ECE 536 Assignment #7

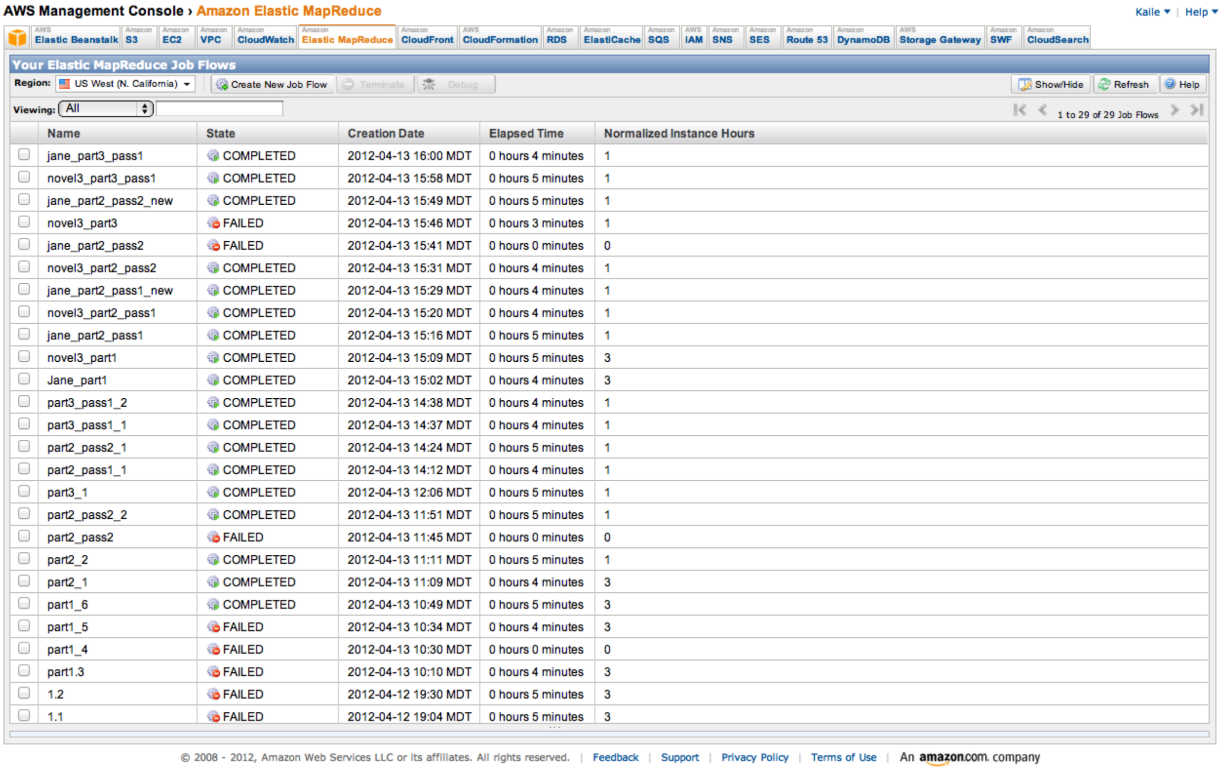
Kaile Liang

April 14, 2012

1 part1

1.1 Screen Capture

Figure 1: Screen Capture of AWS Management Console


AWS Management Console › Amazon Elastic MapReduce
Kaile ▾ | Help ▾
Elastic Beanstalk
S3
EC2
VPC
CloudWatch
Elastic MapReduce
CloudFront
CloudFormation
RDS
ElastiCache
SQS
IAM
SNS
SES
Route 53
DynamoDB
Storage Gateway
SWF
CloudSearch
Your Elastic MapReduce Job Flows
Region: US West (N. California)
Create New Job Flow
Terminate
Debug
Show/Hide
Refresh
Help
Viewing: All
1 to 29 of 29 Job Flows
© 2008 - 2012, Amazon Web Services LLC or its affiliates. All rights reserved. | Feedback | Support | Privacy Policy | Terms of Use | An amazon.com company

	Name	State	Creation Date	Elapsed Time	Normalized Instance Hours
□	jane_part3_pass1	COMPLETED	2012-04-13 16:00 MDT	0 hours 4 minutes	1
□	novel3_part3_pass1	COMPLETED	2012-04-13 15:58 MDT	0 hours 5 minutes	1
□	jane_part2_pass2_new	COMPLETED	2012-04-13 15:49 MDT	0 hours 5 minutes	1
□	novel3_part3	FAILED	2012-04-13 15:46 MDT	0 hours 3 minutes	1
□	jane_part2_pass2	FAILED	2012-04-13 15:41 MDT	0 hours 0 minutes	0
□	novel3_part2_pass2	COMPLETED	2012-04-13 15:31 MDT	0 hours 4 minutes	1
□	jane_part2_pass1_new	COMPLETED	2012-04-13 15:29 MDT	0 hours 4 minutes	1
□	novel3_part2_pass1	COMPLETED	2012-04-13 15:20 MDT	0 hours 4 minutes	1
□	jane_part2_pass1	COMPLETED	2012-04-13 15:16 MDT	0 hours 5 minutes	1
□	novel3_part1	COMPLETED	2012-04-13 15:09 MDT	0 hours 5 minutes	3
□	Jane_part1	COMPLETED	2012-04-13 15:02 MDT	0 hours 4 minutes	3
□	part3_pass1_2	COMPLETED	2012-04-13 14:38 MDT	0 hours 4 minutes	1
□	part3_pass1_1	COMPLETED	2012-04-13 14:37 MDT	0 hours 4 minutes	1
□	part2_pass2_1	COMPLETED	2012-04-13 14:24 MDT	0 hours 5 minutes	1
□	part2_pass1_1	COMPLETED	2012-04-13 14:12 MDT	0 hours 4 minutes	1
□	part3_1	COMPLETED	2012-04-13 12:06 MDT	0 hours 5 minutes	1
□	part2_pass2_2	COMPLETED	2012-04-13 11:51 MDT	0 hours 5 minutes	1
□	part2_pass2	FAILED	2012-04-13 11:45 MDT	0 hours 0 minutes	0
□	part2_2	COMPLETED	2012-04-13 11:11 MDT	0 hours 5 minutes	1
□	part2_1	COMPLETED	2012-04-13 11:09 MDT	0 hours 4 minutes	3
□	part1_6	COMPLETED	2012-04-13 10:49 MDT	0 hours 5 minutes	3
□	part1_5	FAILED	2012-04-13 10:34 MDT	0 hours 4 minutes	3
□	part1_4	FAILED	2012-04-13 10:30 MDT	0 hours 0 minutes	0
□	part1.3	FAILED	2012-04-13 10:10 MDT	0 hours 4 minutes	3
□	1.2	FAILED	2012-04-12 19:30 MDT	0 hours 5 minutes	3
□	1.1	FAILED	2012-04-12 19:04 MDT	0 hours 5 minutes	3

1.2 Choice of 3rd Dataset

I select from the same webset of the first two, a novel named *Gamble with Life*, by Silas K. Hocking. It has 121443 words.

1.3 Outputs:

Alice's Adventures in Wonderland

Listing 1: Output of Alice's Adventures in Wonderland

```
"--SAID 1
"Come   1
"HOW    1
"I      7
"I'll   2
"Project        5
"Such   1
"With   1
"YOU    1
'"--found       1
you!    2
you!'   3
you—all         1
you?    2
you?'   7
young   5
your    62
yourself!'      1
yourself,       1
zigzag, 1
```

Jane Eyre

Listing 2: Output of Jane Eyre

```
"'Go,'  1
"'I     2
"Adele  2
"Ah!    13
"Aire?  1
"Alas!  1
"All,   1
"Amen!  1
"An     2
"And,   2
yourself;       1
youth—only      1
youth?  1
youthful        2
zeal,   1
zealous 1
zenith, 1
{Hush,  1
{I      2
{The    1
```

Gamble With Life

Listing 3: Output of Gamble With Life

```
”’The     1
”’Tis     1
”AND      1
”Admirable        1
”Ah       1
”Ah!      10
”Allow    1
”An       6
”And,     2
”Another          1
young     63
young.    3
young.”   1
younger   4
your      231
yourself,         2
yourself,”        1
youthful          1
zest      3
zigzags;          1
```

2 part2

2.1 Screen Capture

See Figure 1

2.2 Program:

After the first reduce, the size of the file are largely decreased. Considering the requirement, we have to sort all the output, so I decided to use 1 instead of 3 instance to run.

1. Step1: use the result of part1 as input, run first pass of mapreduce, output the unsorted length, word, count.

2. Step2: use the result of Step1 as input, run second pass of mapreduce to sort the file.

Code For Mapper

Listing 4: Mapper used in step1 and step2

```
#!/usr/bin/env python
import sys
# input comes from STDIN (standard input)
for line in sys.stdin:
# remove leading and trailing whitespace
    line = line.strip()
    print line
```

Code for Step1 Reducer

Listing 5: Step1 Reducer

```
#!/usr/bin/env python
from operator import itemgetter
import sys

current_word = None
current_count = 0
current_length = 0
word = None

# input comes from STDIN
for line in sys.stdin:
    # remove leading and trailing whitespace
    line = line.strip()
    # parse the input we got from mapper.py
    word, count = line.split('\t', 1)

    # convert count (currently a string) to int
    try:
        count = int(count)
    except ValueError:
        # count was not a number, so silently
        # ignore/discard this line
        continue

    #filter the word has length [10, 20)
    length = len(word)
    if(length>=10 and length<20):
        if current_word == word:
            current_count += count
        else:
            if current_word:
                current_length = len(current_word)
                print '%s\t%s\t%s' %(len(current_word),current_word,current_count)
            current_count = count
            current_word = word
    else:
        continue


#print the last
current_length = len(current_word)
print '%s\t%s\t%s' %(len(current_word),current_word,current_count)
```

Code for Step2 Reducer

Listing 6: Step2 Recuder

```
#!/usr/bin/env python
from operator import itemgetter
import sys
```

```
word = None
length = 0
# input comes from STDIN
for line in sys.stdin:
    # remove leading and trailing whitespace
    line = line.strip()
    # parse the input we got from mapper.py
    length, word, count = line.split('\t', 2)

    # convert count and length (currently a string) to int
    # print each line
    try:
        length = int(length)
        count = int(count)
        print '%s\t%s\t%s' %(length,word,count)
    except ValueError:
        # count was not a number, so silently
        # ignore/discard this line
        continue
```

2.3 Output

Alice's Adventures in Wonderland

Listing 7: Output of Alice's Adventures in Wonderland

```
10        Turtle—we         1
10        Quadrille,        1
10        Everything        1
10        'Hjckrrh!'        1
10        Fairbanks,        1
10        Pennyworth        2
10        Forty-two.        1
10        Foundation        14
10        associated        7
10        Normans—”         1
17        treacle-well—eh,          1
17        bread-and-butter.         1
17        bread-and-butter,         2
17        WASHING—extra.”'          1
17        WAISTCOAT-POCKET,         1
17        particular—Here,          1
17        gbnewby@pglaf.org         1
18        things—everything         1
19        business@pglaf.org.       1
19        bread-and-butter—'        1
```

Jane Eyre

Listing 8: Output of Jane Eyre

```
10        faith—her         1
10        distress?”        1
```

```
10      distressed      4
10      handling."      1
10      distresses      1
10      handiwork:      1
10      fairy-like      3
10      distribute      5
10      horseback,      1
10      faintness.      1
19      When—how—whither,       1
19      autumn,--Thornfield     1
19      _ignis-fatus_-like,     1
19      keeping,--heirlooms     1
19      imagination,--tall,     1
19      pocket-handkerchief     1
19      instrument—nothing      1
19      proprietor—nothing      1
19      inquisitive-looking     1
19      melancholy-looking.     1
```

Gamble With Life

Listing 9: Output of Gamble With Life

```
10      =Ourselves      1
10      Testament,      1
10      Temperance      2
10      Telephone,      2
10      Teaching.=      1
10      Tabernacle      1
10      Sympathise      1
10      attention.      5
10      Supplement      2
10      =Practical      1
19      time."—-_Nottingham     1
19      story."—-_Newcastle     1
19      unless—unless——"        1
19      study."—-_Ardrossan     1
19      standpoint—'Life's      1
19      teachers."—-_Sunday     1
19      uncommunicativeness     1
19      business@pglaf.org.     1
19      friend."—-_Brighton     1
19      encouraging."—-_The     1
```

3 part3

3.1 Screen Capture

See Figure 1

3.2 Program:

Use 1 instance, and use the result of part2, do another pass of mapreduce.

Code for Mapper

I use the same mapper as the one used in part2.

Code for Reducer

Listing 10: Step1 Reducer

```
#!/usr//bin/env python
from operator import itemgetter
import sys

current_word = None
current_length = 0
word = None
max_count = 0
max_word = None

# input comes from STDIN
for line in sys.stdin:
    # remove leading and trailing whitespace
    line = line.strip()
    # parse the input we got from mapper.py
    length, word, count = line.split('\t', 2)

    # convert length count (currently a string) to int
    try:
        length = int(length)
        count = int(count)
    except ValueError:
        # count was not a number, so silently
        # ignore/discard this line
        continue

    if(length>=10 and length<20):
        if current_length ==length:
            current_word = word
            #get the max count word
            if count > max_count:
                max_count=count
                max_word = current_word
        else:
            if max_count:
                #print out the max count word
                print '%s\t%s\t%s' %(current_length,max_word,max_count)
            current_word = word
            max_word = word
            current_length = length
            max_count = count
    else:
        continue

# do not forget to output the last word if needed!
if current_length ==length:
    print '%s\t%s\t%s' %(current_length,max_word,max_count)
```

3.3 Output

Alice's Adventures in Wonderland

Listing 11: Output of Alice's Adventures in Wonderland

```
10      electronic      27
11      Caterpillar     11
12      Gutenberg-tm    53
13      conversation.   5
14      e—e—evening,    3
15      contemptuously. 2
16      http://pglaf.org        2
17      bread-and-butter,       2
18      things——everything      1
19      business@pglaf.org.     1
```

Jane Eyre

Listing 12: Output of Jane Eyre

```
10      Rochester,      69
11      Rochester's     44
12      Gutenberg-tm    53
13      Brocklehurst,   14
14      circumstances,  8
15      unsophisticated 3
16      accomplishments.        3
17      incomprehensible:       2
18      woman,--impossible      2
19      fashionable-looking     1
```

Gamble With Life

Listing 13: Output of Gamble With Life

```
10      everything      31
11      questioned,     48
12      Gutenberg-tm    53
13      circumstances   13
14      accountability  7
15      disappointment, 6
16      acquaintanceship        3
17      www.gutenberg.org       2
18      cost."--_Pearson's      1
19      http://www.pgdp.net     2
```