





Making Hadoop Easy for a Growing Community

Enabling Big Data for Everyone...

Christophe Bisciglia
Founder christophe@cloudera.com

Overview

Getting Started...

- A Brief History of Hadoop
- Challenges for Existing Users
 - Making Hadoop Easier for Existing Users
- Welcoming New Users to the Hadoop Community
- Challenges for New Users
 - Making Hadoop Easier for New Users
- Putting it All Together
- Next Steps for Cloudera

Growing Up with Hadoop

You've come a long way baby...



Growing Up with Hadoop

You've come a long way baby...

- Early Days
 - 2004: Google Publishes MapReduce/GFS
 - 2005: Hadoop Prototype
 - Doug Cutting and Mike Cafarella
 - 2006: Hadoop Running on 20 nodes
 - Internet Archive and UW



Doug Cutting

Photo Credit: New York Times

Growing Up with Hadoop

You've come a long way baby...

- Formative Years

- 2006: Yahoo! Begins Major Investment
- 2007: Yahoo! Runs Hadoop on 2000 nodes
- 2008: Yahoo! uses Hadoop to claim Terasort Benchmark



Growing Up with Hadoop

You've come a long way baby...

- 3 Major Releases for Hadoop in last year
 - More Reliable
 - More Scalable
 - More Manageable

Growing Up with Hadoop

You've come a long way baby...

- New Sub-Projects Embrace New Users
 - Hive: SQL Data Warehouse for Hadoop
 - Pig: Data Analysis Language



Growing Up with Hadoop

You've come a long way baby...

- Sqoop: Database import for Hadoop
 - Developed by Aaron Kimball, Cloudera
 - Works over JDBC
 - Extensible for better performance

Growing Up with Hadoop

You've come a long way baby...

- RDBMS Vendors Embrace Hadoop
 - MapReduce is great for Analytics
 - Hadoop is the MapReduce Standard

The Vertica logo, consisting of the word "VERTICA" in a bold, blue, sans-serif font with a stylized blue outline around the letters. **VERTICA** integrates directly with Hadoop

Growing Up with Hadoop

You've come a long way baby...

- Adoption Spanning Globe
 - HUGs outside the US
 - Over 10x Companies “PoweredBy”
 - Not Just for Web Companies Anymore

Challenges for Existing Users

What Cloudera has Learned from the Community

- Building Hadoop is Hard
 - Project Split Still Causing Problems
 - Deploying Hadoop is Hard
 - Which Version / Release ? When do you Upgrade?
 - Operating and Administering Hadoop is Hard
 - Lack of Integration with Standard Tools
 - Explaining Hadoop to New Users is Hard
 - Who has had trouble explaining Hadoop to their Manager ?
 - Developing Applications for Hadoop should be Easier
 - Shifting APIs, Difficult Client Configuration, No General UI Tools
-

Build, Deploy, Upgrade, and Operate Now a whole lot easier...



- Cloudera's Distribution for Hadoop (Apache 2 Licensed)
 - Focused on Usability for Operators / Administrators
 - Uses Standard Tools for Packaging, Deployment, Operation, Upgrades, etc
- Stability or New Features: Your Choice
 - Stable Release: Time Tested and Widely Used in Production
 - Testing Release: New Features, Faster Updates
- Write Once, Run Anywhere
 - Cloudera's Distribution Runs on Redhat, Ubuntu, EC2, and more

Cloudera's Distribution for Hadoop (CDH)

Current State of the World...(November 2009)

Current Stable Release: CDH1

Hadoop Base Version(s)

Hadoop 0.18.3

Hive 0.2.0

Pig 0.3.0



Current Testing Release: CDH2

Hadoop Base Version(s)

Hadoop 0.18.3 + 76 Patches

Hadoop 0.20.1 + 152 Patches

Hive 0.4.0 + 14 Patches

Pig 0.4.99 + 7 Patches (5.0 soon)

HBase 0.20.0 (custom build)

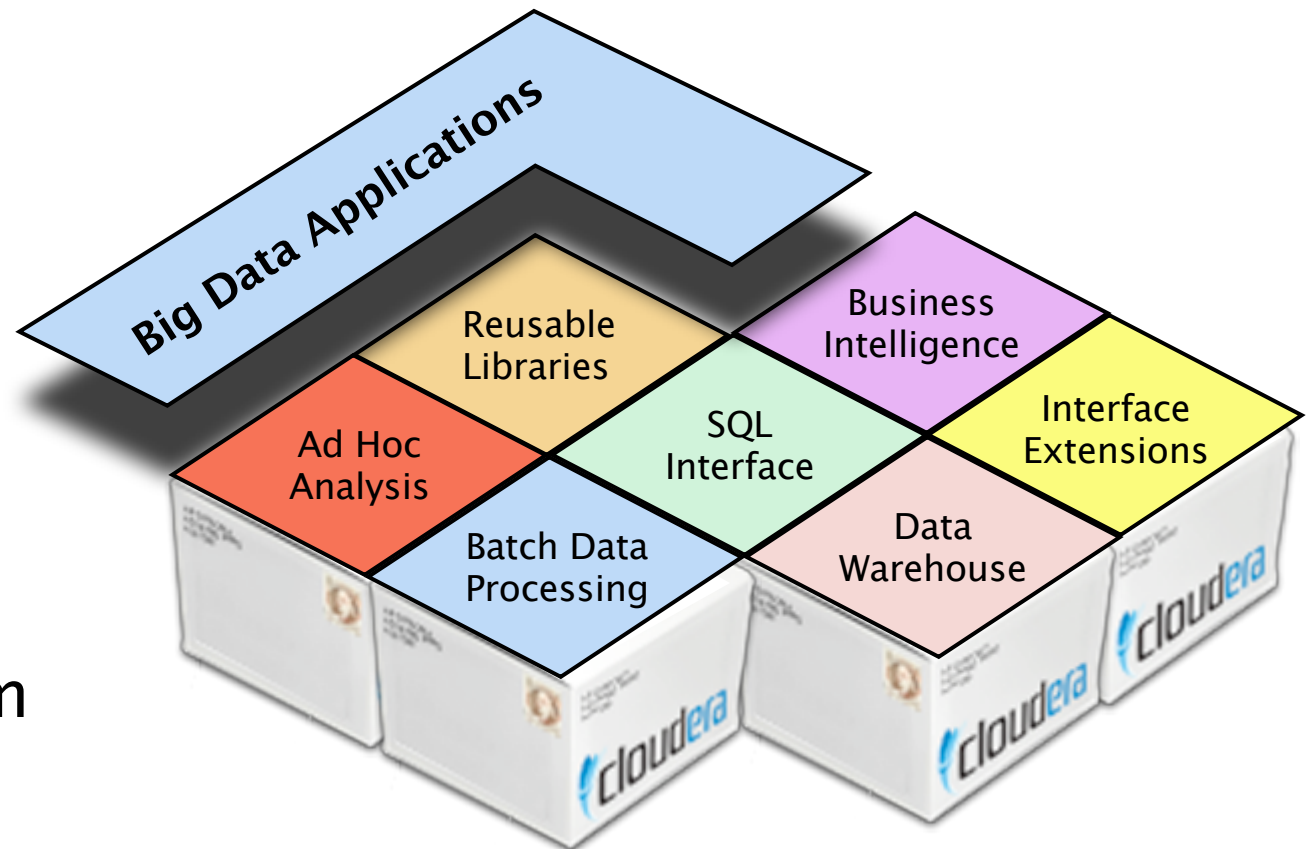
Zookeeper 3.2.1 + 7 Patches



Building an Ecosystem Around Hadoop

Standard Packages Enable Vertical Development

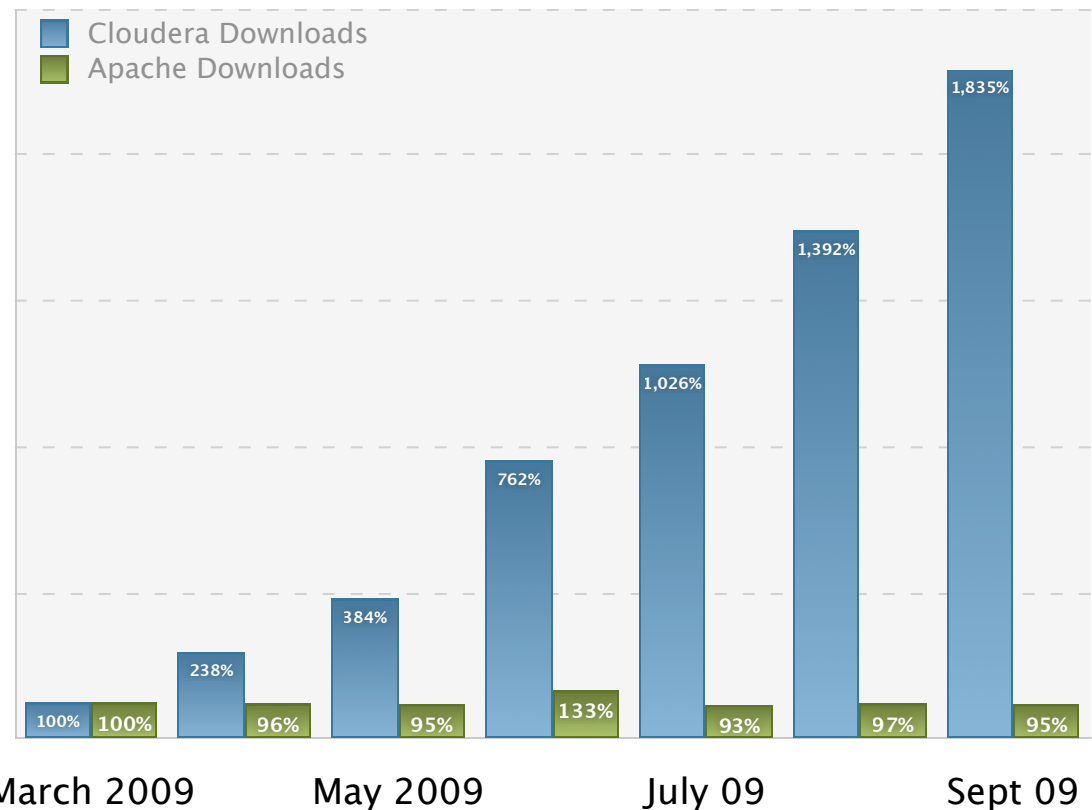
- Open Source
- Modular
- Extensible
- Cross Platform



Standard Packaging Drives Adoption

Cloudera's Distribution Drives New Usage...

- Consistent Downloads from Apache
- Cloudera Packages Drive **New Usage**
- 75% of new users get Hadoop from Cloudera
- Enables New Hadoop Applications



Normalized by unique users accessing hadoop.apache.org/core/releases.html and Cloudera Package Repositories in March 2009. 75% is an estimate based related data. Exact measurement is impossible.

Hadoop's Emerging Community

Not Just Web Companies and Developers Anymore...

Web Properties



Petabyte Scale Data Platform
Faster Iteration, Greater Productivity
Deep Insight into User Behavior

Telecom



Collect All Network Logs (CDR, etc)
Analyze Customer/Network Behavior
Plan for Capacity and Expansion

Utilities



Collect and Analyze Sensor Data
Improve Smart Grid Efficiency
Analytics Shared Across Regions

Financial Services



Build Scalable Risk Models
Leverage More Data for Better Real Time Decisions

Retail Analytics



Recommendations and Micro Targeting
Analyze Buying Behavior / Trends

Biotechnology



Drug Discovery, Genomic Analysis, Sequence Alignment
Improving Researcher Productivity

Hadoop's Emerging Community

Not Just Web Companies and Developers Anymore...

Telecom



- Collect All Network Logs (CDR, etc)
- Analyze Customer/Network Behavior
- Plan for Capacity and Expansion

Utilities



- Collect and Analyze Sensor Data
- Improve Smart Grid Efficiency
- Analytics Shared Across Regions

Financial Services



- Build Scalable Risk Models
- Leverage More Data for Better Real Time Decisions

Biotechnology



- Drug Discovery, Genomic Analysis, Sequence Alignment
- Improving Researcher Productivity

Common Characteristics of New Hadoop Users

- Hadoop Enters Organization Through a Group of Evangelists
- Organizations are New to Web-Scale Infrastructure
- Operations Team Manages a Diverse Set of Systems
- Stability is a Higher Priority
- Hadoop Must Integrate with Existing, Often Legacy, Enterprise Software
- Less Technical Analysts Need Access to Data Stored in Hadoop**

Challenges for New Users

What Cloudera has Learned from Enterprise

- **Stability is More Important than New Features**
 - Cloudera now has Stable and Testing Releases
 - **Integration with Existing Systems is Critical**
 - Enterprises have Large Investments in Existing Systems
 - These Systems are much more Expensive than Hadoop, which leaves a Lot of Room for Leverage.
 - **Linux is Less Common than We Imagine**
 - Most Enterprise Clients Run Windows
 - Accessing Hadoop from Windows is Critical (but can still run Linux on servers)
 - **Enabling Less Technical Users to Access Hadoop is Critical**
 - Cannot require a developer to write every query / question
-

Putting it All Together...

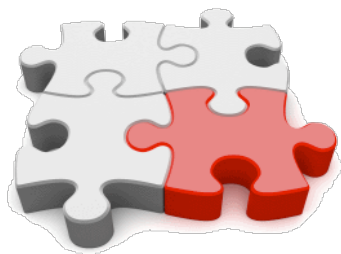
Cloudera is a Software Company. Apache Hadoop is the Core of our Business. The Communities Success Drives our Own...



Cloudera's Distribution for Hadoop
Simplifies Hadoop Deployment and Management



Cloudera's Integration Tools
Tools Like Sqoop Make Working with Existing Data Easy. More Features, Tools, and Extensions to Come



Cloudera's Professional Services
Helps Enterprises Get Moving Quickly with:
Design Consulting, Training, Certification, Support and More



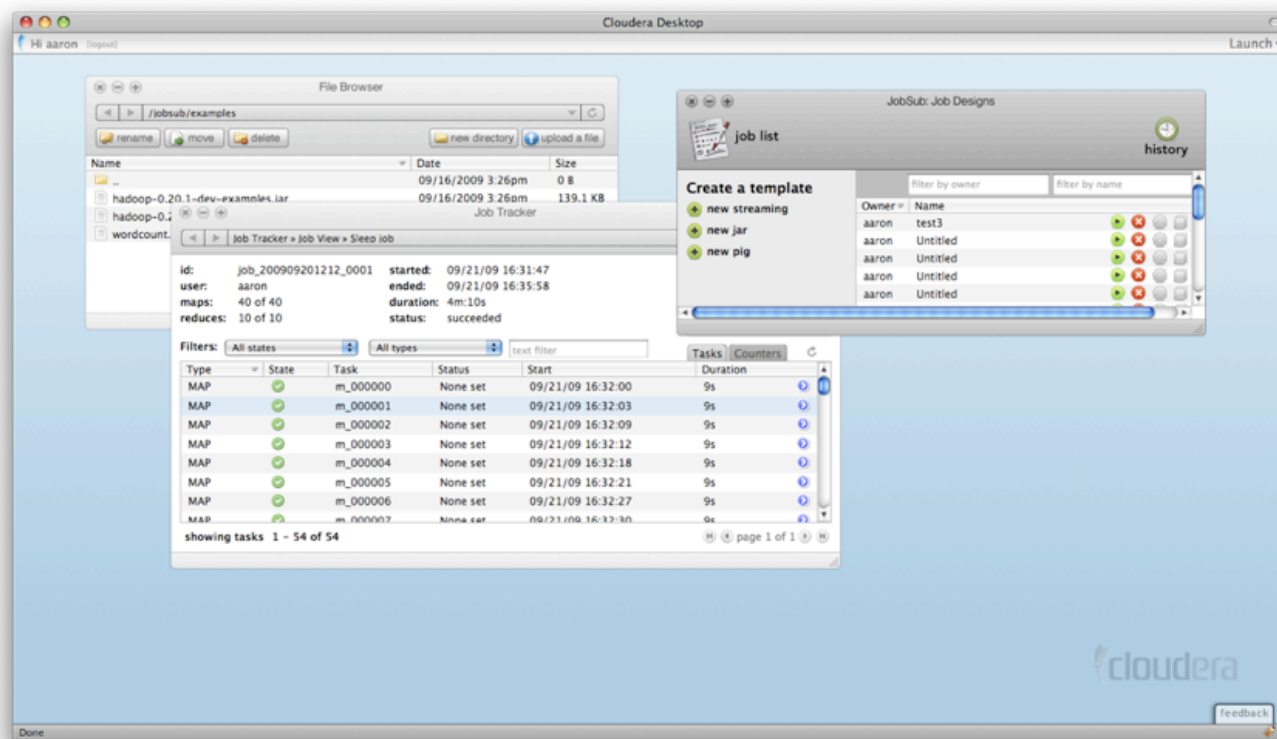
Cloudera Desktop
Enables Users and Operators to Access Hadoop with
Just a Web Browser. Also Provides an Extensible
Application Development Framework.

Looking More Closely at Desktop

Making Hadoop Easy for Everyone...



Looking More Closely at Desktop Making Hadoop Easy for Everyone...

The screenshot shows the Cloudera Desktop interface with several components:

- File Browser:** Located at the top left, showing a directory structure with files like 'hadoop-0.20.1-dfu-examples.jar' and 'wordcount'.
- Job Tracker > Job View > Sleep job:** A central panel displaying job details:
 - id: job_200909201212_0001
 - user: aaron
 - maps: 40 of 40
 - reduces: 10 of 10
 - started: 09/21/09 16:31:47
 - ended: 09/21/09 16:35:58
 - duration: 4m:10s
 - status: succeeded
- JobSub: Job Designs:** A panel on the right with a 'job list' and 'history' section. It includes a 'Create a template' menu with options like 'new streaming', 'new jar', and 'new pig'. Below is a table of job designs:

Owner	Name	Actions
aaron	test3	[Icons]
aaron	Untitled	[Icons]
aaron	Untitled	[Icons]
aaron	Untitled	[Icons]
aaron	Untitled	[Icons]
- Tasks Table:** A table at the bottom showing a list of tasks:

Type	State	Task	Status	Start	Duration
MAP	✓	m_000000	None set	09/21/09 16:32:00	9s
MAP	✓	m_000001	None set	09/21/09 16:32:03	9s
MAP	✓	m_000002	None set	09/21/09 16:32:09	9s
MAP	✓	m_000003	None set	09/21/09 16:32:12	9s
MAP	✓	m_000004	None set	09/21/09 16:32:18	9s
MAP	✓	m_000005	None set	09/21/09 16:32:21	9s
MAP	✓	m_000006	None set	09/21/09 16:32:27	9s
MAP	✓	m_000007	None set	09/21/09 16:32:30	9s



Looking More Closely at Desktop Making Hadoop Easy for Everyone...



**Cloudera Desktop Runs
Entirely Inside the Web
Browser.**

Cloudera Desktop

File Browser

Job Tracker

JobSub: Job Designs

Job list

Create a template

- new streaming
- new jar
- new pig

Owner	Name
aaron	test3
aaron	Untitled
aaron	Untitled
aaron	Untitled
aaron	Untitled

Filters: All states All types text filter

Type	State	Task	Status	Start	Duration
MAP	✓	m_000000	None set	09/21/09 16:32:00	9s
MAP	✓	m_000001	None set	09/21/09 16:32:03	9s
MAP	✓	m_000002	None set	09/21/09 16:32:09	9s
MAP	✓	m_000003	None set	09/21/09 16:32:12	9s
MAP	✓	m_000004	None set	09/21/09 16:32:18	9s
MAP	✓	m_000005	None set	09/21/09 16:32:21	9s
MAP	✓	m_000006	None set	09/21/09 16:32:27	9s
MAP	✓	m_000007	None set	09/21/09 16:32:30	9s

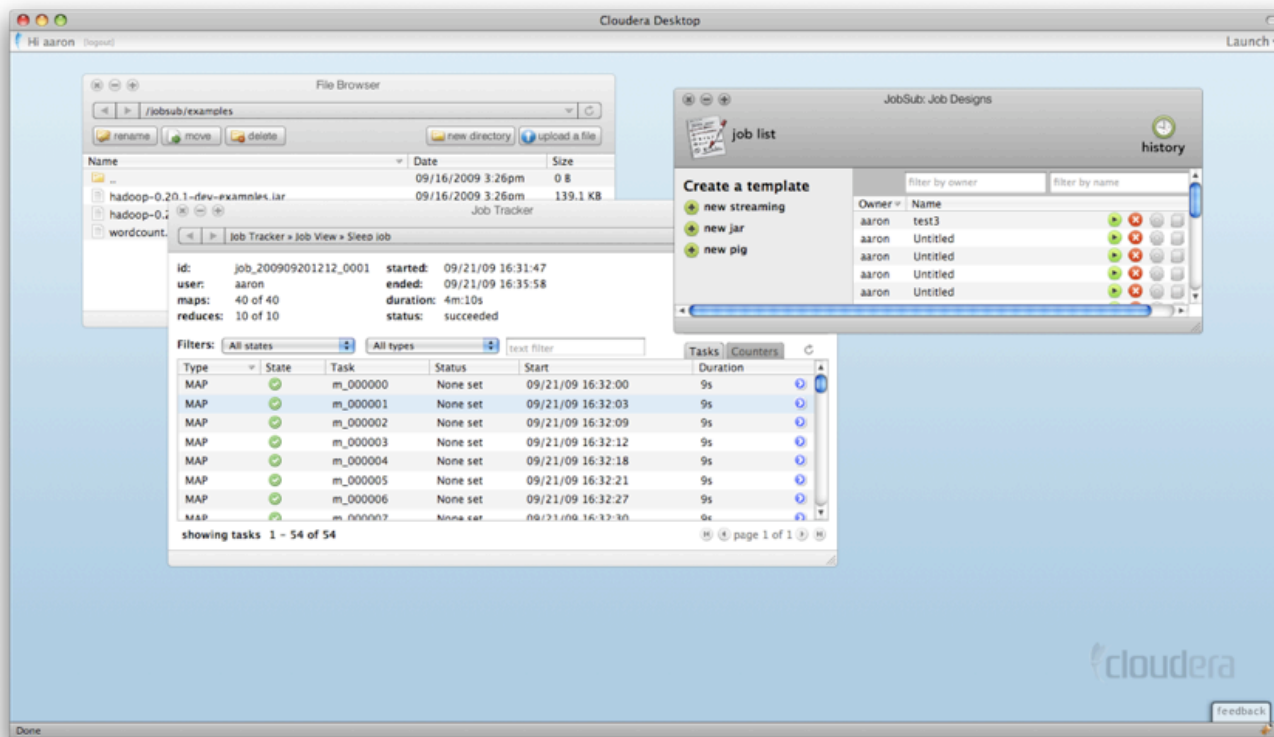
showing tasks 1 - 54 of 54

cloudera

feedback



Looking More Closely at Desktop Making Hadoop Easy for Everyone...

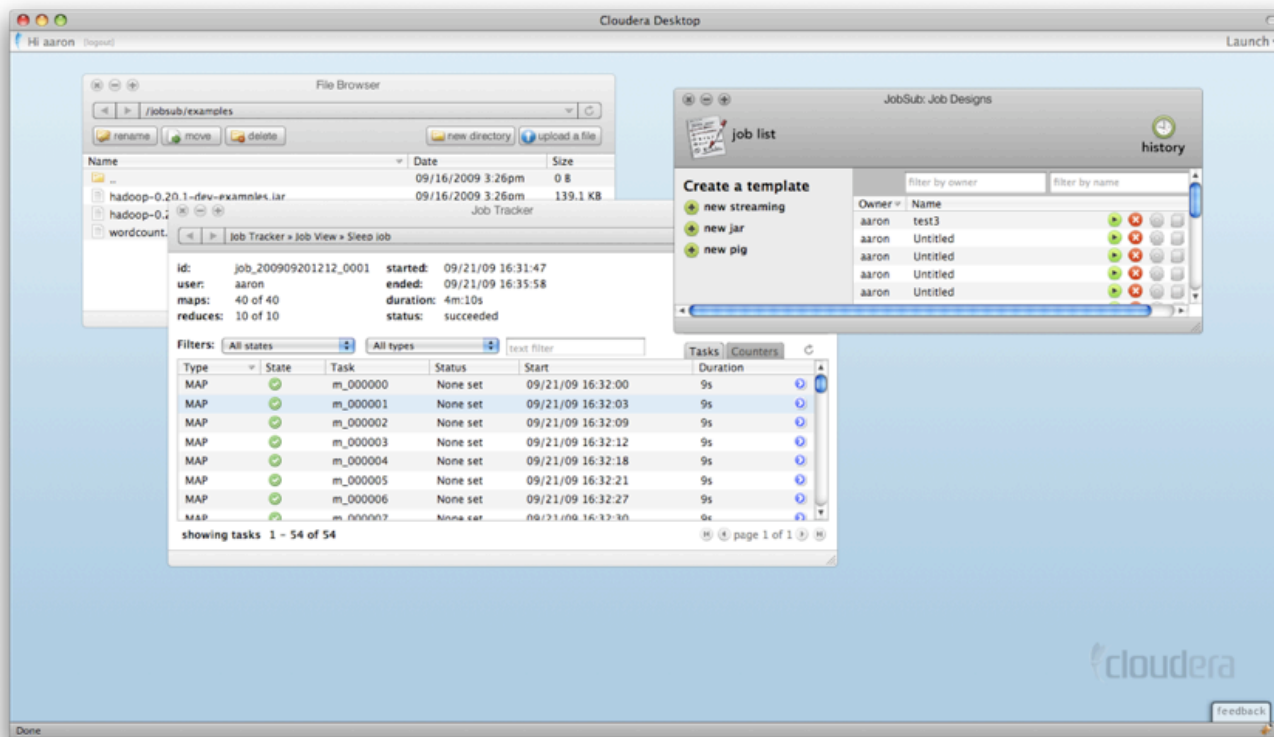


**Cloudera Desktop Runs
Entirely Inside the Web
Browser.**

No Client Side Configuration



Looking More Closely at Desktop Making Hadoop Easy for Everyone...



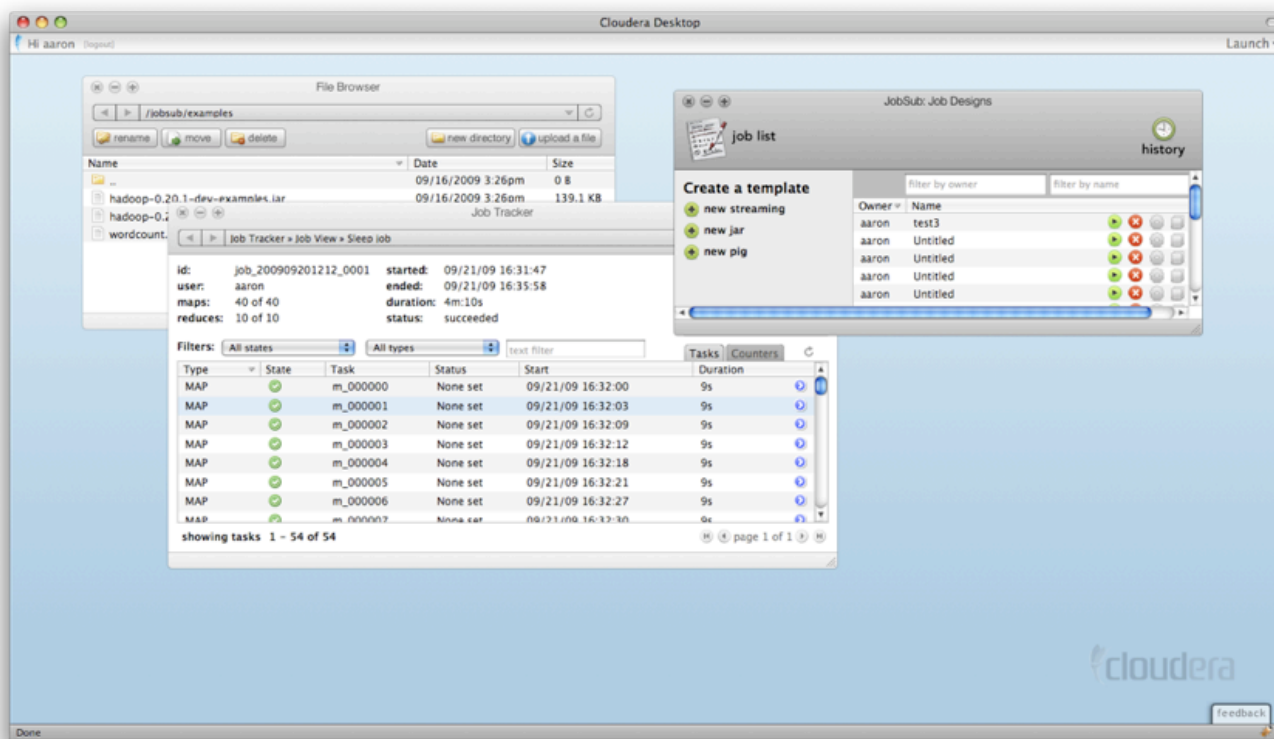
**Cloudera Desktop Runs
Entirely Inside the Web
Browser.**

No Client Side Configuration

**Uses a Desktop Metaphor:
Familiar to Everyone**



Looking More Closely at Desktop Making Hadoop Easy for Everyone...



**Cloudera Desktop Runs
Entirely Inside the Web
Browser.**

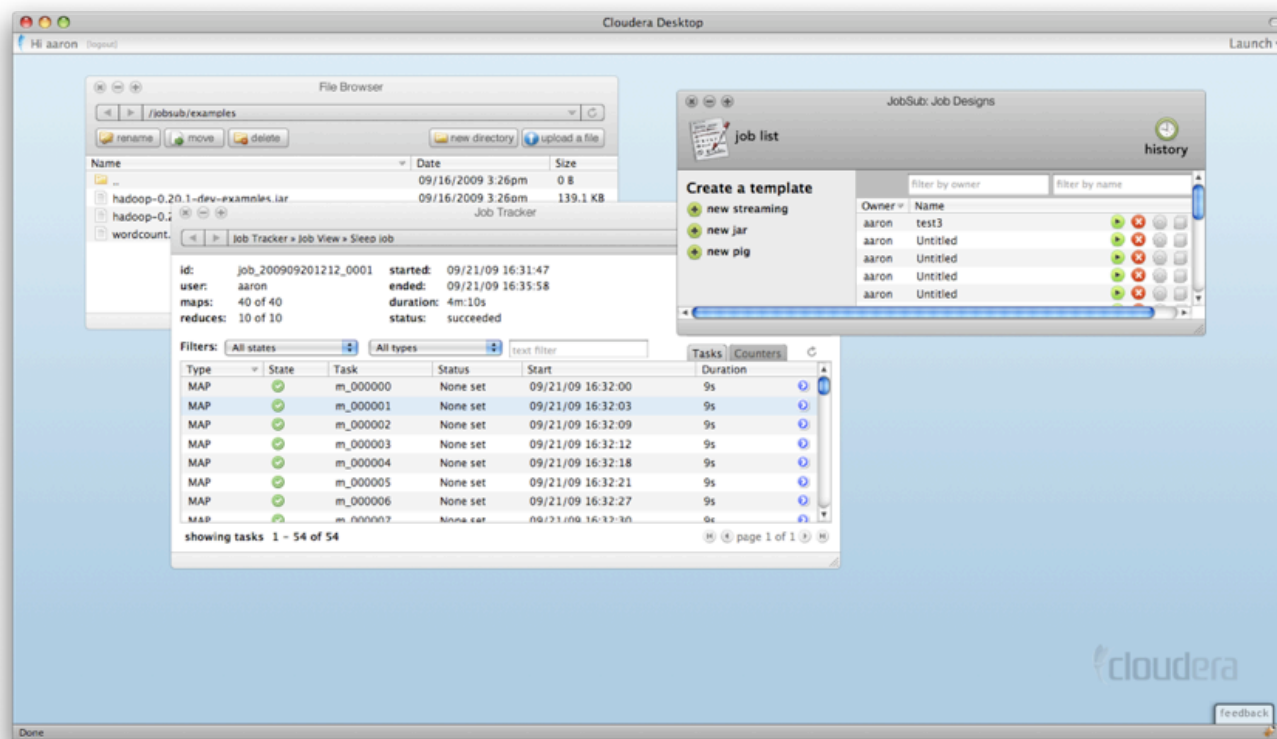
No Client Side Configuration

**Uses a Desktop Metaphor:
Familiar to Everyone**

**Multiple Applications Run
Simultaneously and Interact
With Each Other in Intuitive
Ways**



Looking More Closely at Desktop Making Hadoop Easy for Everyone...



Cloudera Desktop Runs Entirely Inside the Web Browser.

No Client Side Configuration

Uses a Desktop Metaphor: Familiar to Everyone

Multiple Applications Run Simultaneously and Interact With Each Other in Intuitive Ways

Enables Developers to Write Additional Applications and interact with Existing Apps

Looking More Closely at Desktop Working With Files Big and Small



HDFS Web Interface

NameNode 'localhost:54310'

Started: Fri Sep 21 15:49:19 CEST 2007
 Version: 0.14.1, r571288
 Compiled: Thu Aug 30 13:06:02 PDT 2007 by cutting

Browse the filesystem

Cluster Summary

Capacity : 33.64 GB
 Remaining : 30.80 GB
 Used : 8.43 %
[Live Nodes](#) : 1
[Dead Nodes](#) : 0

Live Datanodes : 1

Node	Last Contact	Admin State	Size (GB)	Used (%)	Blocks
localhost	54	In Service	33.64	8.43	8

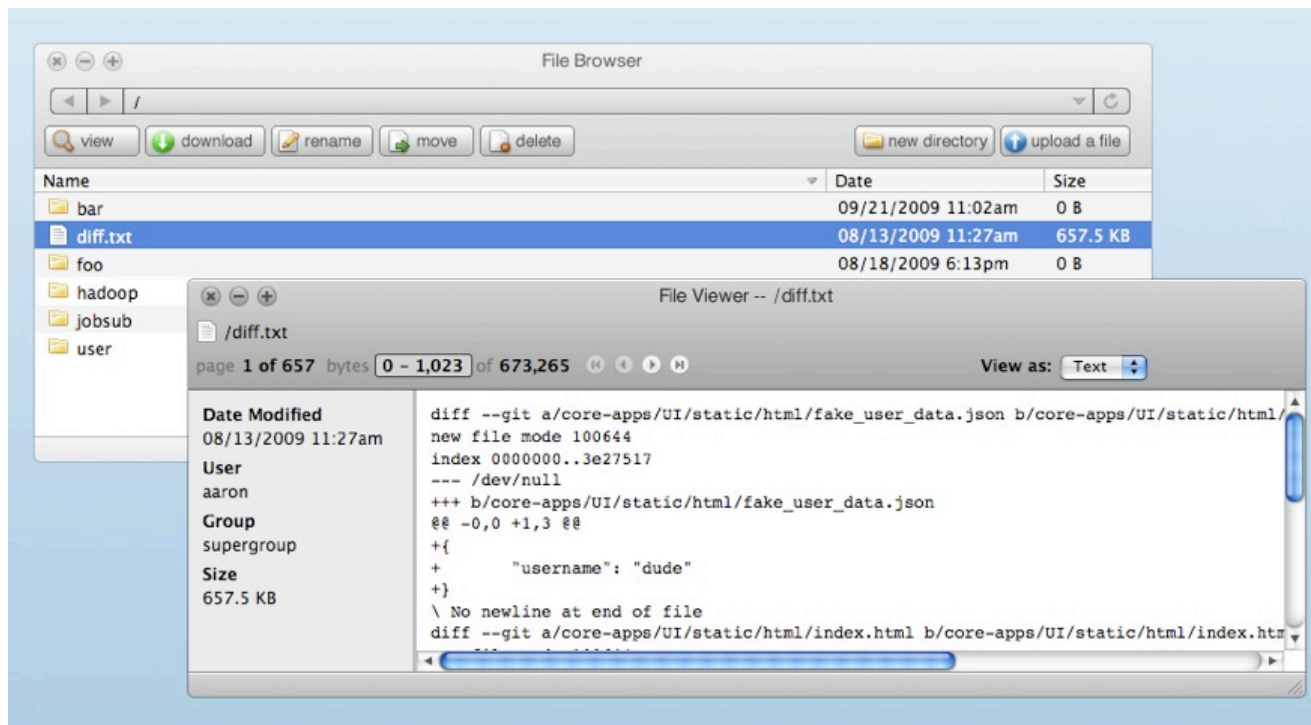
Dead Datanodes : 0

Local logs

[Log](#) directory

[Hadoop](#), 2006.

Looking More Closely at Desktop Working With Files Big and Small

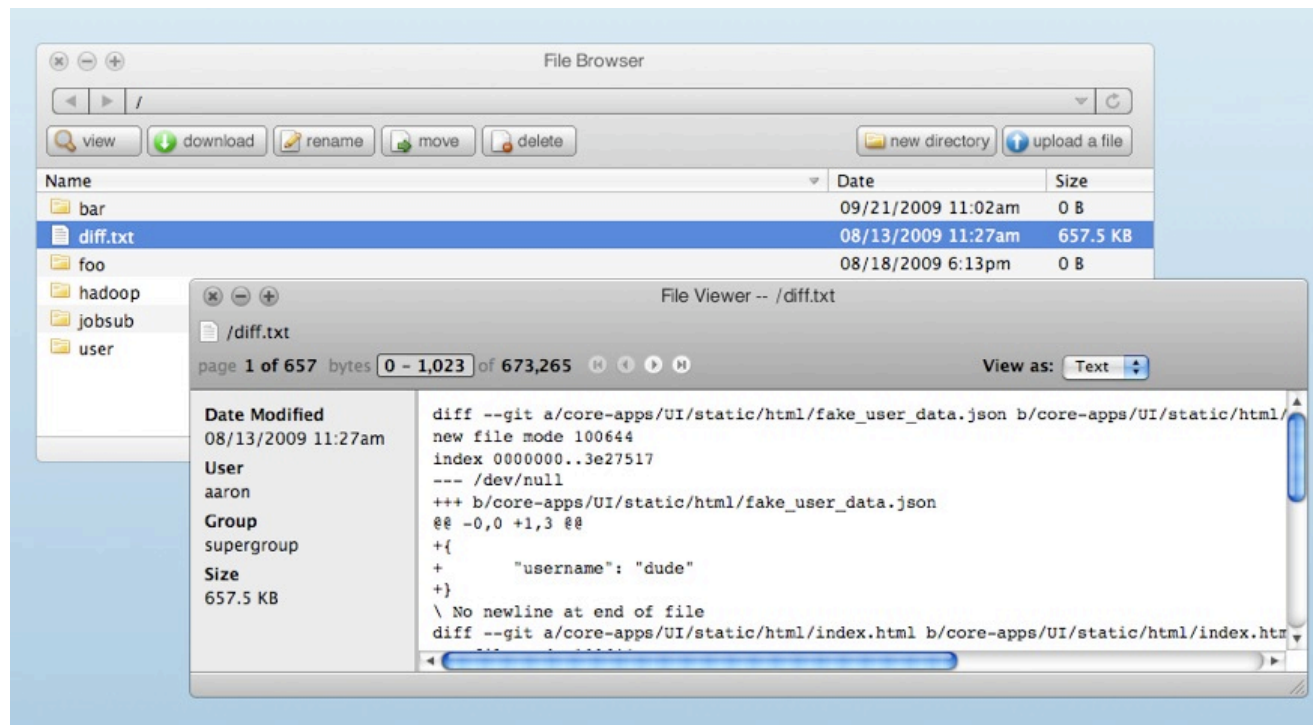


File Browser and File Viewer

Looking More Closely at Desktop Working With Files Big and Small



Upload and Download Files
from Your Web Browser



File Browser and File Viewer

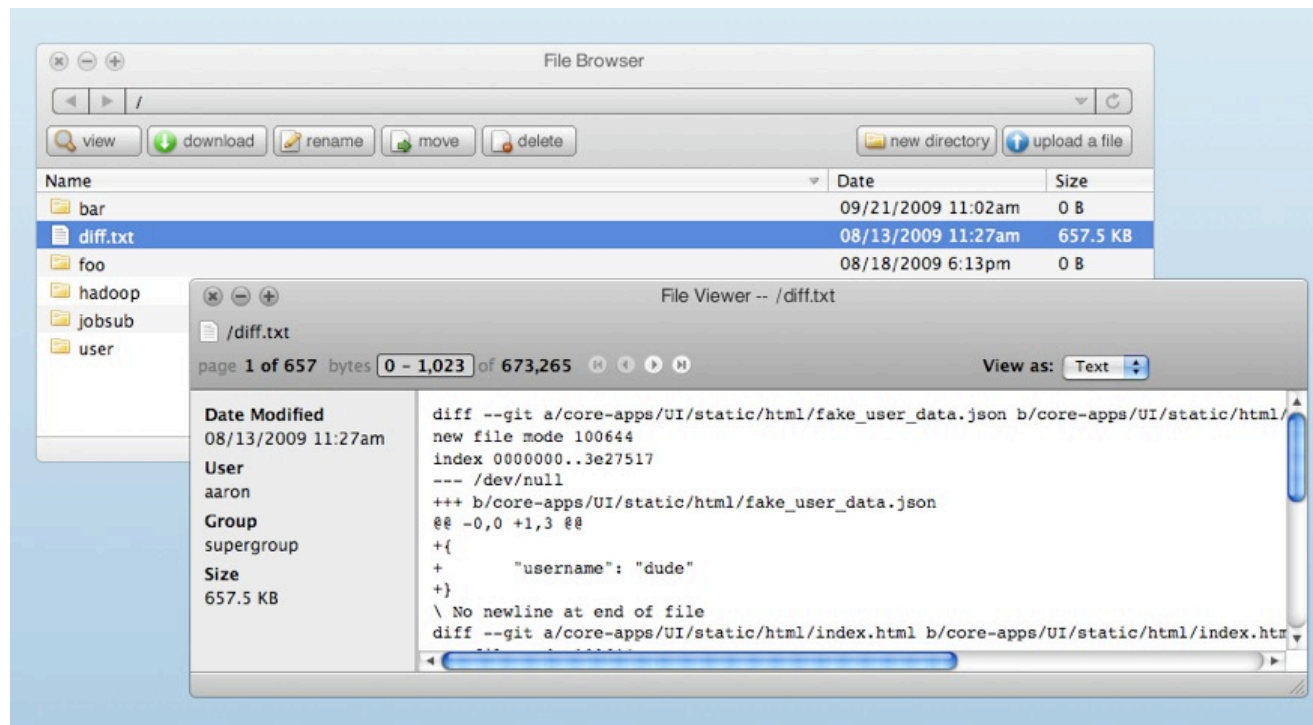


Looking More Closely at Desktop Working With Files Big and Small



Upload and Download Files
from You Web Browser

All Data Goes over HTTP



File Browser and File Viewer



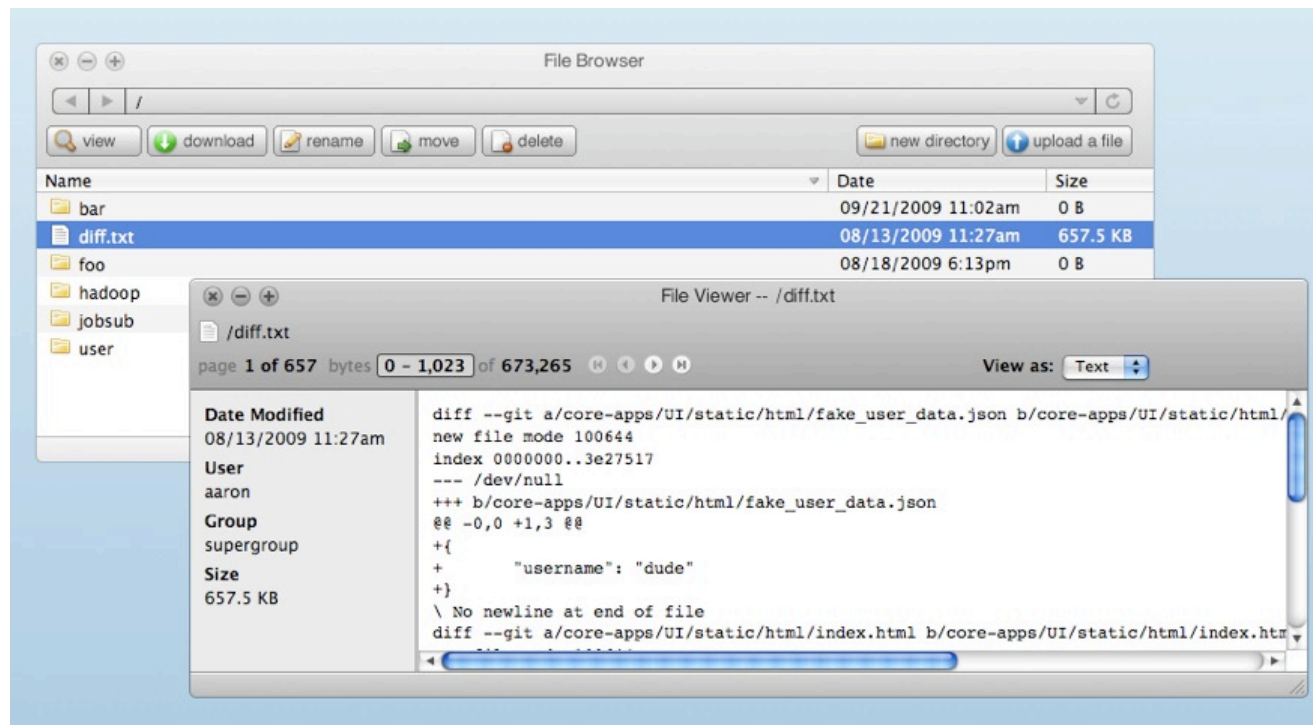
Looking More Closely at Desktop Working With Files Big and Small



Upload and Download Files
from You Web Browser

All Data Goes over HTTP

No Extra Firewall
Configuration



File Browser and File Viewer



Looking More Closely at Desktop Working With Files Big and Small

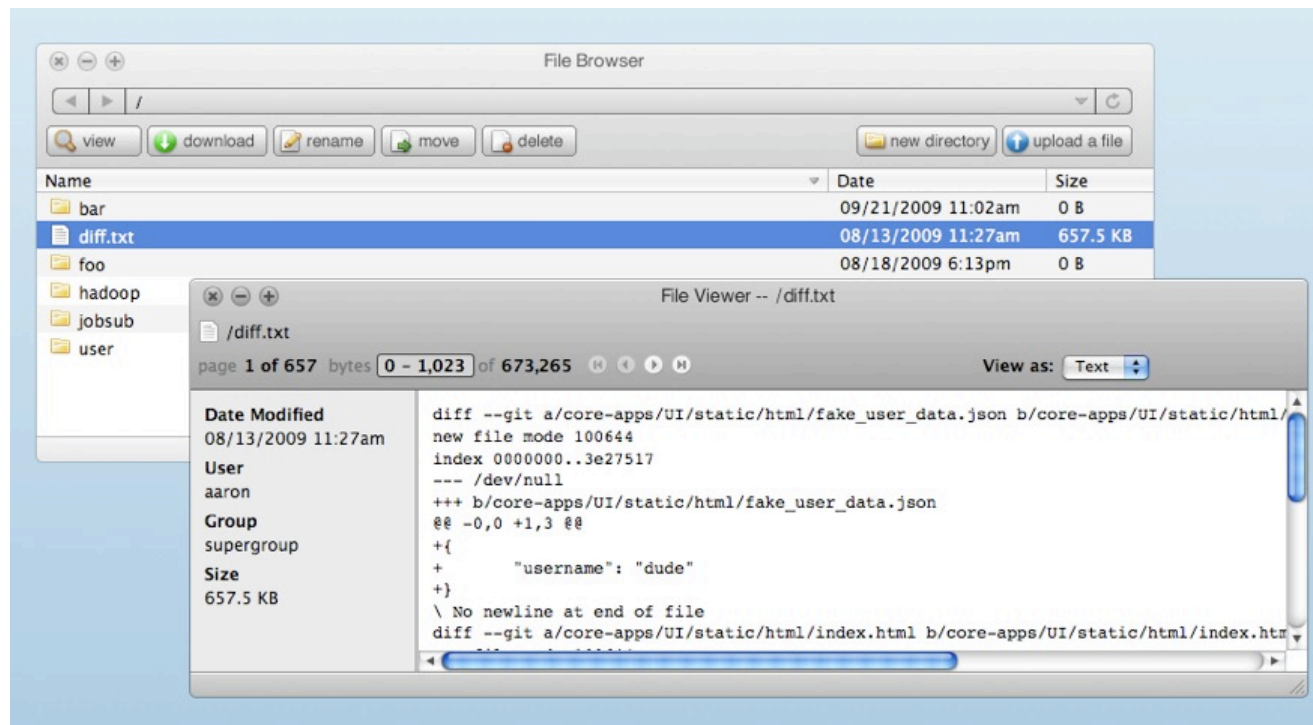


Upload and Download Files
from You Web Browser

All Data Goes over HTTP

No Extra Firewall
Configuration

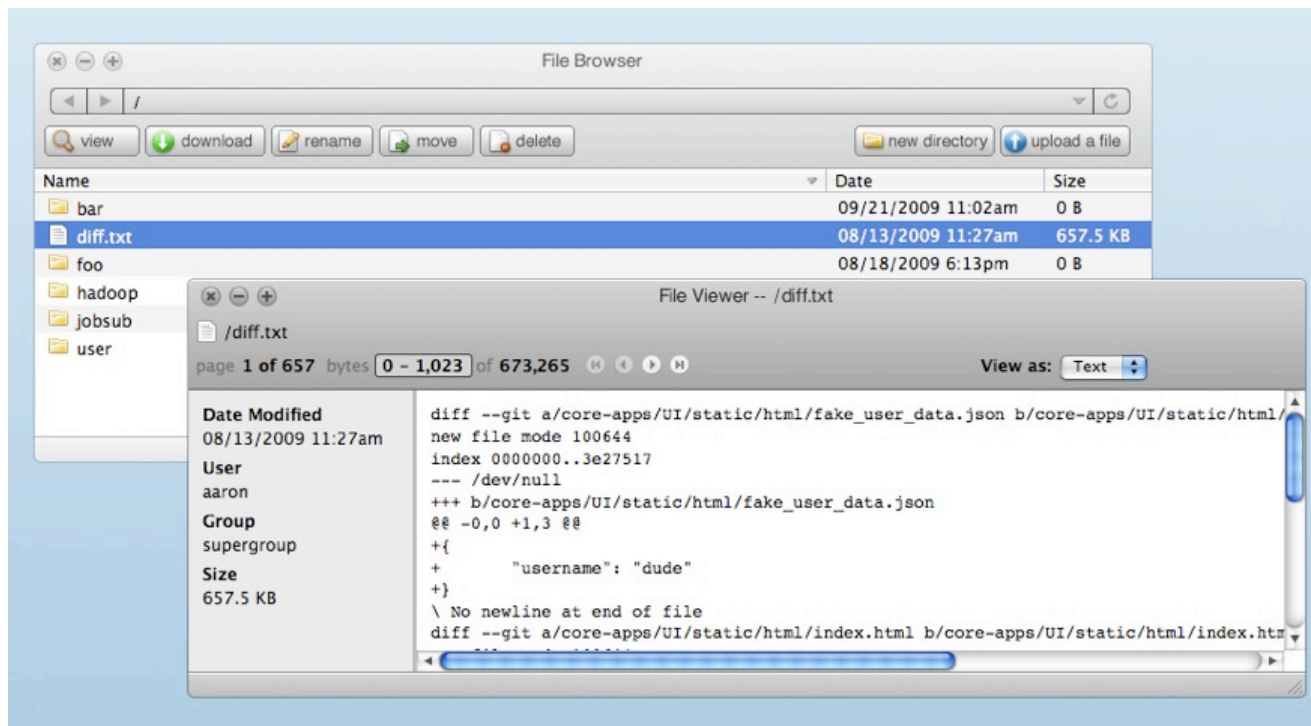
Uses same familiar
Metaphors as Windows
Explorer or Macintosh Finder



File Browser and File Viewer



Looking More Closely at Desktop Working With Files Big and Small



File Browser and File Viewer

**Upload and Download Files
from You Web Browser**

All Data Goes over HTTP

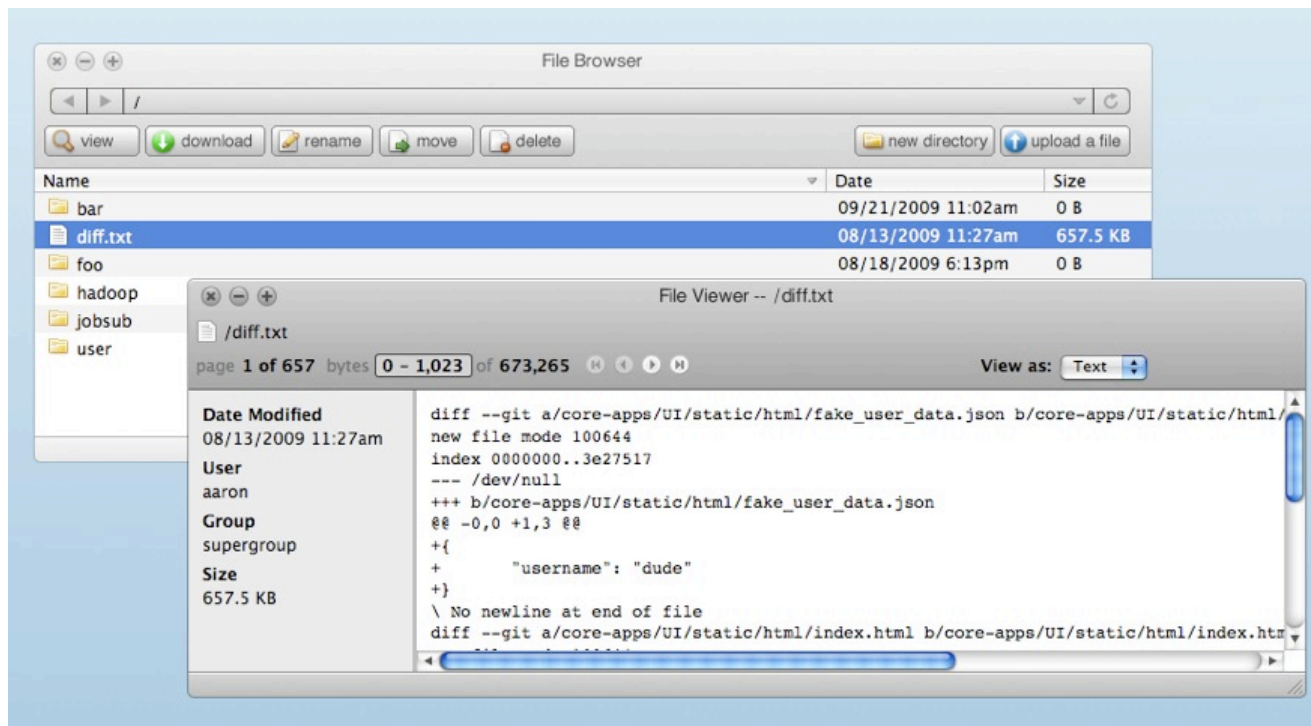
**No Extra Firewall
Configuration**

**Uses same familiar
Metaphors as Windows
Explorer or Macintosh Finder**

**Upload / Download Many
Small Files at Once**



Looking More Closely at Desktop Working With Files Big and Small



File Browser and File Viewer

**Upload and Download Files
from You Web Browser**

All Data Goes over HTTP

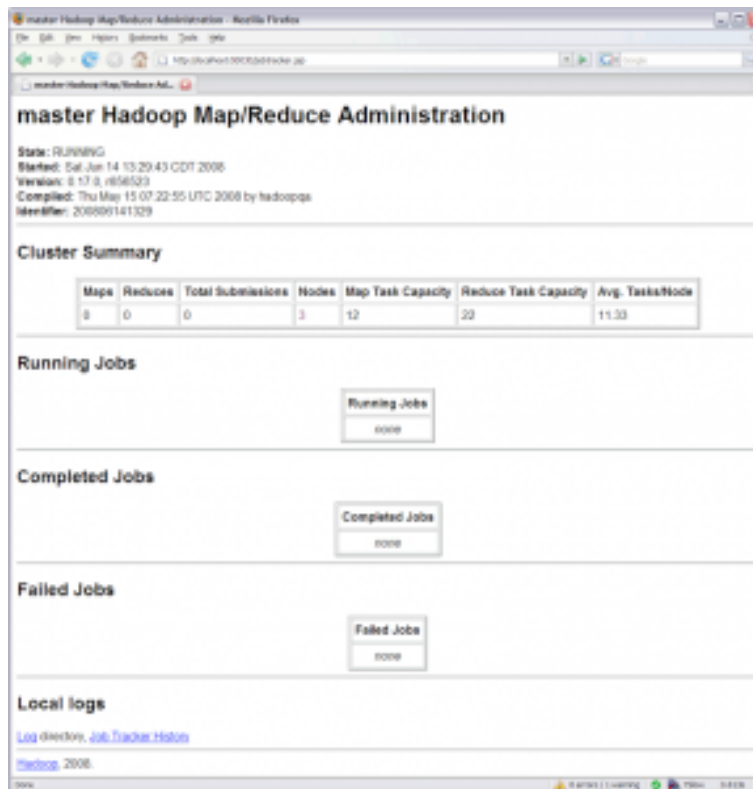
**No Extra Firewall
Configuration**

**Uses same familiar
Metaphors as Windows
Explorer or Macintosh Finder**

**Upload / Download Many
Small Files at Once**

Smart About Big Files

Looking More Closely at Desktop Monitor and Debug Jobs: Easy

master Hadoop Map/Reduce Administration

State: RUNNING
 Started: Sat Jun 14 13:29:43 CDT 2008
 Version: 0.17.8_r656523
 Compiled: Thu May 15 07:22:55 UTC 2008 by hadoopgs
 Identifier: 206699141329

Cluster Summary

Maps	Reduces	Total Submissions	Nodes	Map Task Capacity	Reduce Task Capacity	Avg. Tasks/Node
0	0	0	3	12	20	11.33

Running Jobs

Running Jobs
none

Completed Jobs

Completed Jobs
none

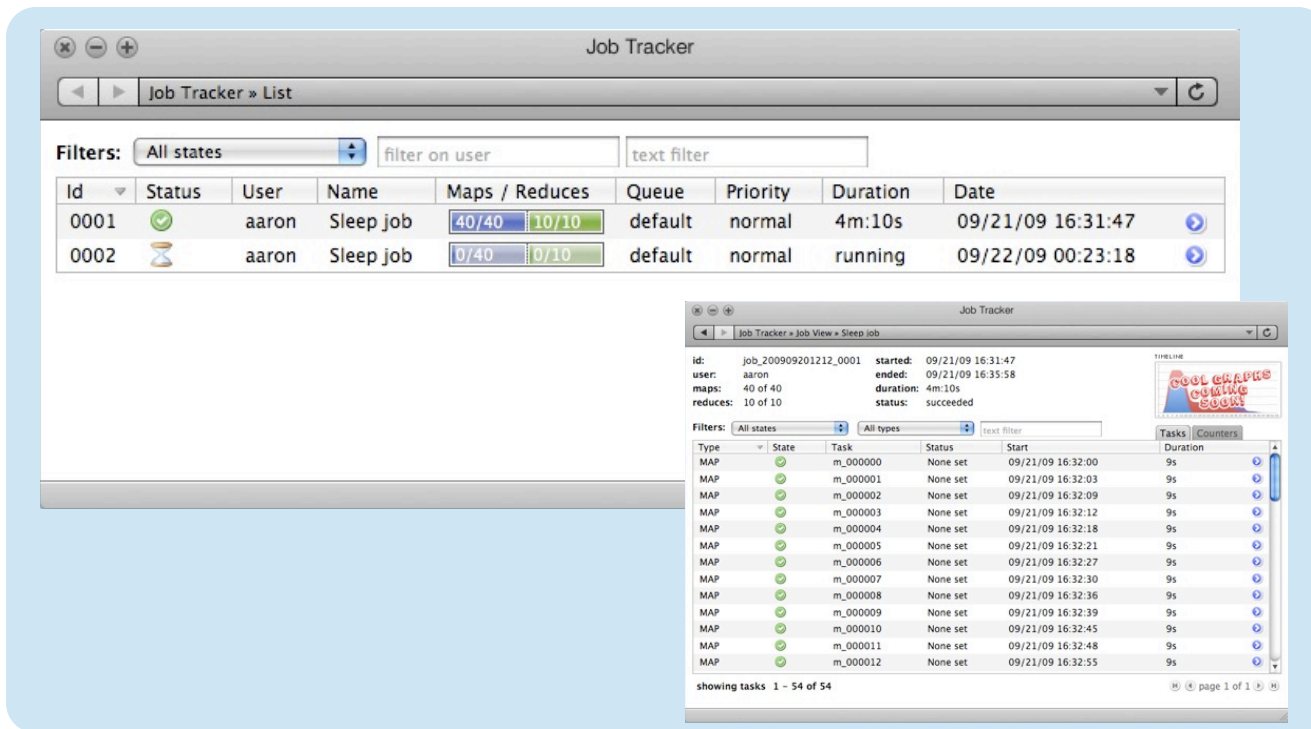
Failed Jobs

Failed Jobs
none

Local logs

[Log Director, Job, Task, Hadoop](#)
[Hadoop, 2008.](#)

Looking More Closely at Desktop Monitor and Debug Jobs: Easy

The screenshot shows the Cloudera Job Tracker web interface. The main window displays a list of jobs with the following data:

Id	Status	User	Name	Maps / Reduces	Queue	Priority	Duration	Date
0001	✓	aaron	Sleep job	40/40 10/10	default	normal	4m:10s	09/21/09 16:31:47
0002	⌚	aaron	Sleep job	0/40 0/10	default	normal	running	09/22/09 00:23:18

An inset window shows a detailed view of job '0001' (Sleep job) with the following summary:

- id: job_200909201212_0001
- user: aaron
- maps: 40 of 40
- reduces: 10 of 10
- started: 09/21/09 16:31:47
- ended: 09/21/09 16:35:58
- duration: 4m:10s
- status: succeeded

The detailed view also includes a table of tasks:

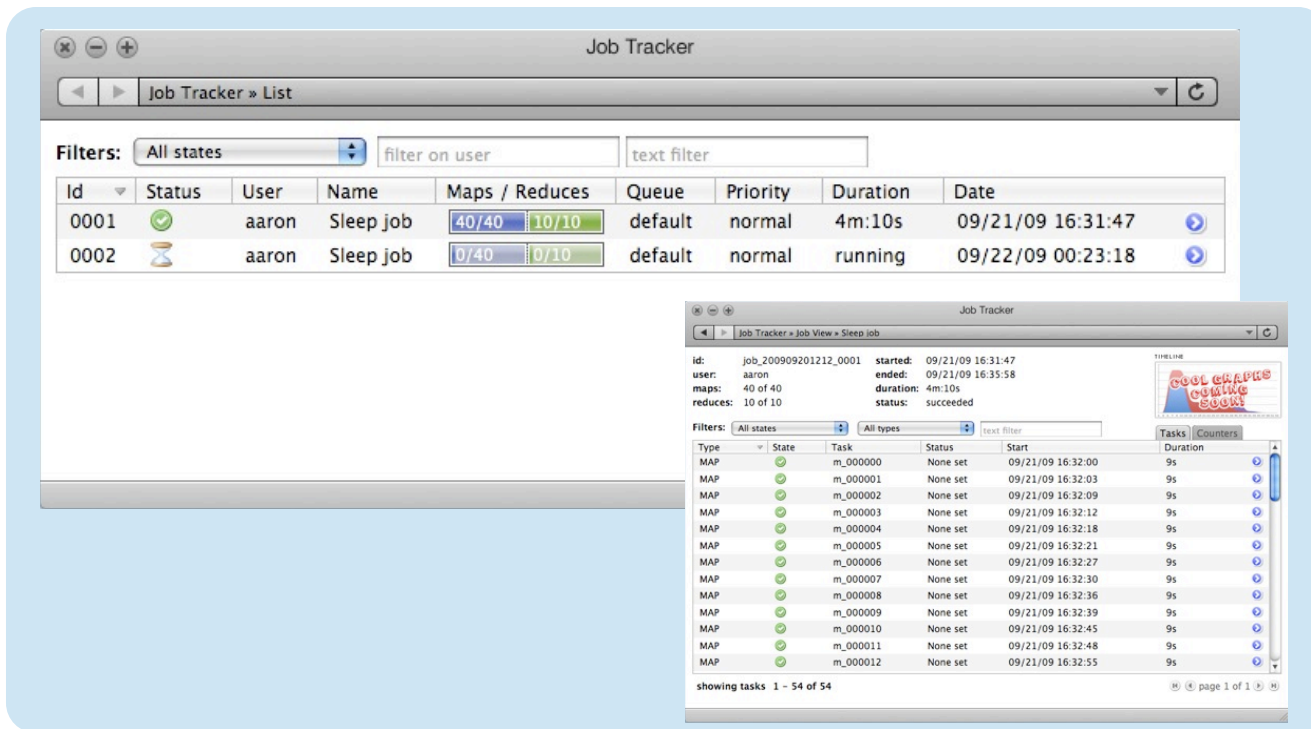
Type	State	Task	Status	Start	Duration
MAP	✓	m_000000	None set	09/21/09 16:32:00	9s
MAP	✓	m_000001	None set	09/21/09 16:32:03	9s
MAP	✓	m_000002	None set	09/21/09 16:32:09	9s
MAP	✓	m_000003	None set	09/21/09 16:32:12	9s
MAP	✓	m_000004	None set	09/21/09 16:32:18	9s
MAP	✓	m_000005	None set	09/21/09 16:32:21	9s
MAP	✓	m_000006	None set	09/21/09 16:32:27	9s
MAP	✓	m_000007	None set	09/21/09 16:32:30	9s
MAP	✓	m_000008	None set	09/21/09 16:32:36	9s
MAP	✓	m_000009	None set	09/21/09 16:32:39	9s
MAP	✓	m_000010	None set	09/21/09 16:32:45	9s
MAP	✓	m_000011	None set	09/21/09 16:32:48	9s
MAP	✓	m_000012	None set	09/21/09 16:32:55	9s

Job Browser

Looking More Closely at Desktop Monitor and Debug Jobs: Easy



Easily See What Jobs Are Running



The screenshot shows the Cloudera Job Tracker interface. The top window displays a list of jobs with the following data:

Id	Status	User	Name	Maps / Reduces	Queue	Priority	Duration	Date
0001	✓	aaron	Sleep job	40/40 10/10	default	normal	4m:10s	09/21/09 16:31:47
0002	⌚	aaron	Sleep job	0/40 0/10	default	normal	running	09/22/09 00:23:18

The bottom window shows a detailed view of job ID 0001, including the following summary:

- id: job_200909201212_0001
- user: aaron
- maps: 40 of 40
- reduces: 10 of 10
- started: 09/21/09 16:31:47
- ended: 09/21/09 16:35:58
- duration: 4m:10s
- status: succeeded

The detailed view also includes a table of individual tasks:

Type	State	Task	Status	Start	Duration
MAP	✓	m_000000	None set	09/21/09 16:32:00	9s
MAP	✓	m_000001	None set	09/21/09 16:32:03	9s
MAP	✓	m_000002	None set	09/21/09 16:32:09	9s
MAP	✓	m_000003	None set	09/21/09 16:32:12	9s
MAP	✓	m_000004	None set	09/21/09 16:32:18	9s
MAP	✓	m_000005	None set	09/21/09 16:32:21	9s
MAP	✓	m_000006	None set	09/21/09 16:32:27	9s
MAP	✓	m_000007	None set	09/21/09 16:32:30	9s
MAP	✓	m_000008	None set	09/21/09 16:32:36	9s
MAP	✓	m_000009	None set	09/21/09 16:32:39	9s
MAP	✓	m_000010	None set	09/21/09 16:32:45	9s
MAP	✓	m_000011	None set	09/21/09 16:32:48	9s
MAP	✓	m_000012	None set	09/21/09 16:32:55	9s

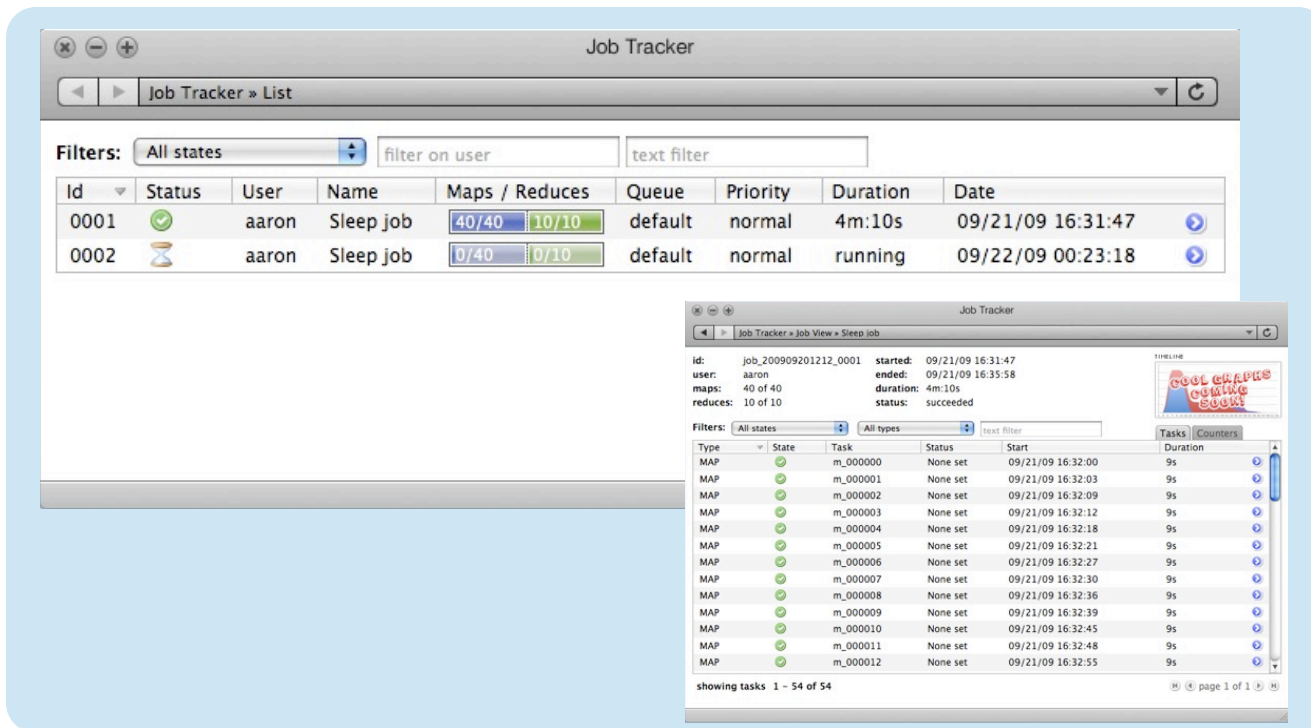
Job Browser

Looking More Closely at Desktop Monitor and Debug Jobs: Easy



Easily See What Jobs Are Running

Click Through To Get More Details



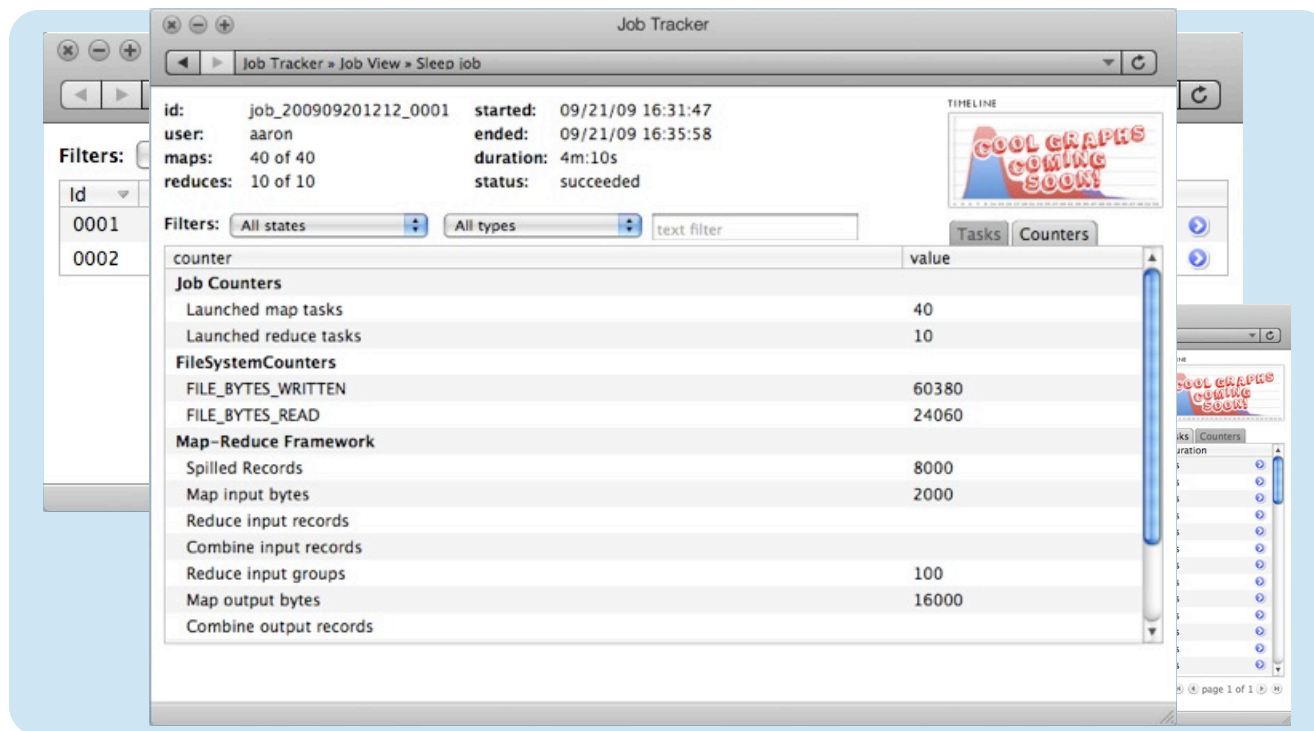
The screenshot shows the Cloudera Job Tracker interface. The top window displays a list of jobs with columns for Id, Status, User, Name, Maps / Reduces, Queue, Priority, Duration, and Date. Two jobs are visible: job 0001 (completed) and job 0002 (running). The bottom window shows a detailed view of job 0001, including its ID, user, start/end times, duration, and status (succeeded). It also displays a list of tasks (MAP) with their individual states and durations.

Id	Status	User	Name	Maps / Reduces	Queue	Priority	Duration	Date
0001	✓	aaron	Sleep job	40/40 10/10	default	normal	4m:10s	09/21/09 16:31:47
0002	⌚	aaron	Sleep job	0/40 0/10	default	normal	running	09/22/09 00:23:18

Type	State	Task	Status	Start	Duration
MAP	✓	m_000000	None set	09/21/09 16:32:00	9s
MAP	✓	m_000001	None set	09/21/09 16:32:03	9s
MAP	✓	m_000002	None set	09/21/09 16:32:09	9s
MAP	✓	m_000003	None set	09/21/09 16:32:12	9s
MAP	✓	m_000004	None set	09/21/09 16:32:18	9s
MAP	✓	m_000005	None set	09/21/09 16:32:21	9s
MAP	✓	m_000006	None set	09/21/09 16:32:27	9s
MAP	✓	m_000007	None set	09/21/09 16:32:30	9s
MAP	✓	m_000008	None set	09/21/09 16:32:36	9s
MAP	✓	m_000009	None set	09/21/09 16:32:39	9s
MAP	✓	m_000010	None set	09/21/09 16:32:45	9s
MAP	✓	m_000011	None set	09/21/09 16:32:48	9s
MAP	✓	m_000012	None set	09/21/09 16:32:55	9s

Job Browser

Looking More Closely at Desktop Monitor and Debug Jobs: Easy

Job Tracker

Job Tracker » Job View » Sleep job

id: job_200909201212_0001 started: 09/21/09 16:31:47
 user: aaron ended: 09/21/09 16:35:58
 maps: 40 of 40 duration: 4m:10s
 reduces: 10 of 10 status: succeeded

Filters: All states All types text filter

counter	value
Job Counters	
Launched map tasks	40
Launched reduce tasks	10
FileSystemCounters	
FILE_BYTES_WRITTEN	60380
FILE_BYTES_READ	24060
Map-Reduce Framework	
Spilled Records	8000
Map input bytes	2000
Reduce input records	
Combine input records	
Reduce input groups	100
Map output bytes	16000
Combine output records	

COOL GRAPHS COMING SOON!

Tasks Counters

page 1 of 1

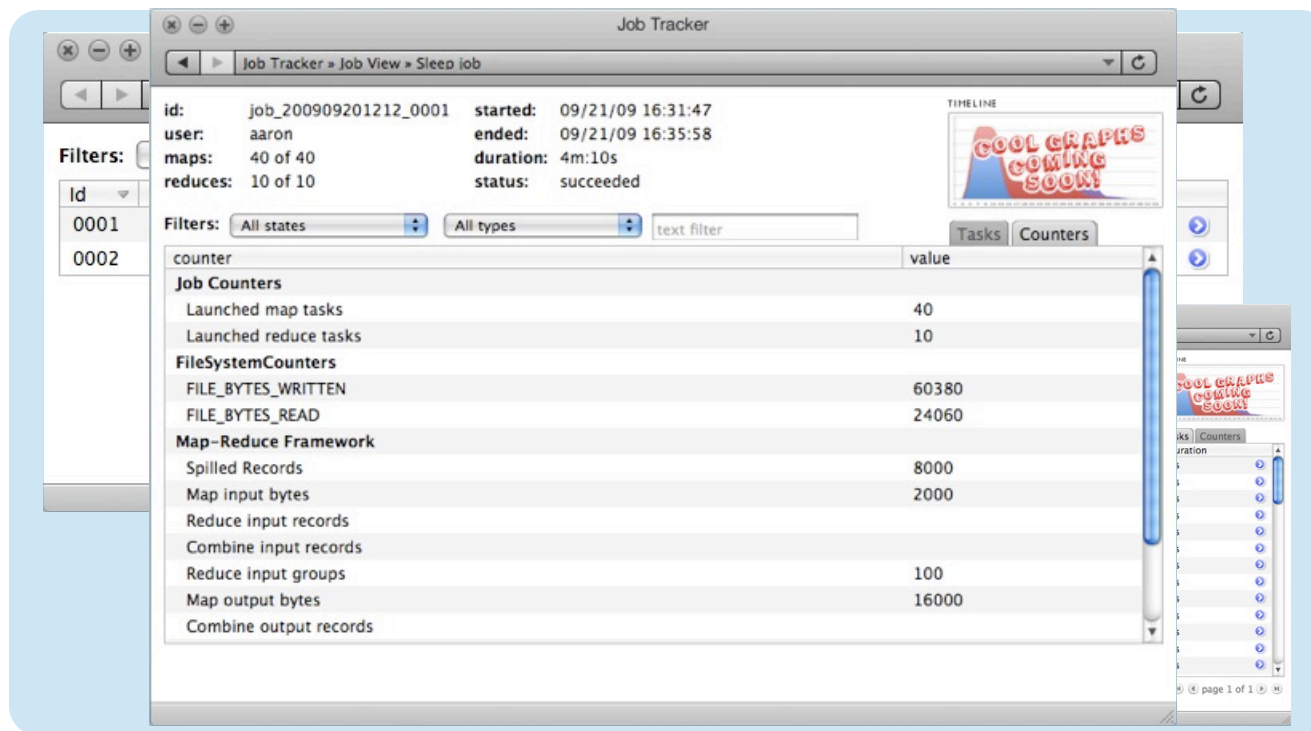
Easily See What Jobs Are Running

Click Through To Get More Details

Cool Graphs Coming Soon!

Job Browser

Looking More Closely at Desktop Monitor and Debug Jobs: Easy



counter	value
Job Counters	
Launched map tasks	40
Launched reduce tasks	10
FileSystemCounters	
FILE_BYTES_WRITTEN	60380
FILE_BYTES_READ	24060
Map-Reduce Framework	
Spilled Records	8000
Map input bytes	2000
Reduce input records	
Combine input records	
Reduce input groups	100
Map output bytes	16000
Combine output records	

Job Browser

Easily See What Jobs Are Running

Click Through To Get More Details

Cool Graphs Coming Soon!

Counters And Other Useful Data Already Available

Looking More Closely at Desktop

Always Keep an Eye on Cluster Health



Looking More Closely at Desktop Always Keep an Eye on Cluster Health



cluster health_
TAKE A STRESS PILL AND THINK THINGS OVER

datanode percent space renaming: 28MB

clusters	●	●	●
HDFS (default)	5	0	1
MR (default)	4	2	0
SUMMARY	malfunction		

Type	State	Task	Status	Start
MAP	●	m_000000	None set	09/21/09
MAP	●	m_000001	None set	09/21/09
MAP	●	m_000002	None set	09/21/09
MAP	●	m_000003	None set	09/21/09 16:32:12
MAP	●	m_000004	None set	09/21/09 16:32:18
MAP	●	m_000005	None set	09/21/09 16:32:21
MAP	●	m_000006	None set	09/21/09 16:32:27
MAP	●	m_000007	None set	09/21/09 16:32:30
MAP	●	m_000008	None set	09/21/09 16:32:36
MAP	●	m_000009	None set	09/21/09 16:32:39
MAP	●	m_000010	None set	09/21/09 16:32:45
MAP	●	m_000011	None set	09/21/09 16:32:48
MAP	●	m_000012	None set	09/21/09 16:32:55

showing tasks 1 - 54 of 54

cluster health_
TAKE A STRESS PILL AND THINK THINGS OVER

```

namenode      1  0  0  view >>
datanodes     5  0  0
-----
capacity      2.6 TB
non-dfs       194.7 GB
fs remaining   2.4 TB

datanode  state  capacity  used
10.10.1.1  ●    537.9 GB  560.0 KB
10.10.1.1  ●    537.9 GB  560.0 KB
10.10.1.1  ●    537.9 GB  560.0 KB
10.10.1.1  ●    537.9 GB  560.0 KB
10.10.1.1  ●    537.9 GB  560.0 KB
10.10.1.1  ●    537.9 GB  560.0 KB
10.10.1.1  ●    537.9 GB  560.0 KB

```

Cluster Health

Looking More Closely at Desktop Always Keep an Eye on Cluster Health



Keep a Cluster Health Window Open while Working

cluster health_
TAKE A STRESS PILL AND THINK THINGS OVER

datanode percent space renaming: 28MB

```
clusters
HDFS (default) 5 0 1
MR (default) 4 2 0
SUMMARY malfunction
```

Type	State	Task	Status	Start
MAP	✓	m_000000	None set	09/21/09
MAP	✓	m_000001	None set	09/21/09
MAP	✓	m_000002	None set	09/21/09
MAP	✓	m_000003	None set	09/21/09 16:32:12
MAP	✓	m_000004	None set	09/21/09 16:32:18
MAP	✓	m_000005	None set	09/21/09 16:32:21
MAP	✓	m_000006	None set	09/21/09 16:32:27
MAP	✓	m_000007	None set	09/21/09 16:32:30
MAP	✓	m_000008	None set	09/21/09 16:32:36
MAP	✓	m_000009	None set	09/21/09 16:32:39
MAP	✓	m_000010	None set	09/21/09 16:32:45
MAP	✓	m_000011	None set	09/21/09 16:32:48
MAP	✓	m_000012	None set	09/21/09 16:32:55

showing tasks 1 - 54 of 54

```
cluster health_
TAKE A STRESS PILL AND THINK THINGS OVER

namenode 1 0 0 view >>
datanodes 5 0 0

capacity 2.6 TB
non-dfs 194.7 GB
fs remaining 2.4 TB

datanode state capacity used
10.10.1.1 ● 537.9 GB 560.0 KB
10.10.1.1 ● 537.9 GB 560.0 KB
10.10.1.1 ● 537.9 GB 560.0 KB
10.10.1.1 ● 537.9 GB 560.0 KB
10.10.1.1 ● 537.9 GB 560.0 KB
10.10.1.1 ● 537.9 GB 560.0 KB
```

Cluster Health

Looking More Closely at Desktop Always Keep an Eye on Cluster Health



Keep a Cluster Health Window Open while Working

Visual Metrics: Easy to Read

cluster health_
TAKE A STRESS PILL AND THINK THINGS OVER

datanode percent space renaming: 28MB

clusters

HDFS (default)	5	0	1
MR (default)	4	2	0
SUMMARY			malfunction

datanode state capacity used

10.10.1.1	●	537.9 GB	560.0 KB
10.10.1.1	●	537.9 GB	560.0 KB
10.10.1.1	●	537.9 GB	560.0 KB
10.10.1.1	●	537.9 GB	560.0 KB
10.10.1.1	●	537.9 GB	560.0 KB
10.10.1.1	●	537.9 GB	560.0 KB

Type	State	Task	Status	Start
MAP	●	m_000000	None set	09/21/09
MAP	●	m_000001	None set	09/21/09
MAP	●	m_000002	None set	09/21/09
MAP	●	m_000003	None set	09/21/09 16:32:12
MAP	●	m_000004	None set	09/21/09 16:32:18
MAP	●	m_000005	None set	09/21/09 16:32:21
MAP	●	m_000006	None set	09/21/09 16:32:27
MAP	●	m_000007	None set	09/21/09 16:32:30
MAP	●	m_000008	None set	09/21/09 16:32:36
MAP	●	m_000009	None set	09/21/09 16:32:39
MAP	●	m_000010	None set	09/21/09 16:32:45
MAP	●	m_000011	None set	09/21/09 16:32:48
MAP	●	m_000012	None set	09/21/09 16:32:55

showing tasks 1 - 54 of 54

Cluster Health

Looking More Closely at Desktop Always Keep an Eye on Cluster Health



Keep a Cluster Health Window Open while Working

Visual Metrics: Easy to Read

Diagnose and Resolve Problems Quickly

datanode	state	capacity	used
10.10.1.1	●	537.9 GB	560.0 KB
10.10.1.1	●	537.9 GB	560.0 KB
10.10.1.1	●	537.9 GB	560.0 KB
10.10.1.1	●	537.9 GB	560.0 KB
10.10.1.1	●	537.9 GB	560.0 KB
10.10.1.1	●	537.9 GB	560.0 KB
10.10.1.1	●	537.9 GB	560.0 KB

Cluster Health

Looking More Closely at Desktop Always Keep an Eye on Cluster Health



cluster health_
TAKE A STRESS PILL AND THINK THINGS OVER

datanode percent space renaming: 28MB

clusters

```

HDFS (default) 5 0 1
MR (default) 4 2 0
SUMMARY malfunction
  
```

Type	State	Task	Status	Start
MAP	✓	m_000000	None set	09/21/09
MAP	✓	m_000001	None set	09/21/09
MAP	✓	m_000002	None set	09/21/09
MAP	✓	m_000003	None set	09/21/09 16:32:12
MAP	✓	m_000004	None set	09/21/09 16:32:18
MAP	✓	m_000005	None set	09/21/09 16:32:21
MAP	✓	m_000006	None set	09/21/09 16:32:27
MAP	✓	m_000007	None set	09/21/09 16:32:30
MAP	✓	m_000008	None set	09/21/09 16:32:36
MAP	✓	m_000009	None set	09/21/09 16:32:39
MAP	✓	m_000010	None set	09/21/09 16:32:45
MAP	✓	m_000011	None set	09/21/09 16:32:48
MAP	✓	m_000012	None set	09/21/09 16:32:55

namenode 1 0 0 view >>

datanodes 5 0 0

capacity 2.6 TB

non-dfs 194.7 GB

fs remaining 2.4 TB

datanode	state	capacity	used
10.10.1.1	●	537.9 GB	560.0 KB
10.10.1.1	●	537.9 GB	560.0 KB
10.10.1.1	●	537.9 GB	560.0 KB
10.10.1.1	●	537.9 GB	560.0 KB
10.10.1.1	●	537.9 GB	560.0 KB
10.10.1.1	●	537.9 GB	560.0 KB

Keep a Cluster Health Window Open while Working

Visual Metrics: Easy to Read

Diagnose and Resolve Problems Quickly

Setup Your Own Metrics with Simple Extension Interface

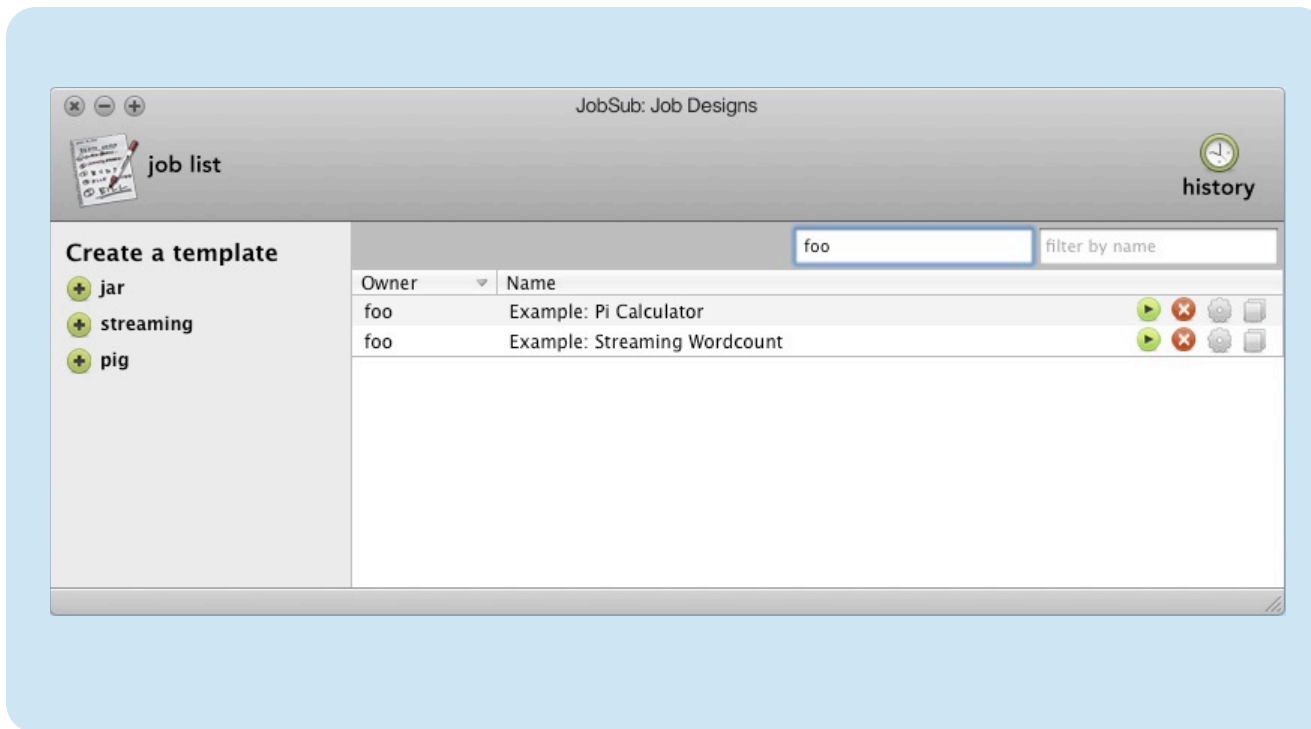
Cluster Health

Looking More Closely at Desktop

Enabling Less Technical Users



Looking More Closely at Desktop Enabling Less Technical Users

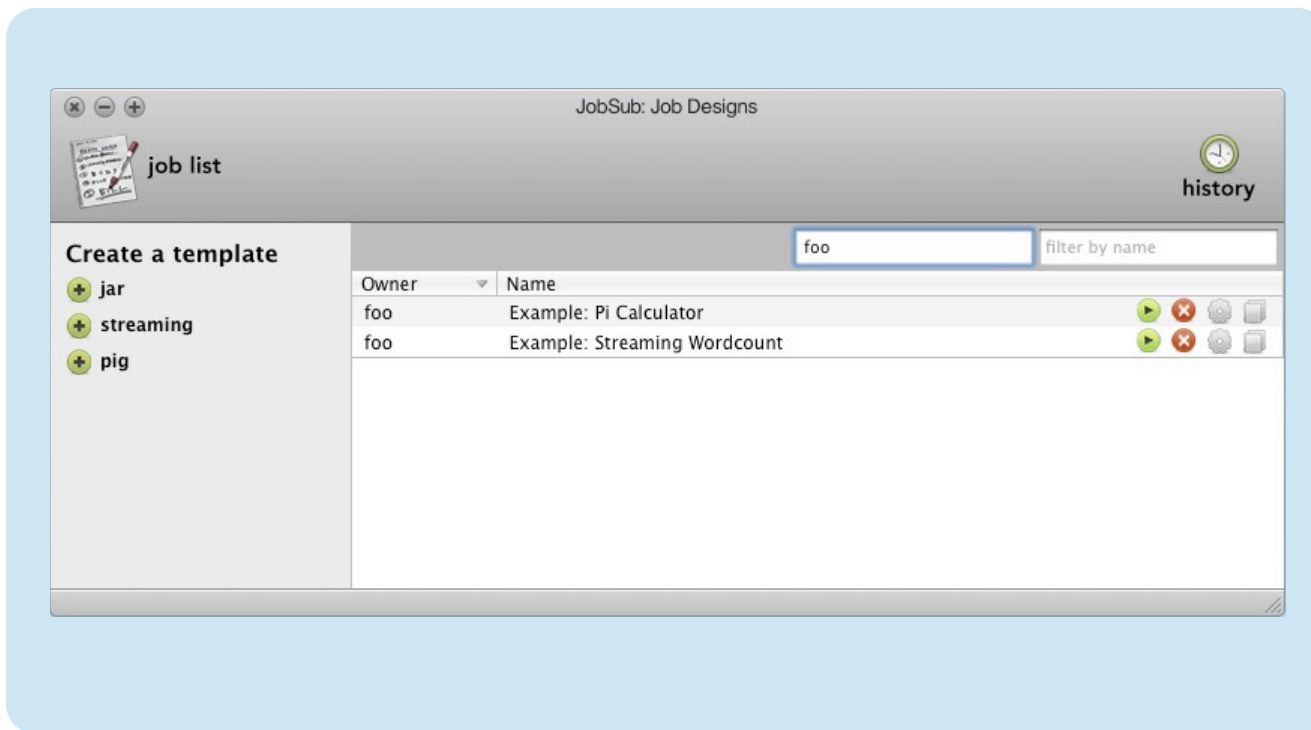


Job Designer

Looking More Closely at Desktop Enabling Less Technical Users



Developers Can Write Java,
Streaming or Pig Jobs (more
coming soon)



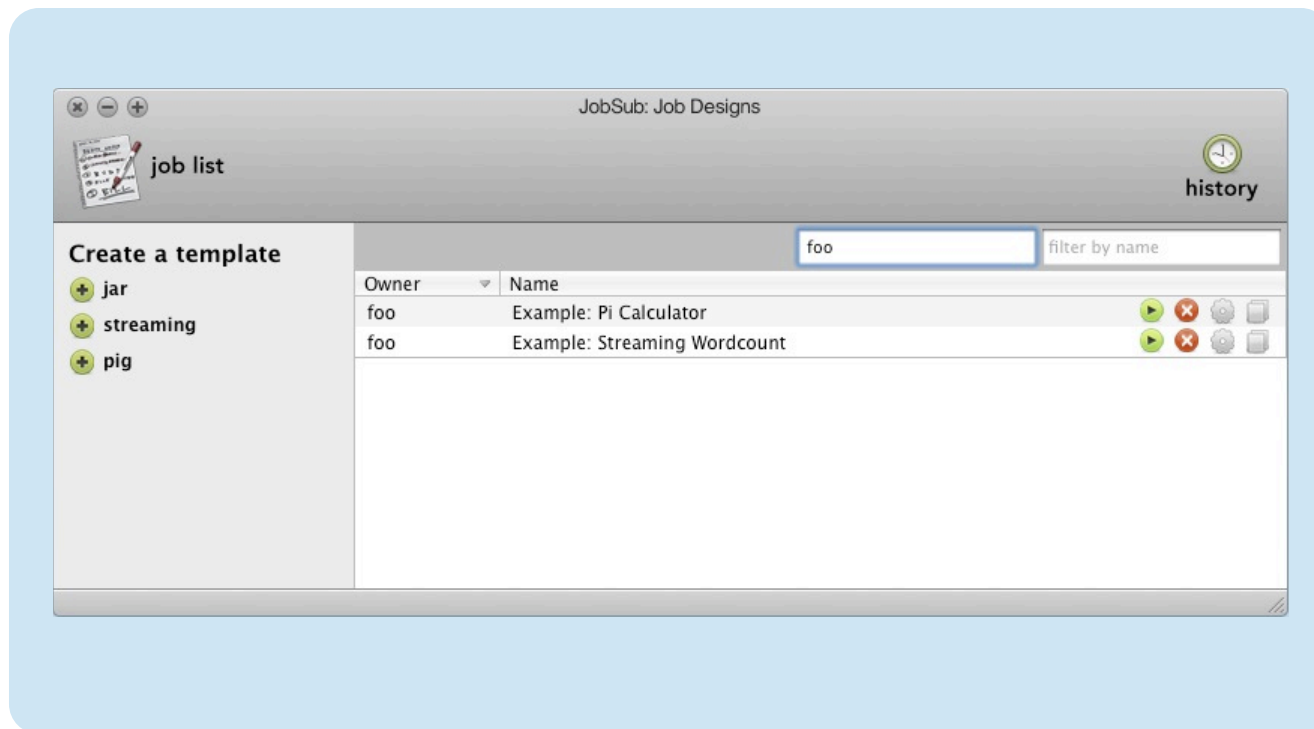
Job Designer

Looking More Closely at Desktop Enabling Less Technical Users



Developers Can Write Java, Streaming or Pig Jobs (more coming soon)

Developers Can Specify Input Parameters in Job Designer: Input Files, Date Range, Output Database, Etc



Job Designer

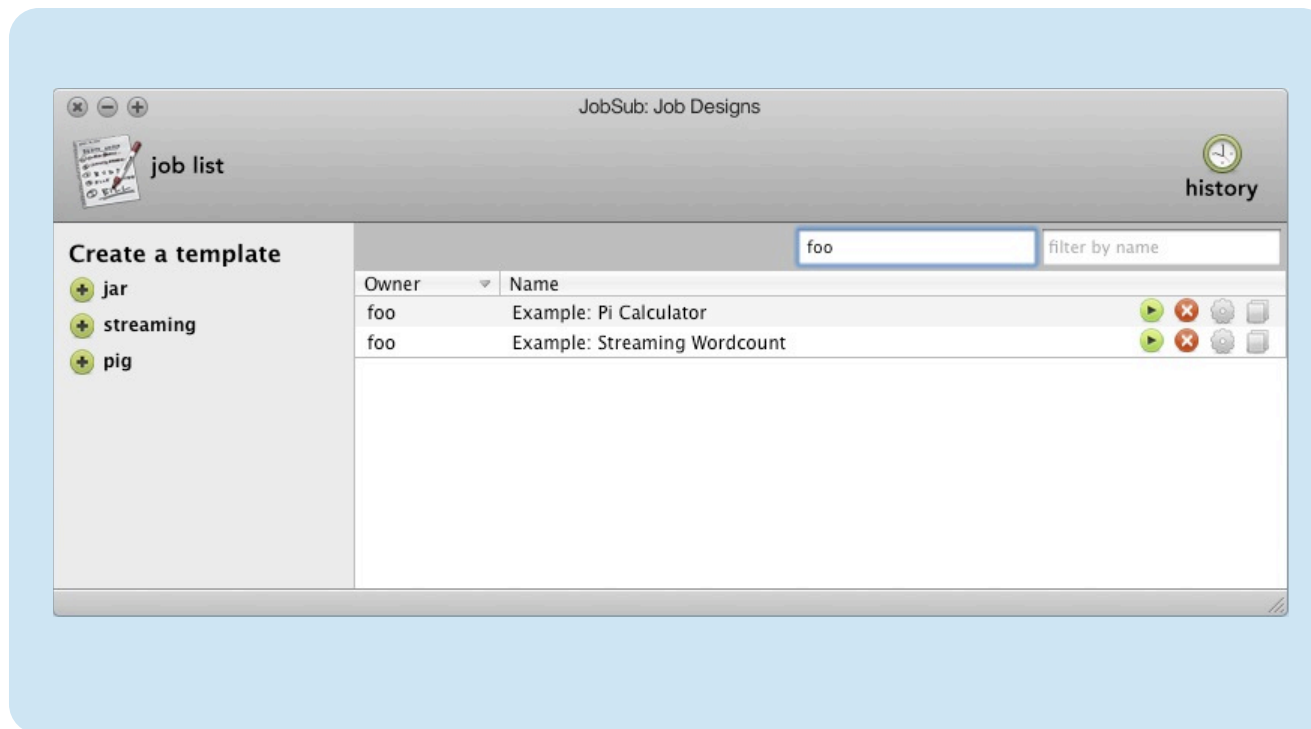
Looking More Closely at Desktop Enabling Less Technical Users



Developers Can Write Java, Streaming or Pig Jobs (more coming soon)

Developers Can Specify Input Parameters in Job Designer: Input Files, Date Range, Output Database, Etc

Analysts and Other Less Technical Users are Prompted for Input Parameters when they Run a Job.



Job Designer



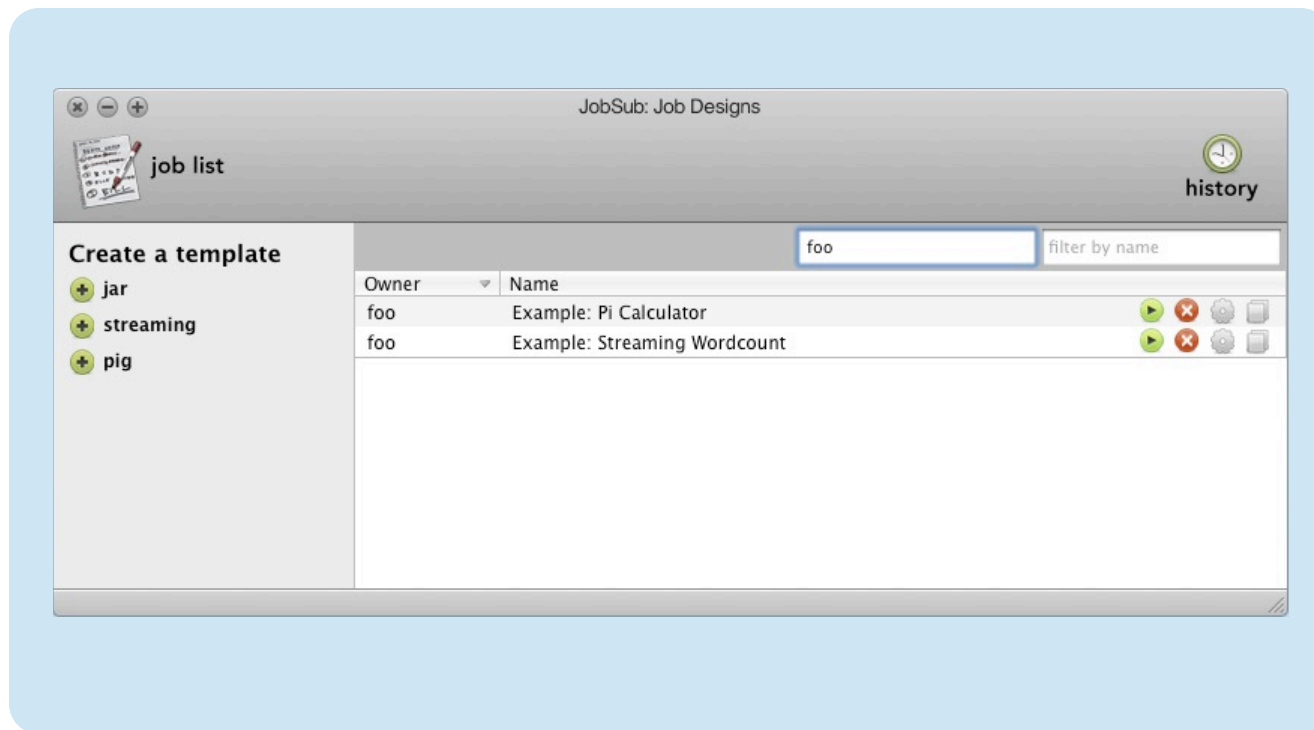
Looking More Closely at Desktop Enabling Less Technical Users

Developers Can Write Java, Streaming or Pig Jobs (more coming soon)

Developers Can Specify Input Parameters in Job Designer: Input Files, Date Range, Output Database, Etc

Analysts and Other Less Technical Users are Prompted for Input Parameters when they Run a Job.

Results Viewable in Web Browser



Job Designer

Cloudera Desktop

Product Details



- Currently Works with Cloudera's Distribution for Hadoop
 - CDH 2: Will Eventually Support All Releases of Hadoop
- Works with Hadoop Clusters in Your Data Center or the Cloud
- Available for Free: Can Modify Locally, No Distribution
- Window Manager is Open Source
 - Part of MooTools Javascript Framework
- Not a Replacement for Command Line Tools
 - Seasoned Developers may still use Familiar Tools
 - Useful for beginners, non-developers and regular developers

Cloudera Desktop API

Build Your Own Desktop Application



- Cloudera is Building a Reusable API for Developing Desktop Apps
- Framework Continues to Stabilize
- Working with a small number of Partners on New Applications
- Hope to Capture Innovation of Ecosystem in a Single Interface

- The API Alpha Will Open Sometime in the Next Few Months:
 - Interested?
 - Email: desktop-api-subscribe@cloudera.com

Hadoop: Learning More Resources from Cloudera

- Get Hadoop: <http://www.cloudera.com/hadoop>
 - Cloudera's Distribution for Hadoop
 - Automatic configuration
 - Easy deployment with standard Linux administration tools
- Get Desktop: <http://www.cloudera.com/desktop>
- Free Online Training: <http://www.cloudera.com/hadoop-training>
 - Complete with lectures and programming exercises
 - Pre-configured virtual machine to get started right away
- Blog: <http://www.cloudera.com/blog>
- Twitter: <http://twitter.com/cloudera>