HUAWEI

Day19 打破数据孤岛

—— 异构数据源联合分析业务创新实践

目录

- 异构数据源商业价值
- 案例分享——企业智能化数据经营分析

HUAWEI

异构数据源商业价值

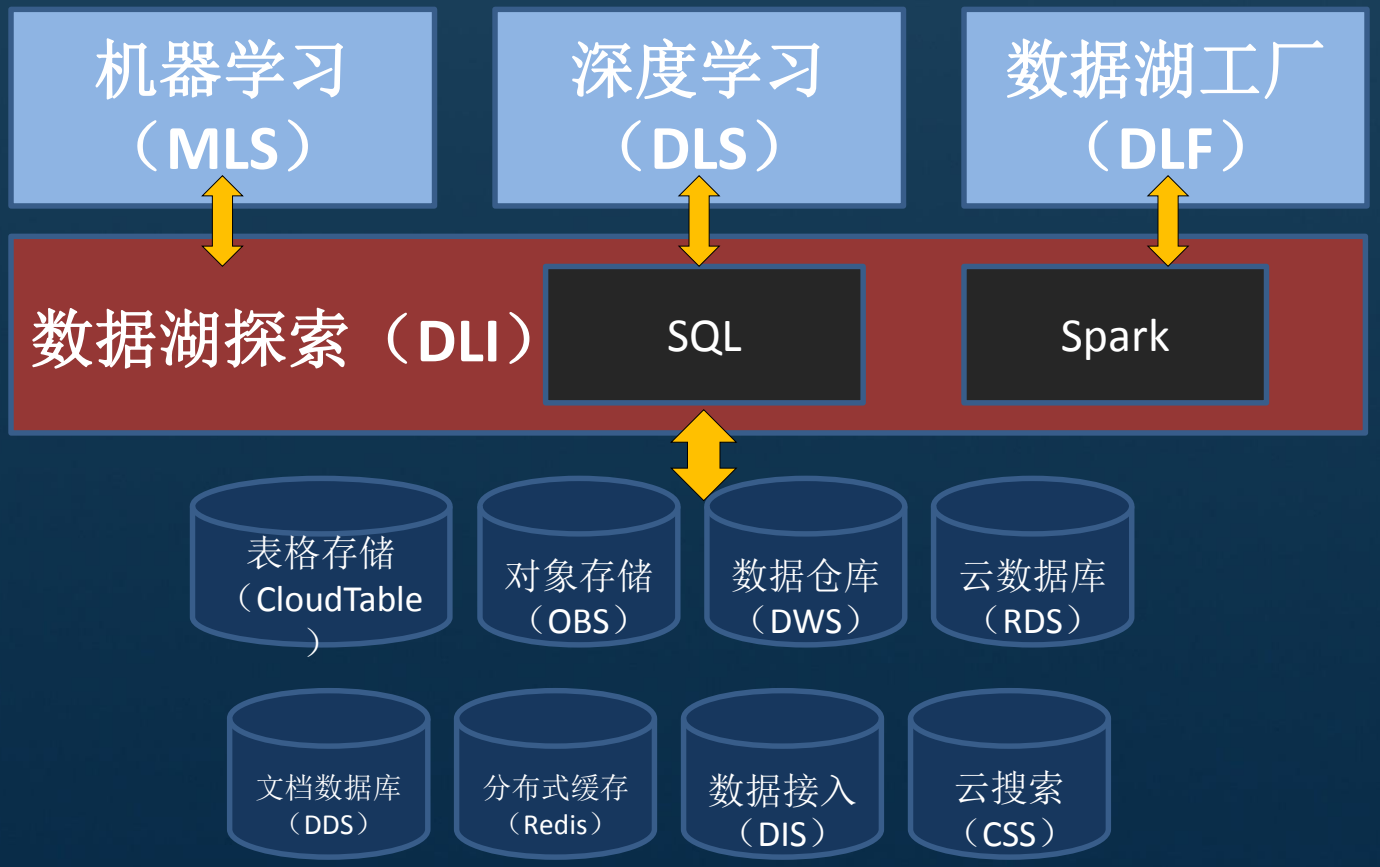
普遍的企业多源数据场景

各种场景下的数据，往往需要使用各种不同的大数据产品进行数据的存储与处理。而不同来源的数据，需要联合分析才能发挥其最大价值。

HUAWEI

数据湖探索（DLI）服务多源数据分析架构

DLI通过Spark的DataSource能力，连接不同数据源，实现不同服务之间的跨源分析计算。同时，DLI提供基于SQL的计算集群与定制化的Spark serverless计算集群。用户可以编写简单的SQL语句，也可以自定义复杂的Spark程序来实现跨源分析。



HUAWEI

HUAWEI

案例分享
——企业智能化数据经营分析实战

案例介绍

背景

企业H需要对其产品、经销商与销售订单进行信息管理。其中，各个数据信息介绍如下：

产品信息包含如下内容：

- **产品编号**
- 产品名称
- 单价

产品信息表（PRODUCT_INFO）的Schema信息

标签	产品编号	产品名称	单价
字段名	ID	NAME	PRICE

经销商信息包含如下内容：

- **经销商编号**
- **经销商名称**

经销商信息表（DEALER_INFO）的Schema信息

标签	经销商编号	经销商名称
字段名	ID	NAME

HUAWEI

案例介绍

销售订单信息包含如下内容：

- **经销商编号**
- **订单编号**
- **订单日期**
- **订购商品名称**
- **订购商品数量**

销售订单信息表（BILL）的Schema信息

经销商编号	订单编号	日期		商品信息		
RowKey前4位	RowKey后4位	time		arts		
		month	day	商品1	商品2	...

考虑到每份订单订购的商品是不同的，因此企业H采用将订单信息以No-SQL方式（HBase）进行存储。其中：

- **RowKey由4位经销商编号+4位订单编号组成**
- **商品信息存放至列簇arts中，以每种商品的名称作为列名，值为商品数量**
- **因业务需要（后续场景将介绍），考虑将订单日期存放至列簇time中，并以“month”表示月份，“day”表示日期**

HUAWEI

样例数据

产品信息数据：

ID	NAME	PRICE
0001	nova	1000
0002	P20	3000
0003	Mate20	5000

经销商信息数据：

ID	NAME
0003	B1
0005	X7

HUAWEI

样例数据

订单信息数据：

```
hbase(main):001:0> scan 'BILL'
ROW                                   COLUMN+CELL
 00030001                             column=arts:Mate20, timestamp=1541678508529, value=\x00\x00\x00\x05
 00030001                             column=arts:nova, timestamp=1541679193771, value=\x00\x00\x00\x0A
 00030001                             column=time:day, timestamp=1541679193771, value=\x00\x00\x00\x01
 00030001                             column=time:month, timestamp=1541679193771, value=Oct
 00030002                             column=arts:Mate20, timestamp=1541679171382, value=\x00\x00\x00\x05
 00030002                             column=arts:P20, timestamp=1541679208503, value=\x00\x00\x00\x05
 00030002                             column=time:day, timestamp=1541679208503, value=\x00\x00\x00\x01
 00030002                             column=time:month, timestamp=1541679208503, value=Nov
 00030003                             column=arts:Mate20, timestamp=1541679175926, value=\x00\x00\x00\x01
 00030003                             column=arts:P20, timestamp=1541679213593, value=\x00\x00\x00\x01
 00030003                             column=arts:nova, timestamp=1541679196942, value=\x00\x00\x00\x02
 00030003                             column=time:day, timestamp=1541679213593, value=\x00\x00\x00\x02
 00030003                             column=time:month, timestamp=1541679213593, value=Nov
 00050001                             column=arts:Mate20, timestamp=1541679182273, value=\x00\x00\x00\x05
 00050001                             column=arts:P20, timestamp=1541679217519, value=\x00\x00\x00\x05
 00050001                             column=arts:nova, timestamp=1541679202374, value=\x00\x00\x00\x14
 00050001                             column=time:day, timestamp=1541679217519, value=\x00\x00\x00\x01
 00050001                             column=time:month, timestamp=1541679217519, value=Oct
 00050002                             column=arts:Mate20, timestamp=1541679186275, value=\x00\x00\x00\x05
 00050002                             column=time:day, timestamp=1541679186275, value=\x00\x00\x00\x14
 00050002                             column=time:month, timestamp=1541679186275, value=Nov
5 row(s) in 0.1910 seconds
```

	time		arts		
rowkey	month	day	nova	P20	Mate20
00030001	Oct	1	10		5
00030002	Nov	1		5	5
00030003	Nov	2	2	1	1
00050001	Oct	1	20	5	5
00050002	Nov	20			5

HUAWEI

场景一：计算经销商B1每月Mate20销量与销售额

解决方案：每个月的销量数据存放在CT的订单信息中，DLI可以使用分组聚合的方式按月计算销量；而商品信息存放在DWS中，要使用DLI的跨源多表联合能力，计算出销售额（销量 * 单价）。

场景一：计算经销商B1每月Mate20销量与销售额

HUAWEI

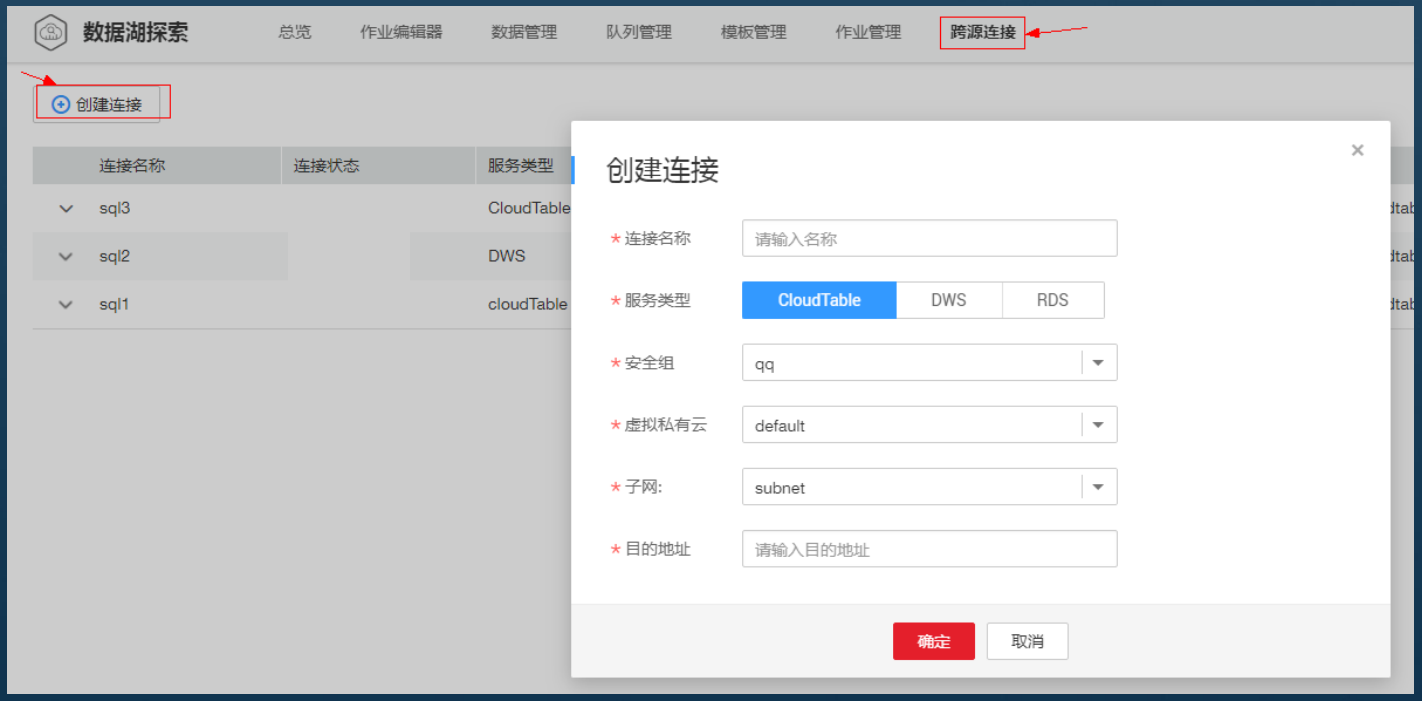
场景一：计算经销商B1每月Mate20销量与销售额

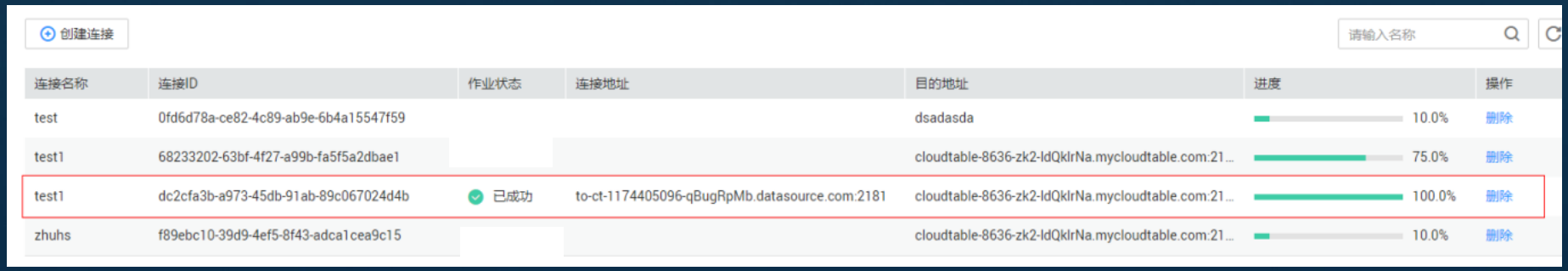
HUAWEI

场景一：计算经销商B1每月Mate20销量与销售额

HUAWEI

场景一：计算经销商B1每月Mate20销量与销售额

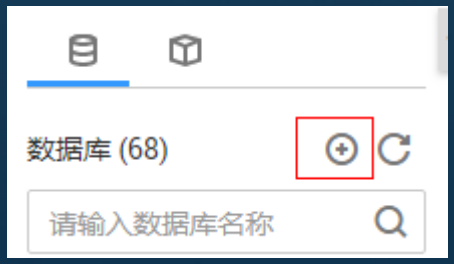


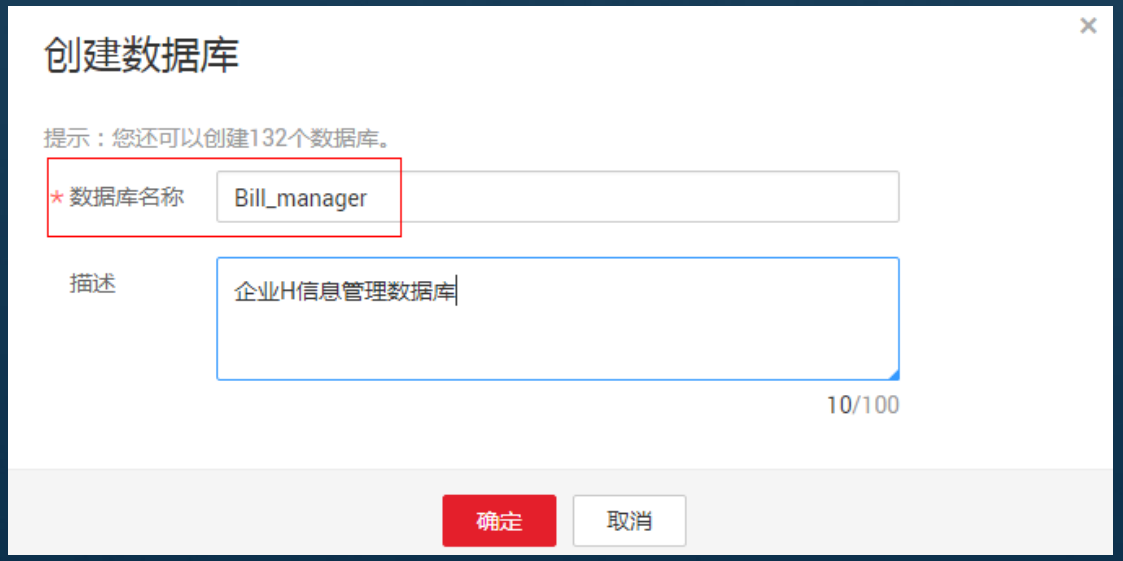


HUAWEI

场景一：计算经销商B1每月Mate20销量与销售额

创建数据库





HUAWEI

场景一：计算经销商B1每月Mate20销量与销售额

创建DLI关联CT表，查询订单详情

SQL 队列 default 数据库 bill_manager 执行 格式化 设置 更多

```
CREATE table CT_Bill(userId string, billId string, month string, day int, num_nova int, num_P20 int, num_Mate20 int) using CLOUDTABLE OPTIONS (
  'ClusterId' = '51bf529e-31ee-4c7f-b639-d494d87a0cfc',
  'TableName' = 'BILL',
  'RowKey' = 'userId:4, billId:4',
  'zkhost' = 'cloudtable-uquery-test002-zk2-nUYFmjxZ.cloudtable.com:2181',
  'Cols' = 'month:time.month, day:time.day, num_nova:arts.nova, num_P20:arts.P20, num_Mate20:arts.Mate20'
)
```

行 1，列 116

Ctrl+R或Ctrl+Enter: 执行SQL，Ctrl+F: 格式化SQL，Ctrl+Z: 回退，Tab: 自动联想

执行成功

作业类型: DDL

参数说明：

ClusterId: CT的集群ID

TableName: CT中的表名

RowKey: RowKey信息，格式为关联表中的字段名:长度

Zkhost: CT集群中的zk链接地址

Cols: 关联表与CT中实表的字段对应关系，格式为关联表字段:列簇.列名

HUAWEI

场景一：计算经销商B1每月Mate20销量与销售额

创建DLI关联CT表，查询订单详情

	time		arts		
rowkey	month	day	nova	P20	Mate20
00030001	Oct	1	10		5
00030002	Nov	1		5	5
00030003	Nov	2	2	1	1
00050001	Oct	1	20	5	5
00050002	Nov	20			5

HUAWEI

场景一：计算经销商B1每月Mate20销量与销售额

创建关于订单商品Mate20的关联表

SQL 队列 default 数据库 bill_manager ▶执行 格式化 设置 更多

```
1  CREATE table CT_Bill_mate20(userId string, billId string, month string, day int, num int) using CLOUDTABLE OPTIONS (
2    'ClusterId' = '51bf529e-31ee-4c7f-b639-d494d87a0cfc',
3    'TableName' = 'BILL',
4    'RowKey' = 'userId:4, billId:4',
5    'zkhost' = 'cloudtable-uquery-test002-zk2-nUYFmjxZ.cloudtable.com:2181,cloudtable-uquery-test002-zk1-uEqMaJ0L.cloudtable.com:2181,cloudtable-uquery-test002-zk3-12wpIJLj.cloudtable.com:2181
6    'Cols' = 'month:time.month, day:time.day, num:arts.Mate20'
7  )
```

行 1，列 11

Ctrl+R或Ctrl+Enter: 执行SQL，Ctrl+F: 格式化SQL，Ctrl+Z: 回退，Tab: 自动联想

执行成功

作业类型: DDL

HUAWEI

场景一：计算经销商B1每月Mate20销量与销售额

Tips. 新增商品的操作

```
hbase(main):011:0* create 'BILL','time','arts'
2018-11-08 19:03:28,842 INFO  [main] client.HBaseAdmin: Created BILL
0 row(s) in 1.2580 seconds

=> Hbase::Table - BILL
```

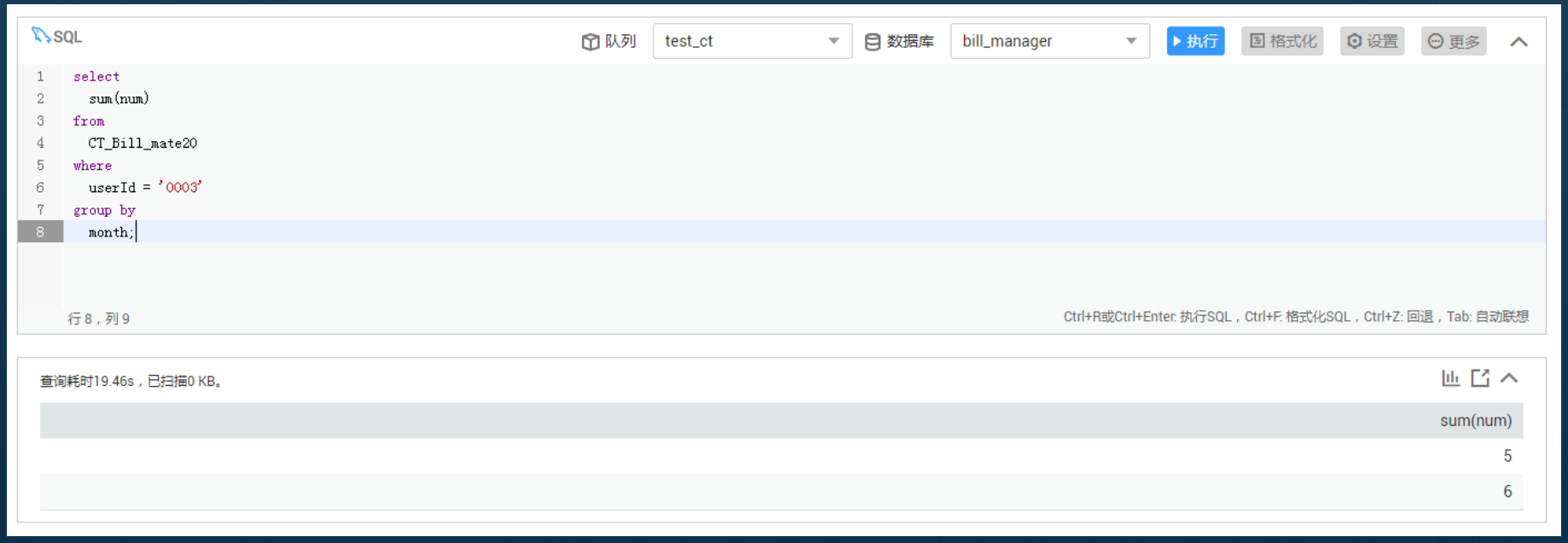

```
hbase(main):002:0> scan 'BILL'
ROW                                   COLUMN+CELL
 00030001                             column=arts:Mate20, timestamp=1541678508529, value=\x00\x00\x00\x05
 00030001                             column=time:day, timestamp=1541678508529, value=\x00\x00\x00\x01
 00030001                             column=time:month, timestamp=1541678508529, value=Oct
1 row(s) in 0.0960 seconds

hbase(main):003:0>
```


HUAWEI

场景一：计算经销商B1每月Mate20销量与销售额

查询经销商B1每月Mate20销量



HUAWEI

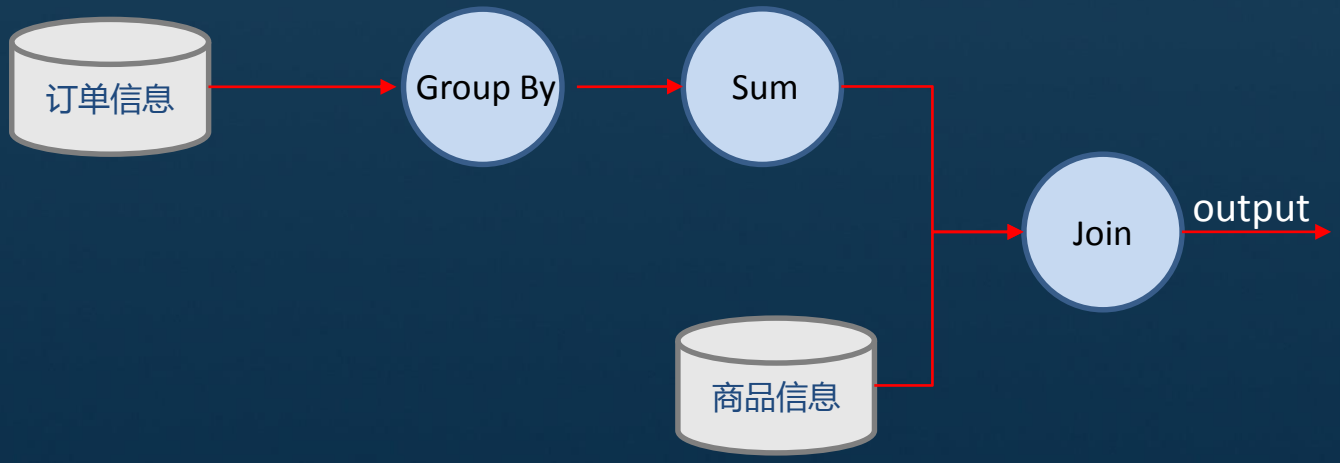
场景一：计算经销商B1每月Mate20销量与销售额

创建DWS关联表，关联商品表

HUAWEI

场景一：计算经销商B1每月Mate20销量与销售额

联合DWS商品表与CT的订单信息表进行计算



HUAWEI

场景一：计算经销商B1每月Mate20销量与销售额

联合DWS商品表与CT的订单信息表进行计算

SQL 队列 test_ct 数据库 bill_manager 执行 格式化 设置 更多

```
1  select
2    t1.month,
3    t1.count_mate20,
4    t1.count_mate20 * DWS_Arts.price as total_price
5  from (
6      select 'Mate20' name, CT_Bill_mate20.month month, sum(CT_Bill_mate20.num) count_mate20
7      from CT_Bill_mate20
8      where CT_Bill_mate20.userId = '0003'
9      group by CT_Bill_mate20.month
10   ) t1, DWS_Arts
11 where DWS_Arts.name = t1.name
```

行 11，列 30　　Ctrl+R或Ctrl+Enter: 执行SQL，Ctrl+F: 格式化SQL，Ctrl+Z: 回退，Tab: 自动联想

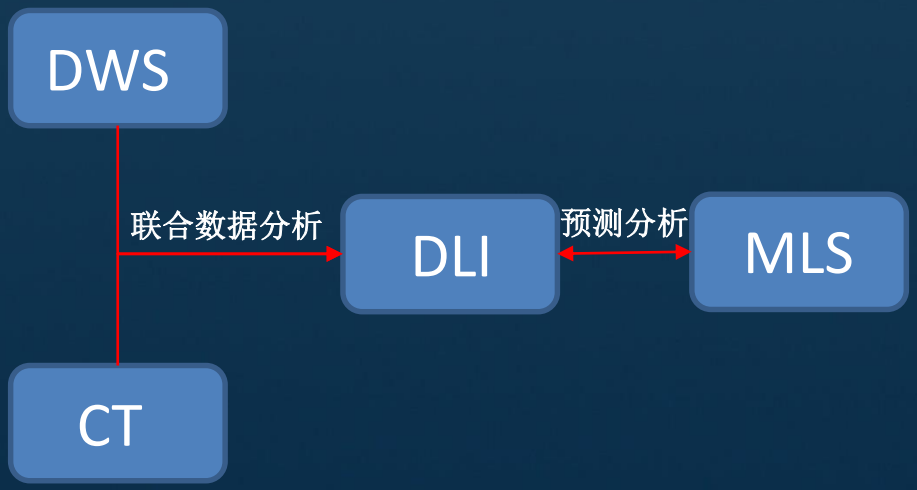
查询耗时4.88s，已扫描0 KB。

month	count_mate20	total_price
Oct	5	25000
Nov	6	30000

HUAWEI

场景二：预测经销商B1下月Mate20销量

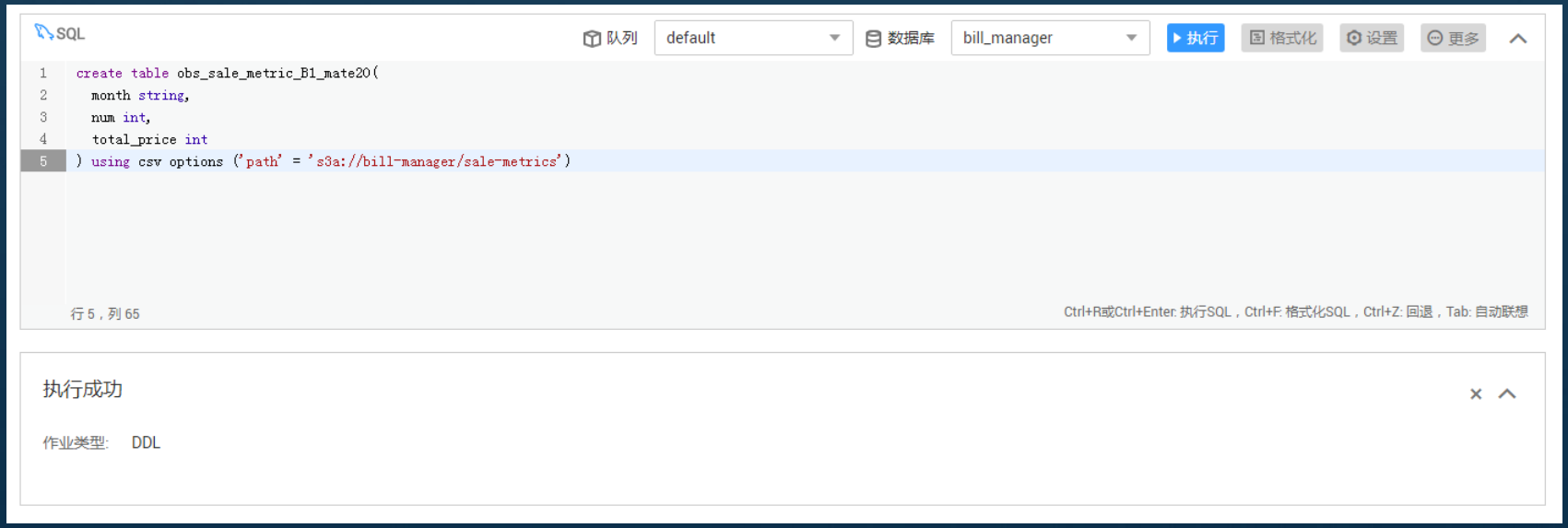
解决方案：MLS是深度集成DLI的机器学习服务，为用户提供简易的工作流操作界面来实现数据的智能分析，如销售预测等。MLS支持通过OBS读取数据并结合创建的工作流，转换成Spark作业提交到DLI后台运行。



HUAWEI

场景二：预测经销商B1下月Mate20销量

创建CSV格式的OBS外表



HUAWEI

场景二：预测经销商B1下月Mate20销量

将场景二的结果导出到OBS外表中

SQL 队列 test_ct 数据库 bill_manager ▶执行 格式化 设置 更多

```
1  insert into bill_manager.obs_sale_metric_b1_mate20 select
2    t1.month,
3    t1.count_mate20,
4    t1.count_mate20 * DWS_Arts.price as total_price
5  from (
6      select 'Mate20' name, CT_Bill_mate20.month month, sum(CT_Bill_mate20.num) count_mate20
7      from CT_Bill_mate20
8      where CT_Bill_mate20.userId = '0003'
9      group by CT_Bill_mate20.month
10   ) t1, DWS_Arts
11 where DWS_Arts.name = t1.name
```

行 11，列 30　　Ctrl+R或Ctrl+Enter: 执行SQL，Ctrl+F: 格式化SQL，Ctrl+Z: 回退，Tab: 自动联想

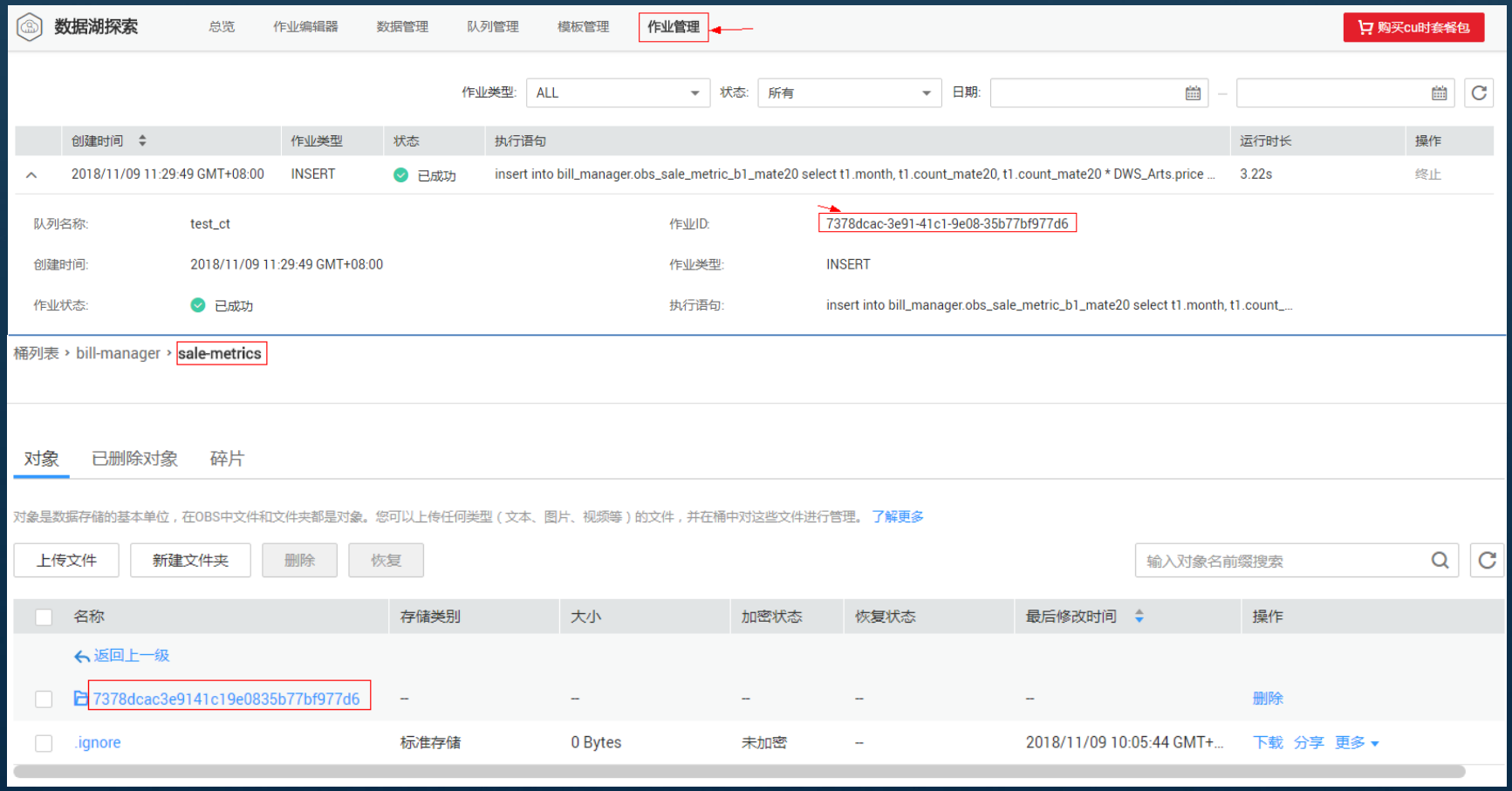
执行成功

作业类型: INSERT

HUAWEI

场景二：预测经销商B1下月Mate20销量

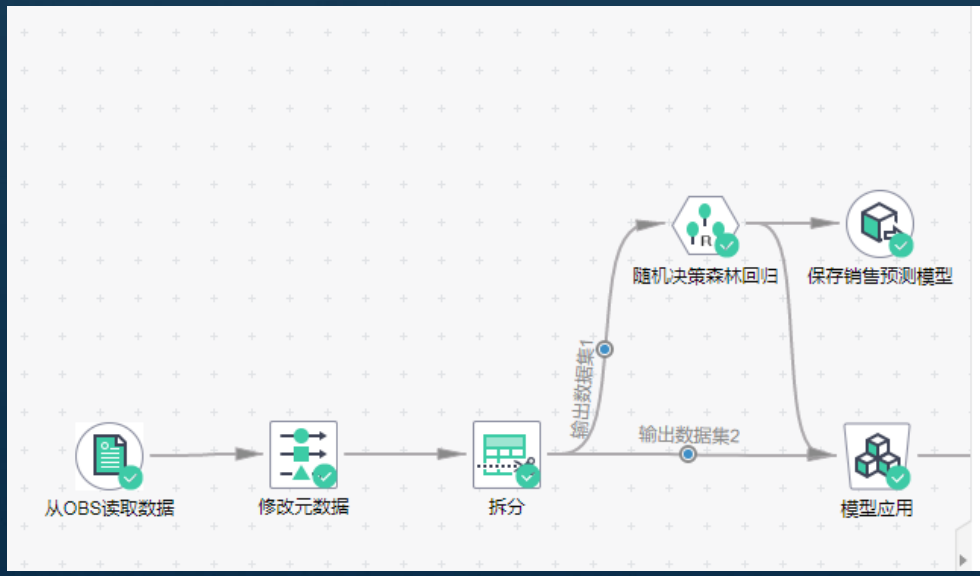
确认结果

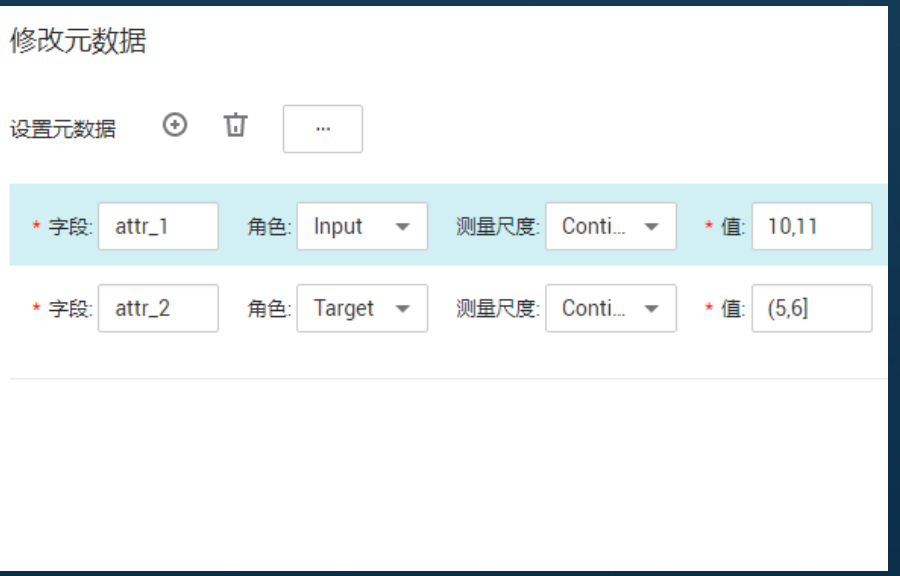


HUAWEI

场景二：预测经销商B1下月Mate20销量

在机器学习服务（MLS）中创建销售预测实例，并对导出到OBS的数据进行模型训练

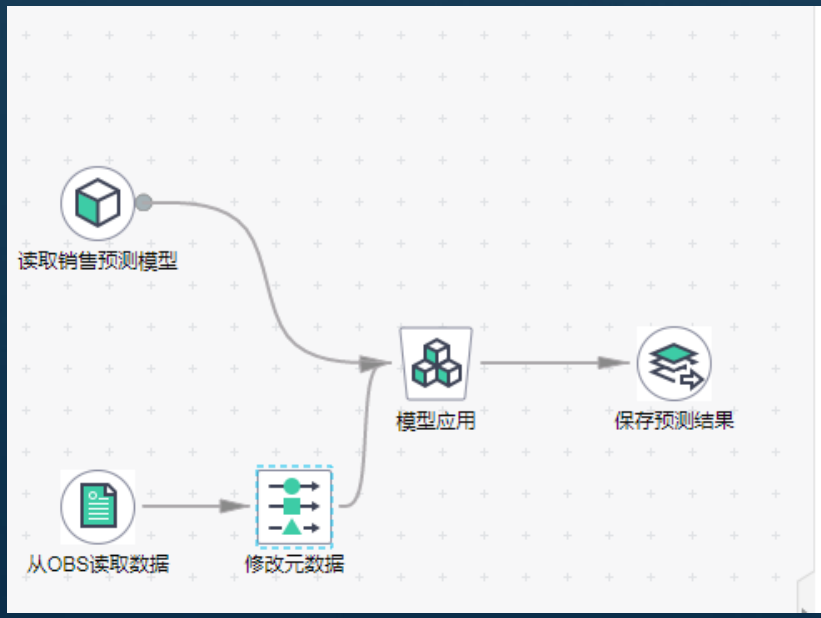


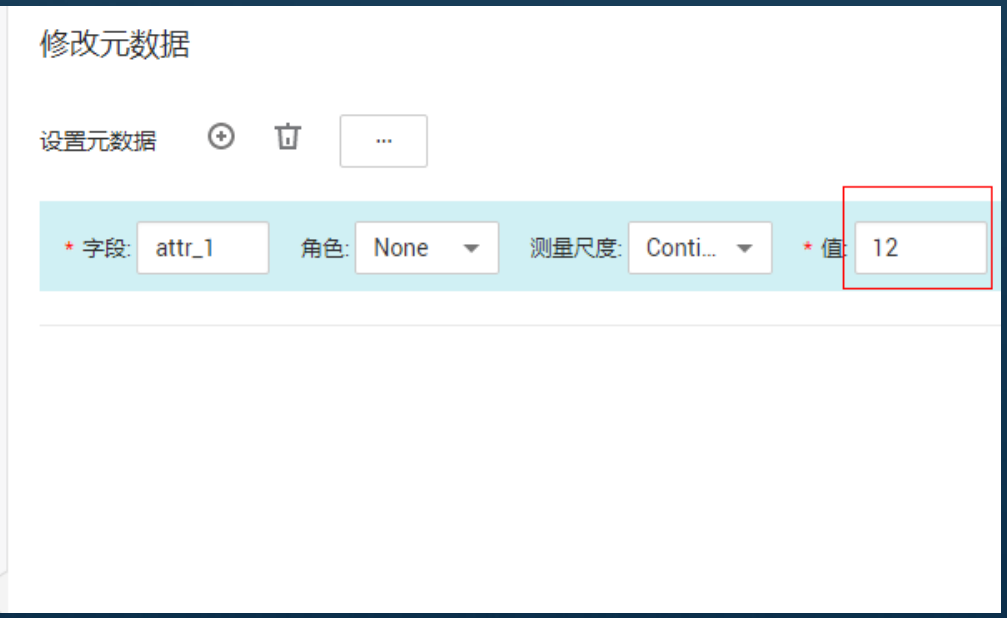


HUAWEI

场景二：预测经销商B1下月Mate20销量

使用训练好的模型进行预测

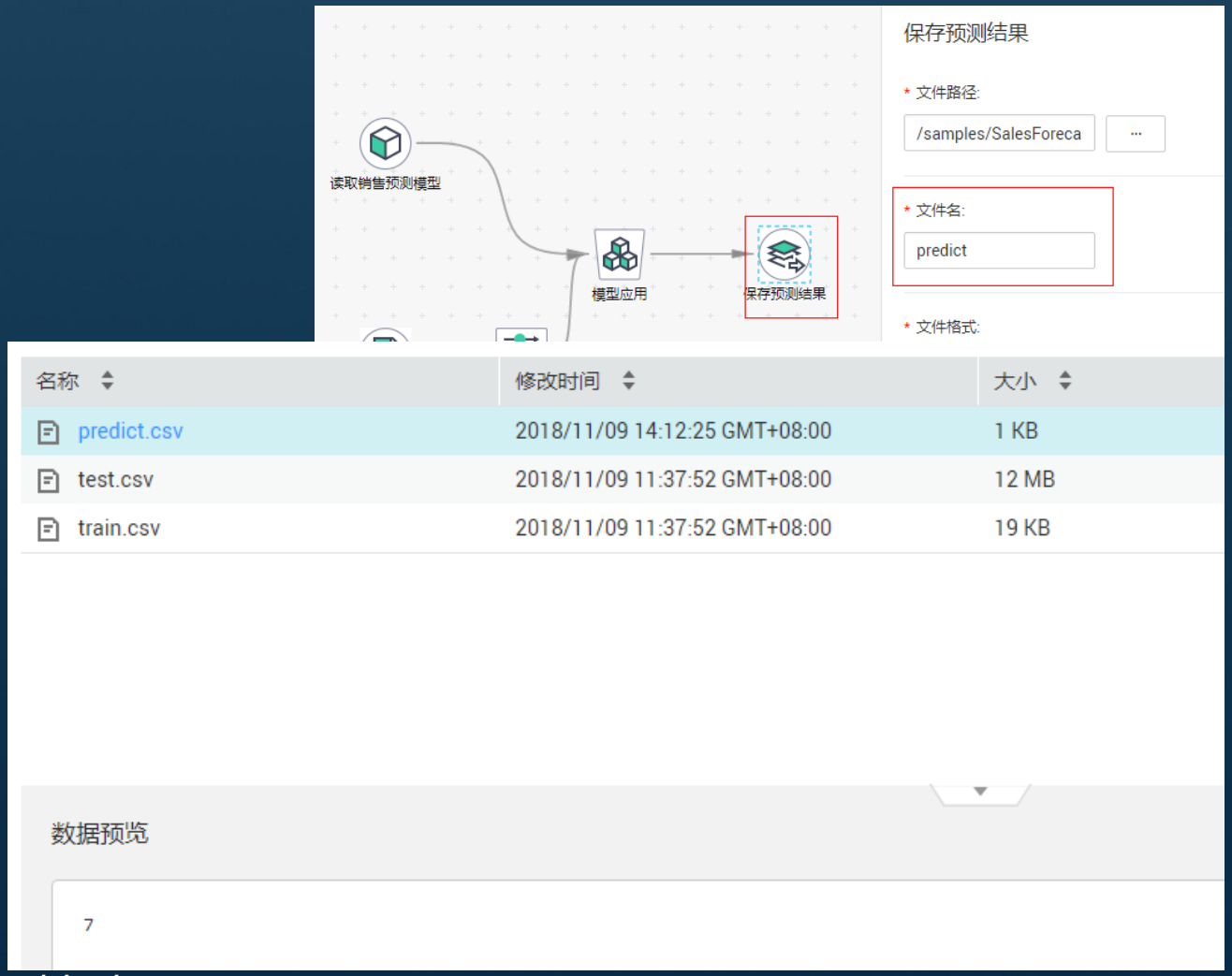




HUAWEI

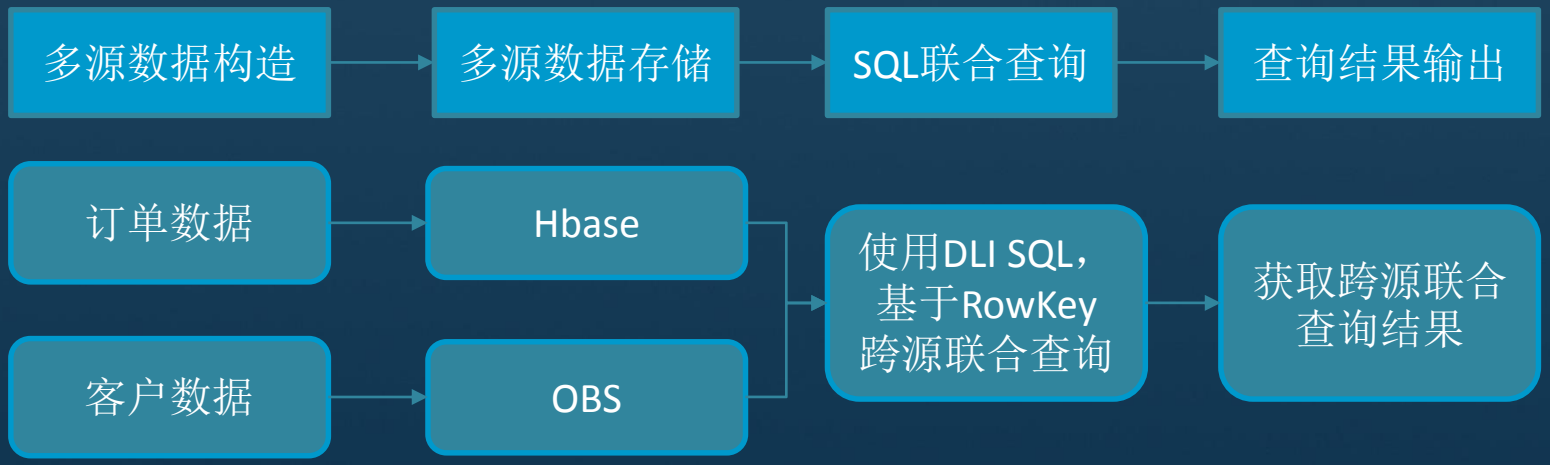
场景二：预测经销商B1下月Mate20销量

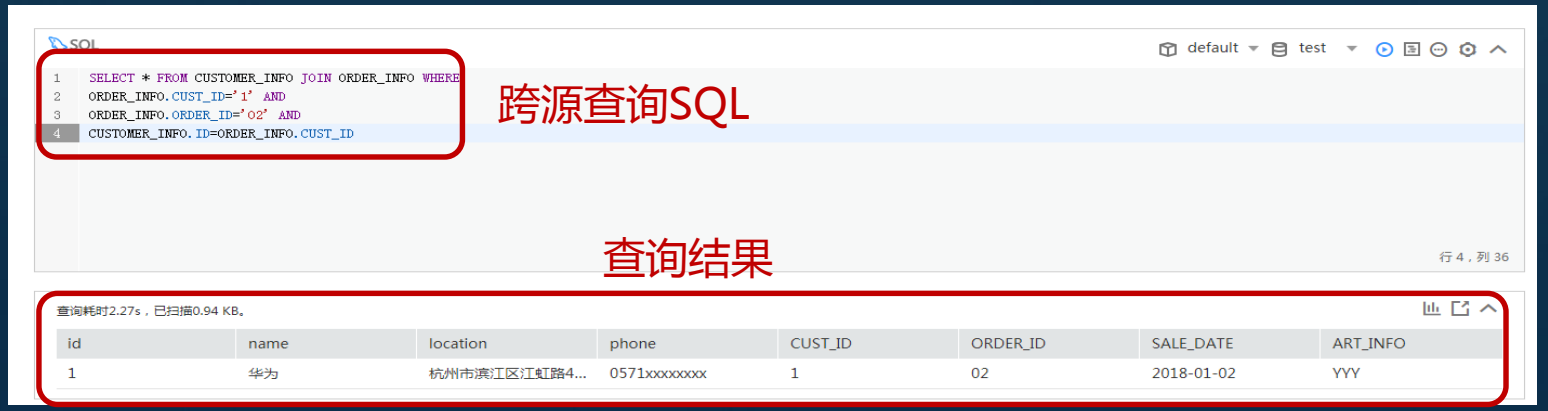
查看预测结果



HUAWEI

使用数据湖探索（DLI）进行多源数据分析实践





HUAWEI

Thank You.

Copyright©2018 Huawei Technologies Co., Ltd. All Rights Reserved.
The information in this document may contain predictive statements including, without limitation, statements regarding the future financial and operating results, future product portfolio, new technology, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. Huawei may change the information at any time without notice.