Alink: 基于Apache Flink的算法平台

Alink: an Algorithm Platform on Apache Flink

公司：阿里巴巴

职位：资深技术专家

演讲者：杨旭

Xu Yang

Senior Staff Engineer at Alibaba Group

What is Alink?

FLINK FORWARD

- PAI 算法平台的一部分，是基于Flink的算法平台。
  Part of the PAI algorithm platform, based on Flink's algorithm platform.

- 同时支持批式 / 流式算法，支持机器学习、统计等方面的二百多种常用算法
  Support batch/streaming algorithms, support more than 200 commonly used algorithms in machine learning, statistics, etc.

- 帮助数据分析和应用开发人员能够从数据探索、模型训练、实时预测、可视化展示，端到端地完成整个流程。
  Help data analytics and application developers complete the process from end to end with data exploration, model training, real-time forecasting, and visual presentation.

What is Alink?

FLINK FORWARD

- 相关名称的公共部分
  Common part of related words

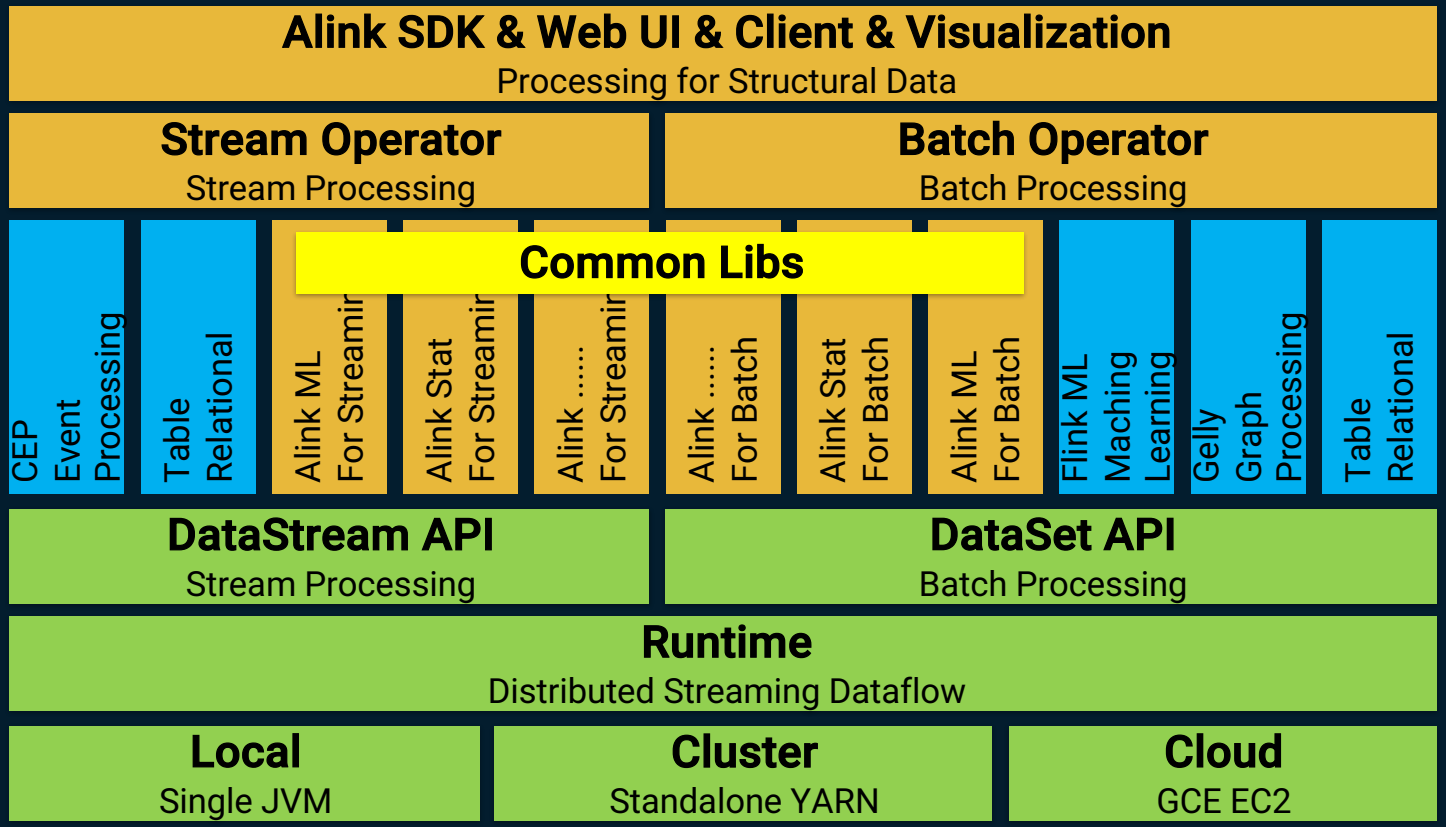
  Alibaba, Algorithm, AI, Flink, Blink

- 各算法功能通过“link”的方式进行链接
  Each algorithm function is linked by means of "link"

```
op1.link(op2)
```

```
op3.linkFrom(op1,op2)
```

Alink 架构(Alink Architecture)

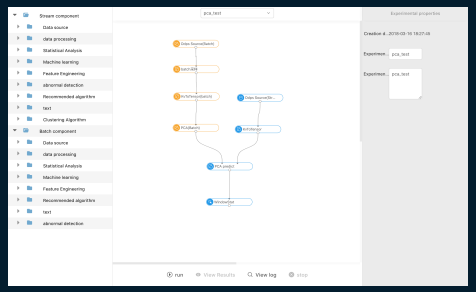


How to use?

FLINK FORWARD

多种调用方式，适合不同用户及场景

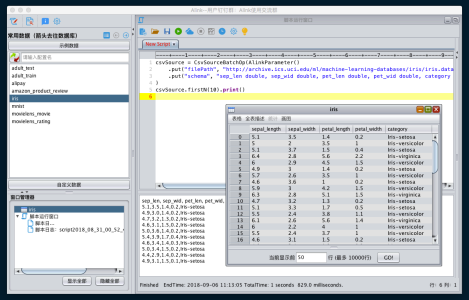
Three calling modes for different users and scenes



网页前端(Web UI)

简单便捷，工作流配置和执行

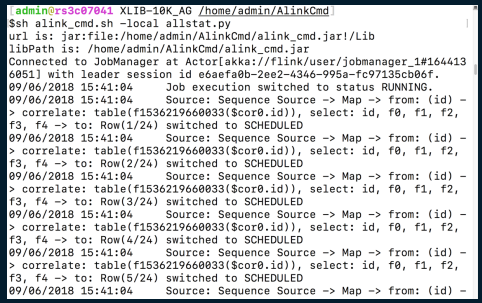
Drag-drop, easy to build workflow



PC客户端(Client)

支持脚本编辑运行，支持本地运行与集群运行

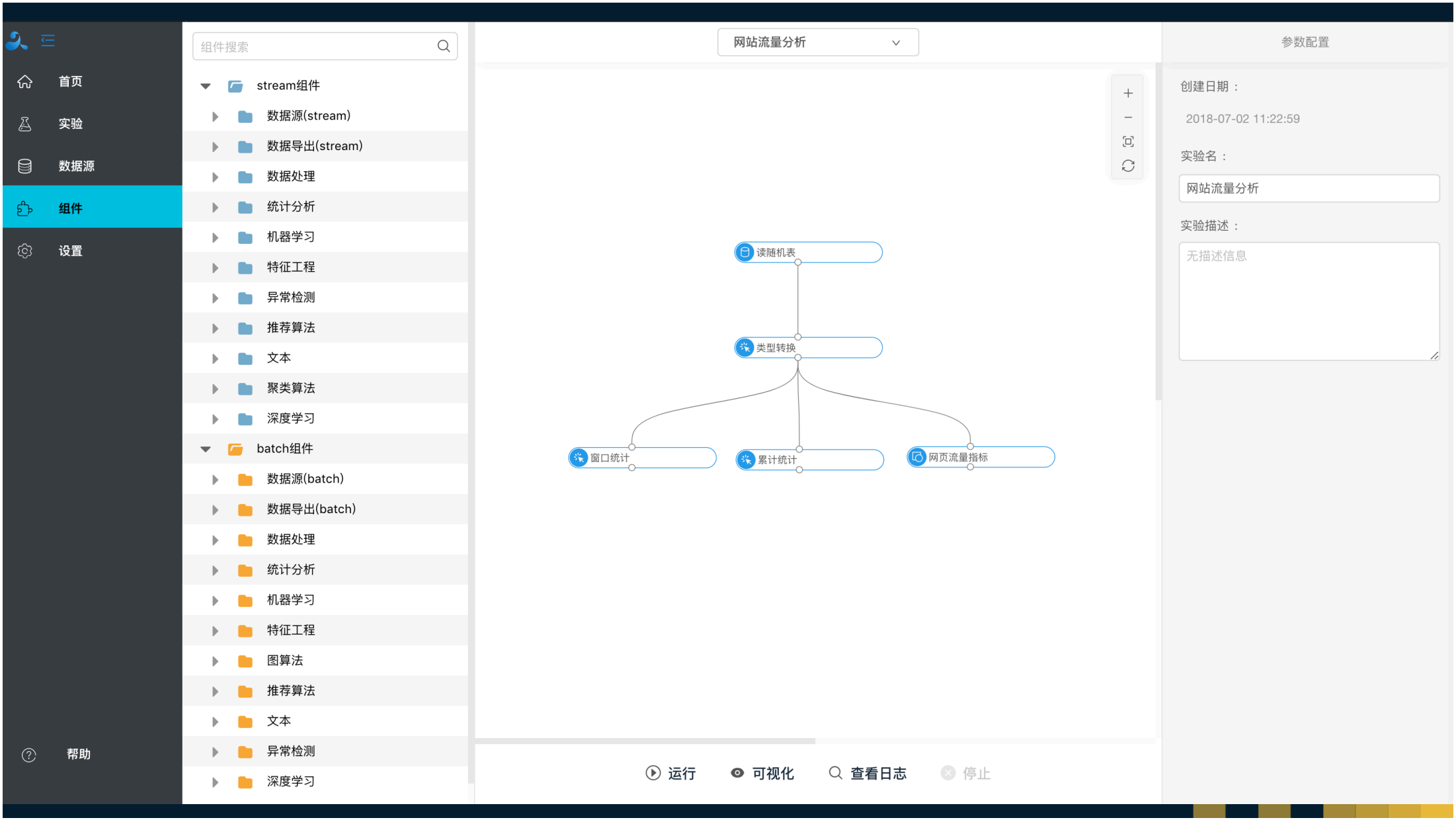
Local run， Edit and run script

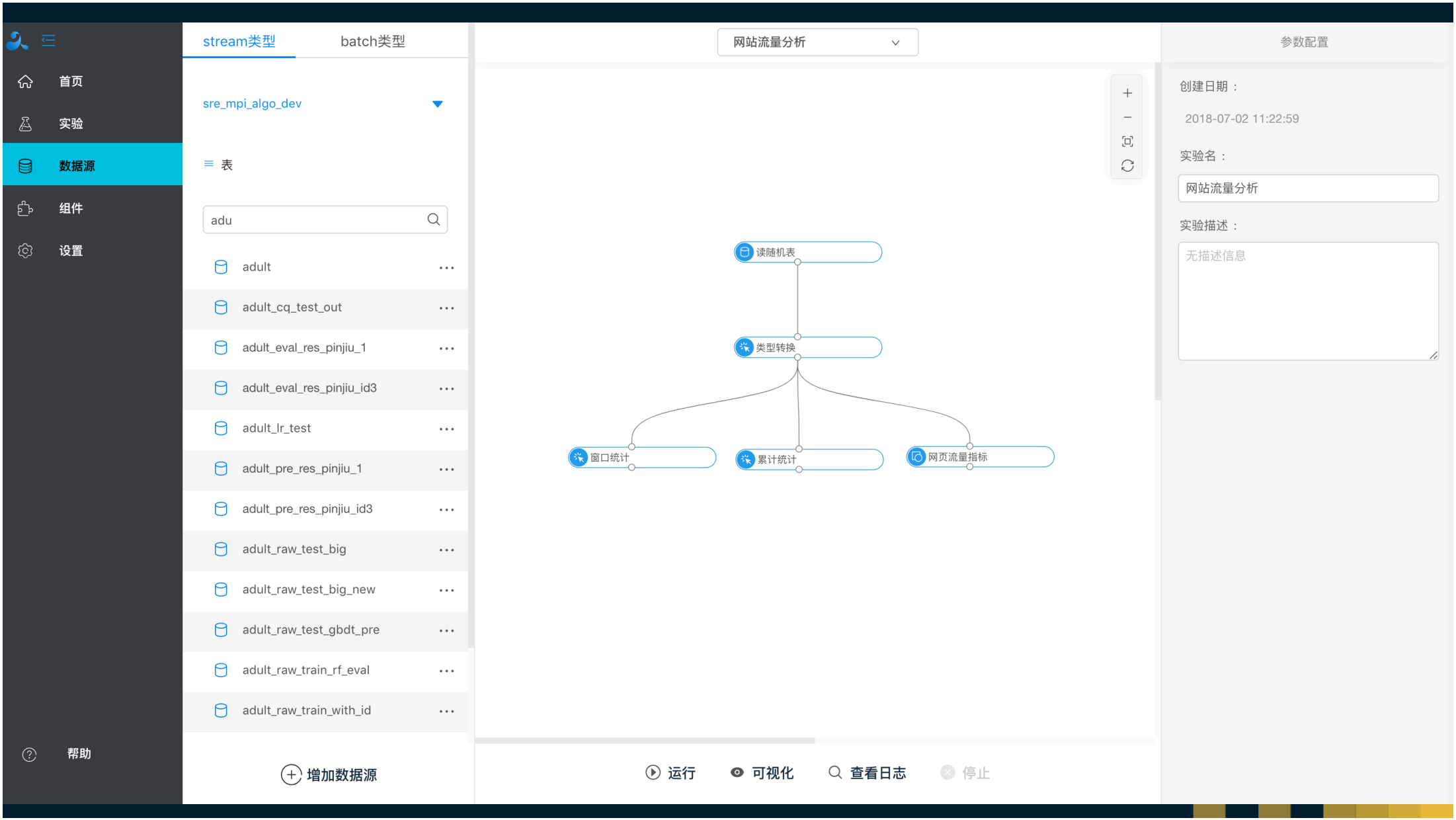


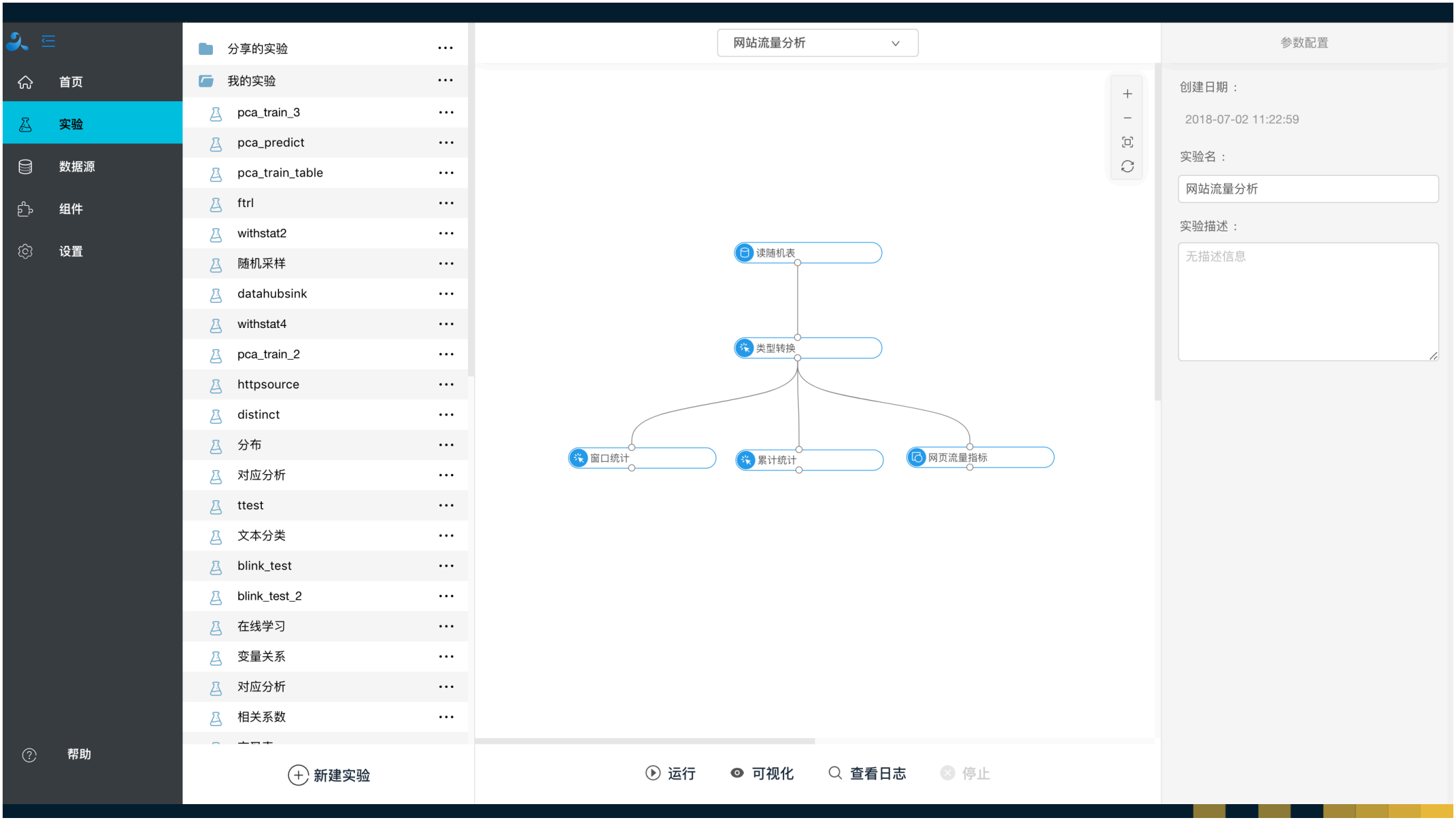
命令行(Console)

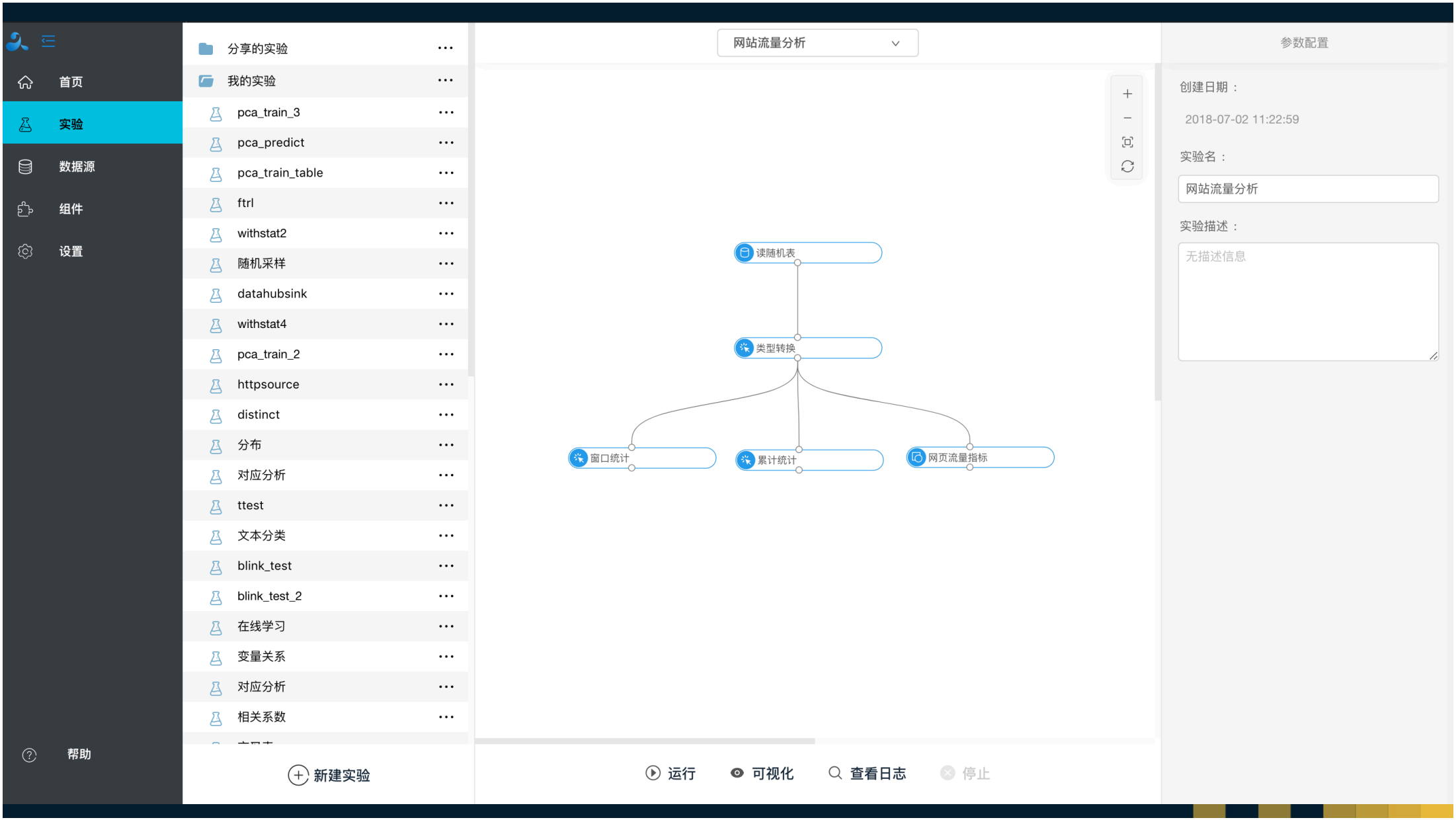
运行Alink脚本

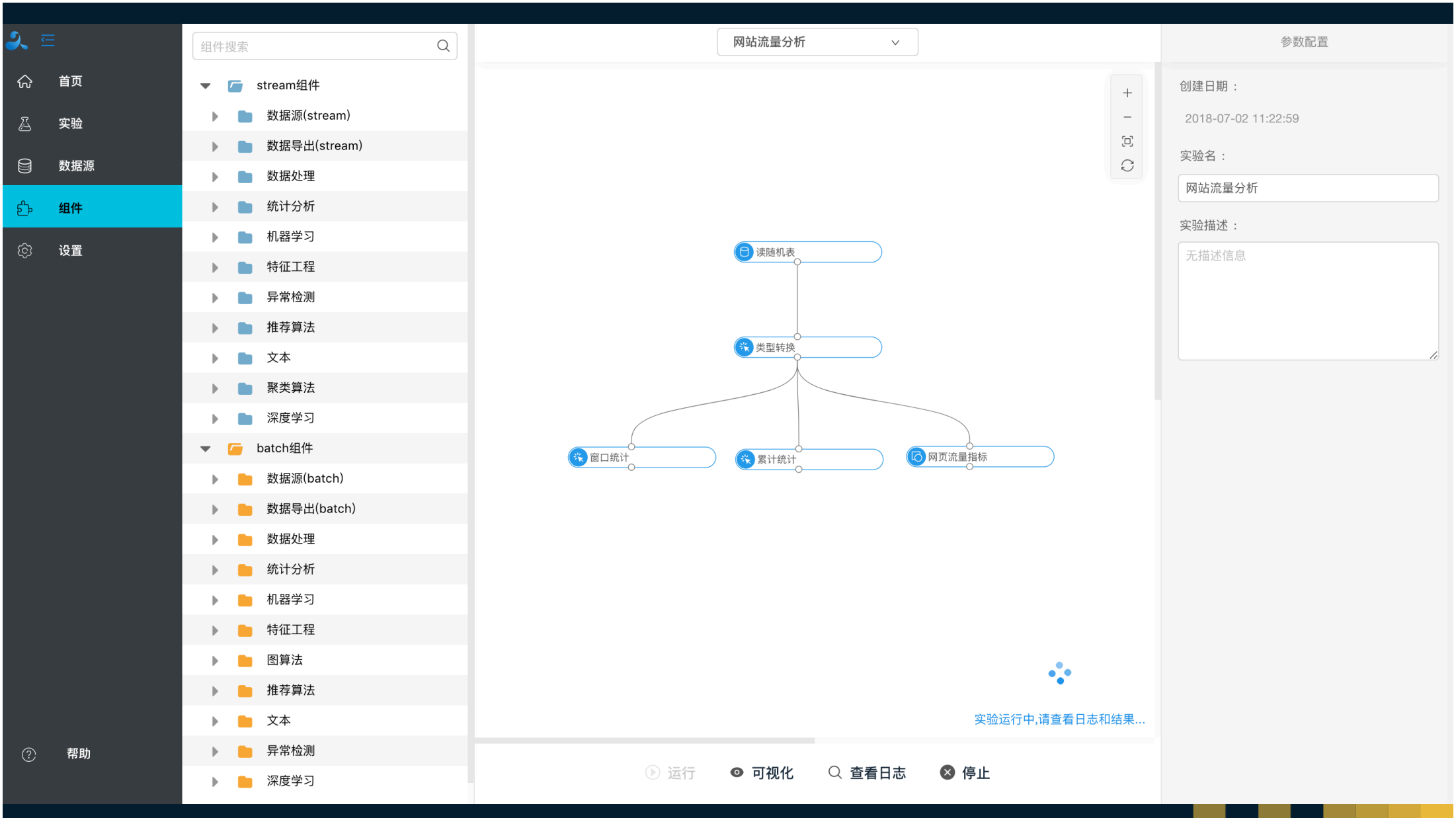
Excute Alink Scripts


首页
实验
数据源
组件
设置
帮助
组件搜索
stream组件
数据源(stream)
数据导出(stream)
数据处理
统计分析
机器学习
特征工程
异常检测
推荐算法
文本
聚类算法
深度学习
batch组件
数据源(batch)
数据导出(batch)
数据处理
统计分析
机器学习
特征工程
图算法
推荐算法
文本
异常检测
深度学习
网站流量分析
读随机表
类型转换
窗口统计
累计统计
网页流量指标
运行
可视化
查看日志
停止
参数配置
创建日期：
2018-07-02 11:22:59
实验名：
网站流量分析
实验描述：
无描述信息

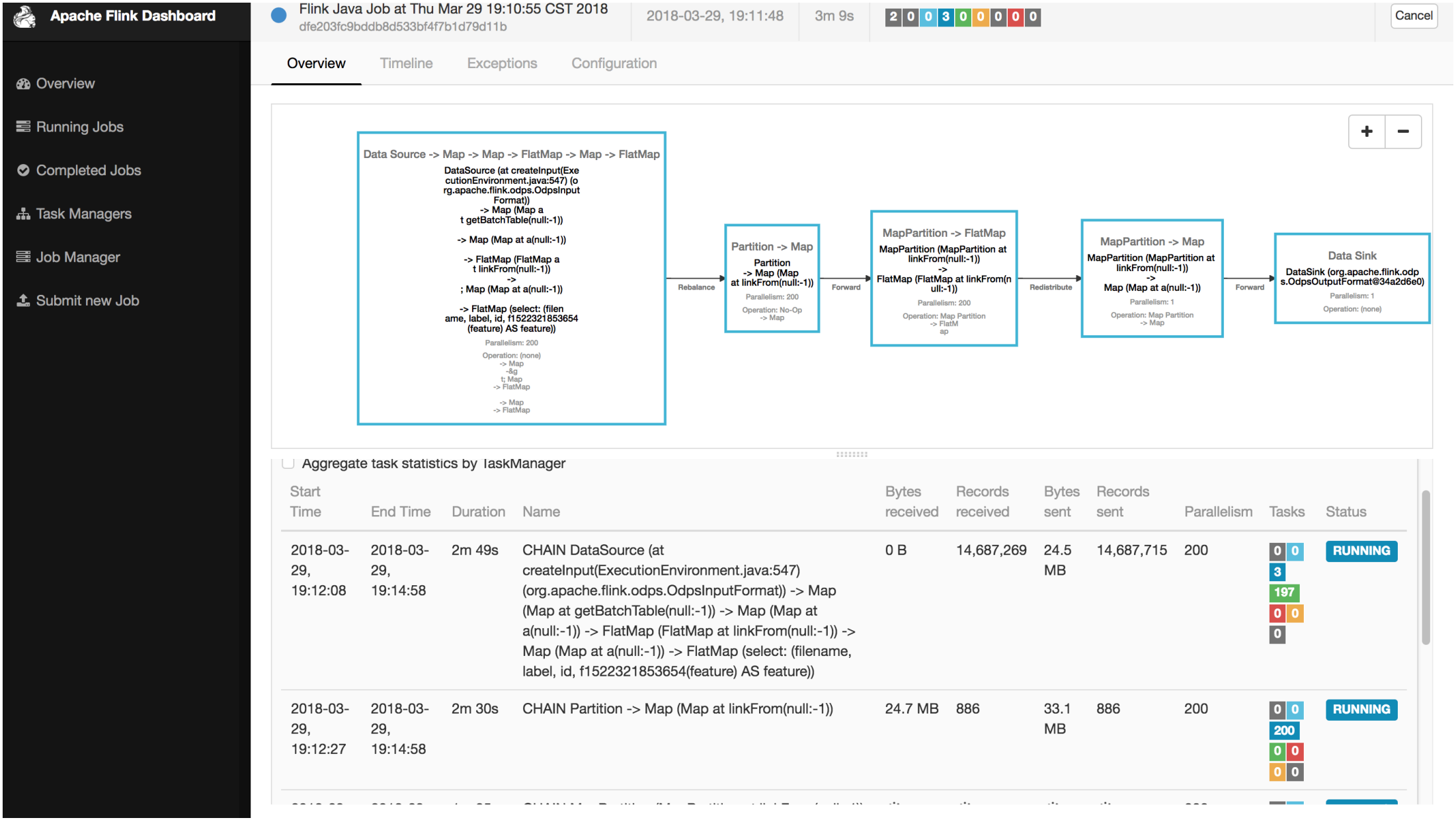

首页
实验
数据源
组件
设置
帮助
stream类型
batch类型
sre_mpi_algo_dev
表
adu
adult
adult_cq_test_out
adult_eval_res_pinjiu_1
adult_eval_res_pinjiu_id3
adult_lr_test
adult_pre_res_pinjiu_1
adult_pre_res_pinjiu_id3
adult_raw_test_big
adult_raw_test_big_new
adult_raw_test_gbdt_pre
adult_raw_train_rf_eval
adult_raw_train_with_id
增加数据源
网站流量分析
读随机表
类型转换
窗口统计
累计统计
网页流量指标
运行
可视化
查看日志
停止
参数配置
创建日期：
2018-07-02 11:22:59
实验名：
网站流量分析
实验描述：
无描述信息

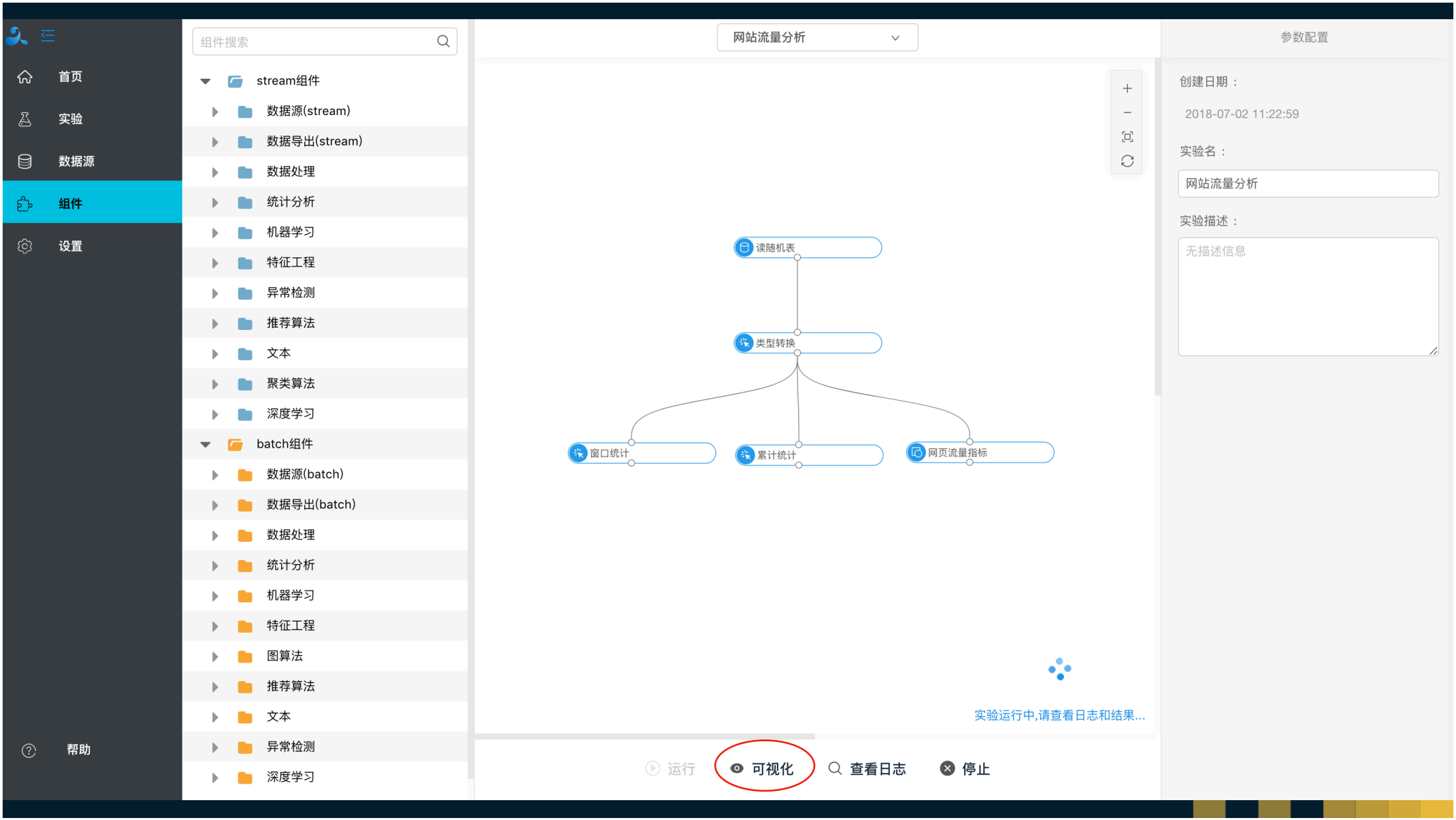

首页
实验
数据源
组件
设置
帮助
分享的实验
我的实验
pca_train_3
pca_predict
pca_train_table
ftrl
withstat2
随机采样
datahubsink
withstat4
pca_train_2
httpsource
distinct
分布
对应分析
ttest
文本分类
blink_test
blink_test_2
在线学习
变量关系
对应分析
相关系数
新建实验
网站流量分析
读随机表
类型转换
窗口统计
累计统计
网页流量指标
运行
可视化
查看日志
停止
参数配置
创建日期：
2018-07-02 11:22:59
实验名：
网站流量分析
实验描述：
无描述信息


首页
实验
数据源
组件
设置
帮助
分享的实验
我的实验
pca_train_3
pca_predict
pca_train_table
ftrl
withstat2
随机采样
datahubsink
withstat4
pca_train_2
httpsource
distinct
分布
对应分析
ttest
文本分类
blink_test
blink_test_2
在线学习
变量关系
对应分析
相关系数
新建实验
网站流量分析
读随机表
类型转换
窗口统计
累计统计
网页流量指标
运行
可视化
查看日志
停止
参数配置
创建日期：
2018-07-02 11:22:59
实验名：
网站流量分析
实验描述：
无描述信息


首页
实验
数据源
组件
设置
帮助
组件搜索
stream组件
数据源(stream)
数据导出(stream)
数据处理
统计分析
机器学习
特征工程
异常检测
推荐算法
文本
聚类算法
深度学习
batch组件
数据源(batch)
数据导出(batch)
数据处理
统计分析
机器学习
特征工程
图算法
推荐算法
文本
异常检测
深度学习
网站流量分析
读随机表
类型转换
窗口统计
累计统计
网页流量指标
实验运行中,请查看日志和结果...
运行
可视化
查看日志
停止
参数配置
创建日期：
2018-07-02 11:22:59
实验名：
网站流量分析
实验描述：
无描述信息

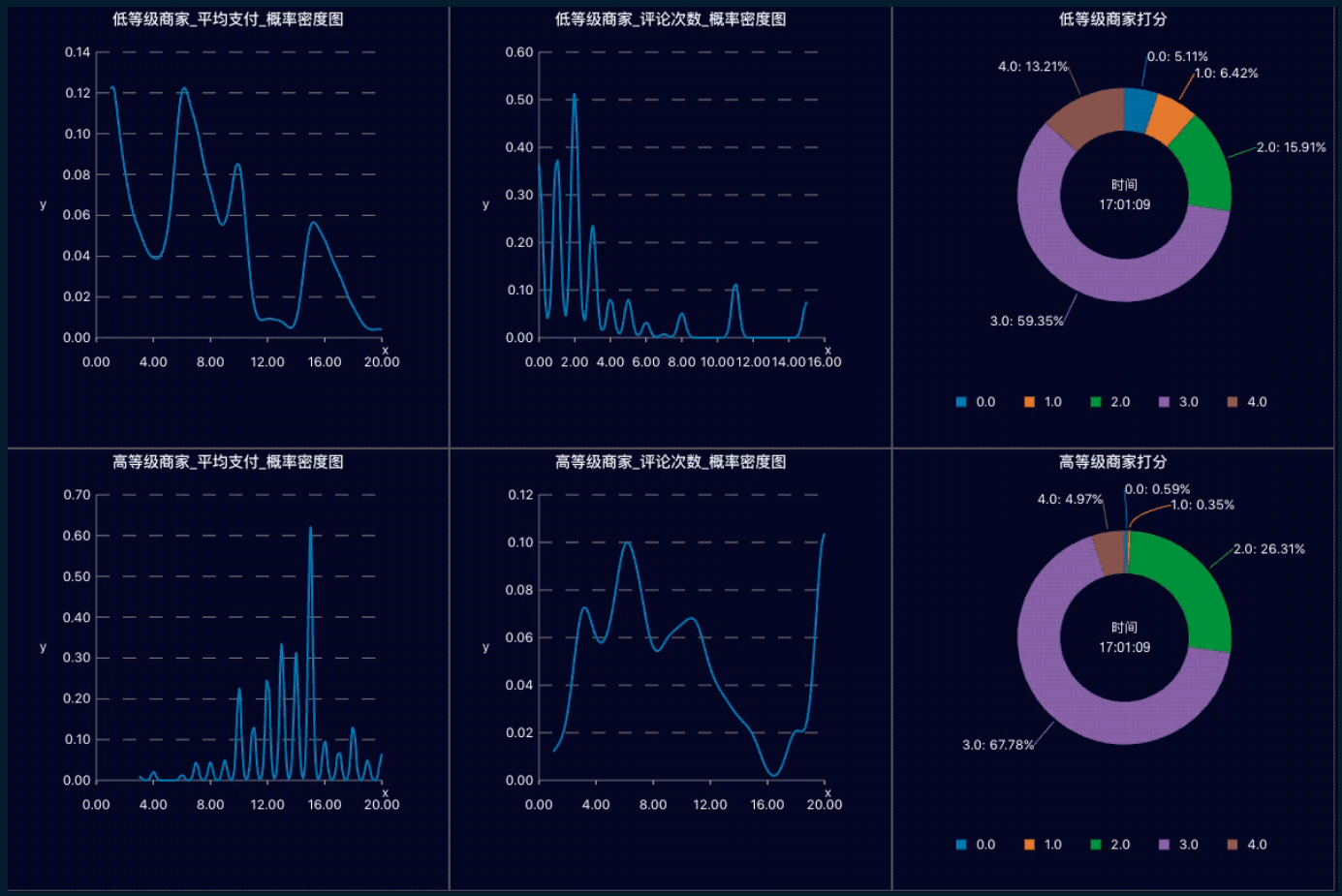
[70].2018-09-05 16:29:47,561 INFO org.apache.flink.yarn.Utils - Copying from file:/home/admin/data/local/cupid_flink/cupid_flink_1bba622031e5903c878736368ebc08e6/lib/flink-dist_2.11-1.4.0.jar to tempresource://sre_mpi_algo_dev/.flink/application_1536136176053_793701766/flink-dist_2.11-1.4.0.jar
[71].2018-09-05 16:29:54,347 INFO org.apache.flink.yarn.Utils - Copying from /tmp/application_1536136176053_793701766-flink-conf.yaml5958475879321536138.tmp to tempresource://sre_mpi_algo_dev/.flink/application_1536136176053_793701766/application_1536136176053_793701766-flink-conf.yaml5958475879321536138.tmp
[72].2018-09-05 16:30:03,942 INFO org.apache.flink.yarn.Utils - Copying from file:/tmp/application_1536136176053_7937017666168014009721836240.tmp to tempresource://sre_mpi_algo_dev/.flink/application_1536136176053_793701766/application_1536136176053_7937017666168014009721836240.tmp
[73].2018-09-05 16:30:09,222 INFO org.apache.flink.yarn.YarnClusterDescriptor - Submitting application master application_1536136176053_793701766
[74].2018-09-05 16:30:48,171 INFO org.apache.hadoop.yarn.client.api.impl.YarnClientImpl - Submitted applicationType Apache Flink application application_1536136176053_793701766 to ResourceManager at instanceId 20180905083010364glv9ah89
[75].2018-09-05 16:30:48,171 INFO org.apache.flink.yarn.YarnClusterDescriptor - Waiting for the cluster to be allocated
[76].2018-09-05 16:30:52,327 INFO org.apache.flink.yarn.YarnClusterDescriptor - YARN application has been deployed successfully.
[77].2018-09-05 16:30:52,328 INFO org.apache.flink.yarn.YarnClusterDescriptor - Application master address: 11.137.198.4
[78].2018-09-05 16:30:52,328 INFO org.apache.flink.yarn.YarnClusterDescriptor - jobview(job web ui): http://jobview.odps.aliyun-inc.com/proxyview/jobview/?h=http://service.odps.aliyun-inc.com/api&p=sre_mpi_algo_dev&i=20180905083010364glv9ah89&t=flink&id=application_1536136176053_793701766&metaname=20180905083010364glv9ah89&token=VFhHd0lXRjFFUUlCMWJBRzdQNzVQRTI3NEkwPSxPRFBTX09CTzoxNjIwOTcwMjYyOTYxMT
[79].2018-09-05 16:30:52,328 INFO org.apache.flink.yarn.YarnClusterDescriptor - Flink tracking url: http://11.137.198.4:44674
[80].2018-09-05 16:30:52,328 INFO org.apache.flink.yarn.YarnClusterDescriptor - InstanceId: 20180905083010364glv9ah89
[81].InstanceId: 20180905083010364glv9ah89
[82].2018-09-05 16:30:52,473 INFO org.apache.flink.yarn.YarnClusterDescriptor - Cupid LogView: http://logview.odps.aliyun-inc.com:8080/logview/?h=http://service.odps.aliyun-inc.com/api&p=sre_mpi_algo_dev&i=20180905083010364glv9ah89&token=akZBR2xqdXVPYnBTRVYrR2VHczhvYkRtSnZ3PSxPRFBTX09CTzoxNjIwOTcwMjYyOTYxMTgyLDE1MzY3NDEwNTIseyJTdGF0ZW1lbnQiOlt7IkFjdGlvbiI6WyJvZHBzOlJlYWQiXSwiRWZmZWN
[83].Cupid LogView: http://logview.odps.aliyun-inc.com:8080/logview/?h=http://service.odps.aliyun-inc.com/api&p=sre_mpi_algo_dev&i=20180905083010364glv9ah89&token=akZBR2xqdXVPYnBTRVYrR2VHczhvYkRtSnZ3PSxPRFBTX09CTzoxNjIwOTcwMjYyOTYxMTgyLDE1MzY3NDEwNTIseyJTdGF0ZW1lbnQiOlt7IkFjdGlvbiI6WyJvZHBzOlJlYWQiXSwiRWZmZWN
[84].2018-09-05 16:30:52,769 INFO org.apache.flink.yarn.YarnClusterDescriptor - Application report for application_1536136176053_793701766 (state RUNNING)
[85].2018-09-05 16:30:57,854 INFO org.apache.flink.yarn.YarnClusterDescriptor - Application report for application_1536136176053_793701766 (state RUNNING)
[86].2018-09-05 16:31:08,015 INFO org.apache.flink.yarn.YarnClusterDescriptor - Application report for application_1536136176053_793701766 (state RUNNING)
[87].2018-09-05 16:31:29,502 INFO org.apache.flink.yarn.YarnClusterDescriptor - Application report for application_1536136176053_793701766 (state RUNNING)
[88].2018-09-05 16:32:11,585 INFO org.apache.flink.yarn.YarnClusterDescriptor - Application report for application_1536136176053_793701766 (state RUNNING)
[89].2018-09-05 16:33:43,102 INFO org.apache.flink.yarn.YarnClusterDescriptor - Application report for application_1536136176053_793701766 (state RUNNING)
[90].2018-09-05 16:36:38,907 INFO org.apache.flink.yarn.YarnClusterDescriptor - Application report for application_1536136176053_793701766 (state RUNNING)
[91].2018-09-05 16:42:22,172 INFO org.apache.flink.yarn.YarnClusterDescriptor - Application report for application_1536136176053_793701766 (state RUNNING)
[92].2018-09-05 16:53:35,516 INFO org.apache.flink.yarn.YarnClusterDescriptor - Application report for application_1536136176053_793701766 (state RUNNING)
[93].2018-09-05 17:16:40,267 INFO org.apache.flink.yarn.YarnClusterDescriptor - Application report for application_1536136176053_793701766 (state RUNNING)
[94].2018-09-05 18:02:14,192 INFO org.apache.flink.yarn.YarnClusterDescriptor - Application report for application_1536136176053_793701766 (state RUNNING)
[95].2018-09-05 18:29:29,839 INFO org.apache.flink.yarn.YarnClusterDescriptor - Application report for application_1536136176053_793701766 (state RUNNING)
[96].2018-09-05 19:32:59,583 INFO org.apache.flink.yarn.YarnClusterDescriptor - Application report for application_1536136176053_793701766 (state RUNNING)
[97].2018-09-05 20:29:31,056 INFO org.apache.flink.yarn.YarnClusterDescriptor - Application report for application_1536136176053_793701766 (state RUNNING)
[98].2018-09-05 22:29:31,905 INFO org.apache.flink.yarn.YarnClusterDescriptor - Application report for application_1536136176053_793701766 (state RUNNING)
[99].2018-09-05 22:35:06,701 INFO org.apache.flink.yarn.YarnClusterDescriptor - Application report for application_1536136176053_793701766 (state RUNNING)
[100].2018-09-06 00:29:36,694 INFO org.apache.flink.yarn.YarnClusterDescriptor - Application report for application_1536136176053_793701766 (state RUNNING)
[101].2018-09-06 02:29:40,554 INFO org.apache.flink.yarn.YarnClusterDescriptor - Application report for application_1536136176053_793701766 (state RUNNING)
[102].2018-09-06 04:29:45,221 INFO org.apache.flink.yarn.YarnClusterDescriptor - Application report for application_1536136176053_793701766 (state RUNNING)
[103].2018-09-06 04:40:28,727 INFO org.apache.flink.yarn.YarnClusterDescriptor - Application report for application_1536136176053_793701766 (state RUNNING)
[104].2018-09-06 06:29:47,470 INFO org.apache.flink.yarn.YarnClusterDescriptor - Application report for application_1536136176053_793701766 (state RUNNING)
[105].2018-09-06 08:29:48,795 INFO org.apache.flink.yarn.YarnClusterDescriptor - Application report for application_1536136176053_793701766 (state RUNNING)
[106].2018-09-06 10:29:50,162 INFO org.apache.flink.yarn.YarnClusterDescriptor - Application report for application_1536136176053_793701766 (state RUNNING)

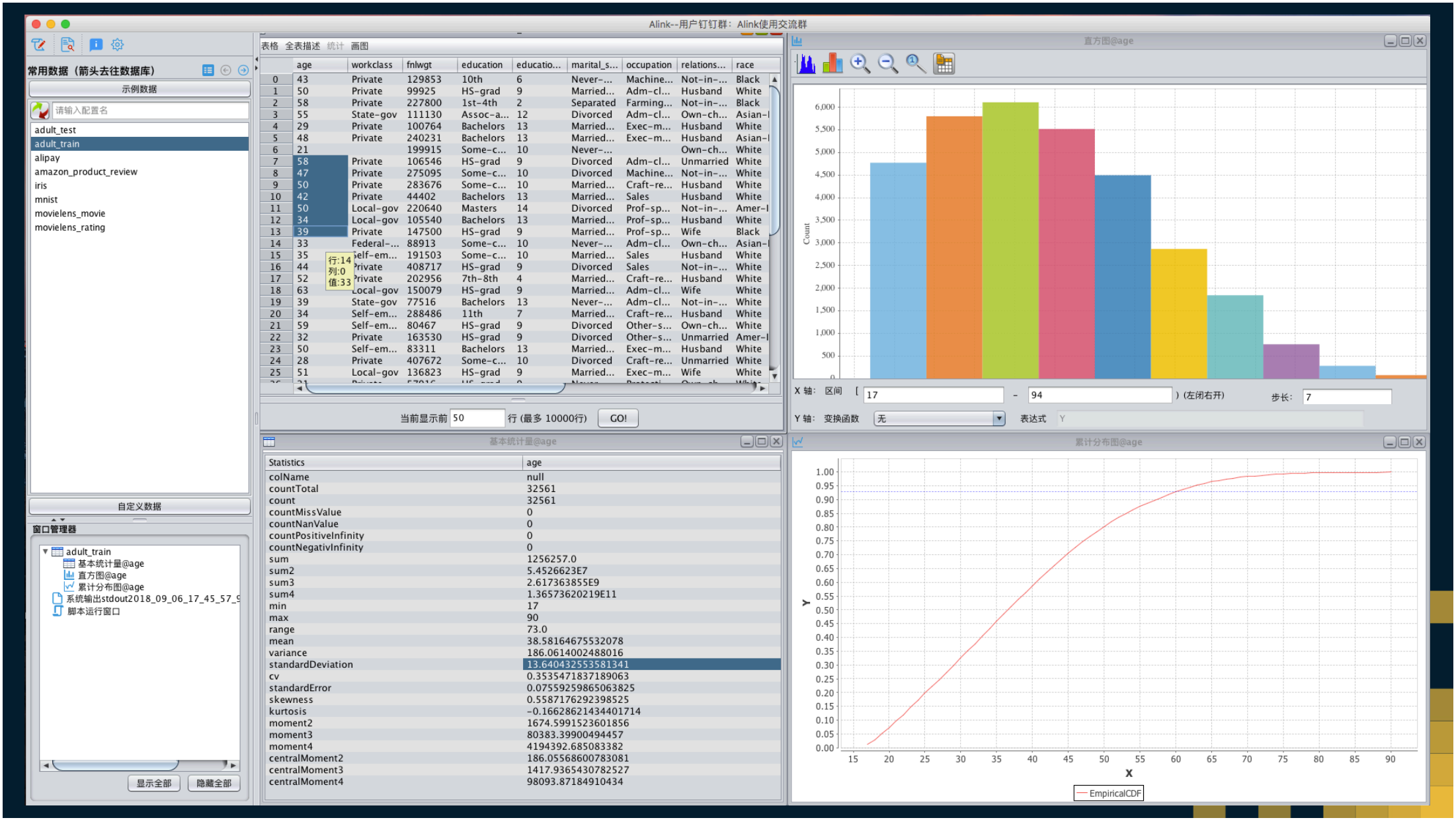

Apache Flink Dashboard
Overview
Running Jobs
Completed Jobs
Task Managers
Job Manager
Submit new Job
Flink Java Job at Thu Mar 29 19:10:55 CST 2018
dfe203fc9bddb8d533bf4f7b1d79d11b
2018-03-29, 19:11:48
3m 9s
2 0 0 3 0 0 0 0 0
Cancel
Overview
Timeline
Exceptions
Configuration
Data Source -> Map -> Map -> FlatMap -> Map -> FlatMap
DataSource (at createInput(ExecutionEnvironment.java:547) (org.apache.flink.odps.OdpsInputFormat))
-> Map (Map at getBatchTable(null:-1))
-> Map (Map at a(null:-1))
-> FlatMap (FlatMap at linkFrom(null:-1))
->
; Map (Map at a(null:-1))
-> FlatMap (select: (filename, label, id, f1522321853654 (feature) AS feature))
Parallelism: 200
Operation: (none)
-> Map
-&g
t; Map
-> FlatMap
-> Map
-> FlatMap
Rebalance
Partition -> Map
Partition
-> Map (Map at linkFrom(null:-1))
Parallelism: 200
Operation: No-Op
-> Map
Forward
MapPartition -> FlatMap
MapPartition (MapPartition at linkFrom(null:-1))
->
FlatMap (FlatMap at linkFrom(null:-1))
Parallelism: 200
Operation: Map Partition
-> FlatM
ap
Redistribute
MapPartition -> Map
MapPartition (MapPartition at linkFrom(null:-1))
->
Map (Map at a(null:-1))
Parallelism: 1
Operation: Map Partition
-> Map
Forward
Data Sink
DataSink (org.apache.flink.odps.OdpsOutputFormat@34a2d6e0)
Parallelism: 1
Operation: (none)
Aggregate task statistics by TaskManager
Start Time
End Time
Duration
Name
Bytes received
Records received
Bytes sent
Records sent
Parallelism
Tasks
Status
2018-03-29, 19:12:08
2018-03-29, 19:14:58
2m 49s
CHAIN DataSource (at createInput(ExecutionEnvironment.java:547) (org.apache.flink.odps.OdpsInputFormat)) -> Map (Map at getBatchTable(null:-1)) -> Map (Map at a(null:-1)) -> FlatMap (FlatMap at linkFrom(null:-1)) -> Map (Map at a(null:-1)) -> FlatMap (select: (filename, label, id, f1522321853654(feature) AS feature))
0 B
14,687,269
24.5 MB
14,687,715
200
0 0 3 197 0 0 0
RUNNING
2018-03-29, 19:12:27
2018-03-29, 19:14:58
2m 30s
CHAIN Partition -> Map (Map at linkFrom(null:-1))
24.7 MB
886
33.1 MB
886
200
0 0 200 0 0 0 0
RUNNING

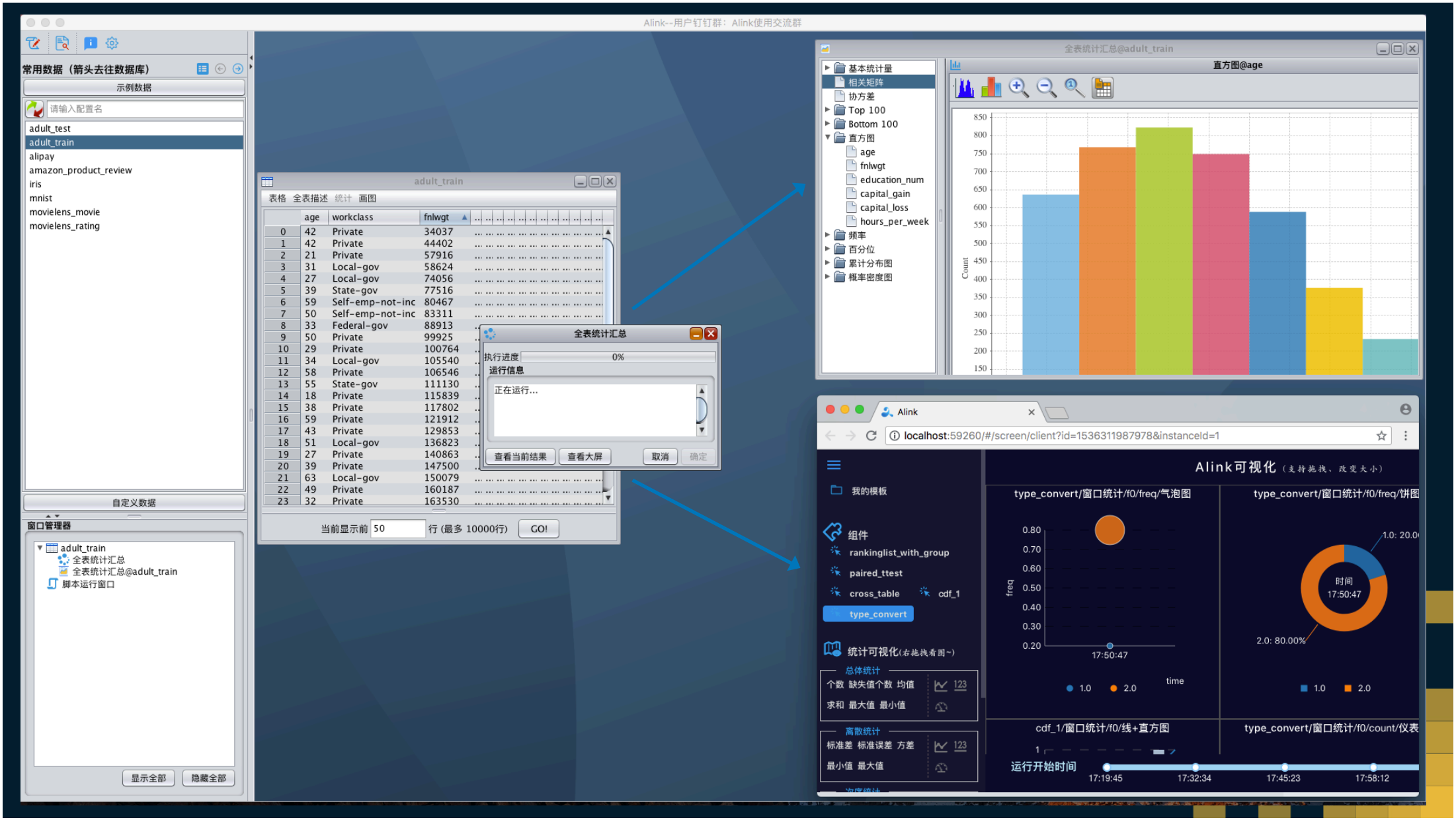

首页
实验
数据源
组件
设置
帮助
组件搜索
stream组件
数据源(stream)
数据导出(stream)
数据处理
统计分析
机器学习
特征工程
异常检测
推荐算法
文本
聚类算法
深度学习
batch组件
数据源(batch)
数据导出(batch)
数据处理
统计分析
机器学习
特征工程
图算法
推荐算法
文本
异常检测
深度学习
网站流量分析
读随机表
类型转换
窗口统计
累计统计
网页流量指标
实验运行中,请查看日志和结果...
运行
可视化
查看日志
停止
参数配置
创建日期：
2018-07-02 11:22:59
实验名：
网站流量分析
实验描述：
无描述信息

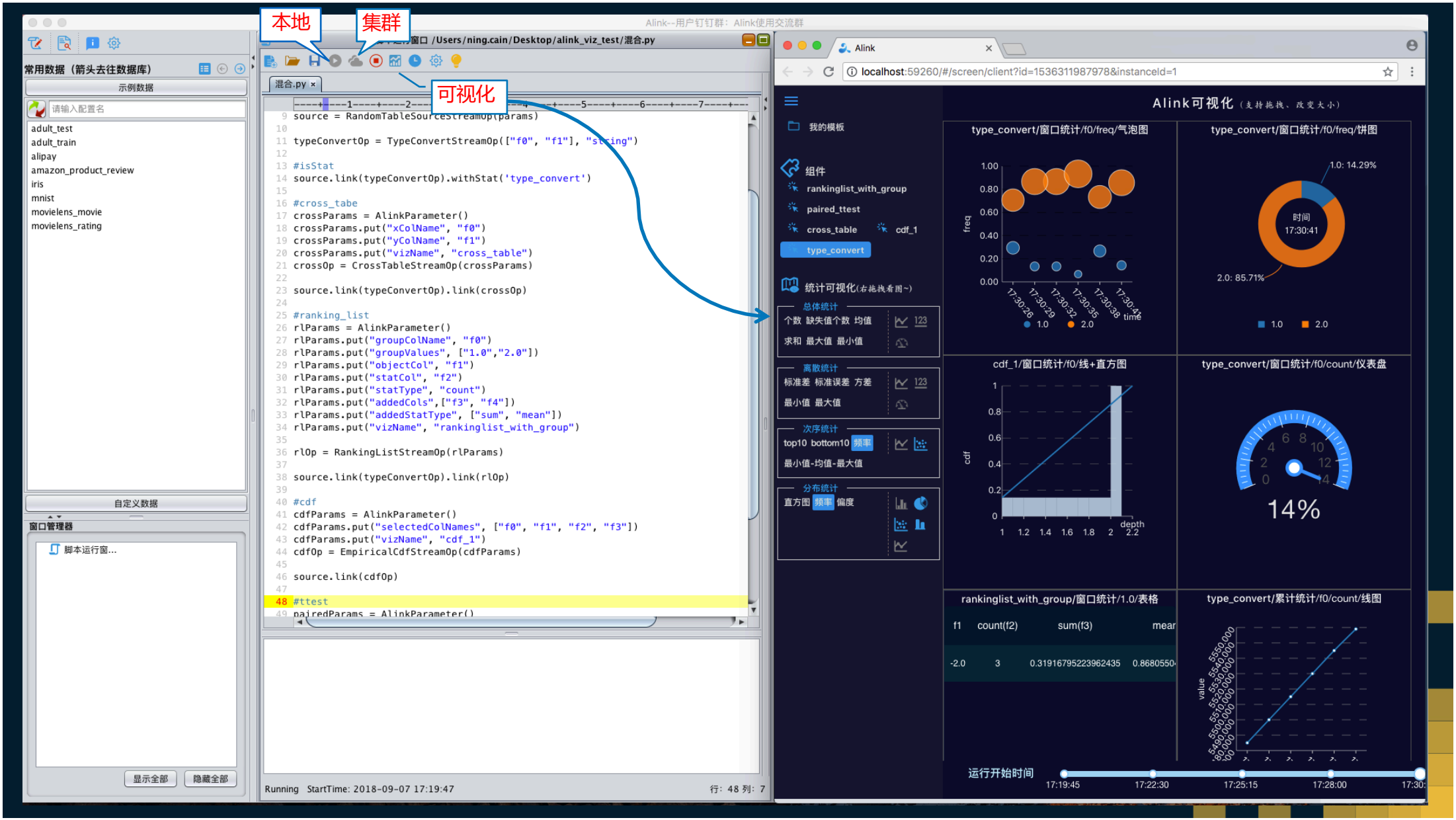

我的模板
组件
类型转换
网页流量指标
窗口统计
累计统计
统计可视化(右拖拽看图~)
总体统计
个数 缺失值个数 均值
求和 最大值 最小值
离散统计
标准差 标准误差 方差
最小值 最大值
次序统计
top10 bottom10 频率
最小值-均值-最大值
分布统计
直方图 频率 偏度
Alink可视化（支持拖拽、改变大小）
freq
time
1.0
2.0
value
sum
1.0: 20.00%
2.0: 80.00%
时间
16:59:12
窗口统计/f0/freq/层叠柱状图
窗口统计/f2/topK/线图
窗口统计/f2/mean/仪表盘
0.56%
运行开始时间
当前时间
16:47:51
16:50:42
16:53:33
16:56:24
16:59:16


WindowStat_shop_level_freq_bubble
weight
0.6
0.5
0.4
0.3
0.2
0.1
14:31:12
14:31:15
x
high
low
medium
WindowStat_shop_level_freq_bar
weight
1
0.8
0.6
0.4
0.2
0
14:31:12
14:31:15
x
high
low
medium
WindowStat_shop_level_count_line
y
490000
480000
470000
460000
450000
440000
430000
420000
410000
400000
14:31:16
时间
count
指标卡_WindowStat_shop_level_count
Value 408255.0000
环比 86.12%
时间 14:31:18
AllStat_shop_level_freq_pie
high: 24.0759%
medium: 48.1635%
low: 27.7606%
时间
14:31:18
high
low
medium
AllStat_shop_level_count_line
y
1400000
1200000
1000000
800000
600000
400000
14:31:16
时间
count


低等级商家_平均支付_概率密度图
y
x
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0.00 4.00 8.00 12.00 16.00 20.00
低等级商家_评论次数_概率密度图
y
x
0.60
0.50
0.40
0.30
0.20
0.10
0.00
0.00 2.00 4.00 6.00 8.00 10.00 12.00 14.00 16.00
低等级商家打分
4.0: 13.21%
0.0: 5.11%
1.0: 6.42%
2.0: 15.91%
时间
17:01:09
3.0: 59.35%
0.0 1.0 2.0 3.0 4.0
高等级商家_平均支付_概率密度图
y
x
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
0.00 4.00 8.00 12.00 16.00 20.00
高等级商家_评论次数_概率密度图
y
x
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0.00 4.00 8.00 12.00 16.00 20.00
高等级商家打分
4.0: 4.97%
0.0: 0.59%
1.0: 0.35%
2.0: 26.31%
时间
17:01:09
3.0: 67.78%
0.0 1.0 2.0 3.0 4.0


Alink--用户钉钉群：Alink使用交流群
常用数据（箭头去往数据库）
示例数据
请输入配置名
adult_test
adult_train
alipay
amazon_product_review
iris
mnist
movielens_movie
movielens_rating
自定义数据
窗口管理器
adult_train
基本统计量@age
直方图@age
累计分布图@age
系统输出stdout2018_09_06_17_45_57_9
脚本运行窗口
显示全部
隐藏全部
表格 全表描述 统计 画图
age workclass fnlwgt education educatio... marital_s... occupation relations... race
0 43 Private 129853 10th 6 Never-... Machine... Not-in-... Black
1 50 Private 99925 HS-grad 9 Married... Adm-cl... Husband White
2 58 Private 227800 1st-4th 2 Separated Farming... Not-in-... Black
3 55 State-gov 111130 Assoc-a... 12 Divorced Adm-cl... Own-ch... Asian-
4 29 Private 100764 Bachelors 13 Married... Exec-m... Husband White
5 48 Private 240231 Bachelors 13 Married... Exec-m... Husband Asian-
6 21 199915 Some-c... 10 Never-... Own-ch... White
7 58 Private 106546 HS-grad 9 Divorced Adm-cl... Unmarried White
8 47 Private 275095 Some-c... 10 Divorced Machine... Not-in-... White
9 50 Private 283676 Some-c... 10 Married... Craft-re... Husband White
10 42 Private 44402 Bachelors 13 Married... Sales Husband White
11 50 Local-gov 220640 Masters 14 Divorced Prof-sp... Not-in-... Amer-
12 34 Local-gov 105540 Bachelors 13 Married... Prof-sp... Husband White
13 39 Private 147500 HS-grad 9 Married... Prof-sp... Wife Black
14 33 Federal-... 88913 Some-c... 10 Never-... Adm-cl... Own-ch... Asian-
15 35 Self-em... 191503 Some-c... 10 Married... Sales Husband White
16 44 Private 408717 HS-grad 9 Divorced Sales Not-in-... White
17 52 Private 202956 7th-8th 4 Married... Craft-re... Husband White
18 63 Local-gov 150079 HS-grad 9 Married... Adm-cl... Wife White
19 39 State-gov 77516 Bachelors 13 Never-... Adm-cl... Not-in-... White
20 34 Self-em... 288486 11th 7 Married... Craft-re... Husband White
21 59 Self-em... 80467 HS-grad 9 Divorced Other-s... Own-ch... White
22 32 Private 163530 HS-grad 9 Divorced Other-s... Unmarried Amer-
23 50 Self-em... 83311 Bachelors 13 Married... Exec-m... Husband White
24 28 Private 407672 Some-c... 10 Divorced Craft-re... Unmarried White
25 51 Local-gov 136823 HS-grad 9 Married... Exec-m... Wife White
行:14
列:0
值:33
当前显示前 50 行 (最多 10000行) GO!
基本统计量@age
Statistics age
colName null
countTotal 32561
count 32561
countMissValue 0
countNanValue 0
countPositiveInfinity 0
countNegativInfinity 0
sum 1256257.0
sum2 5.4526623E7
sum3 2.617363855E9
sum4 1.36573620219E11
min 17
max 90
range 73.0
mean 38.58164675532078
variance 186.0614002488016
standardDeviation 13.640432553581341
cv 0.3535471837189063
standardError 0.07559259865063825
skewness 0.5587176292398525
kurtosis -0.16628621434401714
moment2 1674.5991523601856
moment3 80383.39900494457
moment4 4194392.685083382
centralMoment2 186.05568600783081
centralMoment3 1417.9365430782527
centralMoment4 98093.87184910434
直方图@age
Count
X 轴： 区间 [17 - 94) (左闭右开) 步长： 7
Y 轴： 变换函数 无 表达式 Y
累计分布图@age
y
x
EmpiricalCDF

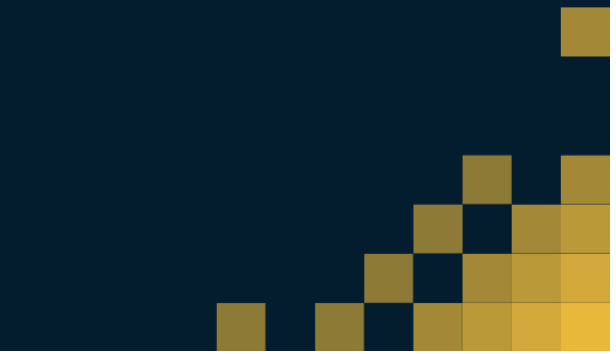

Alink--用户钉钉群：Alink使用交流群
常用数据（箭头去往数据库）
示例数据
请输入配置名
adult_test
adult_train
alipay
amazon_product_review
iris
mnist
movielens_movie
movielens_rating
自定义数据
窗口管理器
adult_train
全表统计汇总
全表统计汇总@adult_train
脚本运行窗口
显示全部
隐藏全部
adult_train
表格 全表描述 统计 画图
age workclass fnlwgt
0 42 Private 34037
1 42 Private 44402
2 21 Private 57916
3 31 Local-gov 58624
4 27 Local-gov 74056
5 39 State-gov 77516
6 59 Self-emp-not-inc 80467
7 50 Self-emp-not-inc 83311
8 33 Federal-gov 88913
9 50 Private 99925
10 29 Private 100764
11 34 Local-gov 105540
12 58 Private 106546
13 55 State-gov 111130
14 18 Private 115839
15 38 Private 117802
16 59 Private 121912
17 43 Private 129853
18 51 Local-gov 136823
19 27 Private 140863
20 39 Private 147500
21 63 Local-gov 150079
22 49 Private 160187
23 32 Private 163530
当前显示前 50 行 (最多 10000行) GO!
全表统计汇总
执行进度 0%
运行信息
正在运行...
查看当前结果 查看大屏 取消 确定
全表统计汇总@adult_train
基本统计量
相关矩阵
协方差
Top 100
Bottom 100
直方图
age
fnlwgt
education_num
capital_gain
capital_loss
hours_per_week
频率
百分位
累计分布图
概率密度图
直方图@age
Count
Alink
localhost:59260/#/screen/client?id=1536311987978&instanceId=1
我的模板
组件
rankinglist_with_group
paired_ttest
cross_table
cdf_1
type_convert
统计可视化(右拖拽看图~)
总体统计
个数 缺失值个数 均值
求和 最大值 最小值
离散统计
标准差 标准误差 方差
最小值 最大值
Alink可视化（支持拖拽、改变大小）
type_convert/窗口统计/f0/freq/气泡图
freq
time
1.0 2.0
type_convert/窗口统计/f0/freq/饼图
时间
17:50:47
2.0: 80.00%
cdf_1/窗口统计/f0/线+直方图
运行开始时间
17:19:45 17:32:34 17:45:23 17:58:12
type_convert/窗口统计/f0/count/仪表


本地
集群
可视化
Alink--用户钉钉群：Alink使用交流群
常用数据（箭头去往数据库）
示例数据
请输入配置名
adult_test
adult_train
alipay
amazon_product_review
iris
mnist
movielens_movie
movielens_rating
自定义数据
窗口管理器
脚本运行窗...
显示全部
隐藏全部
混合.py
source = RandomTableSourceStreamOp(params)
typeConvertOp = TypeConvertStreamOp(["f0", "f1"], "string")
#isStat
source.link(typeConvertOp).withStat('type_convert')
#cross_tabe
crossParams = AlinkParameter()
crossParams.put("xColName", "f0")
crossParams.put("yColName", "f1")
crossParams.put("vizName", "cross_table")
crossOp = CrossTableStreamOp(crossParams)
source.link(typeConvertOp).link(crossOp)
#ranking_list
rlParams = AlinkParameter()
rlParams.put("groupColName", "f0")
rlParams.put("groupValues", ["1.0","2.0"])
rlParams.put("objectCol", "f1")
rlParams.put("statCol", "f2")
rlParams.put("statType", "count")
rlParams.put("addedCols",["f3", "f4"])
rlParams.put("addedStatType", ["sum", "mean"])
rlParams.put("vizName", "rankinglist_with_group")
rlOp = RankingListStreamOp(rlParams)
source.link(typeConvertOp).link(rlOp)
#cdf
cdfParams = AlinkParameter()
cdfParams.put("selectedColNames", ["f0", "f1", "f2", "f3"])
cdfParams.put("vizName", "cdf_1")
cdfOp = EmpiricalCdfStreamOp(cdfParams)
source.link(cdfOp)
#ttest
Running StartTime: 2018-09-07 17:19:47
行：48 列：7
Alink
localhost:59260/#/screen/client?id=1536311987978&instanceId=1
Alink可视化（支持拖拽、改变大小）
我的模板
组件
rankinglist_with_group
paired_ttest
cross_table
cdf_1
type_convert
统计可视化(右拖拽看图~)
总体统计
个数 缺失值个数 均值
求和 最大值 最小值
离散统计
标准差 标准误差 方差
最小值 最大值
次序统计
top10 bottom10 频率
最小值-均值-最大值
分布统计
直方图 频率 偏度
type_convert/窗口统计/f0/freq/气泡图
type_convert/窗口统计/f0/freq/饼图
1.0: 14.29%
时间
17:30:41
2.0: 85.71%
1.0
2.0
cdf_1/窗口统计/f0/线+直方图
type_convert/窗口统计/f0/count/仪表盘
14%
rankinglist_with_group/窗口统计/1.0/表格
f1 count(f2) sum(f3) mean
-2.0 3 0.31916795223962435 0.8680550
type_convert/累计统计/f0/count/线图
运行开始时间
17:19:45
17:22:30
17:25:15
17:28:00

Current Flink ML Library

FLINK FORWARD

- **Supervised Learning**
  - SVM
  - Multiple linear regression
  - Optimization Framework
- **Unsupervised Learning**
  - k-Nearest neighbors join
- **Data Preprocessing**
  - Polynomial Features
  - Standard Scaler
  - MinMax Scaler
- **Recommendation**
  - Alternating Least Squares (ALS)
- **Outlier Selection**
  - Stochastic Outlier Selection (SOS)
- **Utilities**
  - Distance Metrics
  - Cross Validation



Alink Supported Algorithms(1/4)

FLINK FORWARD

- 回归(Regression)

  Multi-Linear Regression, Lasso Regression, Ridge Regression, SVM Regression, Stepwise Linear Regression, Cart, GBDT, Random Forest Regression
- 分类(Classification)

  Logistic Regression, Supported Vector Machine(SVM), Perceptron, Naive Bayes, K-Nearest Neighbor, Tradaboost, Random Forest, ID3, Cart, C45
- 聚类(Clustering)

  KMeans, KModes, DBSCan, AGNES, PIC
- 深度学习(Deep Learning)

  TensorFlow Prediction and Training
- 在线学习(Online Learning)

  FTRL，KMeans， Perceptron， Passive Aggressive (PA)， PA-I， PA-II
- 评估(Evaluation)

  EvalClassification, EvalClustering, EvalRegression

Alink Supported Algorithms(2/4)

FLINK FORWARD

- 数据处理(Data Processing)

  Random Sampling, Stratified Sampling,

  Normalization, Standardization, Fill Missing Value, Type Conversion

  KvToTensor,TableToTensor,TensorToTable,TensorFunction,TensorToTuples

  Velocity Variable, Network traffic indicator, TensorExpandDim, Append ID

  Split a single column into multiple collumns, Select the Column After splitting,

  Extract Json Values, Single Column into Multiple, Multiple Columns into Single

  SqlCmd, As, Select, UnionAll, Where, GroupBy, Distinct,

  Intersect, Join, Minus, Orderby

  Multi-Stream merge, LatestJoin, Lookup

- 特征工程(Feature Engineering)

  One-Hot Coding, Feature Scale Transformation, Feature Anomaly Smoothing, Linear Model Feature importance Analysis

Alink Supported Algorithms(3/4)

FLINK FORWARD

- 基本统计(Basic Statistics)
  Window Statistics, Full Table Statistics, Grouped Window Statistics
  Count, Sum, Mean, Maximum, Minimum, Number of missing values, variance, standard deviation, Standard Error, Kurtosis, Skewness, etc.
  Largest k Values, Smallest k Values

- 变量关系(Variable Relationship)
  Covariance, Correlation Coefficient, Correspondence Analysis, Cross Table, Multicollinearity

- 数据分布(Data Distribution)
  Percentile, Frequency, histogram, PDF, CDF, Empirical PDF, Empirical CDF,
  P-P Plot, Lorenz Curve

- 假设检验(Hypothetical Test)
  T-Test, chi2 Test, AD Test, KS Test

- 数据降维(Reduction)
  Principal Component Analysis(PCA)， tSNE

- 时间序列(Time Series)
  ARIMA, Garch, ArimaGarch

Alink Supported Algorithms(4/4)

FLINK FORWARD

- 异常检测(Outlier Selection)
  SOS, K-Sigma, AVF, Boxplot, AGD, One Class SVM, SMA, EWMA, CDM, G Test, UriNumberDetection, GroupDetection, GroupMFIDetection, BigGraphGeneration

- 推荐算法(Recommendation)
  ALS, Simrank, FM, ItemCF

- 文本分析(Text Analysis)
  Word Count, Word Segmentation, Stop Word Filtering, Tokenizer, New Word Recognition, TF-IDF, Text Feature Generation
  Word2Vec, Text Sensitive Number Capture, Bank Card Information Parsing, ID Card Information Parsing, Word Sequence to ID sequence, String Similarity, Semantic Vector Distance, SimHash

- 图算法(Graph)
  Single Source Shortest Path, Community Detection, Label Propagation, PageRank, HITS, Tree Depth, Connected Graph, Modularity, K-Core, Triangle Count, Second-Degree Neighbor Lookup

Demo for Statistics and Visualization

FLINK FORWARD

- IJCAI-17 Dataset
  - https://tianchi.aliyun.com/datalab/index.htm
  - Trading amounts and locations of Alipay users
  - 19.6 million users, 67 million trades

	user_id	province	datetime	pay	score	comment_cnt	shop_level	catalog
0	15476608	广东	20160916 19:09:00	20	1	1	high	food
1	4981258	北京	20160311 14:03:00	6	4	13	high	food
2	6268699	天津	20160713 13:07:00	12	3	6	high	food
3	1604996	上海	20160222 19:02:00	13	2	2	medium	food
4	4981258	北京	20160311 14:03:00	6	4	13	high	food
5	2952114	江西	20151010 10:10:00	16	2	3	high	food
6	20086010	天津	20160904 19:09:00	20	4	5	low	mall
7	12392093	北京	20160722 17:07:00	19	2	2	low	food
8	17337606	江苏	20161008 18:10:00	15			low	supermarket
9	668310	浙江	20160904 17:09:00	17	4	0	low	supermarket


首页
实验
数据源
组件
设置
csv
stream组件
数据源(stream)
数据导出(stream)
数据处理
统计分析
机器学习
特征工程
异常检测
推荐算法
文本
聚类算法
深度学习
batch组件
数据源(batch)
数据导出(batch)
数据处理
统计分析
机器学习
特征工程
图算法
推荐算法
文本
异常检测
矩阵分解
深度学习
排行榜
运行
部署
可视化
查看日志
停止
实验属性
创建日期 2018-12-10 21:52:45
创建者 036769
运行实验的出错重试次数 Number类型
3
实验名称
排行榜
实验描述
无描述信息

Classification Demo

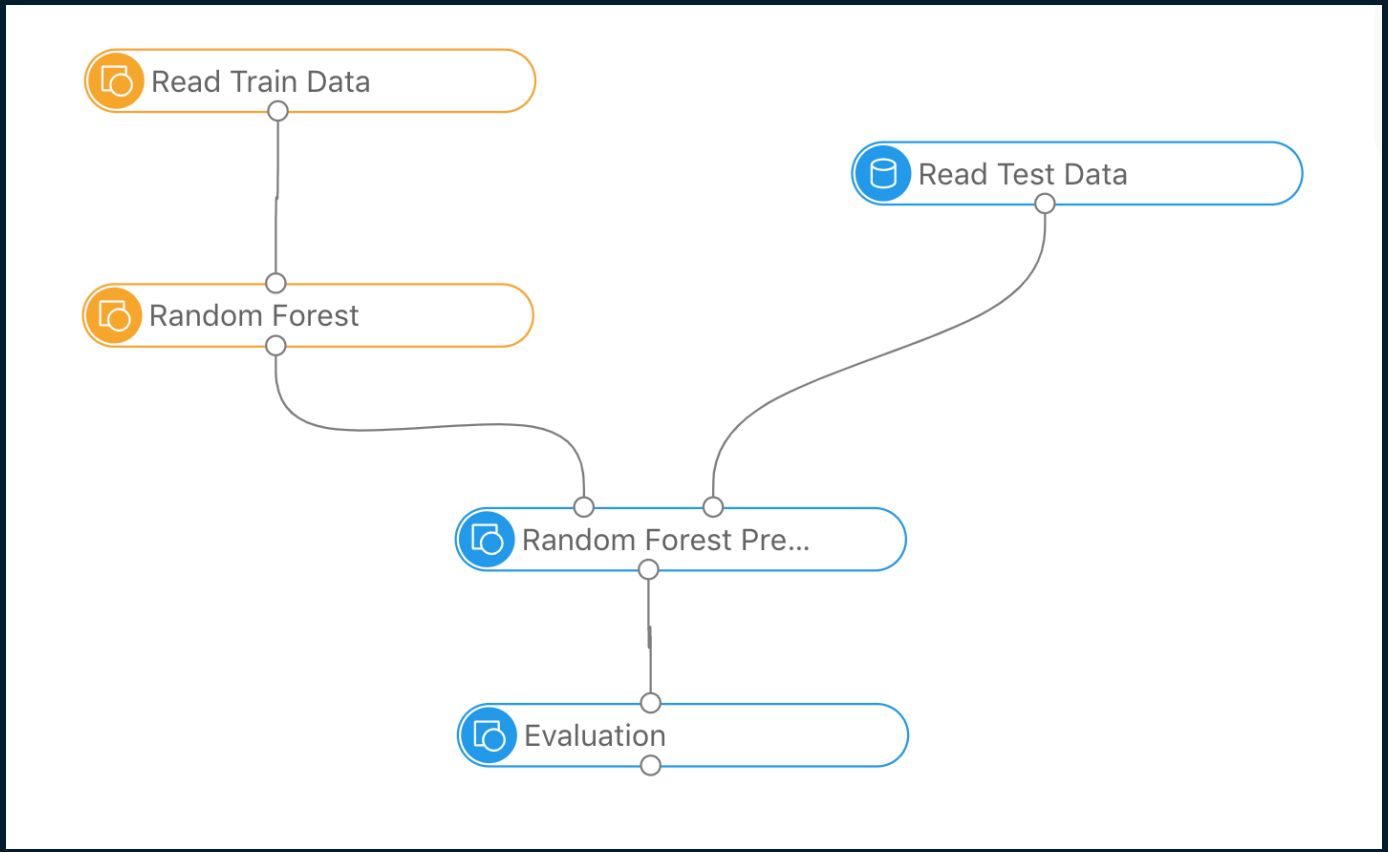
FLINK FORWARD

DataSet

- https://archive.ics.uci.edu/ml/datasets/adult
- Predict whether income exceeds $50K/yr based on census data.
- 48842 instances, 6 continuous attributes, 8 discrete attributes.

age	workclass	fnlwgt	education	educatio...	marital_...	occupation	relations...	race	sex	capital_g...	capital_l...	hours_p...	native_c...	labels
25	Private	226802	11th	7	Never-...	Machin...	Own-ch...	Black	Male	0	0	40	United-...	<=50K
38	Private	89814	HS-grad	9	Married...	Farming...	Husband	White	Male	0	0	50	United-...	<=50K
28	Local-g...	336951	Assoc-...	12	Married...	Protecti...	Husband	White	Male	0	0	40	United-...	>50K
44	Private	160323	Some-c...	10	Married...	Machin...	Husband	Black	Male	7688	0	40	United-...	>50K
18		103497	Some-c...	10	Never-...		Own-ch...	White	Female	0	0	30	United-...	<=50K
34	Private	198693	10th	6	Never-...	Other-s...	Not-in-...	White	Male	0	0	30	United-...	<=50K
29		227026	HS-grad	9	Never-...		Unmarri...	Black	Male	0	0	40	United-...	<=50K
63	Self-em...	104626	Prof-sc...	15	Married...	Prof-sp...	Husband	White	Male	3103	0	32	United-...	>50K
24	Private	369667	Some-c...	10	Never-...	Other-s...	Unmarri...	White	Female	0	0	40	United-...	<=50K
55	Private	104996	7th-8th	4	Married...	Craft-re...	Husband	White	Male	0	0	10	United-...	<=50K
65	Private	184454	HS-grad	9	Married...	Machin...	Husband	White	Male	6418	0	40	United-...	>50K
36	Federal...	212465	Bachelors	13	Married...	Adm-cl...	Husband	White	Male	0	0	40	United-...	<=50K
26	Private	82091	HS-grad	9	Never-...	Adm-cl...	Not-in-...	White	Female	0	0	39	United-...	<=50K
58		299831	HS-grad	9	Married...		Husband	White	Male	0	0	35	United-...	<=50K
48	Private	279724	HS-grad	9	Married...	Machin...	Husband	White	Male	3103	0	48	United-...	>50K
43	Private	346189	Masters	14	Married...	Exec-m...	Husband	White	Male	0	0	50	United-...	>50K
20	State-gov	444554	Some-c...	10	Never-...	Other-s...	Own-ch...	White	Male	0	0	25	United-...	<=50K
43	Private	128354	HS-grad	9	Married...	Adm-cl...	Wife	White	Female	0	0	30	United-...	<=50K
37	Private	60548	HS-grad	9	Widowed	Machin...	Unmarri...	White	Female	0	0	20	United-...	<=50K
40	Private	85019	Doctorate	16	Married...	Prof-sp...	Husband	Asian-P...	Male	0	0	45		>50K

Classification Demo



Classification Demo

Alink查看随机森林模型信息

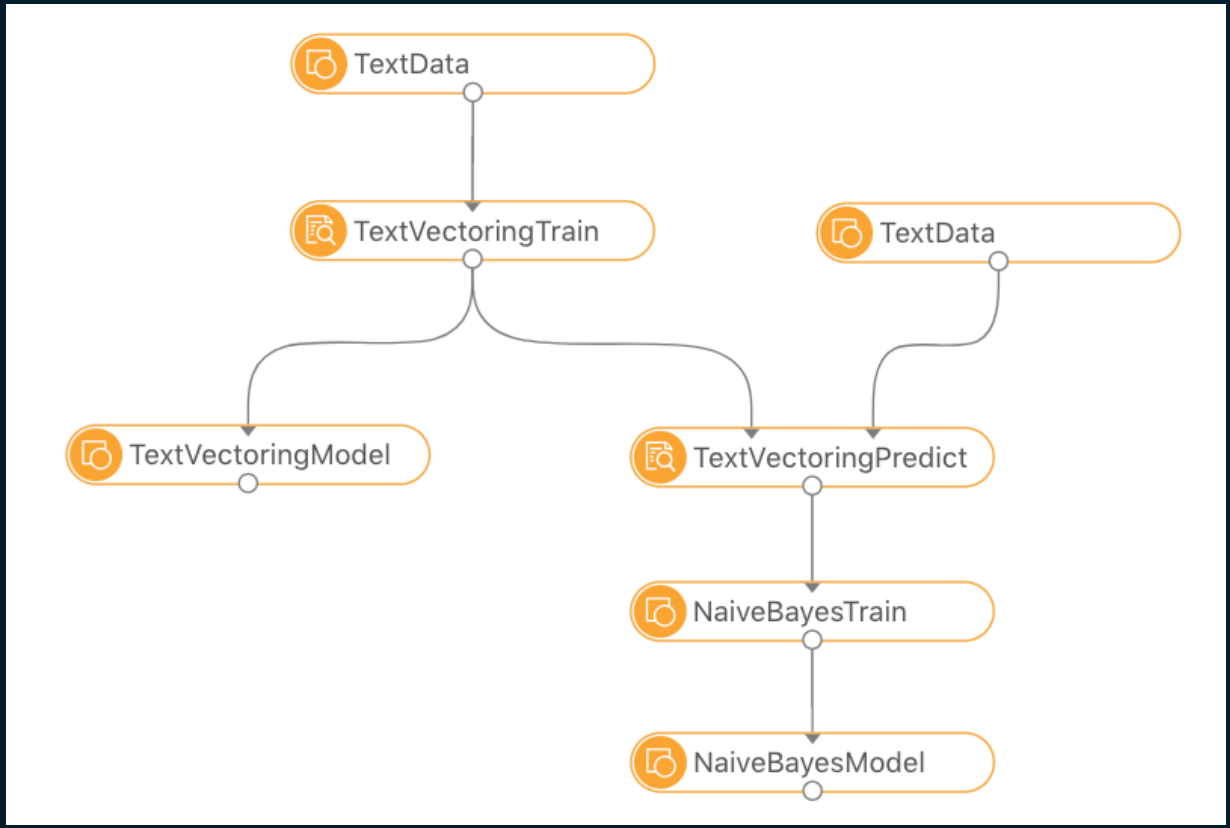
示例：文本分类(Text Classification)

FLINK FORWARD

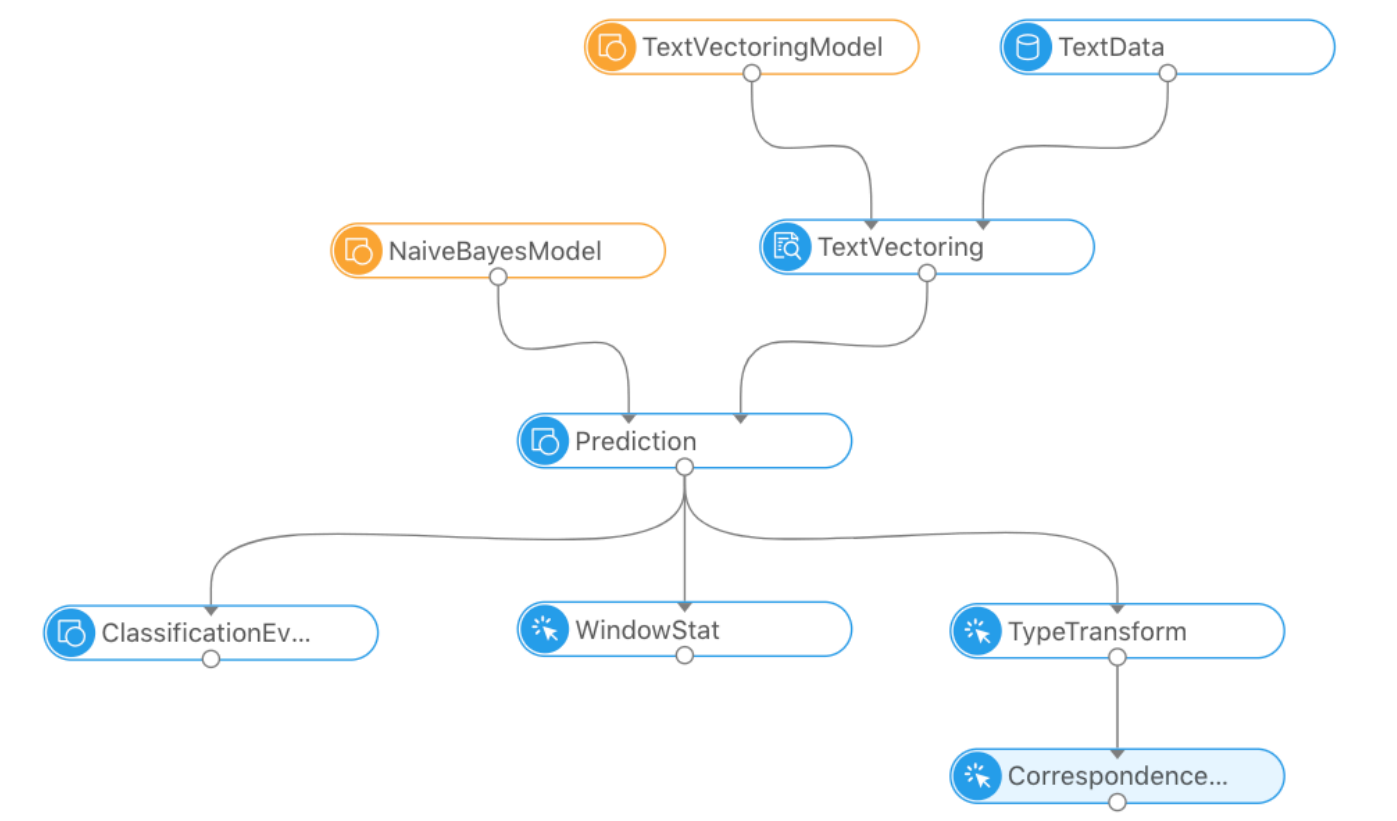
- Dataset
  - http://jmcauley.ucsd.edu/data/amazon/
  - 142.8 million product reviews with rating value (1 ~ 5)
  - Task: From review content, predict rating value

text	rating
One of her best, if you like mysteriesShe gets better and better!	5
I thought it was a bit drawn out for me a wasn't a page turner.	4
Awful one of her worst books waste of money boring .	1
Loved the story but she is one of my favorite authors..	5
This novel kept me on my mental toes. great read	5
am getting back to this author after a long time away. Good start!	3

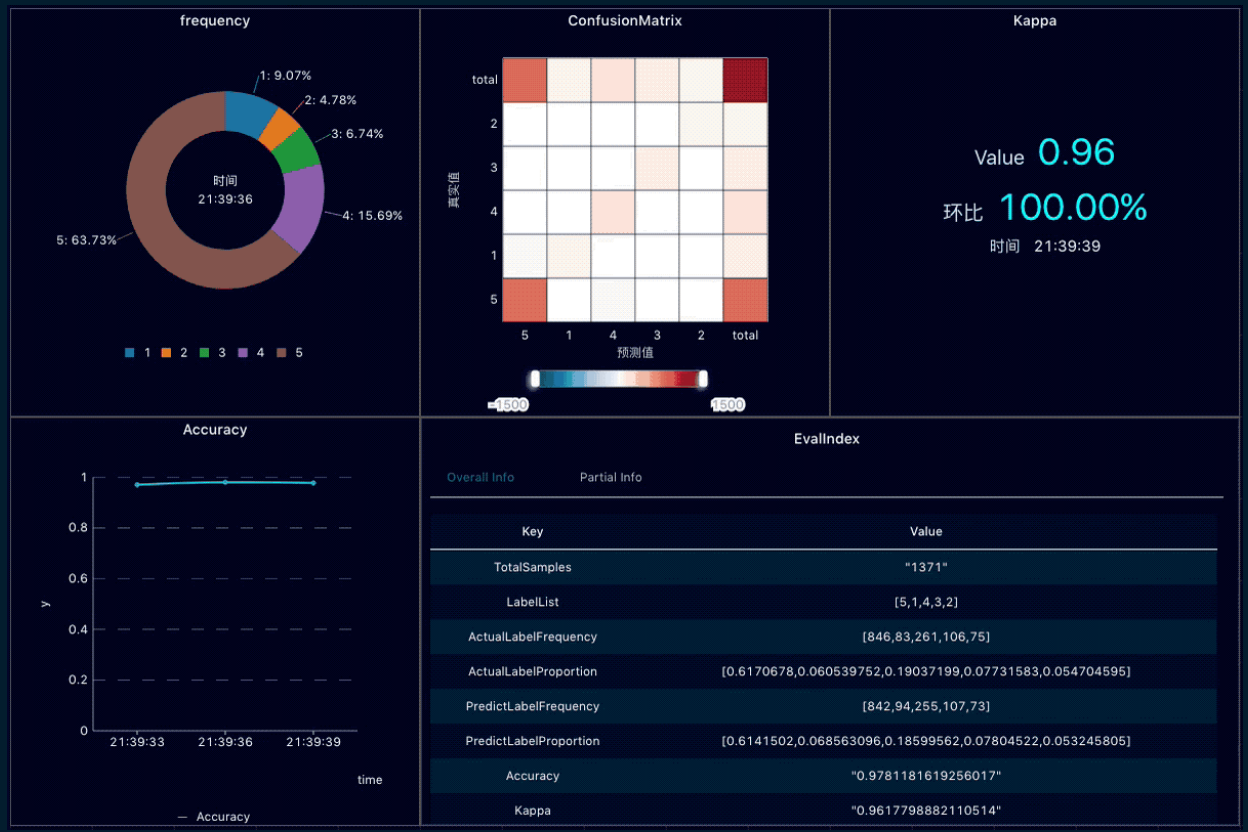
示例：文本分类(Text Classification)



示例：文本分类(Text Classification)



示例：文本分类(Text Classification)



FM 算法

因子分解机 (Factorization Machine, FM) 是由 Steffen Rendle 2010 年提出的一种基于矩阵分解的机器学习算法，常用于大规模的CTR预估。

Factorization Machine (FM) is a matrix-based machine learning algorithm proposed by Steffen Rendle in 2010. It is commonly used in large-scale CTR estimation.

- 线性模型(Linear Model)

$$y = w_0 + \sum_{i=1}^{n} w_i x_i$$

- 二阶多项式模型(Quadratic Polynomial Model)

$$y = w_0 + \sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} w_{ij} x_i x_j$$

- 因子分解机模型(FM Model)

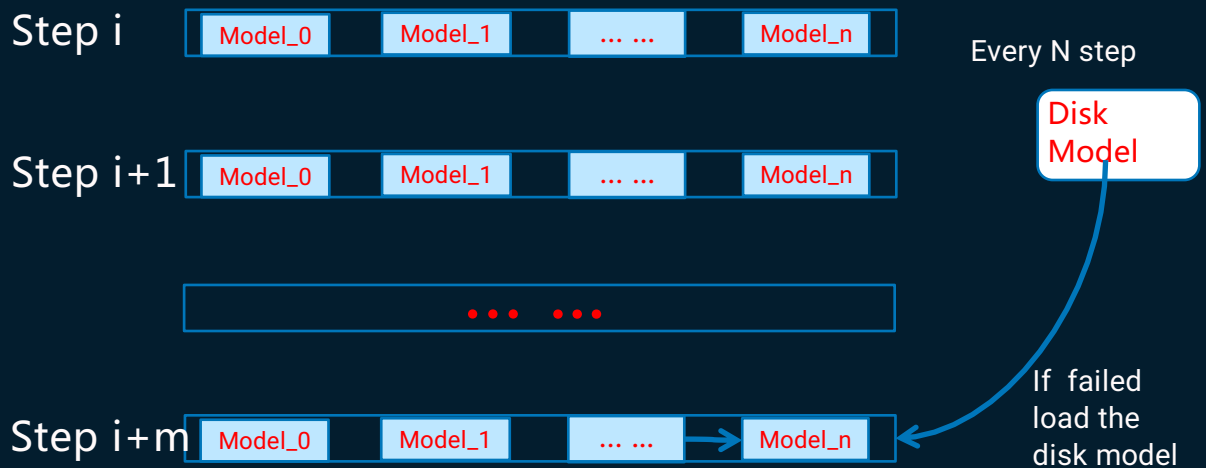
$$y = w_0 + \sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} < v_i, v_j > x_i x_j$$

ALGORITHM 1: Stochastic Gradient Descent (SGD)

Input: Training data S, regularization parameters λ, learning rate η, initialization σ
Output: Model parameters $\Theta = (w_0, \mathbf{w}, \mathbf{V})$
$w_0 \leftarrow 0$; $\mathbf{w} \leftarrow (0, \ldots, 0)$; $\mathbf{V} \sim \mathcal{N}(0, \sigma)$;
repeat
  for $(\mathbf{x}, y) \in S$ **do**
    $w_0 \leftarrow w_0 - \eta \left(\frac{\partial}{\partial w_0} l(\hat{y}(\mathbf{x}|\Theta), y) + 2\lambda^0 w_0 \right)$;
    for $i \in \{1, \ldots, p\} \wedge x_i \neq 0$ **do**
      $w_i \leftarrow w_i - \eta \left(\frac{\partial}{\partial w_i} l(\hat{y}(\mathbf{x}|\Theta), y) + 2\lambda^w_{\pi(i)} w_i \right)$;
      for $f \in \{1, \ldots, k\}$ **do**
        $v_{i,f} \leftarrow v_{i,f} - \eta \left(\frac{\partial}{\partial v_{i,f}} l(\hat{y}(\mathbf{x}|\Theta), y) + 2\lambda^v_{f,\pi(i)} v_{i,f} \right)$;
      end
    end
  end
until *stopping criterion is met*;

Alink FM 算法特色

- 分布式模型存储
- 支持超大规模（百亿）特征
- 支持failover功能

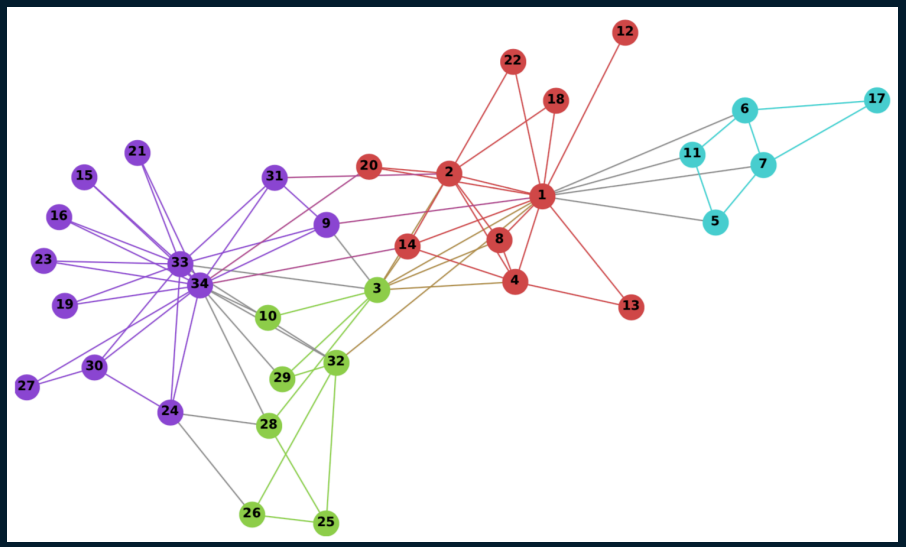


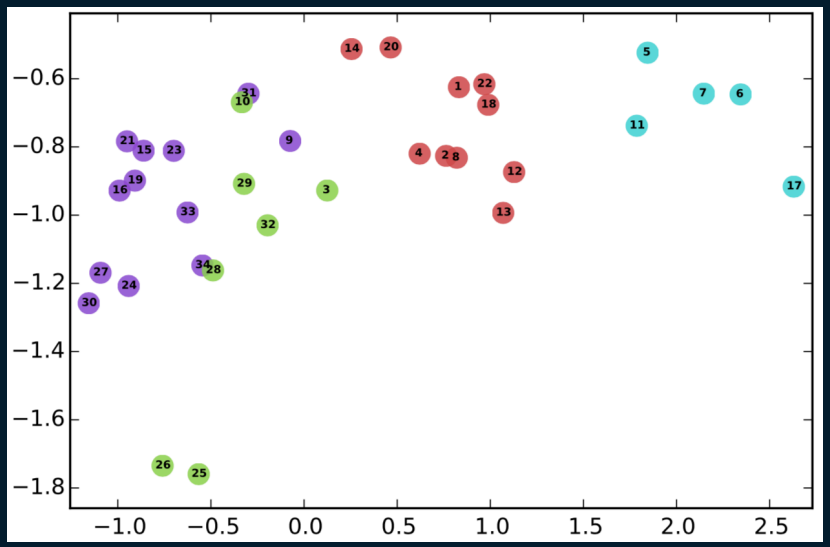
规模测试

样本规模	特征规模	训练时间
500亿	1千万特征，因子数100（十亿参数）	7小时9分
500亿	1亿特征，因子数100（百亿参数）	12小时2分钟

Graph Embedding

FLINK FORWARD





Graph Embedding

FLINK FORWARD

- Huge Graph from Industry

  Facebook: ~2 billion active users

  Wechat: ~1 billion active users

  Amazon: 400M active users, 400M products

  Taobao: 500M active users, 800M products

- Alink supported algorithms with billions of nodes

  1、DeepWalk

  2、Node2Vec

  3、MetaPath2Vec


THANKS