1A)

mydata = sc.textFile("/loudacre/weblogs/")

mydata.map(lambda line: line.split(' ')).filter(lambda fields: "html" in fields[6]).keyBy(lambda fields: (fields[0] + "/" + fields[2])).keys().take(10)

[u'3.94.78.5/69827',

u'19.38.140.62/21475',

u'129.133.56.105/2489',

u'217.150.149.167/4712',

u'209.151.12.34/45922',

u'184.97.84.245/144',

u'233.60.251.2/33908',

u'160.134.139.204/51340',

u'19.209.18.222/13392',

u'230.220.223.28/12643']

mydata.map(lambda line: line.split(' ')).filter(lambda fields: "html" in fields[6]).keyBy(lambda fields: (fields[0] + "/" + fields[2])).keys().saveAsTextFile("loudacre\_weblogs\_html")

hdfs dfs -cat loudacre\_weblogs\_html/part-00000 | head

3.94.78.5/69827

19.38.140.62/21475

129.133.56.105/2489

217.150.149.167/4712

209.151.12.34/45922

184.97.84.245/144

233.60.251.2/33908

160.134.139.204/51340

19.209.18.222/13392

230.220.223.28/12643

1B)

mydata.map(lambda line: line.split(' ')).keyBy(lambda fields: (fields[0] + "/" + fields[2])).count()

1079891 total web log records

mydata.map(lambda line: line.split(' ')).filter(lambda fields: "html" in fields[6]).keyBy(lambda fields: (fields[0] + "/" + fields[2])).count()

474360 HTML page requests

Upload data to HDFS:

$ cd ~/training\_materials/dev1/data/

$ hdfs dfs -mkdir /loudacre/activations/

$ hdfs dfs -put activations/\* /loudacre/activations/.

2A)

The keys are the paths to the files.

mydata.keys().take(2)

[u'hdfs://localhost:8020/loudacre/activations/2008-10.xml',

u'hdfs://localhost:8020/loudacre/activations/2008-11.xml']

The value is the entire contents of the file.

2B)

flatMap()

import xml.etree.ElementTree as ElementTree

def getactivations(s):  
    filetree = ElementTree.fromstring(s)  
    return filetree.getiterator('activation')

xmldata = mydata.flatMap(lambda fields: getactivations(fields[1]))

2C)

def getmodel(activation):

return activation.find('model').text

def getaccount(activation):

      return activation.find('account-number').text

xmldata.map(lambda activation: getaccount(activation) + ":" + getmodel(activation)).saveAsTextFile("/loudacre/account-models")

3A)

3B)