Integrating Job and Cloud Health Information of Apache Aurora with Apache Airavata

Abstract
This project will incorporate Apache Aurora to enable Airavata to launch jobs on large cloud environments, and collect the related information on the health of each job and the cloud resources. The project will also analyze the current micro-services architecture of Airavata and develop code for an updated architecture for modules such as Logging. As a result, another outcome of this project would be the development of a module that will collect all the logging information from the various execution points in an Airavata job's life cycle and provide search and mining capability.

Introduction

Apache Aurora is a service scheduler, that runs on top of Apache Mesos. Aurora is a Mesos framework that schedules jobs on Mesos. This combination enables the use of long-running services that take advantage of Apache Mesos scalability, fault-tolerance, and resource isolation. Apache Mesos is a cluster manager, which provides information about the state of the cluster. Aurora uses that knowledge to make scheduling decisions. For example, when a machine experiences failure Aurora automatically reschedules those previously-running services onto a healthy machine in order to keep them running. Each job is tracked by Aurora to be in one of the following states: pending, assigned, starting, running, and finished.

Apache Aurora requires a configuration file ".aurora" to launch jobs. Following is an example of Aurora configuration file:
import os
hello_world_process = Process(name = 'hello_world', cmdline = 'echo hello world')

hello_world_task = Task(
    resources = Resources(cpu = 0.1, ram = 16 * MB, disk = 16 * MB),
    processes = [hello_world_process])

hello_world_job = Job(
    cluster = 'cluster1',
    role = os.getenv('USER'),
    task = hello_world_task)

jobs = [hello_world_job]

To launch the job with the above configuration we use:

aurora job create cluster1/$USER/test/hello_world hello_world.aurora

This project will develop modules in Airavata to automatically generate the Aurora configuration file to launch a job on an Aurora-managed cluster in a cloud environment. The Aurora user interface, as shown in the web portal, provides detailed information on the job status, job name, start and finish times, the location of the logs, and resource usage. This project will use add a module to Apache Airavata to pull this detailed information using the Aurora HTTP API.

Goals

- This project will investigate how apache Aurora collects information of cluster environment for display on the Aurora web interface. We will study the Aurora HTTP API and retrieve all the information related to the target infrastructure and job health, and make it available to the Airavata job submission module.
• We will address the scenarios wherein the resource requirements of jobs need to be optimally matched with the available resources (memory, cores, and disk) in a Cloud environment, such as Jetstream.
• We will allow a policy-driven approach to determine how jobs should be submitted to the Aurora queue.
• We will process the retrieved information from Aurora and convert the information to a format that can be consumed by the Orchestrator and GFac for further job submissions.
• We will use the appropriate design patterns to integrate the use of Aurora as one of the options for Big Data and Cloud resource frameworks with the Airavata framework.
• We will make the resource information from Aurora available in a database so that the Orchestrator and GFac can use it to determine what resources to request for subsequent jobs and what order to place them in the Aurora queue. This will require updates to the Orchestrator and GFac.

Timeline

<table>
<thead>
<tr>
<th>Timeline</th>
<th>Deliverable(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>May 1st - May 2nd: Community Bonding Period</td>
<td></td>
</tr>
<tr>
<td>May 25th</td>
<td>Start project Development</td>
</tr>
<tr>
<td>---------</td>
<td>----------------------------</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>May 25th - June 3rd</th>
<th>1.5 weeks</th>
</tr>
</thead>
</table>
|                     | • Install and understand how Apache Airavata works using a small cluster in my lab. To ensure success I plan to take help from my labmate who worked on GSoC last summer and has been able to successfully install and work with Airavata over the past year. I can also use that installation directly if needed.  
  • This task has been moved to June 28th - July 1st  
  • Environment setup, including Mesos, reading about requirements for the Java client, design principles, setting up JIRA page to track the progress, setting up git repository  
  • Task completed |

<table>
<thead>
<tr>
<th>June 6th - June 17th</th>
<th>2 weeks</th>
</tr>
</thead>
</table>
|                      | • Understand the job submission module of Airavata with a few sample applications using Airavata. Meet (virtually via google hangout) the Airavata team. I plan to use profilers to track all the method calls involved from the beginning to the end of a task launch. This will help me understand which modules are used for a job launch.  
  • Partially completed  
  • Start working on the Java client for job submissions through Aurora and getting the design corrected to ease further development  
  • Task Completed |
June 20th - July 1st

- Collect requirements of applications run via Apache Airavata and match them to how Aurora’s configuration files need to be created. One risk in this time period is that it may not be able to collect an exhaustive list of requirements from the mentors. I will address it by assuming that Airavata should be able to work with at least two different Cloud and Big Data frameworks such as Marathon and Aurora.
  - Aurora client is being developed and tested, Marathon client still under development
  - Adding more functionality to the Java Client and testing
  - Task Completed

July 1st - July 15th

- Code changes in Airavata to incorporate the creation of the Aurora configuration files. A risk with this timeline is that it may not be possible for the reviewers and mentors to verify all the code I will post. I will address this risk by posting code in small modules all through the project time period.
  - Not Completed
  - I will continue to work on this module
<table>
<thead>
<tr>
<th>August 1st - 12th (2 weeks)</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Develop unit test cases and metrics to determine the effectiveness of the new module in Airavata that interacts with Aurora. One risk with this timeline is ensuring that the Airavata build process does not break due to this integration. I will address this risk by using the available Ansible playbooks for Airavata or develop my own to test installation on a variety of workstations available in my research lab. Additionally, I will submit pull requests and patches throughout the project period.</td>
</tr>
<tr>
<td>• Not Completed</td>
</tr>
<tr>
<td>• Requires the previous task to be completed</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>August 15th - 17th (0.5 weeks)</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Continue with final documentation, testing, and release of the developed software.</td>
</tr>
<tr>
<td>• Task Completed</td>
</tr>
</tbody>
</table>

Additional Work Completed: A library of functions for Apache Marathon has also been completed, essentially will allow airavata to communicate with Apache Aurora and Marathon.
Postings:

Mailing Lists:

- [https://mail-archives.apache.org/mod_mbox/airavata-dev/201606.mbox/%3CCAB75BakRcXvTZXRRC_vyyeEaGxz3gXen_V6ef7VPqDbsMY2vCQ%40mail.gmail.com%3E](https://mail-archives.apache.org/mod_mbox/airavata-dev/201606.mbox/%3CCAB75BakRcXvTZXRRC_vyyeEaGxz3gXen_V6ef7VPqDbsMY2vCQ%40mail.gmail.com%3E)
- [https://mail-archives.apache.org/mod_mbox/airavata-dev/201604.mbox/%3CCAB75BakJLkU38ES2Xnou%3DOFHdyFVZfF4ik0Fm1V0%2BDAzBLVR0A%40mail.gmail.com%3E](https://mail-archives.apache.org/mod_mbox/airavata-dev/201604.mbox/%3CCAB75BakJLkU38ES2Xnou%3DOFHdyFVZfF4ik0Fm1V0%2BDAzBLVR0A%40mail.gmail.com%3E)
- [https://mail-archives.apache.org/mod_mbox/airavata-dev/201606.mbox/%3CCAB75BakJKo5OCNmhsK2BV5fA64QtkqyPw_9E7hECtEK5rKseUQ%40mail.gmail.com%3E](https://mail-archives.apache.org/mod_mbox/airavata-dev/201606.mbox/%3CCAB75BakJKo5OCNmhsK2BV5fA64QtkqyPw_9E7hECtEK5rKseUQ%40mail.gmail.com%3E)
- [https://mail-archives.apache.org/mod_mbox/airavata-dev/201608.mbox/%3CCAB75Bak0tdikyGc1dBU%3DbTdTGU-eU7AG%2BiSYWppHMYLzWHJFDxQ%40mail.gmail.com%3E](https://mail-archives.apache.org/mod_mbox/airavata-dev/201608.mbox/%3CCAB75Bak0tdikyGc1dBU%3DbTdTGU-eU7AG%2BiSYWppHMYLzWHJFDxQ%40mail.gmail.com%3E)
- [https://mail-archives.apache.org/mod_mbox/airavata-dev/201606.mbox/%3CCAB75BamGBQ8r5JgPX AjZaHwinkqR4qgXS14MM0Bb%3D87LRSPxpw%40mail.gmail.com%3E](https://mail-archives.apache.org/mod_mbox/airavata-dev/201606.mbox/%3CCAB75BamGBQ8r5JgPX AjZaHwinkqR4qgXS14MM0Bb%3D87LRSPxpw%40mail.gmail.com%3E)

JIRA:

Airavata

- ![AIRAVATA-1987](https://jira.apache.org/browse/AIRAVATA-1987) - Apache Aurora Integration [OPEN]

Aurora

- ![AURORA-1748](https://jira.apache.org/browse/AURORA-1748) - Documentation not updated on webpage [OPEN]

GIT PR:

- [https://github.com/apache/airavata/pull/39](https://github.com/apache/airavata/pull/39)
Involvement

After the GSOC project, I will continue to work on resource management to improve makespan and power usage of Big Data technologies when used for Scientific Applications in Cloud environments.

Personal Information

Name: Gourav Rattihalli

University: The State University of New York at Binghamton

Major: Pursuing Doctorate in Computer Science (2nd semester)

Expected Graduation: May 2019

Technical Skills: Hands on in depth working experience with Python, C, Java, Javascript, Jquery, CSS, HTML, XML, XSL, Oracle and MySQL database.

Open Source Tools: Experience in working with Apache Mesos, Apache Aurora, Marathon, Docker, Apache Cordova.

Contact Information:

- Email: grattih1@binghamton.edu, gouravrattihalli@outlook.com
- Skype ID: gouravrattihalli@outlook.com

Availability
I am available during the entire GSoC 2016 period. I have no other commitment during this period

Typical Working Hours:

11 am to 1 pm, 2pm to 8 pm

(5 days a week, amounting to 40 hours per week)

Other Commitments: None

Anytime in Summer where you will be away for more than 2 days: None

References

1. Apache Airavarta: https://airavata.apache.org/
3. Twitter: https://blog.twitter.com/2015/all-about-apache-aurora