Q Search this Blog

about

Sorting data with mapreduce

Sorting is simple in mapreduce but a the same time not very intuitive. Sorting large files with **mapreduce** is a step that is essentially step in many graph algorithms including the famous **PageRank**.

The mapper phase of the algorithm takes a key, value pair say (k1,v1) and we want to sort the data based on value. In this case mapper would simply output (v1,k1) as its output. The keys received by the reducer are in sorted order, so the reducer just needs to output its input, i.e. it is an identity reducer.

So to summarize

```
Mapper Phase : Input (k1,v1) ->Output (v1,k1)

Reducer Phase : Input(v1,k1)->Output(v1,k1)
```

To generate a large amount of data I generate a set of random numbers using the below python code.

```
#!/usr/bin/env python
import sys
import numpy
n=1000000
a=numpy.random.random(n)

count = 1
for item in a:
    print "{0}\t{1}".format(count,item)
    count = count + 1;
```

This generates the input for the map-phase of the map-reduce algorithm $% \left(1\right) =\left(1\right) \left(1\right) \left($

The code for the mapper is

```
#!/usr/bin/env python
import sys
import re

for line in sys.stdin:
    line = line.strip()
    arg = line.split('\t')
    print "{0}\t{1}".format(arg[1], arg[0])
The code for the reducer is
```

```
#!/usr/bin/env python
import sys
import re
#identity reducer...

for line in sys.stdin:
    line = line.strip()
    arg = line.split('\t')
    print "{0}\t{1}".format(arg[0],arg[1])
```

You can download the code and script to run the code on **cloudera vm** from **here**

```
Share this: Facebook StumbleUpon Twitter
```

```
Parallel Programming
```

mapreduce sort hadoop cloudera vm

Archives

May 2011 April 2011

March 2011 February 2011

Blogroll

Discuss
Get Inspired
Get Polling
Get Support
Learn WordPress.com

WordPress Planet

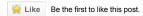
WordPress.com News

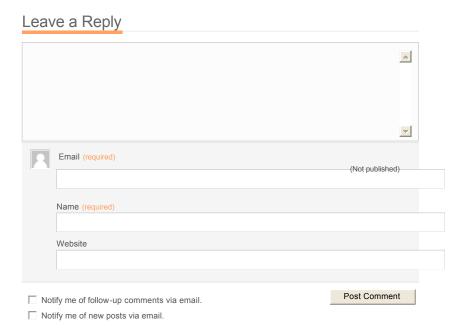
Meta

Register Log in RSS Entries Comments RSS

April 17, 2011

Leave a comment





Blog at WordPress.com.

Theme: Neutra by Artmov.