



在openLookeng中构建一个Connector

黎一泽 <https://gitee.com/armilly>

环节一览

1. 直播小福利
2. 跨源大数据分析查询
3. Connector基本知识
4. 代码讲解-构建csv文件Connector
5. 代码讲解-Connector SQL Query Push Down
6. 问答交流环节

直播间小福利

- openLookEng社区小礼品



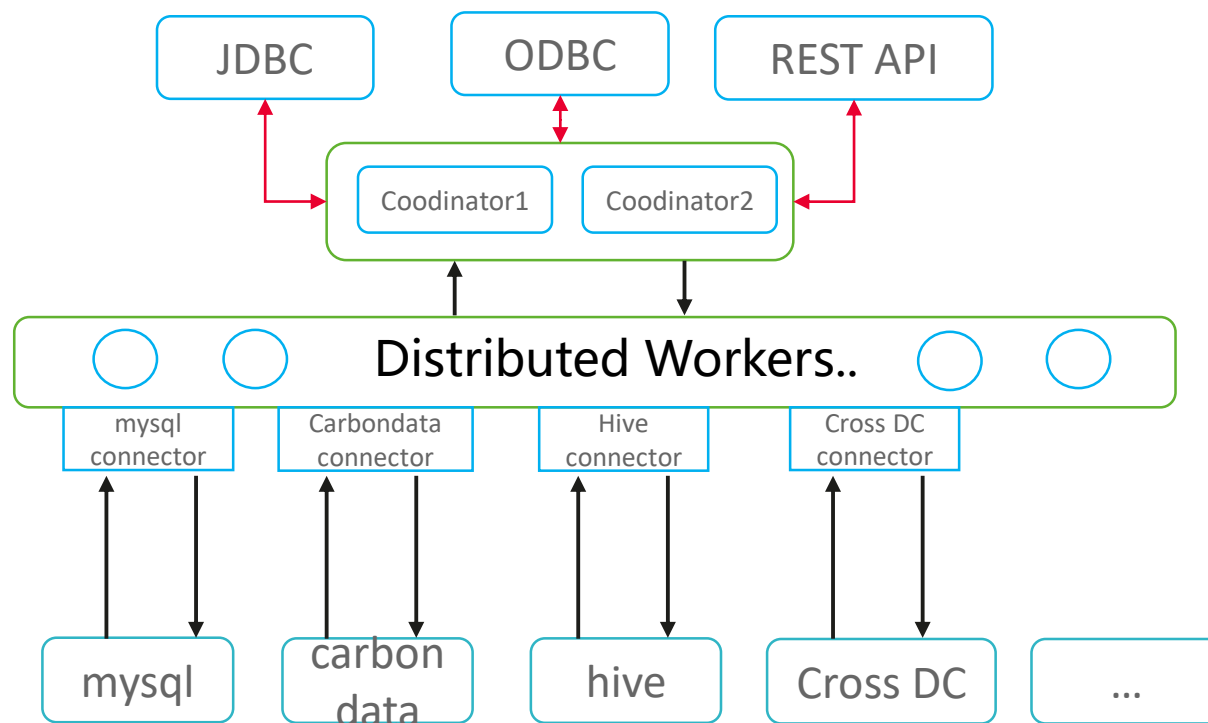
跨源大数据分析查询

- 企业多数据库难题
 - › 在多份数据上需要写多份代码
 - › 分离数据很难碰撞产生新价值
 - › 聚合分析需要ETL操作



跨源大数据分析查询

- openLooKeng特性
 - › 通过SQL 2003提供了所有数据的全局视图
 - › 多样的北向接入方式：JDBC、ODBC、RESTful API
 - › 多样的南向数据源对接：mysql、hive、hbase、Carbondata、Cross DC...



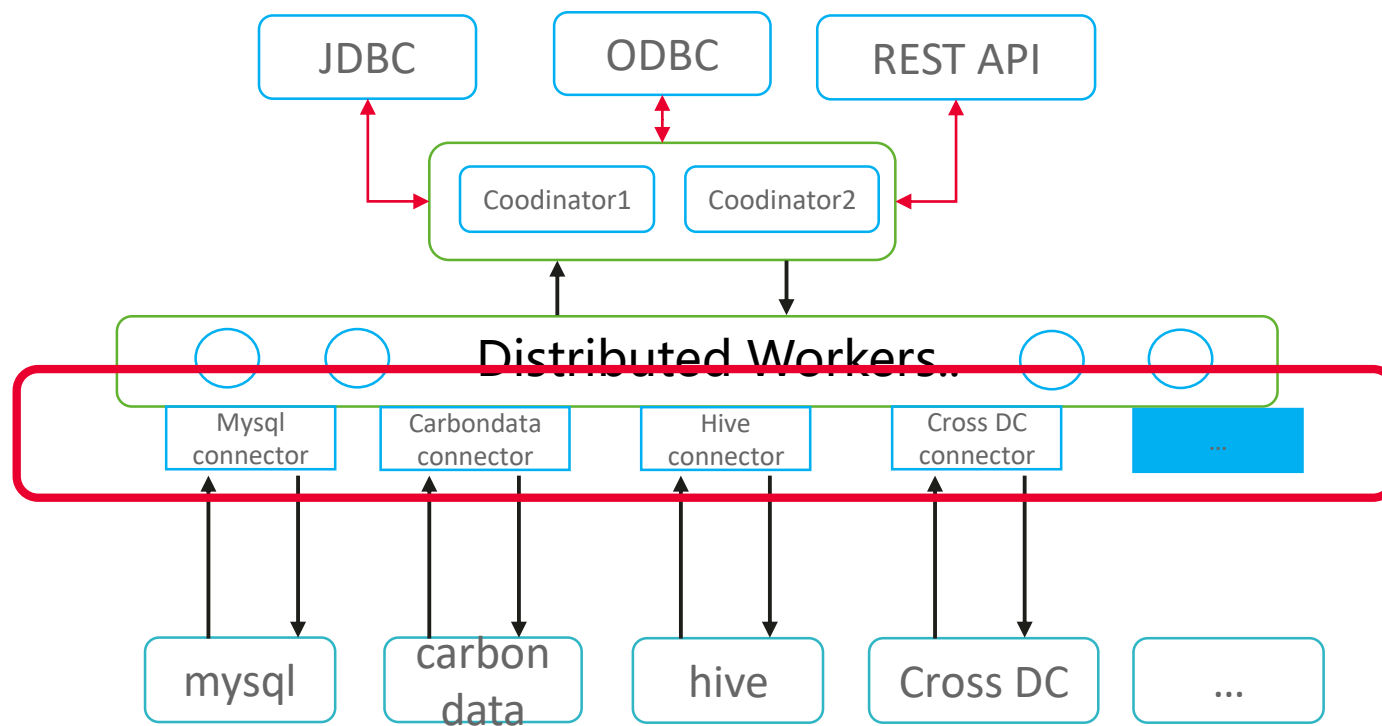
解决跨源大数据分析查询难题-提供统一查询语言

Video https://openlookeng-website.obs.ap-southeast-1.myhuaweicloud.com/highlight1_en.mp4

解决跨源大数据分析查询难题-多数据源对接、轻松扩展

Video https://openlookeng-website.obs.ap-southeast-1.myhuaweicloud.com/highlight2_en.mp4

构建一个Connector



构建一个Connector-基础知识

- 数据源对接

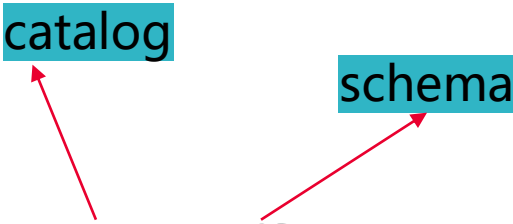
- › openLookEng都通过Connector来对接数据源

- 元数据

- › 在openLookEng中一个数据源的元数据信息有 catalog、schemas、tables、columns

- 数据源数据读写

- › 数据读，在Connector层通过取回数据表格的形式（类似于JDBC的ResultSet）。
 - › 数据写，在Connector层执行具体的操作，例如下发DML SQL语句，更改文件内容等操作。



```
lk> use csv.csv_tables;
USE
lk:csv_tables> show tables;
Table
-----
foo1
foo2
(2 rows)
```

Query 20201105_024937_00009_n4fma, FINISHED, 1 node
Splits: 19 total, 19 done (100.00%)
0:00 [2 rows, 48B] [57 rows/s, 1.34KB/s]

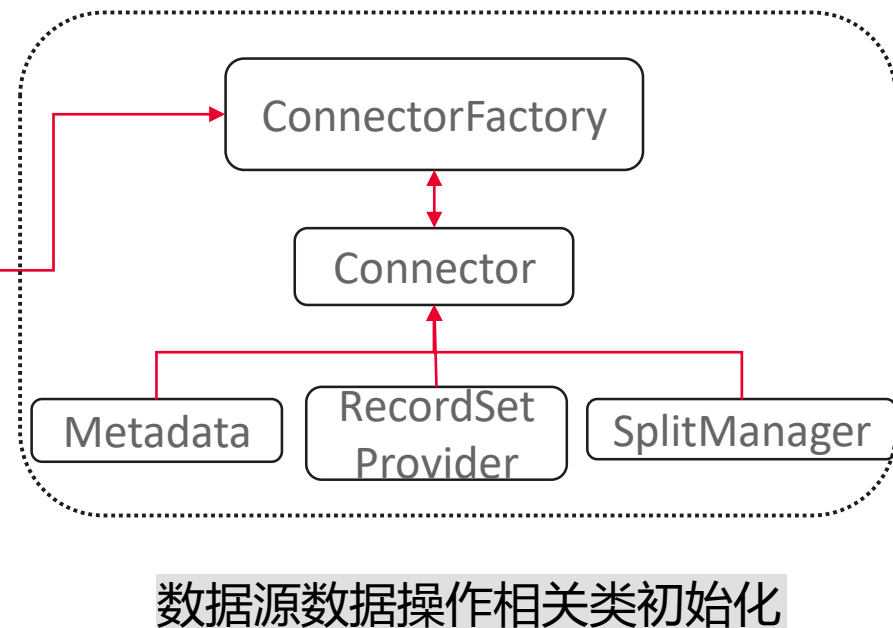
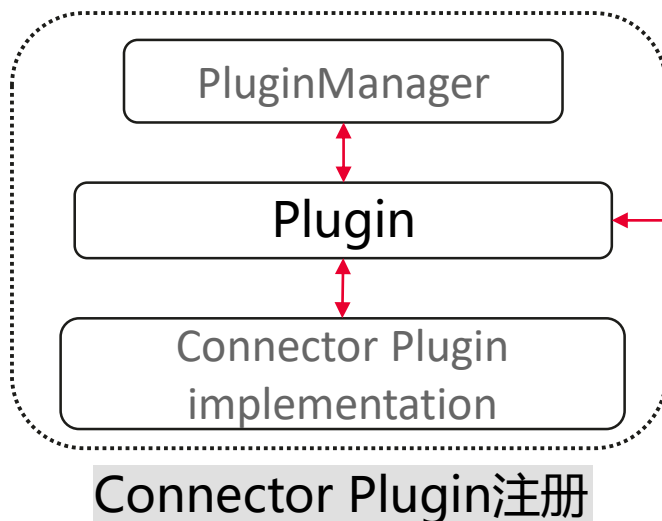
```
lk:csv_tables> desc foo1;
Column | Type | Extra | Comment
-----+-----+-----+-----
col1   | bigint |      | 
col2   | bigint |      | 
(2 rows)
```

Query 20201105_024941_00010_n4fma, FINISHED, 1 node
Splits: 19 total, 19 done (100.00%)
0:00 [2 rows, 126B] [46 rows/s, 2.86KB/s]

```
lk:csv_tables> select * from foo1;
col1 | col2
-----+-----
123  | 3234
111  | 888
222  | 333
(3 rows)
```

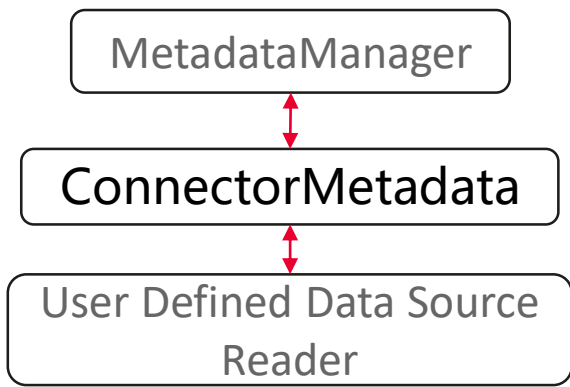
构建一个Connector-基础知识

- Connector 作为Plugin的注册过程
 - 接口、策略模式和配置文件 的设计模式, 通过定制相同的接口
 - java.util.ServiceLoader
- 获取Connector的工厂类
 - ConnectorFactory
- 获取Metadata、SplitManager、RecordSetProvider的接口类
 - Connector

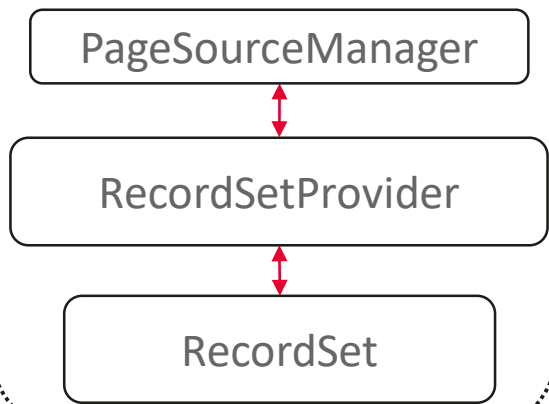


构建一个Connector-基础知识

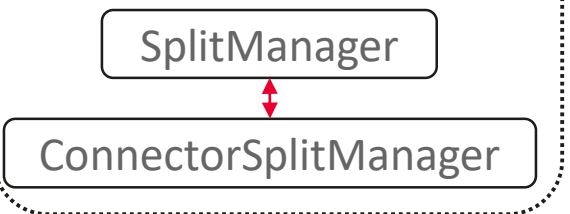
1. Read Metadata



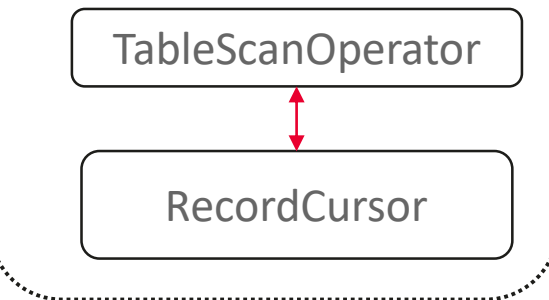
3. Get RecordCursor



2. Get Splits

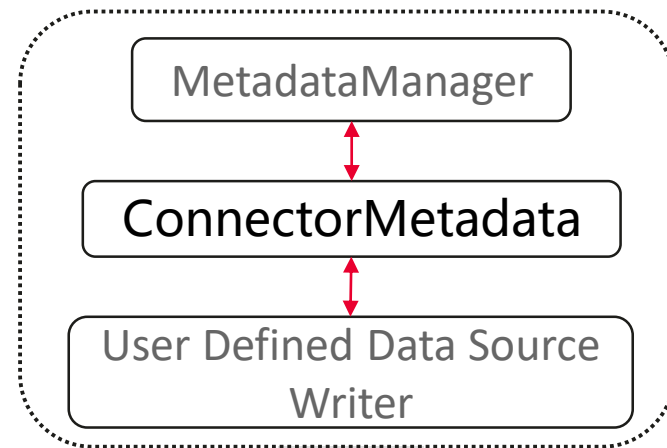


4. Get data columns



读数据

- 通过Connector读写数据
- 读
 - 1. 取元数据
 - 2. 获取Split信息
 - 3. 为每一个Split构建RecordCursor
 - 4. 通过RecordCursor取数据
- 写
 - 修改元数据、数据信息





写数据

构建csv文件Connector

- Plugin实现
 - CsvFilePlugin
- ConnectorFactory实现
 - CsvFileConnectorFactory
 - CsvFileModule
- ConnectorMetadata实现
 - CsvFileMetadata
 - CsvFileTables
 - CsvFileTablesHandle
 - CsvFileColumnHandle
- ConnectorSplit实现
 - CsvFileSplitManager
 - CsvFileSplit
- 数据Cursor实现
 - CsvFileRecordSetProvider
 - CsvFileRecordSet
 - CsvFileRecordCursor
- **User Defined Data Source Reader and Writer**
- 演示



Csv connector代码地址: <https://gitee.com/armilly/hetu-core/tree/feature-csv-connector/>

 **Yize_Li / hetu-core**

forked from [openLookKeng / hetu-core](#) 

[</> 代码](#) [Issues 0](#) [Pull Requests 0](#) [Wiki 3](#) [统计](#) [DevOps](#)

feature-csv-connector ▾

 分支 30  标签 6










+ Pull Request

+ Issue

文件 ▾

Web IDE

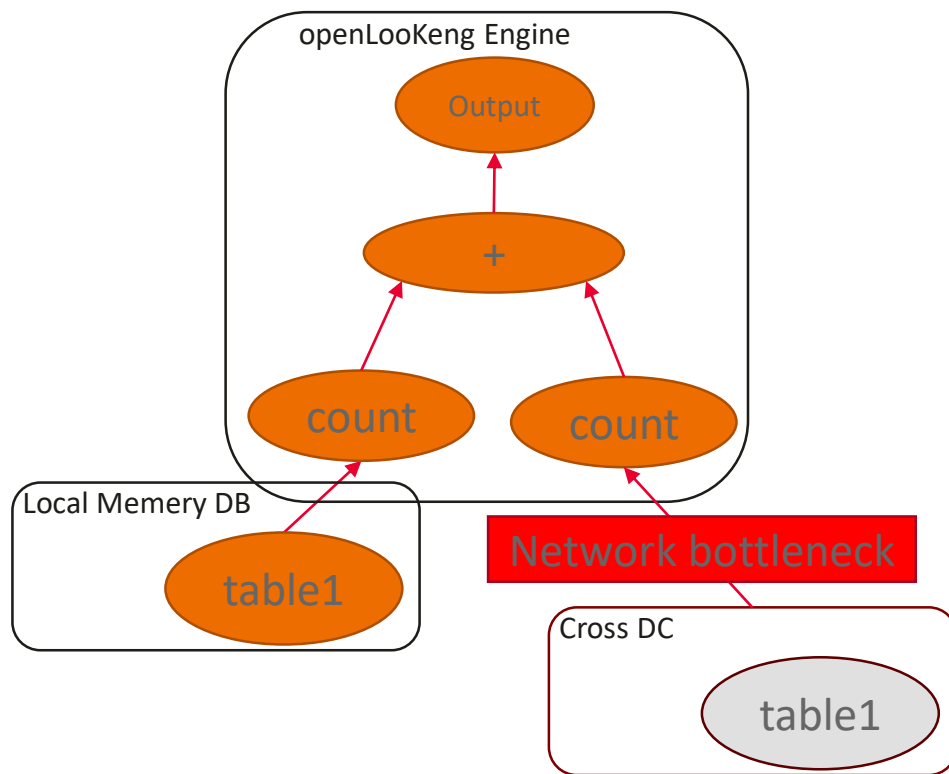
克隆/下载 ▾

 Yize_Li	build csv connector	b1e790a	1小时前	598 次提交
 .gitee	Add issue and PR templates to assist contributors in filling up requir...			4月前
 .github	Add issue and PR templates to assist contributors in filling up requir...			4月前
 .mvn	add the .mvn dir and the relative files.			4月前
 docker	!363 Fix security flaws in run-hetu			6天前
 hetu-carbondata	Json Injection Fix			18天前
 hetu-common	Update parallel index creation logic, improve index exception handling...			18天前
 hetu-csv-file	build csv connector			1小时前
 hetu-datacenter	Change OpenLookKeng version number to have snapshot			1月前

网络瓶颈造成作业执行性能问题

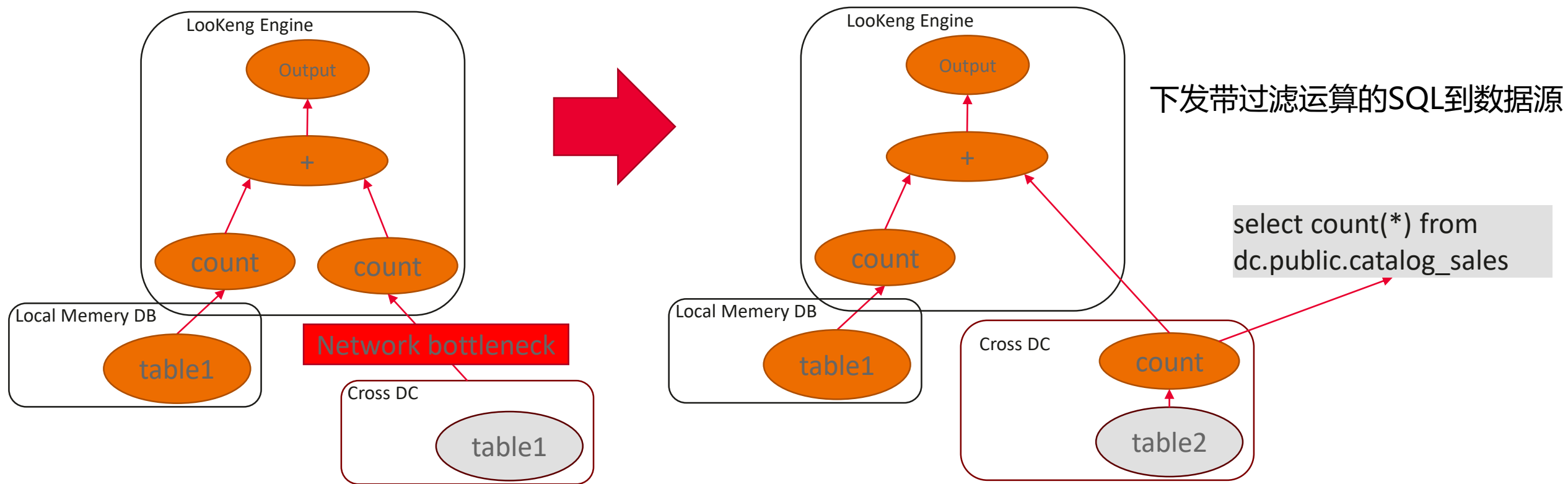
跨源执行SQL语句，网络代价不可忽略

```
select (select count(*) as col2 from tpcds.tiny.catalog_returns) + (select count(*) as col2 from dc.public.catalog_sales)
```



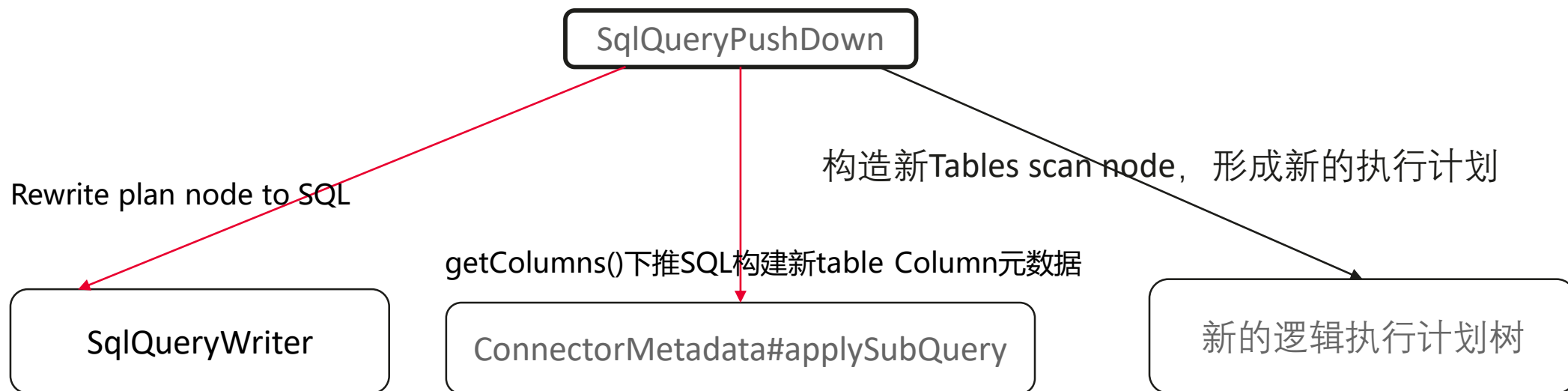
全表取回耗费网络资源，存在网络瓶颈。

Connector SQL Query Push Down



SQL Query Push Down执行计划树优化

Connector SQL Query Push Down



- 实现SQL Query Push Down功能
 - › 实现SqlQueryWriter和ConnectorMetadata#getSqlQueryWrite
 - › 实现ConnectorMetadata#applySubQuery接口
 - › 举个例子：DC connector

Left issues

- 完善CSV file connector demo代码
 - › 当前csv file connector只支持一个固定的schemas，需要支持多schemas(简单)
 - › 当前csv file connector只支持csv文件静态一次初始读入，需要支持动态读取csv文件内容(简单)
 - › 当前csv file connector只支持一种数据类型：bigint，添加多种数据类型的支持(中等)
 - › 当前csv file connector只支持csv文件数据的读入分析，需要添加csv文件的sql表格增删改等修改功能(中等)
 - › 添加代码UT(中等)
- 在MySQL connector中实现Sub Query Push Down功能(简单)
- 欢迎往代码分支合入改进代码，完成后可以微信截图或者其他证据（比如代码连接）告诉小助手，领取社区文化纪念品
 - › Csv file connector demo还在个人仓库：<https://gitee.com/armilly/hetu-core/tree/feature-csv-connector>
 - › 官方Gitee仓库：<https://gitee.com/openlookeng>
 - › 官方Github仓库：<https://github.com/openlookeng>

问答交流环节



微信小助手



微信公众号

主页: <https://openlookeng.io/>

Gitee仓库: <https://gitee.com/openlookeng>

Github仓库: <https://github.com/openlookeng>

Slack: <https://openlookeng.slack.com>

Thank you.