MaxCompute 对开源系统的支持与融合

-

高级专家 艺卓

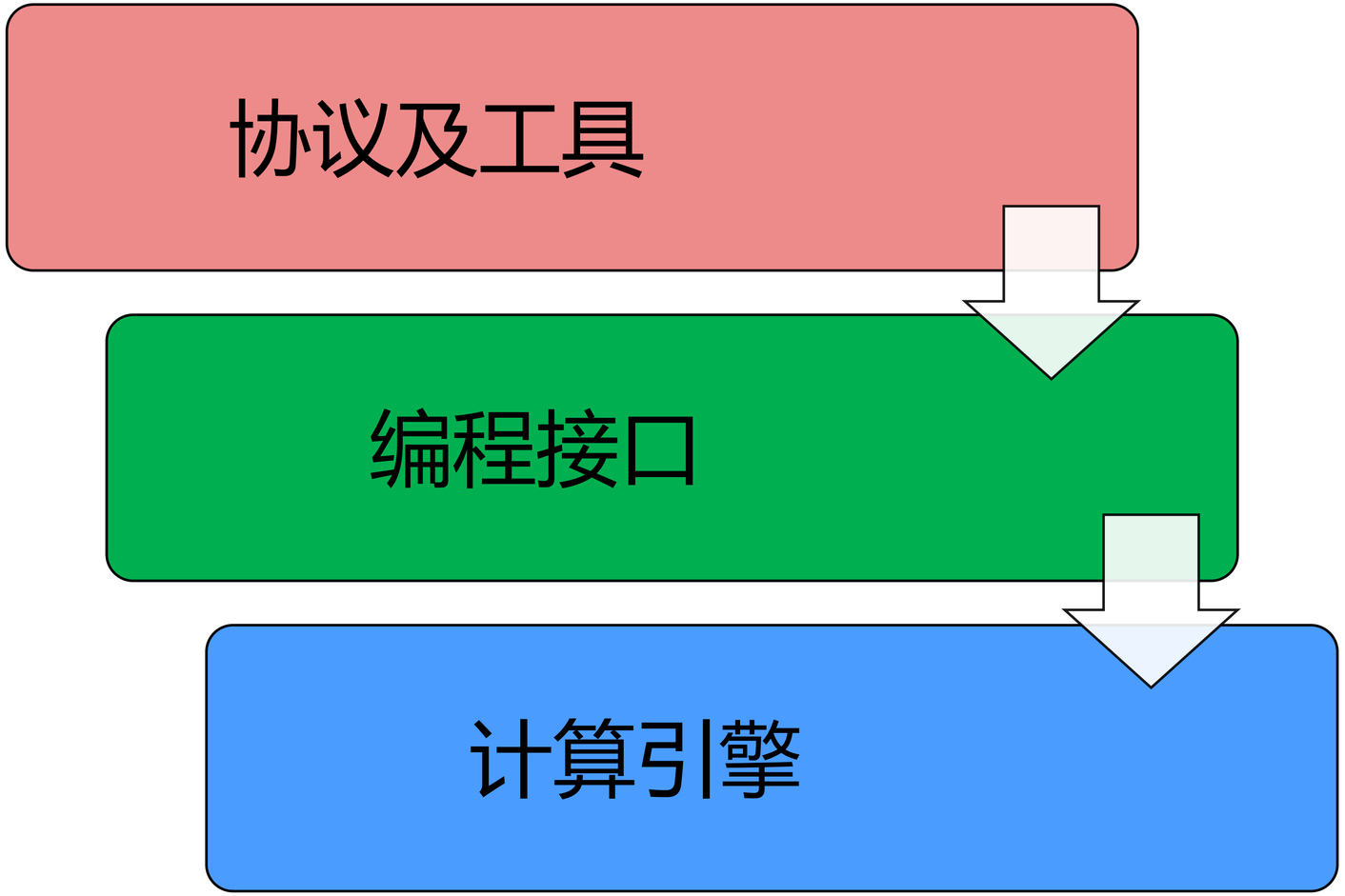
2017

MaxCompute 和开源

- 2017
  阿里自研的一站式大数据解决方案
  MaxCompute 融合 OpenSource

- 2016
  阿里自研的大数据计算平台
  MaxCompute 支持 OpenSource

开源的层次

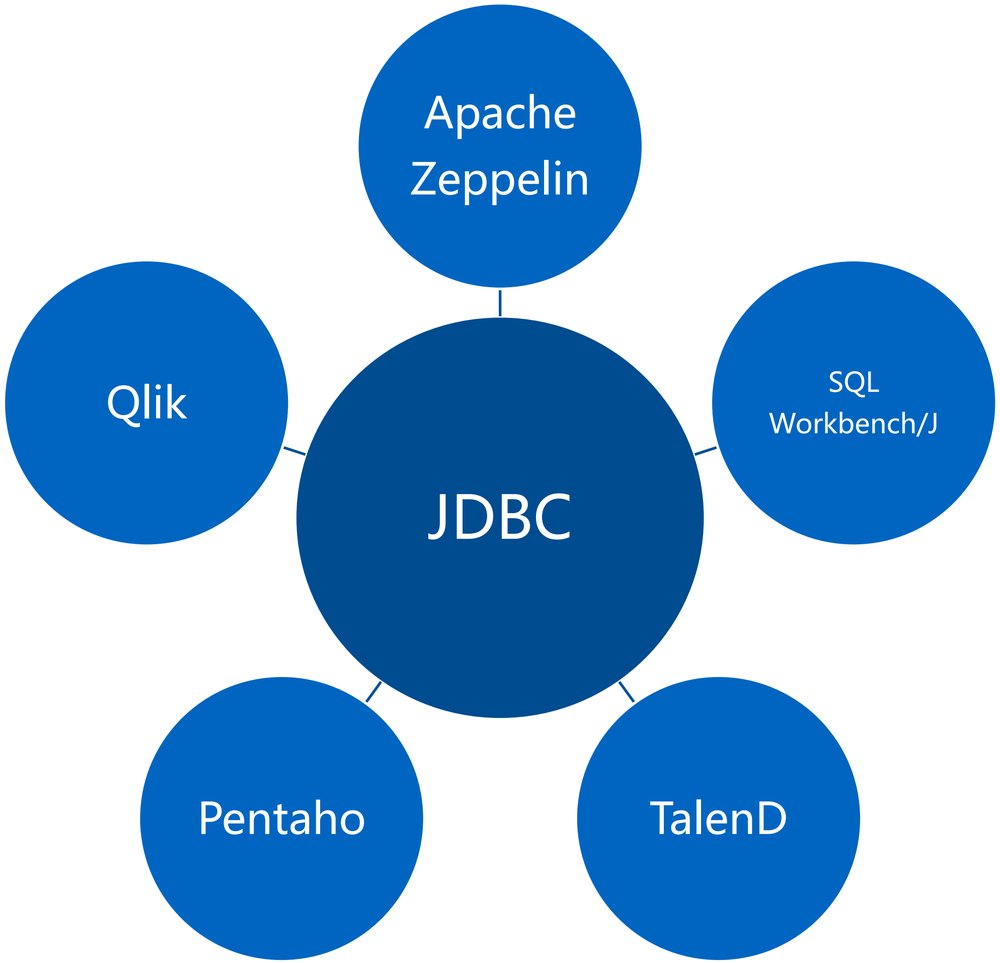


开源的协议及工具

- JDBC
  对接已有软件
  提供标准 JDBC 编程接口

- Hive Proxy
  提供 Hive Thrift 协议兼容接口
  对接 Hive 社区已有工具

- ETL 工具



开源的编程接口

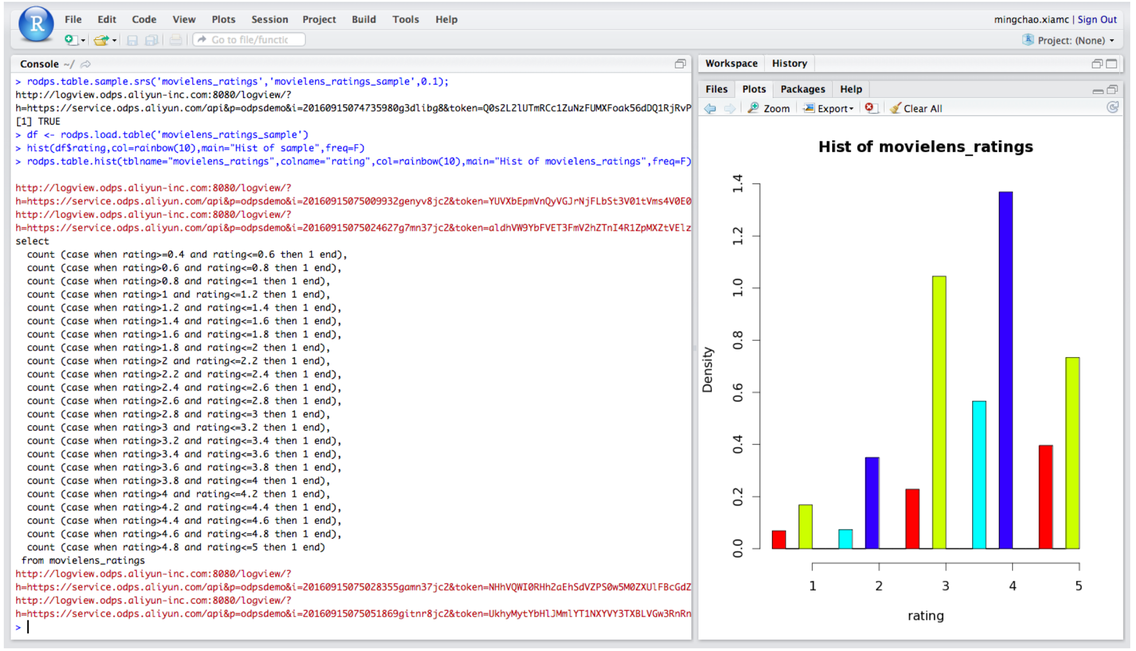
- MaxCompute SQL 2.0

- 兼容 Hive 类型系统
- 兼容 Hive 内建函数
- 兼容 Hive 用户定义函数
- 支持 External Table
- CTE/INSERT/JOIN/UNION 等语句增强

开源的编程接口

- RODPS

- 直接操作 MaxCompute 数据
- 支持 R 生态已有工具及代码库

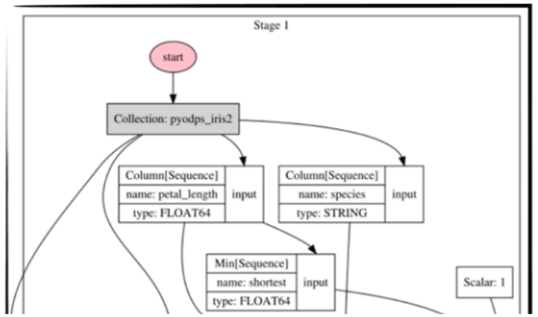


开源的编程接口

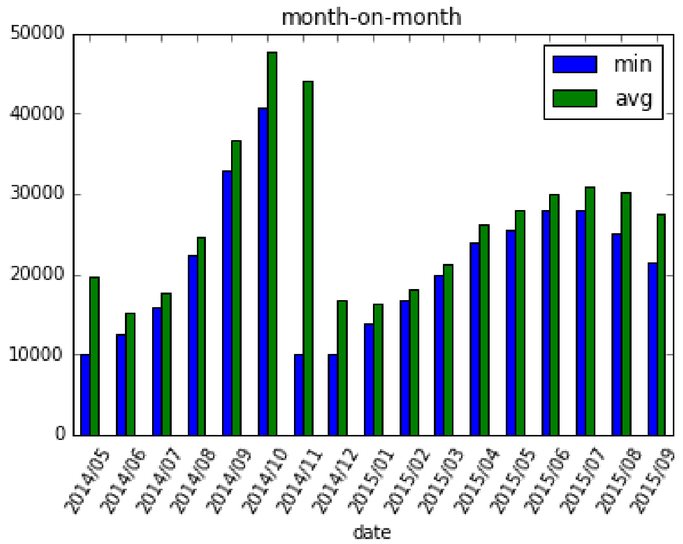
- PyODPS

- 高度兼容 Pandas DataFrame
- 直接赋予社区代码大数据计算能力
- 对接 Jupyter Notebook 等社区生态

```
df = iris.groupby('species').agg(
    shortest=iris.petal_length.min())
df = df['species', df.shortest + 1]
df.visualize()
```



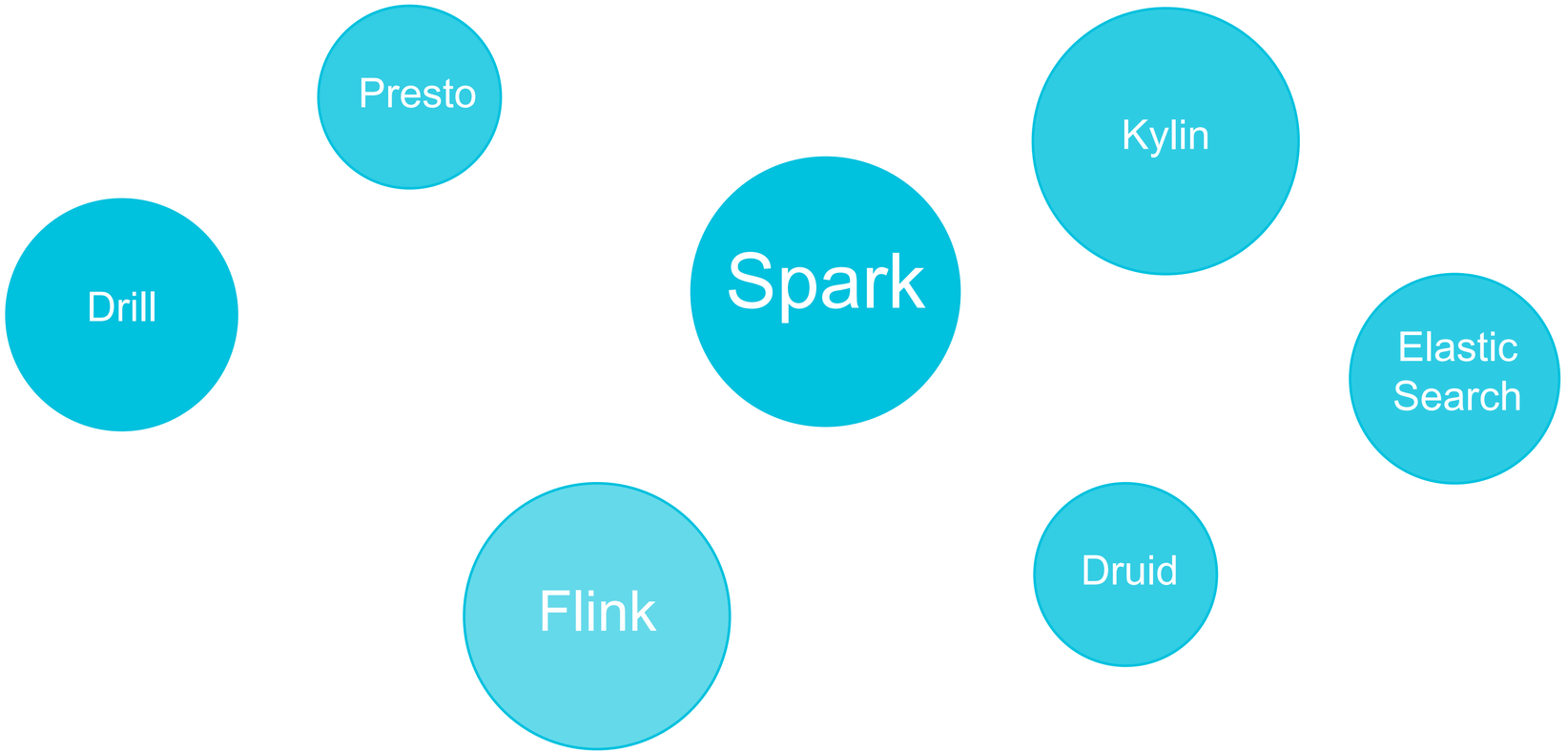

```
In [4]: # 我们按日期排列，并绘制一张柱状图，来看看环比的情况。

        df.sort('d').plot(x='d', kind='bar', xlabel='date', rot=60,

Out[4]: <matplotlib.axes._subplots.AxesSubplot at 0x7fd91329d0d0>
```



开源的计算引擎

优点：
- 快速搭建
- 学习资料
- 保护代码投资

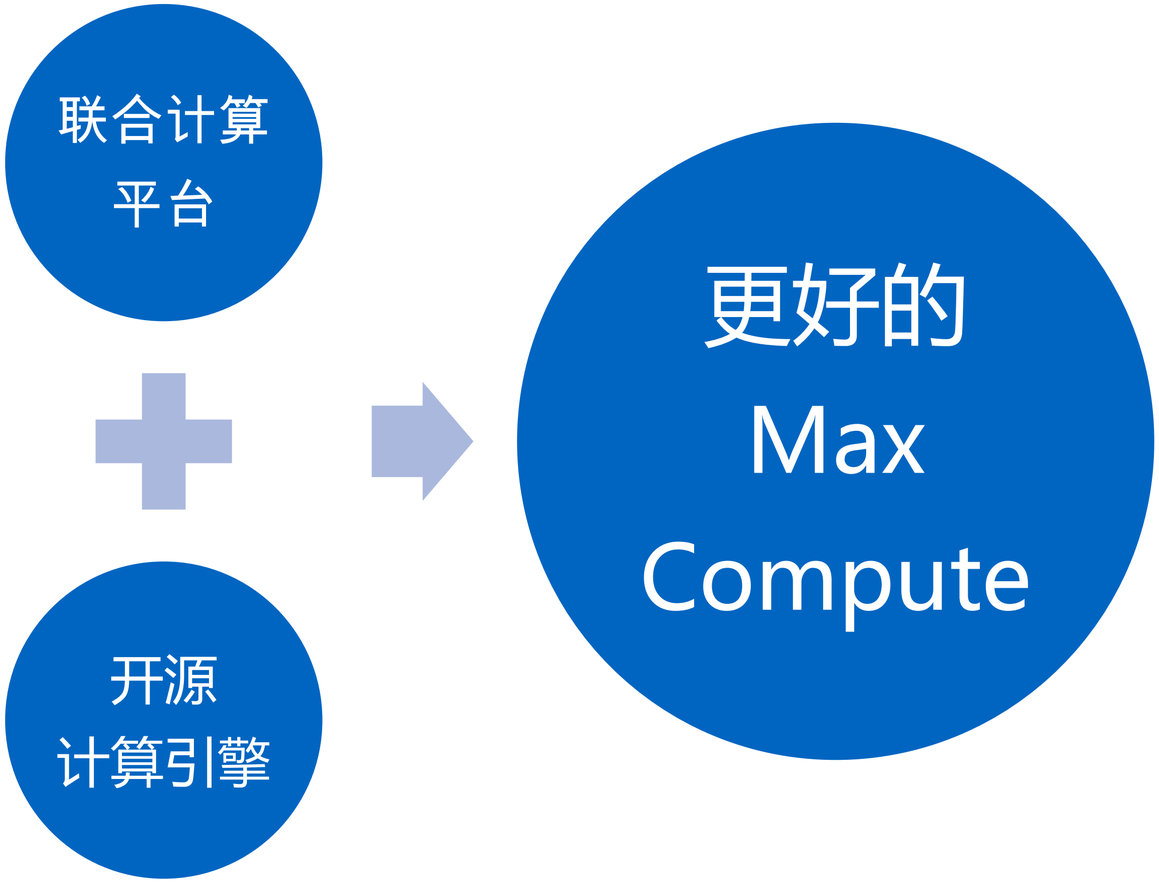
缺点：
- 数据分散
- 数据一致问题
- 资源效率

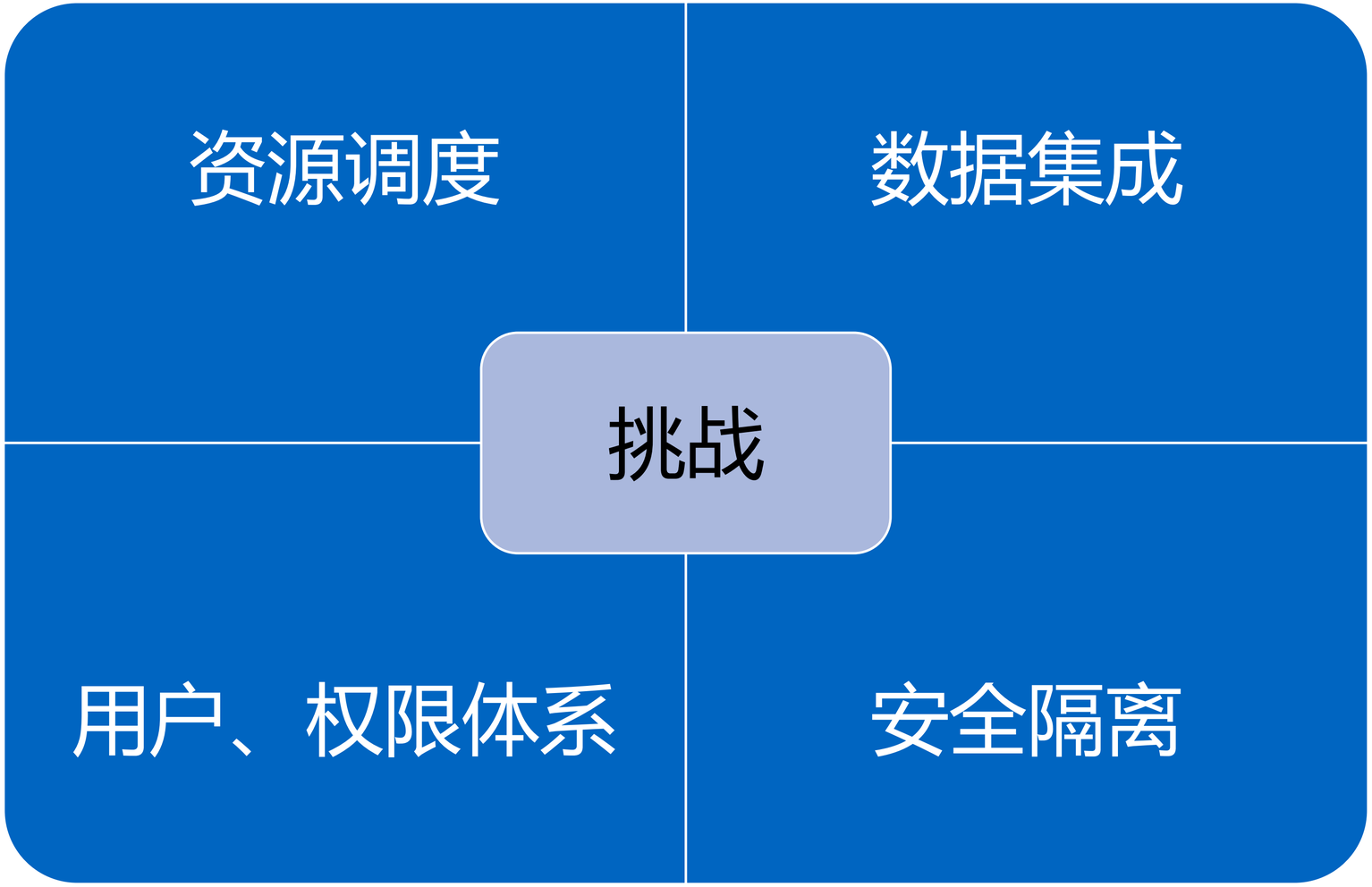


开源的计算引擎

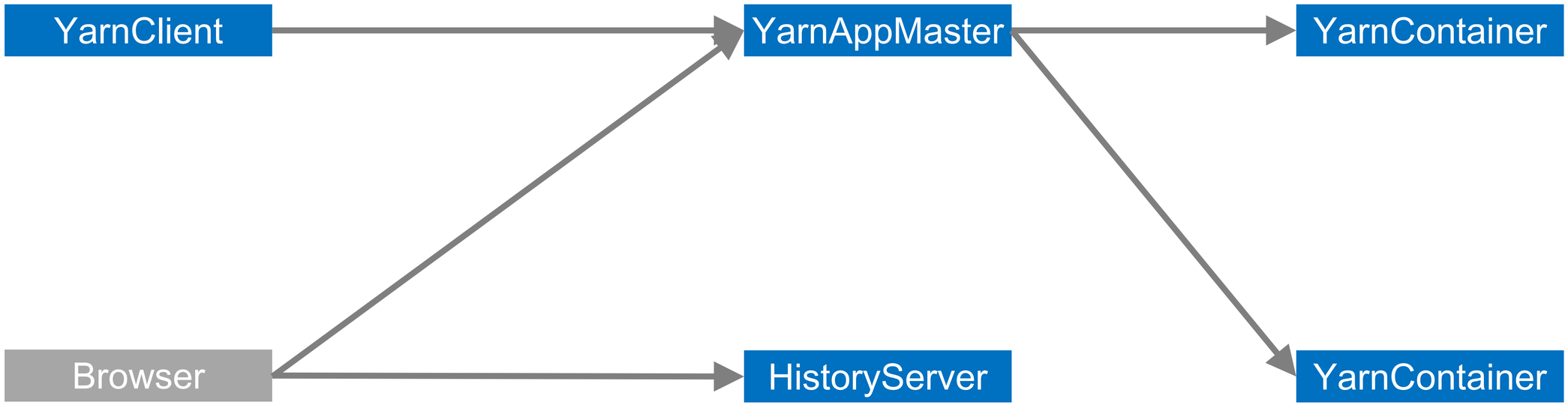
- 保持自研优势
  拥抱开源生态

- 数据存储统一
- 资源调度统一
- 安全控制统一

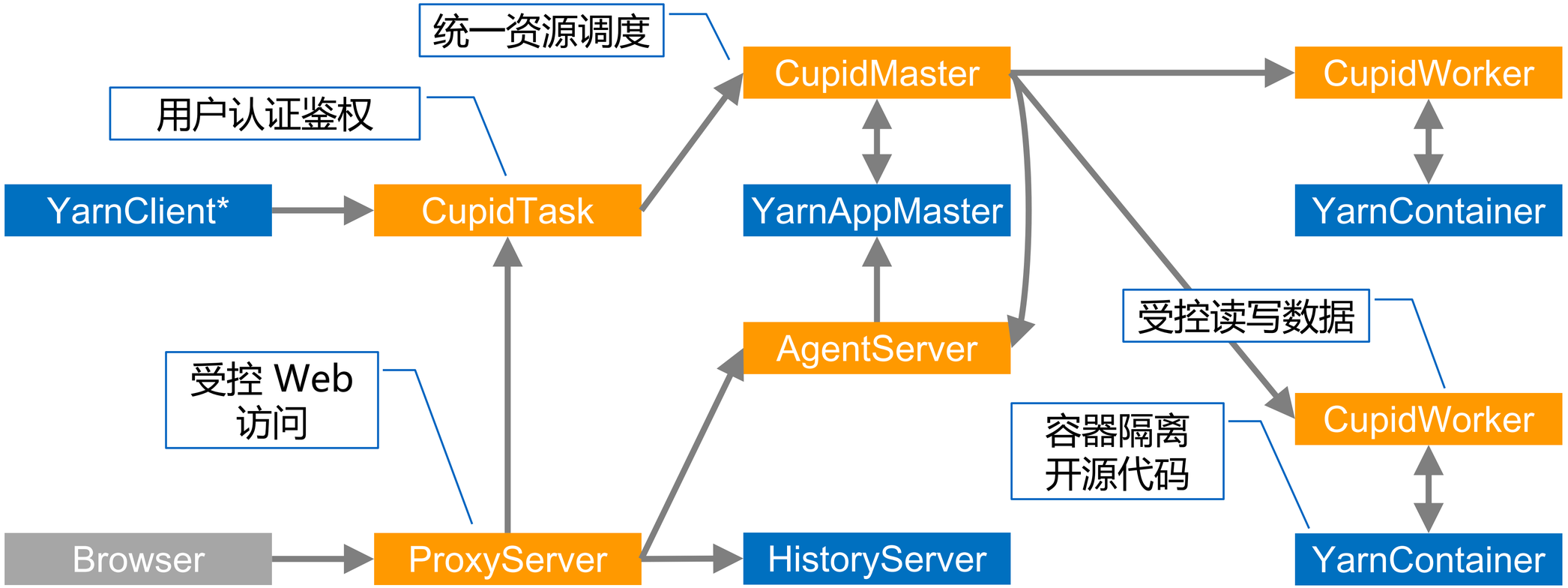



资源调度
数据集成
挑战
用户、权限体系
安全隔离

联合计算平台架构


YarnClient
YarnAppMaster
YarnContainer
Browser
HistoryServer
YarnContainer

联合计算平台架构


统一资源调度
用户认证鉴权
YarnClient*
CupidTask
CupidMaster
YarnAppMaster
CupidWorker
YarnContainer
受控读写数据
AgentServer
受控 Web
访问
容器隔离
开源代码
CupidWorker
Browser
ProxyServer
HistoryServer
YarnContainer

Spark Jobs Stages Storage Environment Executors com.aliyun.odps.sp

Spark Jobs (?)

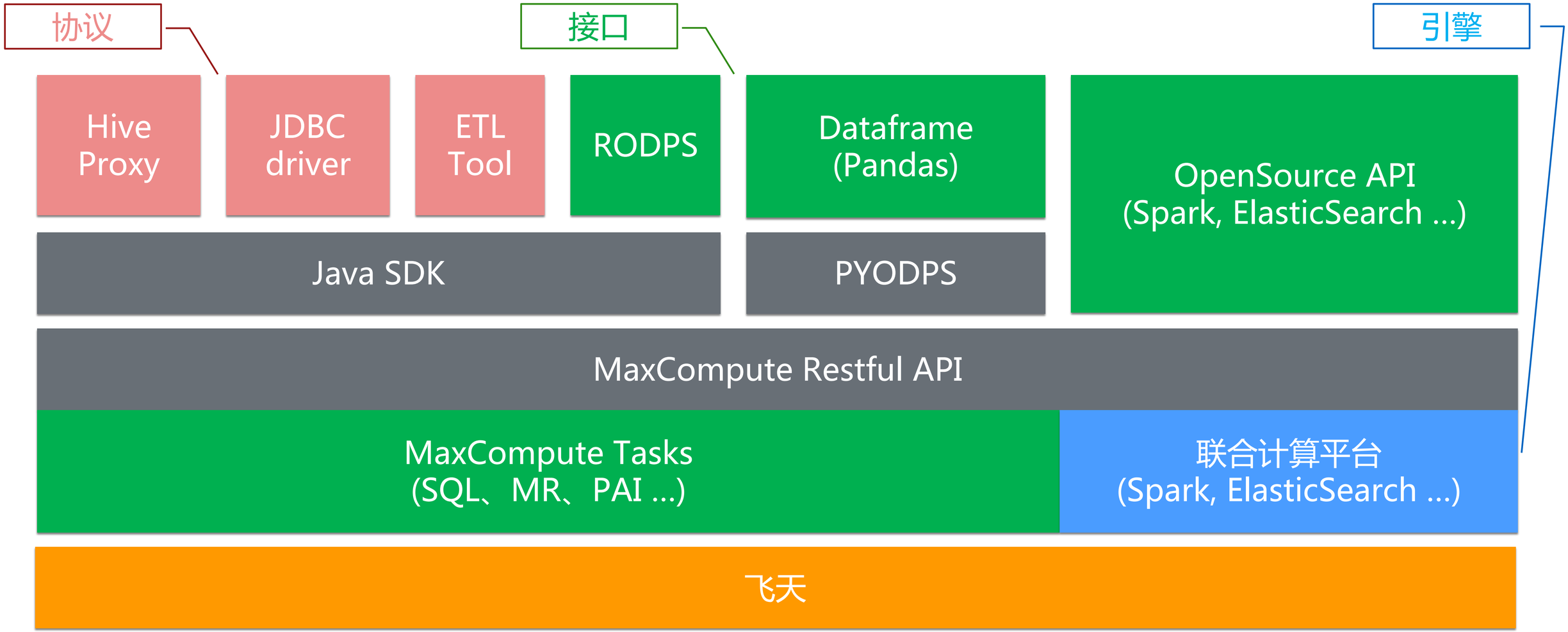
Scheduling Mode: FIFO
Completed Jobs: 3

▸ Event Timeline

Completed Jobs (3)

Job Id	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
2	count at Join.scala:21	2016/11/29 01:17:42	2 s	4/4	40/40
1	count at Join.scala:20	2016/11/29 01:17:42	0.5 s	3/3	30/30
0	count at Join.scala:19	2016/11/29 01:17:37	4 s	3/3	30/30

总结与展望：与开源融合的一站式大数据解决方案



THANK YOU

扫码参与云栖大会调查问卷，赢取大会限量纪念品