Heywhale 和鲸

数 据 科 学 协 同 创 新 平 台

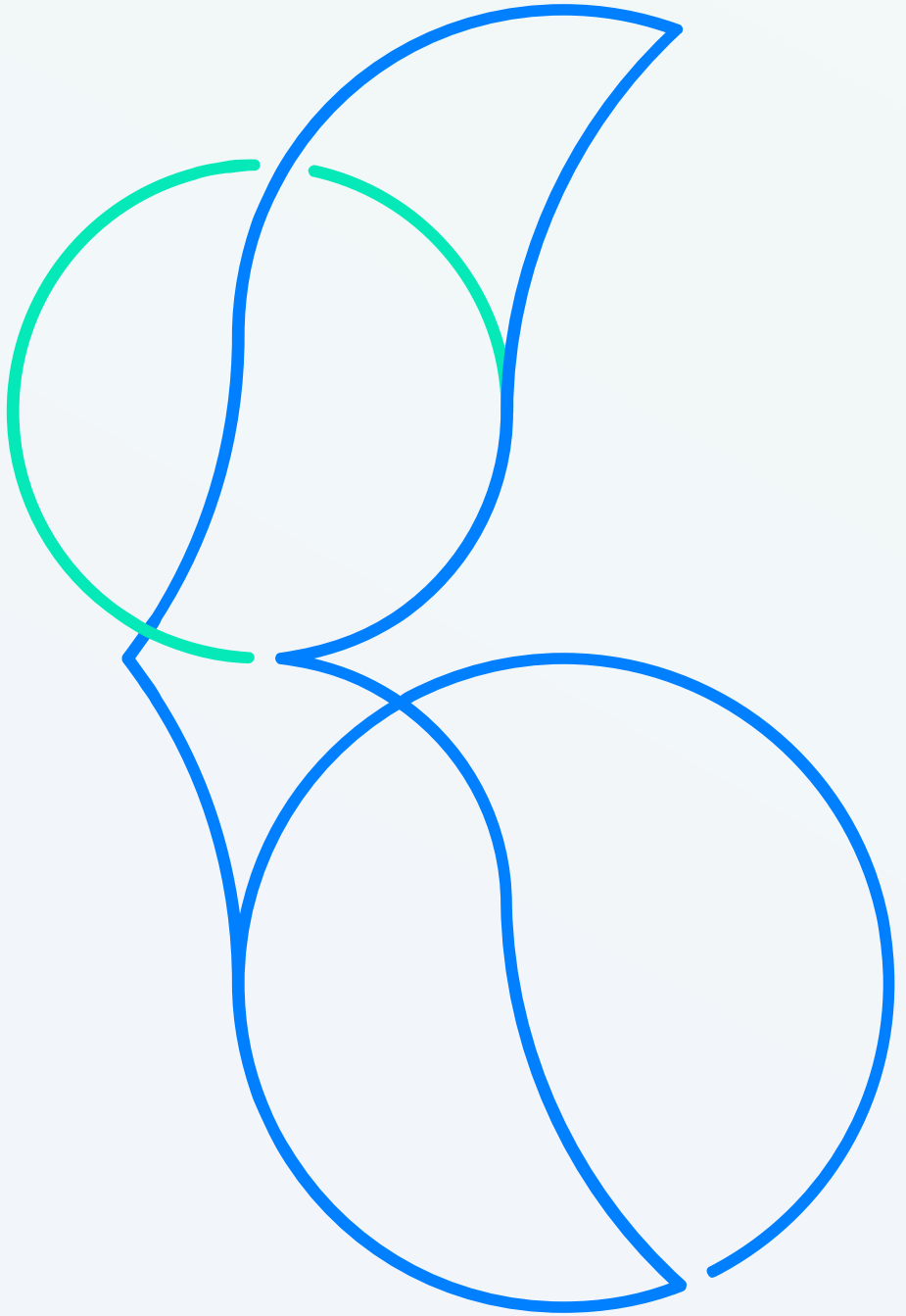
SQLFlow与机器学习和数据运营

让AI像SQL一样简单

Heywhale.com

上 海 和 今 信 息 科 技 有 限 公 司

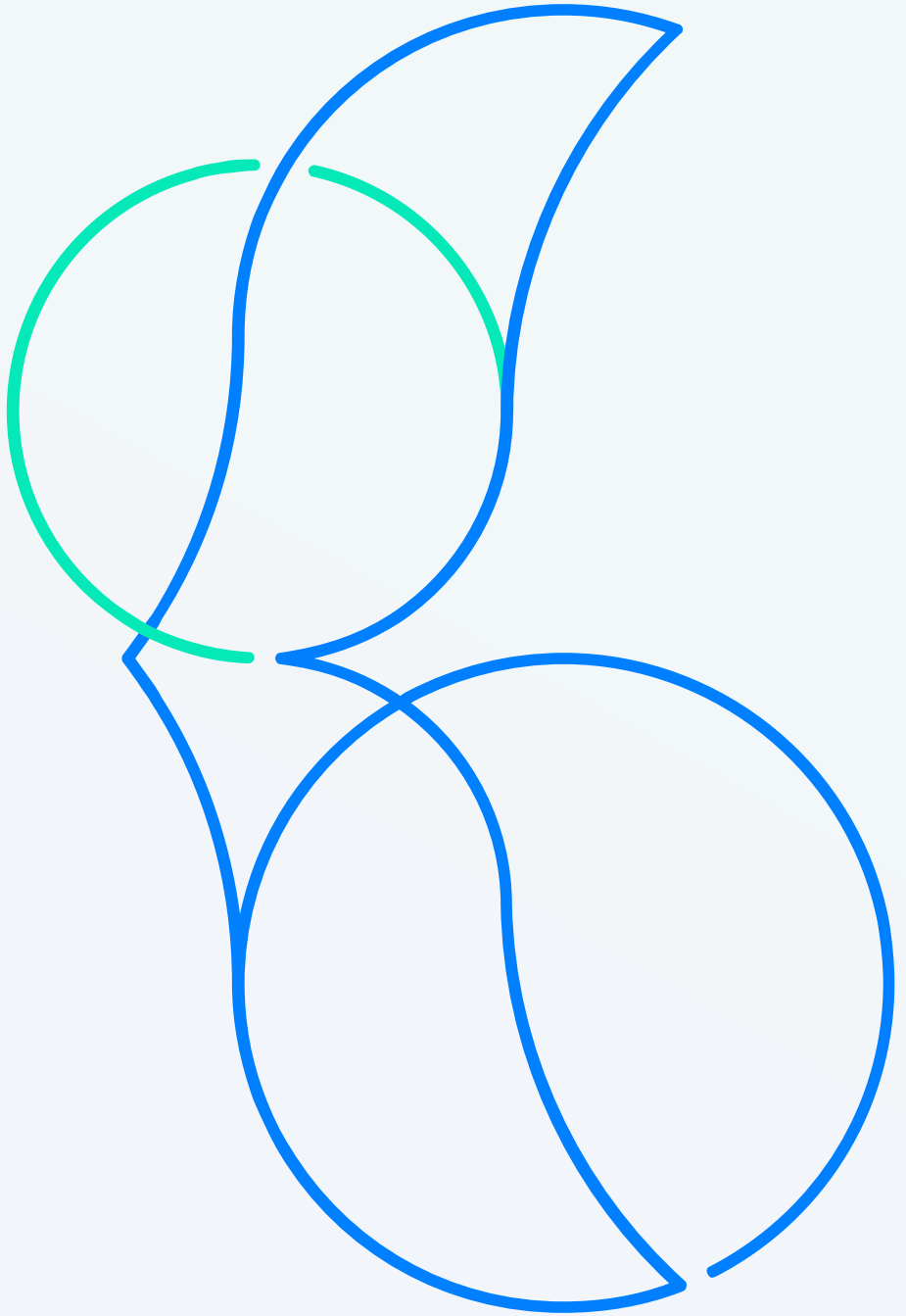
CONTENTS
目录{



1

SQLFlow

一切皆可SQL

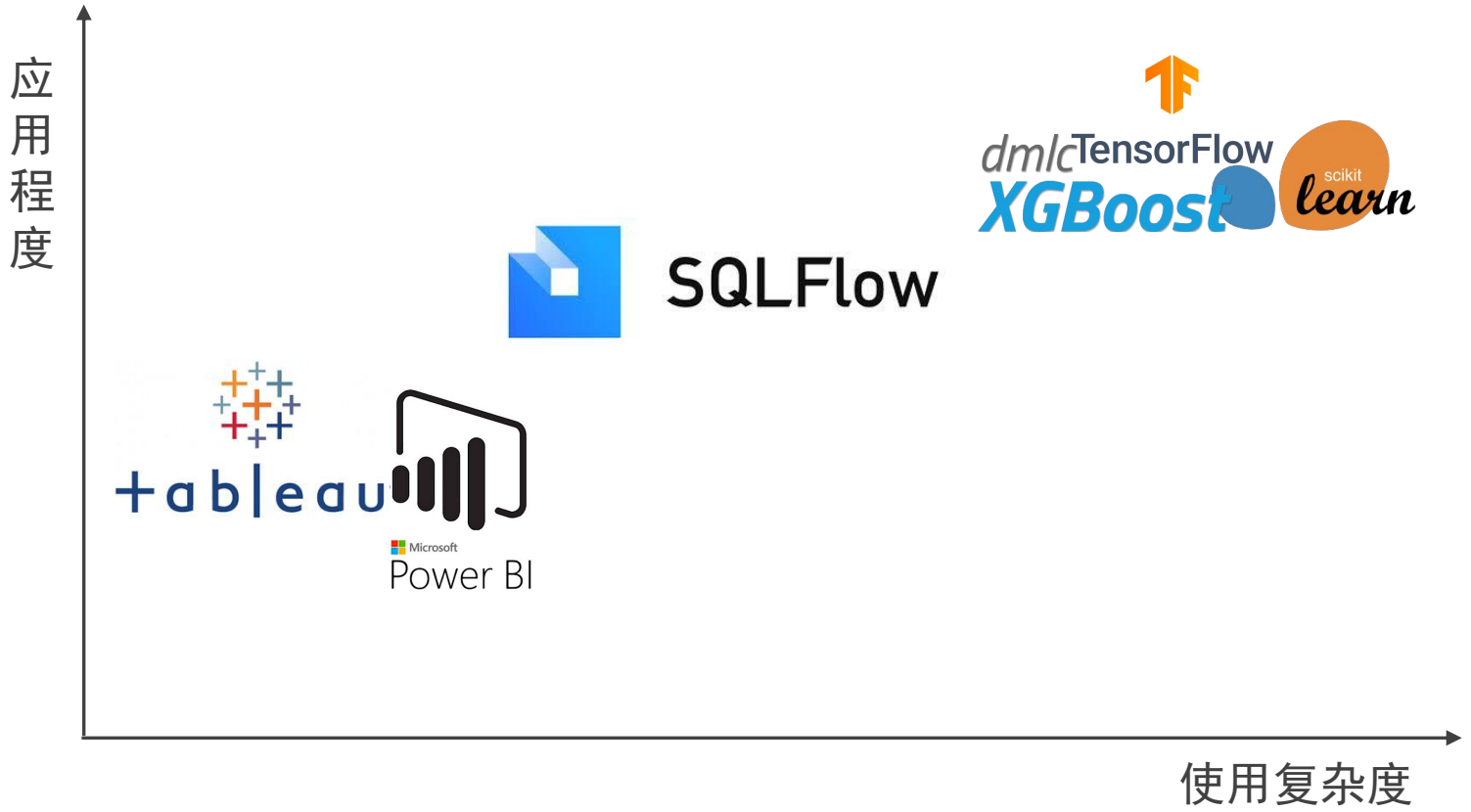


Heywhale.com

上海和今信息科技有限公司

SQLFlow的电梯演讲

让机器学习像SQL一样简单，降低数据分析与机器学习的门槛，提高工程师应用机器学习的效率。



数据应用的三象限

BI软件：易用，但可扩展性差

对于一般的运营专家和数据分析师来说，BI软件可以完成大部分工作，但是BI软件的功能是泛化的通用功能，如果需要在一些常用的统计性描述之外添加新的描述信息就比较难于扩展，系统化方面也做得不好。

SQLFlow：和SQL用法一致，支持扩展新模型

对于大部分的数据分析师和工程师来说，SQL是必备技能，但是Python不是，对于传统数据分析师来说SQL是他们的舒适域，对于其他工程师来说SQL可以更快更有效的进行机器学习工程，不需要从头开始，SQL的数据有schema，对于问题的定义也非常的干净。

机器学习框架：使用Python等编程语言，可自由扩展

自由度高，也没有限定数据源，不光是分布式数据库，分布式文件系统本地文件也好，都可以使用，也不光是可以集成到训练当中作为服务本身对外发布也可以，需要一定的专业基础和编程能力。

从过去受到的启发

MapReduce横空出世

大规模的数据集有了比较理想的处理方案，但是设计虽好，使用起来却特别繁琐，针对一个数据处理问题如何写MapReduce

成了一个复杂的工作。

Hive让MapReduce的普适性变得更强

很多数据分析师需要从原本的Oracle，切换到Hadoop上工作，Hive将SQL翻译成对应的MapReduce执行，完全屏蔽了MapReduce

的主体，使得一大批数据分析师可以基本迎接大数据的时代。

SQL太香了

从SQL到NoSQL到NewSQL，数据业务本身还是希望有schema，有无限扩展能力。也有像Pig这样的方言或者MLLib这样的工具出现。

对于机器学习来说，也是希望数据是标准的，数据处理的能力是可扩展的，SQL就是一个非常好的、纯粹的interface，有优秀的定义和用户基础。

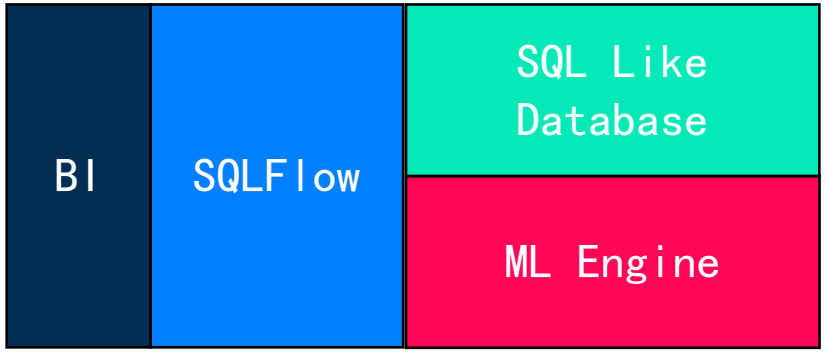
对于SQL的使用和研究也非常成熟有很多现成的工具可以使用。

SQLFlow的象限

SQLFLow负责衔接机器学习任务和SQL数据库，通过将扩展的SQL拆解成SQL的数据提取和机器学习任务，定义了一个非常标准的实时机器学习任务。

SQL 本身具备非常强的生态条件，传统的关系型数据库也好，解释型的类SQL数据库也好，甚至其他的非SQL数据源（MongoDB、S3）都是可以通过类似Hive的机制成为SQLFlow的数据源。

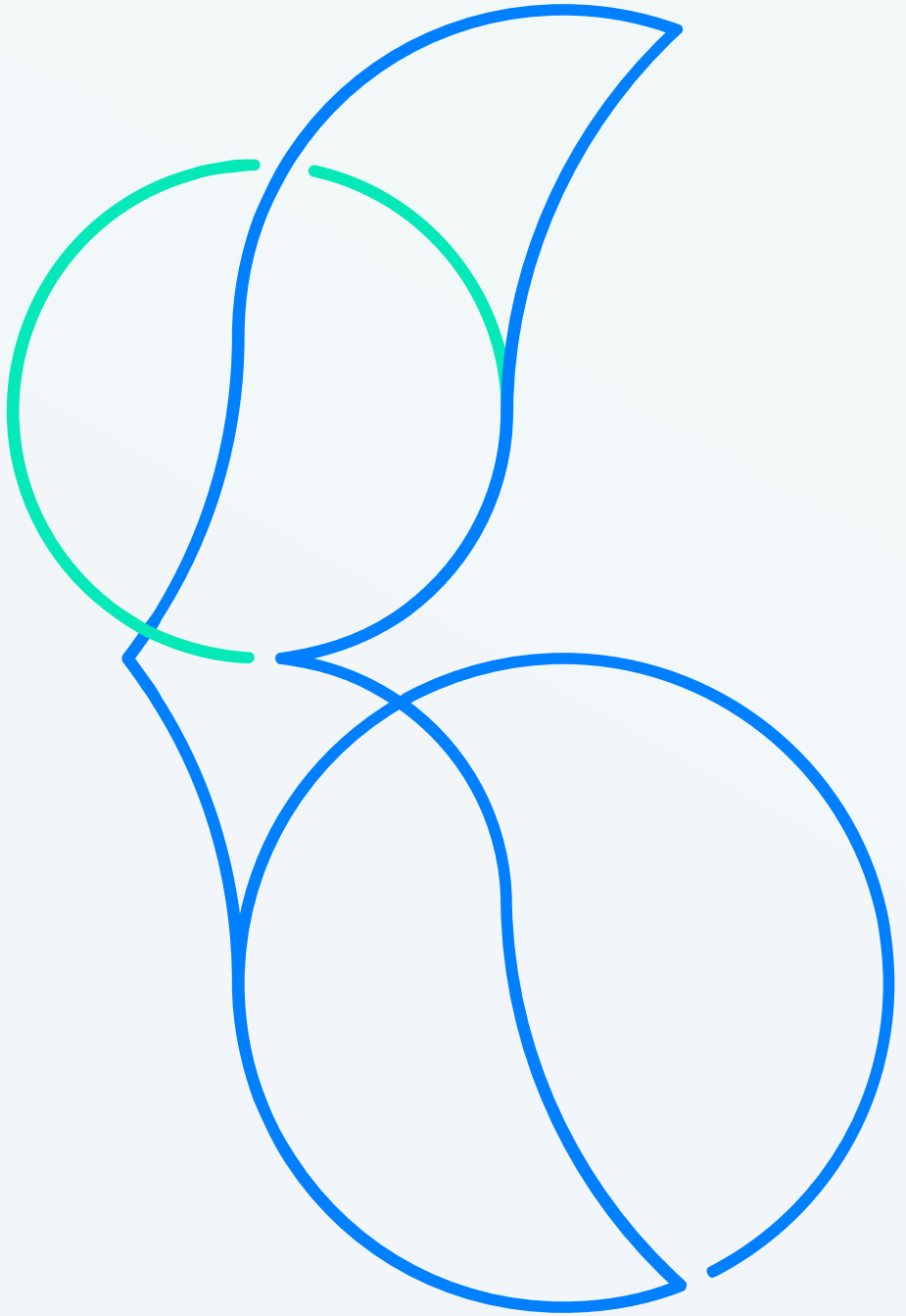
从工程的角度来说，SQLFlow围绕着有schema的数据构建机器学习任务，天然标准化了生成程序需要负责的边界，数据落库以后相对于原始数据会干净很多。




```
SELECT * FROM iris.train
TO TRAIN DNNClassifer
WITH hidden_units = [10, 10], n_classes = 3,
EPOCHS = 10
COLUMN sepal_length, sepal_width, petal_length, petal_width
LABEL class
INTO sqlflow_models.my_dnn_model;
```

2

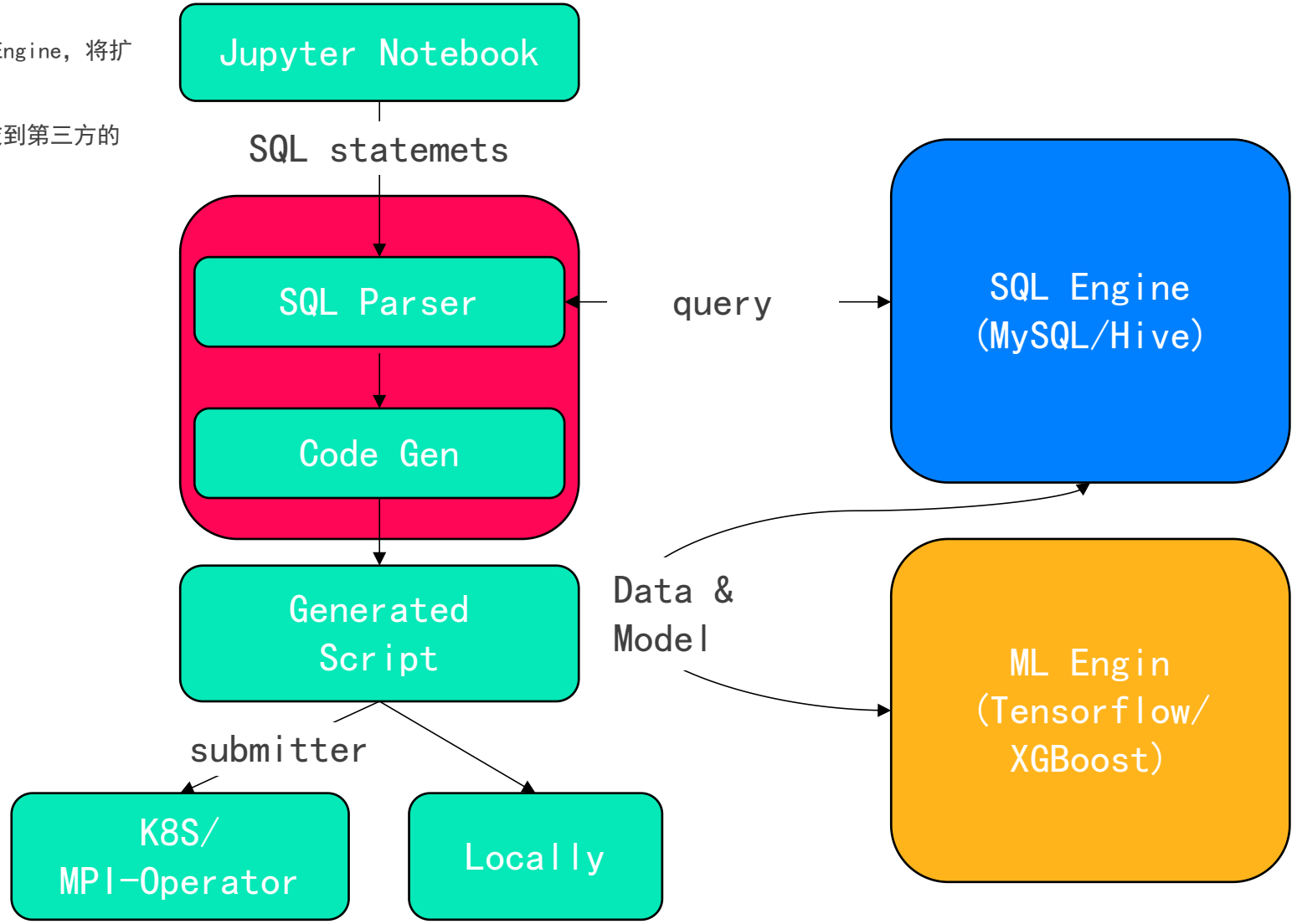
SQLFlow的设计



SQLFlow的整体设计

SQLFLow 整体比较简单，SQL Parser 将标准SQL的部分旁路给SQL Engine，将扩展的部分翻译成机器学习任务。

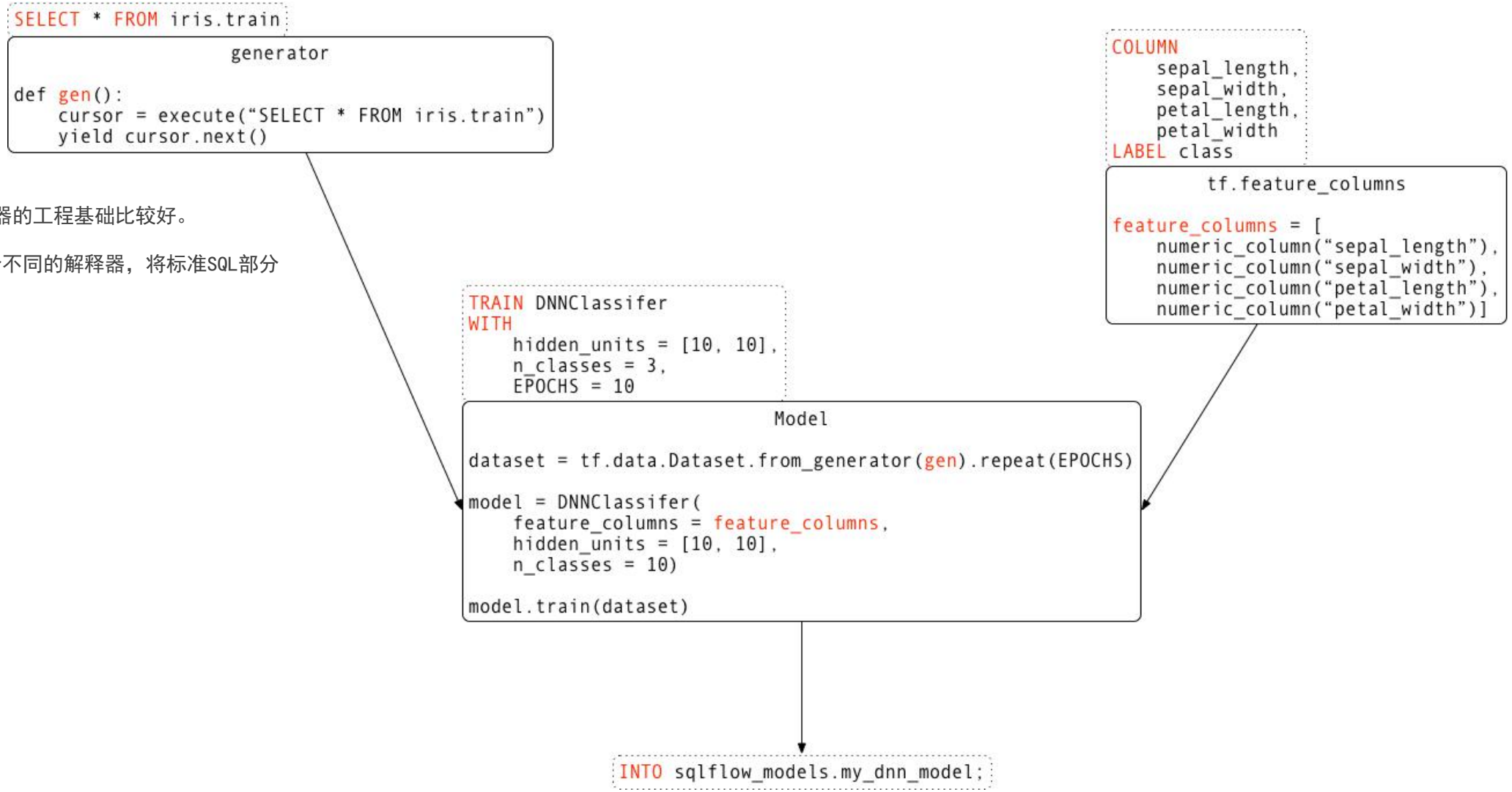
生成程序是相对独立的，可以就地执行也可以通过 submitter 提交到第三方的系统当中去执行。
这部分可以进行比较自由的扩展，而本身的主体结构不需要修改。



SQL Parser

SQL历史悠久，相关解释器的工程基础比较好。

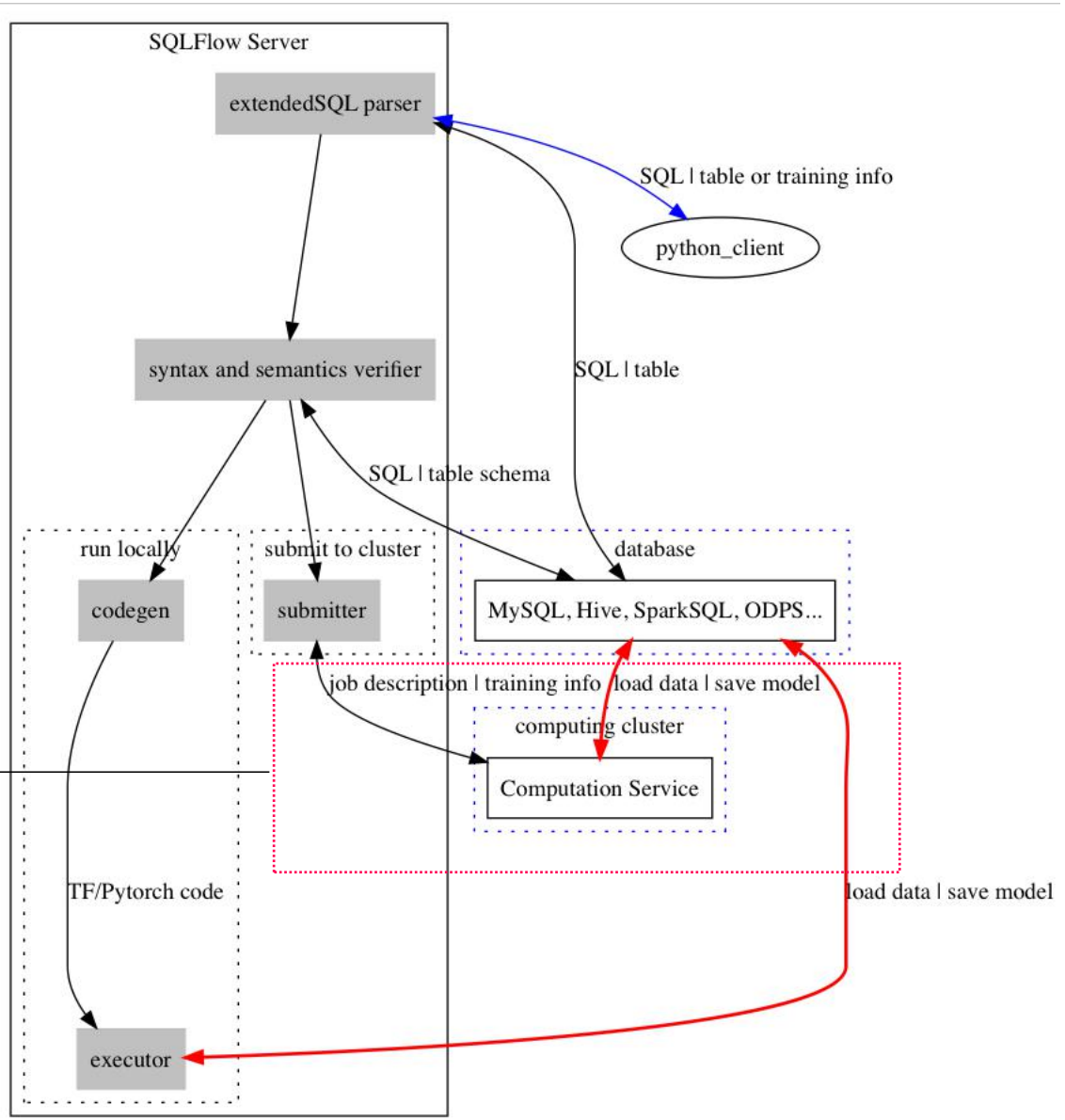
根据不同的 driver 交给不同的解释器，将标准SQL部分旁路给SQL数据库。



SQLFlow Submitter

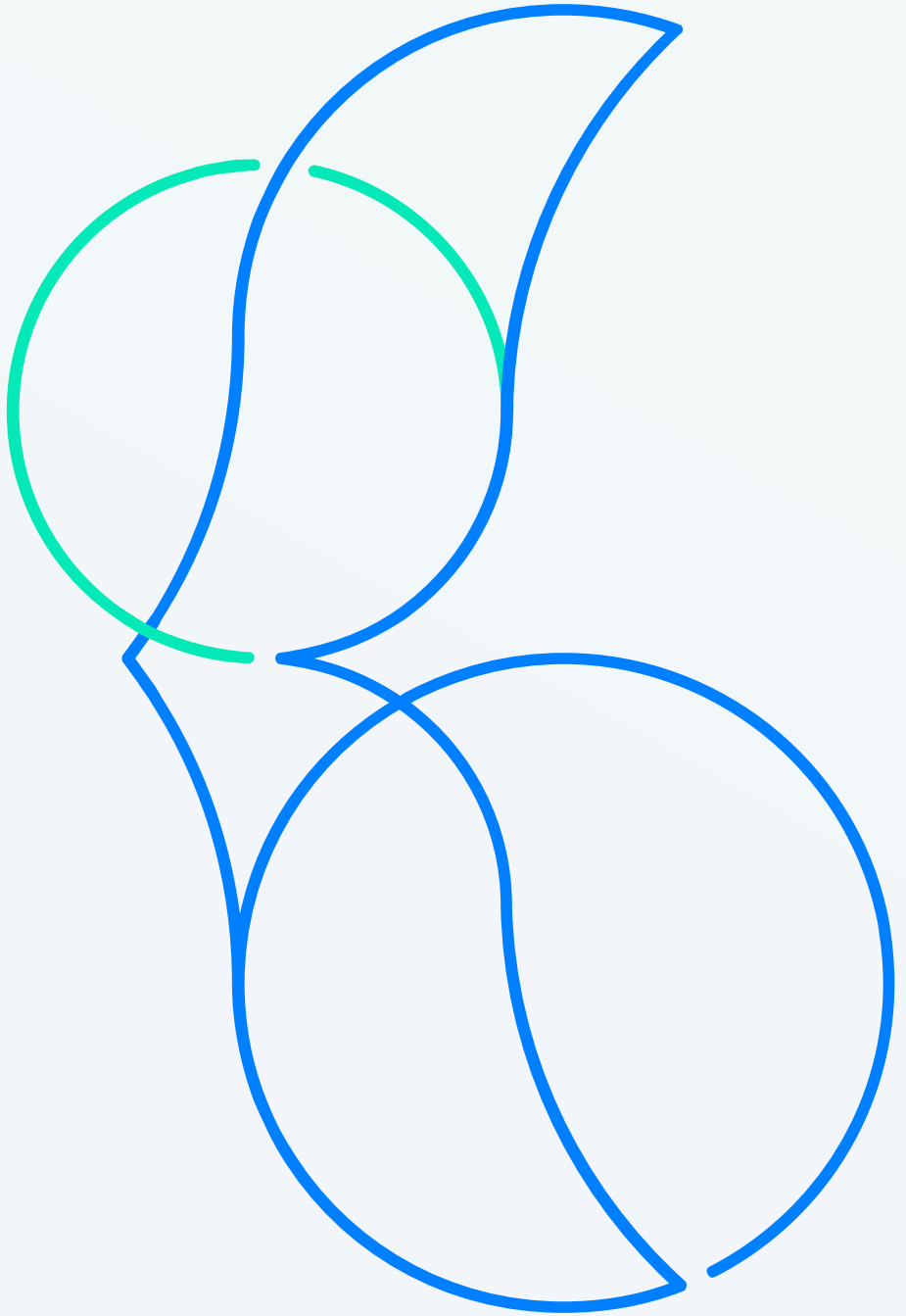
Submitter作为SQLFLow一个模块的形式存在，扮演任务提交者的角色

将机器学习任务提交到第三方的平台进行运行。



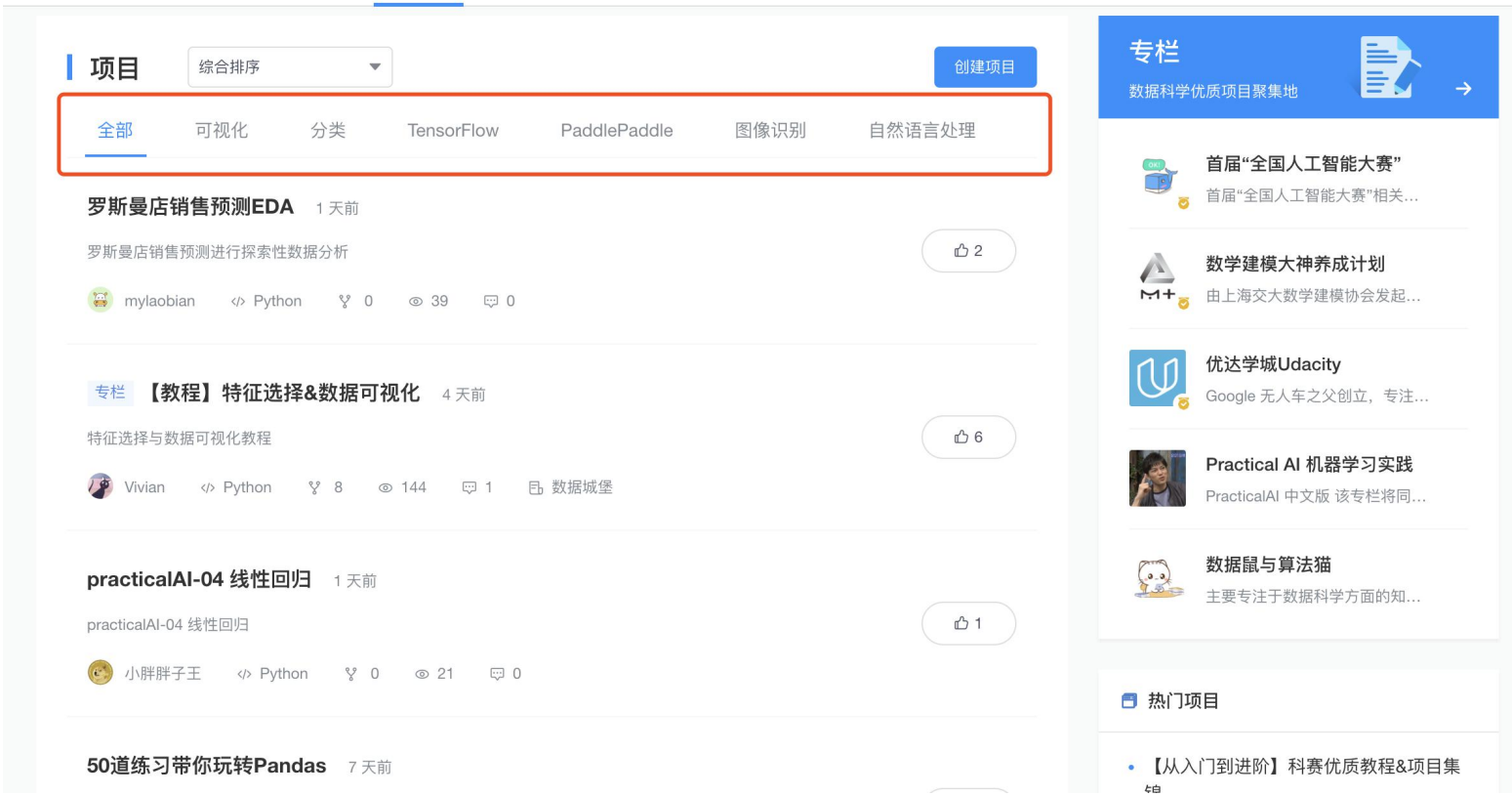
3

应用：网站项目分类



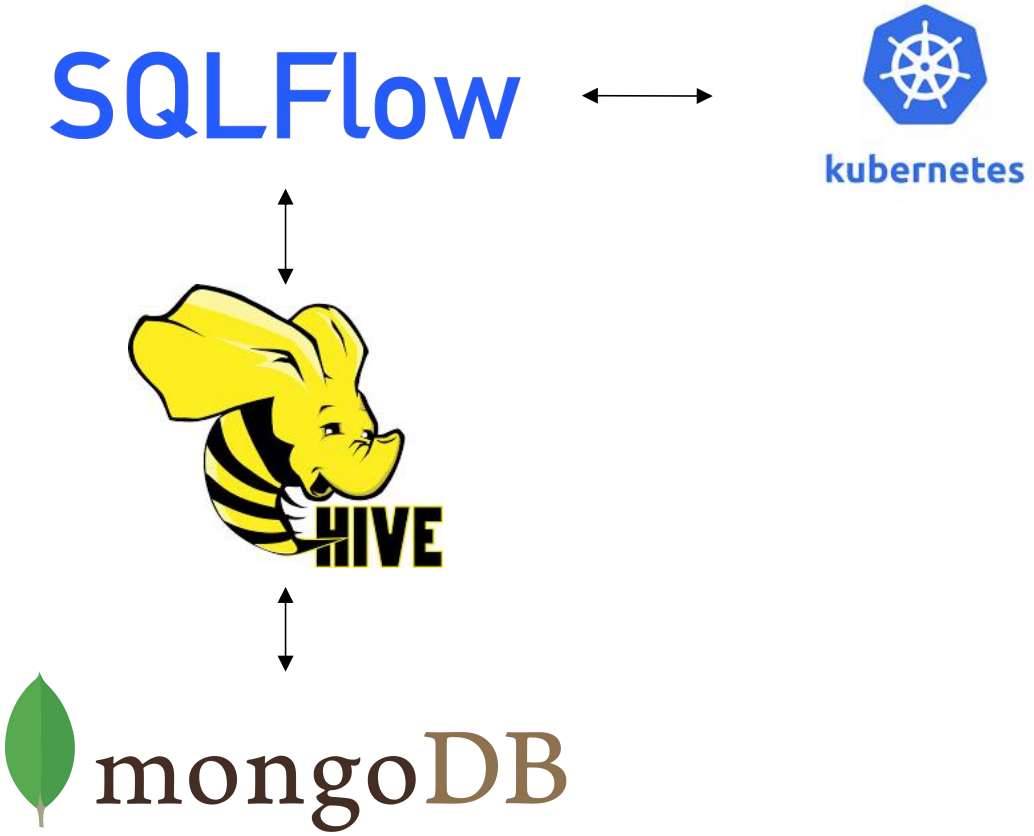
社区项目分类

kesci.com 的数据分析项目需要将有效的质量较好的项目分类与排序，我们对项目的排序就是一个简单的有监督分类。



数据流和处理

项目的原始信息存在MongoDB当中，通过Hive External Table的方式拉出来，SQLFLow 将项目的用于训练的字段通过SELECT作为数据集提取，再将sklearn的训练任务部署在K8S上，定期执行把分类字段写回到业务数据库MongoDB当中。



工作量

一个支持分词和词向量训练的分类模型

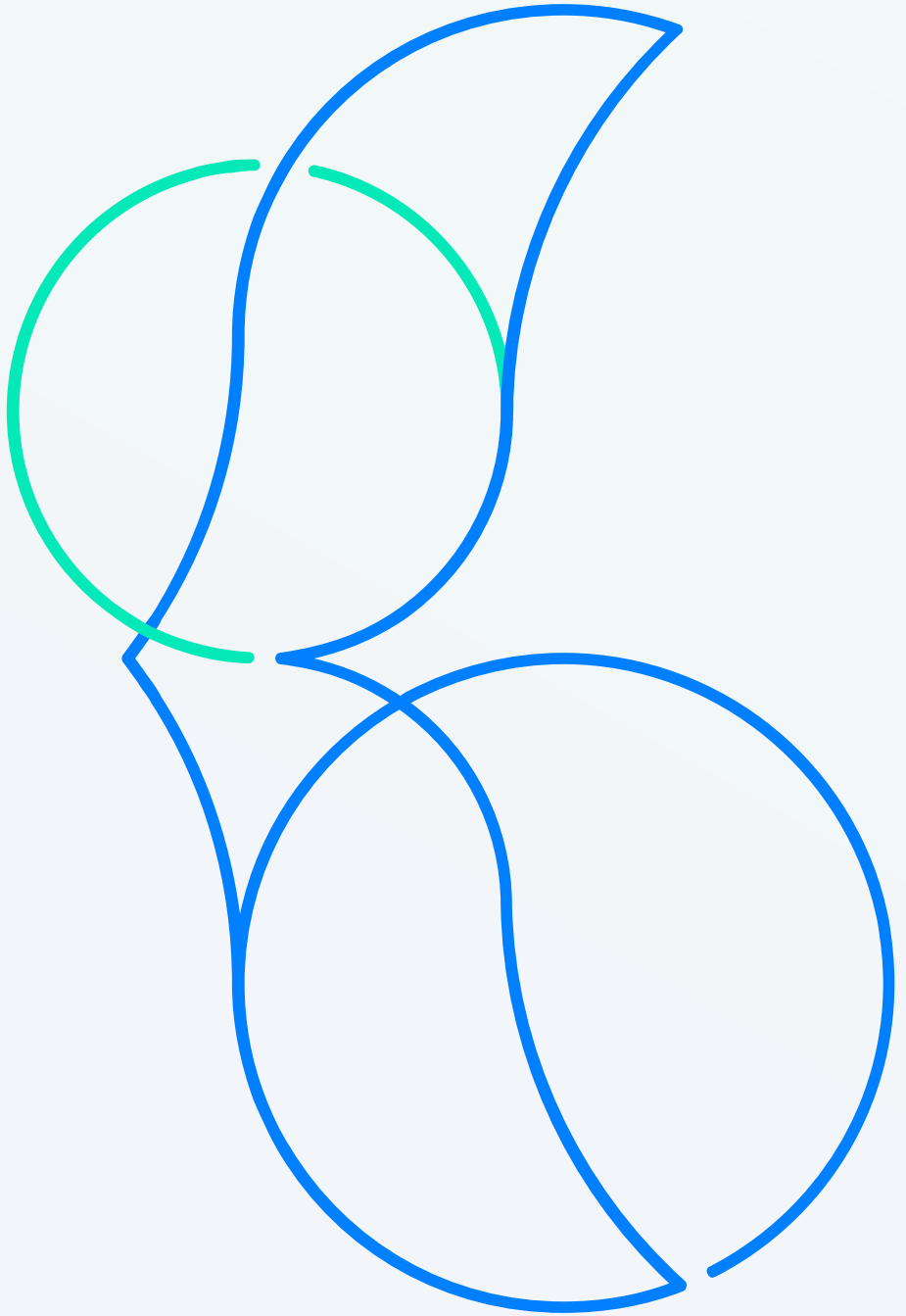
这个工作比较繁琐，但和之前需要读取MongoDB的数据，进行预处理然后写回到MongoDB来说，变成了围绕SQL实现的机器学习任务，将来可以用于类似的文本分类的任务而不需要适配新的数据源。

一条SQL语句

指定WHERE和SELECT以及模型名称就可以进行训练，不用任务其他的工具，如果产品经理或者数据分析师想使用这个模块，就不需要重复受数据工程的折磨。

4

Demo：莺尾花分类



Heywhale 和鲸

数 据 科 学 协 同 创 新 平 台

Heywhale.com

上 海 和 今 信 息 科 技 有 限 公 司

上海　上海市徐汇区宜州路路188号B8栋14层

北京　北京市朝阳区东直门外大街东外56号A座

电话　021 - 8037 0235（转008）

邮箱　business@heywhale.com