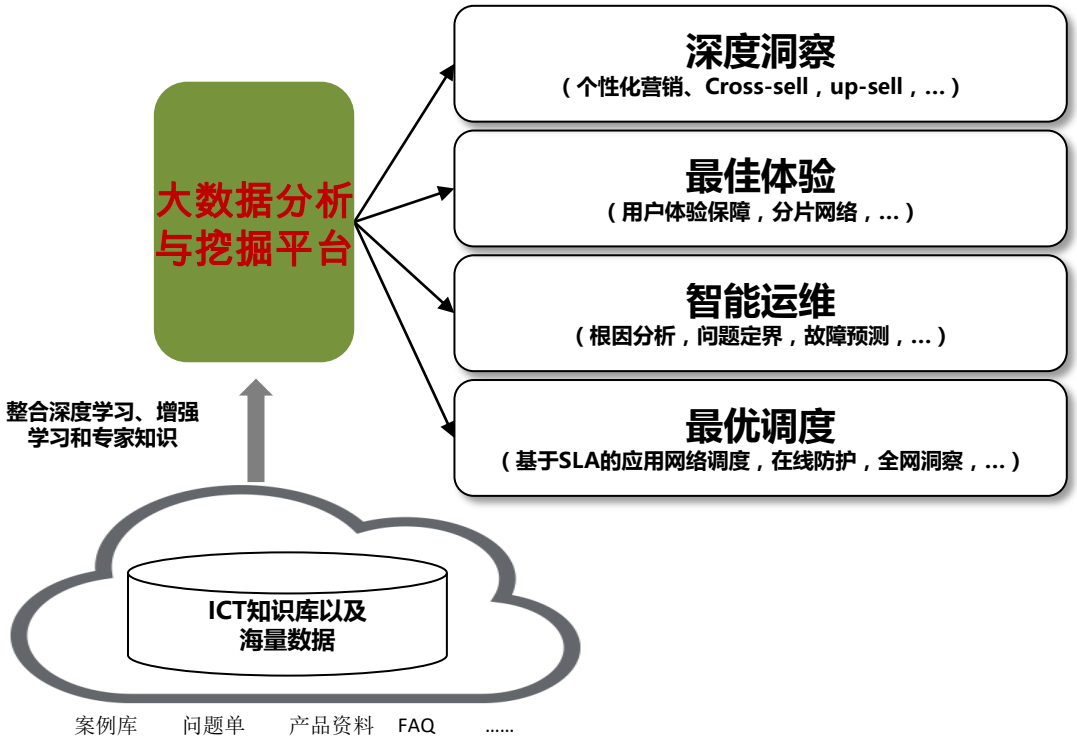
中小数据时代

传统的机器学习研究并不把海量数据作为处理对象，很多**算法是为处理中小规模数据设计**的，直接把这些算法**用于海量数据，效果可能很差**，甚至**可能用不起来**。

大数据时代

新的数据种类不断涌现，对大数据集、高维数据的学习，算法关注点转移到**分布式可扩展、有效利用非标记数据解决训练数据质量问题（半监督学习）、提高学习结果泛化能力（集成学习）、不同领域进行知识迁移（迁移学习）、特征自动学习（深度学习）**等

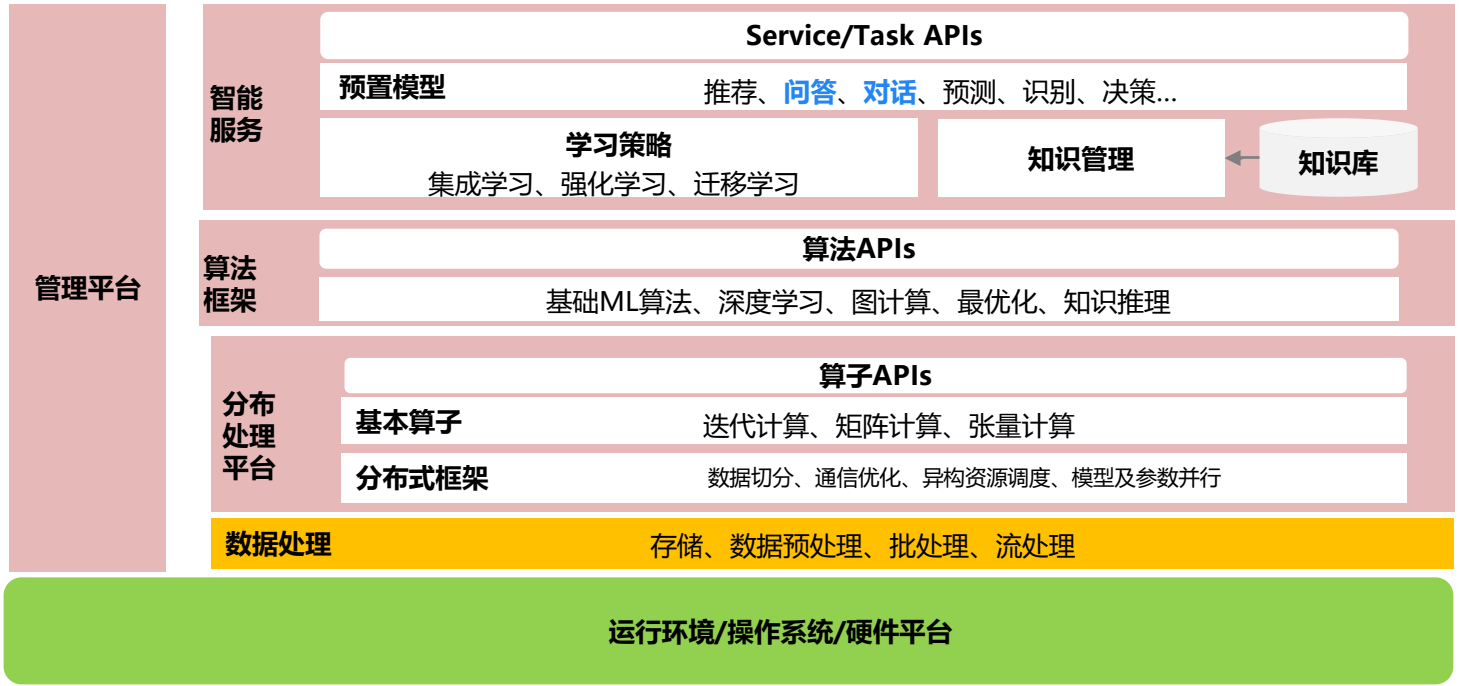
分布式数据分析与人工智能平台使能ICT基础设施智能化



Key Message：

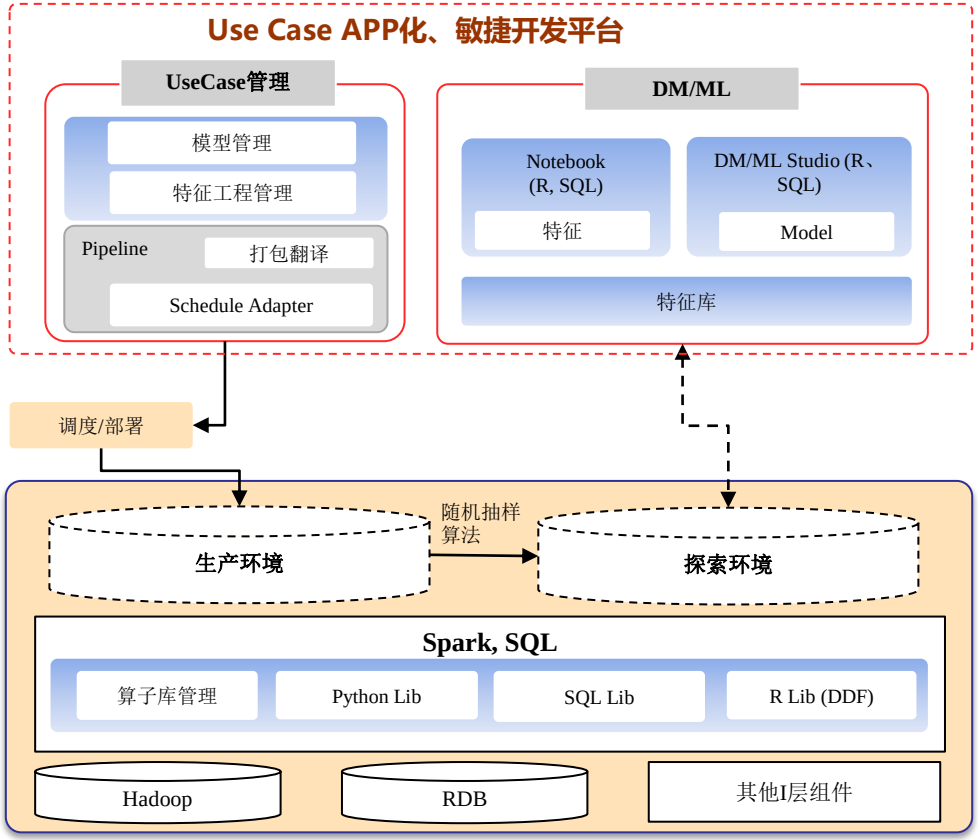
1. 自学习反馈，持续个性化成长
2. 多个深度网络联合推理及根因分析
3. 大规模网络状态分析及最优决策
4. 360度分析，提供个性化业务定制
5. 专家融合知识库

分布式数据分析与人工智能平台



面向ICT领域构建高效分布式大数据分析与人工智能平台，支撑电信、IT、金融、大视频等场景

离线大数据分析平台



解决UseCase 大规模发布与部署问题

- 支持**UseCase 的快速开发**。

提高UseCase开发与定制效率

- **统一的特征集合管理**(基于UseCase沉淀)，一次开发多UseCase共用，减少模型前期数据处理的开发工作量。
- **交互式的数据分析能力**，**多语言交互式探索能力**，提高UseCase数据分析的效率。

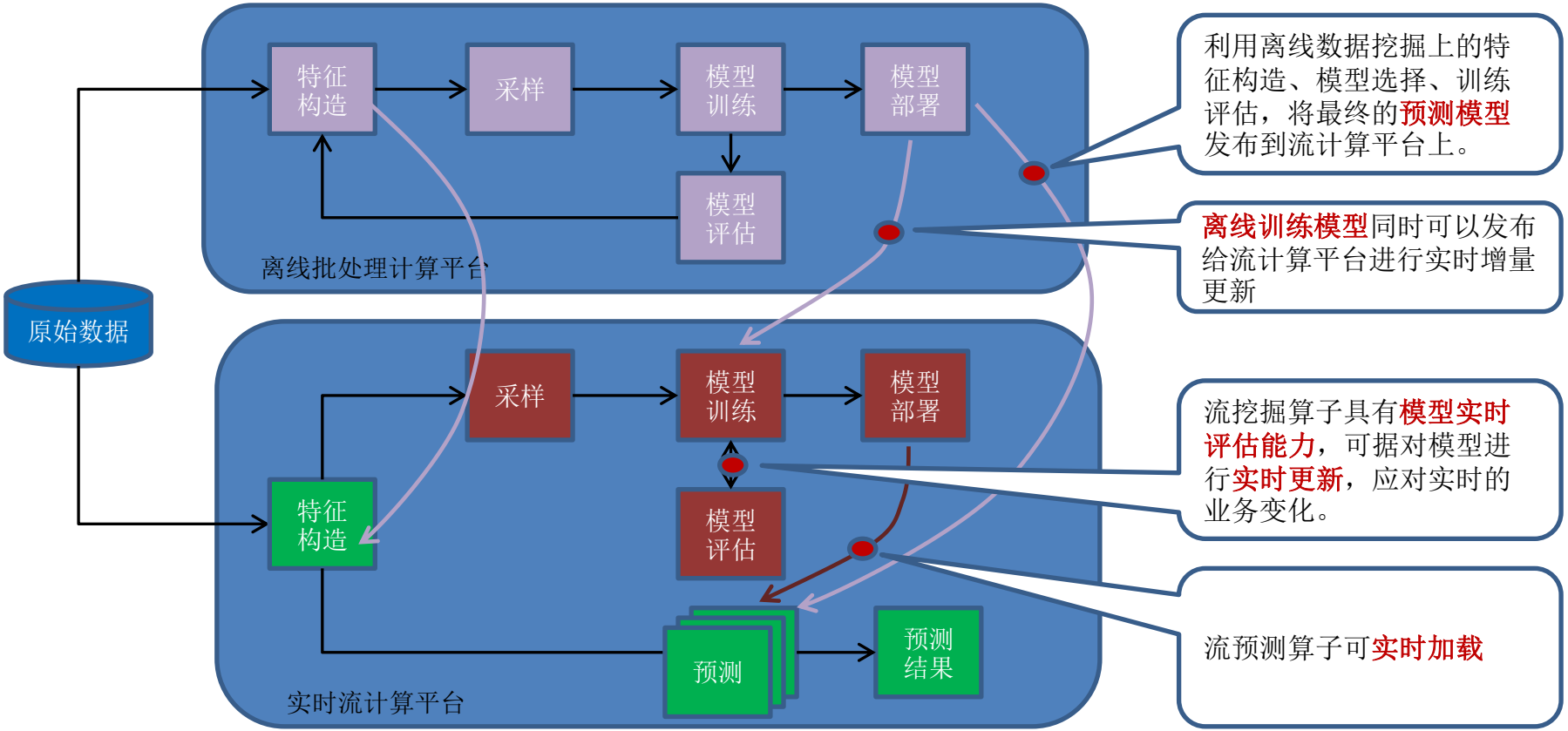
插件式的UseCase管理按需发布

- UseCase开发完成后，通过**Pipeline打包、发布**。
- UseCase Package发布后，**现网快速的导入安装即可运行**，整个安装过程实现自动化流程管理。

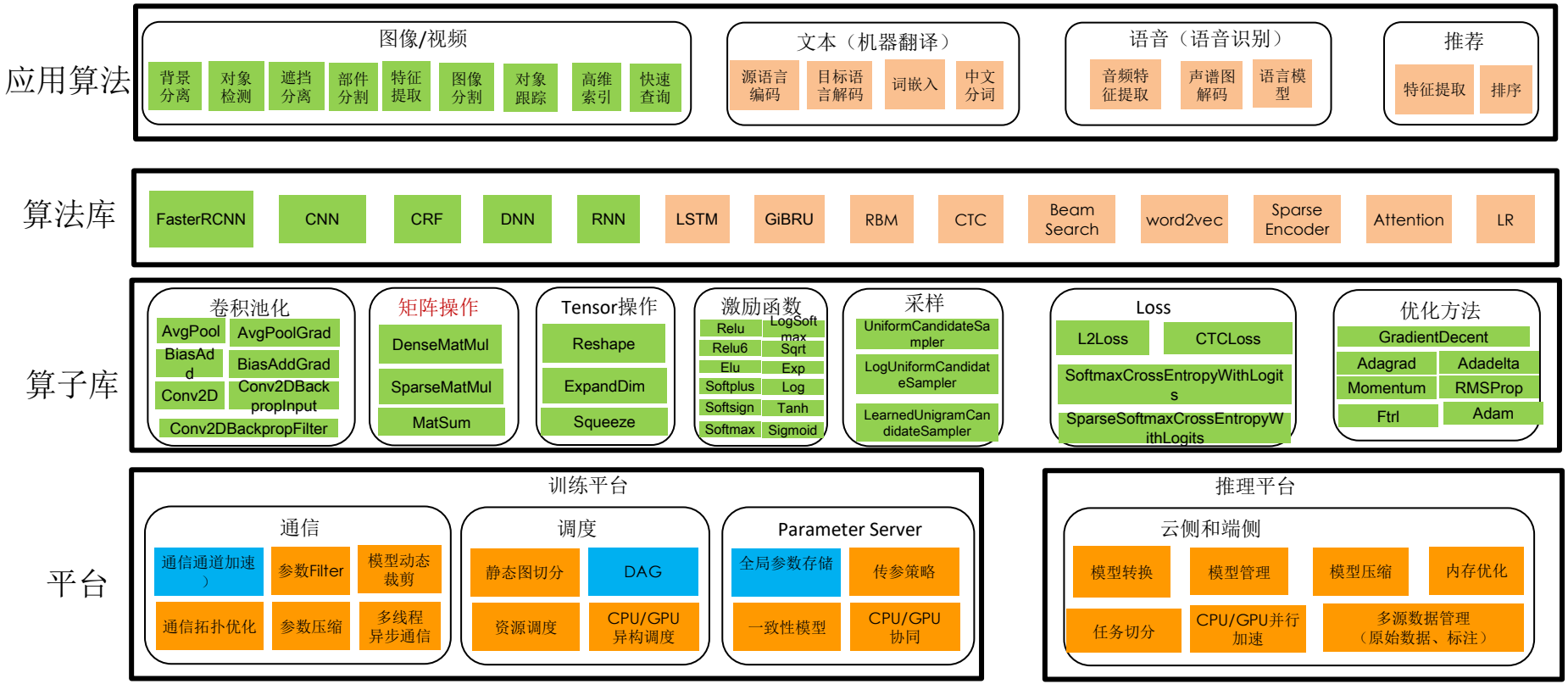
组件式的业务/数学算法沉淀提高模型重用率

- **统一的统计学/机器学习/数据挖掘算法的管理与开放式算法注入**，提高建模的效率。
- **业务模型算法的沉淀**，实现模型的代码重用性，减少重复开发。

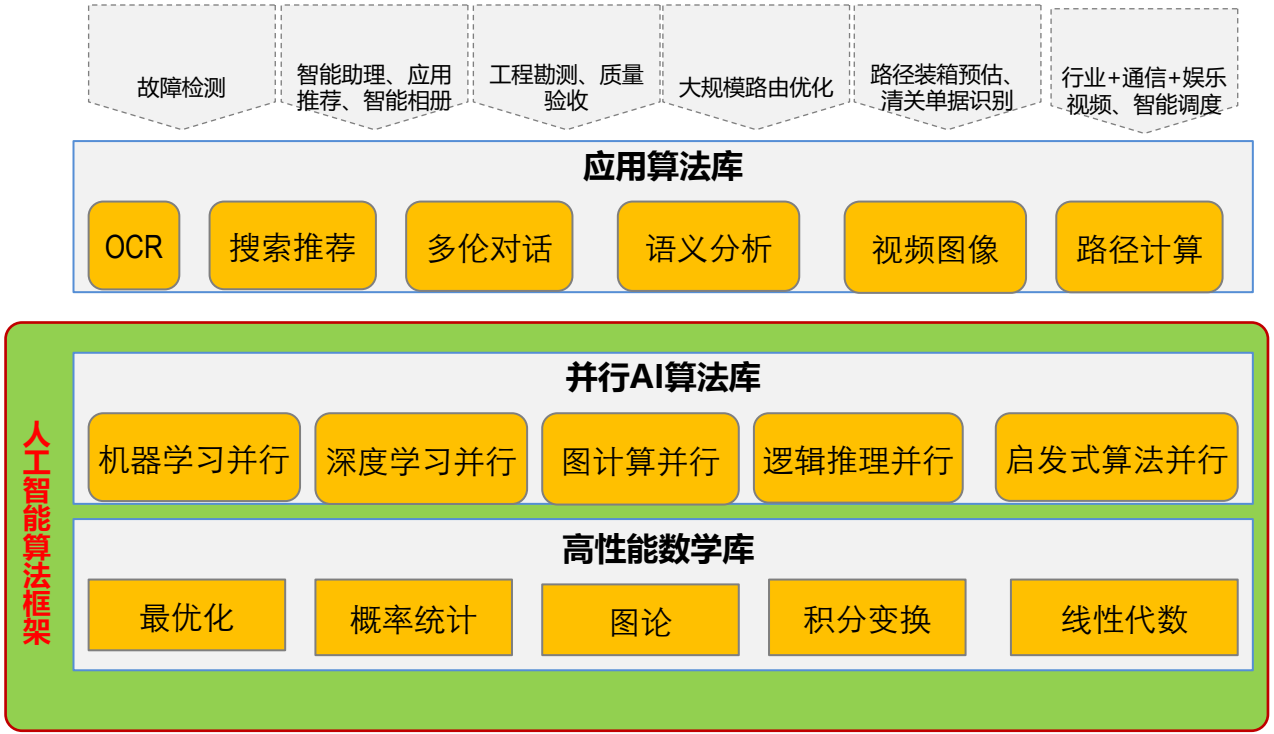
分布式实时流挖掘平台


特征构造
采样
模型训练
模型部署
模型评估
离线批处理计算平台
原始数据
采样
模型训练
模型部署
特征构造
模型评估
预测
预测结果
实时流计算平台
利用离线数据挖掘上的特征构造、模型选择、训练评估，将最终的预测模型发布到流计算平台上。
离线训练模型同时可以发布给流计算平台进行实时增量更新
流挖掘算子具有模型实时评估能力，可据对模型进行实时更新，应对实时的业务变化。
流预测算子可实时加载

分布式深度学习平台架构



分布式并行数据分析和人工智能算法框架：高性能、低功耗、弹性可扩展



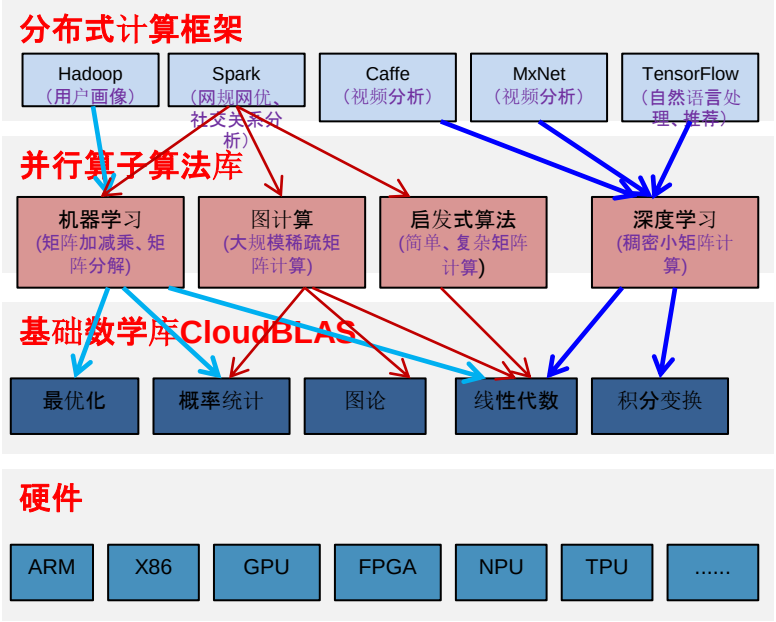
业界现状：

1. ML算法已成熟，随平台硬件演进持续并行加速
2. DL算法理论待突破，训练和预测性能待进一步提升
3. 图计算、逻辑推理、启发式算法待分布式并行加速支撑大规模数据

构建端云协同并行AI算法库和高性能数学库：向下和硬件结合提升底层基础数学库并行性能、降低功耗；向上和智能应用结合提升复杂场景AI算法性能、自适应能力。

高性能分布式并行数学库：面向大数据分析与人工智能场景，实现最佳性能

BLAS是核心控制层，向上对接语言和支撑应用，向下对接硬件



基础数学库的挑战：

1.并行化
2.低功耗
3. 自适应
4.弹性扩展

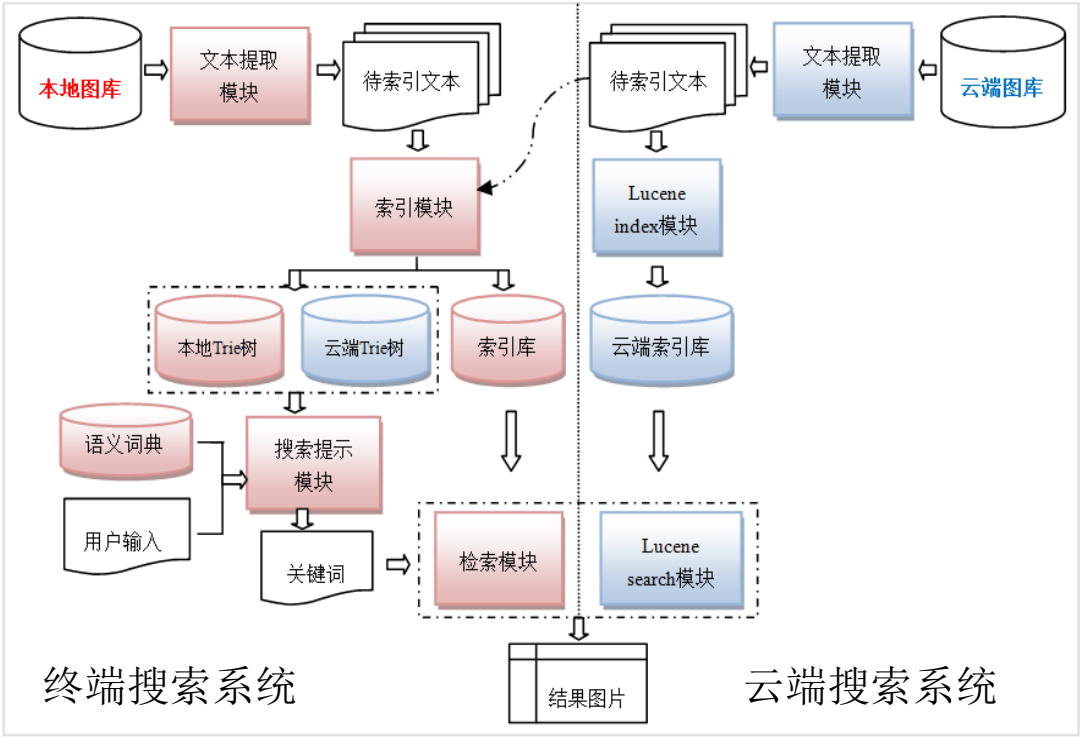
以矩阵计算为例:

1. 大规模稀疏矩阵计算（矩阵切分、分布式计算和通信协同）
2. 海量稠密张量计算（多核并行、指令级并行）
3.复杂矩阵计算分布式（计算强依赖，节点间存在大量的通信开销）

构建高性能CloudBLAS数学库，兼容业界通用芯片，端侧在并行化低功耗自适应上发力，云侧在自适应弹性扩展分布式上发力

大数据分析及人工智能算法分布式并行化挑战

以tag搜图系统和算法：华为手机相册搜索



搜索系统由文本提取、索引、搜索提示以及检索 等模块组成

➢搜索终端图库时将利用本地Trie树实现搜索提示并给出搜索关键词，调用本地检索模块从本地索引库中搜索图片

➢ 搜索云端图库时将利用保存在终端的云端Trie树实现搜索提示并给出搜索关键词，发送到云端调用Lucene search模块从云端索引库中搜索图片

端云协同图片搜索算法，端侧降低搜索算法内存使用和功耗，云端提升搜索算法索引更新效率

基于Spark对Apriori和FPGrowth算法分布式并行实现大数据频繁项挖掘

大数据频繁项挖掘的背景与挑战

•公安同行车辆检测，计算1年连续通过多个(>1)卡口的车辆(百亿级)，帮助预判可能的集体策划行为或犯罪行为。同时常见应用包括商场购物篮分析、Web应用关联度计算等。

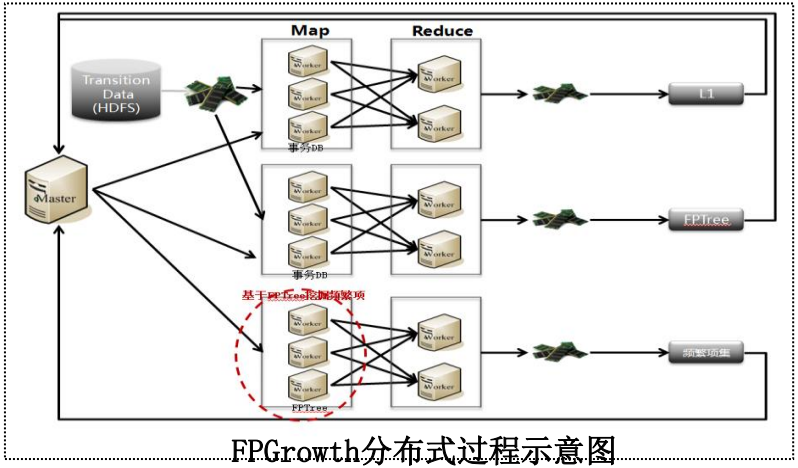
•挖掘频繁项集需爆炸式组合计算，当事务DB(GB)越大、项越多(万以上)、步长(项集长度)越大(>5)，会生成大量的数据集和需要很长的计算时间，对节点的内存开销和运算效率都望待改进与提升。。

关键技术

•Apriori分布式化：Master将事务DB均分到多个Worker节点，各Worker节点多次扫描部分DB、多次MR计算频繁项，最终汇总到Master节点。

•FPGrowth分布式化：阶段1，Master将事务DB均分到多个Worker节点，两次扫描事务DB生成只含频繁项的FPTree；阶段2，将FPTree分散到多个Worker节点进行一次MR计算频繁项，最终汇总到Master节点。

• 二者都使用先验假设原理，并基于Spark框架采用MR技术对算法进行分布式化。 但FPG基于FPTree，扫描DB、MR次数少，因此在数据搜索、节点存储和网络开销上更少，性能更优。



FPGrowth分布式过程示意图

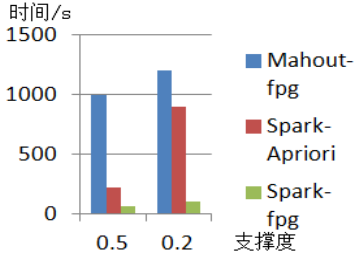
测试结果

测试数据：Web关联分析webdocs，1692082条，5267656items，1.6G。

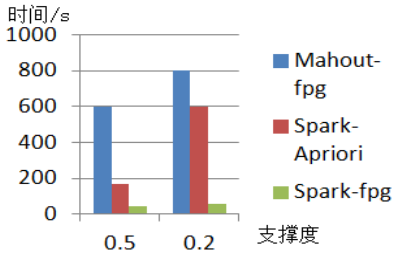
测试环境： 硬件 SUSE3、150G内存、32CPU，软件 Spark1.0、Hadoop2.4、JDK1.7、scala2.1；分别在单机、双机集群：1M+2W测试Spark-Apriori、Spark-fpg、Mahout-fpg的性能。

•分布式Spark-fpg性能提升10+倍。

•增加计算节点，计算时间趋于减半。

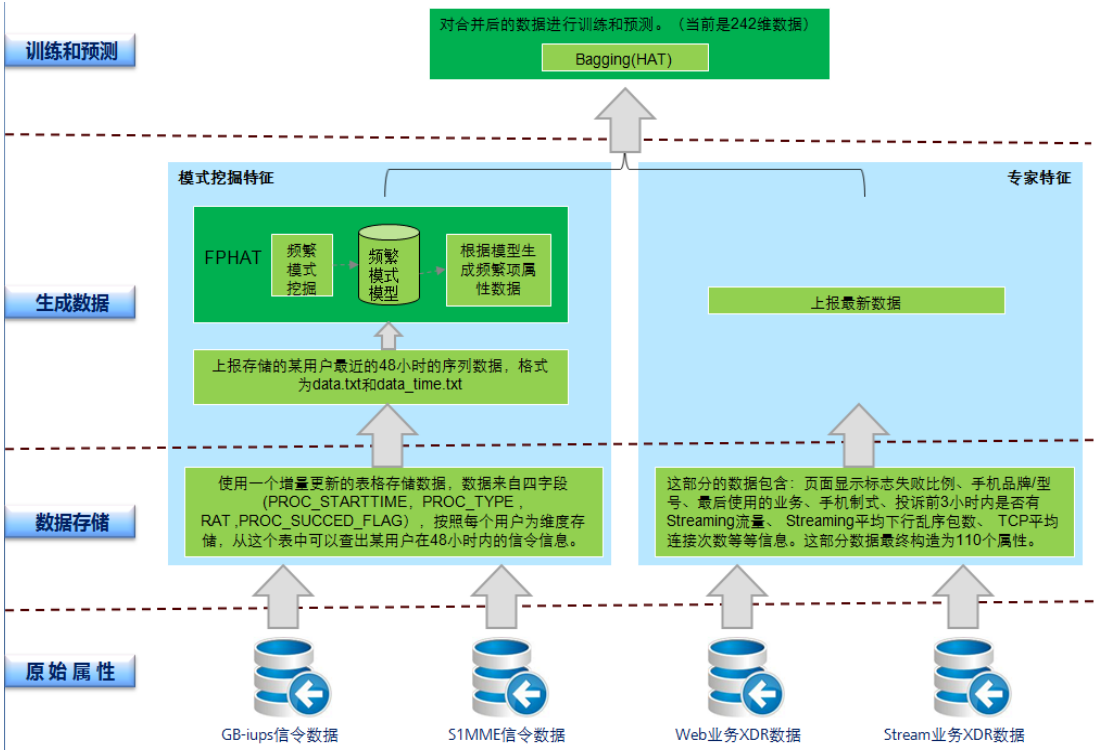


单机环境



双机环境

基于实时流挖掘进行潜在投诉用户预测



利用实时流挖掘算法进行用户投诉预测

- 基于序列频繁模式挖掘算法自动构造特征，利用流挖掘算法实时预测潜在投诉用户

1. 基于CloSpan和PrefixSpan算法挖掘频繁项，利用FPHAT流挖掘算法基于序列数据构造132维特征，结合人工经验构造的110维特征，总共242维特征对投诉和非投诉用户进行刻画。
2. 利用adaptive bagging组合分类算法建模（包括投诉用户和非投诉用户），模型可增量更新
3. 利用建立好的模型，对新的数据进行预测（预测内容为此用户是否为潜在投诉用户）

DBSCAN：提升算法可扩展性/准确率

背景

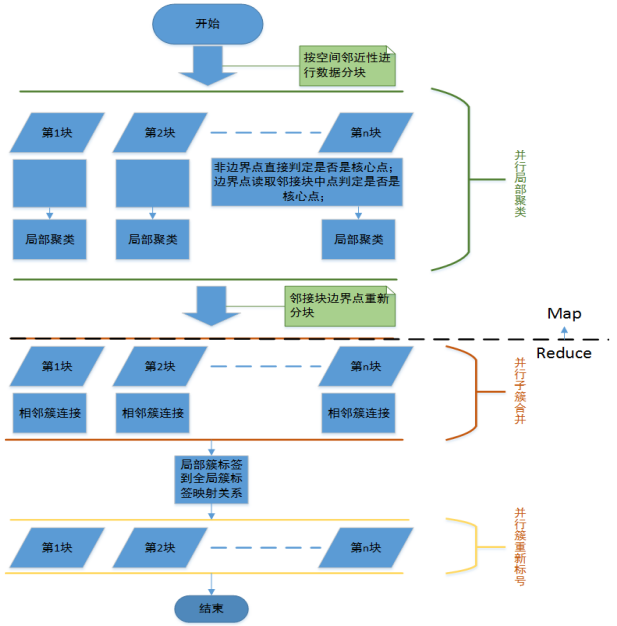
- 目前开源社区MLlib K-means聚类算法需要指定聚类簇个数，且只能发现“类圆形簇”
- DBSCAN是基于密度的聚类方法，不需要指定聚类簇个数，可发现任意形状的簇。可应用于家/工作地聚类、路径分析等场景

挑战

- DBSCAN对于集合中任意一点p计算eps邻域，需要计算数据集中每个点到p的距离，DBSCAN时间复杂度是$O(n^2)$，随数据量增长，串行算法运行时间会迅速增长。**如何对数据按空间切分及加快临界点的计算是分布式并行最大的挑战**

分布式并行思路

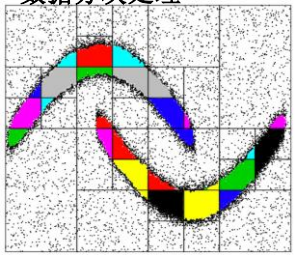
- **数据分块**：对数据按空间区域分层划分(每次选择取值范围最大的维度进行划分),控制数据块最短边(>2eps)、点的数量(>5000)、最大分层(<10),**均衡各节点计算任务**
- **Map**：对每个数据块中的非边界点判定是否是核心点，读取数据块近邻块中所有点判定边界点是否是核心点，**对距离小于eps的密度可达点进行局部聚类**
- **Reduce**：一个簇跨越多个数据块，合并结果：边界点的近邻可能位于其他块中，对邻接块边界点重新分块，如果两个点属于两个相邻的块，并且它们的距离小于eps，并且其中至少一个点是核心点，那么这两个**簇合并**



集群：1Master+3Slave（Memory：20G, CPU：6cores）

测试数据集：2维数据集

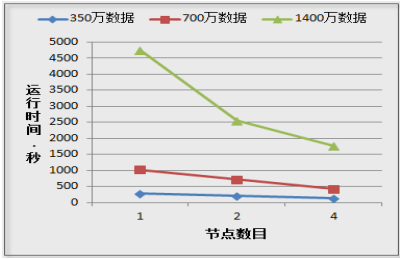
数据分块处理



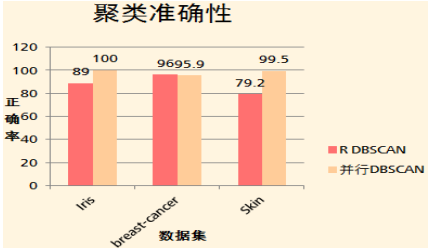
R串行，Spark两台机器

数据量	R （秒）	Spark并行（秒）	加速倍数
1万	11.7	10	1.17
5万	359	17.7	20.28
10万	2004	31.1	64.44
20万	5604	77	72.78
40万	18364	232.9	78.85

D-DBSCA线性加速比



DBSCAN准确率



测试结果：3个benchmark数据集上分布式DBSCAN准确率比R中DBSCAN算法略高，2台机器集群上在40万个2维数据点上性能比R快78倍，4台服务器上测试1400万数据算法线性加速大于0.7，且数据量越大算法线性加速比越高

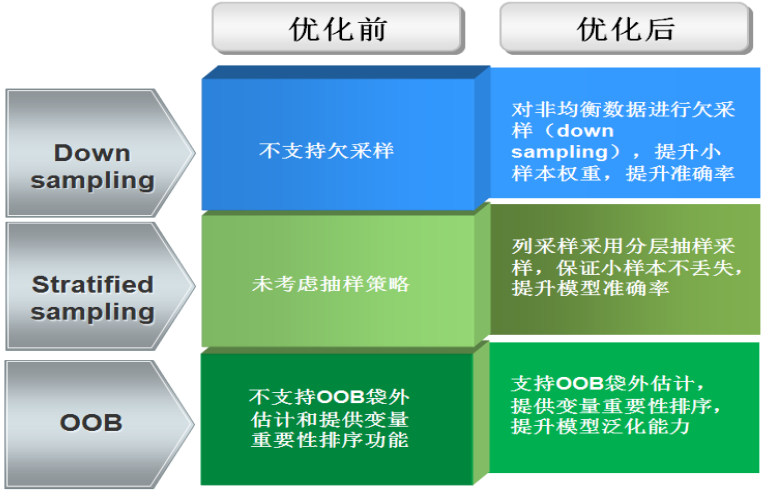
Random Forest:提升模型准确率、泛化能力

背景

- 随机森林基本思想是多棵CART树组合投票决定最后的结果，利用行随机有放回抽样（Bootstrapping）和列随机无放回抽样生成特征。
- RF相比单棵树，具有准确度高、泛化能力强等优点，可用于分类和回归业务场景，如离网预测

开源MLlib RF算法的问题

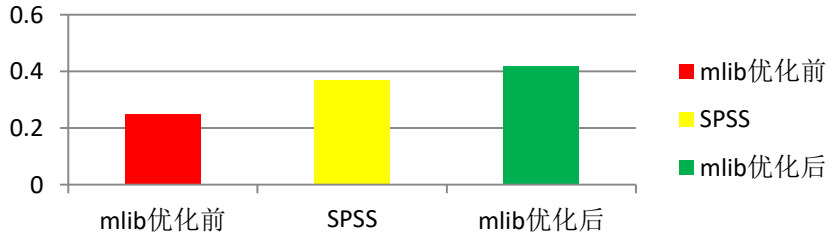
- 对非均衡数据分类准确率比商用SPSS低10个百分点。
- 列采样采用全局随机采样会导致小样本丢失或者严重失衡。
- 不支持OOB袋外估计，对特征按重要性排序（提升算法泛化能力）



集群：1Master+3Slave（Memory：20G+CPU：6cores）

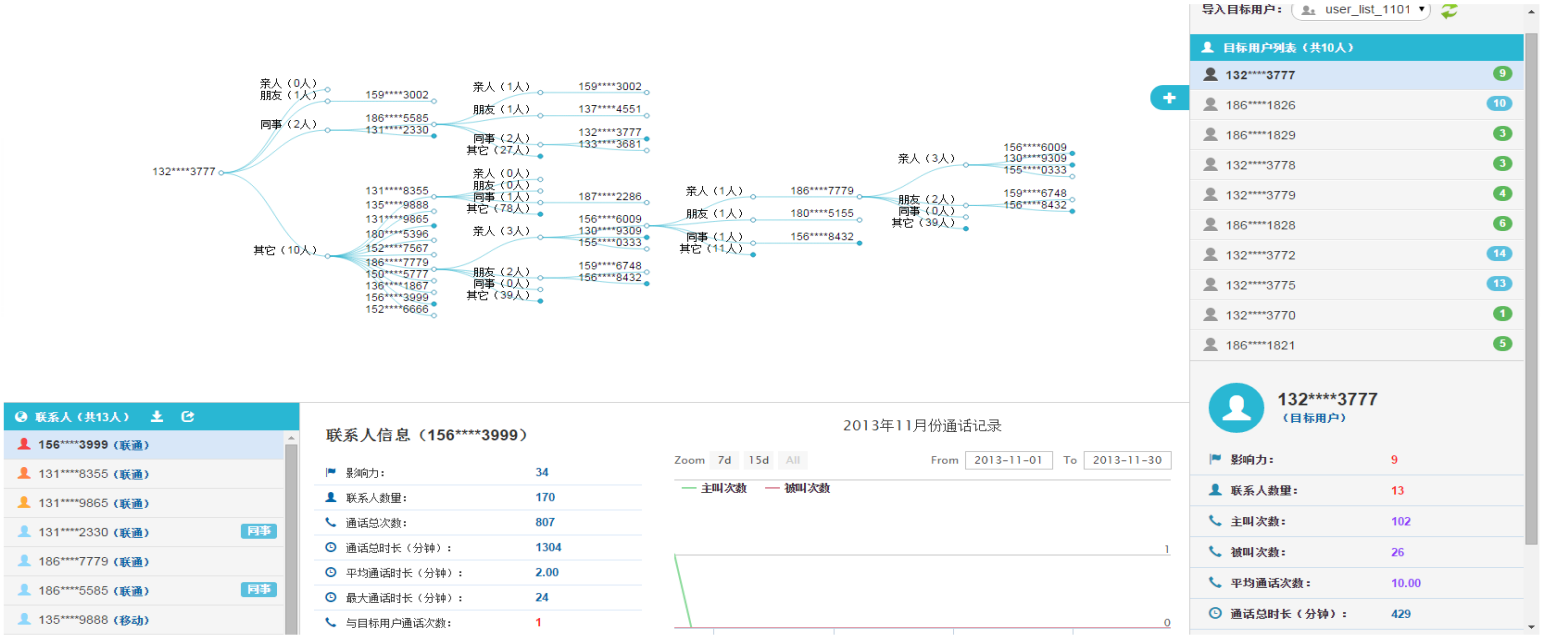
测试数据集：100W用户，300维特征

F1指标： 精确率和召回率的调和均值：F1= 2PR/(P+R)



测试结果：模型准确率F1比开源RF提升17%

社交圈分析：分布式图挖掘PageRank算法



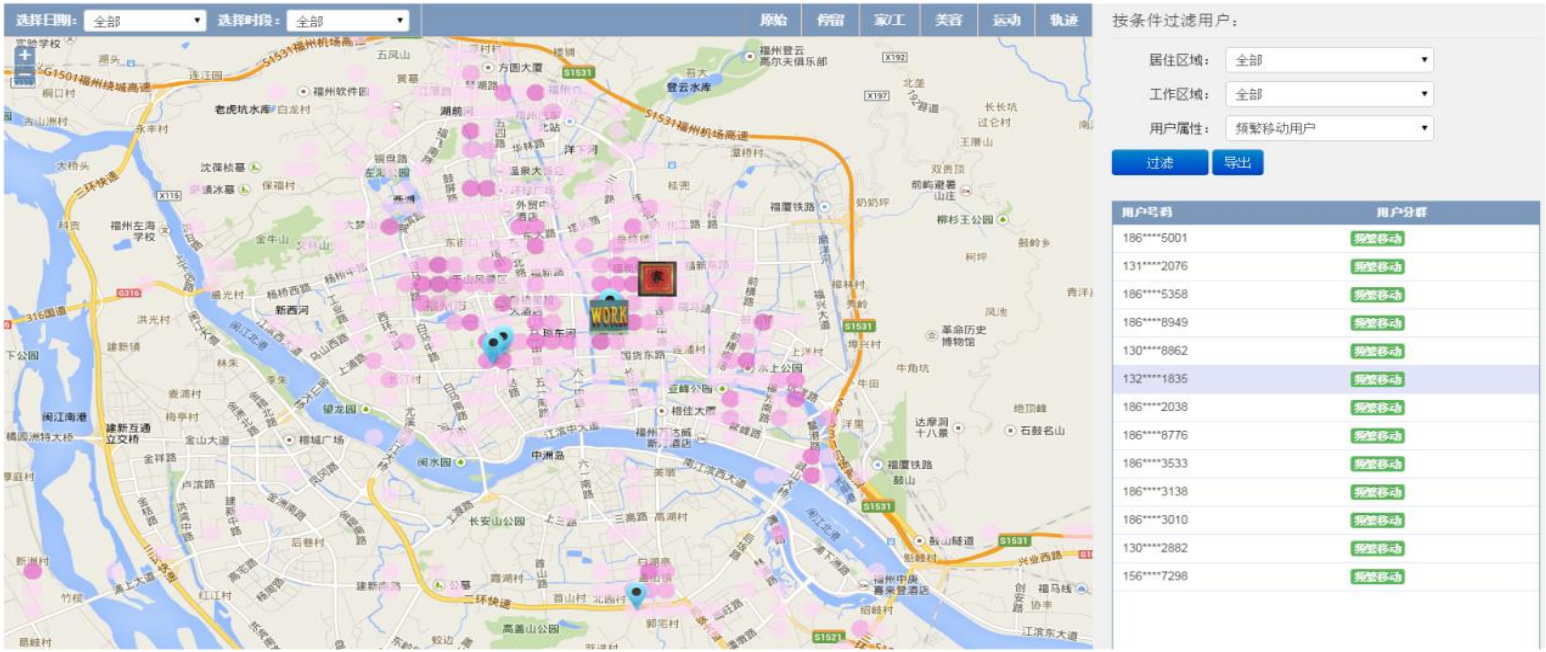
基于用户间通话记录，利用PageRank算法计算用户影响力，分布式PageRank算法性能比Graphx快2-3倍。

基于GIM-V LI行列分块提升矩阵乘算法性能。

GIM-V LI优点：内存占用少，计算速度快，网络传输小，适用于大规模的图计算。

用户分群：分布式Kernel SVM分类算法

简介：通过位置、XDR数据，结合百度POI数据，获取用户在美容院、运动场所、夜店机场等特定场所的驻留时长、上网使用时尚APP流量等信息，构造年轻时尚女性、出租司机、体育运动爱好者的特征；利用Kernel SVM进行用户分群，算法准确性比Linear SVM算法提高10%



THANKS