SPIN Research Paper

Integrating Hadoop into Openstack and setting benchmarks for processing of spatial data

Abstract/Synopsis

The project was abandoned after a while because we realised that there was not much scope for Hadoop to be incorporated into the main system. However, soon after we realised there was not much scope to move forward because of infrastructure support issues. The center was not able to adapt to a totally Hadoop-based system as of yet. For testing purposes, we incorporated Sahara into a test node with select capabilities at first. Then, we tried running jobs on the system to test its capabilities.

Methodology

For the CyberGIS center, we mainly need Hadoop to process vast amounts of spatial data in three forms:

- twitter data,
- vector files,
- or raster files.

Initially, for testing purposes, I created a test environment inside a single instance in OpenStack. This gave me root access to download and delete applications, like Sahara, or other components of OpenStack as and when needed. Inside this test environment, I developed a Sahara environment inside the test node. This environment allows us to run files for testing through the Hadoop cluster on top of Openstack. So, I tried doing that by sending some jobs through the file-system.

So, to ensure the speed and usefulness of Hadoop on the Openstack clusters, there are certain benchmarks that can be run every time we need to test the systems. These benchmarks will be varying in size of files, types of data, including edge cases.

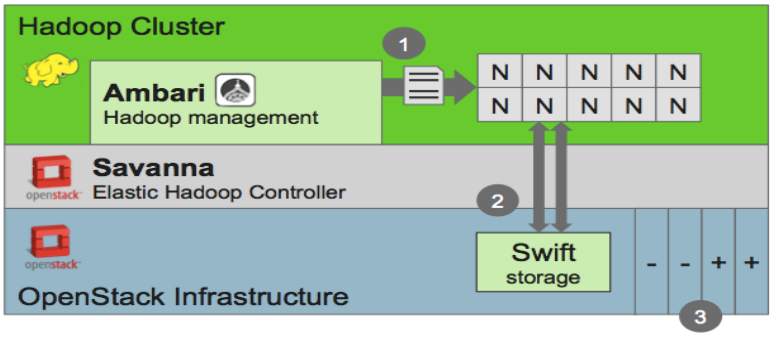
Introduction

Openstack is a popular open source cloud computing software used by corporations to simplify the process of creating virtual machines and using them. One can deploy virtual machines and host them on the cloud using the applications downloaded by OpenStack.

Hadoop is another useful open source software used by corporations to simplify the processing of data. It breaks down data into <key,value> pairs before analysing it in a parallel, distributed manner across several nodes. It allows the faster and more efficient processing of data due to the parallelized processing of data, this, among various other factors make it a very desirable choice

Until now, we would use the same server to process the data hosted in OpenStack, however, using Sahara, in conjunction with Hadoop, allows us to do the data processing on different hadoop clusters, thus making things faster.

This project aims to make processing of large amounts of data easier for the center. A huge amount of information is stored on ROGER, and processing all of it on one compute node in a serialized fashion is not very efficient. Hadoop enables us to change that. The CyberGIS Center will benefit greatly from using an alternate avenue to process data. Files that may take hours or days to run can be done much quicker.



1. **Cluster templates: deploy pre configured Hadoop clusters in seconds from Horizon or Ambari**
2. **HDFS-Swift connectors: move data between HDFS and Swift object storage**
3. **Dynamic infrastructure provides elasticity without any requisite change in cluster configuration**

Literature Review

Some other research I have found has done comparisons between using Hadoop on a virtual machine and a natively configured Hadoop system, This is very similar to what I am doing right now and it has shown that, performance-wise, Hadoop runs faster on a cloud system as opposed to being run natively.

Link:http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=7021017&url=http%3A%2F%2Fieeexplore.ieee.org%2Fxpls%2Fabs_all.jsp%3Farnumber%3D7021017

Another paper I found that talks about the usage of big data on top of OpenStack is: http://iamot2015.com/2015proceedings/documents/P057.pdf. This paper does a good job of detailing the exact infrastructure of Sahara as a service and reaches the conclusion that Sahara drastically reduces the difficulty of using big data services on top of OpenStack directly.

Summary

There was an infrastructure support problem with setting up Hadoop in the system. There was not much availability for the system to be set up even if it did work very well yet. So, we had to wait to actually use Hadoop on the system actively. We were also having some problems debugging some errors in Sahara. By this point though, we had realised that the project itself was not actually viable for the whole of the CyberGIS center to use yet, so, we more or less put it to a side.

Some things I had learnt though:

- Being comfortable with OpenStack and Hadoop; This was one of the biggest takeaways from the project. Since I had not worked with OpenStack or Hadoop before, I had to read a lot of documentation and learn how the cloud really works. I also had to learn to write simple scripts in MapReduce and Hive for testing purposes. This was a bit of a challenge as well because it was my first time understanding how Big Data Processing software runs.
- Setting up test environment in OpenStack: There are plenty of modules available online but the one I used specifically was called DevStack. This is one of the modules that allows us to set up an OpenStack environment in OpenStack itself.
- Using Sahara and running jobs on it: There were a lot of dependency issues with Sahara when installing it and it took me a while to figure out the issues and dissolving them. Again, this taught how to search for dependencies and fix errors using the conf and log files.