腾讯云数据仓库 TCHouse-X

开发指南

腾讯云

文档目录

开发指南
SparkJar 作业开发指南

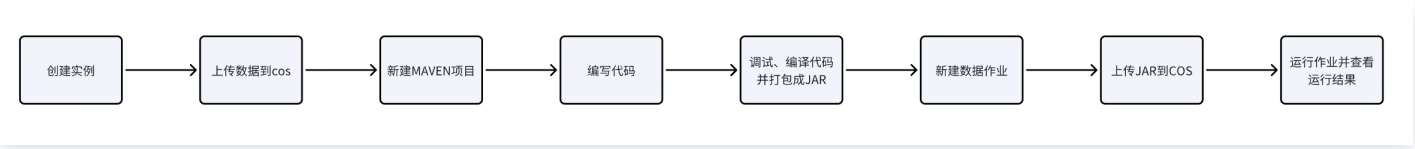
最近更新时间：2026-05-06 16:28:13

应用场景

TCHouse-X 提供完全兼容开源 Apache Spark 的企业级增强 Spark 引擎，支持 Spark 3.5.3 版本，支持用户编写业务程序在 TCHouse-X 上对数据进行读写和分析。本示例演示通过编写 Java 代码在 TCHouse-X 上读写和处理数据的操作。

开发流程

TCHouse-X Spark JAR 作业开发流程如下：



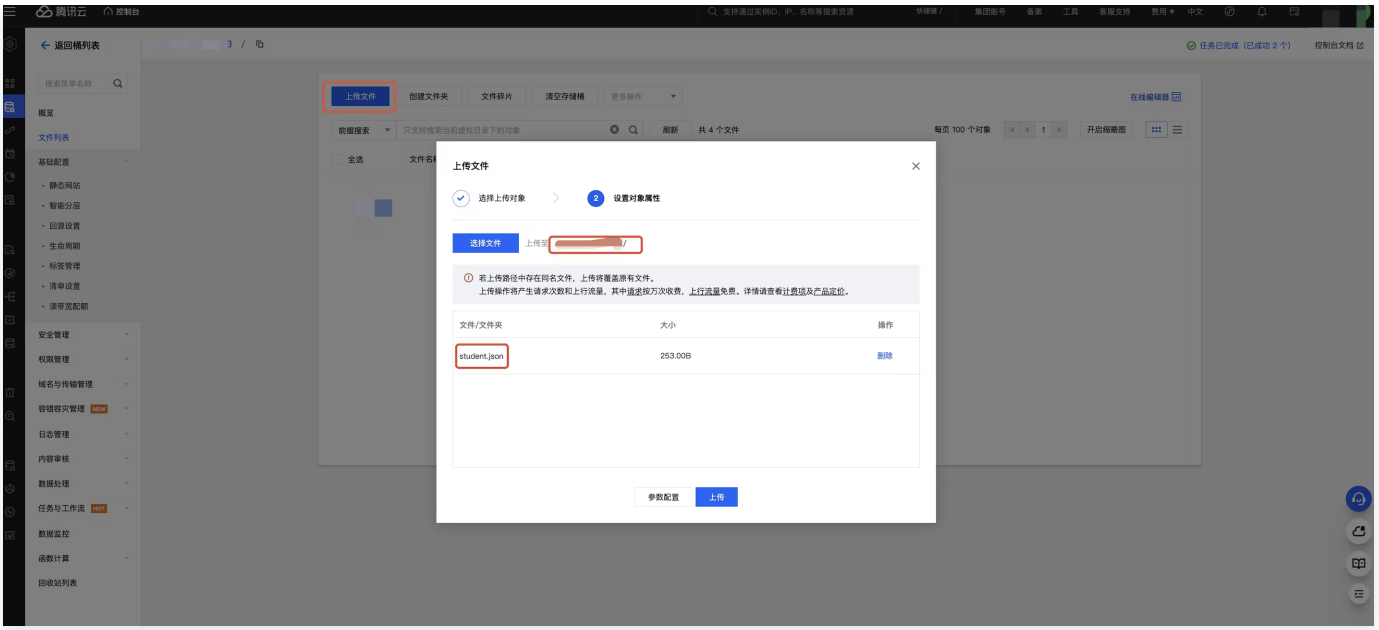
创建实例

首次在 TCHouse-X 上运行作业，需要先创建实例，在实例工作空间下运行 Spark 任务。实例创建流程请参见 实例创建与销毁 。

上传数据到 COS

1. 创建存储桶：请参见 创建存储桶 。
2. 上传文件：

   单击**文件列表** > **上传文件**，选择已下载的本地 student.json 文件上传到所创建桶里，单击上传，完成文件上传。



新建 Maven 项目

1. 通过 IntelliJ IDEA 新建一个名称为“demo”的 Maven 项目。
2. 添加依赖：在 pom.xml 中添加如下依赖：

```
<dependency>
    <groupId>org.apache.spark</groupId>
    <artifactId>spark-core_2.12</artifactId>
    <version>3.5.3</version>
    <scope>provided</scope>
</dependency>
<dependency>
    <groupId>org.apache.spark</groupId>
    <artifactId>spark-sql_2.12</artifactId>
    <version>3.5.3</version>
    <scope>provided</scope>
</dependency>
<dependency>
    <groupId>org.apache.spark</groupId>
    <artifactId>spark-hive_2.12</artifactId>
    <version>3.5.3</version>
    <scope>provided</scope>
</dependency>
```

编写代码

编写代码功能为从 COS 上读写数据和在 TCHouse-X 上建库、建表、查询数据和写入数据。

1. 从 COS 上读写数据代码示例：

```
package com.tencent.tchouse;

import org.apache.spark.sql.Dataset;

import org.apache.spark.sql.Row;

import org.apache.spark.sql.SaveMode;

import org.apache.spark.sql.SparkSession;


public class IOTest {

    public static void main(String[] args) {

        //1.创建SparkSession

        SparkSession spark = SparkSession

                .builder()

                .appName("io test")

                .config("spark.some.config.option", "some-value")

                .getOrCreate();

        //2.读取cos上的 json 文件

        String readPath = "cosn://your_cos_bucket_name/student.json";

        Dataset<Row> readData = spark.read().json(readPath);

        //3.对数据集做业务计算操作生成结果数据，计算支持API和SQL形式，这里生成临时表用sql读数据

        readData.createOrReplaceTempView("student");
```

```
        Dataset<Row> result = spark.sql("SELECT * FROM student limit
10");

        result.show();

        //4.结果数据保存到cos

        String writePath =
"cosn://your_cos_bucket_name/student_output";

        result.write().mode(SaveMode.Append).parquet(writePath);

        //5.关闭session

        spark.stop();
    }
}
```

2. TCHouse-X 上建库、建表、查询数据和写入数据：

```
package com.tencent.tchouse;

import org.apache.spark.sql.SparkSession;

public class SQLTest {
    public static void main(String[] args) {
        //1.创建SparkSession
        SparkSession spark = SparkSession
                .builder()
                .appName("sql test")
                .config("spark.some.config.option", "some-value")
                .enableHiveSupport()
                .getOrCreate();
        // 1.建数据库
        spark.sql("CREATE DATABASE IF NOT EXISTS `test_db` COMMENT
'test db'");

        // 2.建内表
```

```
        spark.sql("CREATE TABLE test_db.student (id INT, name STRING,
age INT) STORED AS PARQUET
TBLPROPERTIES('enable.TCHouseXStorage'='true','TCHouseXStorage.enableU
pdate'='true')");
        // 3.写内数据
        spark.sql("INSERT INTO test_db.student VALUES (1,'Andy',12),
(2,'Justin',3) ");
        // 4.查内数据
        spark.sql("SELECT * FROM test_db.student LIMIT 10").show();

        // 5.建外表
        spark.sql("CREATE EXTERNAL TABLE test_db.ex_student (id INT,
name STRING, age INT) STORED AS PARQUET LOCATION 'cosn://demo-
1301087413/external_data'");
        // 6.写外数据
        spark.sql("INSERT INTO test_db.ex_student VALUES
(1,'Andy',12),(2,'Justin',3) ");
        // 7.查外数据
        spark.sql("SELECT * FROM test_db.ex_student").show();
    }
}
```

> ⚠ **注意：**
> 请将代码中的文件读取地址更改为您存储文件的 COS 地址。

调试、编译代码并打成 JAR 包

通过 IntelliJ IDEA 对 demo 项目编译打包，在项目 target 文件夹下生成 JAR 包 .jar。

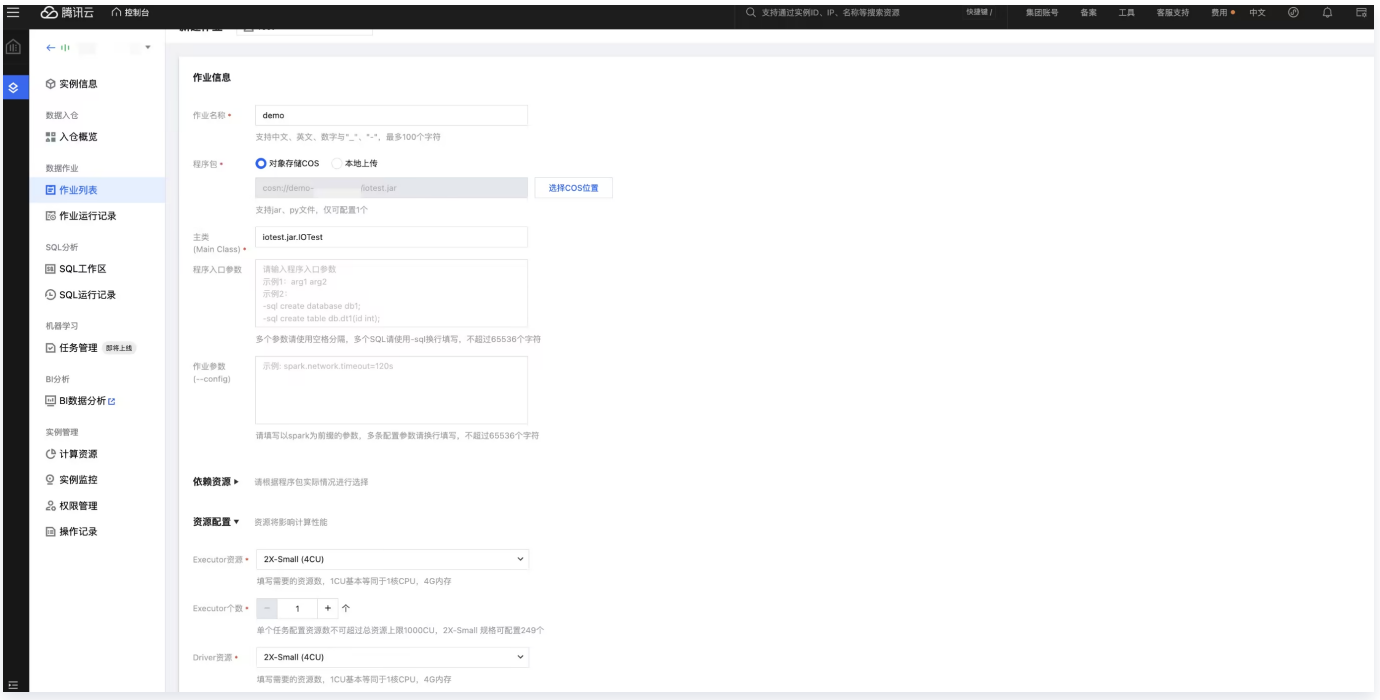
新建数据作业

1. 登录 TCHouse-X 控制台，进入实例，单击左侧菜单**数据作业-作业列表**进入数据作业管理页。
2. 单击**新建作业**按钮，进入创建页。
3. 在作业配置页面，配置作业运行参数，具体说明如下：

配置参数	说明
作业名称	自定义 Spark JAR 作业名称，例如：demo

程序包	上传 JAR 包到 COS 或者使用本地上传
主类（Main Class）	iotest.jar.IOTest
资源配置	配置作业运行所需要的资源，计量单位为 CU，1CU=1核 CPU+4G 内存 • Executor 资源 • Executor 个数 • Driver 资源

其他参数值保持默认。

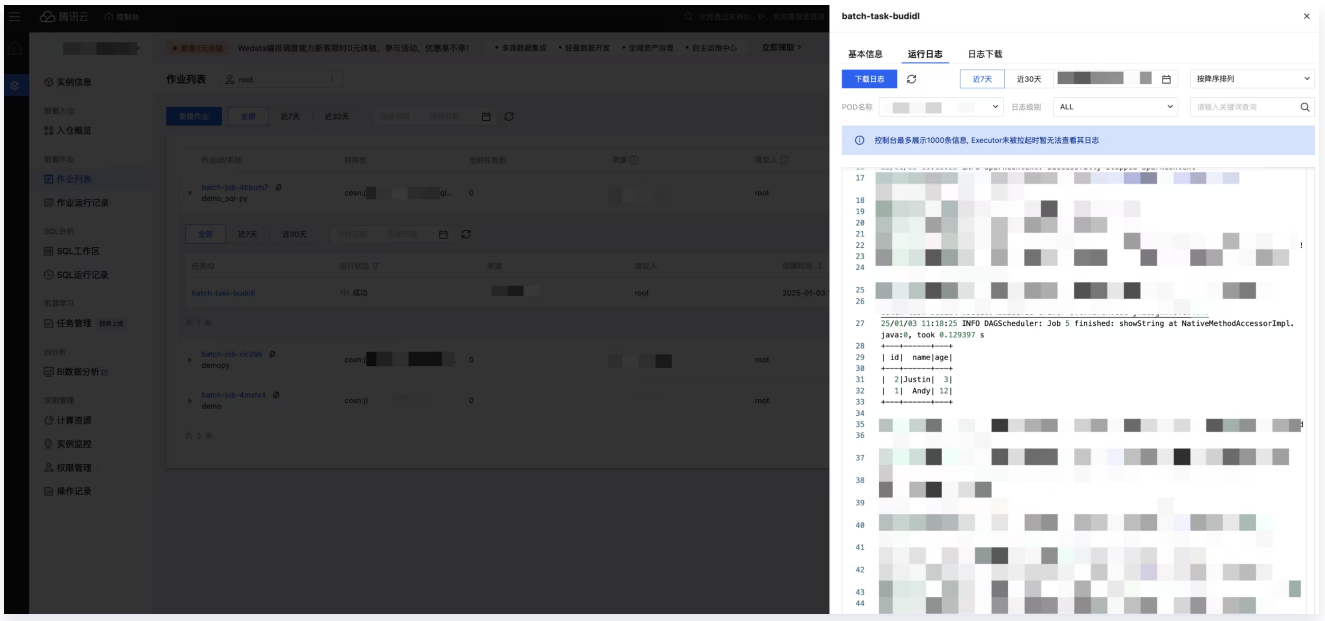


4. 点击创建并启动或者仅创建，在作业列表页面可以看到创建的作业。

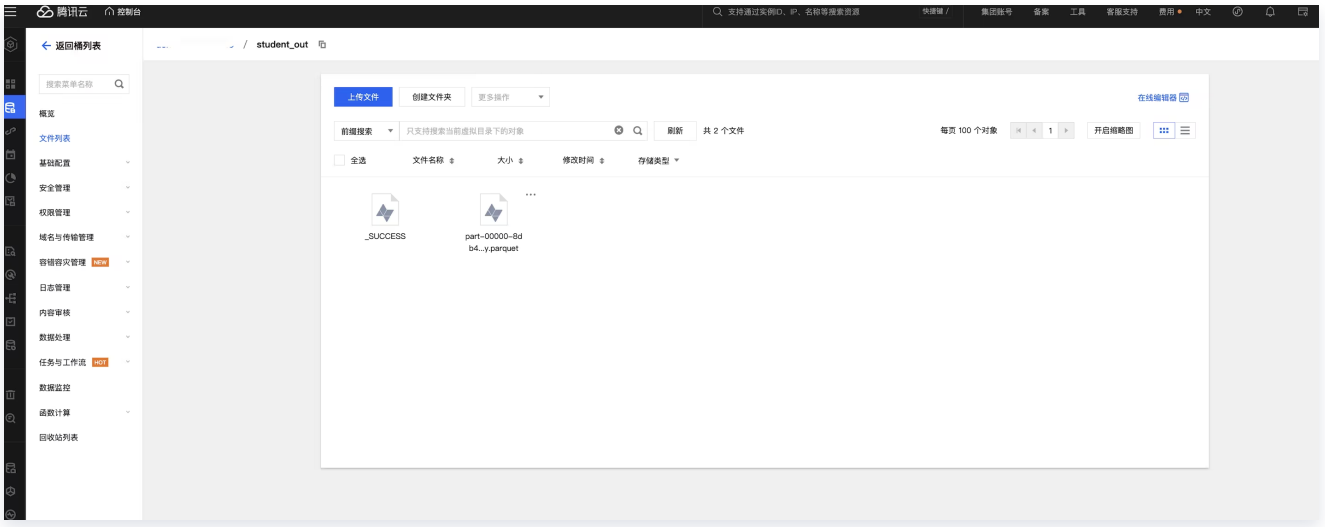
运行并查看作业结果

1. 运行作业：在作业列表页面，找到新建的作业，单击启动，即可运行作业。
2. 作业列表页面对应作业子列表或者作业运行记录页可以查看作业任务
3. 查看作业运行结果：

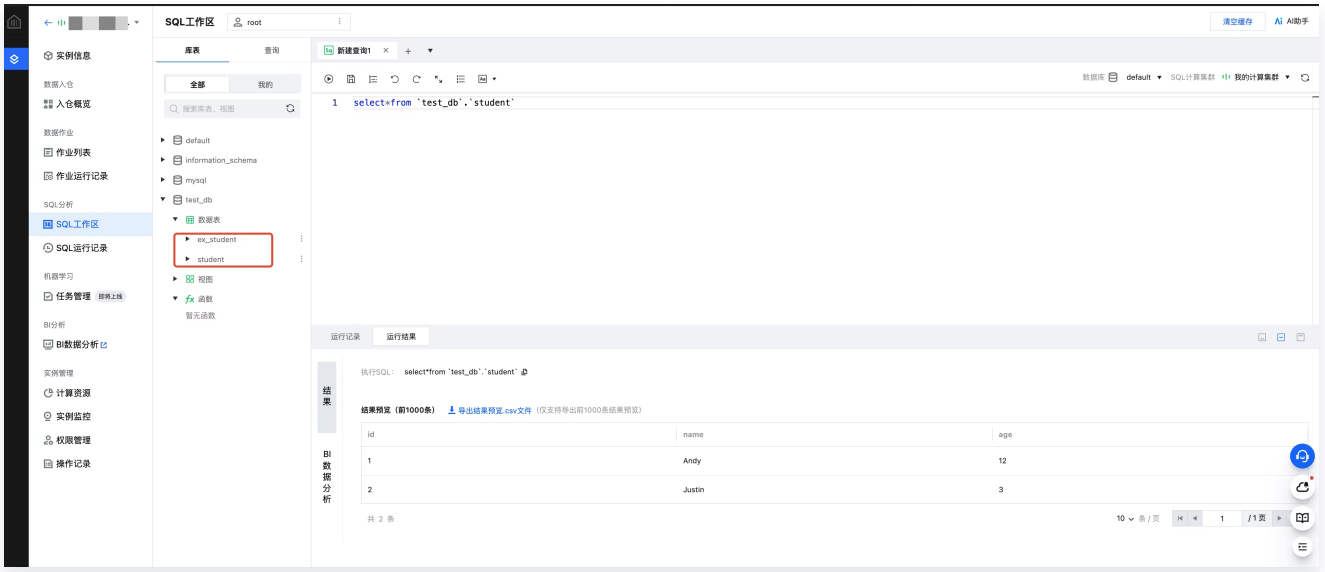
   3.1 查看作业运行日志：点击任务的**查看详情**，进入运行日志页面查看日志



3.2 运行从 COS 读写数据示例，则到 COS 控制台查看数据写入结果。



3.3 运行在 TCHouse-X 上建表、建库，则到 TCHouse-X 的 **SQL 工作区**页面查看建库、建表。



PySpark 作业开发指南

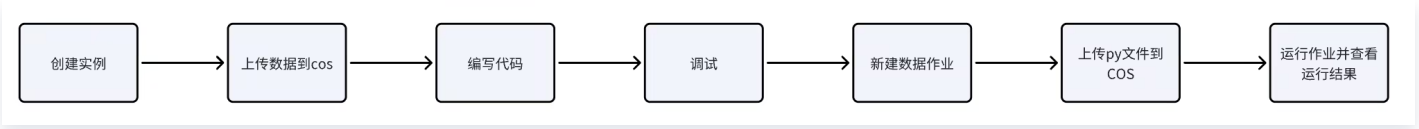
最近更新时间：2026-05-06 16:28:13

应用场景

TCHouse-X 提供完全兼容开源 Apache Spark 的企业级增强 Spark 引擎，支持 Spark 3.5.3 版本，支持用户编写业务程序在 TCHouse-X 上对数据进行读写和分析。本示例演示通过编写 Python 代码在 TCHouse-X 上读写和处理数据的操作。

开发流程

TCHouse-X Spark Python 作业开发流程如下：



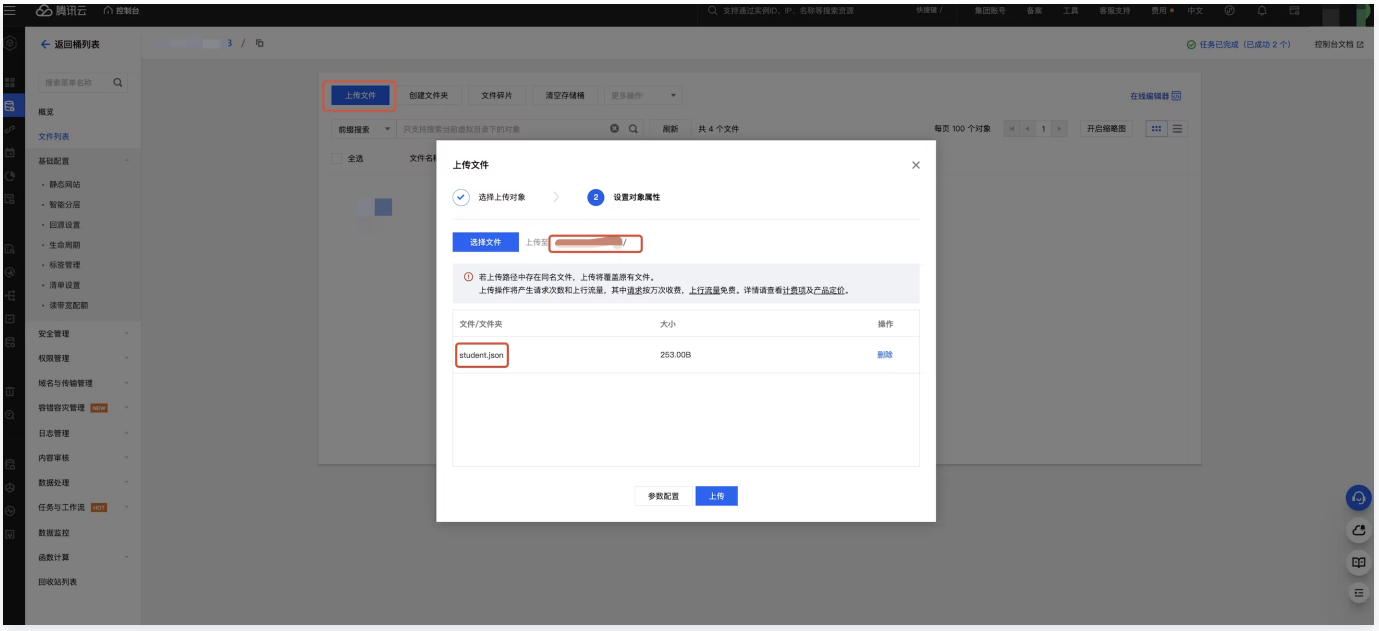
创建实例

首次在 TCHouse-X 上运行作业，需要先创建实例，在实例工作空间下运行 Spark 任务。实例创建流程请参见 实例创建与销毁 。

上传数据到 COS

1. 创建存储桶：请参见 创建存储桶 。
2. 上传文件：

   单击文件列表 > 上传文件，选择已下载的本地 student.json 文件上传到所创建桶里，单击上传，完成文件上传。



构建虚拟环境&安装依赖

```
mkdir spark-dev && cd spark-dev
python3 -m venv venv
source venv/bin/activate
pip install "pyspark[sql]==3.5.3"
```

导入 Python 项目

打开 PyCharm > File > Open，选择 spark-dev 目录。

编写代码

1. 新建 cosio.py 文件，编写代码，功能为从 COS 上读写数据。

```
  from pyspark.sql import SparkSession

if __name__ == '__main__':
    spark = SparkSession \
        .builder \
        .appName("io test") \
        .getOrCreate()

    # 1.读cos上的数据
    read_path = "cosn://your_cos_bucket_name/student.json"
    student_df = spark.read.json(read_path)

    # 2.对数据做操作
    student_df.createOrReplaceTempView("student")
    data_df = spark.sql("SELECT * FROM student limit 10")
    data_df.show()

    # 3.写数据
    write_path = "cosn://your_cos_bucket_name/student_out"
    data_df.write.parquet(path=write_path)

    spark.stop()
```

2. 新建 sql_test.py 文件，编写代码，功能为在 TCHouse-X 上建库、建表、查询数据和写入数据。

```
  from pyspark.sql import SparkSession

if __name__ == '__main__':
    spark = SparkSession \
        .builder \
        .appName("sql test") \
        .enableHiveSupport() \
        .getOrCreate()
    # 1.建数据库
    spark.sql("CREATE DATABASE IF NOT EXISTS test_db COMMENT 'test
db'")

    # 2.建内表
    spark.sql("CREATE TABLE test_db.student (id INT, name STRING, age
INT) USING TCI")

    # 3.写内数据
    spark.sql("INSERT INTO test_db.student VALUES (1,'Andy',12),
(2,'Justin',3) ")
    # 4.查内数据
    spark.sql("SELECT * FROM test_db.student LIMIT 10").show()

    # 5.建外表
    spark.sql("CREATE EXTERNAL TABLE test_db.ex_student (id INT, name
STRING, age INT) USING PARQUET LOCATION
'cosn://your_cos_bucket_name/external_data'")
    # 6.写外数据
    spark.sql("INSERT INTO test_db.ex_student VALUES (1,'Andy',12),
(2,'Justin',3) ")
    # 7.查外数据
    spark.sql("SELECT * FROM test_db.ex_student").show()

    spark.stop()
```

⚠ 注意：
请将代码中的文件读取地址更改为您存储文件的 COS 地址。

调试

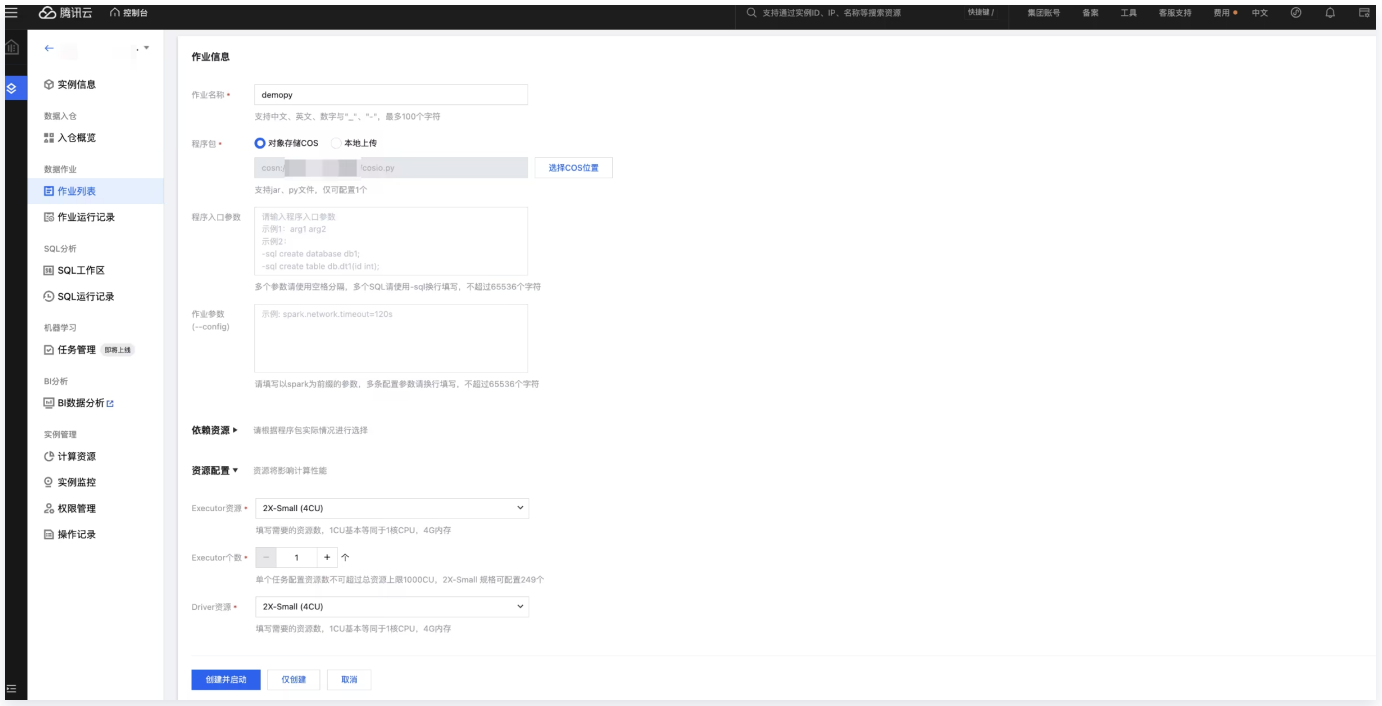
本地 PyCharm 调试无语法错误。

新建数据作业

1. 登录 TCHouse-X 控制台，进入实例，单击左侧菜单**数据作业 > 作业列表**进入数据作业管理页。
2. 单击**新建作业**，进入创建页。
3. 在作业配置页面，配置作业运行参数，具体说明如下：

配置参数	说明
作业名称	自定义 Pyspark 作业名称，例如：demopy
程序包	上传 py 文件到 COS 或者使用本地上传
资源配置	配置作业运行所需要的资源，计量单位为 CU，1CU=1核CPU+4GB 内存 • Executor 资源 • Executor 个数 • Driver 资源

其他参数值保持默认。



4. 单击创建并启动或者仅创建，在作业列表页面可以看到创建的作业。

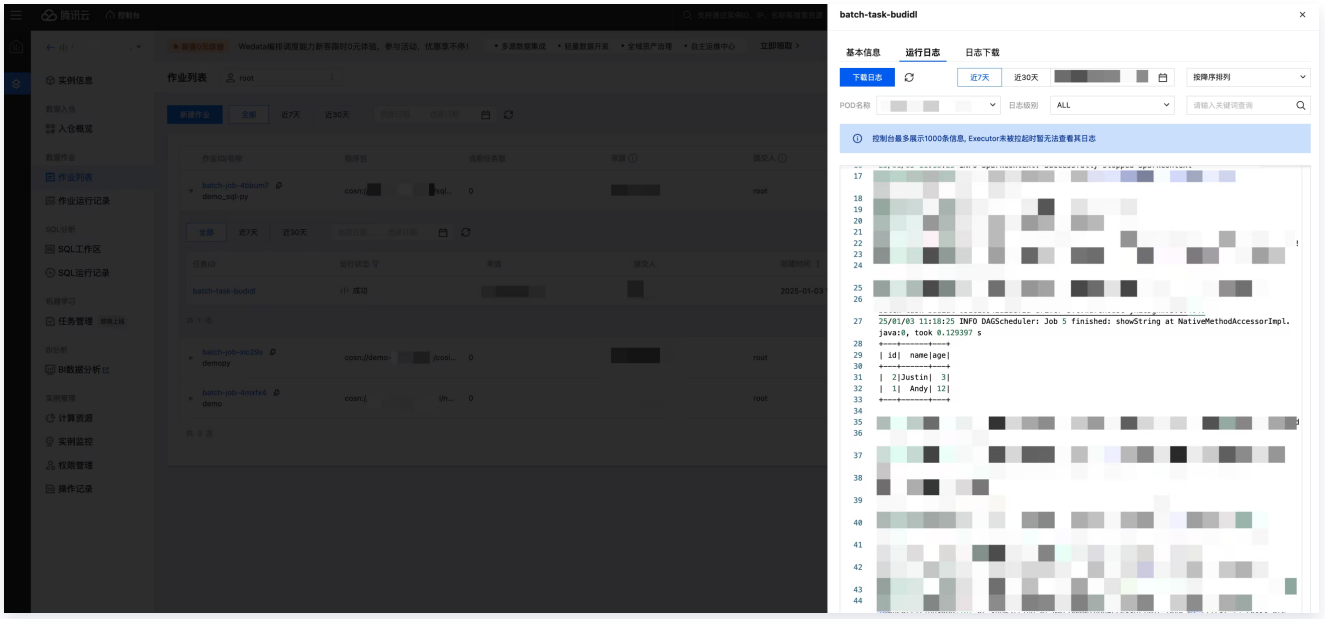
运行并查看作业结果

1. 运行作业：在作业列表页面，找到新建的作业，单击启动，即可运行作业。

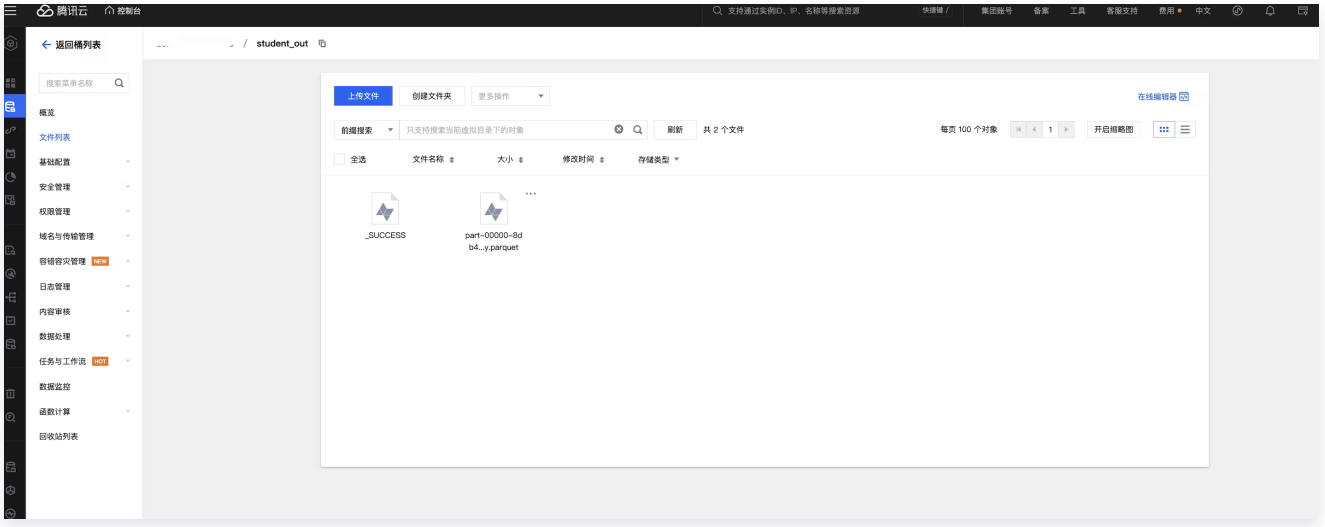
2. 作业列表页面对应作业子列表或者作业运行记录页可以查看作业详情。

3. 查看作业运行结果：

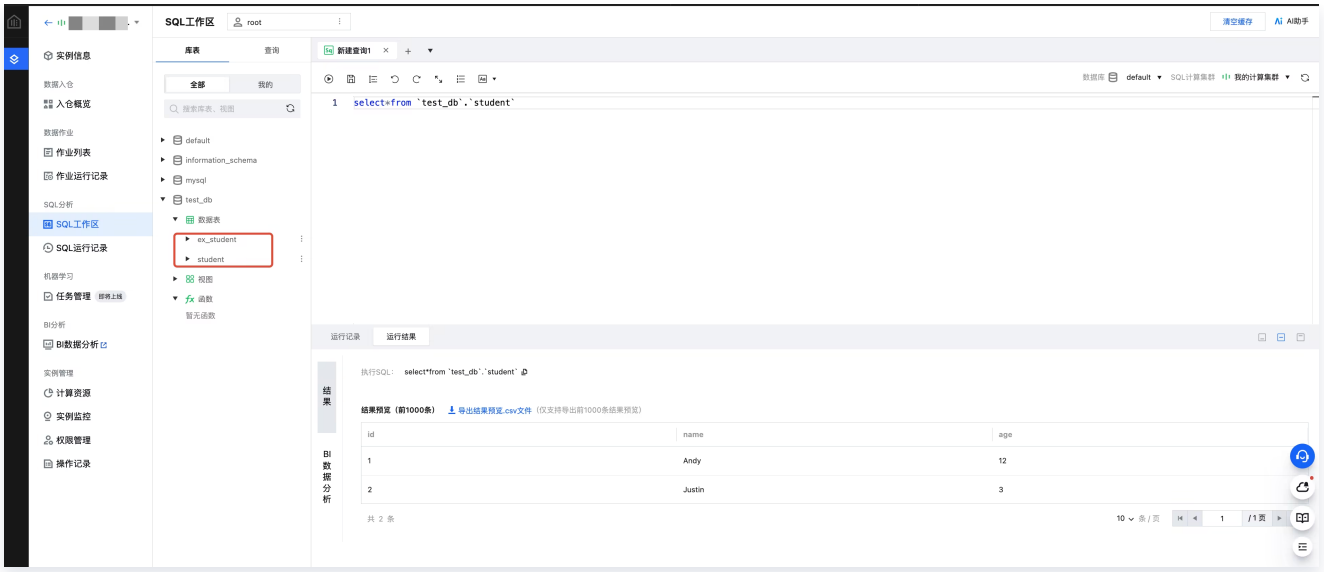
3.1 查看作业运行日志：单击任务的**查看详情**，进入运行日志页面查看日志。



3.2 运行从 COS 读写数据示例，则到 COS 控制台查看数据写入结果。



3.3 运行在 TCHouse-X 上建表、建库，则到 TCHouse-X 的 **SQL 工作区**页面查看建库、建表。


SQL工作区
root
清空缓存
AI助手
实例信息
数据入仓
入仓概览
数据作业
作业列表
作业运行记录
SQL分析
SQL工作区
SQL运行记录
机器学习
任务管理
即将上线
BI分析
BI数据分析
实例管理
计算资源
实例监控
权限管理
操作记录
库表
查询
全部
我的
搜索库表、视图
default
information_schema
mysql
test_db
数据表
ex_student
student
视图
函数
暂无函数
新建查询1
数据库 default
SQL计算集群 我的计算集群
1 select*from `test_db`.`student`
运行记录
运行结果
结果
BI数据分析
执行SQL: select*from `test_db`.`student`
结果预览（前1000条）
导出结果预览.csv文件
（仅支持导出前1000条结果预览）
id
name
age
1
Andy
12
2
Justin
3
共 2 条
10 条 / 页
1
/ 1 页