1 1 help

creating spark data structure from multiline record

I'm trying to read in retrosheet event file into spark. The event file is structured as such.

```
id,TEX201403310
version,2
info,visteam,PHI
info,hometeam,TEX
info,site,ARL02
info,date,2014/03/31
info,number,0
info,starttime,1:07PM
info,daynight,day
info,usedh,true
info,umphome,joycj901
info,attendance,49031
start,reveb001,"Ben Revere",0,1,8
start,rollj001,"Jimmy Rollins",0,2,6
start,utlec001,"Chase Utley",0,3,4
start,howar001,"Ryan Howard",0,4,3
start,byrdm001,"Marlon Byrd",0,5,9
id,TEX201404010
version,2
info,visteam,PHI
info,hometeam,TEX
```

As you can see for each game the events loops back.

I've read the file into a RDD, and then via a second for loop added a key for each iteration, which appears to work. But I was hoping to get some feedback on if there was a cleaning way to do this using spark methods.

```
logFile = '2014TEX.EVA'
event_data = (sc
              .textFile(logfile)
              .collect())

idKey = 0
newevent_list = []
for line in event_dataFile:
    if line.startswith('id'):
        idKey += 1
        newevent_list.append((idKey,line))
    else:
        newevent_list.append((idKey,line))

event_data = sc.parallelize(newevent_list)
```

python apache-spark pyspark

edited Jul 5 '15 at 11:35

zero323
45.5k 11 48 79

asked Jul 5 '15 at 5:17

user1136149
38 3

1 Answer

PySpark since version 1.1 supports Hadoop Input Formats.You can use `textinputformat.record.delimiter` option to use a custom format delimiter as below

```
retrosheet = sc.newAPIHadoopFile(
    '/path/to/retrosheet/file',
    'org.apache.hadoop.mapreduce.lib.input.TextInputFormat',
    'org.apache.hadoop.io.LongWritable',
    'org.apache.hadoop.io.Text',
    conf={'textinputformat.record.delimiter': '\nid,'}
)
(retrosheet
    .filter(lambda (k, v): v)
    .map(lambda (k, v): (
        v if v.startswith('id') else 'id,{0}'.format(v)).splitlines()))
```

edited Jul 5 '15 at 11:43 answered Jul 5 '15 at 11:24

zero323
45.5k 11 48 79

very cool, thanks! – user1136149 Jul 25 '15 at 3:57