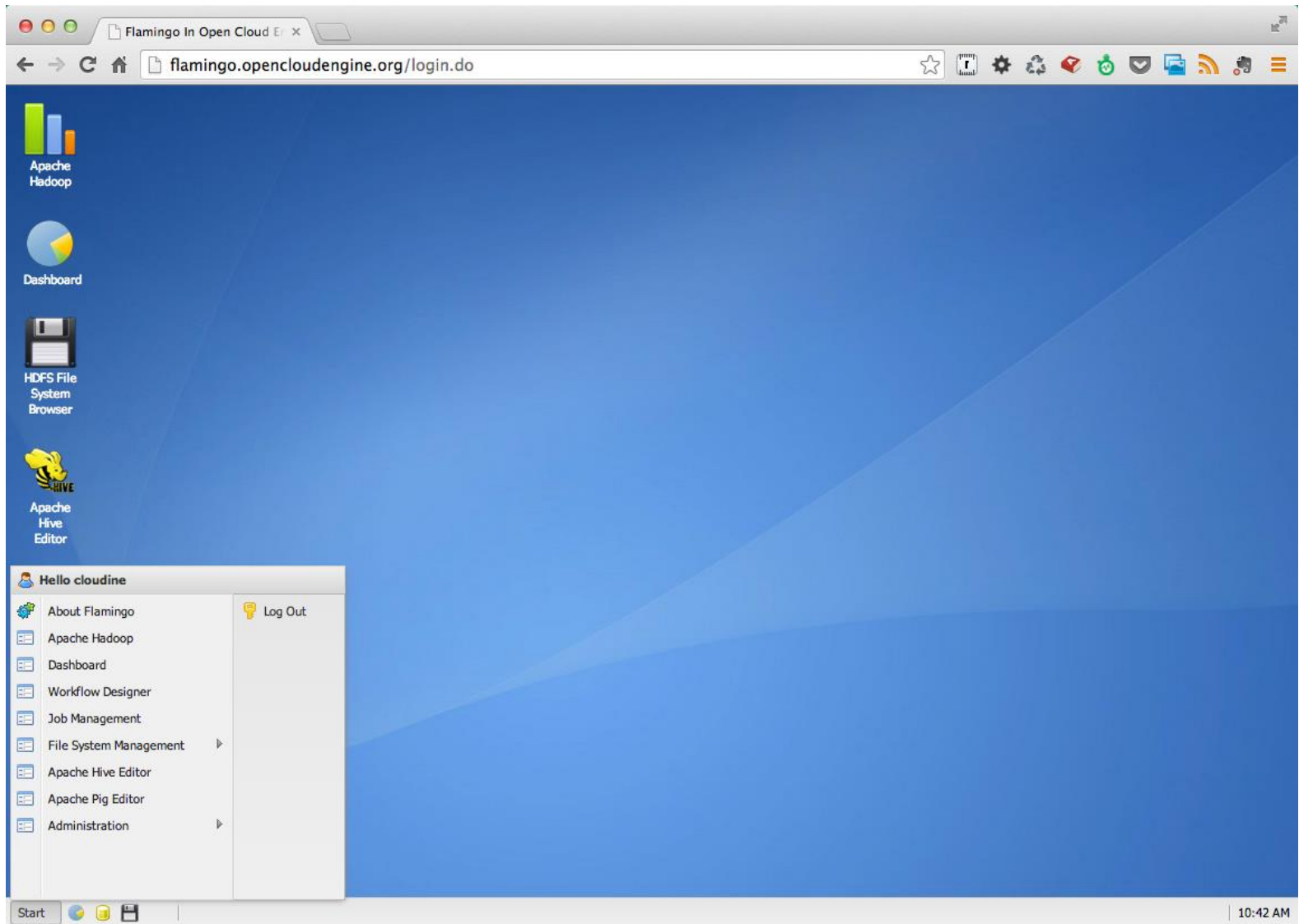


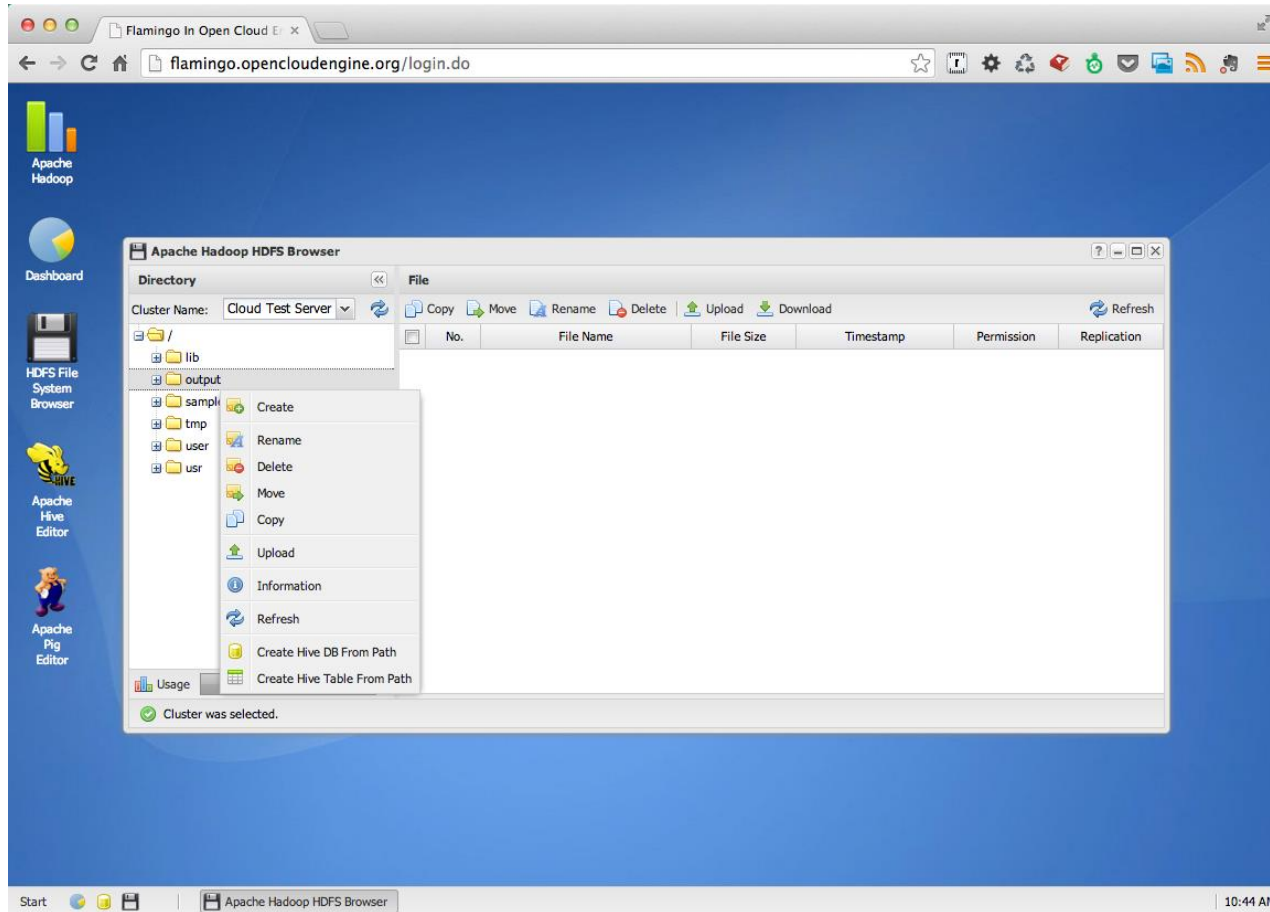
알기 쉬운 Hadoop 기반 빅데이터 플랫폼 아키텍처 및 활용 방안

NIPA Open Frontier Lab
박치완
chiwanpark@icloud.com

Flamingo 환경



HDFS 관리 화면 - 디렉토리 관리, 파일 업로드, 다운로드, Hive Metastore 관리



HDFS에 저장된 디렉토리 또는 파일의 상세 정보 열람

The screenshot displays the Apache Hadoop HDFS Browser interface. The main window shows a directory tree with the following structure:

- /
- lib
- output
- samples
- tmp
- user
- usr

The 'output' directory is selected, and a 'Directory Information' dialog box is open, showing the following details:

Basic

- Name: output
- Path: /output
- Type: File Directory
- Length: 4,210,008,894
- Modification: 2014-02-22 16:28:57

Permission

| | | | |
|--------|--|---|---|
| Owner: | <input checked="" type="checkbox"/> Read | <input checked="" type="checkbox"/> Write | <input checked="" type="checkbox"/> Execute |
| Group: | <input checked="" type="checkbox"/> Read | <input type="checkbox"/> Write | <input checked="" type="checkbox"/> Execute |
| Other: | <input checked="" type="checkbox"/> Read | <input type="checkbox"/> Write | <input checked="" type="checkbox"/> Execute |

Space

| | | | | | |
|------------------|----|------------------------|------|----------------------|----------|
| Block Size: | 0 | Disk Space Quota: | -1 | Disk Consumed Space: | 11.75 GB |
| Replication: | 0 | Number Of Directories: | 0 | | |
| Directory Quota: | -1 | Number Of Files: | 1835 | | |

The interface also shows a sidebar with navigation options: Apache Hadoop, Dashboard, HDFS File System Browser, Apache Hive Editor, and Apache Pig Editor. The status bar at the bottom indicates 'Ready' and '10:44 AM'.

HDFS Browser

HDFS의 파일 처리 이력을 기록해 추후 관리자가 특정 파일의 이력 추적 가능

Apache Hadoop HDFS Audit Log

Cluster Name: Cloud Test Server Start: End: Type: All Path: Find Clear

| No | User | File System | Type | Action | Path | Size | Date |
|----|-------|-------------|-----------|--------|-----------------------------|-------------|---------------------|
| 6 | admin | HDFS | Directory | Delete | /output/samples | 35,229,460 | 2014-02-24 10:48:13 |
| 5 | admin | HDFS | Directory | Delete | /output/samples/delicious | 51,803,127 | 2014-02-24 10:48:09 |
| 4 | admin | HDFS | Directory | Delete | /output/samples/apache-ooce | 341,016,036 | 2014-02-24 10:47:45 |
| 3 | admin | HDFS | Directory | Delete | /output/samples/ankus | 2,168,009 | 2014-02-24 10:47:38 |
| 2 | admin | HDFS | Directory | Copy | /samples » /output | 430,216,632 | 2014-02-24 10:47:28 |
| 1 | admin | HDFS | Directory | Delete | /folder | 4,168,918 | 2014-02-24 08:21:14 |

Page 1 of 1 | Displaying 1 - 6 of 6

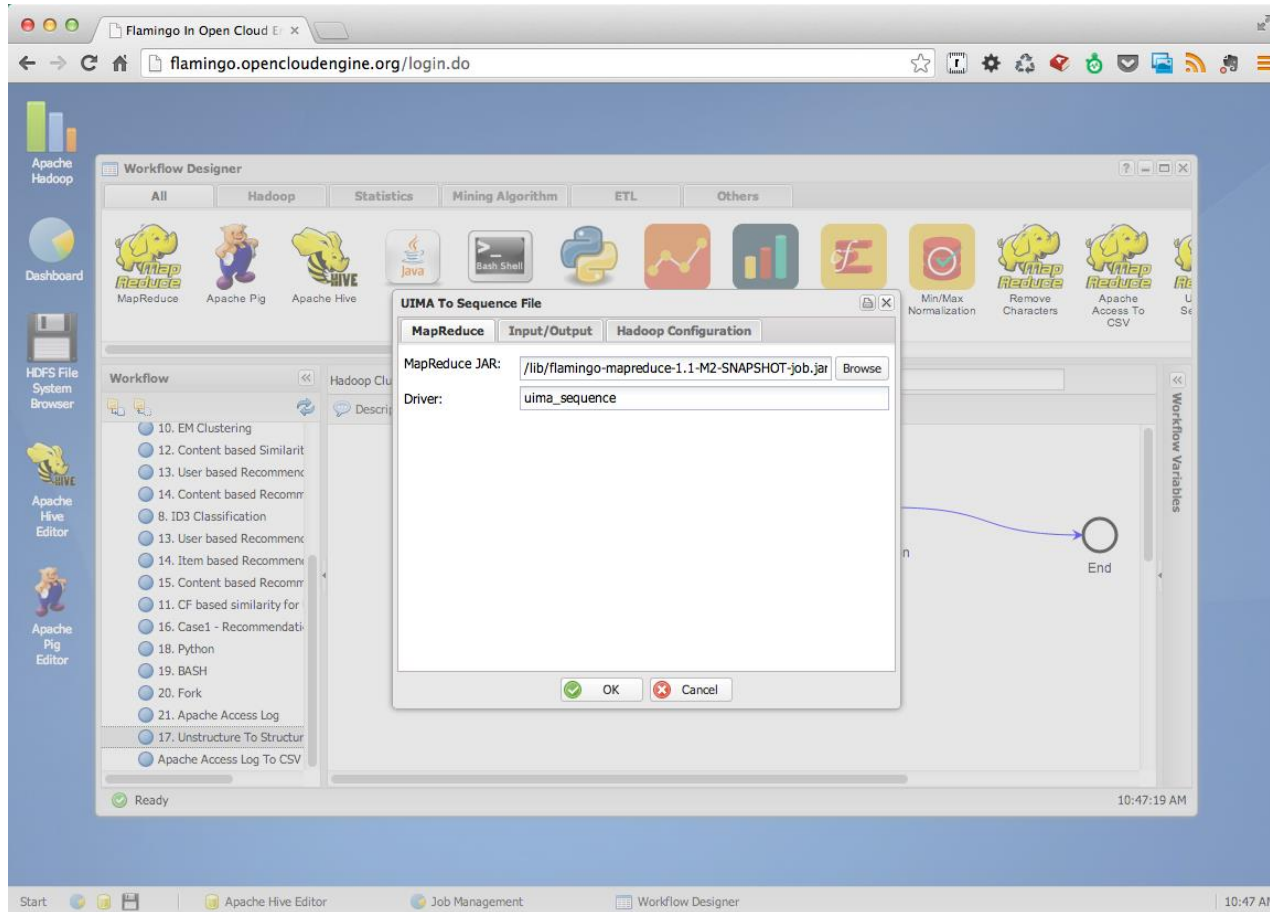
Workflow Designer

데이터 분석 Workflow를 DAG 기반으로 구성하여 작성
- 재사용 가능한 여러 컴포넌트 제공

The screenshot displays the Flamingo Workflow Designer interface within a web browser. The browser address bar shows 'flamingo.opencloudengine.org/login.do'. The interface includes a sidebar with navigation options like 'Apache Hadoop', 'Dashboard', 'HDFS File System Browser', and editors for 'Apache Hive', 'Apache Pig', and 'Apache Pig'. The main workspace is titled 'Workflow Designer' and shows a workflow named '17. Unstructure To Structure' on a 'Cloud Test Server' Hadoop cluster. The workflow is a Directed Acyclic Graph (DAG) starting with a 'Start' node, followed by a 'UIMA To Sequence File' component (MapReduce icon), then a 'UIMA Application' component (MapReduce icon), and finally an 'End' node. The workflow is currently in a 'Ready' state. The bottom of the screen shows a taskbar with 'Start', 'Apache Hive Editor', 'Job Management', and 'Workflow Designer' windows, along with the system time '10:47 AM'.

Workflow Designer

기존에 사용하던 MapReduce도 통합하여 Workflow에 적용 가능



Workflow Dashboard

수행된 Workflow의 기록 열람

The screenshot displays the Flamingo Workflow Dashboard in a web browser. The dashboard is titled "Dashboard" and shows the Hadoop Cluster as "Cloud Test Server". It features a "History" tab and a "Running Jobs" tab. The "History" tab is active, showing a table of workflow actions. The table has columns for No., Workflow Name, Action Name, Start, End, Elapsed, Progress, Status, and User. The actions listed are all completed with a status of "Success".

| No. | Workflow Name | Action Name | Start | End | Elapsed | Progress | Status | User |
|------|------------------------------|-------------|---------------------|---------------------|---------|----------|---------|-------|
| 1475 | 19. BASH | End | 2014-02-24 10:45:00 | 2014-02-24 10:45:02 | 0:02 | 100% | Success | admin |
| 1474 | Python Test | End | 2014-02-24 10:45:00 | 2014-02-24 10:45:00 | 0:00 | 100% | Success | admin |
| 1473 | 17. Unstructure To Structure | End | 2014-02-24 10:45:00 | 2014-02-24 10:45:21 | 0:21 | 100% | Success | admin |
| 1472 | Python Test | End | 2014-02-24 10:44:00 | 2014-02-24 10:44:00 | 0:00 | 100% | Success | admin |
| 1471 | 17. Unstructure To Structure | End | 2014-02-24 10:44:00 | 2014-02-24 10:44:23 | 0:23 | 100% | Success | admin |
| 1470 | 19. BASH | End | 2014-02-24 10:44:00 | 2014-02-24 10:44:02 | 0:02 | 100% | Success | admin |
| 1469 | Python Test | End | 2014-02-24 10:43:00 | 2014-02-24 10:43:00 | 0:00 | 100% | Success | admin |
| 1468 | 19. BASH | End | 2014-02-24 10:43:00 | 2014-02-24 10:43:02 | 0:02 | 100% | Success | admin |
| 1467 | 17. Unstructure To Structure | End | 2014-02-24 10:43:00 | 2014-02-24 10:43:23 | 0:23 | 100% | Success | admin |
| 1466 | 17. Unstructure To Structure | End | 2014-02-24 10:42:00 | 2014-02-24 10:42:22 | 0:22 | 100% | Success | admin |
| 1465 | 19. BASH | End | 2014-02-24 10:42:00 | 2014-02-24 10:42:02 | 0:02 | 100% | Success | admin |
| 1464 | Python Test | End | 2014-02-24 10:42:00 | 2014-02-24 10:42:00 | 0:00 | 100% | Success | admin |
| 1463 | 19. BASH | End | 2014-02-24 10:41:00 | 2014-02-24 10:41:02 | 0:02 | 100% | Success | admin |
| 1462 | 17. Unstructure To Structure | End | 2014-02-24 10:41:00 | 2014-02-24 10:41:21 | 0:21 | 100% | Success | admin |
| 1461 | Python Test | End | 2014-02-24 10:41:00 | 2014-02-24 10:41:00 | 0:00 | 100% | Success | admin |
| 1460 | 17. Unstructure To Structure | End | 2014-02-24 10:40:00 | 2014-02-24 10:40:22 | 0:22 | 100% | Success | admin |

Page 1 of 93 | Displaying 1 - 16 of 1475

Workflow Dashboard

Workflow내 개별 MapReduce Job 마다 수행 Log, Configuration 확인

The screenshot displays the Flamingo Workflow Dashboard interface. A modal window titled "17. Unstructure To Structure - WF_20140214_307740502" is open, showing a table of job actions and their details.

| Workflow ID | Action Name | Start | End | Elapsed | Status |
|-----------------------|-----------------------|---------------------|---------------------|---------|---------|
| WF_20140214_307740502 | End | 2014-02-24 10:45:21 | 2014-02-24 10:45:21 | 0:00 | Success |
| WF_20140214_307740502 | UIMA Application | 2014-02-24 10:45:09 | 2014-02-24 10:45:21 | 0:12 | Success |
| WF_20140214_307740502 | UIMA To Sequence File | 2014-02-24 10:45:00 | 2014-02-24 10:45:09 | 0:09 | Success |

Action Information

Action ID: 5123 Start: 2014-02-24 10:45:09 Elapsed: 0:12
Job ID: 809275413 End: 2014-02-24 10:45:21 Status: Success
Workflow ID: WF_20140214_307740502 Action Name: UIMA Application
Log Path: /tmp/2014/02/24/64/JOB_20140224_104500_64_176163649/953552195/action.log

Log

```
180 14/02/24 10:45:21 INFO mapred.JobClient: org.apache.uima.SentenceAnnotation=9
181 14/02/24 10:45:21 INFO mapred.JobClient: org.apache.uima.TokenAnnotation=14850
182 14/02/24 10:45:21 INFO mapred.JobClient: Map-Reduce Framework
183 14/02/24 10:45:21 INFO mapred.JobClient: Map input records=1
184 14/02/24 10:45:21 INFO mapred.JobClient: Physical memory (bytes) snapshot=232427520
185 14/02/24 10:45:21 INFO mapred.JobClient: Spilled Records=0
186 14/02/24 10:45:21 INFO mapred.JobClient: CPU time spent (ms)=4010
187 14/02/24 10:45:21 INFO mapred.JobClient: Total committed heap usage (bytes)=185729024
188 14/02/24 10:45:21 INFO mapred.JobClient: Virtual memory (bytes) snapshot=3334606848
189 14/02/24 10:45:21 INFO mapred.JobClient: Map input bytes=4833
190 14/02/24 10:45:21 INFO mapred.JobClient: Map output records=1
191 14/02/24 10:45:21 INFO mapred.JobClient: SPLIT_RAW_BYTES=252
192 14/02/24 10:45:21 INFO uima.UIMADriver: UIMADriver completed. Timing: 10554 ms
193
```

구성된 Workflow를 Batch Job으로 등록

The screenshot shows the Flamingo Job Management interface. A dialog box titled "Register batch job" is open, displaying the following information:

Job Information

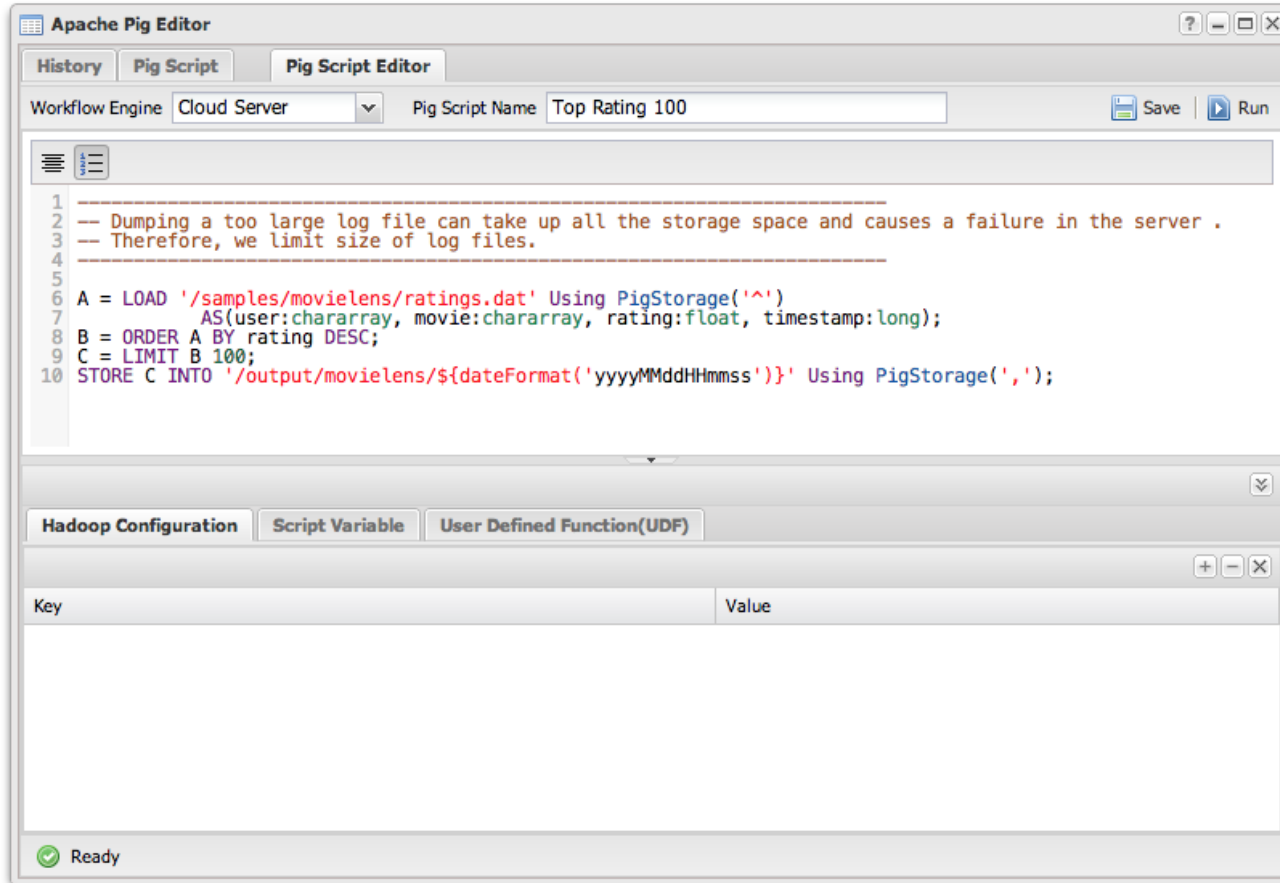
Job Name: 4. Normalization
Cron Expression: 0 0 * * * *
Workflow Identifier: WF_20131125_199850452
Author: admin
Recently Changed Date: 2013-11-25 17:00:03
Identifier: 12
Status Code: REGISTERED

Workflow XML

```
1 <?xml version="1.0" encoding="UTF-8"?>
2 <workflow xmlns="http://www.openflamingo.org/schema/workflow" workflowName="4. Normalization">
3 <description/>
4 <start name="OG_4106_2" description="Start" te="OG_4106_8"/>
5 <action name="OG_4106_8" description="Min/Max Normalization" to="OG_4106_5">
6 <mapreduce>
7 <jar>org.ankus:ankus-core:0.1</jar>
8 <className>Normalization</className>
9 <command>
10 <variable value="-input"/>
11 <variable value="/samples/ankus/Clustering/KMeans/iris.txt"/>
12 <variable value="-output"/>
13 <variable value="/output/20131128/04-01"/>
14 <variable value="-indexList"/>
15 <variable value="1"/>
16 <variable value="-exceptionIndexList"/>
17 <variable value="4"/>
18 <variable value="-remainAllFields"/>
```

Apache Pig Integration

Apache Pig를 통합해, Pig Latin을 통해 추상화 된 데이터 분석



The screenshot shows the Apache Pig Editor interface. The main window displays a Pig Latin script for finding the top 100 movies by rating. The script is as follows:

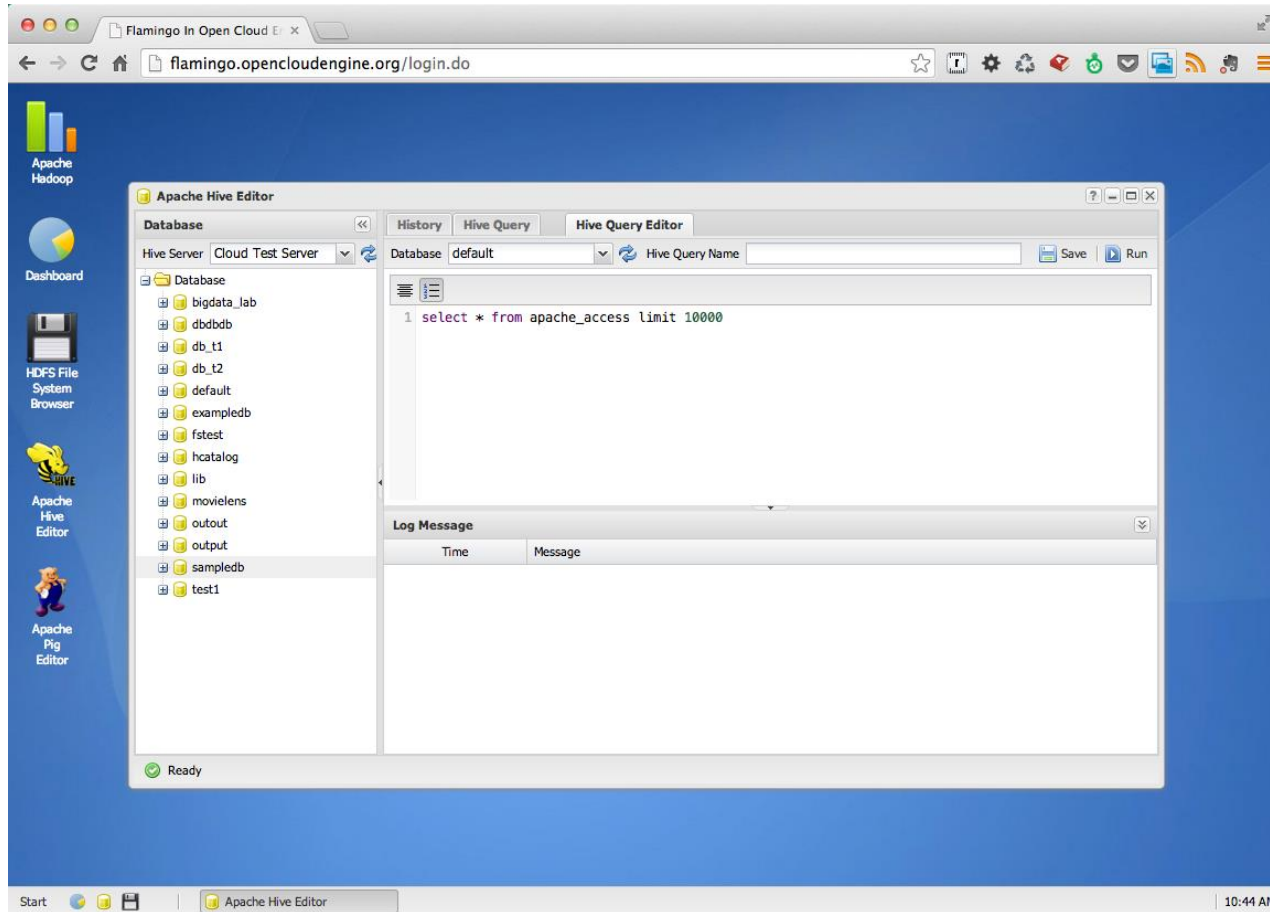
```
1
2 -- Dumping a too large log file can take up all the storage space and causes a failure in the server .
3 -- Therefore, we limit size of log files.
4
5
6 A = LOAD '/samples/movielens/ratings.dat' Using PigStorage('^')
7     AS(user:chararray, movie:chararray, rating:float, timestamp:long);
8 B = ORDER A BY rating DESC;
9 C = LIMIT B 100;
10 STORE C INTO '/output/movielens/${dateFormat('yyyyMMddHHmmss')} ' Using PigStorage(',');
```

Below the script editor, there are tabs for 'Hadoop Configuration', 'Script Variable', and 'User Defined Function(UDF)'. The 'Hadoop Configuration' tab is active, showing a table with 'Key' and 'Value' columns. The status bar at the bottom indicates 'Ready'.

| Key | Value |
|-----|-------|
|-----|-------|

Apache Hive Integration

Apache Hive를 통합해, HiveQL을 통해 SQL과 비슷한 구문으로 데이터 분석 수행



Apache Hive Integration

수행된 Query의 결과를 즉시 열람

The screenshot displays the Apache Hive Editor web interface. The left sidebar contains navigation icons for Apache Hadoop, Dashboard, HDFS File System Browser, Apache Hive Editor, and Apache Pig Editor. The main content area is divided into several sections:

- Database:** A tree view showing the database structure, including 'bigdata_lab', 'dbbdb', 'db_t1', 'db_t2', 'default', 'exampledb', 'fstest', 'hcatalog', 'lib', 'movieiens', 'output', 'sampledb', and 'test1'.
- History:** A table listing recent queries with columns for Execution ID, Database, Query, Length, Elapsed time, Status, and Start time.
- Detail:** A section for viewing the details of a selected query, including the Hive Query and the Result.

The History table shows two successful queries:

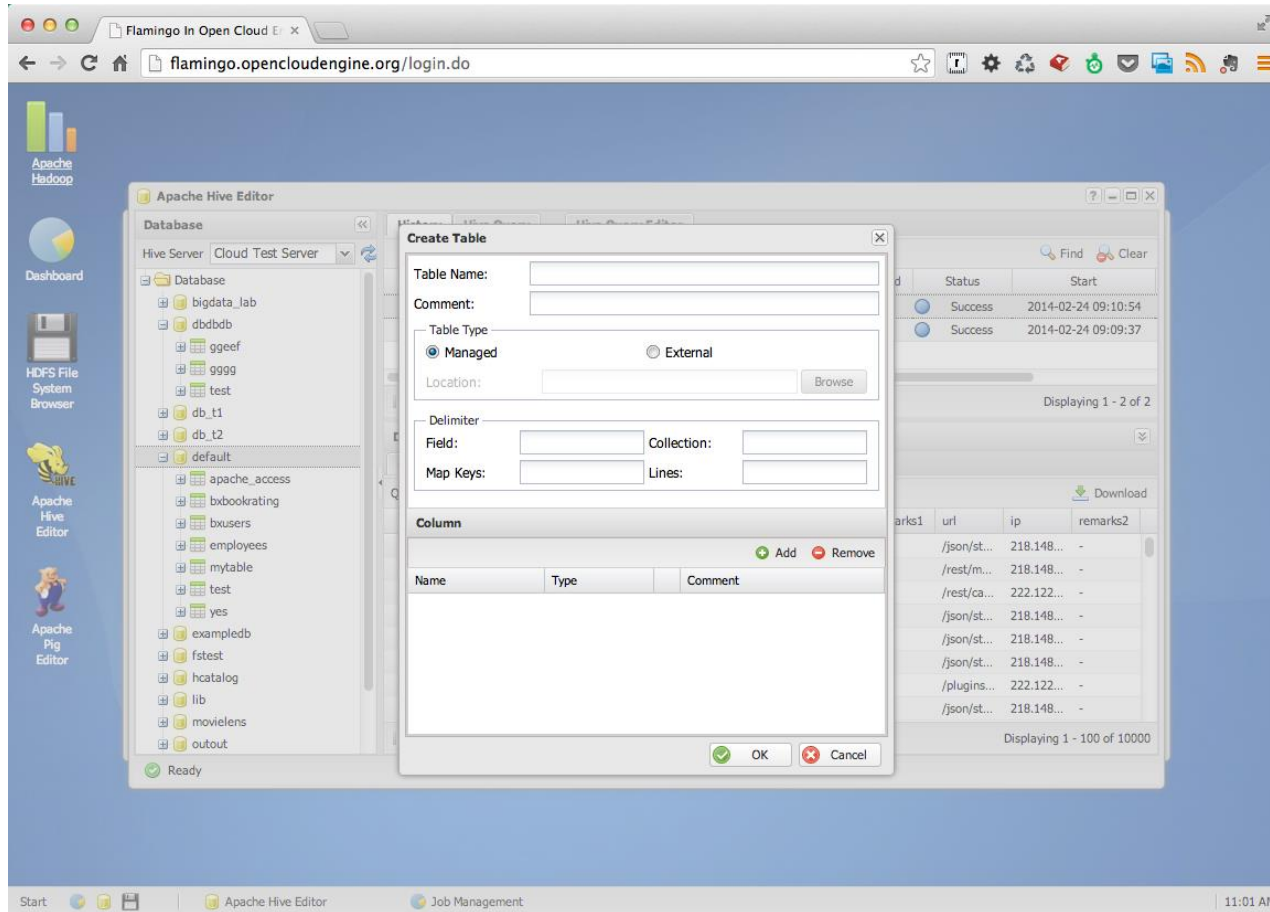
| Execution ID | Database | Qu... | Length | Elapsed | Status | Start |
|--------------------------|----------|--------|-----------|---------|---------|---------------------|
| 20140224091053_966372834 | default | sel... | 2,866,419 | 0:05 | Success | 2014-02-24 09:10:54 |
| 20140224090936_933163063 | default | sel... | 31,529 | 0:00 | Success | 2014-02-24 09:09:37 |

The Detail section shows the Result of a query, which is a table of HTTP requests:

| No | timesta... | remarks3 | status | length | method | agent | remarks1 | url | ip | remarks2 |
|----|------------|-------------|--------|--------|--------|-------------|----------|-------------|------------|----------|
| 1 | 20/Jan/... | http://w... | 200 | 503 | POST | Mozilla/... | - | /json/st... | 218.148... | - |
| 2 | 20/Jan/... | http://w... | 200 | 410 | GET | Mozilla/... | - | /rest/m... | 218.148... | - |
| 3 | 20/Jan/... | - | 200 | 1084 | GET | JIRA-6... | - | /rest/ca... | 222.122... | - |
| 4 | 20/Jan/... | http://w... | 200 | 503 | POST | Mozilla/... | - | /json/st... | 218.148... | - |
| 5 | 20/Jan/... | http://w... | 200 | 503 | POST | Mozilla/... | - | /json/st... | 218.148... | - |
| 6 | 20/Jan/... | http://w... | 200 | 503 | POST | Mozilla/... | - | /json/st... | 218.148... | - |
| 7 | 20/Jan/... | - | 200 | 400 | GET | Java/1.7... | - | /plugins... | 222.122... | - |
| 8 | 20/Jan/... | http://w... | 200 | 503 | POST | Mozilla/... | - | /json/st... | 218.148... | - |

Apache Hive Integration

Hive Table를 Query문 없이 UI 만으로 생성

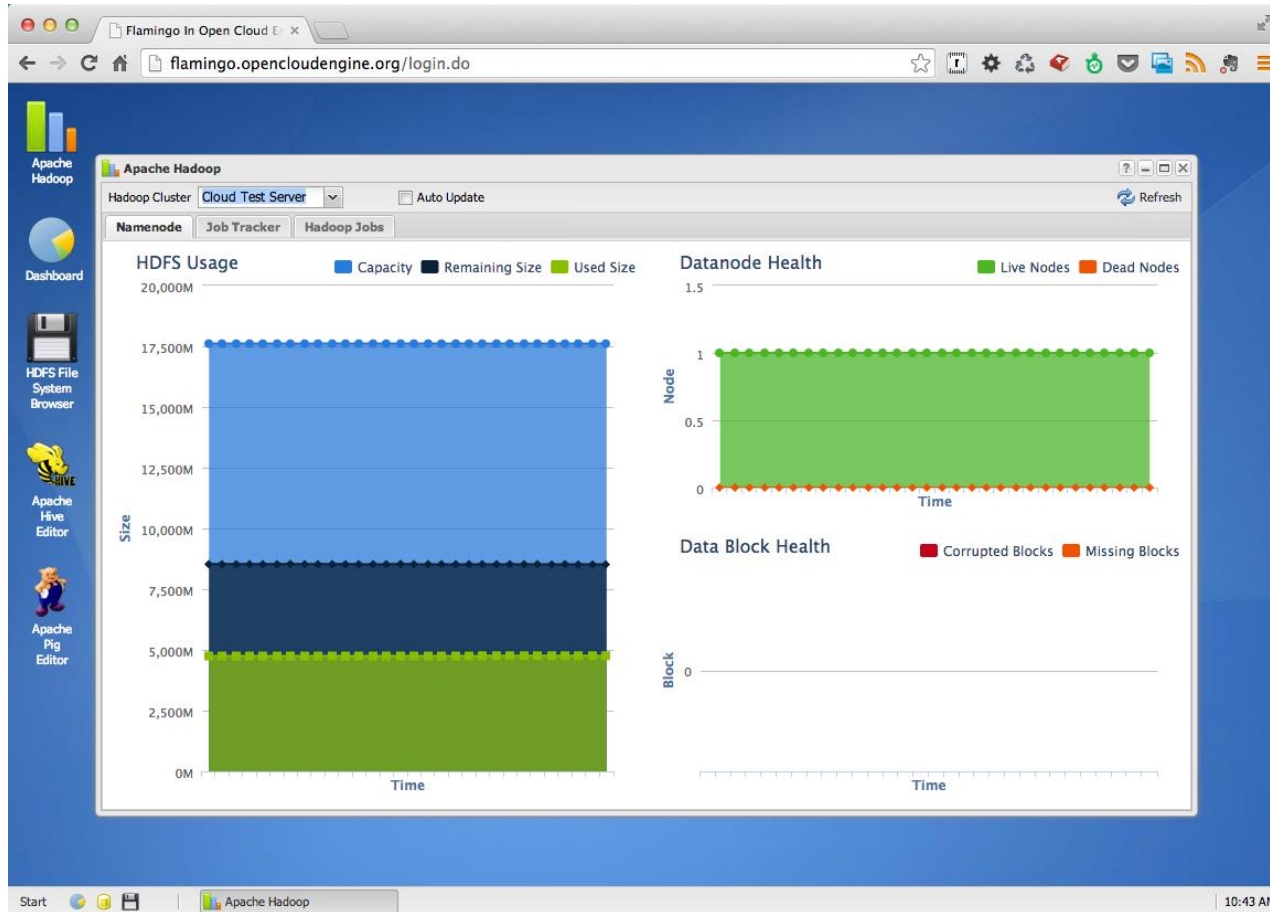


The screenshot displays the Apache Hive Editor interface within a web browser. The main window shows a tree view of databases and tables. A 'Create Table' dialog box is open, allowing users to define a new table without writing a query. The dialog includes the following fields and options:

- Table Name:** A text input field.
- Comment:** A text input field.
- Table Type:** Radio buttons for 'Managed' (selected) and 'External'.
- Location:** A text input field with a 'Browse' button.
- Delimiter:** A section with 'Field:' and 'Collection:' input fields.
- Map Keys:** A section with 'Map Keys:' and 'Lines:' input fields.
- Column:** A table with columns 'Name', 'Type', and 'Comment'. It includes 'Add' and 'Remove' buttons.

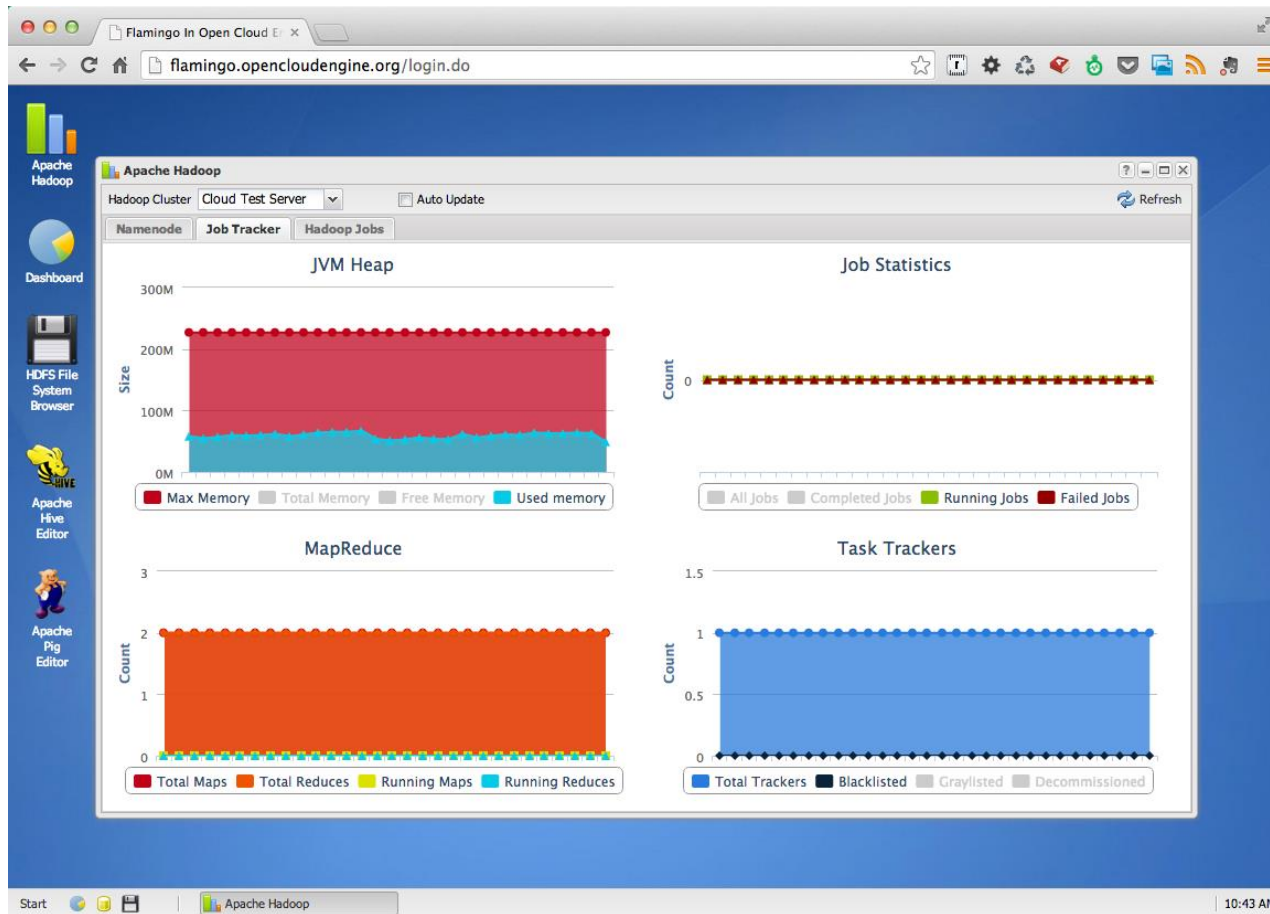
The background interface shows a sidebar with navigation icons for Apache Hadoop, Dashboard, HDFS File System Browser, Apache Hive Editor, and Apache Pig Editor. The main area displays a tree view of databases and tables, including 'bigdata_lab', 'dbdbdb', 'ggeef', '9999', 'test', 'db_t1', 'db_t2', 'default', 'apache_access', 'bxbookrating', 'bxusers', 'employees', 'mytable', 'test', 'yes', 'exampledb', 'fstest', 'hcatalog', 'lib', 'movielens', and 'outout'. The status bar at the bottom indicates 'Ready' and 'Job Management'.

다양한 Monitoring 기능 제공 - HDFS, Datanode Monitoring



Monitoring

JobTracker, Namenode 등 주요 Hadoop Cluster의 Metrics 제공



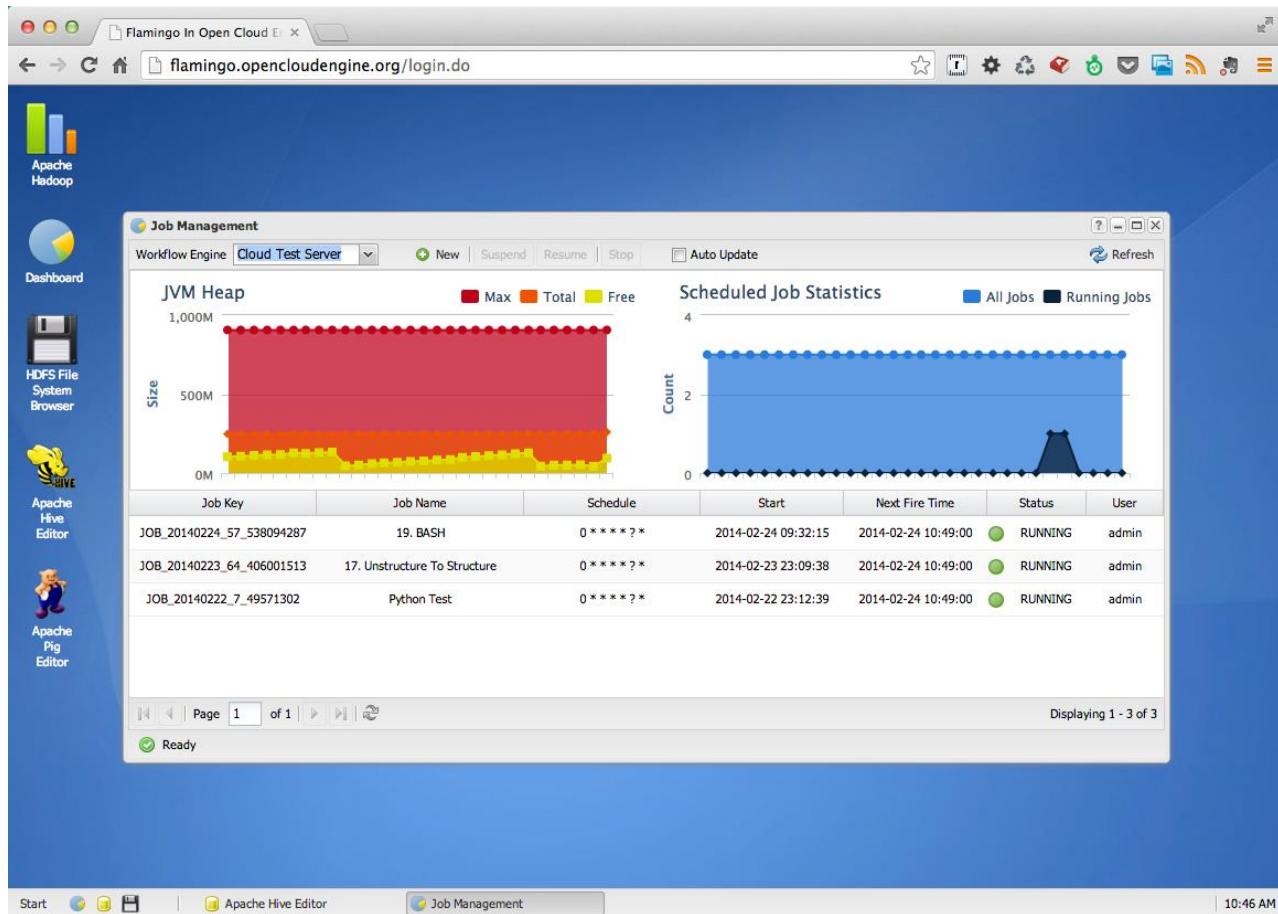
수행한 MapReduce Job에 대한 Monitoring 정보 제공

The screenshot shows the Apache Hadoop Job Tracker interface for a cluster named 'Cloud Test Server'. The interface displays a list of 17 MapReduce jobs. The first job is in a 'Running' state, while the others are 'Success'. The table below provides detailed information for each job.

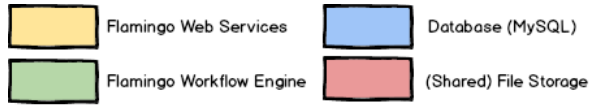
| No. | Hadoop Job ID | Hadoop Job Name | Start | End | Elapsed | Map | Reduce | Status | Hadoop User |
|-----|-----------------------|---------------------------|---------------------|---------------------|---------|------|--------|---------|-------------|
| 1 | job_201402221048_1696 | UIMA Application : /li... | 2014-02-24 10:45:11 | | 0:06 | 0% | 0% | Running | root |
| 2 | job_201402221048_1695 | UIMA : Text To Sequ... | 2014-02-24 10:45:02 | 2014-02-24 10:45:08 | 0:06 | 100% | 100% | Success | root |
| 3 | job_201402221048_1694 | UIMA Application : /li... | 2014-02-24 10:44:12 | 2014-02-24 10:44:22 | 0:10 | 100% | 100% | Success | root |
| 4 | job_201402221048_1693 | UIMA : Text To Sequ... | 2014-02-24 10:44:02 | 2014-02-24 10:44:09 | 0:07 | 100% | 100% | Success | root |
| 5 | job_201402221048_1692 | UIMA Application : /li... | 2014-02-24 10:43:12 | 2014-02-24 10:43:22 | 0:10 | 100% | 100% | Success | root |
| 6 | job_201402221048_1691 | UIMA : Text To Sequ... | 2014-02-24 10:43:02 | 2014-02-24 10:43:09 | 0:07 | 100% | 100% | Success | root |
| 7 | job_201402221048_1690 | UIMA Application : /li... | 2014-02-24 10:42:11 | 2014-02-24 10:42:21 | 0:10 | 100% | 100% | Success | root |
| 8 | job_201402221048_1689 | UIMA : Text To Sequ... | 2014-02-24 10:42:02 | 2014-02-24 10:42:09 | 0:06 | 100% | 100% | Success | root |
| 9 | job_201402221048_1688 | UIMA Application : /li... | 2014-02-24 10:41:11 | 2014-02-24 10:41:20 | 0:09 | 100% | 100% | Success | root |
| 10 | job_201402221048_1687 | UIMA : Text To Sequ... | 2014-02-24 10:41:02 | 2014-02-24 10:41:08 | 0:06 | 100% | 100% | Success | root |
| 11 | job_201402221048_1686 | UIMA Application : /li... | 2014-02-24 10:40:12 | 2014-02-24 10:40:21 | 0:09 | 100% | 100% | Success | root |
| 12 | job_201402221048_1685 | UIMA : Text To Sequ... | 2014-02-24 10:40:02 | 2014-02-24 10:40:09 | 0:07 | 100% | 100% | Success | root |
| 13 | job_201402221048_1684 | UIMA Application : /li... | 2014-02-24 10:39:11 | 2014-02-24 10:39:21 | 0:09 | 100% | 100% | Success | root |
| 14 | job_201402221048_1683 | UIMA : Text To Sequ... | 2014-02-24 10:39:02 | 2014-02-24 10:39:08 | 0:06 | 100% | 100% | Success | root |
| 15 | job_201402221048_1682 | UIMA Application : /li... | 2014-02-24 10:38:11 | 2014-02-24 10:38:21 | 0:09 | 100% | 100% | Success | root |
| 16 | job_201402221048_1681 | UIMA : Text To Sequ... | 2014-02-24 10:38:02 | 2014-02-24 10:38:08 | 0:06 | 100% | 100% | Success | root |
| 17 | job_201402221048_1680 | UIMA Application : /li... | 2014-02-24 10:37:11 | 2014-02-24 10:37:20 | 0:09 | 100% | 100% | Success | root |

Monitoring

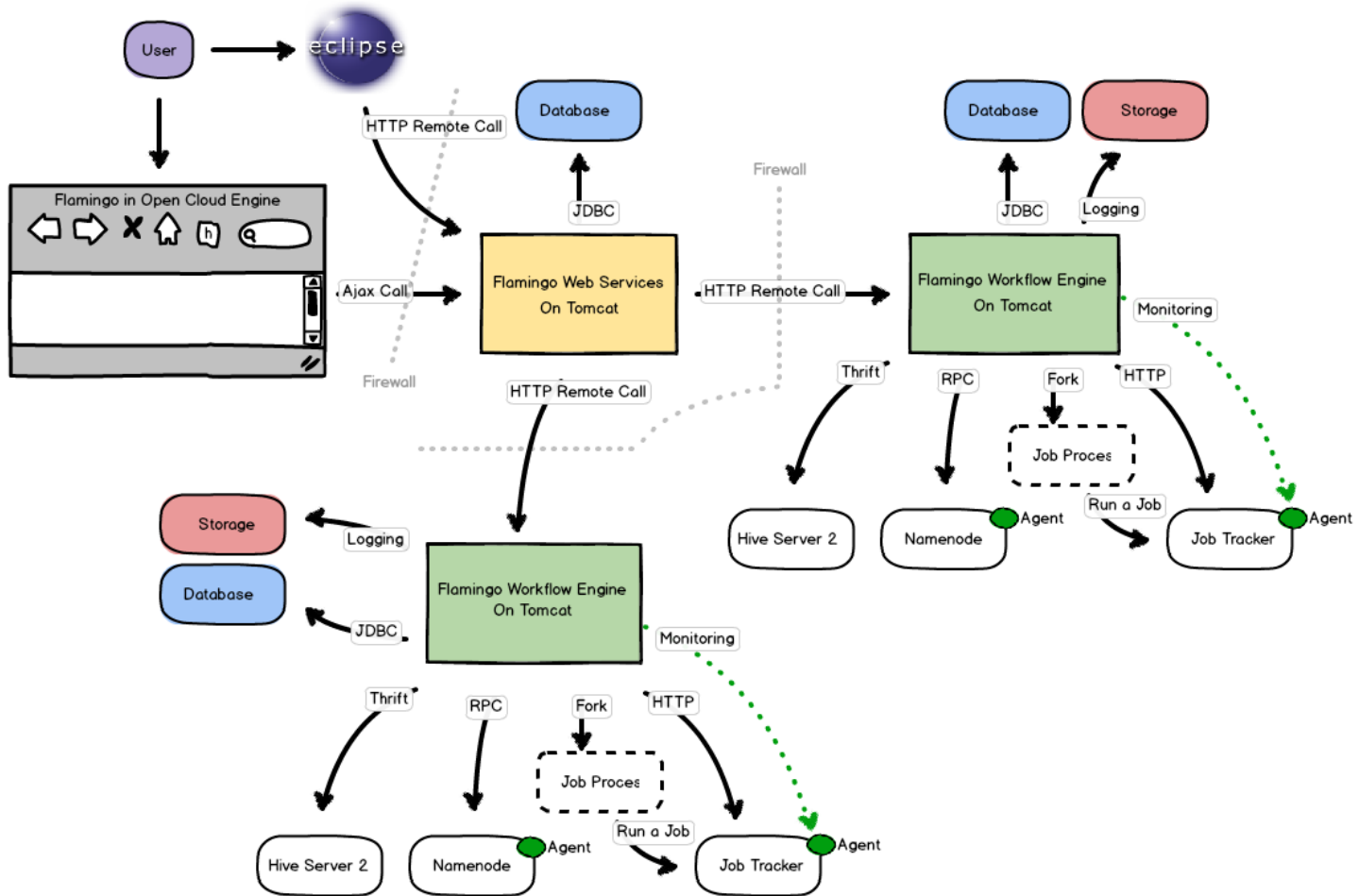
JobTracker, Namenode 등 주요 Hadoop Cluster의 Metrics 제공



Architecture



Architecture for Multi Hadoop Cluster



Future

- 기본적으로 제공되는 Component 보강
 - 데이터 전처리 모듈, 추론 모듈 등
- 사용자 별 Quota 설정 등 개별 제한 기능
- Hadoop 2 지원
- Amazon EMR, Rackspace Hadoop Platform 등 Hadoop 기반 엔터프라이즈 플랫폼 지원

Project Information

- 프로젝트 홈페이지
 - <http://wiki.opencloudengine.org/display/IN/Flamingo>
- Issue Tracker
 - <http://jira.opencloudengine.org>
- Build Server
 - <http://build.opencloudengine.org>
- License
 - Web UI (Ext.JS를 활용해 GPLv3)
 - Engine (Apache License)

DEMO

Q&A