Deployment of a Storm cluster

VASANTH METHKUPALLI[1,*] AND AJIT BALAGA[1,**]

[1] School of Informatics and Computing, Bloomington, IN 47408, U.S.A.
[*] Corresponding authors:mvasanthiiit@gmail.com
[**] Corresponding authors: ajit.balaga@gmail.com

project-P015, April 30, 2017

This project focuses on deployment of Apache Storm using Ansible playbook on Chameleon Cloud VM, future systems cloud and benchmarking of the deployment.

Keywords: Storm, Ansible, Java, Python

https://github.com/cloudmesh/classes/blob/master/project/S17-IR-P015/report/report.pdf

1. INTRODUCTION

Apache Storm is a distributed stream processing computation framework under Apache. In this project, Storm cluster of one or more Chameleon cloud VMs and future sytems cloud is deployed using Ansible playbook and benchmarking is done to measure the time it took for deployement using a TBD benchmarking tool[1][2].

2. APACHE STORM

Apache Storm is a free and open source distributed realtime computation system. Storm makes it easy to reliably process unbounded streams of data, doing for realtime processing what Hadoop did for batch processing. Storm is simple, can be used with any programming language, and is a lot of fun to use!. Storm has many use cases: realtime analytics, online machine learning, continuous computation, distributed RPC, ETL, and more. Storm is fast: a benchmark clocked it at over a million tuples processed per second per node. It is scalable, fault-tolerant, guarantees your data will be processed, and is easy to set up and operate.Storm integrates with the queueing and database technologies you already use. A Storm topology consumes streams of data and processes those streams in arbitrarily complex ways, repartitioning the streams between each stage of the computation however needed. Read more in the tutorial[1][3].

3. MILESTONES

- Performing Analysis on local VM
- Deploying Storm and Hadoop on FutureSystems and Chameleon Cloud
- Analysis on the distributed cloud environment
- Benchmarking
- Final update with report

4. TECHNOLOGIES

- Distributed Computation and Storage:- Storm
- Development:- Python and Java
- Deployment:- Ansible

5. DEPLOYMENT

Ansible Playbook is used as the application and configuration deployment tool. Deploying the hadoop and spark framework into the cluster environment. Ansible will help push configurations to the environment automatically based on playbooks written for various configurations.[4][2]

6. BENCHMARKING

TBD

REFERENCES

[1] "Apache Storm." [Online]. Available: http://storm.apache.org/
[2] "Apache ZooKeeper - Releases." [Online]. Available: http://zookeeper.apache.org/releases.html
[3] "Storm (event processor) - Wikipedia." [Online]. Available: https://en.wikipedia.org/wiki/Storm_(event_processor)
[4] "Apache Storm." [Online]. Available: http://storm.apache.org/