
实验 4 MyJoin_Hive

1. 实验要求

实验任务

使用 MapReduce 完成两张表的 join 操作。

实验数据在 `hdfs://master001:9000/data/hive_myjoin` 目录下。单机测试可以使用 FTP“实验要求”目录下的测试数据集。

1. 输入数据为 `order.txt` 和 `product.txt`。
2. 先将这两个文件通过使用 MapReduce 进行 join 操作，将结果输出到 HDFS 的个人目录上。
3. 进入 SQL On Hadoop 页面，使用 Hive 建表管理上一步输出的结果，输入建表语句：
`create table orders(id int,order_date string,pid string,name string,price int,num int) row format delimited fields terminated by ' ' location '/user/hc01/output/';`
其中 **row format delimited fields terminated by ' '**表示数据集分隔符为空格；
其中 **location '/user/hc01/output/'**表示所管理的数据集在 HDFS 的所处路径；
4. 最后在 Hive 上通过 `show tables` 能查看到名为 `orders` 的表，并且通过 `select` 语句能查出内容。

`product.txt` 文件从左往右分别为“商品 ID”、“商品名称”，“商品单价”。

`order.txt` 文件分别为“订单 ID”、“订单日期”、“商品 ID”、“购买数量”。

合并两个文件数据，并存入到 Hive 中。

要求程序跑完后可通过 Hive 查看生成的表。

输出格式

```
hive> create table orders(id int,order_date string,pid string,name string,price int,num int) row format delimited fields terminated by ' ' location '/user/hc01/output/';
OK
Time taken: 0.12 seconds
hive> select * from orders limit 10;
OK
1522  20190801      1    chuizi  3999   73
1268  20190731      1    chuizi  3999    6
1520  20190801      1    chuizi  3999    6
1519  20190801      1    chuizi  3999   34
1273  20190731      1    chuizi  3999   29
1275  20190731      1    chuizi  3999   64
1279  20190731      1    chuizi  3999   98
1510  20190801      1    chuizi  3999   21
1068  20190731      1    chuizi  3999   62
1069  20190731      1    chuizi  3999   39
Time taken: 2.202 seconds, Fetched: 10 row(s)
hive> [hc01@master01 bin]$ ^C
```

2. 实验数据

文本文件均使用 UTF-8 字符编码，数据之间使用空格分隔。

输入数据的情况如下图所示：



```
product.txt
1 chunzi 3999
2 huawei 3999
3 xiaomi 2999
4 apple 5999

order.txt
1 1001 20190731 1 68
2 1002 20190731 1 32
3 1003 20190731 2 63
4 1004 20190731 1 76
5 1005 20190731 4 5
6 1006 20190731 3 26
7 1007 20190731 1 49
8 1008 20190731 2 42
9 1009 20190731 2 63
10 1010 20190731 2 23
11 1011 20190731 1 31
12 1012 20190731 4 24
13 1013 20190731 3 66
14 1014 20190731 2 29
15 1015 20190731 3 39
16 1016 20190731 2 3
17 1017 20190731 4 65
18 1018 20190731 1 10
19 1019 20190731 2 23
20 1020 20190731 1 75
21 1021 20190731 4 38
22 1022 20190731 1 45
23 1023 20190731 2 73
24 1024 20190731 2 4
25 1025 20190731 1 4
26 1026 20190731 1 31
27 1027 20190731 3 80
28 1028 20190731 2 92
29 1029 20190731 2 44
30 1030 20190731 2 26
31 1031 20190731 4 28
32 1032 20190731 1 14
33 1033 20190731 1 39
34 1034 20190731 3 82
35 1035 20190731 1 81
36 1036 20190731 1 85
37 1037 20190731 3 82
38 1038 20190731 3 51
39 1039 20190731 2 29
40 1040 20190731 2 86
41 1041 20190731 4 76
42 1042 20190731 2 30
43 1043 20190731 1 51
44 1044 20190731 2 99
45 1045 20190731 3 35
46 1046 20190731 1 61
47 1047 20190731 2 6
```

数据集位于集群的 HDFS 存储上，HDFS 存储位置为：hdfs://master001:9000/data/hive_myjoin

注意 最终每个小组的程序必须在课程指定集群上运行，而且输入数据集是全部数据集。结果输出到集群的 HDFS 上。

3. 实验报告要求

在最后提交的压缩包中，除了包含源代码、JAR 包、JAR 包执行方式说明，还需要包含一个实验报告。实验报告中请包含：

1. Map 和 Reduce 的设计思路（含 Key、Value 类型）。

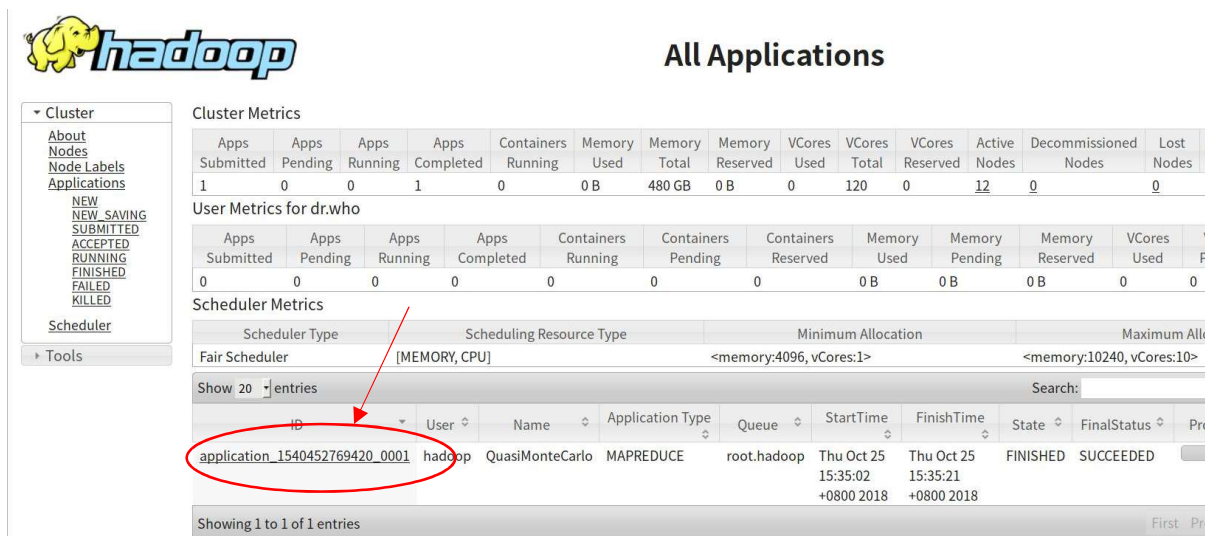
2. MapReduce 中 Map 和 Reduce 的伪代码（或者带注释的实际代码，如果使用实际代码，请做好排版）。
3. hive 输出结果文件的部分截图。
4. 请在报告中包含在集群上执行作业后，Yarn Resource Manager 的 WebUI 执行报告内容。请完整包括执行报告内容，否则影响分数。每个 MapReduce Job 对应一个报告）。执行报告内容示例见下文。

4. WebUI 执行报告

在以后的实验报告中，如果需要在集群上执行 MapReduce Job，请在实验报告中附上相关的 MapReduce Job 的执行报告，以作为评分依据。如果没有执行报告，在评分时将会认为该 MapReduce Job 没有在集群上执行，会影响实验得分。

校园网访问实验平台 **114.212.190.95:8082**

输入小组账户和密码，点击左侧栏“大数据并行计算平台”，再点击“MapReduce 并行计算”可以进入集群监控页面（见下图）。



The screenshot shows the Hadoop Yarn Resource Manager WebUI. The left sidebar contains navigation links: Cluster, About, Nodes, Node Labels, Applications, NEW, NEW SAVING, SUBMITTED, ACCEPTED, RUNNING, FINISHED, FAILED, KILLED, Scheduler, and Tools. The main content area is titled 'All Applications' and displays various metrics and a table of applications.

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	VCores Used	VCores Total	VCores Reserved	Active Nodes	Decommissioned Nodes	Lost Nodes
1	0	0	1	0	0 B	480 GB	0 B	0	120	0	12	0	0

User Metrics for dr.who

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Containers Pending	Containers Reserved	Memory Used	Memory Pending	Memory Reserved	VCores Used
0	0	0	0	0	0	0	0 B	0 B	0 B	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation
Fair Scheduler	[MEMORY, CPU]	<memory:4096, vCores:1>	<memory:10240, vCores:10>

Applications Table

ID	User	Name	Application Type	Queue	StartTime	FinishTime	State	FinalStatus	Progress
application_1540452769420_0001	hadoop	QuasiMonteCarlo	MAPREDUCE	root.hadoop	Thu Oct 25 15:35:02 +0800 2018	Thu Oct 25 15:35:21 +0800 2018	FINISHED	SUCCEEDED	

Showing 1 to 1 of 1 entries

图 1. 集群监控页面

在该页面上，每个 MapReduce Job 都有一项记录，在记录最右侧“Tracking UI”一栏可以访问到该 Job 的执行情况（见上图画圈的位置）。在执行情况页面（见下图）记录的有 Job 的执行时间、执行状态（是否 SUCCEEDED）等信息。

请在实验报告中附上 MapReduce Job 的执行情况页面截屏，以表明该 Job 是在集群上实际执行过的。



图 2. Job 执行情况页面