Monitoring Slurm with a Splunk App

LANL Workload Management Team

Nicole Dobson

18 Sept 2019
The need for a better monitoring tool…

- Maintain functionality
- Multiple systems
- Response time

- Portable and easy for multiple systems
- Quick and easy detection
- Faster diagnosis
Using the command line

Say we just want to monitor utilization:

```bash
-srep-report -t percent cluster util start=00:00:00

Cluster Utilization 2019-09-11T00:00:00 - 2019-09-11T00:59:59
Usage reported in Percentage of Total

<table>
<thead>
<tr>
<th>Cluster Allocate</th>
<th>Down</th>
<th>PLND</th>
<th>Dow</th>
<th>Idle</th>
<th>Reserved</th>
<th>Reported</th>
</tr>
</thead>
<tbody>
<tr>
<td>badger</td>
<td>0.62%</td>
<td>0.00%</td>
<td>3.94%</td>
<td>0.47%</td>
<td>100.00%</td>
<td></td>
</tr>
</tbody>
</table>

bash-4.2$ sinfo --partition any
PARTITION AVAIL TIMELIMIT NODES STATE NODELIST
any up infinite 4 drain* ba[373,429,607,647]
any up infinite 1 drain ba374
any up infinite 7 resv ba[003-006,053,104,613]
any up infinite 642 alloc ba[001,007-052,054-103,105-154,156-335,48-660]
any up infinite 6 idle ba[002,155,336,407-409]
```
Using Splunk

We can expand on that single number to a visual representation that updates in real time.

Now we can easily identify unexpected utilization metrics without having to constantly run a command and verify that the numbers it returns are within normal bounds.
Why we use Splunk

What Splunk does:

• Ingests log messages and other log-like data
• Command box functions similarly to grep
• Allows visuals to be created and auto-updated all from that one command box
• Can create alerts on data and trends

Benefits:

• No need to scp data off of clusters to monitoring server in order to interpret it
• Splunk command box allows for grepping through logs same as command line
• One command to ingest data and create visual, not having to keep track of and maintain multiple scripts
• No need to set up cron job to look for and alert on conditions of interest, can alert on trends
All Clusters Monitoring Dashboard
Single Cluster Dashboard
### Top Non-Zero Exit Code Jobs

<table>
<thead>
<tr>
<th>cluster</th>
<th>exit_code</th>
<th>end_state</th>
<th>user_name</th>
<th>jobname</th>
<th>wallclock_limit (hrs)</th>
<th>approx_duration (hrs)</th>
<th>perc_wall_clock_used</th>
<th>NodeCount</th>
<th>count</th>
<th>severity</th>
</tr>
</thead>
<tbody>
<tr>
<td>grizzly</td>
<td>9:0</td>
<td>FAILED</td>
<td>user_1</td>
<td>job_name_A</td>
<td>16</td>
<td>0.0</td>
<td>0.0%</td>
<td>256</td>
<td>2</td>
<td>512</td>
</tr>
<tr>
<td>grizzly</td>
<td>9:0</td>
<td>FAILED</td>
<td>user_2</td>
<td>job_name_B</td>
<td>1.0</td>
<td>0.0</td>
<td>0.0%</td>
<td>500</td>
<td>1</td>
<td>500</td>
</tr>
<tr>
<td>trinitite-knl</td>
<td>0:15</td>
<td>FAILED</td>
<td>user_3</td>
<td>job_name_C</td>
<td>1.0</td>
<td>0.5</td>
<td>50%</td>
<td>90</td>
<td>3</td>
<td>270</td>
</tr>
<tr>
<td>grizzly</td>
<td>6:0</td>
<td>FAILED</td>
<td>user_1</td>
<td>job_name_A</td>
<td>16</td>
<td>3</td>
<td>19%</td>
<td>128</td>
<td>2</td>
<td>256</td>
</tr>
<tr>
<td>snow</td>
<td>1:0</td>
<td>FAILED</td>
<td>user_4</td>
<td>job_name_D</td>
<td>2.0</td>
<td>1</td>
<td>50%</td>
<td>2</td>
<td>121</td>
<td>242</td>
</tr>
<tr>
<td>grizzly</td>
<td>9:0</td>
<td>FAILED</td>
<td>user_5</td>
<td>job_name_E</td>
<td>7.0</td>
<td>2</td>
<td>25%</td>
<td>114</td>
<td>2</td>
<td>228</td>
</tr>
<tr>
<td>grizzly</td>
<td>9:0</td>
<td>FAILED</td>
<td>user_6</td>
<td>job_name_F</td>
<td>16</td>
<td>1</td>
<td>6.3%</td>
<td>50</td>
<td>4</td>
<td>200</td>
</tr>
<tr>
<td>grizzly</td>
<td>9:0</td>
<td>FAILED</td>
<td>user_6</td>
<td>job_name_F</td>
<td>16</td>
<td>3</td>
<td>19%</td>
<td>50</td>
<td>4</td>
<td>200</td>
</tr>
<tr>
<td>grizzly</td>
<td>59:0</td>
<td>FAILED</td>
<td>user_7</td>
<td>job_name_G</td>
<td>3.0</td>
<td>3</td>
<td>100%</td>
<td>100</td>
<td>2</td>
<td>200</td>
</tr>
</tbody>
</table>

### Average Wall Clock Usage Percentage by User

![Average Wall Clock Usage Percentage by User](image)

### Number of Jobs Submitted Trend on badger by Hour and Day of Week (last 3 months)

![Number of Jobs Submitted Trend on badger by Hour and Day of Week](image)
Log Messages and Data Sources

• slurmd log messages
  • Reservation start and end
  • slurmd running

• Custom made cron script
  • Slurm commands: sinfo, sdiag, squeue, scontrol …
  • Easy to maintain and add to, same across all clusters

• Job completion data
  • Epilog script ran at end of job reporting on data items

• Some supporting logs from other systems or software
Improved our maintenance procedures
Fine-tuned our policies
Quickly get a sense of health, normal pattern of use, and appropriate heartbeats
Questions?

Over 70 years at the forefront of supercomputing
Over 70 years at the forefront of supercomputing