Alluxio 性能测试

存算分离场景

物理机配置

本次测试8台物理机配置相同：

cpu	物理核数	processor	mem	HDD
Intel(R) Xeon(R) Gold 5120 CPU @ 2.20GHz	2	56	256G	1.1T*7

集群规划

> 一个管理节点，四个计算节点，三个存储节点。存储节点和计算节点分离部署，计算节点和`Alluxio`集群并置部署。

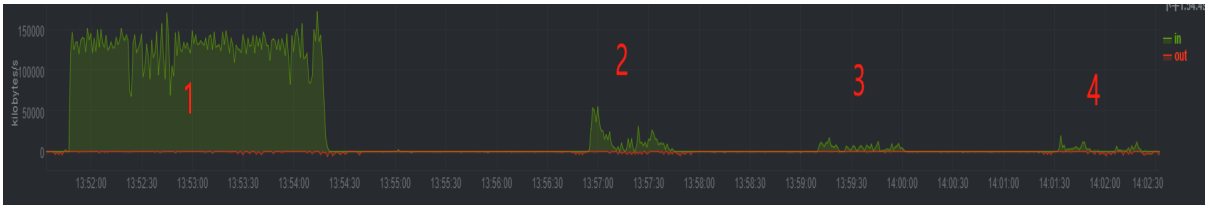
IP	hostname	role	service
192.168.1.3	manager.bigdata	管理节点	Namenode,resourcemanager,spark client,alluxio master
192.168.1.4	calculate.1.bigdata	计算节点	nodemanager,alluxio worker
192.168.1.5	calculate.2.bigdata	计算节点	nodemanager,alluxio worker
192.168.1.6	calculate.3.bigdata	计算节点	nodemanager,alluxio worker
192.168.1.7	calculate.4.bigdata	计算节点	nodemanager,alluxio worker
192.168.1.8	storage.1.bigdata	存储节点	datanode
192.168.1.9	storage.2.bigdata	存储节点	datanode
192.168.1.10	storage.3.bigdata	存储节点	datanode

测试结果

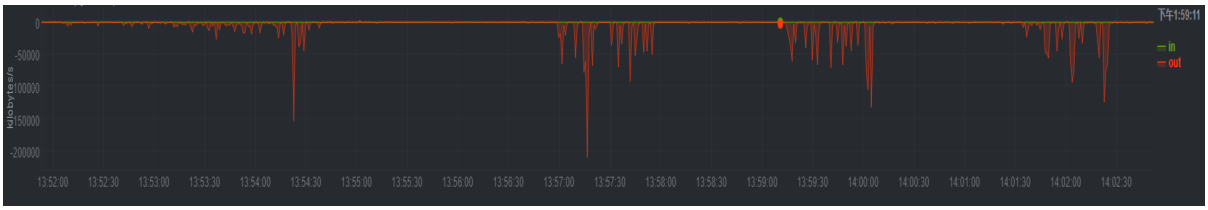
- tpc-ds单条sql：query4

*	第一次(s)	第二次(s)	第三次(s)	第四次(s)
spark without alluxio	189.908	77.69	77.341	77.072
spark with alluxio	249.513	85.426	82.482	78.778

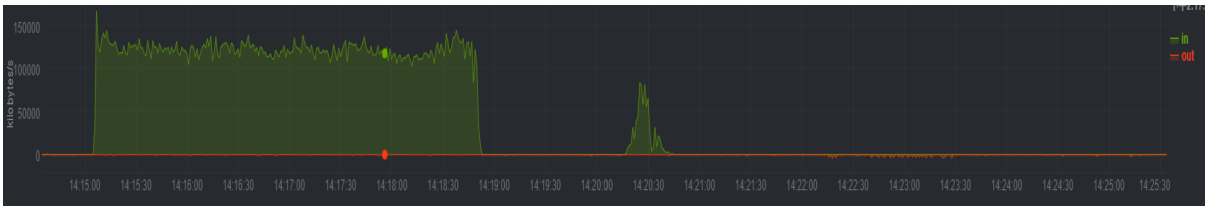
    - 磁盘IO



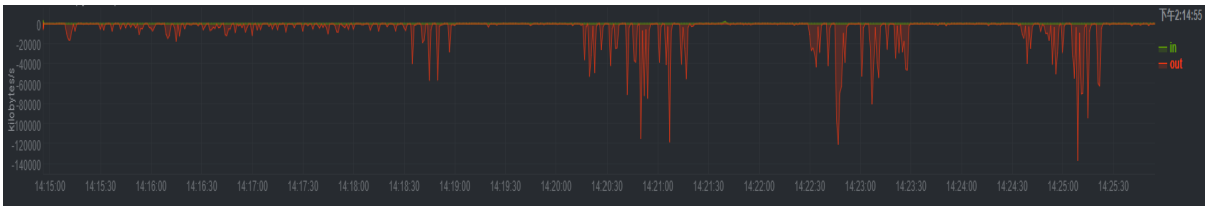
图一 spark without alluxio-存储节点



图二 spark without alluxio-计算节点



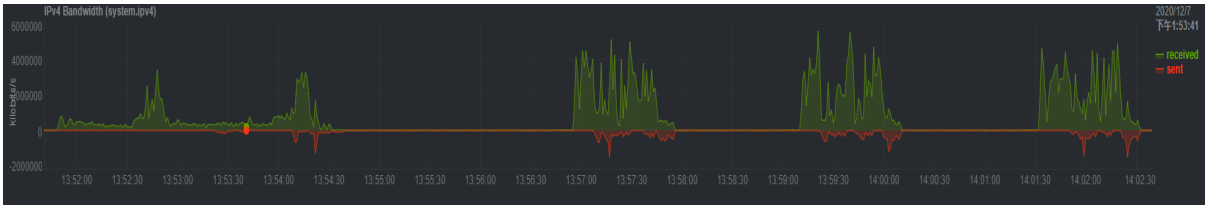
图三 spark with alluxio-存储节点



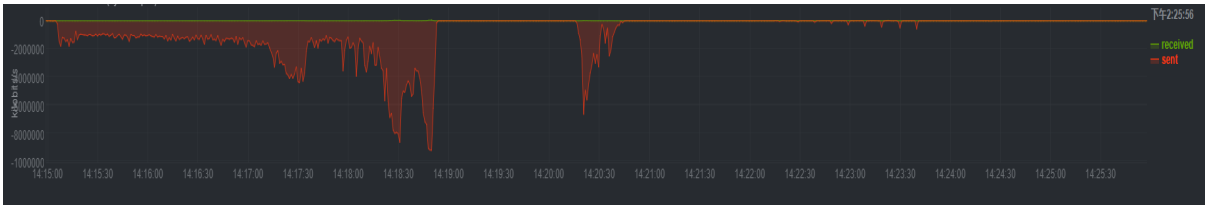
图四 spark with alluxio-计算节点

- 网络IO

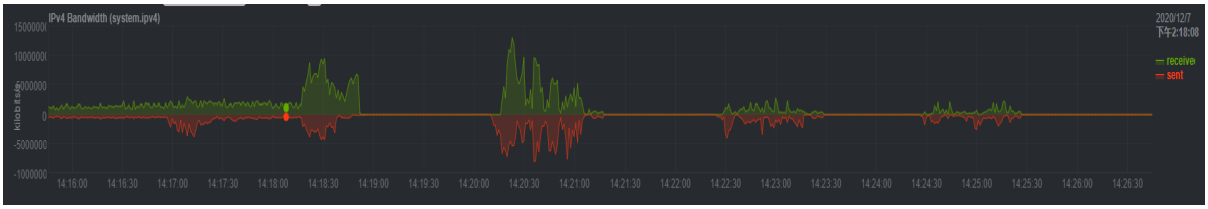
spark without alluxio-存储节点



spark without alluxio-计算节点



spark with alluxio-存储节点



spark with alluxio-计算节点

1. spark without alluxio时：第一次执行查询任务时，数据需要从存储节点的磁盘上读取数据，因此执行时间较长，后面几次由于缓存的作用，数据不再直接从磁盘读取（如图1所示），因此执行时间变短。多次执行之后，查询时间稳定在77s左右。

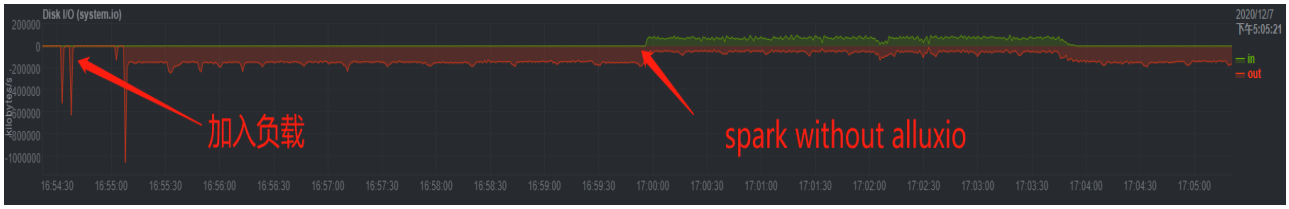
2. spark with alluxio时：第一次执行查询任务时，数据需要通过Alluxio Worker从存储节点的磁盘上读取数据，因此时间也比较长，而且比spark without alluxio查询时间还要长。后面由于数据被存储到Alluxio Worker节点上，因此数据直接从Alluxio Worker上获取数据，因此存储节点的磁盘IO使用率就减少了（如图三所示）。多次执行之后，查询时间稳定在78s左右。

3. 通过对比数据可以发现在集群空载运行时，没有系统缓存时spark with alluxio性能提升141%。有系统缓存后，spark with alluxio对spark查询性能没有提升，甚至有所下降。原因分析：

- Alluxio监控显示没有发生shorr-circuit read，因此数据全部通过网络从远程Alluxio Worker节点传输。由于缓存的作用，spark without alluxio也是通过网络从存储节点的内存中读取数据。因此性能相仿。

HDFS集群有负载时query4查询时间对比：

*	第1次(s)	第2次(s)
spark without alluxio	233.348	98.603
spark with alluxio	95.781	95.001



集群加上负载时磁盘IO

- tppc-ds前10条sql

*	第一次(s)	第二次(s)
spark without alluxio	267.876	166.265
spark with alluxio	172.532	172.803

结论

- 没有系统缓存时，spark with alluxio性能提升明显。
- 系统缓存起作用时，spark with alluxio没有性能提升，主要原因在于没有short-circuit read发生。如果应用程序需要读取的数据已经被缓存在本地Alluxio Worker上。短路读可以避免通过TCP socket传输数据，并能够提供内存级别的数据访问速度。短路读是从Alluxio读取数据的最高性能方式。

问题排查

通过分析spark ui页面提供的信息以及yarn日志，发现部分数据是从本地加载的。如下图所示：

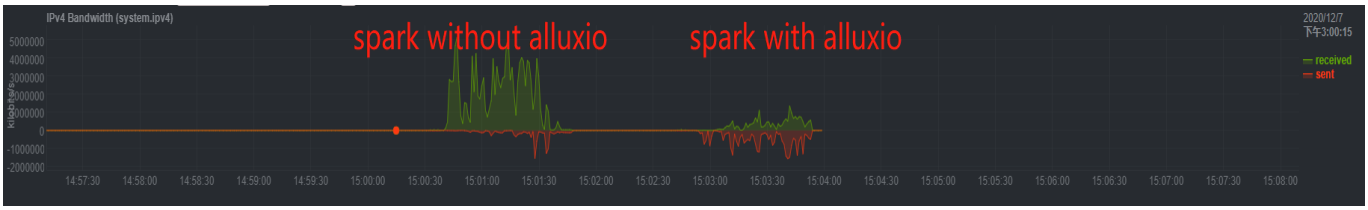
Index	ID	Attempt	Status	Locality Level	Executor ID	Host	Launch Time	Duration	GC Time	Input Size / Records	Write Time	Shuffle Write Size / Records	Errors
0	5626	0	SUCCESS	NODE_LOCAL	7	worker1.bigdata stdout stderr	2020/12/07 14:51:04	8 s	1 s	20.7 MB / 1226343	4 ms	650.6 KB / 33757	
0	6026	1 (speculative)	KILLED	RACK_LOCAL	2	master.bigdata stdout stderr	2020/12/07 14:51:12	0.3 s		0.0 B / 0		0.0 B / 0	another attempt succeeded
1	5628	0	SUCCESS	NODE_LOCAL	13	worker3.bigdata stdout stderr	2020/12/07 14:51:04	8 s	1 s	21.1 MB / 1256411	4 ms	816.5 KB / 44342	
1	6017	1 (speculative)	KILLED	NODE_LOCAL	19	worker1.bigdata stdout stderr	2020/12/07 14:51:12	0.8 s		0.0 B / 0		0.0 B / 0	another attempt succeeded
2	5630	0	SUCCESS	NODE_LOCAL	11	worker1.bigdata stdout stderr	2020/12/07 14:51:04	7 s	1 s	21.0 MB / 1253670	4 ms	583.5 KB / 29594	
3	5632	0	SUCCESS	NODE_LOCAL	23	worker1.bigdata stdout stderr	2020/12/07 14:51:04	10 s	2 s	21.0 MB / 1252462	4 ms	1039.2 KB / 58788	
3	6031	1 (speculative)	KILLED	RACK_LOCAL	4	worker2.bigdata stdout stderr	2020/12/07 14:51:12	3 s		0.0 B / 0		0.0 B / 0	another attempt succeeded
4	5639	0	SUCCESS	NODE_LOCAL	3	worker1.bigdata stdout stderr	2020/12/07 14:51:04	5 s	0.6 s	21.0 MB / 1251220	7 ms	333.3 KB / 14874	
5	5643	0	SUCCESS	NODE_LOCAL	27	worker1.bigdata stdout stderr	2020/12/07 14:51:04	6 s	0.9 s	20.9 MB / 1248177	3 ms	330.9 KB / 14773	
6	5647	0	SUCCESS	NODE_LOCAL	19	worker1.bigdata stdout stderr	2020/12/07 14:51:04	5 s	0.6 s	20.9 MB / 1246487	3 ms	580.6 KB / 29420	
7	5648	0	SUCCESS	NODE_LOCAL	15	worker1.bigdata stdout stderr	2020/12/07 14:51:04	3 s	0.4 s	20.9 MB / 1248205		0.0 B / 0	
8	5624	0	SUCCESS	NODE_LOCAL	4	worker2.bigdata stdout stderr	2020/12/07 14:51:04	8 s	1 s	20.9 MB / 1244284	2 ms	810.8 KB / 43937	
8	6021	1 (speculative)	KILLED	NODE_LOCAL	2	master.bigdata stdout stderr	2020/12/07 14:51:12	0.1 s		0.0 B / 0		0.0 B / 0	another attempt succeeded
9	5653	0	SUCCESS	NODE_LOCAL	7	worker1.bigdata stdout stderr	2020/12/07 14:51:04	8 s	1 s	20.8 MB / 1244204	3 ms	578.6 KB / 29295	
9	6030	1 (speculative)	KILLED	RACK_LOCAL	6	master.bigdata stdout stderr	2020/12/07 14:51:12	0.2 s		0.0 B / 0		0.0 B / 0	another attempt succeeded

spark with alluxio-sparkUI

```
20/12/07 14:31:16 INFO TaskSetManager: Starting task 181.0 in stage 14.0 (TID 8235, worker3.bigdata, executor 6, partition 181, NODE_LOCAL, 9370 bytes)
20/12/07 14:31:16 INFO TaskSetManager: Finished task 15.0 in stage 14.0 (TID 7470) in 4062 ms on worker3.bigdata (executor 6) (51/315)
20/12/07 14:31:16 INFO TaskSetManager: Starting task 182.0 in stage 14.0 (TID 8236, worker3.bigdata, executor 18, partition 182, NODE_LOCAL, 9370 bytes)
20/12/07 14:31:16 INFO TaskSetManager: Finished task 20.0 in stage 14.0 (TID 7473) in 4070 ms on worker3.bigdata (executor 18) (52/315)
20/12/07 14:31:16 INFO TaskSetManager: Starting task 178.0 in stage 14.0 (TID 8237, worker2.bigdata, executor 7, partition 178, NODE_LOCAL, 9370 bytes)
20/12/07 14:31:16 INFO TaskSetManager: Finished task 131.0 in stage 14.0 (TID 8186) in 1118 ms on worker2.bigdata (executor 7) (53/315)
20/12/07 14:31:16 INFO BlockManagerInfo: Removed broadcast_17_piece0 on worker1.bigdata:41644 in memory (size: 6.4 KB, free: 32.0 GB)
20/12/07 14:31:16 INFO TaskSetManager: Starting task 184.0 in stage 14.0 (TID 8238, worker3.bigdata, executor 14, partition 184, NODE_LOCAL, 9370 bytes)
20/12/07 14:31:16 INFO TaskSetManager: Finished task 13.0 in stage 14.0 (TID 7469) in 4121 ms on worker3.bigdata (executor 14) (54/315)
20/12/07 14:31:16 INFO TaskSetManager: Starting task 187.0 in stage 14.0 (TID 8239, worker3.bigdata, executor 22, partition 187, NODE_LOCAL, 9370 bytes)
20/12/07 14:31:16 INFO TaskSetManager: Finished task 9.0 in stage 14.0 (TID 7467) in 4184 ms on worker3.bigdata (executor 22) (55/315)
20/12/07 14:31:16 INFO TaskSetManager: Starting task 188.0 in stage 14.0 (TID 8240, worker3.bigdata, executor 18, partition 188, NODE_LOCAL, 9370 bytes)
20/12/07 14:31:16 INFO TaskSetManager: Finished task 17.0 in stage 14.0 (TID 7472) in 4177 ms on worker3.bigdata (executor 18) (56/315)
20/12/07 14:31:16 INFO TaskSetManager: Starting task 179.0 in stage 14.0 (TID 8241, worker1.bigdata, executor 21, partition 179, NODE_LOCAL, 9370 bytes)
20/12/07 14:31:16 INFO TaskSetManager: Finished task 137.0 in stage 14.0 (TID 8199) in 920 ms on worker1.bigdata (executor 21) (57/315)
20/12/07 14:31:16 INFO TaskSetManager: Starting task 186.0 in stage 14.0 (TID 8242, worker2.bigdata, executor 15, partition 186, NODE_LOCAL, 9370 bytes)
20/12/07 14:31:16 INFO TaskSetManager: Finished task 6.0 in stage 14.0 (TID 7462) in 4248 ms on worker2.bigdata (executor 15) (58/315)
20/12/07 14:31:16 INFO TaskSetManager: Starting task 190.0 in stage 14.0 (TID 8243, worker3.bigdata, executor 22, partition 190, NODE_LOCAL, 9370 bytes)
20/12/07 14:31:16 INFO TaskSetManager: Finished task 169.0 in stage 14.0 (TID 8233) in 237 ms on worker3.bigdata (executor 22) (59/315)
20/12/07 14:31:16 INFO TaskSetManager: Starting task 180.0 in stage 14.0 (TID 8244, worker1.bigdata, executor 1, partition 180, NODE_LOCAL, 9370 bytes)
20/12/07 14:31:16 INFO TaskSetManager: Finished task 142.0 in stage 14.0 (TID 8205) in 965 ms on worker1.bigdata (executor 1) (60/315)
20/12/07 14:31:16 INFO TaskSetManager: Starting task 189.0 in stage 14.0 (TID 8245, worker2.bigdata, executor 3, partition 189, NODE_LOCAL, 9370 bytes)
20/12/07 14:31:16 INFO TaskSetManager: Finished task 139.0 in stage 14.0 (TID 8213) in 953 ms on worker2.bigdata (executor 3) (61/315)
20/12/07 14:31:16 INFO TaskSetManager: Starting task 183.0 in stage 14.0 (TID 8246, worker1.bigdata, executor 1, partition 183, NODE_LOCAL, 9370 bytes)
20/12/07 14:31:16 INFO TaskSetManager: Killing attempt 1 for task 285.1 in stage 10.0 (TID 7408) on master.bigdata as the attempt 0 succeeded on worker1.bigdata
20/12/07 14:31:16 INFO TaskSetManager: Finished task 285.0 in stage 10.0 (TID 6730) in 8764 ms on worker1.bigdata (executor 1) (342/348)
20/12/07 14:31:16 INFO TaskSetManager: Starting task 161.0 in stage 14.0 (TID 8247, master.bigdata, executor 8, partition 161, NODE_LOCAL, 9179 bytes)
20/12/07 14:31:16 WARN TaskSetManager: Lost task 285.1 in stage 10.0 (TID 7408, master.bigdata, executor 8): TaskKilled (another attempt succeeded)
```

spark with alluxio-yarn log

1.通过上面两个图可以看到程序运行过程当中部分数据是从本地加载的，但是程序运行中产生的RDD也会从本地计算节点加载。所以无法判断元数据是否是从本地加载进来的。



计算节点网络接口监控

2.计算节点网络接口监控显示spark without alluxio时，数据通过网络从存储节点传输到计算节点；spark with alluxio时，也有数据通过网络传输到计算节点，但是数据量小很多，所以推测有部分数据确实是通过本地加载的。但是从得到的性能数据上看，数据因该没有命中本地worker，还是通过网络传输的数据。

4.没有发生short-circuit read，可能的原因有两个：

- Alluxio缺少相关配置
- yarn调度策略有问题。

通过hive客户端查询数据时，可以监控到有short-circuit read，但是提交到yarn的所有任务都没有short-circuit read。

目前已经定位到Alluxio中读数据相关代码，待调试后查明原因。