

Open Source AceCon

2021 智能云边开源峰会

AI x Cloud Native x Edge Computing

人工智能 × 云原生 × 边缘计算

MLSQL 开源云原生Data & AI平台

Luke Han | luke.han@kyligence.io

联合创始人 & CEO



Kyligence 公司介绍

自主开源技术，打造开源生态

Open Source AceCon
2021 智能云边开源峰会
AI x Cloud Native x Edge Computing
人工智能 × 云原生 × 边缘计算



- 全球领先的大数据 OLAP 领导者
- 中国首个 Apache 顶级开源项目
- 1500+ 全球生产用户



- 面向 AI 的 SQL 语言
- AI + Big Data 领域开源新秀
- 金融、互联网等行业应用案例

SQL为王，为数据分析师、数据科学家、数据工程师提供统一的全栈平台

Agenda

- 当前落地 Data + AI 所面临的痛点
- MLSQL是什么
- 如何使用MLSQL低成本落地 Data + AI

当前企业落地 Data + AI 所面临的痛点

6. 不同组件任务使用调度连接起来

我不是超人



Machine Learning with Scikit-Learn

1. 使用SQL做一些数据预处理



2. PySpark进一步处理数据



3. 使用机器学习库进行训练

4. 复杂的模型部署

5. 持续的迭代



一线人员的五大痛点

算法落地久



1

- 太多组件
- 需要复杂的调度，运维系统
- 维护成本高

2

- 数据需要在各个组件流转
- 格式，形态都不一致
- 存在多次落盘，产生大量临时数据

3

- 各个地方都需要权限管控
- 容易产生漏洞，且管理繁琐

4

- 组件太多，
- 资源难以利用均衡

5

- 使用者需要学习各个系统，
- 抬高了使用门槛

企业痛点（一）

ROI低



企业发现落地一个算法到具体的某个场景的成本远高于其带来的收益

企业痛点（二）

招募难

企业

我需要一打

人才池子太小

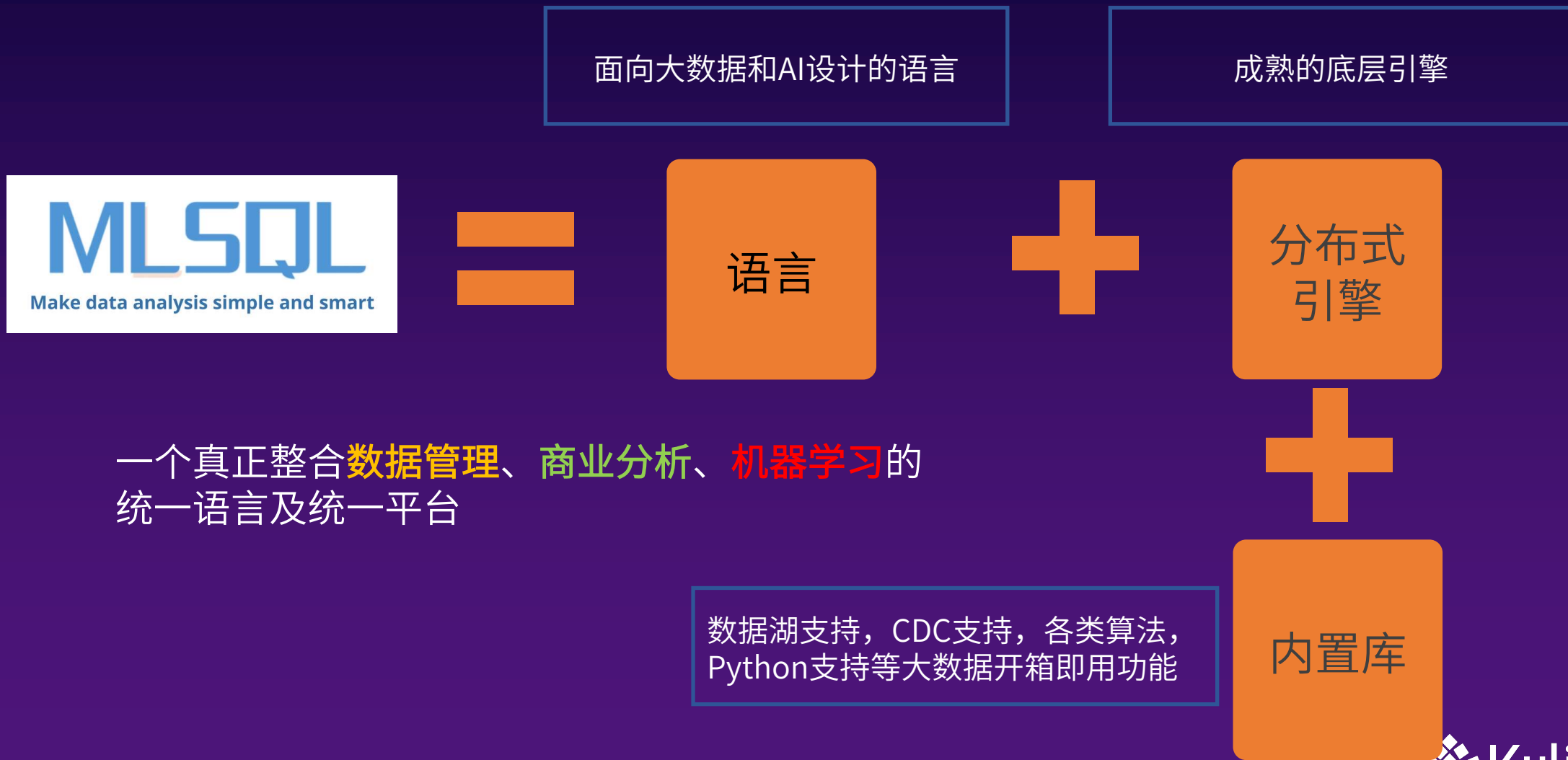


企业愿意花钱，但依然难以构建完整平台和招募到足够的研发和数据科学家

MLSQL 能够解决上述难题

那么 MLSQL 是什么?

MLSQL三大组成



MLSQL 实现从桌面到云端的覆盖



像使用 SAS 软件一样，在笔记本开箱即用，无需云端亦可工作。

一行配置使用云端
算力和存储：

`engine.url=https://m1:9003`



可充分利用云原生带来的算力和存储
突破本地限制

桌面版

The screenshot displays the AceCon desktop application interface. On the left is a file explorer (EXPLORER) showing the project structure for 'MLSQl-EXAMPLE-PROJECT'. The 'src' directory is expanded, showing 'analysis/example/cifar10' and 'common'. The 'UserBehavior.mlsqInb' file is selected. The main editor area shows the 'UserBehavior.mlsqInb' file with a title '使用MLSQl对1亿条淘宝用户行为数据分析'. The content includes a description of the CSV file, a link to the dataset, and a code snippet for loading the data. The terminal at the bottom shows the execution of the 'load csv' command.

EXPLORER

- OPEN EDITORS 1 UNSAVED
- MLSQl-EXAMPLE-PROJECT
 - __mlsql__
 - .result.tmp
 - .vscode
 - settings.json
 - 627f0c17-d0b1-4e1c-a63b-eade30ab1be1
 - data
 - docs
 - example-data
 - logs
 - spark-warehouse
 - src
 - analysis/example
 - cifar10
 - DistributeTFTraining.mlsqInb
 - ResizeImage.mlsqInb
 - ResizeImage.mlsqInb
 - UserBehavior.mlsqInb
 - common
 - CommandExample.mlsqInb
 - ExcelExample.mlsqInb
 - IfElseExample.mlsqInb
 - MySQLConnectExample.mlsqInb
 - PP.mlsqInb
 - PublicModuleIncludeExample.mlsqInb
 - PythonScriptExample.mlsqInb
 - SimpleMLExample.mlsqInb
 - .gitignore
 - .mlsql.config
 - cifar.tgz
 - README.md

Code Editor

src > analysis > example > UserBehavior.mlsqInb > 使用MLSQl对1亿条淘宝用户行为数据分析 > load csv. ./example-data

+ Code + Markdown ▶ Run All ≡ Clear Outputs ...

使用MLSQl对1亿条淘宝用户行为数据分析

一个CSV文件，时间区间为 2017-11-25 到 2017-12-03，总计 100,150,807 条记录，大小为 3.5 G。下载地址：

1. 阿里云：<https://tianchi.aliyun.com/dataset/dataDetail?dataId=649&userId=1#1>

也可参考 [bigdata_analyse](#) 获取数据。

下载后解压放到 `./example-data/custom-download` 目录下。

分析过程中，会有全量去重等比较消耗软件内存的操作，需要4G内存才能运行，请在 `.mlsql.config` 中添加如下配置：

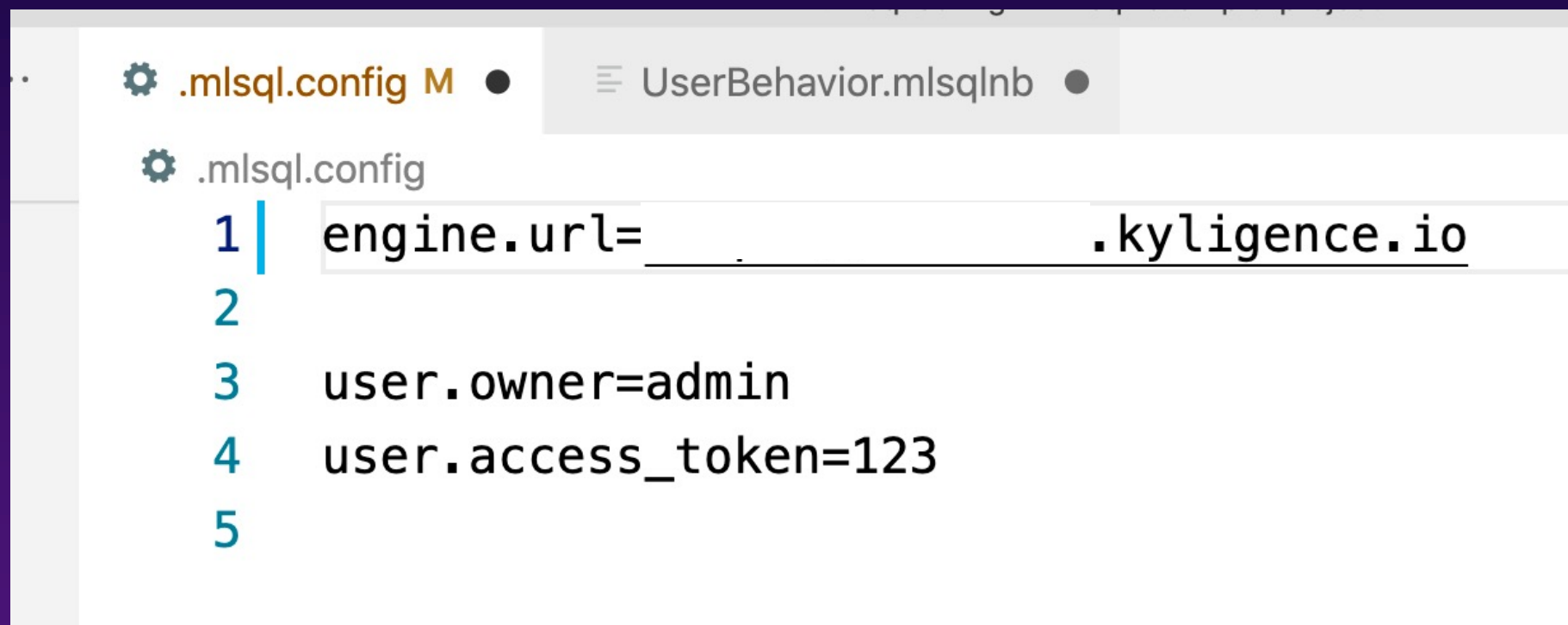
```
engine.memory=4048m
```

Terminal

```
1 -- 设置下Python环境。因为我们会用到Python做图形绘制
2 -- 用户如果需要可以进文件进行修改
3 include project.`./src/common/PyHeader.mlsqInb`;

1 load csv.`./example-data/custom-download/UserBehavior.csv`
2 where header="false"
3 as raw_user_behavior;
4
5 select cast(_c0 as long) as user_id,
6 cast(_c1 as long) as item_id,
7 cast(_c2 as long) as catagory_id,
8 _c3 as behavior_type,
9 cast(_c4 as long) as `timestamp`
10 from raw_user_behavior
11 as user_behavior;
```

桌面版+云端服务



```
.mlsql.config M • UserBehavior.mlsqlInb •  
• .mlsql.config  
1 | engine.url= .kyligence.io  
2  
3 user.owner=admin  
4 user.access_token=123  
5
```

MLSQ 命令行

TERMINAL SQL CONSOLE: MESSAGES PROBLEMS OUTPUT DEBUG CONSOLE

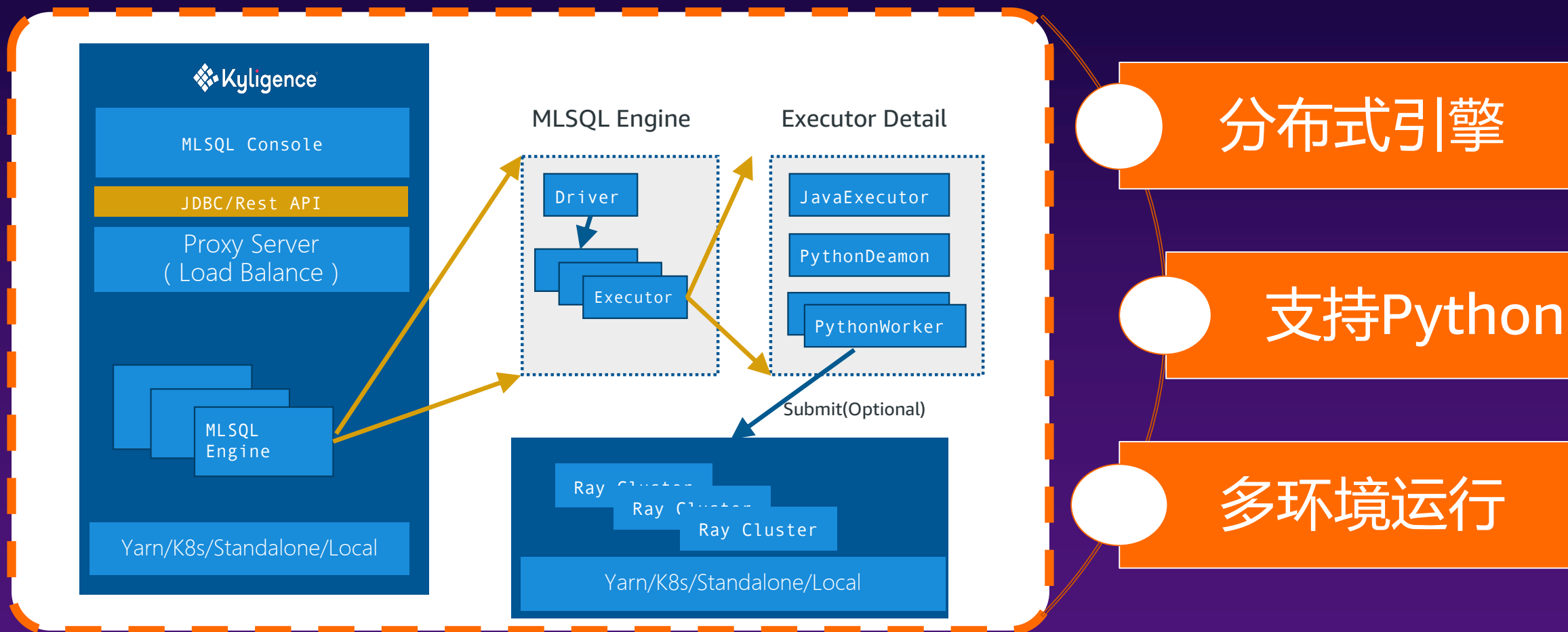
```
(dev) [w@me mlsq-example-project]$ export MLSQL_HOME=/Users/allwefantasy/projects/mlsql/src/mlsql-lang/mlsql-app_2.4-2.1.0-SNAPSHOT
(dev) [w@me mlsq-example-project]$ export PATH=${MSQL_HOME}/bin:$PATH
(dev) [w@me mlsq-example-project]$
(dev) [w@me mlsq-example-project]$ mlsq run ./src/common/mock_data.mlsq
2021/09/06 10:00:18.825018 mlsq[25369] <INFO>: [-xmx4048m -cp /Users/allwefantasy/projects/mlsql/src/mlsql-lang/mlsql-app_2.4-2.1.0-SNAPSHOT/main/*:/Users/allwefantasy/projects/mlsql/src/mlsql-lang/mlsql-app_2.4-2.1.0-SNAPSHOT/libs/*:/Users/allwefantasy/projects/mlsql/src/mlsql-lang/mlsql-app_2.4-2.1.0-SNAPSHOT/plugin/*:/Users/allwefantasy/projects/mlsql/src/mlsql-lang/mlsql-app_2.4-2.1.0-SNAPSHOT/spark/* streaming.core.StreamingApp -streaming.master local[*] -streaming.name MLSQL-desktop -streaming.rest false -streaming.thrift false -streaming.platform spark -streaming.spark.service false -streaming.job.cancel true -streaming.datalake.path ./data/ -streaming.driver.port 9003 -streaming.plugin.clznames tech.mlsq.plugins.ds.MLSQExcelApp,tech.mlsq.plugins.shell.app.MLSQLShell -streaming.platform_hooks tech.mlsq.runtime.SparkSubmitMLSQScriptRuntimeLifecycle -streaming.mlsq.script.path ./src/common/mock_data.mlsq -streaming.mlsq.script.owner admin -streaming.mlsq.scrip.jobName mlsq-cli]
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/Users/allwefantasy/projects/mlsql/src/mlsql-lang/mlsql-app_2.4-2.1.0-SNAPSHOT/main/streamingpro-mlsql-spark_2.4_2.11-2.1.0-SNAPSHOT.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/Users/allwefantasy/projects/mlsql/src/mlsql-lang/mlsql-app_2.4-2.1.0-SNAPSHOT/spark/slf4j-log4j12-1.7.16.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
log4j:WARN No such property [rollingPolicy] in org.apache.log4j.RollingFileAppender.
21/09/06 10:00:19 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
21/09/06 10:00:21 INFO MLSQStreamManager: Start streaming job monitor....
21/09/06 10:00:23 INFO SnapshotTimer: Scheduler MLSQ state every 3 seconds
21/09/06 10:00:24 INFO MLSQExcelApp: Load ds: tech.mlsq.plugins.ds.MLSQExcel
21/09/06 10:00:26 INFO JobManager: JobManager started with initialDelay=30 checkTimeInterval=5
21/09/06 10:00:26 INFO FileInputFormat: Total input paths to process : 1
21/09/06 10:00:26 INFO DefaultMLSQJobProgressListener: [owner] [admin] [groupId] [6fabcb5f-0e08-45c1-9251-85d1aa1de15c] __MMMMMM__ Total jobs: 1 current job:1 job script:load jsonStr.`jsonStr` as mock_data
```

features	label
[5.1, 3.5, 1.4, 0.2]	0
[5.1, 3.5, 1.4, 0.2]	1
[5.1, 3.5, 1.4, 0.2]	0
[4.4, 2.9, 1.4, 0.2]	0
[5.1, 3.5, 1.4, 0.2]	1
[5.1, 3.5, 1.4, 0.2]	0
[5.1, 3.5, 1.4, 0.2]	0
[4.7, 3.2, 1.3, 0.2]	1
[5.1, 3.5, 1.4, 0.2]	0
[5.1, 3.5, 1.4, 0.2]	0

私有服务
一键部署

Deploying MLSQL Engine on k8s

```
## K8S config file resides in ~/.kube/config by default.  
## chncaesar/mlsql-engine-k8s:3.0-2.1.0-SNAPSHOT is a pre-built K8S image  
./mlsql-deploy run \  
  --kube-config ~/.kube/config \  
  --engine-name mlsql-k8s \  
  --engine-image chncaesar/mlsql-engine-k8s:3.0-2.1.0-SNAPSHOT \  
  --engine-executor-core-num 2 \  
  --engine-executor-num 1 \  
  --engine-executor-memory 2048 \  
  --engine-driver-core-num 2 \  
  --engine-driver-memory 2048 \  
  --engine-access-token mlsql \  
  --engine-jar-path-in-container local:///home/deploy/mlsql/mlsql-engine_3.0-2.1.0-SNAPSHOT/libs/stru  
  --storage-name jfs \  
  --storage-meta-url redis://127.0.0.1:6379/1
```



可覆盖人群

数据科学家

大数据工程师

产品运营

应用API (譬如业务后台分析相关功能)

Apps(Notebook, Works, Scripts等)

MLSQL Engines

MLSQL如何帮助企业 低成本落地 Data + AI?

MLSQL四大特性



MLSQL是开源的

Unwatch 115 Unstar 1.3k Fork 470

About

The Programming Language Designed For Big Data and AI

mssql.ai

machine-learning bigdata mssql

sql-like-dsl

Readme

Apache-2.0 License

Contributors 30

+ 19 contributors

Languages

JavaScript 53.5%	Scala 15.3%
CSS 12.3%	HTML 10.1%
Less 5.3%	Java 1.9%
Other 1.6%	

云上/云下 皆可

开源社区保障

商业 (Kyligence) 选择

统一的语言

```
load excel.`/tmp/excel-example/triage-patient.xlsx` where useHeader="true" as triagePatient;
load excel.`/tmp/excel-example/master-email.xlsx` where useHeader="true" as masterEmail;

-- select date_format(cast (UNIX_TIMESTAMP(date, 'dd/MM/yy') as TIMESTAMP), 'dd/MM/yy') as x, date

select date as x, 0 as y1, patientNum as y2 from triagePatient where triage="皮肤科"
union all
select date as x, patientNum as y1, 0 as y2 from triagePatient where triage="眼科"
as tempTable;

select x, sum(y1) as `眼科`, sum(y2) as `皮肤科`,
-- 告诉系统需要生成图表
"line" as dash
from tempTable where x is not null group by x order by x asc
as finalTable;

select tp.*, me.email from triagePatient as tp left join masterEmail as me on tp.master==me.master
as triagePatientWithEmail;

select first(email) as x,
avg(patientNum) as patientEveryDay, master, first(email) as email,
"bar" as dash
from triagePatientWithEmail
group by master
order by patientEveryDay desc
as output;

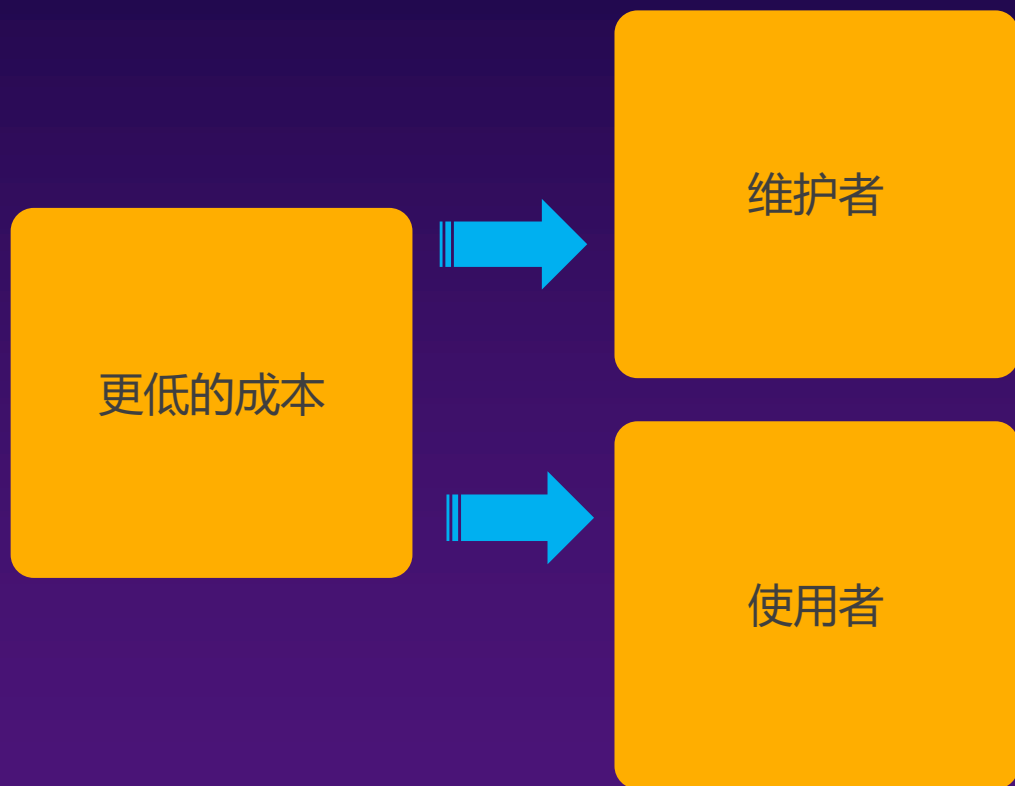
save overwrite triagePatientWithEmail as excel.`/tmp/triagePatientWithEmail.xlsx`
where key="value" ;
```

统一的引擎

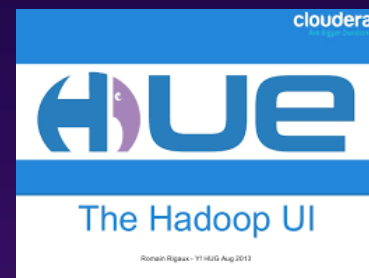
Index of /2.1.0-SNAPSHOT/

../		
mysql-console-2.1.0-SNAPSHOT.tar.gz	17-Jun-2021 09:25	102M
mysql-engine_2.4-2.1.0-SNAPSHOT.tar.gz	31-May-2021 17:42	119M
mysql-engine_3.0-2.1.0-SNAPSHOT.tar.gz	31-May-2021 17:51	118M

统一的价值



告别



简单

几天入门

满足不同层次人
群诉求

代码易于自动化
生成

从运营产品到数据科学家，数据工程师，
都可以使用功能Mysql完成自己的工作

释放大数据+AI生产力

公司不担心招人了

一线人员效率提高

价值

提升了Data+AI给
企业带来的效率

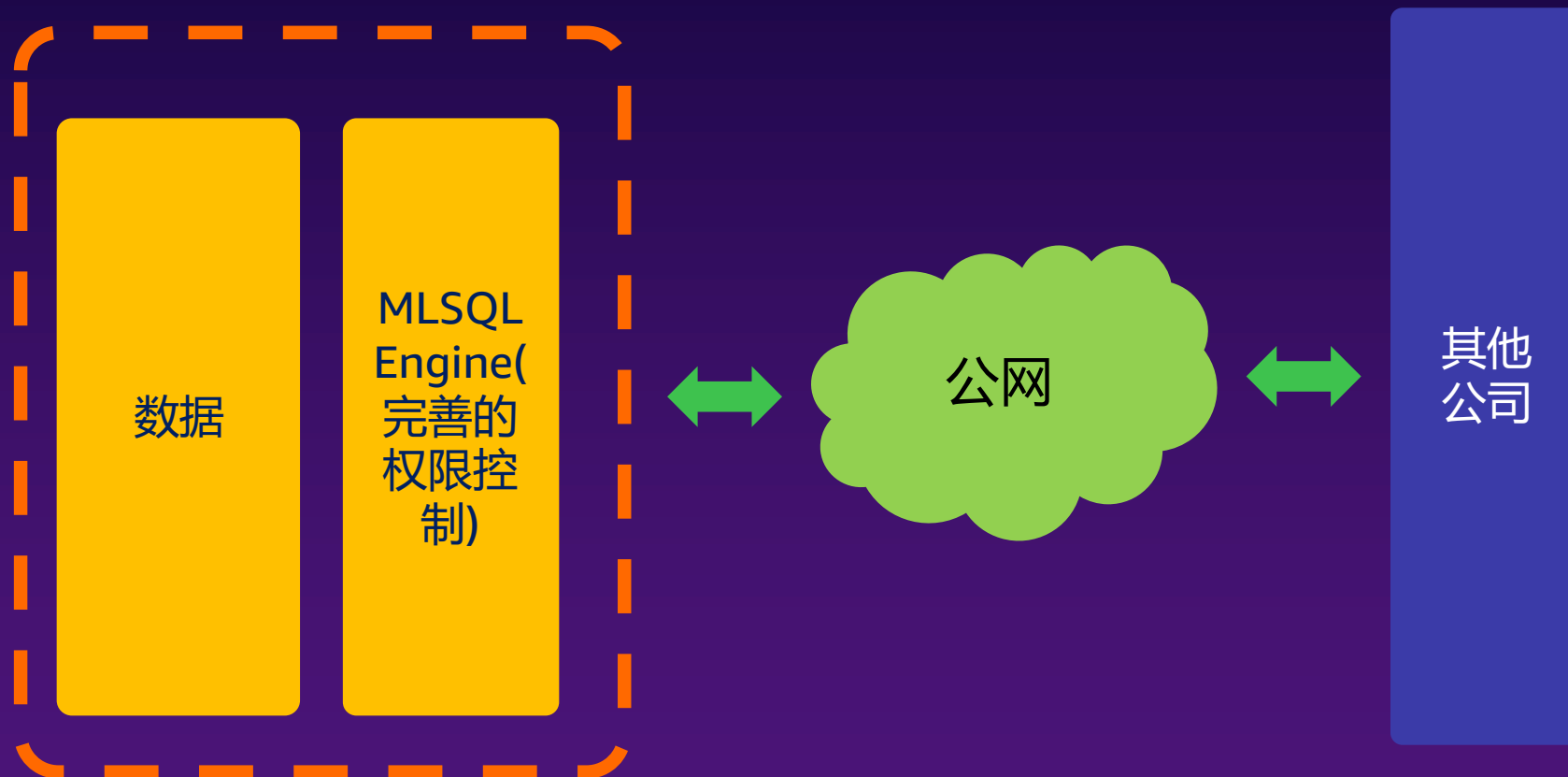
数据安全

1. 对底层数据源无侵入性
2. 支持表，列，行级别权限控制
3. 表级别权限可秒级校验，避免运行等待

语言安全

针对每个用户控制可使用语言功能

1. 是不是能使用某个模块、插件
2. 是不是能使用自定义UDF (Scala/Java)
3. 是不是能使用Python
4. Python 为沙箱，只能访问用户有权限的数据
5. 还有更多。。。。



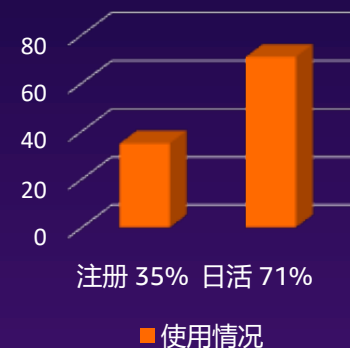
1. 数据不搬家
2. 强大的权限控制
3. 强大的计算能力

典型场景1-某消费金融公司

易于使用

产品，运营，研发，数据科学家等等

使用情况



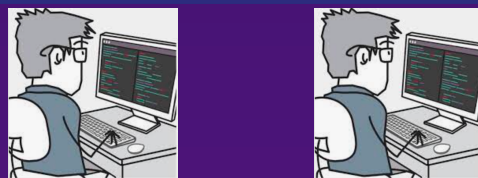
全公司200人

平台能力强大

MLSQL
Make data analysis simple and smart

持续运行：➡ 三年
累计任务：➡ 700万
数据规模：➡ TB级

易于维护



支撑团队：2人

典型场景2-厦门某信息公司

客户30个模型

500+任务



4人一个月

MLSQL

Make data analysis simple and smart



客户一个模型2周搞不定，
堆人也没用

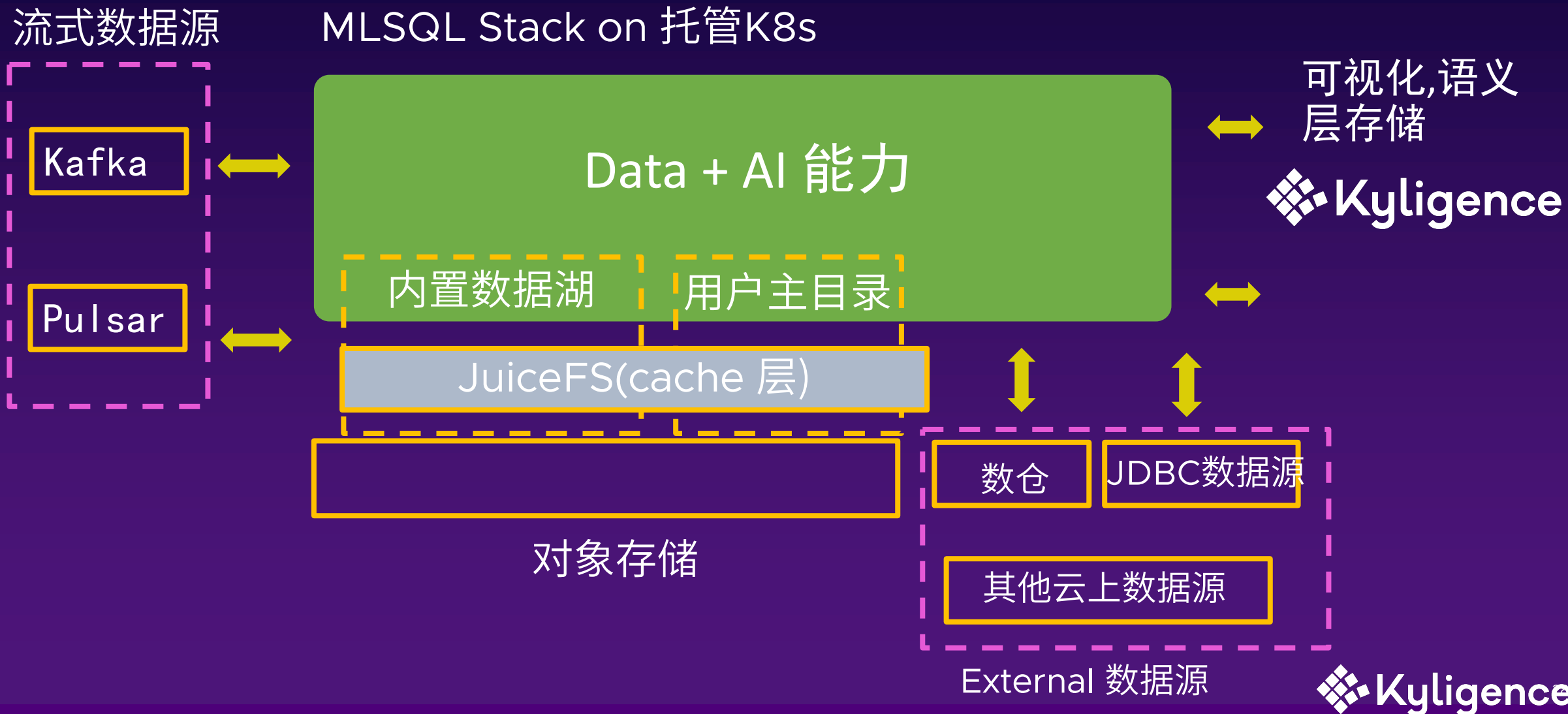


开发效率15倍提升

Kettle难以满足越来越复杂的数据处理，如异构join

MLSQL 脚本支持现场、远程开发部署

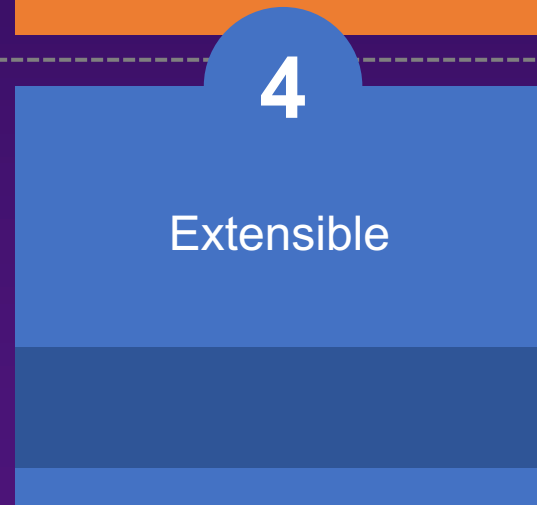
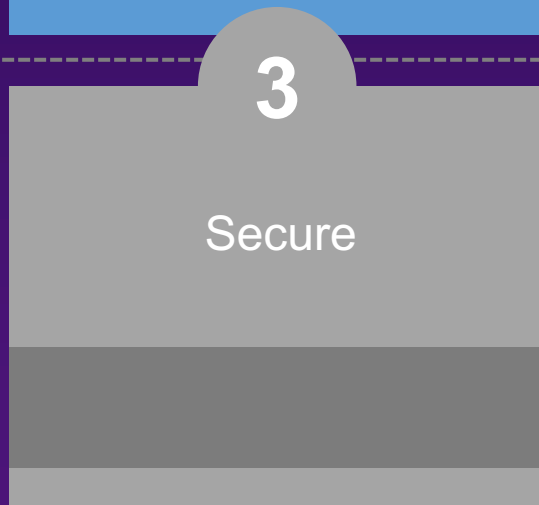
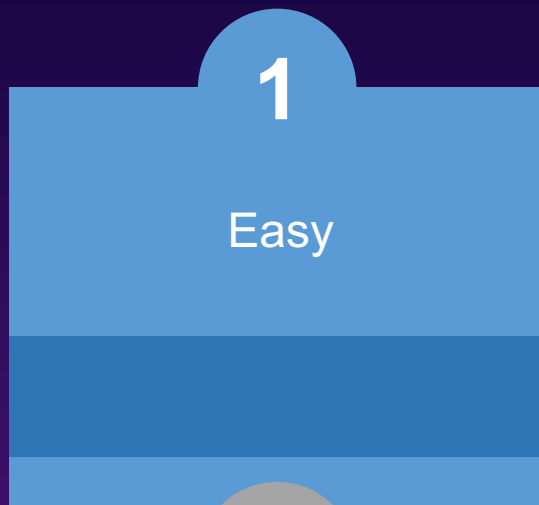
客户现场仅需一人维护即可



MLSQL Summary

- ◆ In Notebook
- ◆ All about SQL
- ◆ Seamless SQL & Python
- ◆ No PySpark

- ◆ Non-intrusive, out-of-box data ACL
- ◆ Security on Plugin, Algorithm, Data and Directory
- ◆ Custom desensitization



- ◆ Analyze and explore multiple data sources
- ◆ Support algorithms and feature engineering, support Python ecosystem
- ◆ Support Kylin and other analytical engines
- ◆ UDF and UDAF hot deployed
- ◆ Pluggable architecture
- ◆ User defined extension



Open Source AceCon

2021

智能云边开源峰会

AI x Cloud Native x Edge Computing

人工智能 × 云原生 × 边缘计算

Thank You

Luke Han | luke.han@kyligence.io

联合创始人 & CEO