Large Data Analysis Using Rhipe/RHadoop

赵扬 (Kevin)
Behavioral Insights and Science Team
11/2/2013
1. **What is Rhipe/RHadoop?**

   *Introduce the basic structure of R+Hadoop platform*

2. **Why do we use Rhipe/RHadoop?**

   *Advantages/Disadvantages of using Rhipe*

3. **How to use Rhipe/RHadoop?**

   *Share 2 specific user cases about data analysis using Rhipe*

4. **How to learn R+Hadoop more efficiently**

   *Learn Rhipe step by step, easy and fast*
1. What is Rhipe/RHadoop?
What is Rhipe/RHadoop?

- Rhipe/RHadoop is just a R package which provides an API (应用程序接口) to use Hadoop.
- RHadoop and Rhipe ("hree-pay")
  - RHadoop is comprised of three packages and developed by Revolution Analytics.
    1. RHDFS
    2. RHBASE
    3. rmr
  
  - RHIVE is developed by my classmate Saptarshi Guha in Purdue University:
    1. Same idea of Rhadoop, different API design
    3. I have used Rhipe to do project for 2 years when I pursuing my Phd degree, it works perfect!
What is Rhipe/RHadoop?

Who is the prefect user of Rhipe/RHadoop?

1. Familiar with R
2. Know some basic statistical knowledge (mean, max, min)
3. Get general idea about Hadoop MapReduce framework
Idea of Hadoop

What is Hadoop? Hadoop is used for parallel computing? (HDFS+MapReduce)
1. Divide(map) and Recombine(reduce)
2. Every mapper/Reducer has key-value pairs

Concrete user case: Parallel compute average student weights in a school by gender
2. Why do we use Rhiphe?
Why do we use Rhipe/RHadoop?

Advantages of Rhipe

<table>
<thead>
<tr>
<th></th>
<th>Rhipe</th>
<th>Hadoop(java)</th>
<th>Pig/scoobi/cascading</th>
<th>Snowfall/multicore (R parallel packages)</th>
</tr>
</thead>
<tbody>
<tr>
<td>User Friendly</td>
<td>✔</td>
<td>✗</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Handle Large data set</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✗</td>
</tr>
<tr>
<td>Apply statistical algorithm</td>
<td>✔</td>
<td>✗</td>
<td>✗</td>
<td>✔</td>
</tr>
<tr>
<td>Large Data visualization</td>
<td>✔</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
</tr>
</tbody>
</table>
Why do we use Rhipe/RHadoop?

Drawbacks of Rhipe

1. Need to write R code in map reduce format (learning curve)

2. Not as mature and stable as Hadoop (Java), pig.

3. Need more formal R packages for statistical algorithm computation in large data set.
3. How to use Rhipe?
How to use Rhip/RHadoop?

- Main page is here: [http://www.datadr.org/index.html](http://www.datadr.org/index.html)
- Rhipe installation
  1. Install Hadoop on clusters
  2. Install R as a shared library. This must be either installed on each of the nodes, or packaged as a zip to be passed to the nodes for each job.
  3. Install Google Protocol Buffers
  4. Set Environment Variables like HADOOP Path and R Path
  5. Install Rhipe package
Word Count example using Rhipe

1. example=list()
2. # map step
3. example$map = expression(
4.   words = unlist(strsplit(unlist(map.values), " "))
5.   lapply(words,function(r){rhcollect(r,1)})
6. )]
7. #reduce step
8. example$reduce = expression(
9.   pre = { total = 0 },
10. reduce = { total = total + sum(unlist(reduce.values)) },
11. post = { rhcollect(reduce.key,total) }
12. )
13. #set arguments and run
14. example$ifolder = input_path
15. example$ofolder = output_path
16. example$input = c("text","sequence")
17. example$jobname = "wordCount"
18. example$zips = zips
19. mr = do.call("rhmr",example)
20. ex = rhex(mr)
More complex case to use Rhipe

- Background: We monitored DNS transactions from J-root server in US.
- Data set:
  - 3 days period (Do analysis on 6 months finally)
  - 109,341,181 transactions.
- Goal:
  - Monitor data status
  - Find out Hacker attacks if any
More complex case to use Rhipe

Steps to do data analysis in Rhipe for very large data set (3 steps):

1. Read raw text files into HDFS and create R data base using Rhipe

2. Grab key-value pairs from data base and do map-reduce jobs to get summary results

3. Data visualization
More complex case to use Rhipe

Mosaic due to confidential issues
My previous projects using Rhipe

Arrival Rate (every 30 second) vs Arrival Time.
4. How to learn Rhipe more efficiently?
How to learn R+Hadoop more efficiently

1. start with Ryan Hafen’s manual (easy to interpret):
   http://ml.stat.purdue.edu/rhafen/rhipe/
How to learn R+Hadoop more efficiently

2. Go through examples from Rhipe manual written by Saptarshi Guha

http://www.datadr.org/doc/index.html
How to learn R+Hadoop more efficiently

3. Use Rhipe on your own project.

Thanks