CSE 512 – Distributed Database System – Phase 2

Team Name: DNS

Deepak Soundararajan – 1211181124 - dsounda2@asu.edu
Narendrakumar Sampath Kumar -121106860 - nsampat1@asu.edu
Swathi Kanchan – 1211213195 - skanchan@asu.edu

Requirement:

To create a custom Java Function to perform Naive Spatial Join Query using Cartesian Product Algorithm with the given GeoSpark Jar and compare the Execution time/ Average memory / Average CPU utilization of the entire cluster (one master and two worker nodes) for the set of programs.

Project Description:

The given data set contains the points and the rectangular boundary in the geographical area. The project finds all the points from each of the rectangle of the query window. For every rectangular point, the method combines all the points. (i.e. Cartesian Product)

1. The function should be integrated into GeoSpark source code. The source code should be forked or downloaded from GeoSpark Github master branch LATEST.

2. The result format should be same as GeoSpark spatial join query: JavaPairRDD<Envelope, HashSet<Point>>.

This phase is continuation of the following tasks which was implemented during Phase 1

- Create GeoSpark SpatialRDD (PointRDD).
- Spatial Range Query: Query the PointRDD using this query window [x1(35.08),y1(-113.79),x2(32.99),y2(-109.73)].
    - Query the PointRDD
    - Build R-Tree index on PointRDD then query this PointRDD.
- Spatial KNN query: Query the PointRDD using this query point [x1(35.08),y1(-113.79)].
    - Query the PointRDD and find 5 Nearest Neighbors.
    - Build R-Tree index on PointRDD then query this PointRDD again.
- Spatial Join query: Create a GeoSpark RectangleRDD and use it to join PointRDD
    - Join the PointRDD using Equal grid without R-Tree index.
    - Join the PointRDD using Equal grid with R-Tree index.
    - Join the PointRDD using R-Tree grid without R-Tree index

Implementation:

The data is preloaded to Hadoop File System and the spark is clustered with one master node and two worker node. The method '`SpatialJoinQueryUsingCartesianProduct`' is implemented in 'JoinQuery' class in GeoSpark library with the following algorithm (Source : https://gist.github.com/aniquetahir/acb2a781a55cf76a5d2a32d4f0a4d5d6).

```
1.Create a PointRDD objectRDD;
2.Create a RectangleRDD queryWindowRDD;
3.Collect rectangles from queryWindowRDD to one Java List L;
4. For each rectangle R in L
    do RangeQuery.SpatialRangeQuery(objectRDD, queryEnvelope, 0, true);
End;
5.Collect all results; //"Collect" is a standard function under SparkContext.
6.Parallelize the results to generate a RDD in this format: JavaPairRDD<Envelope, HashSet<Point>>.;//"Parallelize" is a standard function under SparkContext.
7.Return the result RDD;
```

We used Ganglia as cluster monitoring software to identify the statistics. We get the Execution time, Memory utilization, CPU utilization from Ganglia.

Task 1: Comparing Task 2a & Task 2b of Phase 1

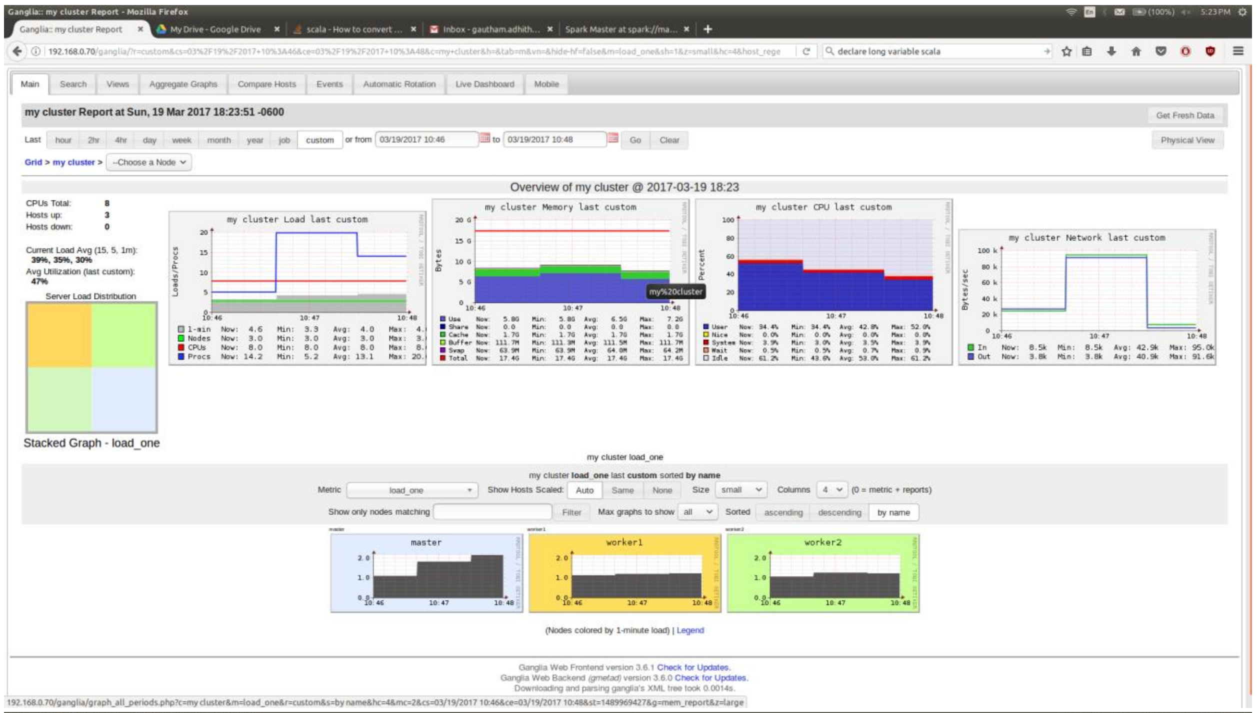
In Phase 1, Task2a, we perform Spatial Range Query using the given query window. In Task 2b, We built R-Tree index on PointRDD then query the given PointRDD again. We ran it for 10 iteration and obtained the execution time, Average CPU Utilization, Average Memory Utilization. The statistics of task 2a and task 2b is compared as shown in below table.

The **Execution time** for 2b is lesser than 2a because of R-Tree Indexing

The **Average CPU utilization** for 2b is lesser than 2a because of R-Tree Indexing

The **Average Memory Utilization** is higher for 2b is higher than 2a as Indexing consumes slightly more memory.

	Spatial Range Query (2a)	Spatial Range Query (Index) (2b)
Execution Time	5.1 Seconds	4.7 Seconds
Average CPU Usage	43.7%	30.9%
Average Memory Usage	5.7G	5.9G


Ganglia:: my cluster Report - Mozilla Firefox

192.168.0.70/ganglia/?r=custom&cs=03%2F19%2F2017+10%3A46&ce=03%2F19%2F2017+10%3A48&c=my+cluster&h=&tab=m&vn=&hide-hf=false&m=load_one&sh=1&z=small&hc=4&host_rege

my cluster Report at Sun, 19 Mar 2017 18:23:51 -0600

Last: custom, or from 03/19/2017 10:46 to 03/19/2017 10:48

Overview of my cluster @ 2017-03-19 18:23

CPUs Total: 8
Hosts up: 3
Hosts down: 0

Current Load Avg (15, 5, 1m): 39%, 35%, 30%
Avg Utilization (last custom): 47%

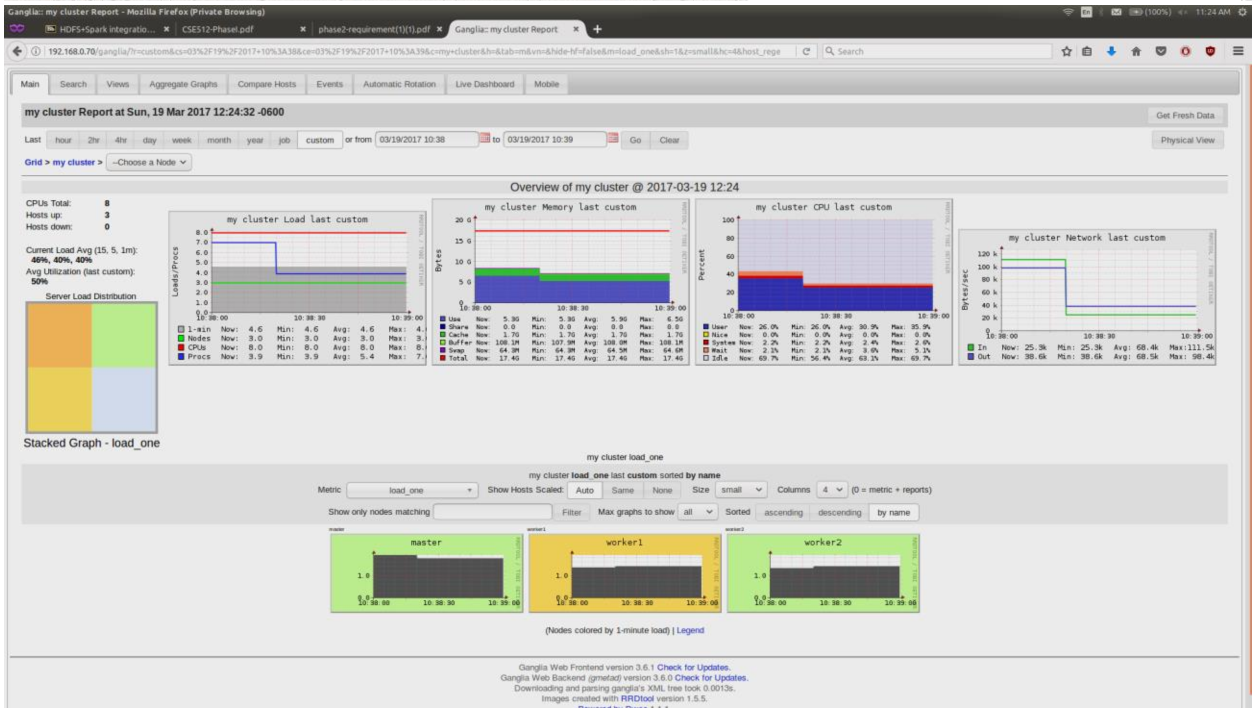
Server Load Distribution
Stacked Graph - load_one
my cluster Load last custom
my cluster Memory last custom
my cluster CPU last custom
my cluster Network last custom
my cluster load_one
master
worker1
worker2

my cluster load_one last custom sorted by name

(Nodes colored by 1-minute load) | Legend

Ganglia Web Frontend version 3.6.1 Check for Updates.
Ganglia Web Backend (gmetad) version 3.6.0 Check for Updates.
Downloading and parsing ganglia's XML tree took 0.0014s.

192.168.0.70/ganglia/graph_all_periods.php?c=my cluster&m=load_one&r=custom&s=by name&hc=4&mc=2&cs=03/19/2017 10:46&ce=03/19/2017 10:48&st=1489969427&g=mem_report&z=large


Ganglia:: my cluster Report - Mozilla Firefox (Private Browsing)

192.168.0.70/ganglia/?r=custom&cs=03%2F19%2F2017+10%3A38&ce=03%2F19%2F2017+10%3A39&c=my+cluster&h=&tab=m&vn=&hide-hf=false&m=load_one&sh=1&z=small&hc=4&host_rege

my cluster Report at Sun, 19 Mar 2017 12:24:32 -0600

Last: custom, or from 03/19/2017 10:38 to 03/19/2017 10:39

Overview of my cluster @ 2017-03-19 12:24

CPUs Total: 8
Hosts up: 3
Hosts down: 0

Current Load Avg (15, 5, 1m): 46%, 40%, 40%
Avg Utilization (last custom): 50%

Server Load Distribution
Stacked Graph - load_one
my cluster Load last custom
my cluster Memory last custom
my cluster CPU last custom
my cluster Network last custom
my cluster load_one
master
worker1
worker2

my cluster load_one last custom sorted by name

(Nodes colored by 1-minute load) | Legend

Ganglia Web Frontend version 3.6.1 Check for Updates.
Ganglia Web Backend (gmetad) version 3.6.0 Check for Updates.
Downloading and parsing ganglia's XML tree took 0.0013s.
Images created with RRDtool version 1.5.5.

Task 2: Comparing Task 3a & 3b of Phase 1

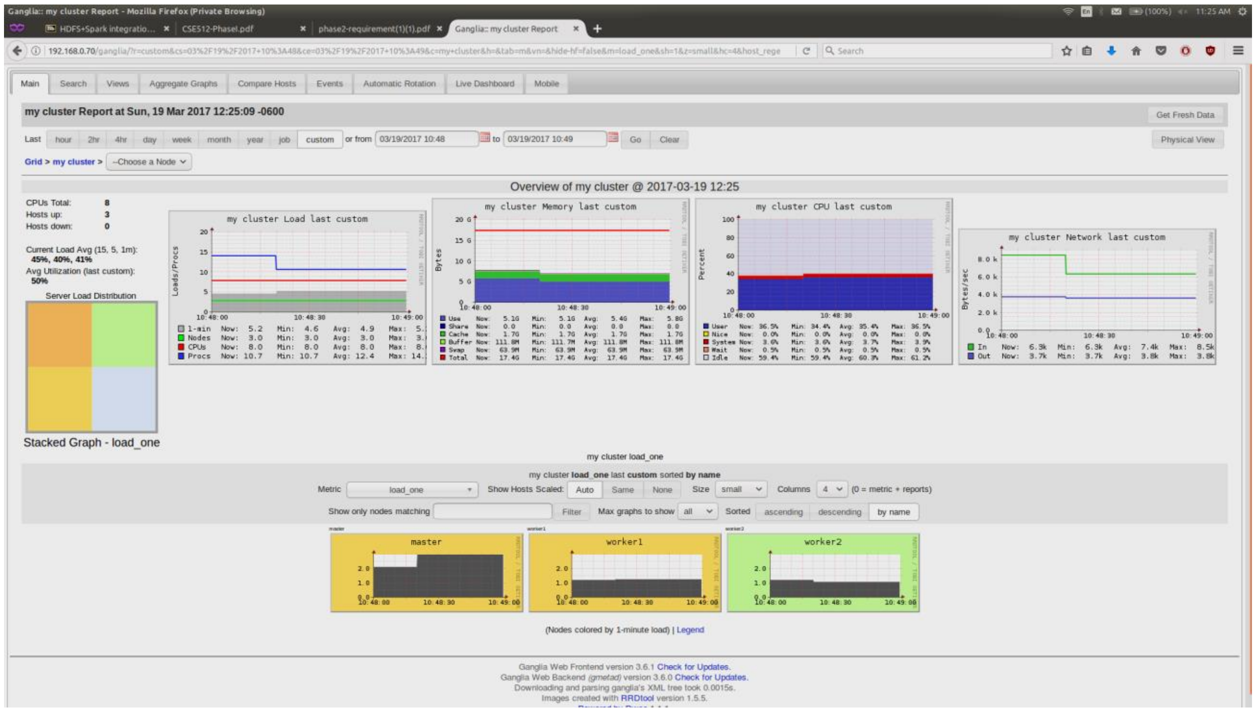
In Task 3a and Task 3b of Phase 1, we perform Spatial KNN query. With the given query point, we find 5 Nearest Neighbors in Task 3a and we built R-Tree index on PointRDD then query this PointRDD again. We ran it for 10 iteration and obtained the execution time, Average CPU Utilization, Average Memory Utilization. The statistics of task 3a and task 3b is compared as shown below. The following are the observations made.

The **Execution time** to find 5 neighbors is slightly higher than finding neighbors built with R-Tree Index.

The **CPU utilization** is slightly higher for 3a when compared to 3b because 3b uses R-Tree Index

Average Memory Utilization of 3b is slightly higher than 3a because of usage of Indexing

	Spatial KNN Query	Spatial KNN Query (Index)
Execution Time	5.9 Seconds	5.1 Seconds
Average CPU Usage	35.4%	33.6%
Average Memory Usage	5.1G	6.5G



Ganglia:: my cluster Report - Mozilla Firefox (Private Browsing)
HDFS+Spark integratio...
CSE512-Phasel.pdf
phase2-requirement(1)(1).pdf
Ganglia:: my cluster Report
192.168.0.70/ganglia/?r=custom&cs=03%2F19%2F2017+10%3A48&ce=03%2F19%2F2017+10%3A49&c=my+cluster&h=&tab=m&vn=&hide-hf=false&m=load_one&sh=1&z=small&hc=4&host_rege
Main
Search
Views
Aggregate Graphs
Compare Hosts
Events
Automatic Rotation
Live Dashboard
Mobile
Get Fresh Data
Physical View
Last
hour
2hr
4hr
day
week
month
year
job
custom
Go
Clear
Grid > my cluster >
--Choose a Node
Server Load Distribution
Stacked Graph - load_one
my cluster Load last custom
my cluster Memory last custom
my cluster CPU last custom
my cluster Network last custom
Loads/Procs
Bytes
Percent
Bytes/sec
master
worker1
worker2
Filter
ascending
descending
by name

my cluster Report at Sun, 19 Mar 2017 12:25:09 -0600

or from 03/19/2017 10:48 to 03/19/2017 10:49

Overview of my cluster @ 2017-03-19 12:25

CPUs Total: 8
Hosts up: 3
Hosts down: 0

Current Load Avg (15, 5, 1m):
45%, 40%, 41%
Avg Utilization (last custom):
50%

```
1-min   Now:  5.2   Min:  4.6   Avg:  4.9   Max:  5.
Nodes   Now:  3.0   Min:  3.0   Avg:  3.0   Max:  3.
CPUs    Now:  8.0   Min:  8.0   Avg:  8.0   Max:  8.
Procs   Now: 10.7   Min: 10.7   Avg: 12.4   Max: 14.
```

```
Use     Now:   5.1G   Min:   5.1G   Avg:   5.4G   Max:   5.8G
Share   Now:   0.0    Min:   0.0    Avg:   0.0    Max:   0.0
Cache   Now:   1.7G   Min:   1.7G   Avg:   1.7G   Max:   1.7G
Buffer  Now: 111.8M   Min: 111.7M   Avg: 111.8M   Max: 111.8M
Swap    Now:  63.9M   Min:  63.9M   Avg:  63.9M   Max:  63.9M
Total   Now:  17.4G   Min:  17.4G   Avg:  17.4G   Max:  17.4G
```

```
User    Now: 36.5%   Min: 34.4%   Avg: 35.4%   Max: 36.5%
Nice    Now:  0.0%   Min:  0.0%   Avg:  0.0%   Max:  0.0%
System  Now:  3.6%   Min:  3.6%   Avg:  3.7%   Max:  3.9%
Wait    Now:  0.5%   Min:  0.5%   Avg:  0.5%   Max:  0.5%
Idle    Now: 59.4%   Min: 59.4%   Avg: 60.3%   Max: 61.2%
```

```
In    Now:  6.3k   Min:  6.3k   Avg:  7.4k   Max:  8.5k
Out   Now:  3.7k   Min:  3.7k   Avg:  3.8k   Max:  3.8k
```

my cluster load_one

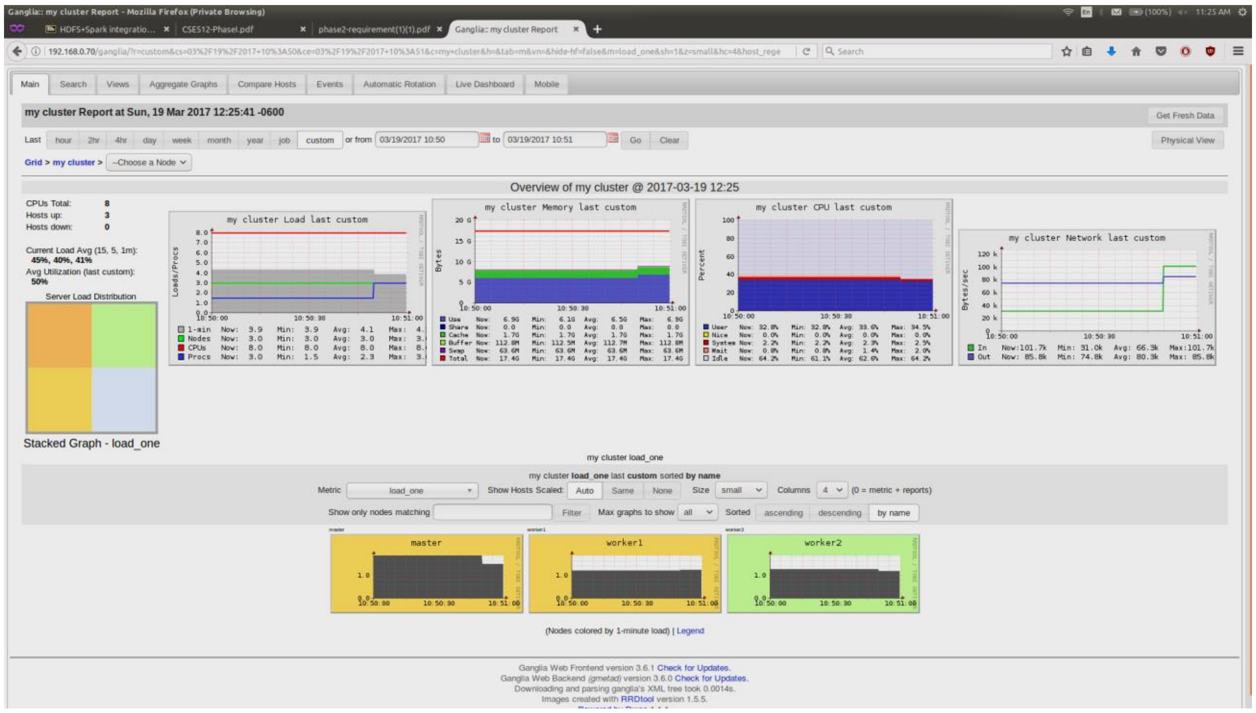
my cluster **load_one** last **custom** sorted **by name**

Metric: load_one Show Hosts Scaled: Auto Size: small Columns: 4 (0 = metric + reports)

Show only nodes matching Max graphs to show: all Sorted: by name

(Nodes colored by 1-minute load) | Legend

Ganglia Web Frontend version 3.6.1 Check for Updates.
Ganglia Web Backend *(gmetad)* version 3.6.0 Check for Updates.
Downloading and parsing ganglia's XML tree took 0.0015s.
Images created with RRDtool version 1.5.5.



my cluster Report at Sun, 19 Mar 2017 12:25:41 -0600

Last: custom, or from 03/19/2017 10:50 to 03/19/2017 10:51

Overview of my cluster @ 2017-03-19 12:25

CPUs Total: 8
Hosts up: 3
Hosts down: 0

Current Load Avg (15, 5, 1m): **45%, 40%, 41%**
Avg Utilization (last custom): **50%**

```
my cluster Load last custom
1-min   Now: 3.9  Min: 3.9  Avg: 4.1  Max: 4.
Nodes   Now: 3.0  Min: 3.0  Avg: 3.0  Max: 3.
CPUs    Now: 8.0  Min: 8.0  Avg: 8.0  Max: 8.
Procs   Now: 3.0  Min: 1.5  Avg: 2.3  Max: 3.

my cluster Memory last custom
Use     Now: 6.9G    Min: 6.1G    Avg: 6.5G    Max: 6.9G
Share   Now: 0.0     Min: 0.0     Avg: 0.0     Max: 0.0
Cache   Now: 1.7G    Min: 1.7G    Avg: 1.7G    Max: 1.7G
Buffer  Now: 112.8M  Min: 112.5M  Avg: 112.7M  Max: 112.8M
Swap    Now: 63.6M   Min: 63.6M   Avg: 63.6M   Max: 63.6M
Total   Now: 17.4G   Min: 17.4G   Avg: 17.4G   Max: 17.4G

my cluster CPU last custom
User    Now: 32.8%  Min: 32.8%  Avg: 33.6%  Max: 34.5%
Nice    Now:  0.0%  Min:  0.0%  Avg:  0.0%  Max:  0.0%
System  Now:  2.2%  Min:  2.2%  Avg:  2.3%  Max:  2.5%
Wait    Now:  0.8%  Min:  0.8%  Avg:  1.4%  Max:  2.0%
Idle    Now: 64.2%  Min: 61.1%  Avg: 62.6%  Max: 64.2%

my cluster Network last custom
In   Now:101.7k  Min: 31.0k  Avg: 66.3k  Max:101.7k
Out  Now: 85.8k  Min: 74.8k  Avg: 80.3k  Max: 85.8k
```

Server Load Distribution

Stacked Graph - load_one

my cluster load_one

my cluster **load_one** last **custom** sorted **by name**

Metric: load_one; Show Hosts Scaled: Auto; Size: small; Columns: 4 (0 = metric + reports); Max graphs to show: all; Sorted: by name

(Nodes colored by 1-minute load) | Legend

Ganglia Web Frontend version 3.6.1 Check for Updates.
Ganglia Web Backend (gmetad) version 3.6.0 Check for Updates.
Downloading and parsing ganglia's XML tree took 0.0014s.
Images created with RRDtool version 1.5.5.

Task 3: Comparison of Task 4(a) & (b) of Phase 1

In the first part of this task 4, we compared the Execution times, Average memory, Average CPU utilization for spatial join query implementation of PointRDD using Equal grid without R-Tree index and PointRDD using Equal grid with R-Tree index. We ran the program for 10 iteration and suitable data is obtained.

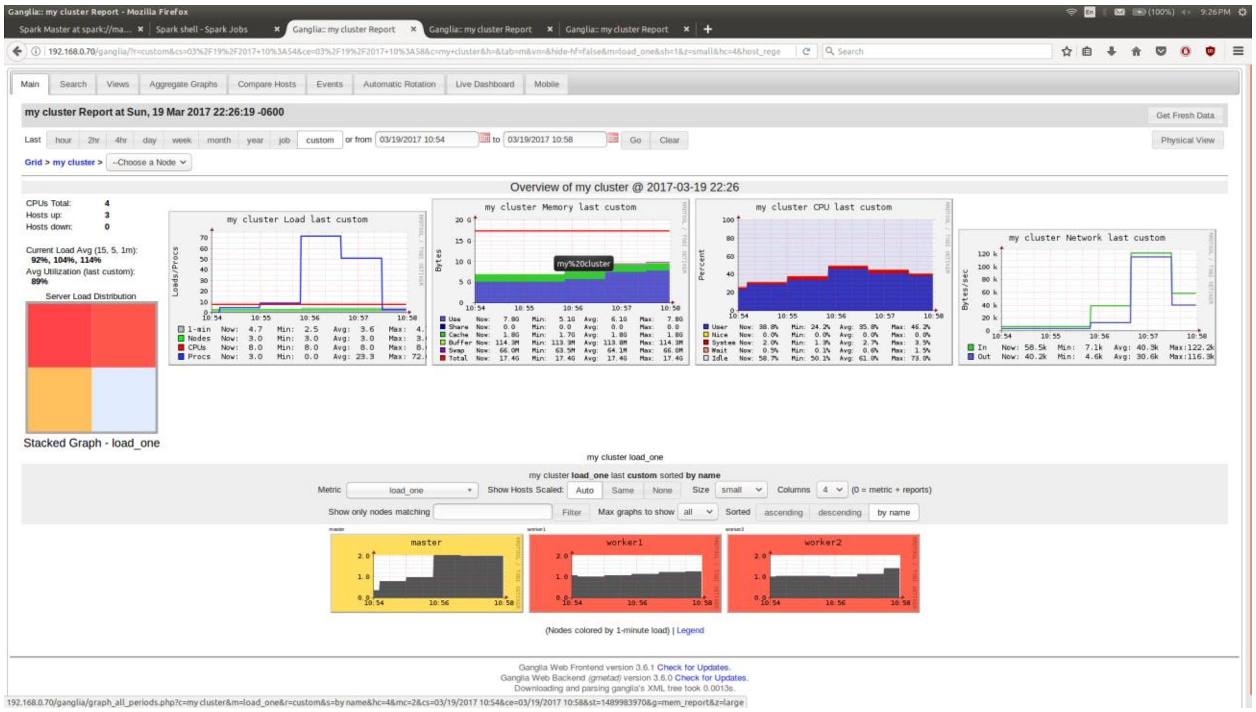
We found that **Execution time** of Spatial join, pointRDD having equal grid without R-Tree index was higher than Spatial join, pointRDD having equal grid with R-Tree index.

The **Average memory utilization** increases slightly when used with Index. This is because of R-Tree indexing.

The **Average CPU utilization** of 4a is little higher than 4b because of the usage of R-Tree Index.

The difference can be seen in the table below,

	Spatial Join Query (Equal Grid) (4a)	Spatial Join Query (Equal Grid with Index) (4b)
Average Runtime	313 Seconds	227 Seconds
Average CPU Usage	36.8%	35.8%
Average Memory Usage	6.1G	7.3G



Ganglia:: my cluster Report - Mozilla Firefox

192.168.0.70/ganglia/?r=custom&cs=03%2F19%2F2017+10%3A54&ce=03%2F19%2F2017+10%3A58&c=my+cluster&h=&tab=m&vn=&hide-hf=false&m=load_one&sh=1&z=small&hc=4&host_rege

my cluster Report at Sun, 19 Mar 2017 22:26:19 -0600

Last: custom, or from 03/19/2017 10:54 to 03/19/2017 10:58

Grid > my cluster > --Choose a Node

Overview of my cluster @ 2017-03-19 22:26

CPUs Total: 4
Hosts up: 3
Hosts down: 0

Current Load Avg (15, 5, 1m):
92%, 104%, 114%
Avg Utilization (last custom):
89%

Server Load Distribution

Stacked Graph - load_one

```
my cluster Load last custom
1-min   Now: 4.7   Min: 2.5   Avg: 3.6   Max: 4.
Nodes   Now: 3.0   Min: 3.0   Avg: 3.0   Max: 3.
CPUs    Now: 8.0   Min: 8.0   Avg: 8.0   Max: 8.
Procs   Now: 3.0   Min: 0.0   Avg: 23.3  Max: 72.

my cluster Memory last custom
Use     Now:   7.8G  Min:   5.1G  Avg:   6.1G  Max:   7.8G
Share   Now:   0.0   Min:   0.0   Avg:   0.0   Max:   0.0
Cache   Now:   1.8G  Min:   1.7G  Avg:   1.8G  Max:   1.8G
Buffer  Now: 114.3M  Min: 113.3M  Avg: 113.8M  Max: 114.3M
Swap    Now:  66.0M  Min:  63.5M  Avg:  64.1M  Max:  66.0M
Total   Now:  17.4G  Min:  17.4G  Avg:  17.4G  Max:  17.4G

my cluster CPU last custom
User    Now: 38.8%  Min: 24.2%  Avg: 35.8%  Max: 46.2%
Nice    Now:  0.0%  Min:  0.0%  Avg:  0.0%  Max:  0.0%
System  Now:  2.0%  Min:  1.3%  Avg:  2.7%  Max:  3.5%
Wait    Now:  0.5%  Min:  0.1%  Avg:  0.6%  Max:  1.5%
Idle    Now: 58.7%  Min: 50.1%  Avg: 61.0%  Max: 73.0%

my cluster Network last custom
In   Now: 58.5k  Min: 7.1k  Avg: 40.3k  Max:122.2k
Out  Now: 40.2k  Min: 4.6k  Avg: 30.6k  Max:116.3k
```

my cluster load_one

my cluster **load_one** last **custom** sorted **by name**

Metric: load_one; Show Hosts Scaled: Auto; Size: small; Columns: 4 (0 = metric + reports)

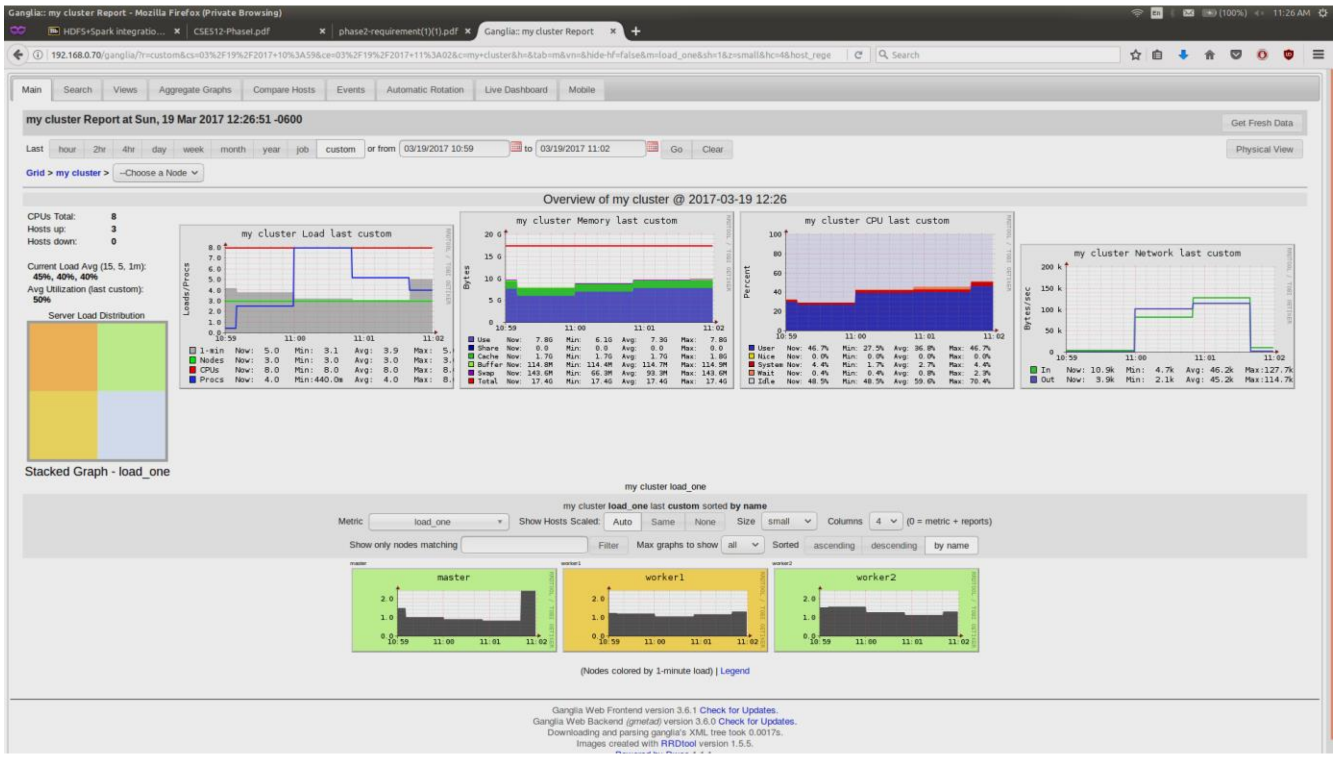
Show only nodes matching; Max graphs to show: all; Sorted: by name

master, worker1, worker2

(Nodes colored by 1-minute load) | Legend

Ganglia Web Frontend version 3.6.1 Check for Updates.
Ganglia Web Backend *(gmetad)* version 3.6.0 Check for Updates.
Downloading and parsing ganglia's XML tree took 0.0013s.

192.168.0.70/ganglia/graph_all_periods.php?c=my cluster&m=load_one&r=custom&s=by name&hc=4&mc=2&cs=03/19/2017 10:54&ce=03/19/2017 10:58&st=1489983970&g=mem_report&z=large



Ganglia:: my cluster Report - Mozilla Firefox (Private Browsing)

192.168.0.70/ganglia/?r=custom&cs=03%2F19%2F2017+10%3A59&ce=03%2F19%2F2017+11%3A02&c=my+cluster&h=&tab=m&vn=&hide-hf=false&m=load_one&sh=1&z=small&hc=4&host_rege

my cluster Report at Sun, 19 Mar 2017 12:26:51 -0600

Last: custom, or from 03/19/2017 10:59 to 03/19/2017 11:02

Grid > my cluster > --Choose a Node

Overview of my cluster @ 2017-03-19 12:26

CPUs Total: 8
Hosts up: 3
Hosts down: 0

Current Load Avg (15, 5, 1m):
45%, 40%, 40%
Avg Utilization (last custom):
50%

Server Load Distribution

Stacked Graph - load_one

```
my cluster Load last custom
1-min   Now: 5.0   Min: 3.1     Avg: 3.9   Max: 5.
Nodes   Now: 3.0   Min: 3.0     Avg: 3.0   Max: 3.
CPUs    Now: 8.0   Min: 8.0     Avg: 8.0   Max: 8.
Procs   Now: 4.0   Min:440.0m   Avg: 4.0   Max: 8.

my cluster Memory last custom
Use     Now:   7.8G  Min:   6.1G  Avg:   7.3G  Max:   7.8G
Share   Now:   0.0   Min:   0.0   Avg:   0.0   Max:   0.0
Cache   Now:   1.7G  Min:   1.7G  Avg:   1.7G  Max:   1.8G
Buffer  Now: 114.8M  Min: 114.4M  Avg: 114.7M  Max: 114.9M
Swap    Now: 143.6M  Min:  66.3M  Avg:  93.3M  Max: 143.6M
Total   Now:  17.4G  Min:  17.4G  Avg:  17.4G  Max:  17.4G

my cluster CPU last custom
User    Now: 46.7%  Min: 27.5%  Avg: 36.8%  Max: 46.7%
Nice    Now:  0.0%  Min:  0.0%  Avg:  0.0%  Max:  0.0%
System  Now:  4.4%  Min:  1.7%  Avg:  2.7%  Max:  4.4%
Wait    Now:  0.4%  Min:  0.4%  Avg:  0.8%  Max:  2.3%
Idle    Now: 48.5%  Min: 48.5%  Avg: 59.6%  Max: 70.4%

my cluster Network last custom
In   Now: 10.9k  Min: 4.7k  Avg: 46.2k  Max:127.7k
Out  Now:  3.9k  Min: 2.1k  Avg: 45.2k  Max:114.7k
```

my cluster load_one

my cluster **load_one** last **custom** sorted **by name**

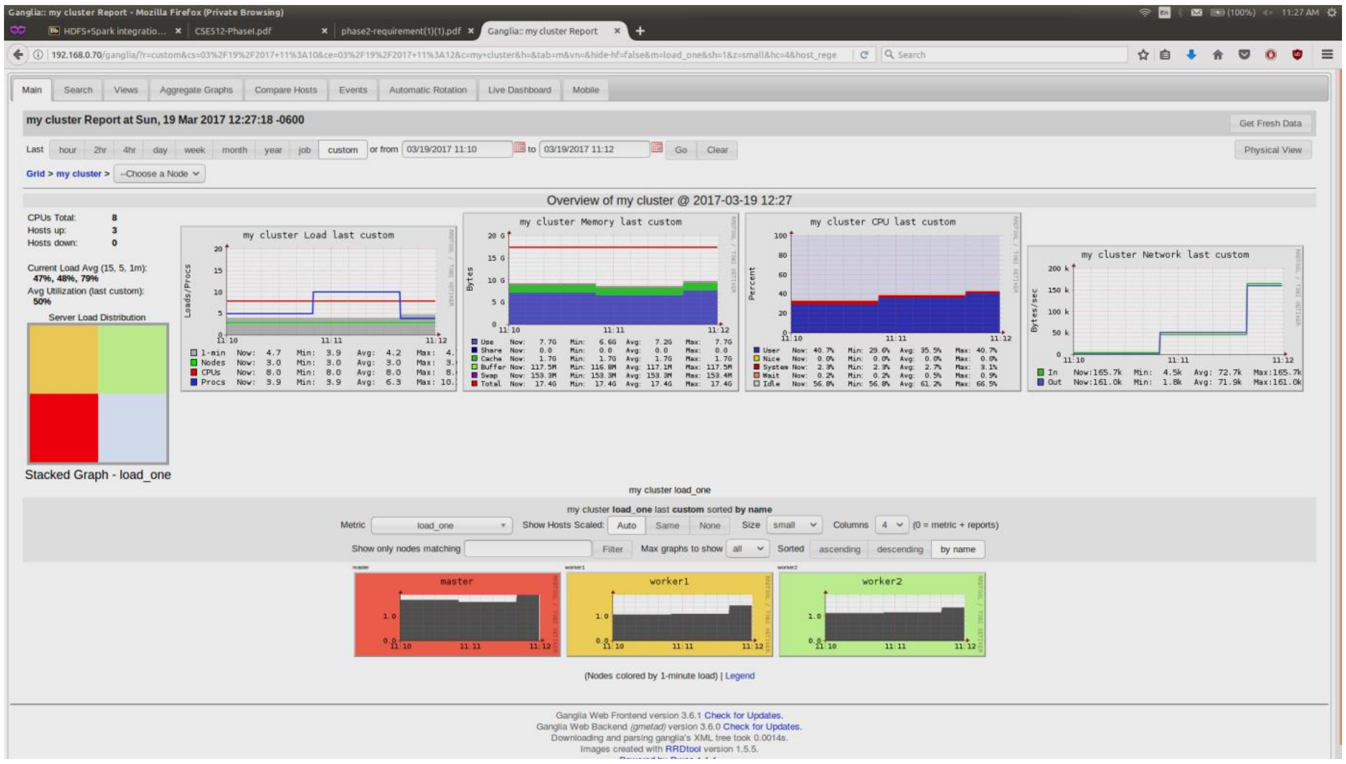
Metric: load_one; Show Hosts Scaled: Auto; Size: small; Columns: 4 (0 = metric + reports)

Show only nodes matching; Max graphs to show: all; Sorted: by name

master, worker1, worker2

(Nodes colored by 1-minute load) | Legend

Ganglia Web Frontend version 3.6.1 Check for Updates.
Ganglia Web Backend *(gmetad)* version 3.6.0 Check for Updates.
Downloading and parsing ganglia's XML tree took 0.0017s.
Images created with RRDtool version 1.5.5.



my cluster Report at Sun, 19 Mar 2017 12:27:18 -0600

Last: custom, from 03/19/2017 11:10 to 03/19/2017 11:12

Grid > my cluster > --Choose a Node

Overview of my cluster @ 2017-03-19 12:27

CPUs Total: 8
Hosts up: 3
Hosts down: 0

Current Load Avg (15, 5, 1m): **47%, 48%, 79%**
Avg Utilization (last custom): **50%**

```
my cluster Load last custom
1-min   Now:  4.7   Min:  3.9   Avg:  4.2   Max:  4.
Nodes   Now:  3.0   Min:  3.0   Avg:  3.0   Max:  3.
CPUs    Now:  8.0   Min:  8.0   Avg:  8.0   Max:  8.
Procs   Now:  3.9   Min:  3.9   Avg:  6.3   Max: 10.

my cluster Memory last custom
Use    Now:   7.7G  Min:   6.6G  Avg:   7.2G  Max:   7.7G
Share  Now:   0.0   Min:   0.0   Avg:   0.0   Max:   0.0
Cache  Now:   1.7G  Min:   1.7G  Avg:   1.7G  Max:   1.7G
Buffer Now: 117.5M  Min: 116.8M  Avg: 117.1M  Max: 117.5M
Swap   Now: 153.3M  Min: 153.3M  Avg: 153.3M  Max: 153.4M
Total  Now:  17.4G  Min:  17.4G  Avg:  17.4G  Max:  17.4G

my cluster CPU last custom
User   Now: 40.7%  Min: 29.6%  Avg: 35.5%  Max: 40.7%
Nice   Now:  0.0%  Min:  0.0%  Avg:  0.0%  Max:  0.0%
System Now:  2.3%  Min:  2.3%  Avg:  2.7%  Max:  3.1%
Wait   Now:  0.2%  Min:  0.2%  Avg:  0.5%  Max:  0.9%
Idle   Now: 56.8%  Min: 56.8%  Avg: 61.2%  Max: 66.5%

my cluster Network last custom
In   Now:165.7k  Min:  4.5k  Avg: 72.7k  Max:165.7k
Out  Now:161.0k  Min:  1.8k  Avg: 71.9k  Max:161.0k
```

Stacked Graph - load_one

my cluster load_one

my cluster **load_one** last **custom** sorted **by name**

Metric: load_one Show Hosts Scaled: Auto Size: small Columns: 4 (0 = metric + reports)
Show only nodes matching Max graphs to show: all Sorted: by name

(Nodes colored by 1-minute load) | Legend

Ganglia Web Frontend version 3.6.1 Check for Updates.
Ganglia Web Backend *(gmetad)* version 3.6.0 Check for Updates.
Downloading and parsing ganglia's XML tree took 0.0014s.
Images created with RRDtool version 1.5.5.

Task 3: Comparing Task 4 (a) & (c) of Phase 1

In the second part of this task 3, we compared the Execution times, Average memory, Average CPU utilization for spatial join query implementation of PointRDD using Equal grid without R-Tree index. and PointRDD using R-Tree grid without R-Tree index. We ran for 10 iteration and suitable statistics is obtained.

We found that **Execution time** of 4a is much higher than the execution time of 4c, since 4a uses Equal grid and 4c uses R-Tree grid

The **Average memory utilization** increases slightly from 4a to 4c indicating that the R-Tree grid utilizes slightly higher memory compared to Equal grid.

The **Average CPU utilization** of 4c is lesser when compared to 4a and 4b. The difference is because of the R-Tree grid usage whereas in 4a Equal grid is used.

The difference is seen in the table

	Spatial Join Query (Equal Grid) (4a)	Spatial Join Query (RTree Grid with Index) (4c)
Average Runtime	313 Seconds	243 Seconds
Average CPU Usage	36.8%	35.5%
Average Memory Usage	6.1G	7.2G

2. **Comparison of Phase1 Task4 (a) (b) (c) with Phase 2 Task A:**

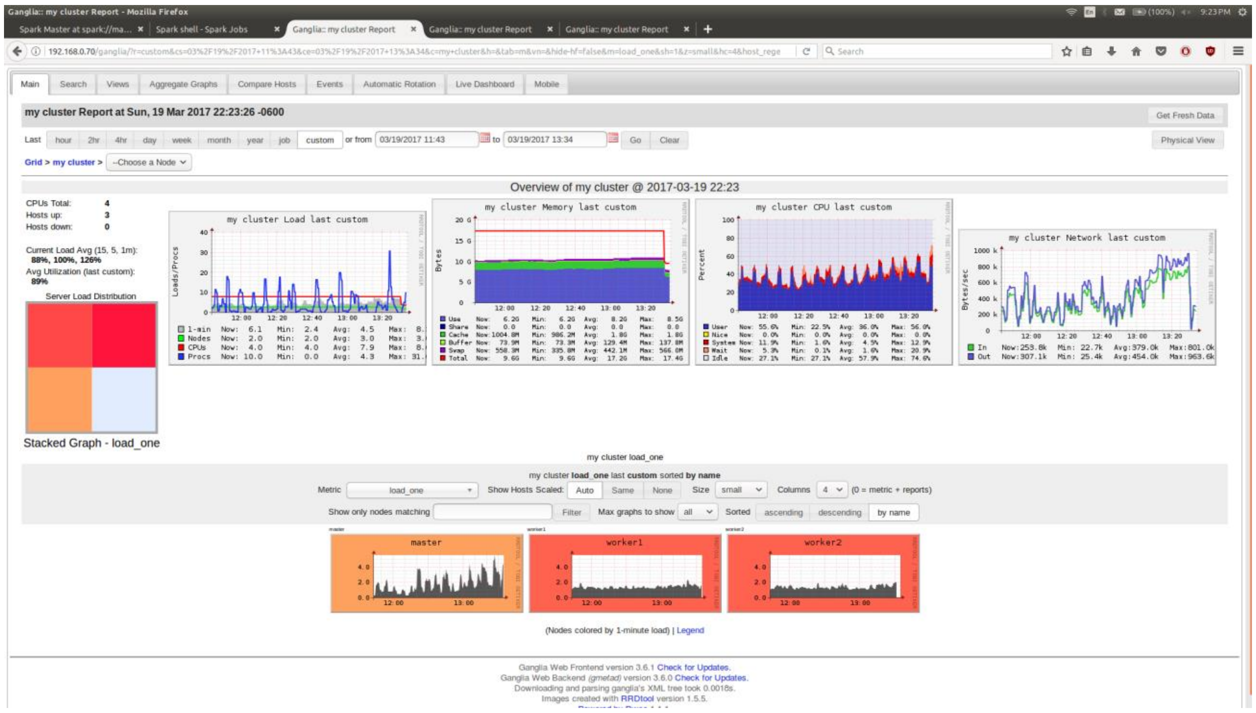
The task 4 of the Phase 1 which uses the normal join from existing GeoSpark Library and task 1 of the Phase 2, where we implemented a Naïve Spatial Join Query using Cartesian Product algorithm. This means that it is running using all the queryEnvelope (RectangleRDD) which takes huge amount of time. The Execution time is much higher than the Spatial Join Query executed in 4 (a) (b) (c)

The **Average memory utilization** of task1 of Phase 2 is much higher than any of the 4 (a) (b) (c) because of the Cartesian product algorithm.

The **Average CPU Utilization** is similar any of the 4 (a) (b) (c) but in phase2Task1 the *Maximum CPU usage goes up to 59.2%* whereas in 4 (a) (b) (c) it was equal to 40%. This is because of the Cartesian product algorithm.

	Spatial Join Query (Equal Grid without Index)4 (a)	Spatial Join Query (Index) 4 (b)	Spatial Join Query (R-Tree Grid without Index) 4 (c)	Spatial Join Query (Cartesian) (Phase 2)
Average Runtime	313 Seconds	227 Seconds	243 Seconds	6450 Seconds
Average CPU Usage	36.8%	35.8%	35.5%	36%
Average Memory Usage	6.1G	7.3G	7.2G	8.2G

Phase 2 Task 1, Ganglia Statistics:


Ganglia:: my cluster Report - Mozilla Firefox
Spark Master at spark://ma...
Spark shell - Spark Jobs
Ganglia:: my cluster Report
192.168.0.70/ganglia/?r=custom&cs=03%2F19%2F2017+11%3A43&ce=03%2F19%2F2017+13%3A34&c=my+cluster&h=&tab=m&vn=&hide-hf=false&m=load_one&sh=1&z=small&hc=4&host_rege
Search
Main
Search
Views
Aggregate Graphs
Compare Hosts
Events
Automatic Rotation
Live Dashboard
Mobile
Get Fresh Data
Physical View
Last hour 2hr 4hr day week month year job custom
Go
Clear
Grid > my cluster > --Choose a Node
Stacked Graph - load_one
Server Load Distribution

my cluster Report at Sun, 19 Mar 2017 22:23:26 -0600

Last: custom, or from 03/19/2017 11:43 to 03/19/2017 13:34

Overview of my cluster @ 2017-03-19 22:23

CPUs Total:	4
Hosts up:	3
Hosts down:	0

Current Load Avg (15, 5, 1m):
88%, 100%, 126%
Avg Utilization (last custom):
89%

```
my cluster Load last custom
1-min   Now:  6.1   Min:  2.4   Avg:  4.5   Max:  8.
Nodes   Now:  2.0   Min:  2.0   Avg:  3.0   Max:  3.
CPUs    Now:  4.0   Min:  4.0   Avg:  7.9   Max:  8.
Procs   Now: 10.0   Min:  0.0   Avg:  4.3   Max: 31.
```

```
my cluster Memory last custom
Use     Now:    6.2G   Min:   6.2G   Avg:   8.2G   Max:   8.5G
Share   Now:    0.0    Min:   0.0    Avg:   0.0    Max:   0.0
Cache   Now: 1004.8M   Min: 986.2M   Avg:   1.8G   Max:   1.8G
Buffer  Now:   73.9M   Min:  73.3M   Avg: 129.4M   Max: 137.8M
Swap    Now:  558.3M   Min: 335.8M   Avg: 442.1M   Max: 566.0M
Total   Now:    9.6G   Min:   9.6G   Avg:  17.2G   Max:  17.4G
```

```
my cluster CPU last custom
User    Now: 55.6%   Min: 22.5%   Avg: 36.0%   Max: 56.0%
Nice    Now:  0.0%   Min:  0.0%   Avg:  0.0%   Max:  0.0%
System  Now: 11.9%   Min:  1.6%   Avg:  4.5%   Max: 12.9%
Wait    Now:  5.3%   Min:  0.1%   Avg:  1.6%   Max: 20.9%
Idle    Now: 27.1%   Min: 27.1%   Avg: 57.9%   Max: 74.6%
```

```
my cluster Network last custom
In    Now:253.8k   Min: 22.7k   Avg:379.0k   Max:801.0k
Out   Now:307.1k   Min: 25.4k   Avg:454.0k   Max:963.6k
```

my cluster load_one

my cluster **load_one** last **custom** sorted **by name**

Metric: load_one | Show Hosts Scaled: Auto | Size: small | Columns: 4 (0 = metric + reports)

Show only nodes matching | Max graphs to show: all | Sorted: by name

master | worker1 | worker2

(Nodes colored by 1-minute load) | Legend

Ganglia Web Frontend version 3.6.1 Check for Updates.
Ganglia Web Backend *(gmetad)* version 3.6.0 Check for Updates.
Downloading and parsing ganglia's XML tree took 0.0018s.
Images created with RRDtool version 1.5.5.