Kyligence Cloud 华为云 平台快速上手

本文将为您介绍如何在 华为云 平台上快速部署并上手 Kyligence Cloud。本教程通过将使用内置的 New York Taxi 数据集和模型，并通过内置的可视化数据分析工具 Kyligence Insight 制作可视化图表，推荐您使用 Chrome (64.0.* 或更高版本) 浏览器来进行接下来的操作，每个小节标明适合阅读的角色，您可按照您的需要阅读对应小节：

您将掌握的内容

在本教程中，您将了解以下内容：

- 在 华为云 平台快速部署 Kyligence Cloud
- 在 Kyligence Cloud 中创建工作区和项目
- 导入内置的样例数据集和模型
- 部署可视化数据分析工具 Kyligence Insight 并创建分析图表
- 清理 Kyligence Cloud 相关资源

基本概念

- 工作区：工作区是 Kyligence Cloud 下的一级管理单位，每个工作区使用不同的集群，彼此逻辑隔离且数据不共享，例如您可按照开发、测试、生产要求，部署 3 套工作区来满足您的需求。
- 项目：项目是每个工作区下的一级管理单位，同一个工作区下的项目将共享同一套集群资源，您可以在一个工作区内创建多个项目，并服务于不同的业务范围，在一个项目中，您可以设计多个模型并进行查询分析。
- 数据目录：数据目录（Data Catalog）是 Kyligence Cloud 中的元数据管理服务，通过数据目录读取**云对象存储**（Blob, ADLS Gen2, S3, OBS等）中的文件并定义其表结构。数据源类型为**对象存储**的工作区会创建并使用数据目录。每个工作区会创建一个数据目录，数据目录被工作区内所有项目共享，对数据资产进行统一的发布及管理。
- 同步表：由于数据目录是工作区下所有项目共享的，在用户进行数据分析前需要将数据目录中的表同步到当前项目中。
- 模型：模型，也是逻辑语义层。模型是一组表以及它们间的关联关系 (Join Relationship)。模型中定义了事实表、维度表、度量、维度、和一组索引。模

型和其中的索引定义了加载数据时要执行的预计算，当前支持星型模型和雪花模型。

- 索引：在数据加载时将构建索引，索引将被用于加速查询。索引分为聚合索引（Aggregate Index）与明细索引（Table Index）。聚合索引本质是多个维度和度量的组合，适合回答聚合查询，比如某年的销售总额；明细索引本质是大宽表的多路索引，适合回答精确到记录的明细查询，比如某用户的最近 100 笔交易。
- 加载数据：为了加速查询，需要将数据从源表加载入模型，在此过程中也将构建索引，整个过程即是数据的预计算过程。每一次数据加载将产生一个 Segment。载入数据后的模型可以服务于查询，由于预计算，在模型上执行的查询将获得极大的加速。
  - 增量数据加载（Incremental Load）：在事实表上可以定义一个分区日期或时间列。根据分区列，可以按时间范围对超大数据集做增量加载。
  - 全量加载（Full Load）：如果没有定义分区列，那么源表中的所有数据将被一次性加载。
  - 重建索引（Build Index）：用户可以随时调整模型和索引的定义。对于已加载的数据，其上的索引需要按新的定义重新构建。如果用户要求加速某些查询，系统也可能优化模型和索引，进而触发重建索引。
- 查询加速：指通过自动优化模型和索引来加速查询的能力。系统可以依据历史查询模式和数据集特征来自动优化模型和索引。这样可以大量节省用户手工设计模型和索引的时间。

前期准备

Kyligence Cloud 使用 华为云 应用注册功能获得操作授权用来部署需要的资源，请确保您使用的 华为云 订阅具有创建以下服务的权限，如您不确认是否具有以下权限，请联系您的云平台管理员：

资源	资源类型	版本	提供者
虚拟私有云 VPC	网络	-	华为云
网络安全组	网络	-	华为云
弹性负载均衡 ELB	网络	-	华为云
云数据库 MySQL	数据库	MySQL 5.7.32	华为云
弹性云服务器 ECS	计算	OS: CentOS 7.9	华为云
对象存储服务 OBS	存储	-	华为云
云硬盘 EVS	存储	-	华为云

同时，你需要准备下列资料：

- Kyligence Cloud 许可证
- 华为云访问密钥
  - 访问密钥 ID
  - 私有访问秘钥

> 没有 Kyligence Cloud 许可证？您可参考前往 Kyligence 官网 申请试用许可证

部署 Kyligence Cloud 服务

此小节适用角色：IT 管理员

1. 前往 Kyligence Cloud 自动部署引导页面。根据您的需求选择「**华为云 中国站**」或者「**华为云 国际站**」，然后单击 **下一步** 按钮。

2. 在授权信息页面输入以下信息：

   - 部署地区：选择部署 Kyligence Cloud 服务的区域
   - 访问密钥
   - 私有访问密钥

1. 在部署基本信息页面，您需要提供以下信息：

   - 可用区：设置您想要创建 Kyligence Cloud 的区域。
   - 堆栈名称：我们将使用「华为云」应用编排服务 (AOS) 进行部署，请输入该堆栈的名称
   - ECS 机型：选择部署 Kyligence Cloud 服务的 ECS 机型
   - 密钥对：选择远程访问 Kyligence Cloud 服务的 ECS 机型的 SSH 密钥
   - RDS 类型：选择保存 Kyligence Cloud 元数据的数据库类型
   - Zookeeper ECS 机型
   - 访问规则：配置可访问 Kyligence Cloud 服务的 IP 范围，如需配置多个 CIDR，请参考 安全组配置，前往 华为云 控制台，前往 Kyligence Cloud

实例的安全组手动添加入站规则。

> 我们使用华为应用编排(AOS)为您自动部署 Kyligence Cloud，所以我们为会为您创建一个名为[yourstackname]_tmplxxxx的AOS模版。您可以在应用编排服务->我的模版中找到模版详情。

1. （非必须）点击 **下一步**。如果您需要对资源进行管理，可以在此界面为 Kyligence Cloud 部署创建的资源添加标签。

> 标签允许您从不同的标准 (例如用途、所有者或环境) 对资源进行分类管理。当你拥有很多相同类型的资源时候，通过便签则可以进行更高效的管理。
>
> 我们建议您针对每类资源设计一组标签，以满足您的需要。使用一组连续的标签键，管理资源时会更加轻松。
>
> 您可以根据添加的标签搜索和筛选资源。有关如何实施有效的资源标记策略的更多信息，请参阅 华为云 白皮书标记最佳实践 。

您可以点击添加标签然后填写标签名和标签值。 您可以将标签的值设为空的字符串，但是不能将其设为空值。如果您添加的标签的值与该实例上现有标签的值相同，新的值就会覆盖旧值。如果删除资源，资源的所有标签也会被删除。

1. 点击 **部署** 即启动 Kyligence Cloud 自动部署，页面会自动跳转到「部署进度页面」。

> Kyligence Cloud 将在您所选资源组内自动创建所需的资源和服务，整个过程大概需要 20 分钟左右，具体时间可能因不同服务环境的网络条件有所偏差。

2. 部署成功后点击页面上的 Kyligence Cloud 服务地址，即可进入 Kyligence Cloud 主页面。首次进入会提示您输入 Kyligence Cloud 许可证，请上传您的 License 文件。

   > 由于华为云的资源部署限制，您可能需要额外等待10-20分钟，等待期服务会显示 `502 Bad Gateway` 错误。如果您的服务在30分钟内仍未部署成功，请联系 Kyligence 技术支持 。
   >
   > 如果您是首次试用 Kyligence Cloud, 您可以在 Kyligence 官网 申请试用 License，然后在**试用欢迎邮件**中下载您的试用 License 文件。

3. 输入下列初始用户名和密码，登录 Kyligence Cloud 。登录后建议您立即修改登录密码。

   - **用户名**：ADMIN
   - **密码**：KYLIN

创建工作区

此小节适用角色：IT 管理员

- 您需要在创建工作区页面输入以下信息：
  - 工作区名称：请输入工作区名称
  - 数据源类型：您都可以选择 **华为云对象存储服务 OBS** 作为数据源。
  - 查询引擎 SSH 密钥：用来访问您的 Kyligence Cloud 的查询引擎的 SSH 密钥
  - 云对象存储服务 OBS 存储桶：请选择您用于存放 Kyligence Cloud 数据的存储桶
  - 集群配置：请输入您的总加载数据量，Kyligence Cloud 将为您推荐集群配置。您也可以点击“启用自定义配置”来调整集群配置

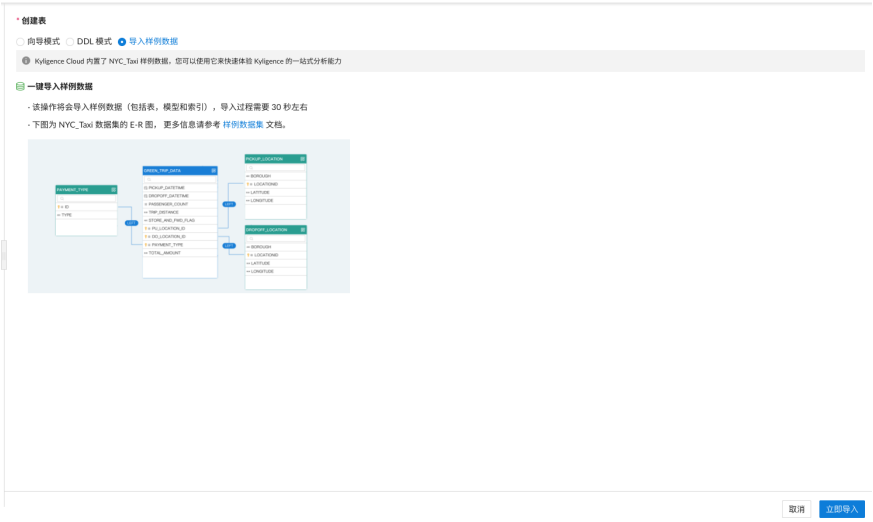
  填写完表单后点击右下角的**审核+创建**，Kyligence Cloud 将自动创建 Spark 集群，创建过程大约需要 5-10 分钟

创建项目、创建表、同步表

此小节适用角色：数据工程师

工作区创建完成后，您需要在工作区内创建项目。点击创建的项目名称，前往**数据源-创建表**页面，点击左侧的 **+ 创建表**，即可创建表到数据目录中。当创建表完成后，即可在**数据源-同步表**页面，将表同步至当前项目，用于后续的建模和分析。

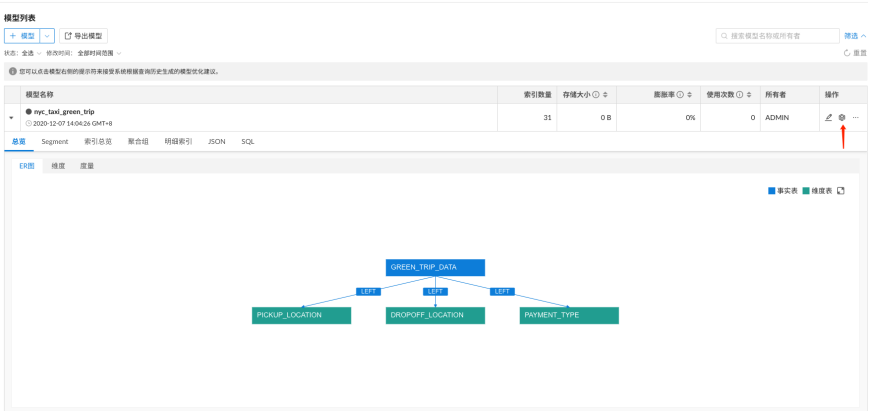
您可选择使用**向导模式**或**DDL模式**来创建表，在本例中我们将使用**导入样例数据**来导入样例数据（包括表，模型和索引），以快速上手。Kyligence Cloud 内置了 NYC_Taxi 数据集，包含了绿色出租车 2019 年 1 月份的出行数据。在本例中，使用此数据集进行分析。在工作区列表点击进入已创建好的工作区，选中左侧 **数据源-创建表** 菜单，点击 **创建表** 按钮，在添加表页面选择 **导入样例表**，点击 **立即导入** 按钮进行模型的导入。关于样例模型的说明及数据字典，请参考NYC_Taxi 数据集。



当页面中出现"导入成功"的提示后，请打开模型界面，您可以看到 **nyc_taxi_green_trip** 即为导入的样例模型。

> 如果您需要使用样例表进行建模，请参考模型章节来创建模型

为了加速查询，请您单击 nyc_taxi_green_trip 模型旁的 **构建索引** 按钮，为模型加载数据，用于接下来的分析。



您可查看该模型下的维度和度量，构建完成后如果查询该模型下的维度和度量时查询将被加速。首次构建数据大概需要 5-8 分钟，您可以点击左侧导航栏中的 **任务** 页面查询进度，构建完成后即可使用该数据集进行分析。构建完成后可点击左侧导航栏中的 **查询**，使用 SQL 进行数据查询。

分析数据

此小节适用角色：数据分析师

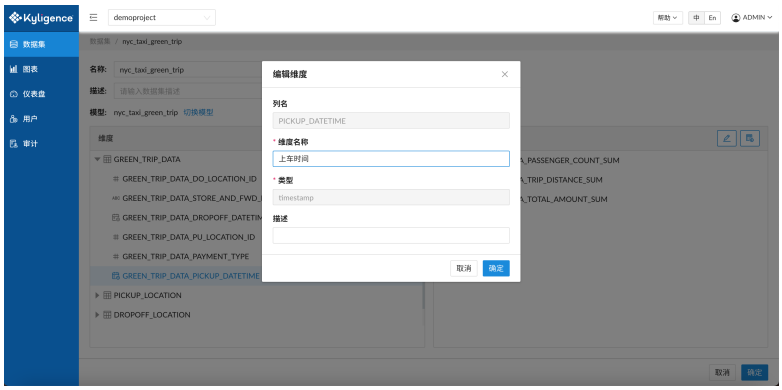
Kyligence Cloud 内置一个可视化分析工具 Kyligence Insight，本节以 Kyligence Insight 为例介绍数据分析的过程，具体步骤如下:

安装并启动 Kyligence Insight

在 **连接 BI** 页面点击连接内置的 Kyligence Insight 下方的 **安装并启动**，等待安装成功后打开 Kyligence Insight。

创建数据集

1. 新建数据集：选择 sample 项目，进入数据集页面。然后点击左上角的 **+数据集** 按钮，并选择数据集用途为 **SQL数据集**。
2. 定义数据集：首先在 **基本信息** 中输入数据集名称为 “Nyc_Taxi”，点击下一步。在 **定义关系** 中将所需的模型拖拽到右侧，然后点击下一步
3. 定义语义：在 **定义语义** 中，您可以进行如下定义:
   - 点击维度名称右侧的编辑按钮，以更改维度名称，例如，将 *GREEN_TRIP_DATA* 表中的 *PICKUP_DATETIME* 重命名为“上车时间”



   - 点击度量名称的编辑按钮，以更改度量名称

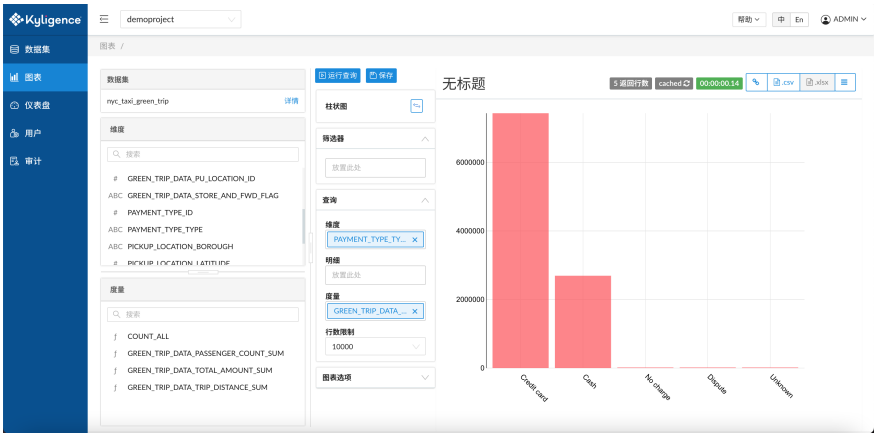
在所有的定义完成之后，点击右下角的保存按钮即可保存新建的数据集。

创建分析图表

点击导航栏中的 **图表**，然后点击左上角的 **+图表** 按钮，选择新建的 “Nyc_Taxi” 数据集，点击 **确认** ，进入分析界面

拖拽所需的维度和度量到右侧，然后点击左上角的运行查询，即可运行查询，得到分析图表

例如，首先点击页面中的切换可视化类型按钮，选择可视化类型为“柱状图”。将“PYAMENT_TYPE_TYPE”拖入到维度中，将”GREEN_TRIP_DATA_TOTAL_AMOUNT_SUM”拖入到度量中，即可得到每个类型对应的订单金额总数



得到分析结果集之后，您可以点击页面中的 **保存** 按钮以保存您的图表，也可以点击右上角的 **导出CSV** 按钮，将查询结果集下载到本地。

清理资源

此小节适用角色：IT 工程师

您可根据您的需求选择以下方式释放您的云上资源以节省成本。

- 方式1 ——停止工作区：如您之后还需要继续使用此工作区的数据和服务，您可以在工作区列表页面 **停止** 工作区，工作区停止后此工作区部署的计算资源将被删除，但数据存储服务仍会保留，您在此工作中的报表、模型、索引均会保留，但在停止状态下此工作区无法提供查询和构建服务，当您有需要时可以随时 **启动** 此工作区。
- 方式2——删除工作区：如您不在需要使用此工作区您可以选择 **删除** 工作区，删除后此工作区的计算和存储资源均会删除，删除后数据将无法恢复。

样例数据集

Kyligence Cloud 内置以下样例数据集，样例数据集数据字典如下：

NewYork Taxi 样例数据集

SSB 样例数据集

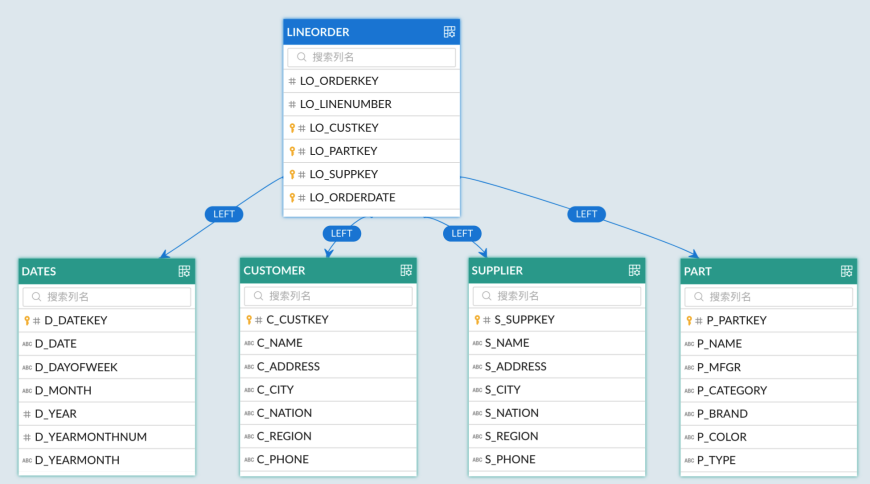
SSB 数据集

Kyligence 为您一套标准的 SSB 数据集用于试用，该数据集的事实表包含 6 亿行数据，本章节将为您介绍 SSB 数据集的数据字典。

SSB 数据集包含 5 张表：

- **LINEORDER** 事实表，描述销售订单的明细信息，每一行对应着一笔交易订单，包含了客户、供应商、订单金额、销售日期等信息。
- **CUSTOMER** 维度表，描述用户的信息，包含用户名称、地址、城市等。
- **SUPPLIER** 维度表，描述供应商的详细介绍，例如供应商名称、地址、电话等。
- **DATE** 维度表，描述了近七年的日期信息。如某个日期所在的年份、月份、星期等。
- **PART** 维度表，描述了零件信息，如零件的名称、类别、颜色、型号等。

这 5 张表共同构成了星型模型的结构。下面是它们的关系图：



实例关系图

数据字典

表	字段	意义
LINEORDER	ORDERKEY	订单ID
LINEORDER	CUSTKEY	顾客ID
LINEORDER	PARTKEY	零件ID
LINEORDER	SUPPKEY	供应商ID
LINEORDER	ORDERDATE	订单日期
LINEORDER	ORDERPRIORITY	订单优先级
LINEORDER	SHIPPRIORITY	交易优先级
LINEORDER	QUANTITY	数量
LINEORDER	EXTENDEDPRICE	额外费用
LINEORDER	ORDTOTALPRICE	订单总额
LINEORDER	DISCOUNT	折扣
LINEORDER	REVENUE	收入
LINEORDER	SUPPLYCOST	供应成本
LINEORDER	TAX	税率
LINEORDER	COMMITDATE	交易日期
LINEORDER	SHIPMODE	交易模式
CUSTOMER	CUSTKEY	客户ID
CUSTOMER	NAME	客户名称
CUSTOMER	ADDRESS	客户地址
CUSTOMER	CITY	客户城市
CUSTOMER	NATION_PREFIX	国家代号
CUSTOMER	NATION	国家
CUSTOMER	REGION	区域
CUSTOMER	PHONE	电话
CUSTOMER	MKTSEGMENT	市场部门
SUPPLIER	SUPPKEY	供应商ID
SUPPLIER	NAME	供应商名称
SUPPLIER	ADDRESS	供应商地址
SUPPLIER	CITY	供应商城市

表	字段	意义
SUPPLIER	NATION_PREFIX	国家代号
SUPPLIER	NATION	国家
SUPPLIER	REGION	区域
SUPPLIER	PHONE	电话
DATE	DATEKEY	日期ID
DATE	DATE	日期
DATE	DAYOFWEEK	星期几
DATE	MONTH	月份
DATE	YEAR	年份
DATE	YEARMONTHNUM	年份数
DATE	YEARMONTH	年月数
DATE	DAYNUMINWEEK	周天数
DATE	DAYNUMINMONTH	月天数
DATE	DAYNUMINYEAR	年天数
DATE	MONTHINYEAR	年月数
DATE	WEEKNUMINYEAR	年周数
DATE	SELLINGSEASON	出售季节
DATE	LASTDAYINWEEKFL	星期最后一天
DATE	LASTDAYINMONTHFL	月份最后一天
DATE	HOLIDAYFL	假日
DATE	WEEKDAYFL	工作日
PART	PARTKEY	零件ID
PART	NAME	零件名称
PART	MFGR	生产商
PART	CATEGORY	种类
PART	BRAND	品牌
PART	COLOR	颜色
PART	TYPE	类型
PART	SIZE	型号

表	字段	意义
PART	CONTAINER	容量

NYC_Taxi 数据说明

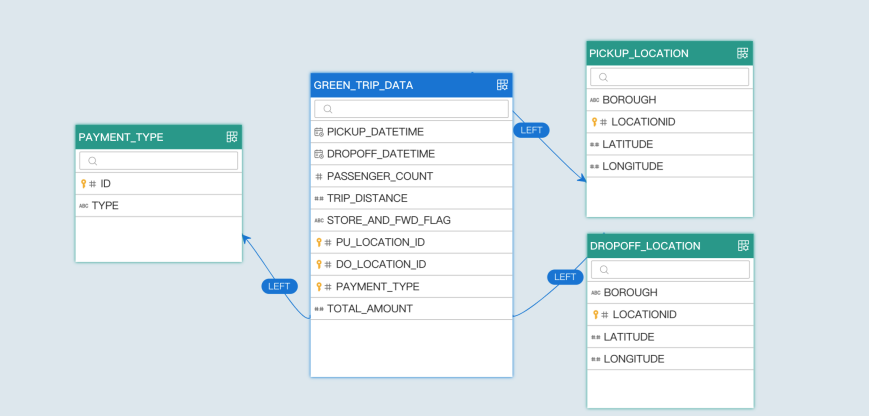
Kyligence Cloud 内置了 NYC_Taxi 数据集，包含了绿色出租车2019年1月份的出行数据(事实表为 63 万行)。如果要快速上手 Kyligence Cloud，可以使用此数据集进行分析。

NYC_Taxi 数据集包含三个表：

- **GREEN_TRIP_DATA** 事实表，包含行程记录的详细信息。它包括上车时间，下车时间，行驶距离以及驾驶员报告的乘客人数等字段。
- **PAYMENT_TYPE** 维度表，描述了付款类型。
- **LOCATION** 维度表，描述上下车的位置，包括区域，经度，纬度。

这三个表共同构成了星型模型的结构。下面是它们的关系图：

> LOCATION 表被使用了两次，分别与 GREEN_TRIP_DATA 表的 PULOCATIONID 字段、DOLOCATIONID 字段 join。



数据字典

表名	字段名	含义
GREEN_TRIP_DATA	PICKUP_DATETIME	上车时间
GREEN_TRIP_DATA	DROPOFF_DATETIME	下车时间
GREEN_TRIP_DATA	PASSENGER_COUNT	驾驶员报告的乘客人数
GREEN_TRIP_DATA	TRIP_DISTANCE	行驶距离
GREEN_TRIP_DATA	STORE_AND_FWD_FLAG	出行记录是否在发送给供应商之前已保存
GREEN_TRIP_DATA	PU_LOCATION_ID	上车位置ID
GREEN_TRIP_DATA	DO_LOCATION_ID	下车位置ID
GREEN_TRIP_DATA	PAYMENT_TYPE	付款类型ID
GREEN_TRIP_DATA	TOTAL_AMOUNT	总费用
PAYMENT_TYPE	ID	付款类型ID
PAYMENT_TYPE	TYPE	付款类型
LOCATION	LOCATIONID	位置ID
LOCATION	BOROUGH	区域
LOCATION	LATITUDE	纬度
LOCATION	LONGITUDE	经度

Kyligence Cloud 在线试用

Kyligence Cloud 提供免费在线试用，您无需部署任何服务即可免费使用，
Kyligence 提供快速入门教程，助您体验端到端分析流程：

- 数据准备：内置样例数据集，您也可支持连接至云上数据湖或上传您的 CSV 数据集
- 智能建模：导入常用 SQL 语句，AI 引擎自动完成数据模型开发和优化
- 高性能分析：高性能 OLAP 加速大数据 BI 分析，标准接口支持各种分析工具

通过您可在 Kyligence 官网申请免费在线试用