

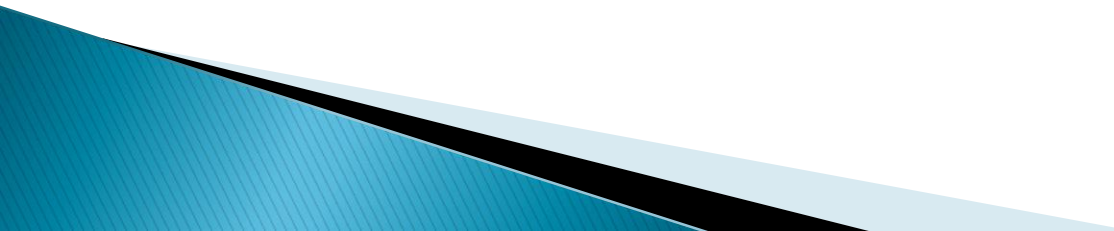
云存储

If your data are worth keeping, then they are
worth keeping online and sharing

报告人：沈洋斌

单 位：国家天文台-昆明理工大学计算机重点实验室
联合培养研究生

主要内容

- ▶ 应用背景
 - ▶ 分布式文件系统（云存储）
 - 概述
 - 主流的分布式文件系统
 - ▶ 系统设计
 - ▶ 原型展示
- 

应用背景

- ▶ 数据的重要性
- ▶ 数据存储现状
 - 大量历史数据离线保存，需要较大的Effort才能得到自己想要的数据。
- ▶ 数据管理现状
 - 较多的人工干预
 - 硬盘损坏，数据丢失

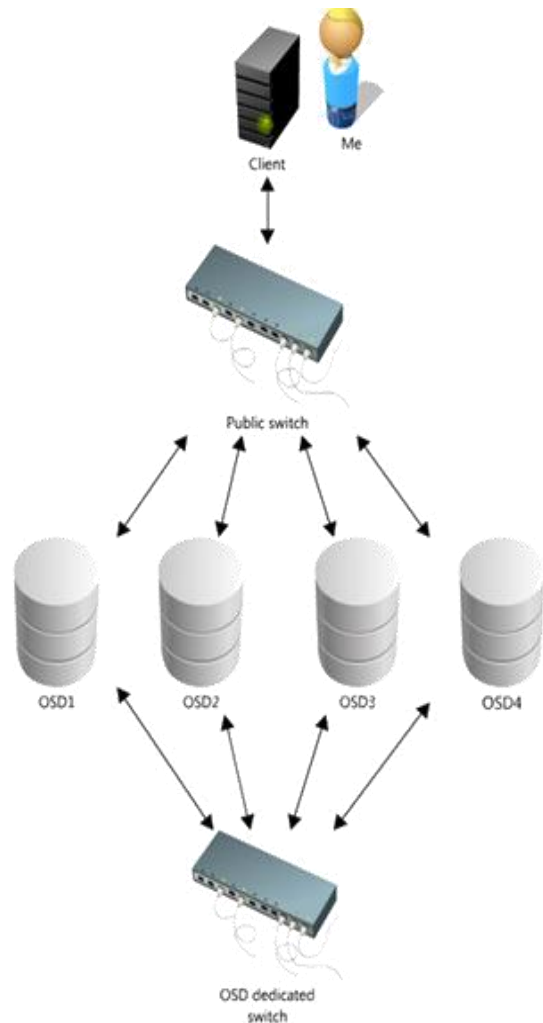
How To Make Our Life Easier?

私有云存储

分布式文件系统



- ▶ 由多台物理计算机节点形成的网络文件系统
 - 存储节点可以动态添加，所以存储容量可以动态提高（Scalability）
 - 同一份数据在多个节点上存在备份，提供数据安全性（Reliability）
 - 数据分割存储，提高读写效率
 - 提供抽象底层硬盘的统一视图



现有开源分布式文件系统（1）

	Lustre	HDFS	GlusterFS	MooseFS	MogileFS	Ceph
Metadata Server	有。存在单点故障。MDS使得用户可以访问到存储在一个或多个MDT上的元数据。每个MDS管理着Lustre中的文件名和目录，为一个或者多个MDT提供网络请求处理。	有。存在单点故障。NameNode是一个中心服务器，负责管理文件系统的命名空间以及客户端对文件的访问。集群中的DataNode负责管理它所在节点上的存储。	无。不存在单点故障。靠运行在各个节点上的动态算法来代替MDS,不需同步元数据,无硬盘I/O瓶颈。	单个MDS。存在单点故障和由单个Master Server带来的性能瓶颈。100万文件大约需要300M内存。25百万份文件大约需要8GiB内存和25GiB硬盘空间。	无单点故障。系统由Client、数据库、Trackers Server和Storage Server四种角色组成。基于GFS的一种开源实现。	多个MDS，不存在单点故障和瓶颈。MDS可以扩展，不存在瓶颈。
POSIX兼容	是	不完全。为了实现对流式读取	是	是	不兼容POSIX	是
用户 / 组的配额	支持	支持	不详	支持	不详	不支持
权限控制	Access Control List (ACL)	独立实现的一个和POSIX类似的文件和目录的权限模型	不详	Access Control List (ACL)	不支持ACL	Access Control List (ACL)
快照	Lustre可以创建所有卷的快照，并且在快照系统中把他们集合在一起，以供Lustre挂载。	利用快照，可以让HDFS在数据损坏时恢复到过去一个已知正确的时间点。HDFS目前还不支持快照功能，但计划在将来的版本支持。	不支持	可以对整个文件甚至在正在写入的文件创建文件的快照。	不详	支持。用户可以创建快照。
网络支持	可支持各种网络，如 Infiniband, TCP/IP, Quadrics Elan, Myrinet (MX and GM) 和 Cray。	只支持TCP/IP	支持很多种网络，如以太网和光纤Infiniband。	支持多种网络。	MogileFS客户端可以通过NFS或HTTP来和MogileFS的存储节点来通信，但首先需要告知跟踪器。	支持多种网络。

现有开源分布式文件系统（2）

	Lustre	HDFS	GlusterFS	MooseFS	MogileFS	Ceph
文件分割	采用RAID 0模式，将数据分割到一定数量的对象中去，每个对象包含很多数据块。当一个被写入的数据块超过了这个对象的容量时，下一次写入将被存储到下一个目标中。Lustre可以把文件最多分布到160个目标存储上。	一个典型的数据块大小是64MB。因而，HDFS中的文件总是按照64M被切分成不同的块，每个块尽可能地存储于不同的Datanode中。	不支持。故适合于存放小文件。	文件被分片，数据块保存在不同的存储服务器上。	不支持。故不适合存储超大文件。	文件被分片，每个数据块是一个对象。对象保存在不同的存储服务器上。
备份及恢复	提供两个备份工具，一个用于扫描文件系统，一个用于打包备份和加压恢复。	支持数据复制。采取一定的策略，将文件的多个副本存放到不同的节点。在读取副本的时候，系统会自动选择最近的副本。	支持数据复制，提供全局的命名空间。多个文件的多个副本可以被放置到不同的host上去。这意味着，host以及所有与其相关的磁盘可以因为任何原因而下线，而由其它弄得host来担任文件的完整副本（例如网络分割，甚至只是系统维护）。读取副本时，系统会选择最近的副本。	多副本，手动恢复。由数据的多副本提供可靠性。	文件是基于他们的“类”，文件可以自动的在多个存储节点上复制，这是为了尽量少的复制，才使用“类”的。假如有三份JPEG图片的拷贝，但实际上只有1或2份拷贝，那么Mogile可以重新建立遗失的拷贝数。用这种办法，MogileFS(不做RAID)可以节约在磁盘，否则你将存储同样的拷贝多份，完全没有必要。	多副本，当节点失效时，自动迁移数据、重新复制副本。由数据的多副本提供可靠性。
技术支持	Sun Microsystems（现为Oracle）	Apache Hadoop项目组	Gluster	Core Technology sp. z o.o.	Danga	UCSC、Linux内核（since 2.6.34）

现有开源分布式文件系统（3）

	Lustre	HDFS	GlusterFS	MooseFS	MogileFS	Ceph
故障自救	Lustre中的MDS可以被配置成一个主动/被动对，而OSS通常被配置成主动/主动对，这样可以提供没有任何开销的冗余。通常，热备的MDS是另一个LustreFS的活跃MDS，所以在集群之中不会出现空闲的设备。	心跳信号检测机制。每个Datanode节点周期性地向Namenode发送心跳信号。Namenode是HDFS集群中的单点故障所在。如果Namenode机器故障，是需要手工干预的。目前自动重启或在另一台机器上做Namenode故障转移的功能还没实现。	当系统中出现永久性损坏的时候（如一台服务器的磁盘发生故障，导致本地文件无法读取），系统会更换主机并将其加入集群，并自动开始恢复过程。Gluster对一些假定的故障（如硬件故障，磁盘故障，网络被分割，以外断电）都有处理预案，系统会自动处理这些故障，不需要管理员的参与。	可通过第三方软件Drbd+Heartbeat+Pacemaker的方式实现心跳检测机制。	在一个非存储区域网络的RAID（non-SAN RAID）的建立中，磁盘是冗余的，但主机不是，如果你整个机器坏了，那么文件也将不能访问。MogileFS在不同的机器之间进行文件复制，因此文件始终是可用的。	Cluster Monitors识别机器故障，当节点失效时，自动迁移数据、重新复制副本。
商业应用	为全球100大高性能存储计算HPC集群中的40%多提供支持。目前已经拥有成千上万的客户端系统，PB级的存储，数百GB/s的I/O吞吐量。	包括中国移动、百度、网易、淘宝、腾讯、金山和华为等众多公司都在研究和用它，另外还有中科院、暨南大学、浙江大学等众多高校在研究它。	世界各地有100多个地区在测试或者使用Guster。大多集中在欧洲和美国地区。亚洲及中国用户较少。	国内较多	51.com、豆瓣、yupoo等网站	无
动态扩容	件数据存在OSS上的对象中。Lustre的容量以及总的存储带宽可以在不中断任何操作的情况下，通过增加新的带有OST的OSS来实现动态扩展。	不详	支持动态扩容。可以通过简单的操作动态增加存储，服务器和客户端。	支持动态扩容。可以在线扩容。	不详	支持动态扩容。

现有开源分布式文件系统（4）

	Lustre	HDFS	GlusterFS	MooseFS	MogileFS	Ceph
扩展能力	高扩展性。在产业化环境中，大多数集群的客户端节点在1万到2万左右，最多可以支持到2.6万个客户端节点。目前足以支持40PB的文件系统。	在一个集群里可扩展到数百个节点。一个单一的HDFS实例应该能支撑数以千万计的文件。	支持线性扩展。可以轻松地扩展到数百PB的量级。	增加存储服务器，可以提高容量和文件操作性能。但是由于不能增加MDS，因此元数据操作性能不能提高，是整个系统的瓶颈。	不详	可以增加元数据服务器和存储节点。容量可扩展。文件操作性能可扩展。元数据操作性能可扩展。
安装	略微复杂	简单	简单。在Ubuntu等发行版Linux中有内嵌的软件源的支持。	简单	简单	简单
开发语言	C语言	Java	C语言	C语言	Perl	C++
I/O特性	Lustre在产业化环境中的部署目前可以提供100 GB/s的性能。Lustre单独客户端的节点的吞吐量最大可以达到2 GB/s，而OSS最大可以达到2.5 GB/s。在Oak Ridge National Laboratories, Lustre运行在Spider File System上，达到了240 GB/s的性能。在千兆以太网中，写的速度保持在100 MB/s左右。	实际环境中通过Client读的速率反而比写的速率慢。在千兆网络中写速度维持在100MB/s左右，读速度为89~90MB/s	千兆以太网中，写的速度保持在65 MB/s左右。读的速度保持在117 MB/s左右。但对小文件的写入则只有3 MB/s左右的速度。官方测试最高带宽达到32GB/s，号称比lustre还要好。	适合于存储小文件，读写速度受磁盘IO小文件存储速度的限制。	适合于存储小文件，读写速度受磁盘IO小文件存储速度的限制。但对于海量小文件，效率要比MooseFS高。	

分布式文件系统选择

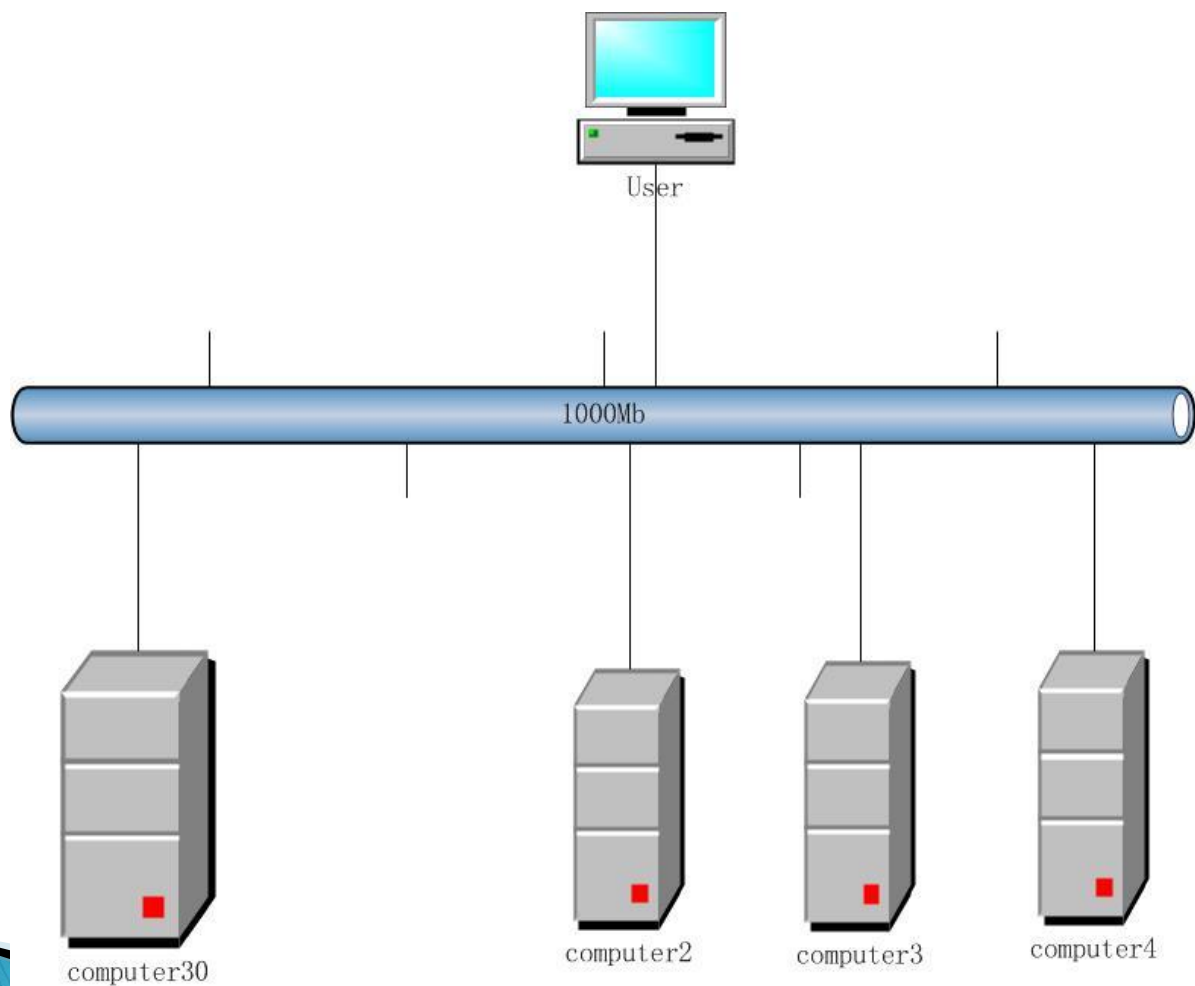
	Lustre	HDFS	GlusterFS	MooseFS	MogileFS	Ceph
单点故障			解决		解决	解决
数据可靠性	☆☆	☆☆☆☆	☆☆☆☆	☆☆	☆☆☆☆	☆☆☆☆
I/O效率	☆☆☆☆	☆☆	☆☆	☆☆	☆☆	☆☆
技术支持	☆☆	☆☆	☆☆	☆☆	☆☆	☆☆
发展潜力	☆☆	☆☆☆☆	☆☆	☆	☆☆	☆☆☆☆
POSIX兼容	是		是	是		是
S3 webservice接口						有
开发语言	C语言	Java	C语言	C语言	Perl	C++

小结

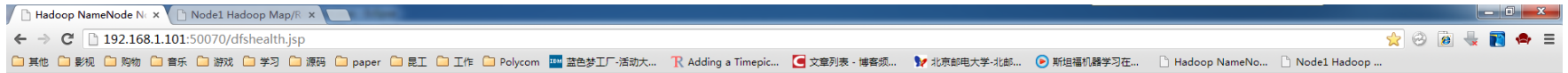
- ▶ CEPH和HDFS都有很高的数据可靠性，较快的读写效率和很好的发展潜力，比较适合我们的实际需要。

系统设计——基于Hadoop的分布式存储

分布式存储测试环境 (30TB)



HDFS



NameNode 'Node1:9000'

Started: Tue Nov 05 21:06:46 CST 2013
Version: 1.2.1, r1503152
Compiled: Mon Jul 22 15:23:09 PDT 2013 by mattf
Upgrades: There are no upgrades in progress.

[Browse the filesystem](#)
[Namenode Logs](#)

Cluster Summary

120 files and directories, 424 blocks = 544 total. Heap Size is 179.88 MB / 1.74 GB (10%)

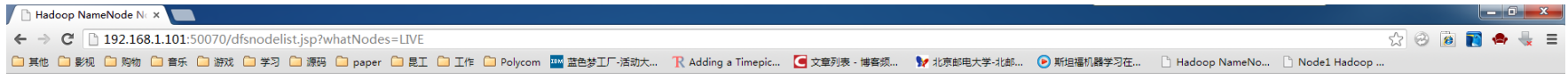
Configured Capacity	:	30.59 TB
DFS Used	:	25.95 GB
Non DFS Used	:	773.15 GB
DFS Remaining	:	29.81 TB
DFS Used%	:	0.08 %
DFS Remaining%	:	97.45 %
Live Nodes	:	4
Dead Nodes	:	0
Decommissioning Nodes	:	0
Number of Under-Replicated Blocks	:	8

NameNode Storage:

Storage Directory	Type	State
NAME_DIR	IMAGE_AND_EDITS	Active

This is [Apache Hadoop](#) release 1.2.1

HDFS



NameNode 'Node1:9000'

Started: Tue Nov 05 21:06:46 CST 2013
Version: 1.2.1, r1503152
Compiled: Mon Jul 22 15:23:09 PDT 2013 by mattf
Upgrades: There are no upgrades in progress.

[Browse the filesystem](#)
[Namenode Logs](#)
[Go back to DFS home](#)

Live Datanodes : 4

Node	Last Contact	Admin State	Configured Capacity (TB)	Used (TB)	Non DFS Used (TB)	Remaining (TB)	Used (%)	Used (%)	Remaining (%)	Blocks
Node1	1	In Service	9.3	0.01	0.21	9.08	0.08	<input type="text"/>	97.71	140
Node2	0	In Service	9.31	0.01	0.21	9.09	0.08	<input type="text"/>	97.71	123
Node3	2	In Service	9.31	0	0.21	9.1	0.05	<input type="text"/>	97.74	85
Node4	2	In Service	2.68	0.01	0.14	2.54	0.22	<input type="text"/>	94.63	103

This is Apache Hadoop release 1.2.1

文件上传

- ▶ HDFS适合大文件存储
 - 智能分析上传目录，组合成大文件
 - 标准化fits格式
- ▶ 现阶段稳定版本HDFS不支持Append操作
 - 以“.+数字”的后缀区分同一天数据的不同上传操作
- ▶ 优化处理
 - 上传过程中，提取fits头信息生成MySQL记录
 - 真实上传前检查MySQL记录，查看该文件是否已经上传

HDFS原始数据格式

Contents of directory /warehouse

Goto :

[Go to parent directory](#)

Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	Group
FULL_20130716_14713920.0	file	687.58 MB	1	64 MB	2013-10-12 11:02	rw-r--r--	hadoop	supergroup
FULL_20130720_14713920.0	file	1.06 GB	1	64 MB	2013-10-12 11:02	rw-r--r--	hadoop	supergroup
FULL_20130819_14713920.0	file	238.55 MB	1	64 MB	2013-10-12 11:02	rw-r--r--	hadoop	supergroup
FULL_20130820_14713920.0	file	56.13 MB	1	64 MB	2013-10-12 11:02	rw-r--r--	hadoop	supergroup
FULL_20130821_14713920.0	file	14.03 MB	1	64 MB	2013-10-12 11:02	rw-r--r--	hadoop	supergroup
FULL_20130830_14713920.0	file	1.25 GB	1	64 MB	2013-10-12 11:02	rw-r--r--	hadoop	supergroup
FULL_20130831_14713920.0	file	2.86 GB	1	64 MB	2013-10-06 16:13	rw-r--r--	hadoop	supergroup
FULL_20130926_14713920.0	file	8 GB	1	64 MB	2013-10-06 16:13	rw-r--r--	hadoop	supergroup
FULL_20131008_14713920.0	file	1 GB	1	64 MB	2013-10-12 11:02	rw-r--r--	hadoop	supergroup
FULL_20131010_14713920.0	file	1.26 GB	1	64 MB	2013-10-12 11:02	rw-r--r--	hadoop	supergroup
FULL_20131011_14713920.0	file	4.43 GB	1	64 MB	2013-10-12 11:02	rw-r--r--	hadoop	supergroup
REGION_20130720_267840.0	file	78.67 MB	1	64 MB	2013-10-12 11:03	rw-r--r--	hadoop	supergroup

[Go back to DFS home](#)

Local logs

[Log directory](#)

This is [Apache Hadoop](#) release 1.2.1

85% 0.7K/S
66.8K/S

文件下载

- ▶ 通过网页形式展现下载接口
- ▶ 根据数据库记录进行对数据进行查询
- ▶ 得到原始数据文件名，及fits文件的偏移量，通过HDFS接口下载对应部分数据
- ▶ 通过Applet提供多文件实时下载，可以同时部署多个站点，实现用户下载接口的负载均衡

用户访问接口

The screenshot displays the HSOS Full Magnetic Fits web interface. The browser address bar shows the URL `192.168.1.4:8080/corona/`. The main content area features a welcome message: "Welcome to HSOS data storage system!" and a sub-header: "You can click files in file list to see details of this file, or double click to download it on your computer." Below this is a 2x4 grid of solar images. The left sidebar contains a file list under "FULL-DISK SOLAR-IMAGE" and "LOCAL-DISK SOLAR-IMAGE". The right sidebar shows a "General" metadata panel with fields for Obs Time, Width, Height, Wave Leng..., Gain, Dark Level, Seeing, Size, Bitpix, Comments, and Origin. At the bottom, there is a table with columns for File Path, Observation Time, Resolution X, Resolution Y, Gain, Dark Level, Exposure, Seeing, Origin, and Comments. The table is currently empty. The bottom status bar shows page navigation (1 of 10,000) and a copyright notice.

General

Obs Time: 2013-07-16 09:31:33

Width: 2712

Height: 2712

Wave Leng...: 6563 + 0.000 A

Gain: 458

Dark Level: 0

Seeing: 0

Size: 14713920

Bitpix: 16

Comments: null

Origin: Full Disk Ha Telesco

Additional Properties

No additional properties available yet!

File Path	Observation Time	Resolution X	Resolution Y	Gain	Dark Level	Exposure	Seeing	Origin	Comments

1 of 10,000

Copyright (c) 2007 MySite - <http://www.mysite.com>

用户访问接口

192.168.1.4:8080/corona/

FULL-DISK SOLAR-IMAGE

- 2013-07-16 (49)
- 2013-07-20 (77)
- 2013-08-19 (17)
- 2013-08-20 (4)
- 2013-08-21 (1)
- 2013-08-30 (91)
- 2013-08-31 (209)

HrHa130820015204full.

HrHa130820015704full.

HrHa130820020204full.

HrHa130820020704full.

HrHa130831041238full.

HrHa130831041241full.

HrHa130831041244full.

HrHa130831041247full.

HrHa130831041250full.

HrHa130831041253full.

HrHa130831041256full.

HrHa130831041259full.

Welcome to HSOS data storage system!

You can click files in file list to see details of this file, or double click to download it on your computer.

General

Obs Time: 2013-08-31 04:12:41.

Width: 2712

Height: 2712

Wave Leng": 6563 + 0.000 A

Gain: 0

Dark Level: 0

Seeing: 0

Size: 14713920

Bitpix: 16

Comments: null

Origin: Full Disk Ha Telesco

Additional Properties

AVG Energy: 1280.71

File Path	Observation Time	Resolution X	Resolution Y	Gain	Dark Level	Exposure	Seeing	Origin	Comments

<< 1 >>

1-50 of 10,000

用户访问接口

The screenshot displays the HSOS Full Magnetic Fits web interface. A central dialog box titled "Fits File Search" is open, allowing users to filter files based on Resolution Range, Observation Time Range, and Instrument Parameter Range. The background shows a file tree on the left and a data table at the bottom.

Fits File Search Dialog:

- Resolution Range: resolution X: [] ~ [], resolution Y: [] ~ []
- Observation Time Range: DateTime: 2013-8-1 ~ 2013-11-1
- Instrument Parameter Range: Gain: [] ~ [], Dark Level: [] ~ [], Exposure: [] ~ []
- Seeing(0 is most ...): []
- Buttons: Reset, Search

Data Table:

File Path	Observation Time	Resolution X	Resolution Y	Gain	Dark Level	Exposure	Seeing	Origin	Comments
HrHa130831041238full.fit	2013-08-31 04:12:38.0	2712	2712	0	0	200	0	Full Disk Ha Telescope	
HrHa130926065730full.fit	2013-09-26 06:57:30.0	2712	2712	1004	0	100	0	Full Disk Ha Telescope	
HrHa130831041241full.fit	2013-08-31 04:12:41.0	2712	2712	0	0	200	0	Full Disk Ha Telescope	
HrHa130831041244full.fit	2013-08-31 04:12:44.0	2712	2712	0	0	200	0	Full Disk Ha Telescope	
HrHa130926065733full.fit	2013-09-26 06:57:33.0	2712	2712	1004	0	100	0	Full Disk Ha Telescope	
HrHa130831041247full.fit	2013-08-31 04:12:47.0	2712	2712	0	0	200	0	Full Disk Ha Telescope	
HrHa130926065734full.fit	2013-09-26 06:57:34.0	2712	2712	1004	0	100	0	Full Disk Ha Telescope	

General Properties:

- Obs Time: 2013-08-31 04:12:41
- Width: 2712
- Height: 2712
- Wave Leng: 6563 + 0.000 A
- Gain: 0
- Dark Level: 0
- Seeing: 0
- Size: 14713920
- Bitpix: 16
- Comments: null
- Origin: Full Disk Ha Telesco
- AWG Energy: 1280.71

Page Footer:

Copyright (c) 2007 MySite - http://www.mysite.com

用户访问接口

192.168.1.4:8080/corona/

FULL-DISK SOLAR IMAGE

Welcome to HSOS data storage system!

You can click files in file list to see details of this file, or double click to download it on your computer.

General

Obs Time: 2013-08-31 04:12:41

Width: 2712

Height: 2712

Wave Leng...: 6563 + 0.000 A

Gain: 0

Dark Level: 0

Seeing: 0

Size: 14713920

Bitpix: 16

Comments: null

Origin: Full Disk Ha Telesco

Additional Properties

AWG Energy: 1280.71

File Path	Observation Time	Resolution X	Resolution Y	Gain	Dark Level	Exposure	Seeing	Origin	Comments
HrHa130831041238full.fit	2013-08-31 04:12:38.0	2712	2712	0	0	200	0	Full Disk Ha Telescope	
HrHa130926065730full.fit	2013-09-26 06:57:30.0	2712	2712	1004	0	100	0	Full Disk Ha Telescope	
HrHa130831041241full.fit	2013-08-31 04:12:41.0	2712	2712	0	0	200	0	Full Disk Ha Telescope	
HrHa130831041244full.fit	2013-08-31 04:12:44.0	2712	2712	0	0	200	0	Full Disk Ha Telescope	
HrHa130926065733full.fit	2013-09-26 06:57:33.0	2712	2712	1004	0	100	0	Full Disk Ha Telescope	
HrHa130831041247full.fit	2013-08-31 04:12:47.0	2712	2712	0	0	200	0	Full Disk Ha Telescope	
HrHa130926065734full.fit	2013-09-26 06:57:34.0	2712	2712	1004	0	100	0	Full Disk Ha Telescope	

4 selected

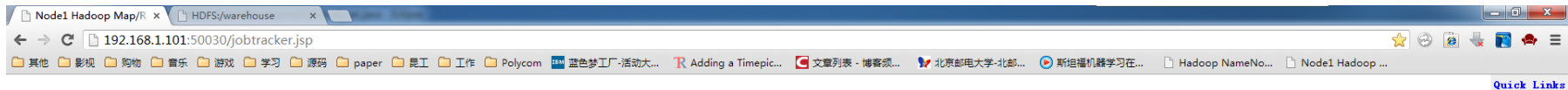
<< 1 >>

1-50 of 1,394

内容提取

- ▶ 使用MapReduce进行分布式计算，可以对全部或者部分文件进行图像内容分析，并把得到的结果插入数据库，供后期进行基于内容的检索项

分布式计算



Node1 Hadoop Map/Reduce Administration

State: RUNNING
Started: Tue Nov 05 21:06:48 CST 2013
Version: 1.2.1, r1503152
Compiled: Mon Jul 22 15:23:09 PDT 2013 by mattf
Identifier: 201311052106
SafeMode: OFF

Cluster Summary (Heap Size is 167.56 MB/1.74 GB)

Running Map Tasks	Running Reduce Tasks	Total Submissions	Nodes	Occupied Map Slots	Occupied Reduce Slots	Reserved Map Slots	Reserved Reduce Slots	Map Task Capacity	Reduce Task Capacity	Avg. Tasks/Node	Blacklisted Nodes	Graylisted Nodes	Excluded Nodes
8	0	1	4	8	0	0	0	8	8	4.00	0	0	0

Scheduling Information

Queue Name	State	Scheduling Information
default	running	N/A

Filter (Jobid, Priority, User, Name)

Example: 'user:smith 3200' will filter by 'smith' only in the user field and '3200' in all fields

Running Jobs

Jobid	Started	Priority	User	Name	Map % Complete	Map Total	Maps Completed	Reduce % Complete	Reduce Total	Reduces Completed	Job Scheduling Information	Diagnostic Info
job_201311052106_0002	Tue Nov 05 22:05:03 CST 2013	NORMAL	hadoop	Fits Analyze: hdfs://192.168.1.101:9000/warehouse/FULL_20130716_14713920.0	<div style="width: 15.38%;"><div style="background-color: #0070C0; height: 10px;"></div></div> 15.38%	39	6	<div style="width: 0.00%;"><div style="background-color: #0070C0; height: 10px;"></div></div> 0.00%	0	0	NA	NA

Retired Jobs

[none](#)

Local Logs

[Log directory](#), [Job Tracker History](#)

This is [Apache Hadoop](#) release 1.2.1



MapRed Log

- 13/11/05 22:05:26 WARN mapred.JobClient: Use GenericOptionsParser for parsing the arguments. Applications should implement Tool for the same.
- 13/11/05 22:05:27 INFO input.FileInputFormat: Total input paths to process : 4
- 13/11/05 22:05:28 INFO mapred.JobClient: Running job: job_201311052106_0002
- 13/11/05 22:05:29 INFO mapred.JobClient: map 0% reduce 0%
- 13/11/05 22:05:38 INFO mapred.JobClient: map 5% reduce 0%
- 13/11/05 22:05:41 INFO mapred.JobClient: map 15% reduce 0%
- 13/11/05 22:05:42 INFO mapred.JobClient: map 20% reduce 0%
- 13/11/05 22:05:43 INFO mapred.JobClient: map 25% reduce 0%
- 13/11/05 22:05:45 INFO mapred.JobClient: map 30% reduce 0%
- 13/11/05 22:05:46 INFO mapred.JobClient: map 35% reduce 0%
- 13/11/05 22:05:47 INFO mapred.JobClient: map 43% reduce 0%
- 13/11/05 22:05:48 INFO mapred.JobClient: map 46% reduce 0%
- 13/11/05 22:05:49 INFO mapred.JobClient: map 51% reduce 0%
- 13/11/05 22:05:51 INFO mapred.JobClient: map 53% reduce 0%
- 13/11/05 22:05:52 INFO mapred.JobClient: map 64% reduce 0%
- 13/11/05 22:05:53 INFO mapred.JobClient: map 66% reduce 0%
- 13/11/05 22:05:54 INFO mapred.JobClient: map 74% reduce 0%
- 13/11/05 22:05:55 INFO mapred.JobClient: map 76% reduce 0%
- 13/11/05 22:05:56 INFO mapred.JobClient: map 84% reduce 0%
- 13/11/05 22:05:57 INFO mapred.JobClient: map 92% reduce 0%
- 13/11/05 22:05:58 INFO mapred.JobClient: map 97% reduce 0%
- 13/11/05 22:05:59 INFO mapred.JobClient: map 100% reduce 0%
- 13/11/05 22:06:00 INFO mapred.JobClient: Job complete: job_201311052106_0002
- 13/11/05 22:06:01 INFO mapred.JobClient: Counters: 19
- 13/11/05 22:06:01 INFO mapred.JobClient: Job Counters
- 13/11/05 22:06:01 INFO mapred.JobClient: SLOTS_MILLIS_MAPS=186378
- 13/11/05 22:06:01 INFO mapred.JobClient: Total time spent by all reduces waiting after reserving slots (ms)=0
- 13/11/05 22:06:01 INFO mapred.JobClient: Total time spent by all maps waiting after reserving slots (ms)=0
- 13/11/05 22:06:01 INFO mapred.JobClient: Rack-local map tasks=2
- 13/11/05 22:06:01 INFO mapred.JobClient: Launched map tasks=39
- 13/11/05 22:06:01 INFO mapred.JobClient: Data-local map tasks=37
- 13/11/05 22:06:01 INFO mapred.JobClient: SLOTS_MILLIS_REDUCES=0
- 13/11/05 22:06:01 INFO mapred.JobClient: File Output Format Counters
- 13/11/05 22:06:01 INFO mapred.JobClient: Bytes Written=0
- 13/11/05 22:06:01 INFO mapred.JobClient: FileSystemCounters
- 13/11/05 22:06:01 INFO mapred.JobClient: HDFS_BYTES_READ=2162953338
- 13/11/05 22:06:01 INFO mapred.JobClient: FILE_BYTES_WRITTEN=2207217
- 13/11/05 22:06:01 INFO mapred.JobClient: File Input Format Counters
- 13/11/05 22:06:01 INFO mapred.JobClient: Bytes Read=2162946240
- 13/11/05 22:06:01 INFO mapred.JobClient: Map-Reduce Framework
- 13/11/05 22:06:01 INFO mapred.JobClient: Map input records=147
- 13/11/05 22:06:01 INFO mapred.JobClient: Physical memory (bytes) snapshot=5668392960
- 13/11/05 22:06:01 INFO mapred.JobClient: Spilled Records=0
- 13/11/05 22:06:01 INFO mapred.JobClient: CPU time spent (ms)=104180
- 13/11/05 22:06:01 INFO mapred.JobClient: Total committed heap usage (bytes)=4591779840
- 13/11/05 22:06:01 INFO mapred.JobClient: Virtual memory (bytes) snapshot=37774553088
- 13/11/05 22:06:01 INFO mapred.JobClient: Map output records=0
- 13/11/05 22:06:01 INFO mapred.JobClient: SPLIT_RAW_BYTES=7098

Related Work

基于Ceph的分布式存储系统

管理员接口

S3 Browser 3-8-7 - Free Version (for non-commercial use only)

Accounts Buckets Files Tools Upgrade to Pro! Help

Path: / 1994/ 05/ 05/

New bucket

- 35Mag
- 35mag
- my-second-bucket
- my-third-bucket

File	Size	Type	Last Modified	Storage Class
..				
image0.fits	10.00 MB	FITS 文件	2013/5/10 18:15:42	STANDARD
image1.fits	10.00 MB	FITS 文件	2013/5/10 18:15:49	STANDARD
image10.fits	10.00 MB	FITS 文件	2013/5/10 18:16:55	STANDARD
image11.fits	10.00 MB	FITS 文件	2013/5/10 18:17:02	STANDARD
image12.fits	10.00 MB	FITS 文件	2013/5/10 18:17:11	STANDARD
image13.fits	10.00 MB	FITS 文件	2013/5/10 18:17:23	STANDARD
image14.fits	10.00 MB	FITS 文件	2013/5/10 18:17:29	STANDARD
image15.fits	10.00 MB	FITS 文件	2013/5/10 18:17:33	STANDARD
image16.fits	10.00 MB	FITS 文件	2013/5/10 18:17:38	STANDARD
image17.fits	10.00 MB	FITS 文件	2013/5/10 18:17:44	STANDARD
image18.fits	10.00 MB	FITS 文件	2013/5/10 18:17:53	STANDARD
image19.fits	10.00 MB	FITS 文件	2013/5/10 18:17:58	STANDARD
image2.fits	10.00 MB	FITS 文件	2013/5/10 18:15:56	STANDARD
image20.fits	10.00 MB	FITS 文件	2013/5/10 18:18:00	STANDARD
image21.fits	10.00 MB	FITS 文件	2013/5/10 18:18:05	STANDARD
image22.fits	10.00 MB	FITS 文件	2013/5/10 18:18:11	STANDARD
image23.fits	10.00 MB	FITS 文件	2013/5/10 18:18:15	STANDARD

Download Upload Delete New Folder Refresh

100 files (1000.00 MB)

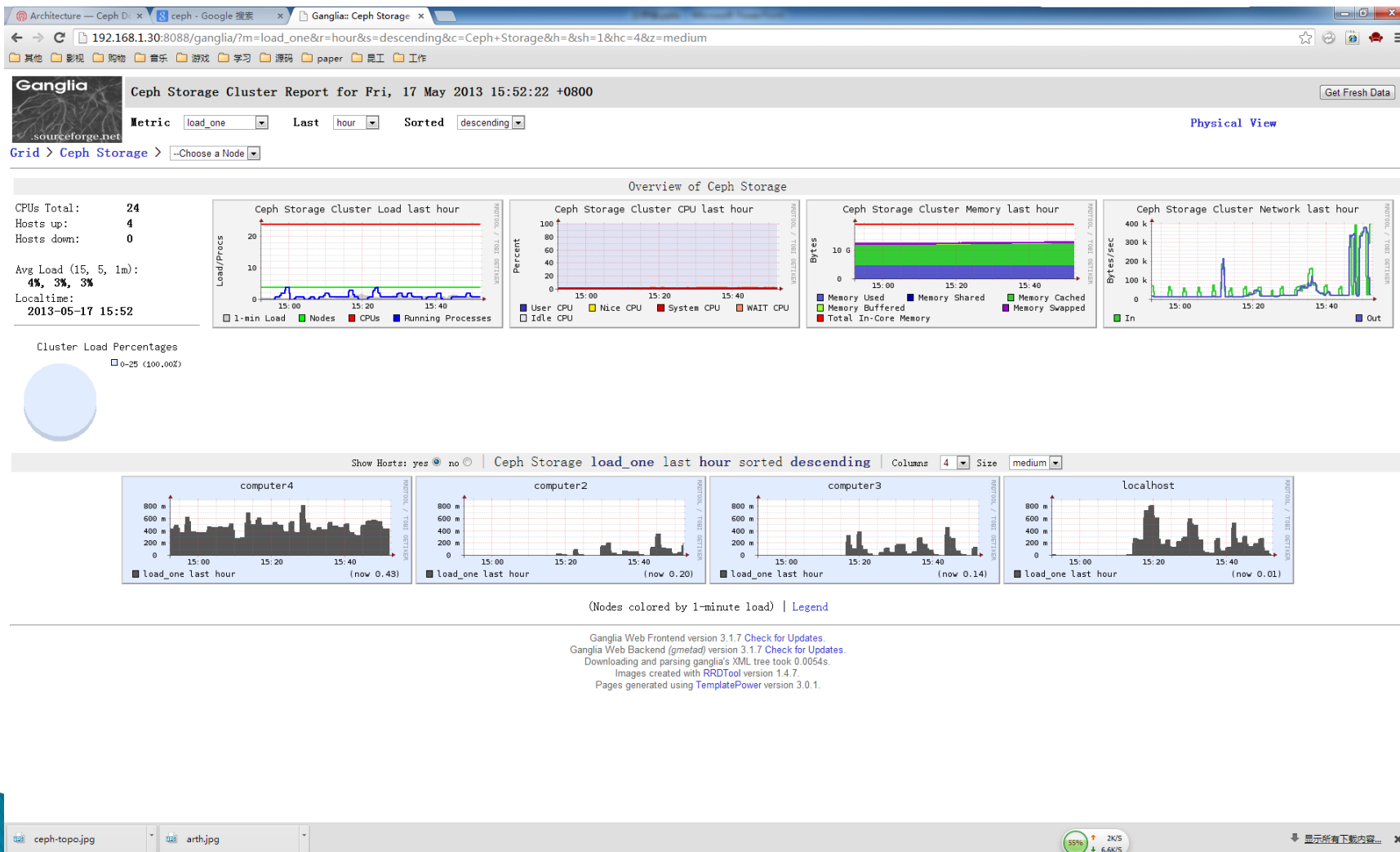
Tasks Permissions **Http Headers** Properties Preview Versions EventLog

URL: <http://192.168.1.30/35mag/1994/05/05/image13.fits> Copy

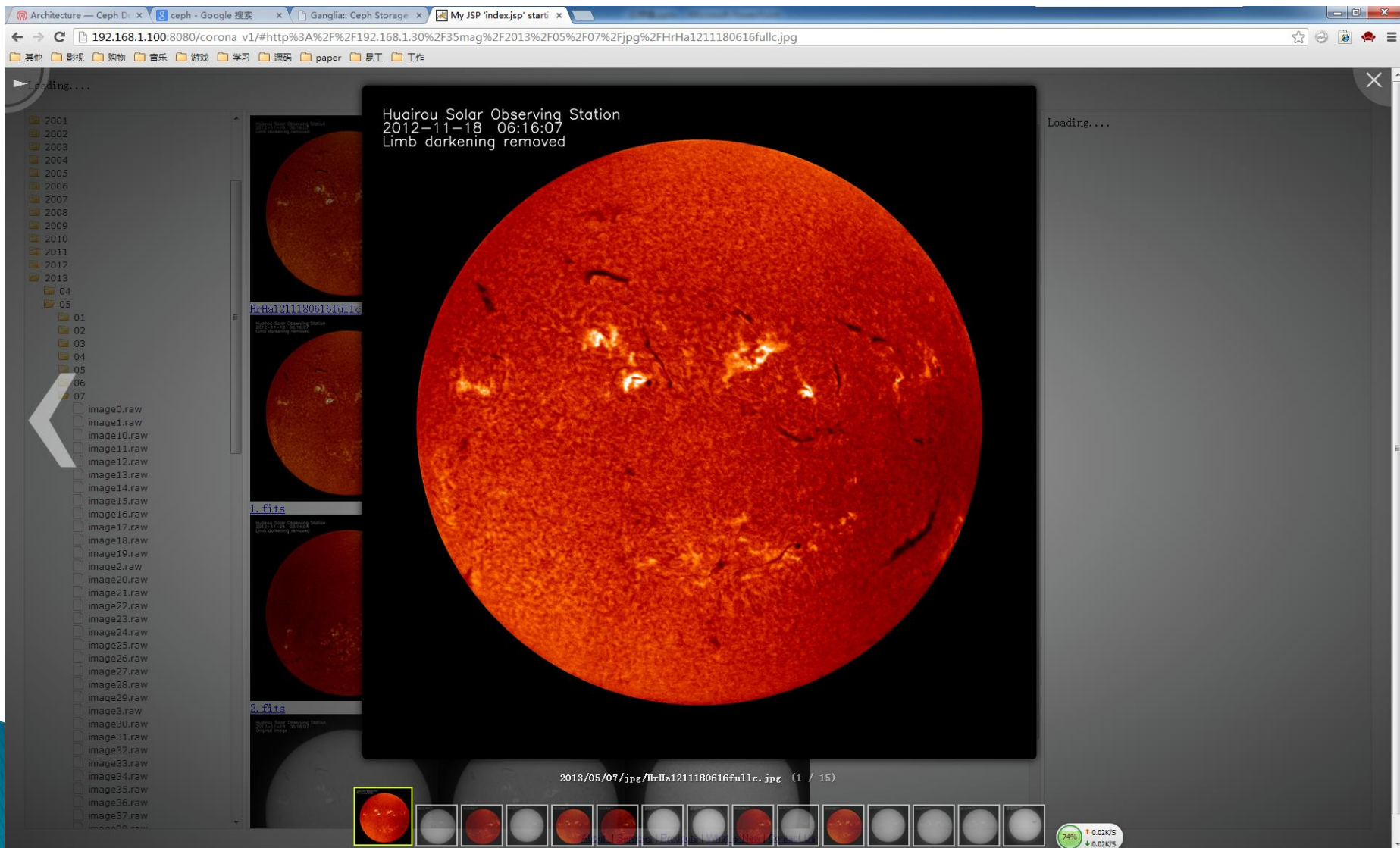
User Name	Full Control	Read	Write	Read Permissions	Write Permissions
Owner (s3user)	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Authenticated Users	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
All Users	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Make public Make private More.. Apply changes Reload

集群监视系统



用户界面



小结

- ▶ 基本解决了开篇提出的问题
 - 构建了一个总容量30TB的可用分布式存储环境
 - 可以在分布式文件系统中上传下载数据
 - 用户可以通过网页浏览/下载数据
- ▶ 相关实验

集群的动态扩容	OK
ftp的动态扩容	OK
集群机器的关机重启	OK
集群单个机器的突然断电	OK
集群全部机器的突然断电	OK

以后的工作

- ▶ 将观测站的数据同步到云中
 - 通过公有云进行同步（3.5GB/8H）
- ▶ 系统性能优化
- ▶ 加入更多的真实数据

总结

- ▶ 提供了针对天文数据特殊性进行优化的大规模数据的存储方案
- ▶ 系统可以实时批量上传和下载大规模天文数据
- ▶ 系统可以基于fits头信息检索，同时提供动态增加对图像内容分析的索引项
- ▶ 设计了分布式图像内容处理框架

谢谢

DISCUSSION