


(/apps/redi
utm_sourc
banner-clic

Spark优雅的操作Redis



贪恋清晨de阳光 (/u/32f6fe29fdc4) + 关注

2017.03.29 11:33* 字数 1066 阅读 11341 评论 9 喜欢 4

(/u/32f6fe29fdc4)

Spark的优势在于内存计算，然而在计算中难免会用到一些元数据或中间数据，有的存在关系型数据库中，有的存在HDFS上，有的存在HBase中，但其读写速度都和Spark计算的速度相差甚远，而Redis基于内存的读写则可以完美解决此类问题，下面介绍Spark如何与Redis交互。

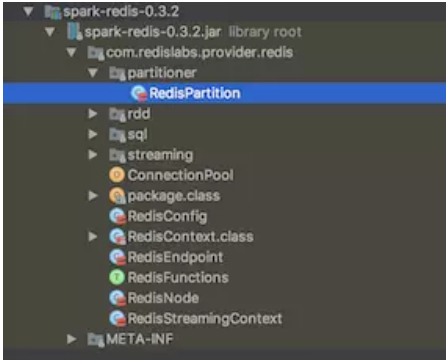
在Spark计算的时候如何加载Redis中的数据，其实官方有现成的包和文档，文档是全英文，好在东西不多，下面介绍如何使用。

首先把jar包引入工程，在maven上居然找不到这个包。。。所以使用Maven和SBT的同学自行解决。下载地址 (<https://spark-packages.org/package/RedisLabs/spark-redis>)(打不开的同学可以尝试翻墙)

2018-11-06更新：最新maven依赖已经被中央仓库收录

```
<dependency>
  <groupId>RedisLabs</groupId>
  <artifactId>spark-redis</artifactId>
  <version>0.3.2</version>
</dependency>
```

(<https://dspclick.youda.com/slot=30edcf3b6-496b-0a451c81c290606218>)



可以看出提供的功能还是挺全面的，有单独的redis分区，redisRDD，SQLAPI以及StreamingAPI

下面我们一点一点来做一个示例：

在这里先看看官方包中的一部分源码：



```
/**
 * 官方提供源码包中解析Redis配置需要的字段
 */
case class RedisEndpoint(val host: String = Protocol.DEFAULT_HOST,
                          val port: Int = Protocol.DEFAULT_PORT,
                          val auth: String = null,
                          val dbNum: Int = Protocol.DEFAULT_DATABASE,
                          val timeout: Int = Protocol.DEFAULT_TIMEOUT)
    extends Serializable {

  /**
   * 源码中获取配置的字段名及来源，可以看出是从SparkConf中读取到相应字段，所以连接redis只需要在
   */
  def this(conf: SparkConf) {
    this(
      conf.get("redis.host", Protocol.DEFAULT_HOST),
      conf.getInt("redis.port", Protocol.DEFAULT_PORT),
      conf.get("redis.auth", null),
      conf.getInt("redis.db", Protocol.DEFAULT_DATABASE),
      conf.getInt("redis.timeout", Protocol.DEFAULT_TIMEOUT)
    )
  }

  ...
}
```

(/apps/redi
utm_sourc
banner-clic

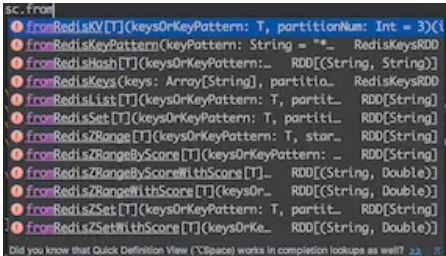
现在我们启动 SparkContext

```
先引入Redis相关的隐式转换
import com.redislabs.provider.redis._

//这里直接使用yarn-cluster模式
val conf = new SparkConf().setMaster("yarn-cluster").setAppName("sparkRedisTest")
conf.set("redis.host", "10.1.11.70") //host,随便一个节点，自动发现
conf.set("redis.port", "6379") //端口号，不填默认为6379
//conf.set("redis.auth", "null") //用户权限配置
//conf.set("redis.db", "0") //数据库设置
//conf.set("redis.timeout", "2000") //设置连接超时时间
val sc = new SparkContext(conf)
```

(https://dsp
click.youde
slot=30edc
f3b6-496b-
0a451c81c
290606218

之后可以看到IDEA给出的提示，sc通过导入的隐式转换可以调出的读取Redis的方法，都是以 fromRedis 开头的，都是redis可以存储的数据结构，这里以常见的KV进行示例



还是先扒一下源码看看：



```
def fromRedisKV[T](keysOrKeyPattern: T,
                  partitionNum: Int = 3)
  (implicit redisConfig: RedisConfig = new RedisConfig(new RedisEndpoint))
  RDD[(String, String)] = {
    keysOrKeyPattern match {
      case keyPattern: String => fromRedisKeyPattern(keyPattern, partitionNum)(redisConfig)
      case keys: Array[String] => fromRedisKeys(keys, partitionNum)(redisConfig).getKVs
      case _ => throw new scala.Exception("KeysOrKeyPattern should be String or Array[String]")
    }
  }
}
```

(/apps/redis-
utm_source=cli-
banner-clip

先看传入的参数：

1. 泛型类型 keysOrKeyPattern

从的模式匹配代码中可以看出，这里的 `T` 可是是两种类型，一个是 `String`，另一个是 `Array[String]`，如果传入其他类型则会抛出运行时异常，其中 `String` 类型的意思是匹配键，这里可以用通配符比如 `foo*`，所以返回值是一个结果集 `RDD[(String, String)]`，当参数类型为 `Array[String]` 时是指传入key的数组，返回的结果则为相应的结果集，RDD的内容类型也是KV形式。

2. Int 类型 partitionNum

生成RDD的分区数，默认为 3，如果传入的第一个参数类型是 `Array[String]`，这个参数可以这样设置，先预估一下返回结果集的大小，使用 `keyArr.length / num + 1`，这样则保证分区的合理性，以防发生数据倾斜。若第一个参数类型为 `String`，能预估尽量预估，如果实在没办法，比如确实在这里发生了数据倾斜，可以尝试考虑使用 `sc.fromRedisKeys()` 返回 key 的集合，提前把握返回结果集的大小，或者根据集群机器数量，把握分区数。

(https://dsp-
click.youdao.com/
slot=30edc
f3b6-496b-
0a451c81c
290606218

3. 柯里化形式隐式参数 redisConfig

由于我们之前在 `sparkConf` 里面set了相应的参数，这里不传入这个参数即可。如要调整，则可以按照源码中的方式传入，其中 `RedisEndpoint` 是一个 `case class` 类，而且很多参数都有默认值（比如 6379 的端口号），所以自己建立一个 `RedisEndpoint` 也是非常方便的。

了解了参数之后来继续完成测试代码：

```
/*这里标出了resultSet的类型*/
val resultSet: RDD[(String, String)] = sc.fromRedisKV("to*")
//找出键以`to`开头的键值对，这里就不进行计算了，直接保存到HDFS看结果如何，同时合并分区便于观察
resultSet.coalesce(1).saveAsTextFile("HDFSpath")
```

现在往redis里面随便set几个数据



```
127.0.0.1:6379> set too 111
OK
127.0.0.1:6379> set abc 222
OK
127.0.0.1:6379> set together 333
OK
```

Redis shell

(/apps/redi
utm_sourc
banner-clic

打包之后运行，命令为：

```
spark-submit --master yarn --deploy-mode cluster --class test.SparkRedis --jars jedis-2.9.0.jar,spark-redis-0.3.2.jar,commons-pool2-2.2.jar
```

命令中指明了依赖的资源包: jedis-2.9.0.jar,spark-redis-0.3.2.jar,commons-pool2-2.2.jar 其中 commons-pool2-2.2.jar 是spark-redis依赖的包，如果集群环境为CDH发行版，可在 /opt/cloudera/parcels/CDH/jars/commons-pool2-2.2.jar 下找到该包，，而且yarn的运行环境里面没有默认引入该包；如果为自建环境，则需要自行下载该包，Maven上搜索 commons-pool2 即可。

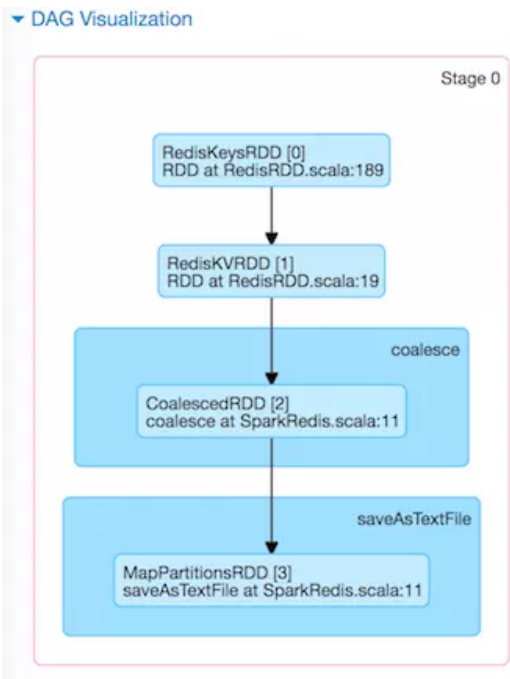
等待执行成功

Logs ID	Name	Application Type	Status	User	Maps	Reduces	Queue	Priority	Duration	Submitted
1486345219544_0205	test.SparkRedis	SPARK	SUCCEEDED	root	100%	100%	root.users.root	N/A	11s	03/29/17 10:28:00

执行UI

(https://dsp
click.youda
slot=30edc
f3b6-496b-
0a451c81c
290606218

再看一下DAG图



DAG图

运行结果



Home / spark / test / redisResult / part-00000

```
(too,111)
(tought,555)
(together,333)
```

运行结果

和预期的一样，以 to 开头的数据都被找到。

如果传入的是key数组

```
val keys = Array[String]("high", "abc", "together")
sc.fromRedisKV(keys).coalesce(1).saveAsTextFile("hdfs://nameservice1/spark/test/redisResult2")
```

结果如下：

Home / spark / test / redisResult2 / part-00000

```
(high,444)
(abc,222)
(together,333)
```

运行结果

完整代码：

```
import org.apache.spark.{SparkConf, SparkContext}
import com.redislabs.provider.redis._

object SparkRedis extends App {
  val conf = new SparkConf().setMaster("yarn-cluster").setAppName("sparkRedisTest")
  conf.set("redis.host", "10.1.11.70")
  val sc = new SparkContext(conf)
  val keys = Array[String]("high", "abc", "together")
  sc.fromRedisKV(keys).coalesce(1).saveAsTextFile("hdfs://nameservice1/spark/test/redisResult2")
}
```

下面看如何写入Redis

```
sc.toRedis
1 toRedisKV(kvs: RDD[(String, String)], ttl: Int = 0)(implicit
2 toRedisFixedLIST(vs: RDD[String], listName: String, ... Unit
3 toRedisHASH(kvs: RDD[(String, String)], hashName: S... Unit
4 toRedisLIST(vs: RDD[String], listName: String, ttl:... Unit
5 toRedisSET(vs: RDD[String], setName: String, ttl: I... Unit
6 toRedisZSET(kvs: RDD[(String, String)], zsetName: S... Unit
Did you know that Quick Documentation View (F1) works in completion lookups as well? >> ↵
```

还是以常见的KV为例

源码中是这样处理的，接收两个参数，RDD类型为 RDD[(String, String)]，第二个为失效时间

(/apps/redi
utm_sourc
banner-clic

(https://dsp
click.youda
slot=30edc
f3b6-496b-
0a451c81c
290606218



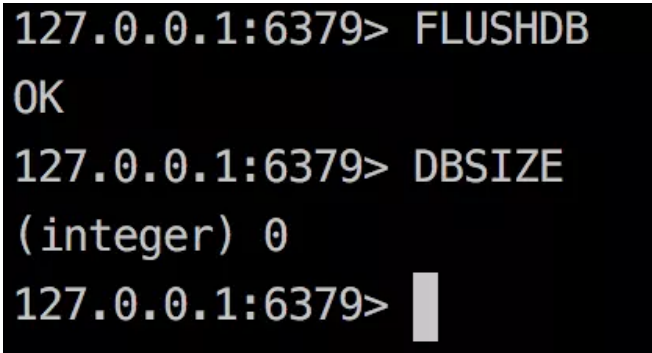
```
def toRedisKV(kvs: RDD[(String, String)], ttl: Int = 0)
    (implicit redisConfig: RedisConfig = new RedisConfig(new RedisEndpoint
        kvs.foreachPartition(partition => setKV(partition, ttl, redisConfig))
    }
```

(/apps/redi
utm_sourc
banner-clic

测试代码为：

```
val data = Seq[(String,String)](("high","111"), ("abc","222"), ("together","333"))
val redisData:RDD[(String,String)] = sc.parallelize(data)
sc.toRedisKV(redisData)
```

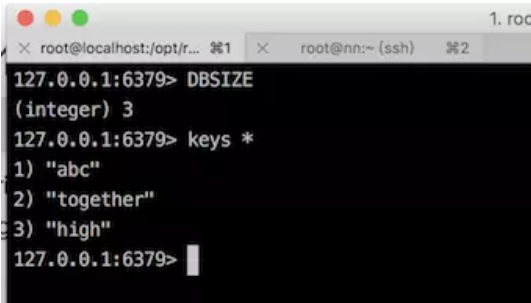
先清空一下redis



redis shell

(https://dsp
click.youda
slot=30edc
f3b6-496b-
0a451c81c
290606218

打包后按相同的命令提交到集群并执行成功后即可看到数据



redis shell

完整代码：



```
import org.apache.spark.{SparkConf, SparkContext}
import com.redislabs.provider.redis._
import org.apache.spark.rdd.RDD

object SparkRedis extends App {
  val conf = new SparkConf().setMaster("yarn-cluster").setAppName("sparkRedisTest")
  conf.set("redis.host", "10.1.11.70")
  val sc = new SparkContext(conf)
  val data = Seq[(String,String)](("high","111"), ("abc","222"), ("together","333"))
  val redisData:RDD[(String,String)] = sc.parallelize(data)
  sc.toRedisKV(redisData)
}
```

(/apps/redi
utm_sourc
banner-clic

原创文章@贪恋清晨de阳光 (http://weibo.com/534212040)

小礼物走一走，来简书关注我

赞赏支持

📖 Spark开发笔记 (/nb/11264799)

举报文章 © 著作权归作者所有

(https://dsp
click.youde
slot=30edc
f3b6-496b-
0a451c81c
290606218



贪恋清晨de阳光 (/u/32f6fe29fdc4)

+ 关注

写了 1066 字，被 3 人关注，获得了 4 个喜欢

(/u/32f6fe29fdc4)

后端攻城狮一枚，在学机器学习，却被卡在了数学，所以目前在学数学。。

喜欢 | 4



开发10年

全记在这本Java进阶宝典了

Spring源码分析

分布式架构

微服务架构

JVM性能优化

高效DevOps

多线程并发编程

点击领取

(/p/428251ede1aa)



登录 (/sign) 发表评论 (source=desktop&utm_medium=not-signed-in-comr



9条评论 只看作者

按时间倒序 按时间正序



望山不是山 (/u/df7a40bf8a24)

6楼 · 2018.11.29 14:56

(/u/df7a40bf8a24)

使用`sc.fromRedisKV("a").foreachPartititon(it=>{it.foreach(println)})`这个job居然用了8秒中，这速度也太慢了吧

赞 回复

(/apps/redi
utm_sourc
banner-clc



青岩虚谷 (/u/5217dc0beef3)

5楼 · 2018.11.02 19:54

(/u/5217dc0beef3)

maven工程，可以用这个依赖，能把jar包下下来。

```
<dependency>
```

```
<groupId>com.redislabs</groupId>
```

```
<artifactId>spark-redis</artifactId>
```

```
<version>2.3.1-M1</version>
```

```
</dependency>
```

赞 回复

贪恋清晨de阳光 (/u/32f6fe29fdc4) : 👍

2018.11.06 18:35 回复

添加新评论

(https://dsp
click.youda
slot=30edc
f3b6-496b-
0a451c81c
290606218



青岩虚谷 (/u/5217dc0beef3)

4楼 · 2018.11.02 19:51

(/u/5217dc0beef3)

然而没有删除功能。

赞 回复



爱遗忘在了五月天 (/u/cfaa6932928d)

3楼 · 2018.09.10 14:29

(/u/cfaa6932928d)

我的jar包下下来怎么用不了

赞 回复



流浪猫的王子 (/u/03cd023081b3)

2楼 · 2017.07.11 11:32

(/u/03cd023081b3)

用maven的怎么整，大神指条明路啊

赞 回复

贪恋清晨de阳光 (/u/32f6fe29fdc4) : maven仓库里面没有这个包，只能自己编译，或者直接在工程里面加进这个代码

2017.07.24 16:21 回复


贪恋清晨de阳光 (/u/32f6fe29fdc4) : @Luis_afa7 (/users/1ae19dc45e77) 用Spark2环境编译一下就行了，代码不多，好处是在分布式环境下，可以和redis的分区对应上


2017.11.15 18:02 回复

添加新评论 | 还有1条评论，展开查看



被以下专题收入，发现更多相似内容


 Spark原理用法 (/c/62b878972bad?utm_source=desktop&utm_medium=notes-included-collection)

 Python爬... (/c/1c860922ba12?utm_source=desktop&utm_medium=notes-included-collection)

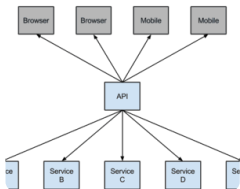
(/apps/redi
utm_sourc
banner-clc

Apache Spark 2.2.0 中文文档 - Spark SQL, DataFrames... (/p/3a7e3480c...

Spark SQL, DataFrames and Datasets Guide Overview SQL Datasets and DataFrames 开始入门 起始点:
SparkSession 创建 DataFrames 无类型的Dataset操作 (aka Dat...

 片刻_ApacheCN (/u/a5d135d71592?
utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendatio


(/p/46fd0faecac1?



(https://dsp
click.youde
slot=30edc
1306-456b
0a451c81c
290606218


utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendatio
Spring Cloud (/p/46fd0faecac1?utm_campaign=maleskine&utm_conte...

Spring Cloud为开发人员提供了快速构建分布式系统中一些常见模式的工具（例如配置管理，服务发现，断
路器，智能路由，微代理，控制总线）。分布式系统的协调导致了样板模式。使用Spring Cloud开发人员可...

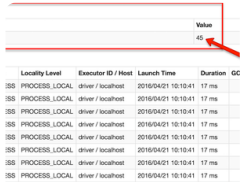
 卡卡罗2017 (/u/d90908cb0d85?
utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendatio

Apache Spark 2.2.0 中文文档 - Spark SQL, DataFrames... (/p/238427535...

Spark SQL, DataFrames and Datasets Guide Overview SQL Datasets and DataFrames 开始入门 起始点:
SparkSession 创建 DataFrames 无类型的Dataset操作 (aka Dat...

 Joyyx (/u/5d6219efd1b8?
utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendatio


(/p/d43ab8f3b779?



Locality Level	Executor ID / Host	Launch Time	Duration	GC
SSS	PROCESS_LOCAL driver / localhost	2016/04/21 10:10:41	17 ms	
SSS	PROCESS_LOCAL driver / localhost	2016/04/21 10:10:41	17 ms	
SSS	PROCESS_LOCAL driver / localhost	2016/04/21 10:10:41	17 ms	
SSS	PROCESS_LOCAL driver / localhost	2016/04/21 10:10:41	17 ms	
SSS	PROCESS_LOCAL driver / localhost	2016/04/21 10:10:41	17 ms	
SSS	PROCESS_LOCAL driver / localhost	2016/04/21 10:10:41	17 ms	
SSS	PROCESS_LOCAL driver / localhost	2016/04/21 10:10:41	17 ms	
SSS	PROCESS_LOCAL driver / localhost	2016/04/21 10:10:41	17 ms	

utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendatio
Apache Spark 2.2.0 中文文档 - Spark 编程指南 | ApacheCN (/p/d43ab8f3...

Spark 编程指南 概述 Spark 依赖 初始化 Spark 使用 Shell 弹性分布式数据集 (RDDs) 并行集合 外部
Datasets (数据集) RDD 操作 基础 传递 Functions (函数) 给 Spark 理解闭包 示例 Local (本地) vs. cl...

 片刻_ApacheCN (/u/a5d135d71592?
utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendatio


(/p/22c450a71328?



(/apps/redi
utm_sourc

utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendatio
Apache Spark 2.2.0 中文文档 - Spark Streaming 编程指南 ... (/p/22c450a...


Spark Streaming 编程指南 概述 一个入门示例 基础概念 依赖 初始化 StreamingContext Discretized Streams
(DStreams) (离散化流) Input DStreams 和 Receivers (接收器) DStreams...

 片刻_ApacheCN (/u/a5d135d71592?

utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendatio

《辐射小马国：七罪之咒》第二十三话（上）： (/p/2fd194daa77f?utm_ca...


清晨 在餐厅里，小皮生气的吃着饭，月华奎灵和薇薇·莱米在一边非常尴尬的看着小皮。 敬心在厨房里看着
月华奎灵她们笑着说到，“小皮火气还没下来呢.....毕竟昨天小皮刚设计好风锯想庆祝下结果你们两个去幽...

 月华奎灵 (/u/9b6de836a211?

utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendatio

爱在旁若无人的太空 (/p/69728e5cd0d6?utm_campaign=maleskine&utm...

杨千嬅 爱在旁若无人的太空 <http://www.xiami.com/song/1774477879>

 seveneawa (/u/bfbe6807a83b?


utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendatio
(https://dsp
click.youde
slot=30edc
f3b6-496b-
0a451c81c
290606218

(/p/2384ec209362?



utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendatio
后发制人！利物浦2比1逆转斯托克城 (/p/2384ec209362?utm_campaign=...

北京时间4月8日22:00，英超第32轮的一场比赛，红军利物浦做客挑战斯托克城。上半场比赛，乔-阿伦伤
退，沙奇里助攻沃尔特斯头球破门；下半场比赛，洛夫伦头球击中横梁，库蒂尼奥、菲尔米诺3分钟内连进...

 英超热点 (/u/e3db5505d25f?


utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendatio

(/p/97f01b06f8a4?



utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendatio
手机 (/p/97f01b06f8a4?utm_campaign=maleskine&utm_content=note&...

一个小小匡 无嘴能说话 时代骄宠儿 朝思暮想它 一刻不舍离 悲喜手中拿 时代海洛因 限你思想家 有它无弊处
只因任务大 行走是距离 难比屏幕划 唐诗宋词有 元曲明清戏 经典诗词会 优美句子多 曲折难创作 极简信息...

 豫视西影 (/u/55809ddcfef5?

utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendatio



(/apps/redi
utm_sourc
banner-clic

(https://dsp
click.youda
slot=30edc
f3b6-496b-
0a451c81c
290606218

