Spark企业级大数据项目实战 第10课

DATAGURU专业数据分析社区

法律声明

【声明】本视频和幻灯片为炼数成金网络课程的教学资料，所有资料只能在课程内使用，不得在课程以外范围散播，违者将可能被追究法律和经济责任。

课程详情访问炼数成金培训网站

http://edu.dataguru.cn

炼数成金逆向收费式网络课程

- Dataguru（炼数成金）是专业数据分析网站，提供教育，媒体，内容，社区，出版，数据分析业务等服务。我们的课程采用新兴的互联网教育形式，独创地发展了逆向收费式网络培训课程模式。既继承传统教育重学习氛围，重竞争压力的特点，同时又发挥互联网的威力打破时空限制，把天南地北志同道合的朋友组织在一起交流学习，使到原先孤立的学习个体组合成有组织的探索力量。并且把原先动辄成千上万的学习成本，直线下降至百元范围，造福大众。我们的目标是：低成本传播高价值知识，构架中国第一的网上知识流转阵地。
- 关于逆向收费式网络的详情，请看我们的培训网站 http://edu.dataguru.cn

本课内容

- 多级分区问题分析、解决方案
- Spark读取中文乱码问题
- Spark Streaming监控文件目录开发、问题分析
- Livy
- Spark jobserver

1 多级分区问题分析、解决方案

分区规则： /houseid=?/dayid=?/hourid=?/minu5=?

改造前的数据清洗流程：

1. 数据入HDFS目录，5分钟一个目录
2. Spark清洗这个5分钟的目录， 数据入HDFS

```
-bash-4.1$  hdfs dfs -ls /hadoop/srcdata/ | more
Found 23 items
drwxr-xr-x   - hdfs hdfs          0 2018-04-08 01:00 /hadoop/srcdata/201803160400
drwxr-xr-x   - hdfs hdfs          0 2018-03-26 18:43 /hadoop/srcdata/201803160405
drwxr-xr-x   - hdfs hdfs          0 2018-04-08 01:01 /hadoop/srcdata/201803160410
drwxr-xr-x   - hdfs hdfs          0 2018-04-08 01:01 /hadoop/srcdata/201803160415
drwxr-xr-x   - hdfs hdfs          0 2018-04-08 01:01 /hadoop/srcdata/201803160420
drwxr-xr-x   - hdfs hdfs          0 2018-04-08 01:01 /hadoop/srcdata/201803160425
drwxr-xr-x   - hdfs hdfs          0 2018-04-08 01:01 /hadoop/srcdata/201803160430
drwxr-xr-x   - hdfs hdfs          0 2018-04-08 01:01 /hadoop/srcdata/201803160435
drwxr-xr-x   - hdfs hdfs          0 2018-04-08 01:01 /hadoop/srcdata/201803160440
drwxr-xr-x   - hdfs hdfs          0 2018-04-08 01:01 /hadoop/srcdata/201803160445
drwxr-xr-x   - hdfs hdfs          0 2018-04-08 01:01 /hadoop/srcdata/201803160450
drwxr-xr-x   - hdfs hdfs          0 2018-04-08 01:03 /hadoop/srcdata/201803160455
drwxr-xr-x   - hdfs hdfs          0 2018-04-08 01:03 /hadoop/srcdata/201803160500
drwxr-xr-x   - hdfs hdfs          0 2018-04-08 01:03 /hadoop/srcdata/201803160505
drwxr-xr-x   - hdfs hdfs          0 2018-04-08 01:04 /hadoop/srcdata/201803160510
drwxr-xr-x   - hdfs hdfs          0 2018-04-08 01:04 /hadoop/srcdata/201803160515
drwxr-xr-x   - hdfs hdfs          0 2018-04-08 01:04 /hadoop/srcdata/201803160520
```

1 多级分区问题分析、解决方案

分区规则： /houseid=?/dayid=?/hourid=?/minu5=?

存在问题：

1. 收敛参数coalesce失效， 小文件特别多
   不同house的数据差异很大， 数据时间分布差异很大，导致清洗后产生大量的小文件。
   100G的原始文件， 清洗后产生了10万多个小文件。
2. 数据入库延迟大
   HDFS的数据有延迟几个小时， 直接导致清洗后数据分布跨多个时间分区。 这也会导致小文件。
3. 数据倾斜
   不同的house的数据差异很大， 数据量分布从几M到20G之间。
4. 清洗效率
   小文件太多， 直接导致NameNode压力负载过大， 清洗效率非常慢

```
$ hdfs dfs -ls -h /hadoop/data/output/data/houseid=10001/dayid=20180316/hourid=12/minu5=00
Items
-   3 hdfs hdfs      3.6 M 2018-03-26 11:33 /hadoop/data/output/data/houseid=10001/dayid=20180316/hourid=12/minu5=00/201803161205-43829.orc
-   3 hdfs hdfs      4.7 M 2018-03-26 11:32 /hadoop/data/output/data/houseid=10001/dayid=20180316/hourid=12/minu5=00/201803161205-43830.orc
-   3 hdfs hdfs      4.3 M 2018-03-26 11:33 /hadoop/data/output/data/houseid=10001/dayid=20180316/hourid=12/minu5=00/201803161205-43831.orc
-   3 hdfs hdfs      4.7 M 2018-03-26 11:33 /hadoop/data/output/data/houseid=10001/dayid=20180316/hourid=12/minu5=00/201803161205-43832.orc
-   3 hdfs hdfs      4.8 M 2018-03-26 11:31 /hadoop/data/output/data/houseid=10001/dayid=20180316/hourid=12/minu5=00/201803161205-43833.orc
-   3 hdfs hdfs      9.0 M 2018-03-26 11:31 /hadoop/data/output/data/houseid=10001/dayid=20180316/hourid=12/minu5=00/201803161205-43834.orc
```

1 多级分区问题分析、解决方案

```
-bash-4.1$ hdfs dfs -ls /hadoop/mydata_new/201803161205 | more
Found 59 items
drwxr-xr-x   - slview supergroup          0 2018-04-03 16:27 /hadoop/mydata_new/201803161205/1000
drwxr-xr-x   - slview supergroup          0 2018-04-03 16:27 /hadoop/mydata_new/201803161205/10007
drwxr-xr-x   - slview supergroup          0 2018-04-03 16:27 /hadoop/mydata_new/201803161205/10008
drwxr-xr-x   - slview supergroup          0 2018-04-03 16:28 /hadoop/mydata_new/201803161205/10009
drwxr-xr-x   - slview supergroup          0 2018-04-03 16:30 /hadoop/mydata_new/201803161205/1002
drwxr-xr-x   - slview supergroup          0 2018-04-03 16:30 /hadoop/mydata_new/201803161205/10024
drwxr-xr-x   - slview supergroup          0 2018-04-03 16:31 /hadoop/mydata_new/201803161205/1003
drwxr-xr-x   - slview supergroup          0 2018-04-03 16:33 /hadoop/mydata_new/201803161205/1005
drwxr-xr-x   - slview supergroup          0 2018-04-03 16:35 /hadoop/mydata_new/201803161205/1006
drwxr-xr-x   - slview supergroup          0 2018-04-03 16:37 /hadoop/mydata_new/201803161205/1007
drwxr-xr-x   - slview supergroup          0 2018-04-03 16:38 /hadoop/mydata_new/201803161205/1008
drwxr-xr-x   - slview supergroup          0 2018-04-03 16:39 /hadoop/mydata_new/201803161205/1009
drwxr-xr-x   - slview supergroup          0 2018-04-03 16:40 /hadoop/mydata_new/201803161205/1010
drwxr-xr-x   - slview supergroup          0 2018-04-03 16:41 /hadoop/mydata_new/201803161205/1012
drwxr-xr-x   - slview supergroup          0 2018-04-03 16:44 /hadoop/mydata_new/201803161205/1014
drwxr-xr-x   - slview supergroup          0 2018-04-03 16:46 /hadoop/mydata_new/201803161205/1017
drwxr-xr-x   - slview supergroup          0 2018-04-03 16:47 /hadoop/mydata_new/201803161205/1018
drwxr-xr-x   - slview supergroup          0 2018-04-03 16:47 /hadoop/mydata_new/201803161205/1019
drwxr-xr-x   - slview supergroup          0 2018-04-03 16:49 /hadoop/mydata_new/201803161205/1020
```


改造后的数据清洗流程：

1. 数据入库的效率改造
   确保数据入HDFS延时不超过1个小时。

2. 数据入HDFS目录，按照时间和house分目录：/${5分钟目录}/${houseid}

3. Spark读取数据源目录，遍历5分钟目录的每个house目录。
   每个house目录生成一个DataFrame， 根据每个house目录的大小确定收敛参数(coalesce)大小为partitionNum。
   （1）、当前时间1个小时的数据，转换为DataFrame， 分区数大小为partitionNum
   （2）、过滤1个小时前的数据，转换为DataFrame, 分区数大小为partitionNum的三分之一。

4. 使用union将每个目录生成的DataFrame合并成一个DataFrame， 将这个DataFrame写入分区

根据每个house目录的大小确定收敛参数(coalesce)大小解决了小文件问题， 同时保证收敛后每个分区的数据基本均衡。

2 Spark读取中文乱码问题

- spark的textFile读取文本文件， 但是这个方法只支持UTF8格式，
  如果GBK格式的文件含有中文， 就会乱码
- 文本格式中文乱码解决
- JSON格式中文乱码问题

3 Spark Streaming监控文件目录开发、问题分析

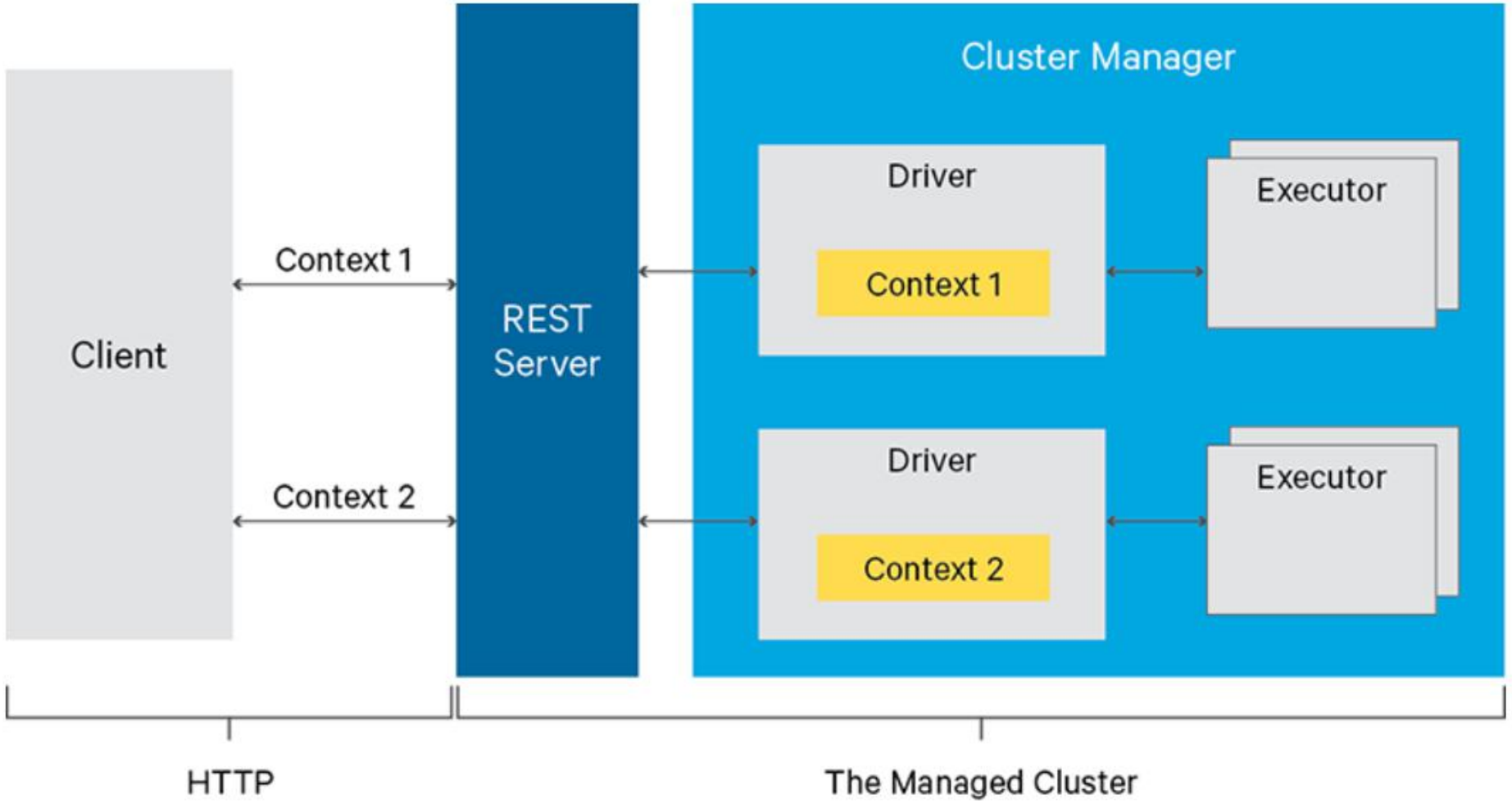
1. 无法解决小文件问题

2. Spark Streaming程序重启数据丢失

3. 大文件入监控目录， 入库时间长， Spark Streaing程序只会读取部分数据

生产上面不建议使用Spark Streaming监控文件目录

4 Livy


Client
Context 1
Context 2
REST Server
Cluster Manager
Driver
Context 1
Executor
Driver
Context 2
Executor
HTTP
The Managed Cluster

4 Livy

- 交互式Scala，Python和R shell
- 批量提交Scala，Java，Python
- 多个用户可以共享同一台Server
- 可用于使用REST从任何位置提交作业
- 不需要对的程序进行任何代码更改
- 共享Spark Contexts和RDDs

4 Livy案例实操

- Livy安装配置
- Livy的使用--Session
- Livy的使用--Batch
- 基于python的rest方式

4 Livy Session 的好处

1. 加速Spark任务的启动速度。 Session第一次创建比较耗时， 后续任务提交的代码片段都是立即执行的。

2. 共享RDD、DataFrame。

5 JobServer

- "Spark as Service"：针对 job 和 contexts 的各个方面提供了 REST 风格的 api 接口进行管理
- 支持 SparkSQL、Hive、Streaming Contexts/jobs 以及定制 job contexts
- 通过集成 Apache Shiro 来支持 LDAP 权限验证
- 通过长期运行的job contexts支持亚秒级别低延迟的任务
- 可以通过结束 context 来停止运行的作业(job)
- 分割 jar 上传步骤以提高 job 的启动
- 异步和同步的 job API，其中同步 API 对低延时作业非常有效
- 支持 Standalone Spark 和 Mesos、yarn
- Job 和 jar 信息通过一个可插拔的 DAO 接口来持久化
- 对RDD或DataFrame对象命名并缓存，通过该名称获取RDD或DataFrame。这样可以提高对象在作业间的共享和重用
- 支持 Scala 2.10 版本和 2.11 版本

4. JobServer案例实操

- JobServer安装配置
- JobServer创建上下文
- JobServer资源分配
- SparkJob API编程

Thanks

FAQ时间