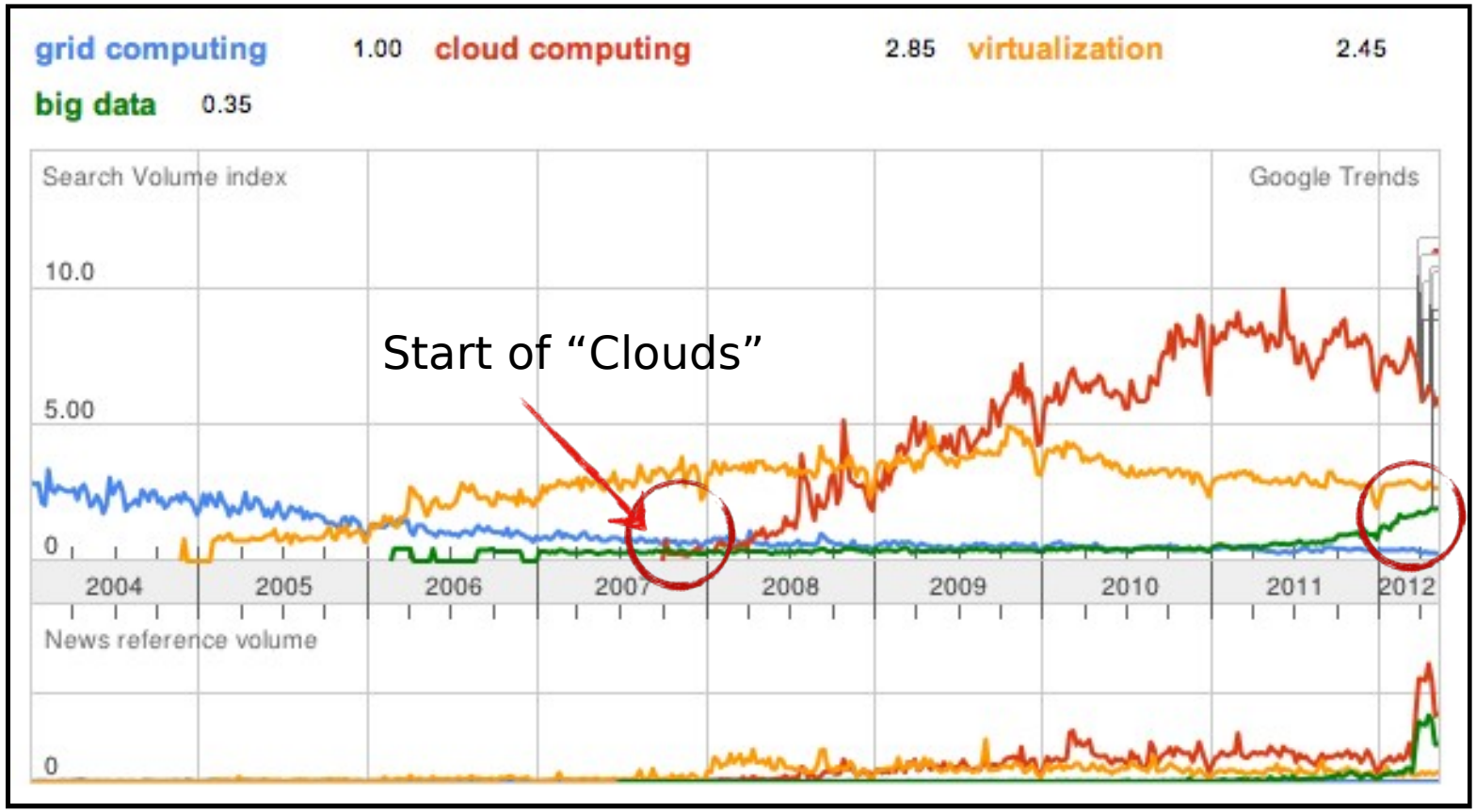
CloudStack and Big Data

Sebastien Goasguen
@sebgoa

May 22nd 2013

LinuxTag, Berlin

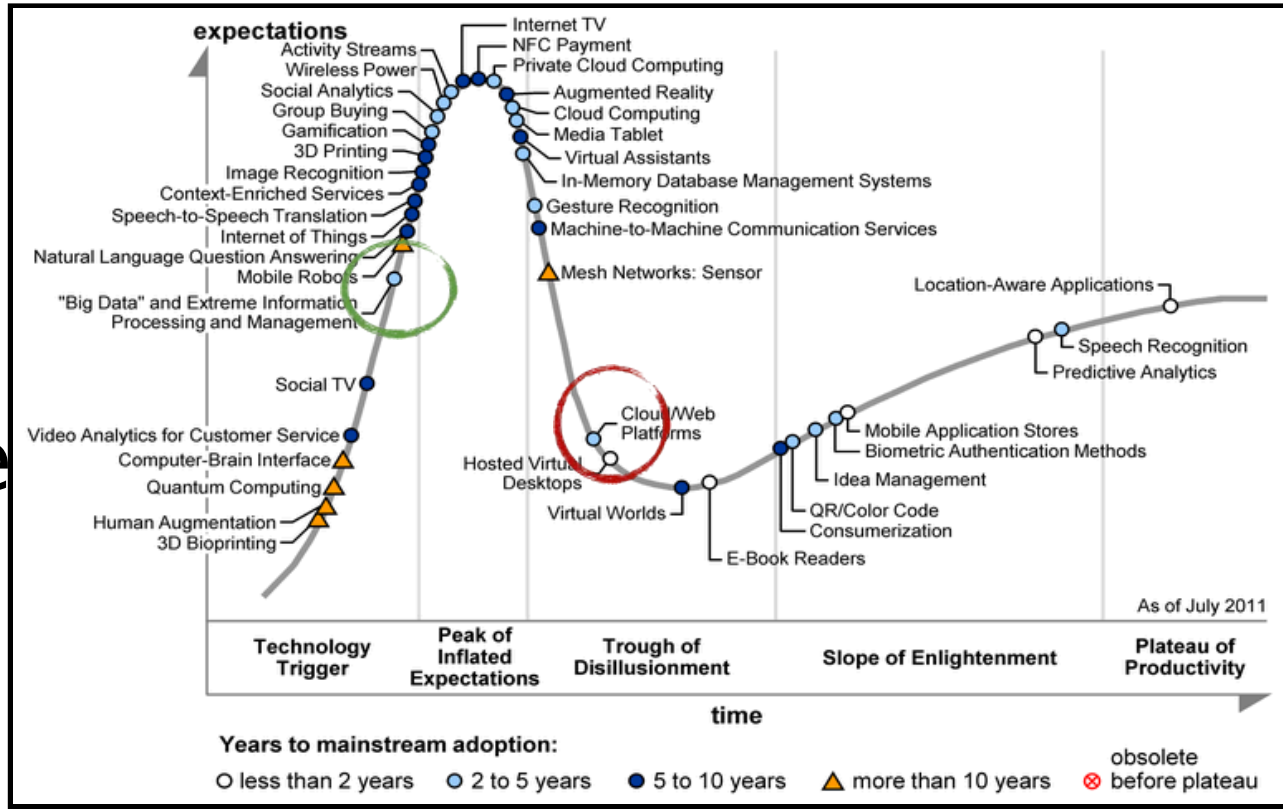
Google trends



- Cloud computing trending down, while “Big Data” is booming. Virtualization

BigData on the Trigger

- Cloud Computing Going down to the “through of Disillusionment”
- “Big Data” on the Technology Trigger



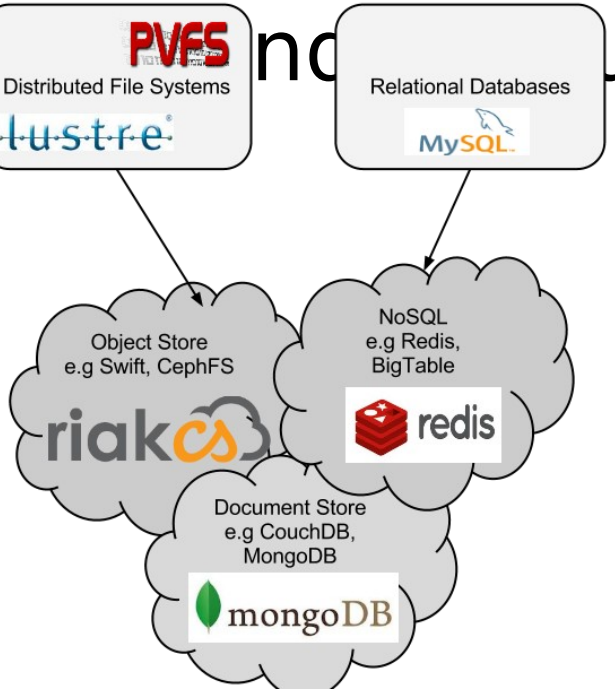
- **Big Data**

What is Big Data ?

- Large scale datasets
  - From scientific instrume
  - From Web apps logs
  - From Health records...
- Complex datasets
  - Not necessarily large.
  - E.g Unstructured data
  - E.g Natural Language
  - E.g IBM Watson

A natural evolution

- From traditional file systems and databases
- To large scale object store and nosql movement designed to handle massive s nc urrency

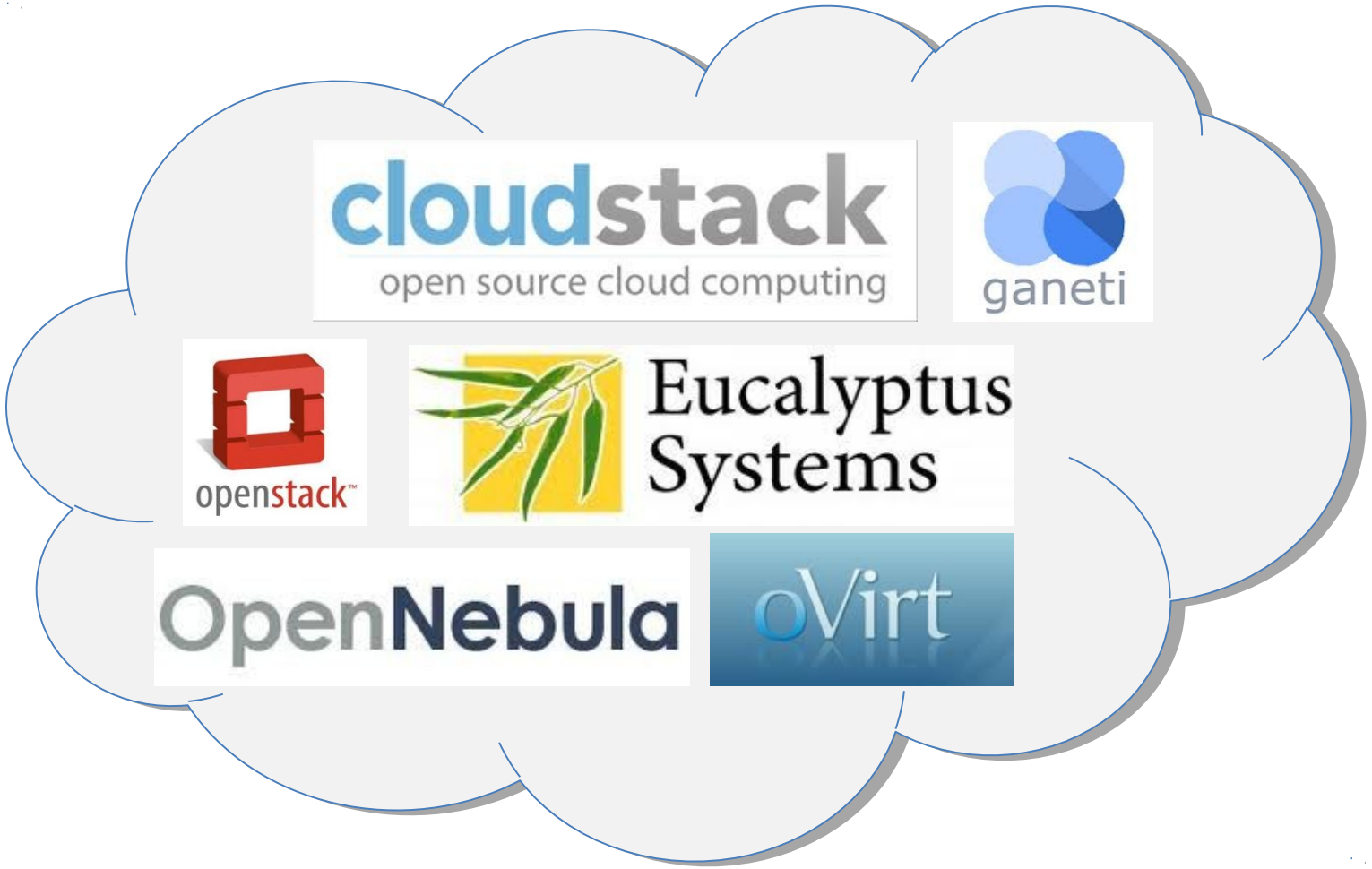


BigData and map-reduce

- While BigData is often associated with HDFS, Map-Reduce is the algorithm used to parallelize data processing.
- BigData ≠ Map-Reduce ≠ HDFS
- Map-reduce is a way to express embarrassingly parallel work easily.
- You can do Map-Reduce without HDFS.
- E.g Basho map-reduce on riackCS

- **CloudStack**

How about IaaS ?



IaaS is really:

- **A Data Center Orchestrator**
  - Data storage
  - Data movement
  - Data processing
- **That can:**
  - Handle failures
  - Support large scale
  - Be programmed

What is CloudStack ?

- Open source Infrastructure as a Service (IaaS) solution.
- **“Programmable” Data Center orchestrator**
- Hypervisor agnostic (with addition of bare metal provisioning)
- Support scalable storage (Ceph, RIAK CS…)
- Support complex enterprise networking (e.g Firewall, load

A bit of History

- Original company VMOPs (2008)
  - Founded by Sheng Liang former lead dev on JVM
- Open source (GPLv3) as CloudStack
- Acquired by Citrix (July 2011)
- Relicensed under ASL v2 April 3, 2012
- Accepted as Apache Incubating Project April 16, 2012
- First Apache (ACS 4.0) released

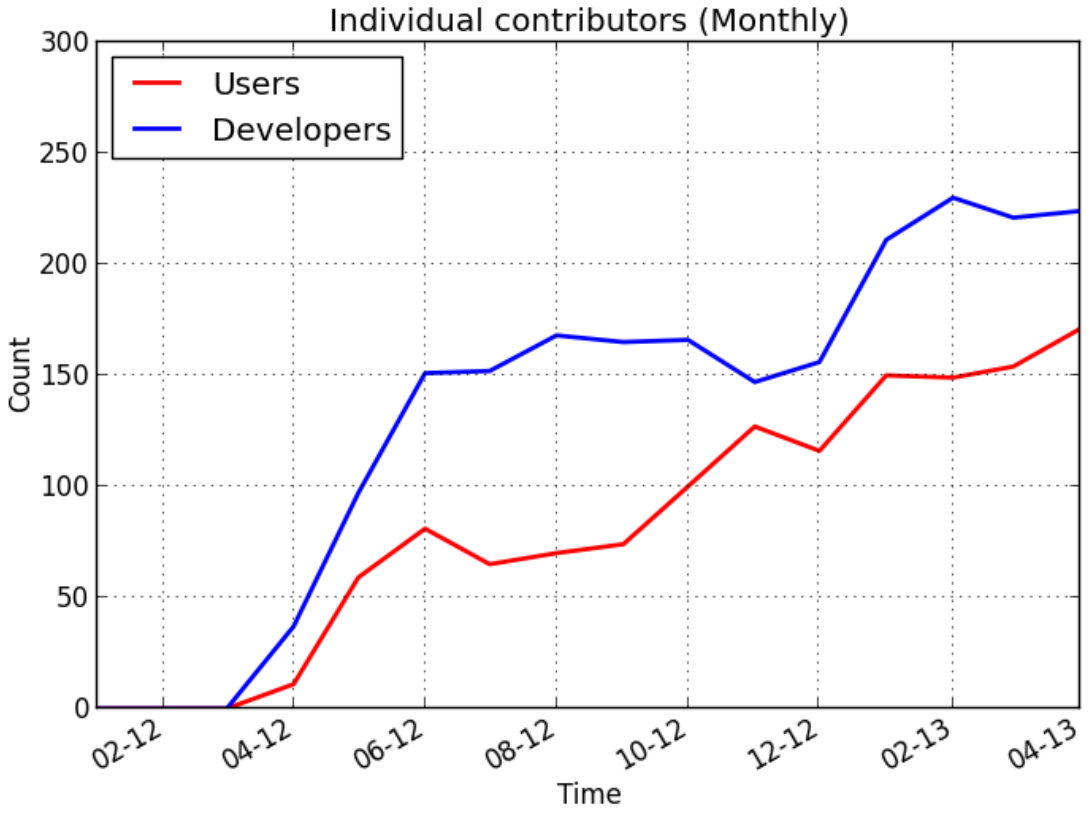
Why ASF ?

- Open Sourced CloudStack to:
  - Build a community
  - Facilitate the building of an ecosystem
  - Faster time to market

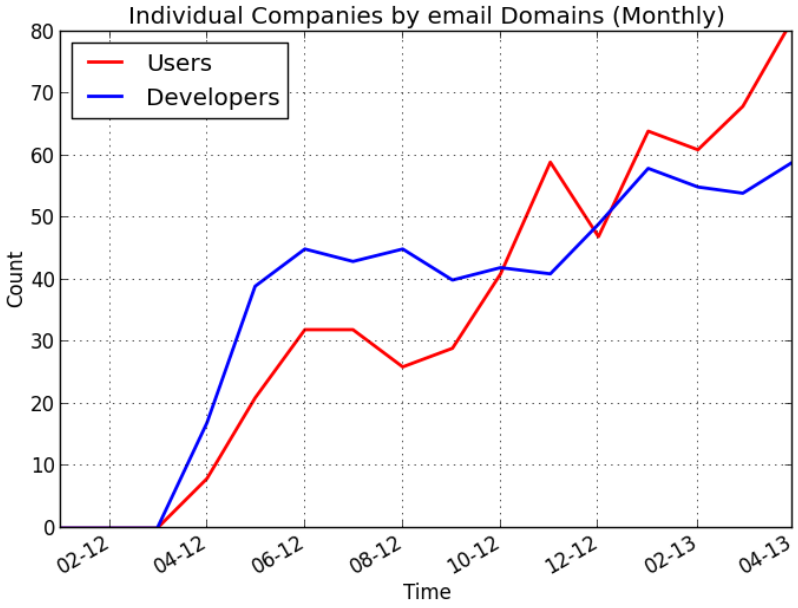
- ASF highly recognized OSS foundation.
- ASF clear processes
- Individual contributions, companies have no standing

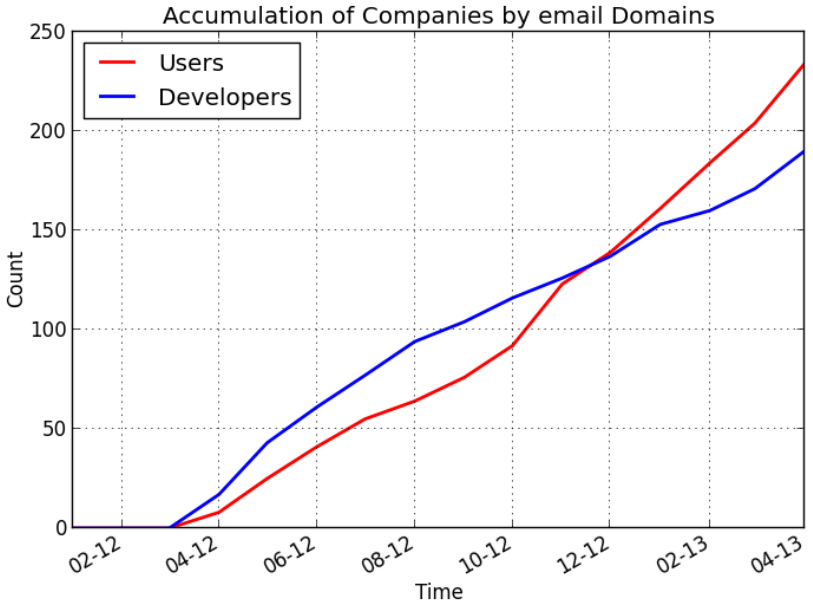
cloudstack
open source cloud computing

Monthly Contributors



Companies


Individual Companies by email Domains (Monthly)
Users
Developers
Count
Time
0
10
20
30
40
50
60
70
80
02-12
04-12
06-12
08-12
10-12
12-12
02-13
04-13


Accumulation of Companies by email Domains
Users
Developers
Count
Time
0
50
100
150
200
250
02-12
04-12
06-12
08-12
10-12
12-12
02-13
04-13

Multiple Contributors

Sungard: Announced last week that 6 developers were joining the Apache project

Schuberg Philis: Big contribution in building/packaging and Nicira support

Go Daddy: Maven building

Caringo: Support for own object store

Basho: Support for RiackCS

- **The Apache Software Foundation**

Apache Software Foundation

We consider ourselves
not simply a group of projects sharing a server, but rather a community of developers and users.

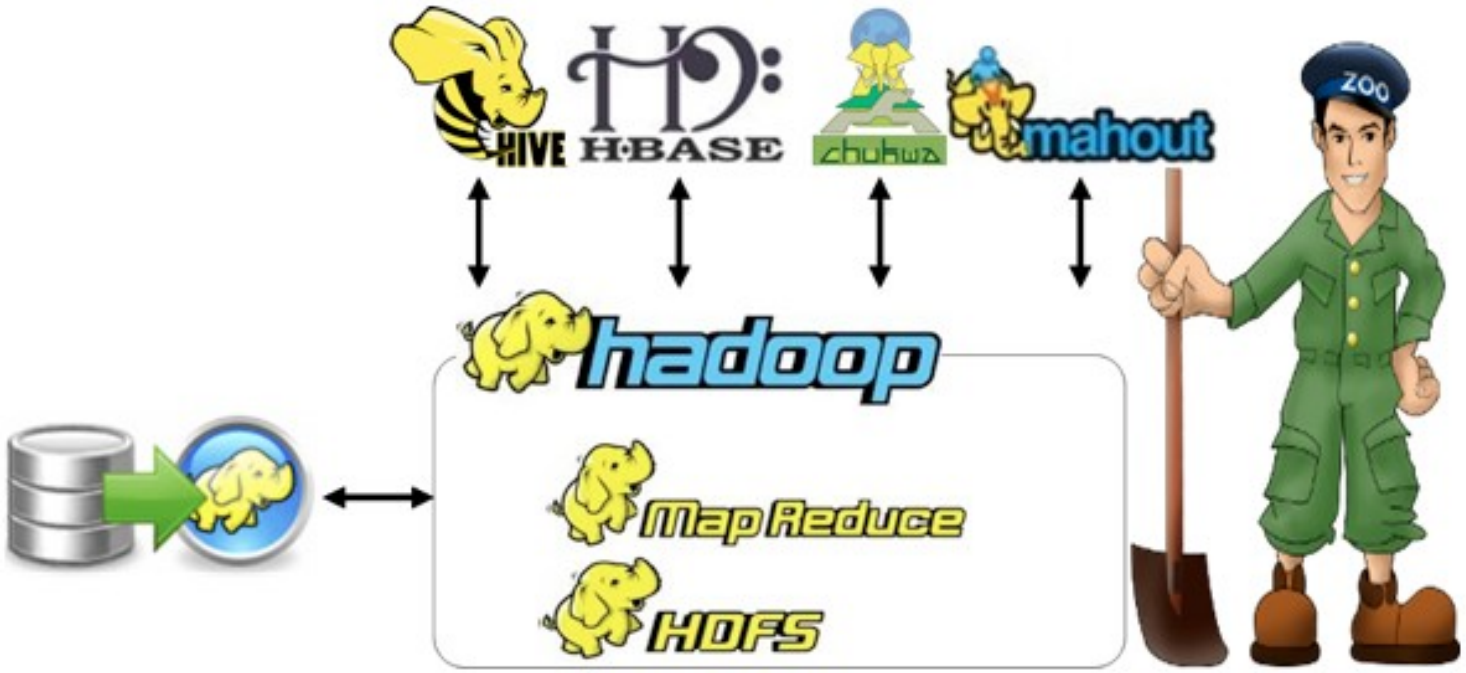
The Apache Software Foundation
provides support for the Apache community of open-source software projects, which provide software products for the public good.

The Apache projects are defined
by collaborative consensus based processes, an open, pragmatic software license and a desire to create high quality software that leads the way in its field.

- 35 projects in incubation:
  - 11 Hadoop related (including Apache provisonr)
  - ~30% Big Data related
  - +jclouds
- 116 top level projects:
  - ~14 cloud or bigdata +10%
  - Deltacloud, Libcloud, Whirr
  - Hadoop, couchdb, cassandra
  - Bigtop, accumulo, lucene, UIMA

Hadoop Ecosystem

- Complex ecosystem to perform data processing on big-data
- Software components can be managed in VMs via CloudStack



- **BigData and CloudStack**

CloudStack and BigData

- Apache CloudStack is a data center orchestrator
- BigData solutions as **storage backends** for image catalogue and large scale instance storage.
- BigData solutions as **workloads** to CloudStack based clouds.

Storage

- **Primary Storage:**
  - Anything that can be mounted on the node of a cluster.
  - Cluster LVM, iSCSI, NFS, Ceph
  - Holds disk images of running VMs and user block stores.
- **Secondary Storage:**
  - Available across the zone
  - Holds snapshots and templates (image repo)
  - **Can use multiple object stores (Gluster , Ceph, riackCS, Swift, Caringo)**

Big Data and CloudStack

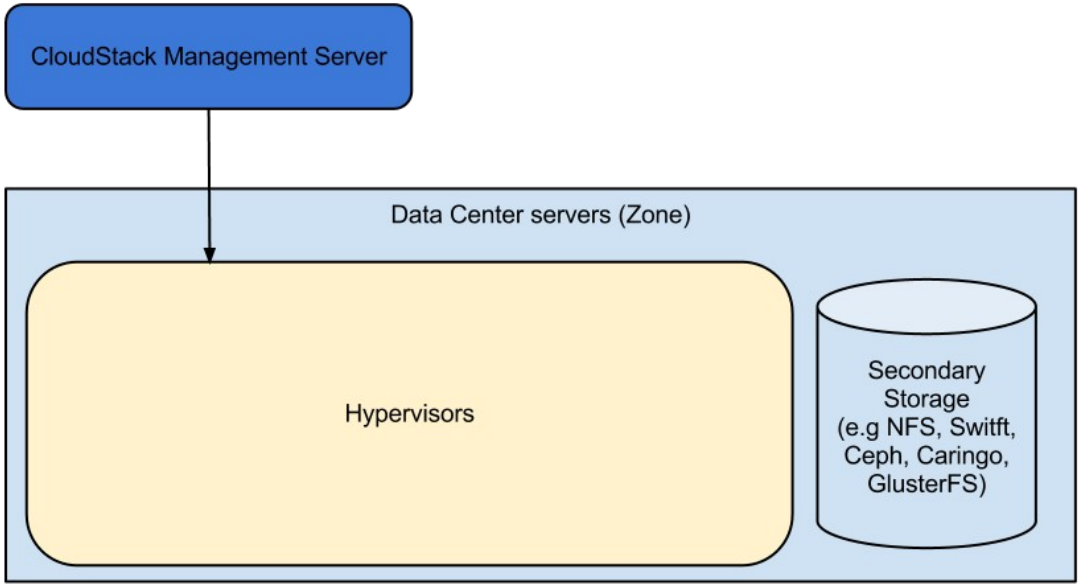
- **“Big Data” solutions can be used as secondary storage** (OpenStack swift, Caringo, CephFS, Gluster FS, RiackCS...).
- Used to deploy a large scale storage backend to manage user images, and user data volumes.
- Primary intent is **not** to use it inside the VMs for data processing.

CloudStack and Baremetal

- CS supports baremetal provisioning.
- This opens the door to multiple scenarios for Big-Data store, Clouds
  - Provision Hadoop cluster on baremetal
  - Operate “Hybrid” cloud: part Hypervisor for VM provisioning, part baremetal for data store.
  - Reconfigure entire cloud on-demand

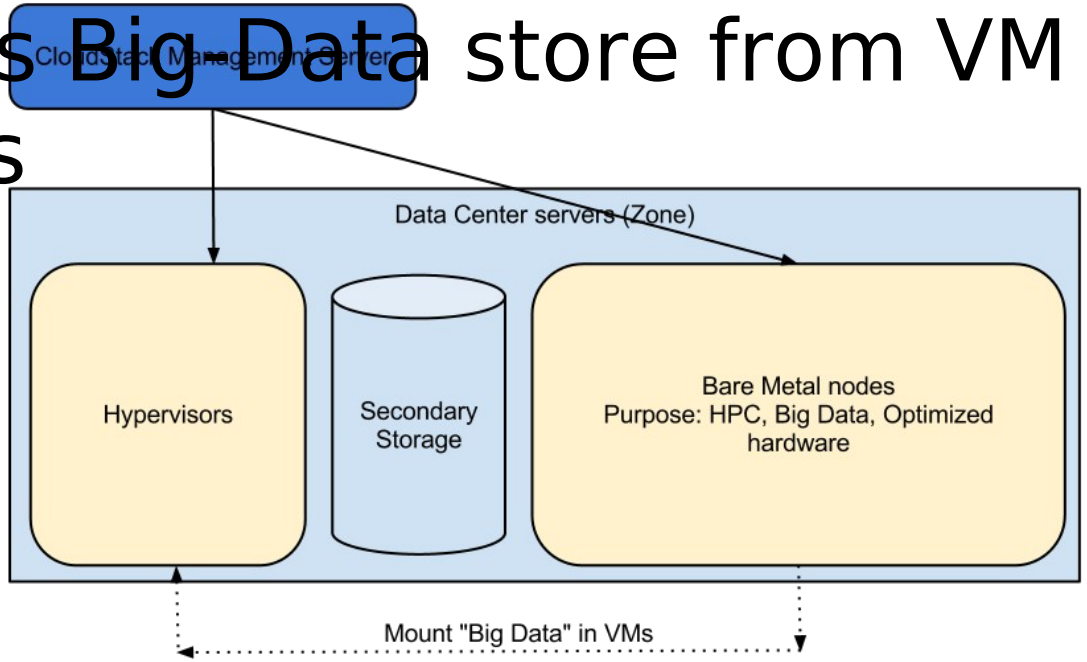
“Traditional” CS deployment

- Farm of hypervisors, separate secondary storage to store VM images and data volumes.



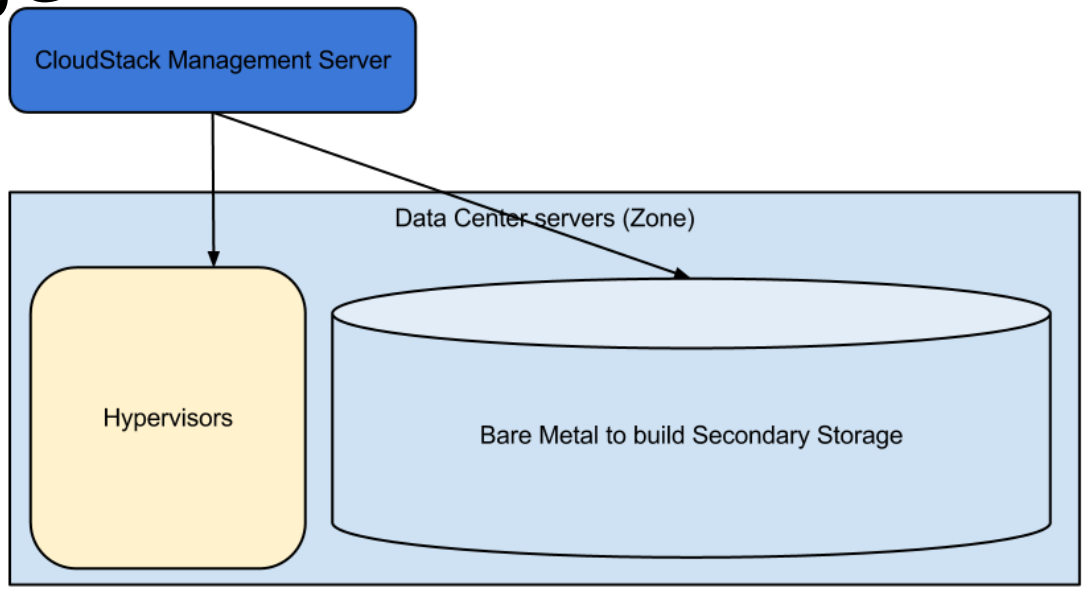
“Bare Metal” Hybrid deployment

- Set of hypervisors, stand-alone secondary storage, bare metal cluster with specialized hardware or software.
- Access Big Data store from VM guests



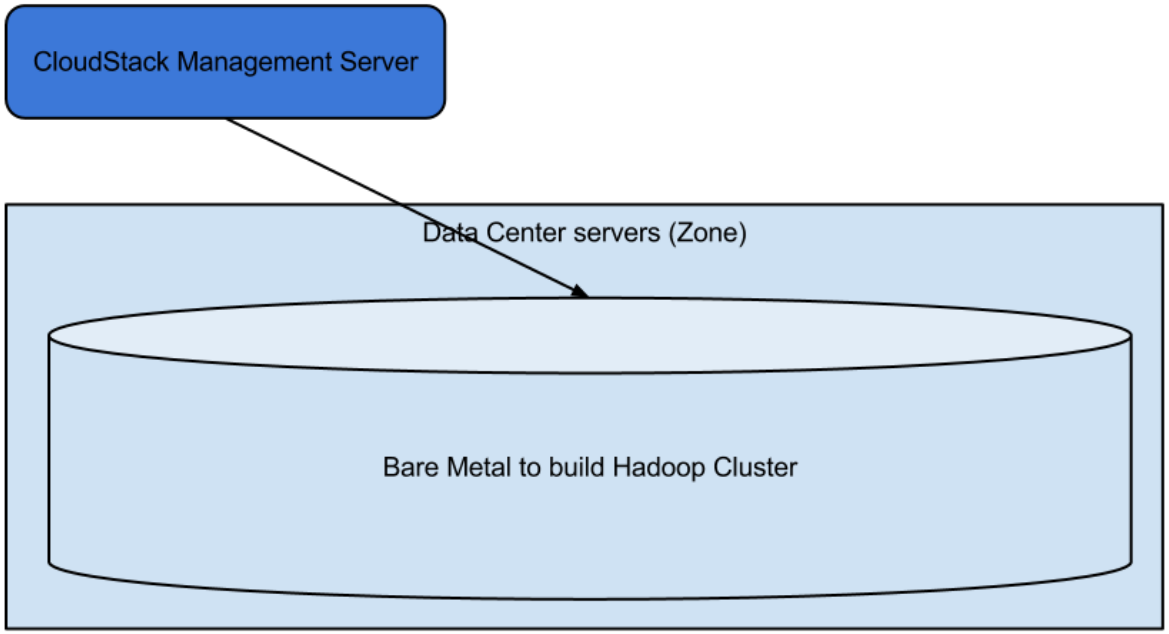
“Bare metal” cluster as secondary storage

- Use bare-metal provisioning to manage larges-scale secondary storage



“Pure” Big-Data store

- Use CS as a traditional data center provisioning system and build a Big-Data store on-demand



Combinations

- CloudStack offers the possibility to switch between these modes on-demand
- **An elastic reconfigurable cloud**

- Just be careful not to override your data

Big Data as a Workload to the Cloud

tools and demo…

Apache Whirr

- Big Data Provisioning tool
- Deploys Hadoop, cdh, Hbase, Yarn, etc in the Cloud
- Use jclouds
- Works with multiple cloud providers including CloudStack

jClouds

- Under Incubation at the Apache Software Foundation (ASF)
- Wrapper to multiple cloud providers

Whirr Configuration

```
whirr.cluster-name=myhadoopcluster
whirr.instance-templates=1 hadoop-jobtracker+hadoop-
namenode,1 hadoop-datanode+hadoop-tasktracker
whirr.provider=cloudstack
whirr.private-key-file=${sys:user.home}/.ssh/id_rsa
whirr.public-key-file=${sys:user.home}/.ssh/id_rsa.pub
whirr.env.repo=cdh4
whirr.hadoop.install-function=install_cdh_hadoop
whirr.hadoop.configure-function=configure_cdh_hadoop
whirr.hardware-id=b6cd1ff5-3a2f-4e9d-a4d1-8988c1191fe8
whirr.endpoint=https://api.exoscale.ch/compute
whirr.image-id=1d16c78d-268f-47d0-be0c-b80d31e765d2

whirr.identity=<your access key>
whirr.credential=<your secret key>
```

- **Demo ?**

Other tools

- Brooklyn

brooklyn®

Overview Download Getting Started Walkthrough User Guide Examples Contributing Search: type & hit enter

Using brooklyn » Examples » Whirr Hadoop Cluster

Whirr Hadoop Cluster

- Apache Provisionr incubating

Apache Provisionr

Apache incubator

Home | Source Code | Issue Tracker | Mailing lists | Releases | Team

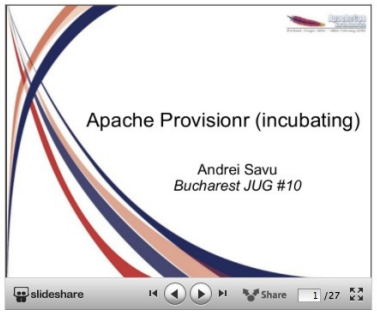
Quick introduction

Provisionr is an effort to develop a service that can be used to create and manage pools of virtual machines on multiple clouds. Our focus is on semi-automated workflows, cloud portability and on providing a robust foundation for configuration and higher level management.

Objectives

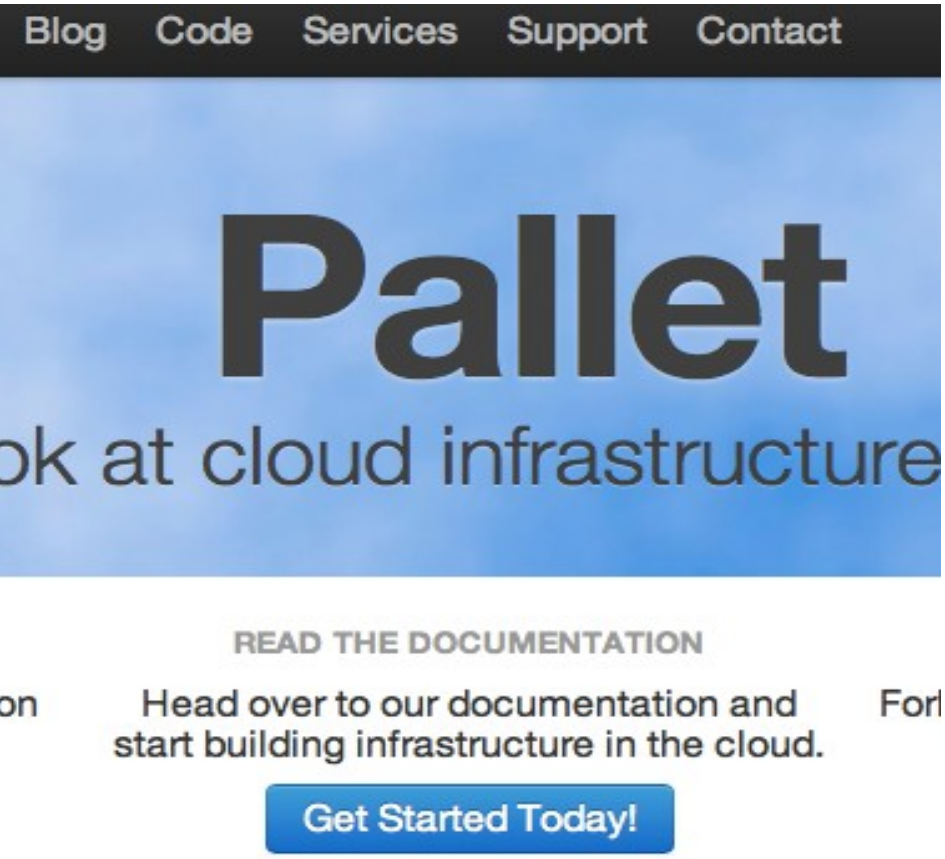
Provisionr solves the problem of cloud portability by hiding completely the APIs and only focusing on building a cluster that matches the same set of assumptions on all clouds, assumptions like: a specific OS, pre-installed packages and binaries, sane dns settings, ssh & vpn access etc. - think a solid foundation for configuration.

As a secondary goal Provisionr will also provide primitives for building automatic or semi-automatic workflows for configuring and monitoring services, workflows that assume that all the machines share a common set of characteristics as described above.



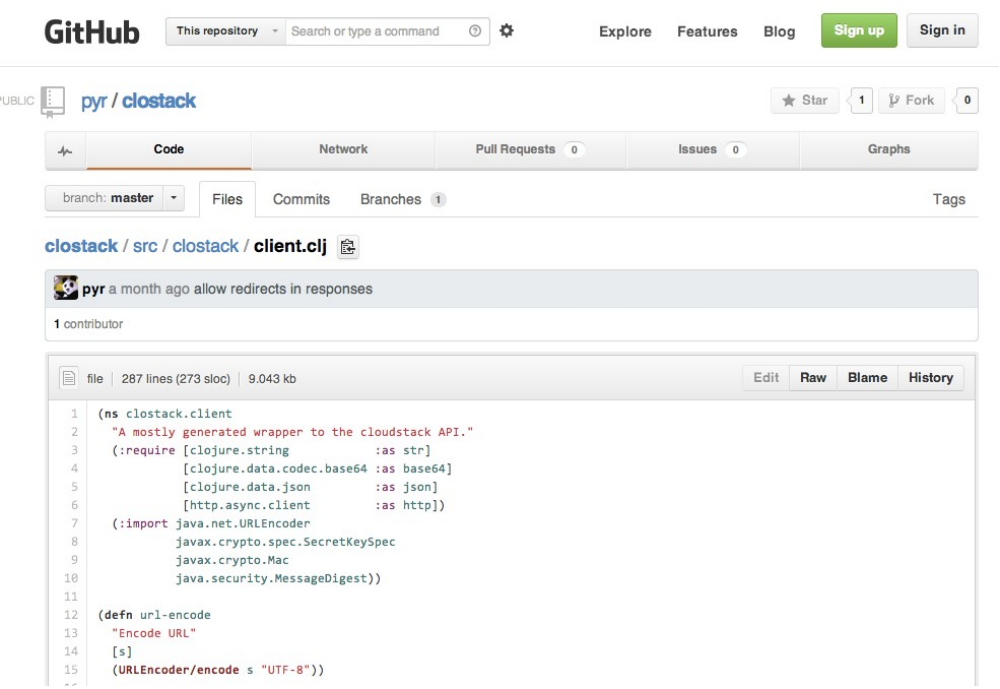
Apache Provisionr (incubating) - Bucharest JUG 10
from Andrei Savu

Others: Pallet

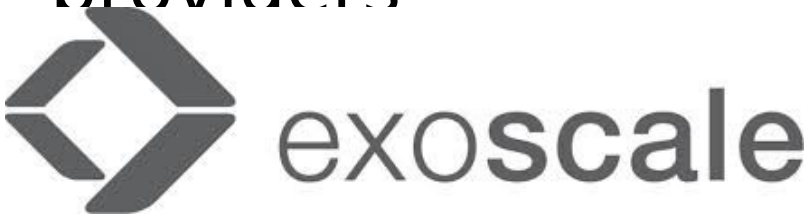


- Clojure based provisioning tool
- Provisions Hadoop clusters in the cloud.
- Equivalent to Whirr but in clojure

CloStack



- Clojure client for CloudStack
- Uses native CloudStack API
- Developed by @pyr at exoscale.ch , a CloudStack based public cloud providers



cloudstack
open source cloud computing

More than hadoop

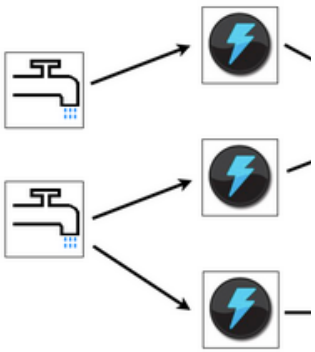
Storm

Distributed and fault-tolerant realtime computation

about documentation blog

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 analy
continuous computation, distributed RPC, E
benchmark clocked it at over **a million tup
node**. It is scalable, fault-tolerant, guarant
processed, and is easy to set up and opera



GitHub This repository Search or type a command Explore Features Blog Sign up Sign in

PUBLIC nathanmarz / storm ★ Star 5,920 Fork 901

Code Network Pull Requests 19 Issues 174 Wiki Graphs

Distributed and fault-tolerant realtime computation: stream processing, continuous computation, distributed RPC, and more — Read more

http://storm-project.net

Clone in Mac ZIP HTTP SSH Git Read-Only https://github.com/nathanmarz/storm.git Read-Only access

branch: master Files Commits Branches 43 Tags

storm / 1000+ commits

Merge pull request #570 from d2r/d2r-config-final-strings

nathanmarz authored 5 days ago latest commit e4a59d35c7

bin	a month ago	convert to multimodule setup (everything in essentially one module bu... [nathanmarz]
conf	2 months ago	keep zmq in clojure [afeng]
logback	21 days ago	access.log should rollover separately [Derek Dagit]
storm-console-logging	a month ago	convert to multimodule setup (everything in essentially one module bu... [nathanmarz]
storm-core	6 days ago	Make Config Strings final [Derek Dagit]
storm-lib	a month ago	convert to multimodule setup (everything in essentially one module bu... [nathanmarz]
.gitignore	a month ago	convert to multimodule setup (everything in essentially one module bu... [nathanmarz]

On-Going Big-Data development

- Hadoop being an Apache project written in Java, there is great potential synergy between CloudStack and Hadoop:

  e.g Develop Elastic Map-Reduce mechanisms to provide map-reduce processing in CS backed by HDFS. Implementation of AWS EMR API.

- Integration of Basho map-reduce (coming in 4.2 release)

GSoC

- ASF is a mentoring organization for GSoC
- CloudStack has several proposals under consideration
  - Improved CloudStack support in Apache Whirr and Provisionr
  - Integration of Apache Mesos with

Info

- Apache Top Level project
- http://www.cloudstack.org
- #cloudstack on irc.freenode.net
- @cloudstack on Twitter
- http://www.slideshare.net/cloudstack
- http://cloudstack.apache.org/mailing-lists.html

Welcoming contributions and feedback, Join the fun !