ps6

Baoyue Liang

10/25/2018

Problem2

```
library(RSQLite)
drv <- dbDriver("SQLite")
dbFilename <- 'stackoverflow-2016.db'
dir <- "/Users/lby/Desktop/ps6"
db <- dbConnect(drv, dbname = file.path(dir, dbFilename))

dbGetQuery(db, "
           select count(*) from
           (SELECT distinct ownerid
           from (questions_tags join questions on questions_tags.questionid = questions.questionid)
           WHERE tag LIKE '%apache-spark%'

           except

           SELECT distinct ownerid
           from (questions_tags join questions on questions_tags.questionid = questions.questionid)
           WHERE tag like '%python%')

          ")
```

```
##   count(*)
## 1     4647
```

Problem3

code

```
# use bash in terminal
srun -A ic_stat243 -p savio2 --nodes=4 -t 3:00:00 --pty bash
module load java spark/2.1.0 python/3.5
source /global/home/groups/allhands/bin/spark_helper.sh
spark-start
pyspark --master $SPARK_URL --conf "spark.executorEnv.PYTHONHASHSEED=321"  --executor-memory 60G
```

```
# python code
import re
from operator import add
from pyspark import SparkContext
lines = sc.textFile('/global/scratch/paciorek/wikistats_full/dated')
def find(line, regex = "Day_of_the_Dead", language = None):
    vals = line.split(' ')
    if len(vals) < 6:
```

```
        return(False)
    if (int(vals[0]) >= 20081201):
        return(False)
    tmp = re.search(regex, vals[3])
    if tmp is None or (language != None and vals[2] != language):
        return(False)
    else:
        return(True)
travel = lines.filter(find)
def stratify(line):
    vals = line.split(' ')
    return(vals[0] + '-' + vals[2], int(vals[4]))
counts = travel.map(stratify).reduceByKey(add)
def transform(vals):
    key = vals[0].split('-')
    return(",".join((key[0], key[1], str(vals[1]))))
outputDir = '/global/home/users/byliang/travel-counts'
counts.map(transform).repartition(1).saveAsTextFile(outputDir)
```

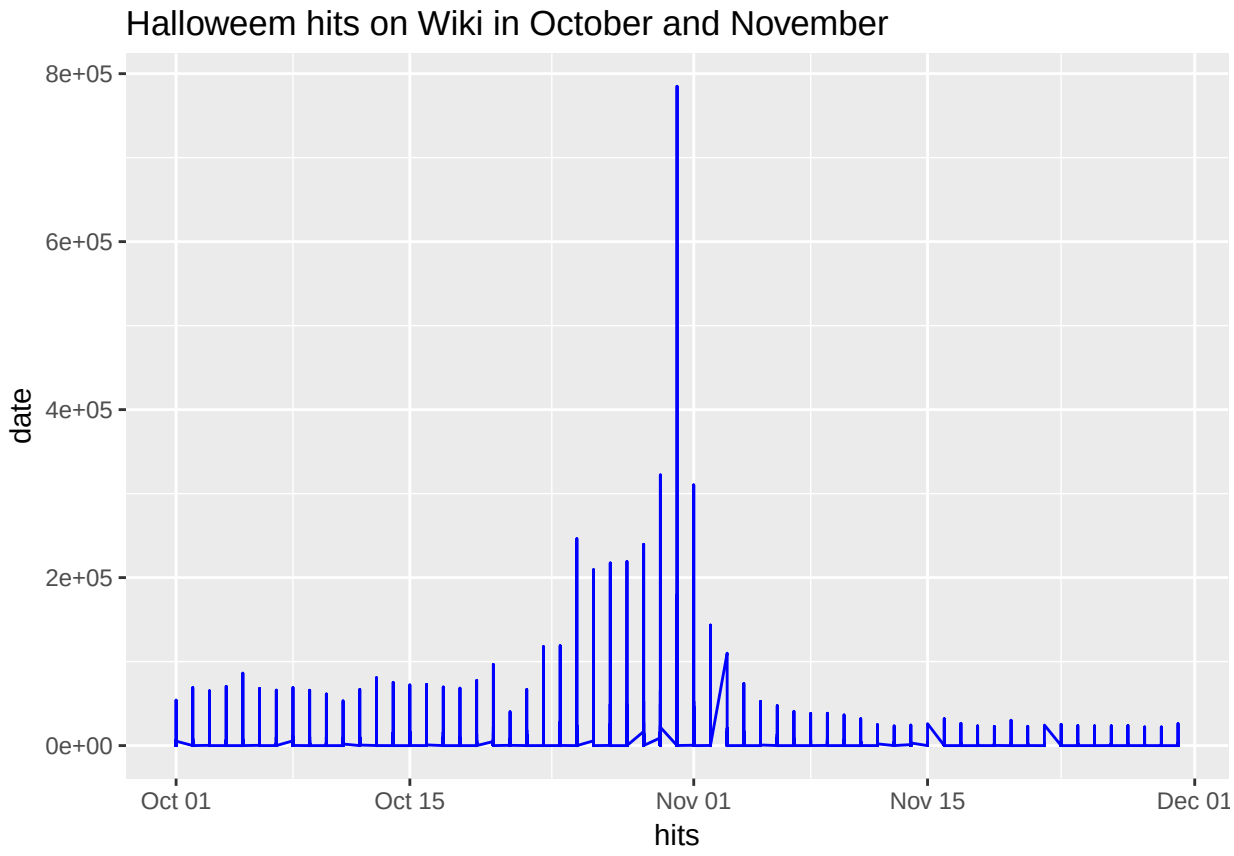
Data analysis

```
library(ggplot2)

dataHal = read.table("part-00000",header = FALSE,sep = ",")
colnames(dataHal) = c("date","content","hits")
dataHal$date = as.Date(as.character(dataHal$date), "%Y%m%d")
```

```
## Warning in strptime(x, format, tz = "GMT"): unknown timezone 'default/
## America/Los_Angeles'
```

```
ggplot(dataHal, aes(x=date)) +
  geom_line(aes(y=hits), colour="blue") +
  labs(x = "hits", y = "date", title = "Halloweem hits on Wiki in October and November")
```



Problem4

a

```
# use bash in terminal
srun -A ic_stat243 -p savio2 --nodes=1 -t 2:00:00 --pty bash
module load r r-packages
R
```

```
# r code
library(parallel)
library(doParallel)
library(foreach)

registerDoParallel(Sys.getenv("SLURM_CPUS_ON_NODE"))

n = 959
re = foreach(i = 0:n,
             .packages = c("readr","stringr","dplyr"),
             .combine = rbind,
             .verbose = TRUE) %dopar% {
        filepath = paste("/global/scratch/paciorek/wikistats_full/dated_for_R/part-",
        str_pad(i,width = 5,side ='left',pad = "0"),sep = "")
        data = read_delim(filepath,delim = " ",col_names = FALSE )
        Obama = data[grep("Barak_Obama",data$X4),]
```

```
        Obama
             }
dim = dim(Obama)
write.table(dim, file = "/global/home/users/byliang/result.txt")
write.table(Obama, file = "/global/home/users/byliang/Obama.txt")
```

b

I ran the code on only one fourth of the files and it takes me around 20 minutes. Suppose I can achieve perfect scalability, it will takes me around 80 minutes to run on all files on one node (20 minutes on 4 nodes), which is slower than pyspark.