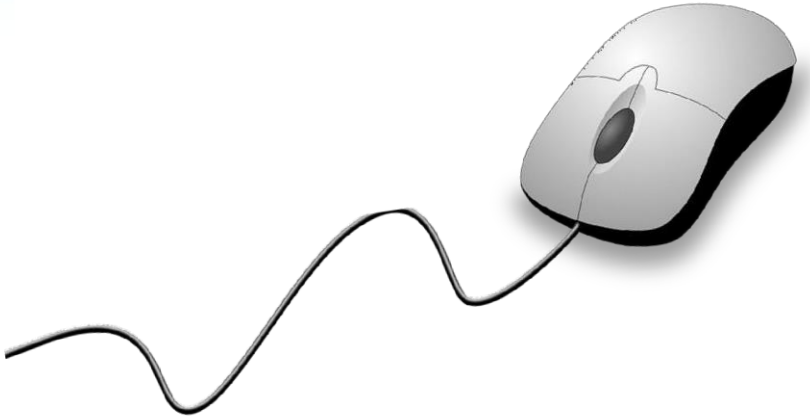
공개SW 솔루션 설치 & 활용 가이드

시스템SW > 데이터관리



ankus 제대로 배워보자

How to Use Open Source Software

Open Source Software Installation & Application Guide

OPENUP

오픈소스 소프트웨어 통합지원센터

Open Source Software Support Center

CONTENTS

1. 개요
2. 기능요약
3. 실행환경
4. 설치 및 실행
5.기능소개
6 . 활용예제
7 . FAQ
8 . 용어정리

1. 개요

소개	• 데이터 마이닝 / 기계 학습(Machine Learning) 지원 분석 도구		
주요기능	• 기초 통계, 분류/군집, 유사/상관, 개인화 추천 등 25개 분석 알고리즘 구성 • Hadoop MapReduce 기반 빅데이터 분산병렬 처리 지원 • 데이터 마이닝/기계 학습 기반 고차원 지능형 예측 분석 지원		
대분류	• 시스템 SW	**소분류**	• 데이터관리
라이선스 형태	• GPL v3	**사전설치 솔루션**	• CentOS 6.x • Jdk 1.7.x • Hadoop 2.6.x • Cubrid 9.3.x
실행 하드웨어	• Intel, AMD, PowerPC 등 • 램 16GB 이상의 하드웨어 (Hadoop cluster 구성서버3대이상) • 최소 100GB 이상의 디스크 공간	**버전**	• 1.0 (2018년 10월 기준)
특징	• 분석 알고리즘과 함수로 구성된 분석 라이브러리(Library) • Hadoop(대용량 분산 병렬 처리 시스템) 기반 최적화 구현 • 데이터 마이닝 / 머신 러닝 기능 지원 • 제품/서비스에 분석 기능 필요 시 활용 적합 (예측 분석, 개인화 추천 지원)		
보안취약점	• N/A		
개발회사/커뮤니티	• ONYCOM, INC. / ankus Community		
공식 홈페이지	• http://openankus.org		

1. 개요

- ankus Analyzer G 개요

★ 분석 알고리즘과 함수로 구성된 분석 라이브러리(Library)

★ Hadoop(대용량 분산 병렬 처리 시스템) 기반 최적화 구현

★ 데이터마이닝/머신러닝 기능 지원

★ 제품/서비스에 분석 기능 필요 시 활용 적합 (예측 분석, 개인화 추천 지원)

2. 기능요약

- **ankus Analyzer G 의 주요 기능**

주요기능	지원여부
데이터 마이닝	지원
머신 러닝 기능	지원
Hadoop	지원
기초 통계알고리즘	지원
분류/군집 알고리즘	지원
유사/상관 알고리즘	지원
개인화 추천 알고리즘	지원
워크플로우 기능	지원

3. 실행환경

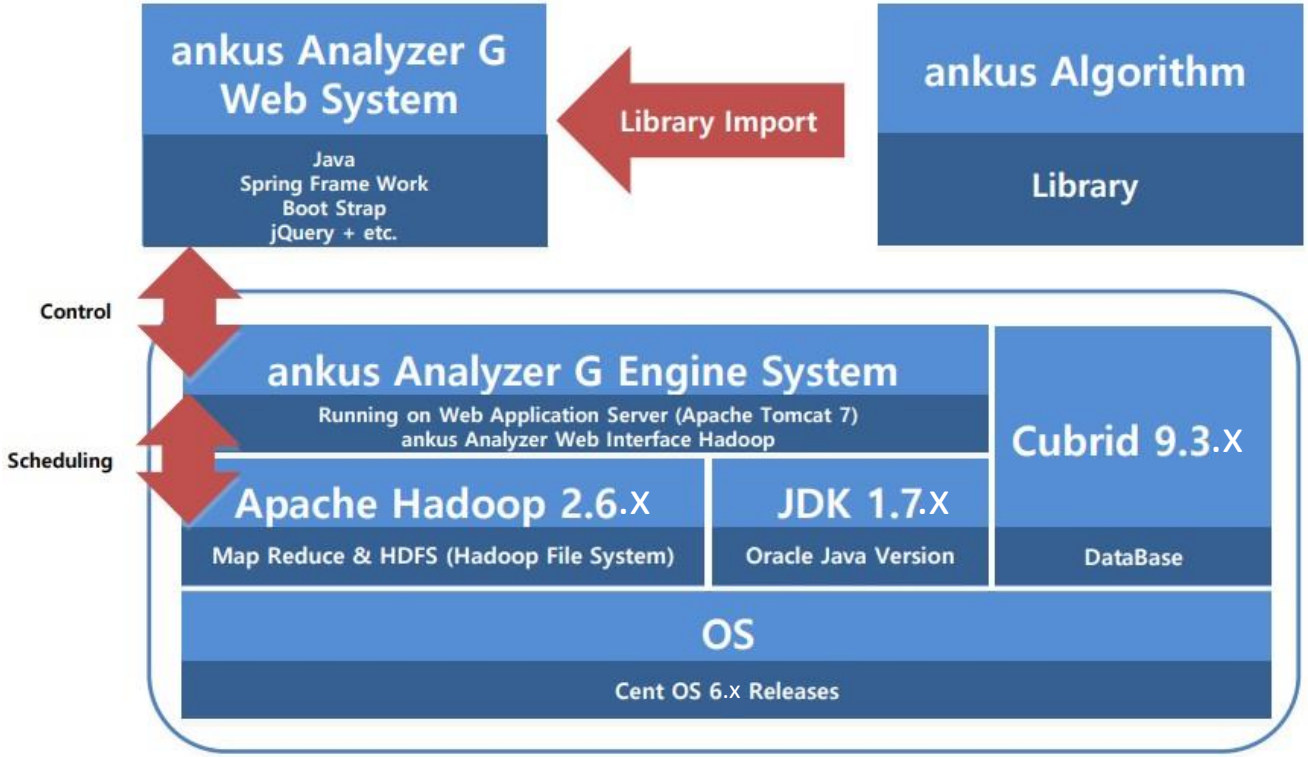
- **설치 요구 사항**

1. OS

- CentOS Linux 6.x

2. 설치 솔루션

- Jdk 1.7.x
- Hadoop 2.6.x
- Cubrid 9.3.x



ankus Analyzer G System Stack

3. H/W

- Hadoop 클러스터 환경 구축을 위한 최소 3대 이상의 PC

4. 설치 및 실행

세부 목차

1. 설치 이미지 준비
2. 설치 파일의 구성 및 개요
3. 서버 설정 및 Hadoop 설치
4. ankus Analyzer G 엔진 모듈 설치 및 Cubrid 설치
5. ankus Analyzer G 웹 모듈 설치
6. 설치 완료

4. 설치 및 실행

4.1 설치 파일 준비

- 본 가이드은 CentOS 6.x 일반 데스크탑 모드로 설치된 서버 환경에서 작성 및 확인

- 필요에 따라 일부 과정을 생략하고 진행할 수 있으며, 이 경우 컴퓨터 설정에 따라 추가적인 작업이 필요할 수 있음

  - 본 가이드은 멀티 노드에서의 Hadoop 설치를 가정하고 작성

- ankus Analyzer G 는 Hadoop 과 Cubrid 가 필요하며,

  - 본 가이드에는 Hadoop 과 Cubrid 설치과정이 포함되어 있음

4. 설치 및 실행

4.2 설치 파일의 구성 및 개요

- ankus
  - 01. hadoop-install
    - 0_server_setting.sh
    - 1_ssh_setting.sh
    - 2_java_install.sh
    - 3_hadoop_install.sh
    - config.sh
    - core-site.xml
    - hadoop-2.6.0.tar.gz
    - hdfs-site.xml
    - jdk-8u131-linux-x64.tar.gz
    - jdk1.7.0_67-ankus.tar.gz
    - mapred-site.xml
    - yarn-site.xml
  - 02. ankus-engine-install
    - 01_cubrid_installdb.sh
    - 02_ankus-engine.sh
    - ankus_cu.sql
    - ankus-site-engine.xml
    - CUBRID-9.3.6.0002-linux.x86_64.sh
    - CUBRIDManager-9.3.6.007-linux-x86_64.tar.gz
    - engine-ROOT.war
    - tomcat_server_28080.xml
    - tomcat-7.0.55-ankus.tar.gz
  - 03. ankus-web-install
    - ankus-library
    - 01_ankus-web.sh
    - ankus-site-web.xml
    - tomcat_server_8080.xml
    - tomcat-7.0.55-ankus.tar.gz
    - web-ROOT.war

"01. 서버 설정 및 Hadoop 설치" 에서 사용될 정보 파일 및 쉘 스크립트 실행 파일들

"02. ankus 엔진 모듈 설치"에서 사용될 정보 파일들 및 쉘 스크립트 실행 파일들

"03. ankus 웹 모듈 설치" 에서 사용될 정보 파일들 및 쉘 스크립트 실행 파일들

4.3 서버 설정 및 Hadoop 설치(1/4)

- **실행 요건**

  - 서버의 호스트 정보 및 방화벽 서비스 등 운영체제 관련 정보 설정을 위해 모든 서버에서 'root'계정으로 본 설정을 실행
  - 쉘 스크립트 파일 실행을 위해 모든 .sh파일에 실행 권한을 부여해야 함 (예) chmod +x *.sh)
  - ankus Analyzer G 설치 및 운영을 위한 uid 로 'ankus' 라는 계정을 생성

```
[root@localhost /]$ adduser ankus
```

4. 설치 및 실행

4.3 서버 설정 및 Hadoop 설치(2/4)

- **운영체제 내 호스트 정보에 사용할 서버들의 IP 및 호스트명 추가**

  - /etc/hosts 파일에 Hadoop 클러스터로 운영될 모든 노드의 정보를 추가

  - ip 와 호스트명을 자신의 정보에 맞춰 입력

```
[ankus@test-master 01. hadoop-install]$ vi /etc/hosts
86.             master
87.             node1
88.             node2
```

- **서버의 방화벽 서비스 중지 (Hadoop 및 ankus 를 위한 네트워크 포트 열기)**

  - 본 가이드에서는 네트워크 포트허용을 일괄적으로 수행하기 위하여 방화벽 서비스를 중지함

```
[ankus@test-master 01. hadoop-install]$ /etc/init.d/iptables stop  iptabl
es: Setting chains to pilicy ACCEPT: filter                    [    OK    ]
iptables: Flushing firewall rules:                             [    OK    ]
iptables: Unloading modules:                                   [    OK    ]
```

  * 관리 정책에 따라 방화벽 서비스를 중지 할 수 없는 경우, Hadoop 및 ankus에서 사용하는 네트워크 포트를 방화벽에서 허용하도록 개별적으로 설정해야 함

4.3 서버 설정 및 Hadoop 설치(3/4)

- **1단계) 서버 세팅 파일 작성/수정 – config.sh**

  - Hadoop을 설치할 (Hadoop의 마스터 노드 및 슬레이브 노드들로 사용할) 서버들의 IP주소 및 호스트명 정보가 다음과 같이 적혀있는 config.sh 파일을 본인의 정보에 맞게 수정

```
uid=ankus  node_mas
ter=master
node_slaves=( 'master" "node1" "node2")
node_ips=("172.30.80.86" "172.30.80.87" "172.30.80.88")
echo [info] uid and master/slave names are set
```

- **2단계) 쉘 스크립트 파일을 실행하여 설정 진행**
  - 0_server_setting.sh 파일 실행 – **root 계정**
  - 1_ssh_setting.sh 파일 실행 – **ankus 계정**

마스터 노드 설정 부분

슬레이브 노드의 서버 수 만큼 반복

```
[ankus@test-master 01. hadoop-install]$ ./1_ssh_setting.sh
[info] uid and master/slave names are set..
Generating public/private rsa key pair.
Enter file in which to save the key (/home/ankus/.ssh/id_rsa): 엔터키 입력
Created directory '/home/ankus/.ssh'.
Enter passphrase (empty for no passphrase): 엔터키 입력
Enter same passphrase again: 엔터키 입력
Your identification has been saved in /home/ankus/.ssh/id_rsa.
                                                                    ...
[info] ssh info for this host is set..
ankus@test-data's password: 서버 패스워드 입력
Enter file in which to save the key (/home/ankus/.ssh/id_rsa): 엔터키 입력
Enter passphrase (empty for no passphrase): 엔터키 입력
Enter same passphrase again: 엔터키 입력
Genetating public/private rsa key pair.
                                                                    ...
[info] ssh info between hosts is shared..
```

  - 2_java_install.sh 파일 실행 – **ankus 계정**
  - 3_hadoop_install.sh 파일 실행 – **ankus 계정**

4. 설치 및 실행

4.4 ankus 엔진 모듈 설치 및 Cubrid 설치(1/7)

- **본 설치는 ankus 엔진 모듈 설치 및 운영을 위한 것으로 다음과 같은 설치가 진행됨**

  - ankus 엔진 모듈 운영을 위한 **Cubrid DB설치** 및 초기 DB 입력

  - ankus 엔진 모듈 웹 서비스를 위한 톰캣 설치 및 설정 변경

- **실행 요건**

  - ankus 엔진 모듈은 Hadoop의 마스터 노드가 설치된 서버에 설치가 되어야 함,

    따라서 본 설치는 반드시 Hadoop의 마스터 노드에서 실행해야 함

  - 가능한 Hadoop 설치 및 운영 계정으로 진행하는 것을 권장함

  - 쉘 스크립트 파일 실행을 위해 모든 .sh파일에 실행 권한을 부여해야 함

    (예) chmod +x *.sh)

4.4.1 Cubrid 설치

- **ankus 엔진 모듈은 Hadoop과의 연계 및 Hadoop 모니터링 등을 위하여 Cubrid을 사용하기 때문에 ankus 엔진 모듈 운영을 위해서는 반드시 Cubrid을 설치 해야 함**

  - ankus 엔진 모듈 운영을 위한 DB 연결은 Cubrid 9.3.6 버전에서만 테스트 되었으므로, 본 설치에서는 Cubrid 9.3.x 버전을 설치 함

- **실행 : Cubrid DB를 위한 sql 파일 수정 – ankus_cu.sql**

  - 본 단계에서 설치 하는 Cubrid을 ankus 엔진 모듈을 위해서만 구동 시에는, 별도의파일수정을하지않음
  - 다만, 설치 Cubrid을 ankus 웹 모듈과 함께 사용할 시에는 (즉, ankus 웹모듈을ankus 엔진 모듈이 설치 되는 Hadoop 마스터 노드에 설치할 때에는) **22페이지** 실행 1단계를 참고하여 본 파일을수정해야함

4.4.1 Cubrid 설치

- 01_cubrid_installdb.sh 파일 실행

```
[ankus@test-master 02. ankus-engine-install]#./01_cubrid_installdb.sh
...
Do you agree to the above license terms? (yes or no) : yes
Do you want to install this software(CUBRID) to the default(/home/ankus/CUBRID) directory? (yes or no) [Default: yes] : yes
Install CUBRID to '/home/ankus/CUBRID' ...
In case a different version of the CUBRID product is being used in other machines, please note that the CUBRID 9.0 servers are only compatible with the CUBRID 9.3.6 clients and vice versa.
Do you want to continue? (yes or no) [Default: yes] : yes
Copying old .cubrid.sh to .cubrid.sh.bak...
...
```

4.4.1 Cubrid 설치

- Cubrid 설치 종료 후, 정상적으로 구동 되었는지 확인

```
[ankus@test-master 02. ankus-engine-install]$ ps -ef | grep cub_
ankus 15200 1 0 18:57 ? 00:00:00 cub_master
ankus 15205 1 0 18:57 pts/17 00:00:00 cub_broker
ankus 15210 1 0 18:57 pts/17 00:00:00 query_editor_cub_cas_1
ankus 15211 1 0 18:57 pts/17 00:00:00 query_editor_cub_cas_2
ankus 15212 1 0 18:57 pts/17 00:00:00 query_editor_cub_cas_3
ankus 15213 1 0 18:57 pts/17 00:00:00 query_editor_cub_cas_4
ankus 15214 1 0 18:57 pts/17 00:00:00 query_editor_cub_cas_5
ankus 15217 1 0 18:57 pts/17 00:00:00 cub_broker
ankus 15222 1 0 18:57 pts/17 00:00:00 broker1_cub_cas_1
ankus 15223 1 0 18:57 pts/17 00:00:00 broker1_cub_cas_2
ankus 15224 1 0 18:57 pts/17 00:00:00 broker1_cub_cas_3
ankus 15225 1 0 18:57 pts/17 00:00:00 broker1_cub_cas_4
ankus 15226 1 0 18:57 pts/17 00:00:00 broker1_cub_cas_5
ankus 15229 1 0 18:57 ? 00:00:00 cub_auto start
ankus 15232 1 0 18:57 ? 00:00:00 cub_js start
...
```

4.4.2 ankus 엔진 모듈 설치

- **ankus 엔진 모듈은 톰캣 기반의 웹 서비스를 하는 모듈로써, 톰캣을 설치하고 ankus 엔진 모듈을 웹 서비스 디렉토리에 포팅하는 방법으로 설치 함**

  - 단, ankus 엔진 모듈은 톰캣 7 버전대에서만 개발/운영이 테스트되어, 본 설치에서는 톰캣 7버전을 이용함

  - ankus 엔진 모듈은 네트워크의 28080포트를 이용하여 서비스되도록 자동 설정 됨

- **1단계) 엔진 모듈 운영을 위한 설정 파일 내용 변경 – ankus-site-engine.xml**

  - 파일의 104번째 라인을 본인의 정보에 맞는 시스템 사용자(Hadoop 운영 계정) 홈 디렉토리 경로로 수정

  - 8페이지 실행 1단계 Hadoop 운영 사용자 계정명 참고 (vi 편집기 명령어로 파일 수정)

```
<property>
   <name>file.system.user.home</name>
   <value>/home/ankus/</value>
   <description>user's home directory</description>
   <expose>false</expose>
</property>
```

4.4.2 ankus 엔진 모듈 설치

- 파일의 111번째 라인을 본인의 정보에 맞는 Hadoop 운영 계정 명으로 수정

  * 8페이지 실행 1단계 Hadoop 운영 사용자 계정명 참고 (vi 편집기 명령어로 파일 수정)

```
<property>
   <name>hadoop.user.name</name>
   <value>ankus</value>
   <description>System user account for Hadoop Security</description>
   <expose>false</expose>
</property>
```

- 파일의 118번째 라인을 본인의 정보에 맞는 java 설치 디렉토리로 수정

  * 본 가이드를 통해 java를 설치 한 경우는 사용자 계정 디렉토리 내에java가설치되므로,

  'echo $JAVA_HOME' 으로 사용자 경로 확인 (vi 편집기 명령어로 파일 수정)

```
<property>
   <name>java.home</name>
   <value>/home/ankus/jdk1.8.0_131</value>
   <description>JDK path</description>
   <expose>false</expose>
</property>
```

4.4.2 ankus 엔진 모듈 설치

- 파일의 125번째 라인을 본인의 정보에 맞는 Hadoop 설치 디렉토리로 수정

  * 본 가이드를 통해 hadoop을 설치한 경우 사용자 계정, 'echo 디렉토리내에Hadoop이설치되므로 $HADOOP_HOME'으로 사용자 경로 확인 (vi 편집기명령어로파일수정)

```
<property>
  <name>hadoop.home</name>
  <value>/home/ankus/hadoop</value>
  <description>Hadoop home directory.</description>
  <expose>false</expose>
</property>
```

- **2단계) ankus 엔진 모듈 설치 쉘 스크립트 실행 및 톰캣 서버 구동**

- ./02_ankus_engine.sh 파일 실행

```
[ankus@test-master 02. ankus-engine-install]$ ./02_ankus-engine.sh
--2017-07-03 10:37:38-- tar -xzf tomcat-7.0.55-ankus.tar.gz
...
mv tomcat-7.0.55-ankus ~/ankus-engine
...
Tomcat started.
```

- 톰캣이 정상 구동되지 않는 경우에는 로그를 통해 문제점 파악

```
[ankus@test-master ~]$ tail -f ankus-engine/logs/catalina.out
2017-07-06 04:03:00,001 INFO [org.ankus.engine.monitoring.job.MonitoringQuartzJob:43] Free Heap=168908360, Total Heap=386400256, Max Heap=892862464, Processors=4
```

4. 설치 및 실행

4.5 ankus 웹 모듈 설치(1/4)

- **본 설치는 ankus 웹 모듈 설치 및 운영을 위한 것으로 다음과 같은 설치가 진행됨**

  - ankus 웹 모듈 운영을 위한 Cubrid DB설치 및 초기 DB 입력

  - ankus 웹 모듈 웹 서비스를 위한 톰캣 설치 및 설정 변경

- **실행 요건**

  - ankus 웹 모듈은 ankus 엔진 모듈과 서로 다른 별도의 서버에서구동이되어도상관없음

  - 쉘 스크립트 파일 실행을 위해 모든 .sh파일에 실행 권한을 부여해야 함

  (예) chmod +x *.sh)

4.5.1 Cubrid 설치

- **ankus 웹 모듈은 사용자 계정 및 기타 관리 정보 저장을 위하여 Cubrid을 사용하기 때문에 ankus 웹 모듈 운영을 위해서는 반드시 Cubrid을 설치 해야 함**
  - ankus 웹 모듈 운영을 위한 DB 연결은 Cubrid 에서만 테스트 되었으므로, 본 설치에서는 Cubrid를 설치 함
- **1단계) Cubrid DB를 위한 sql 파일 수정 – ankus_cu.sql**

- 해당라인의관리자계정정보수정(vi 편집기 명령어로 수정 후, cat 명령어로수정내용확인)

```
[CR_NON_EXPIRED], [AUTHORITY])
VALUES('admin', 'admin', 'ankus', 'ankus@onycom.com', 1, 1, 1, 1, 'ROLE_ADMIN');
```

관리자

- **2단계) Cubird 설치**

- 16페이지 실행 4단계를 참고하여 Cubrid 설치

4.5.2 ankus 웹 모듈 설치

- **ankus 웹 모듈은 톰캣 기반의 웹 서비스를 하는 모듈로써, 톰캣을 설치하고 ankus 웹 모듈을 웹 서비스 디렉토리에 포팅하는 방법으로 설치 함**

  - 단, ankus 웹 모듈은 톰캣 7 버전대에서만 개발/운영이 테스트되어, 본 설치에서는 톰캣 7버전을 이용함

  - ankus 웹 모듈은 네트워크의 8080포트를 이용하여 서비스되도록 자동 설정 됨

- **1단계) 웹 모듈 운영을 위한 설정 파일 내용 변경 – ankus-site-web.xml**

  - 파일 다운로드 최대 크기(29번째 라인), 파일 업로드 최대 크기(36번째 라인)를 필요에 따라 수정

```
<property>
   <name>fs.browser.download.max.file.size</name>
   <value>100000000</value>
   <description>Max size of download files in File System Browser</description>
   <expose>true</expose>
</property>
<property>
   <name>fs.browser.upload.max.file.size</name>
   <value>100000000</value>
   <description>Max size of upload files in File System Browser</description>
   <expose>true</expose>
</property>
```

4.5.2 ankus 웹 모듈 설치

- **2단계) ankus 웹 모듈 설치 쉘 스크립트 실행 및 톰캣 서버 구동**

  - ./02_ankus-web.sh 파일 실행

```
[ankus@test-master 03. ankus-web-install]$ ./02_ankus-web.sh
--2017-07-03 10:56:26-- tar -xzf tomcat-7.0.55-ankus.tar.gz
...
mv tomcat-7.0.55-ankus ~/ankus-web
...
Tomcat started.
```

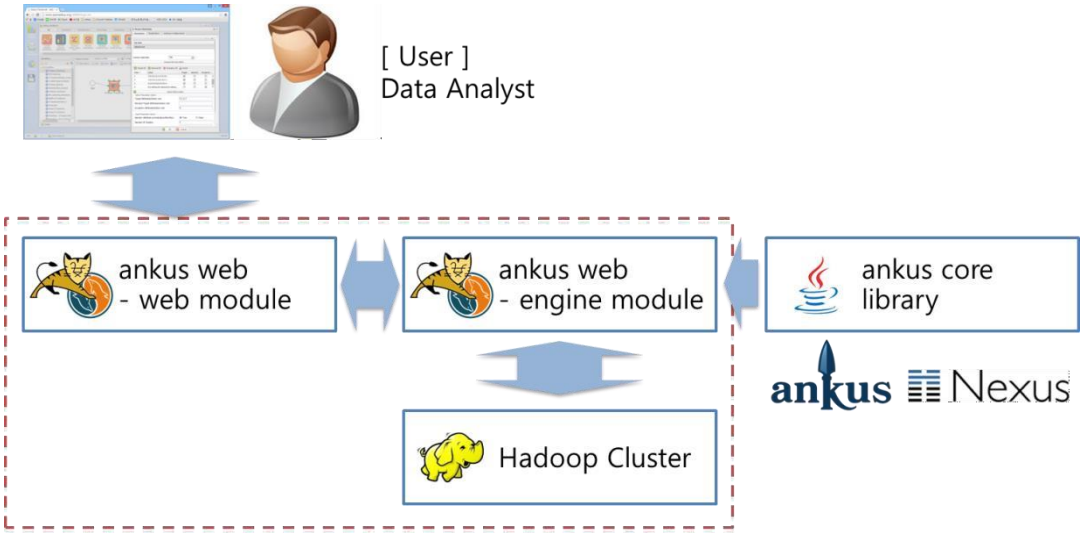
- **ankus 설치 및 정상 구동 확인**

  - 인터넷을 통해 ankus 웹 모듈이 설치된 서버의 8080포트로 접속하여 다음과 같은 로그인 페이지가 뜨면 정상 설치 완료

4. 설치 및 실행

4.6 설치 완료

- **ankus 웹-엔진 연동, 엔진-Hadoop 연동 및 정상 설치 여부 확인**
- **ankus의 웹 모듈과 엔진 모듈 및 Hadoop은 다음과 같은 구조로 연동되어 사용할 수 있는 구조**



- **ankus Analyzer G에 알고리즘은 포함되어 있지 않음. 해당 코어를 다운로드 받아 사용자 가이드를 참조하여 알고리즘 추가를 해야 알고리즘을 사용 가능**
- **다운로드 URL**

  - ankus 알고리즘 : https://github.com/onycom-ankus/ankus_library/raw/master/ankus-core2-onycom-alpha1-1.1.0.jar
  - ankus 차트도구 : https://sourceforge.net/projects/ankus/files/ankus-visualization/1.0.1/ankus-visualization-1.0.1.jar

5. 기능 소개

세부 목차

1. 데이터 분석 절차
2. 관리자도구
3. 워크플로우
4. 스케줄러
5. 진행과정관리
6. HDFS 브라우저
7. 모니터링
8. 차트도구
9. Hive Query 도구

5. 기능소개

5.1 데이터 분석 절차

- **ankus Analyzer G 를 이용해 분석을 수행하는 순서**

01
ankus Analyzer G 접속

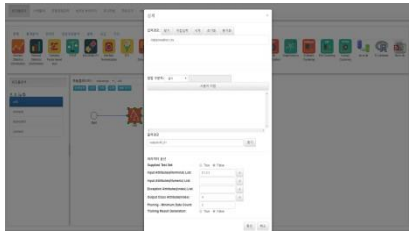
사용자 ID/PW를 이용해 ankusAnalyzerG시스템에접속

02
데이터 업로드

HDFS 브라우저를 이용해 데이터를 Hadoop HDFS에 업로드

03
워크플로우 생성

ankus Analyzer G 에서 Drag & Drop 방식으로 분석 알고리즘을 선택하여 워크플로우를 설계함



04
파라미터 설정

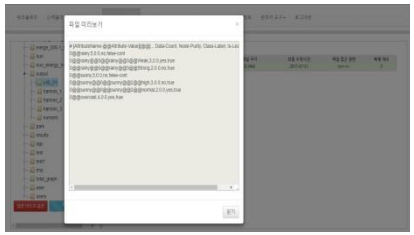
데이터 선택, 저장 경로, 변수 선택 등 알고리즘 별로 분석에 필요한 파라미터를 설정

05
분석 저장 및 실행

작성된 워크플로우를 저장 후 분석을 수행

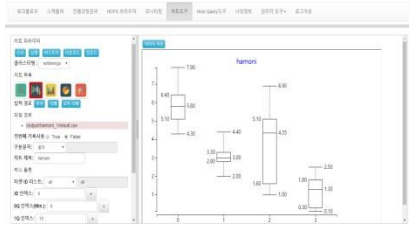
06
진행과정관리에서 작업 확인

진행과정관리를 통해서 Hadoop에서의 작업 진행상황 확인



07
분석 결과 확인

HDFS 브라우저에서 분석 저장 경로로 들어가 분석 결과 다운로드



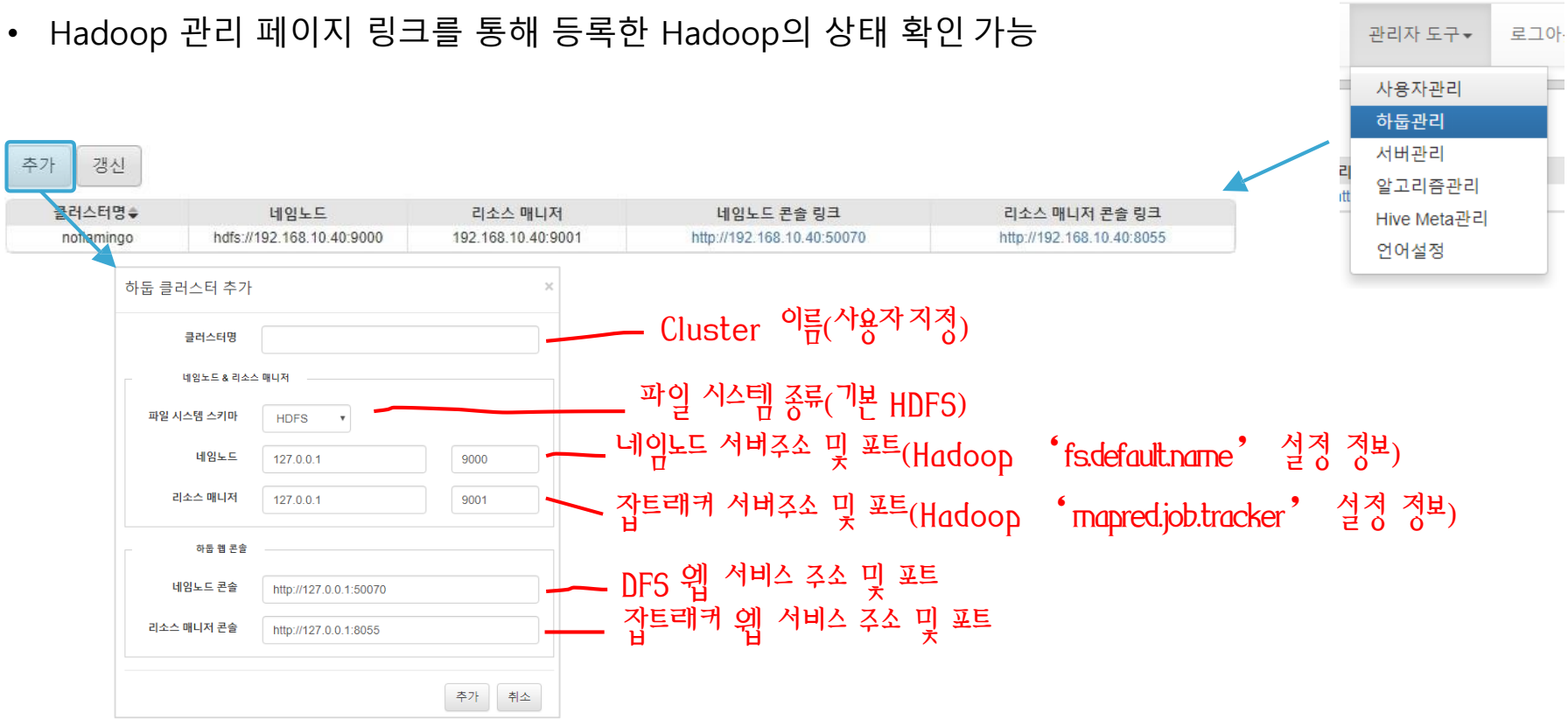
08
D3.js를 이용한 시각화

차트도구에서 분석 결과에 따라 원하는차트를선택하여 D3.js 기반의 시각화 수행

5. 기능 소개

5.2 관리자 도구(1/3)

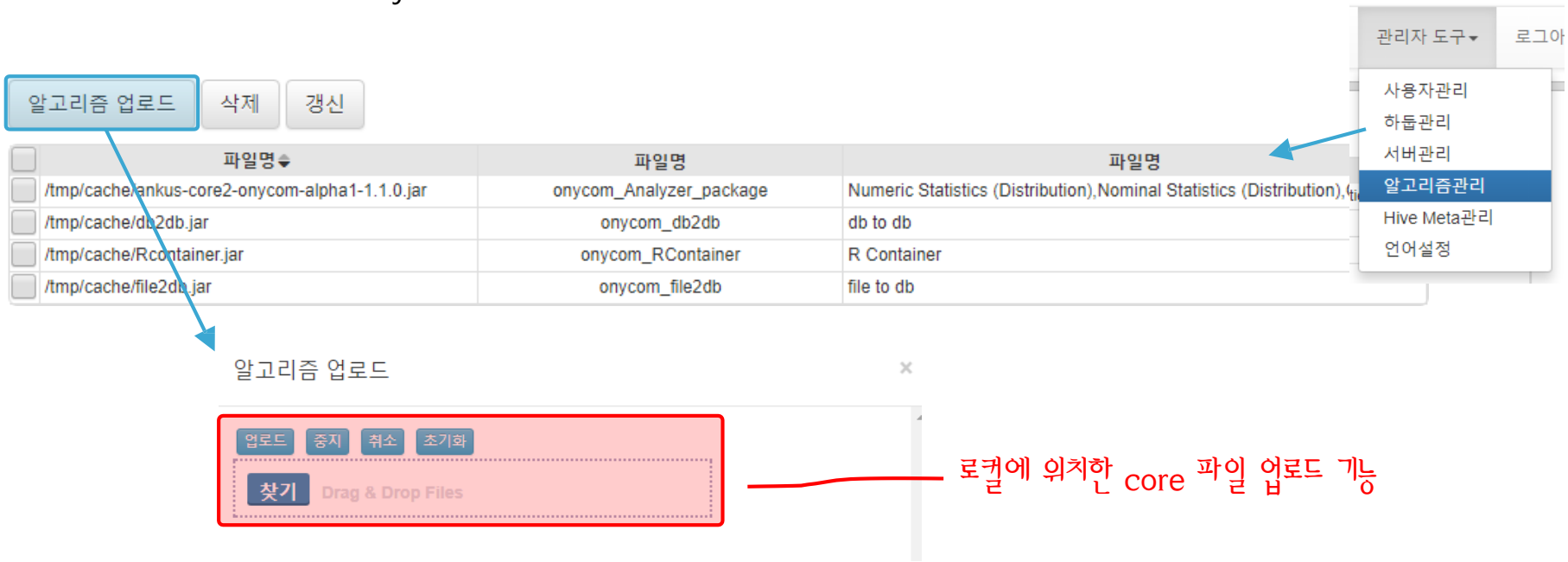
- Hadoop 관리 기능 제공
- ankus Analyzer G 와 연동할 Hadoop 클러스터 정보를 입력하여 다수의 클러스터 관리 가능
- 등록을 완료하면 Hadoop 관리 페이지 링크와 함께 목록 표시 확인
- Hadoop 관리 페이지 링크를 통해 등록한 Hadoop의 상태 확인 가능



5. 기능 소개

5.2 관리자 도구(2/3)

- 알고리즘 관리 기능 제공
- ankus Analyzer G Core를 업로드 하여 관리 함
- Ankus Analyzer G는 알고리즘은 포함되어 있지 않음. Core 파일을 다운로드 받아 업로드 해야함
  - ankus 알고리즘 : https://github.com/onycom-ankus/ankus_library/raw/master/ankus-core2-onycom- alpha1-1.1.0.jar
  - ankus 차트도구 : https://sourceforge.net/projects/ankus/files/ankus-visualization/1.0.1/ankus - visualization-1.0.1.jar



5. 기능 소개

5.2 관리자 도구(3/3)

- Hive Meta 관리 기능 제공
- Ankus Analyzer G 는 사용자가 DataWarehouse(DW)를 쉽게 구축할 수 있도록 Apache Tajo를 연동하여 제공
- Tajo는 대용량 분산 DW 시스템으로서, SQL-on-Hadoop 기능을 제공
- 사용자는 엑셀로 다루기 힘든 큰 데이터를 변환/별도의 Load 없이 바로 SQL로 처리 가능

5. 기능 소개

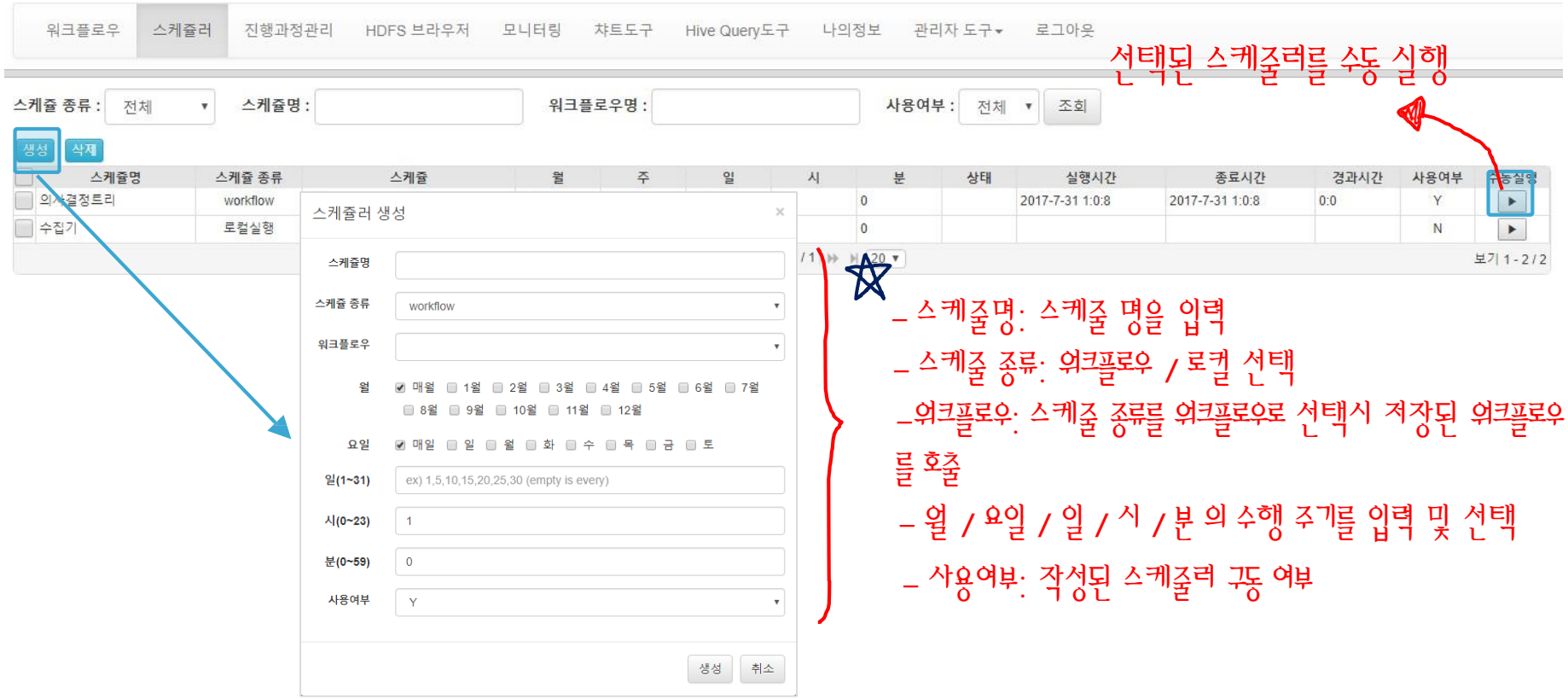
5.3 워크플로우

- 워크플로우 설계 기능을 제공
- ankus Analyzer G 분석 라이브러리에서 원하는 알고리즘을 선택해 분석 워크플로우를 설계
- 설계한 워크플로우를 저장 가능하며, 왼쪽 워크 플로우 목록을 통해 다시 불러오기 가능
- 워크플로우의 세부 설정 기능을 제공하여, 디테일한 분석 설정 가능

5. 기능 소개

5.4 스케줄러

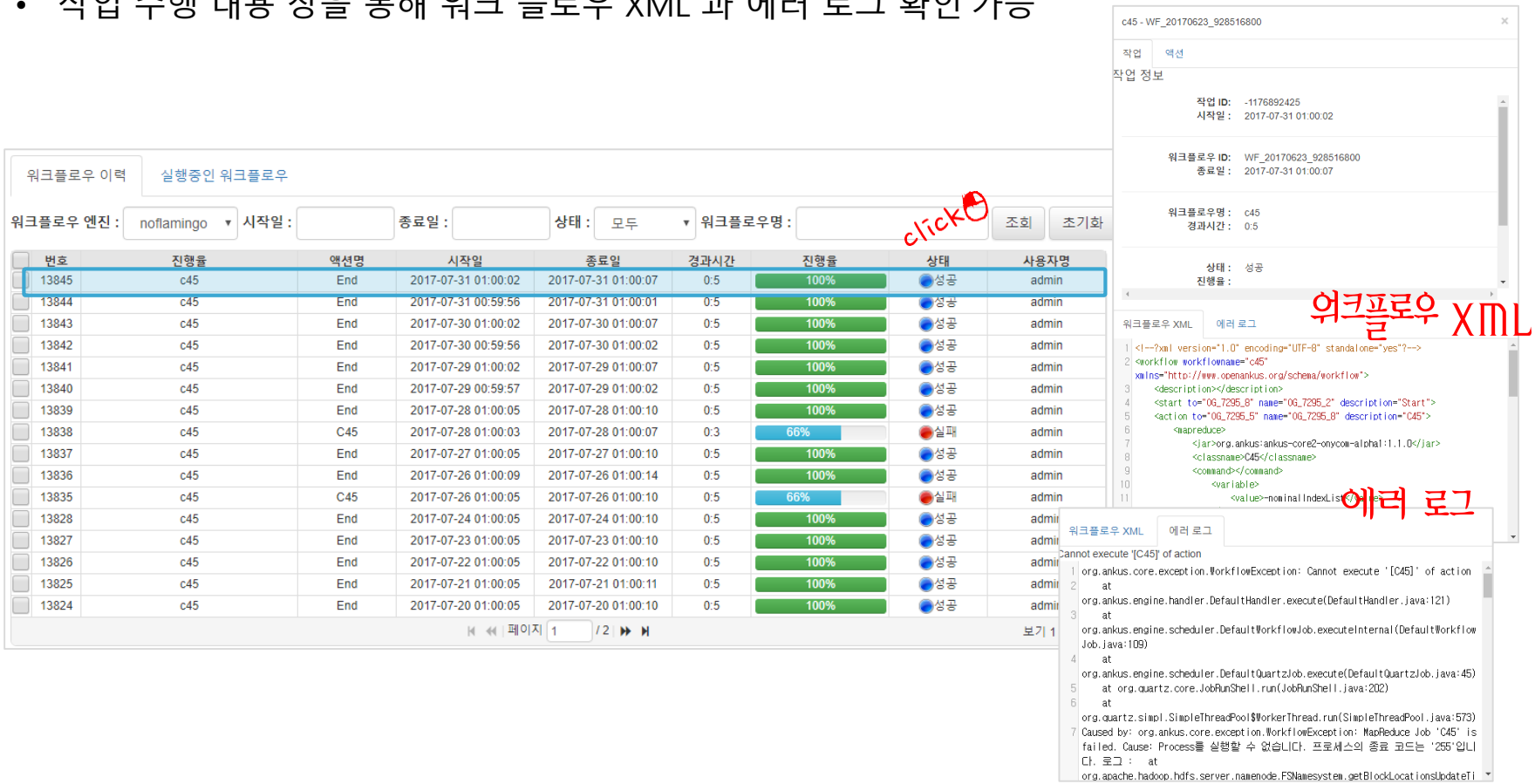
- 워크플로우 및 로컬 프로그램을 스케줄러로 등록하고 주기적으로 실행 가능
- 스케줄 종류(workflow / 로컬)로 구분하며 사용여부로 가동 여부를 결정
- 월간/ 주간/ 일간/ 시간/ 분간 으로 구동 시간을 설정 가능



5. 기능 소개

5.5 진행과정관리

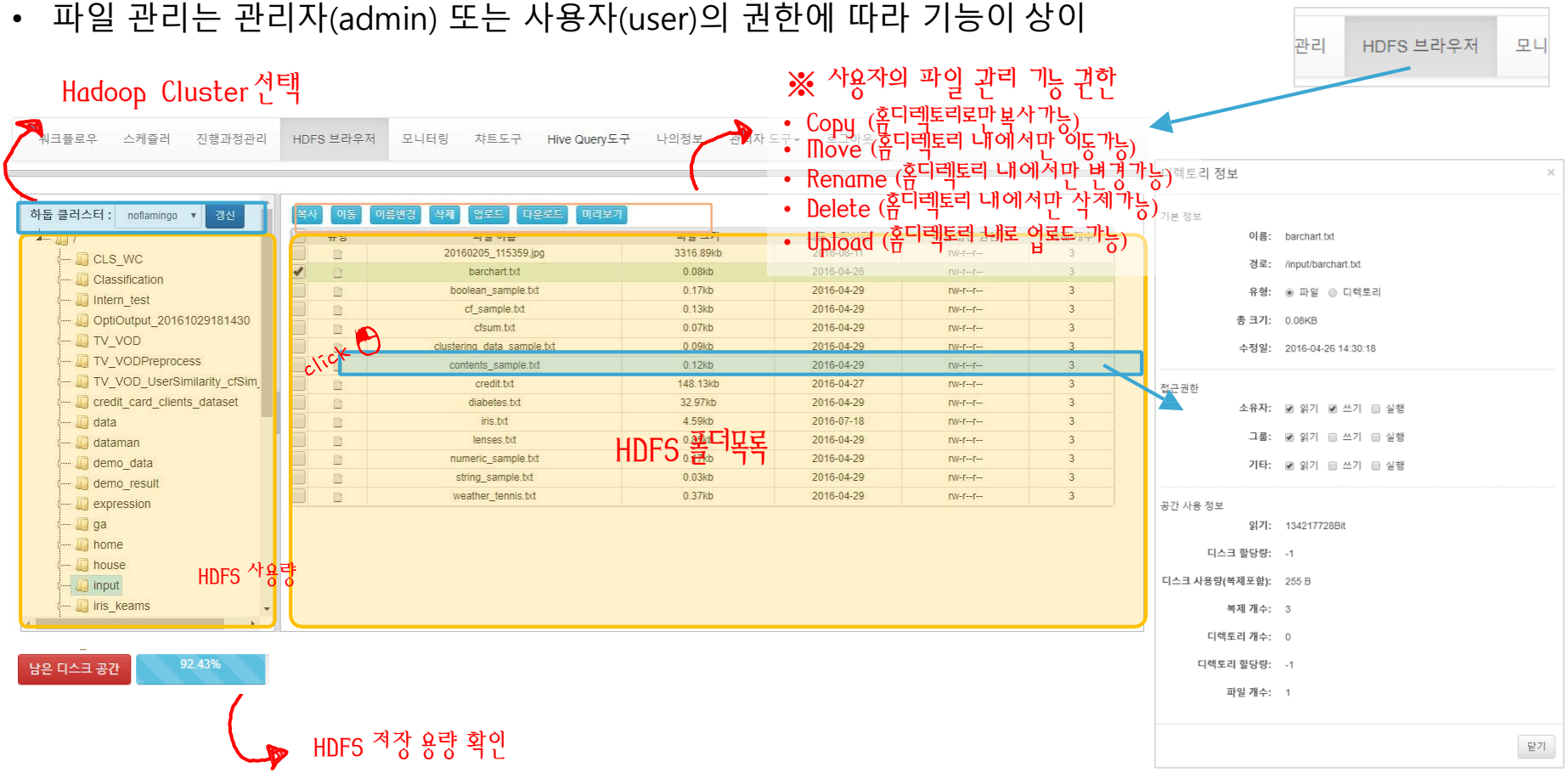
- 분석 작업 상태 모니터링 기능 제공
- 워크플로우 엔진을 선택하고 해당 엔진에서 수행한 분석 작업 목록을 확인 가능
- 상태를 통해 작업 성공/실패 여부를 확인
- 작업 수행 내용 창을 통해 워크 플로우 XML 과 에러 로그 확인 가능



5. 기능 소개

5.6 HDFS 브라우저

- HDFS 브라우저를 제공하여 Hadoop 파일 시스템(HDFS) 관리 가능
- Hadoop 클러스트에서 등록한 Engine을 선택하여 HDFS 에 접근 가능
- 폴더, 파일 업로드 및 다운로드가 가능하며, 파일 미리보기 기능 제공
- 파일 관리는 관리자(admin) 또는 사용자(user)의 권한에 따라 기능이 상이



5. 기능 소개

5.7 모니터링

- 웹 서버 및 Hadoop System 의 상태 정보를 모니터링하는 기능 제공
- Hadoop 클러스터 별로 모니터링 가능
- Hadoop 노드와 Web Server의 CPU 로드율, memory 사용량, disk 사용량 정보 확인 가능

1. Hadoop 클러스터 선택 및 갱신 시간 조정

2. Hadoop 정보의 전체 상태 정보 제공

3. Web server의 cpu, memory, disk 정보 제공

4. Master node의 cpu, memory, disk 정보 제공

5. data node List

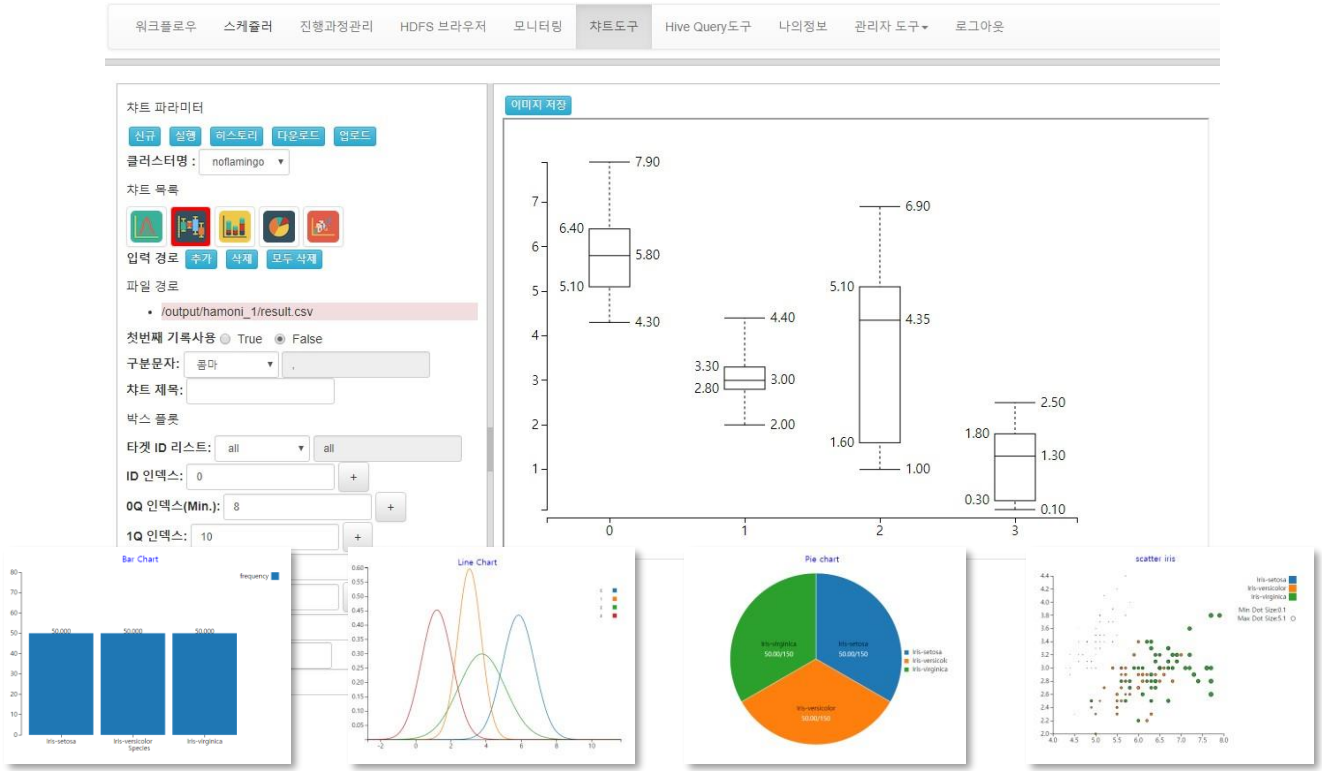
6. 선택된 data node 의 cpu, memory, disk등의 정보 제공

5. 기능 소개

5.8 시각화 - 차트도구

- 데이터 분석을 통한 인사이트 및 직관성을 확보하기 위하여 분석 결과를 시각화 하는 기능 제공
- 분석 결과에 따라 원하는 차트를 그릴 수 있음
- ankus Analyzer G 는 현재 버전에서 총 5종의 차트를 제공
  (PDF Graph, Box Plot, Bar Chart, Pie Chart, Scatter Plot)

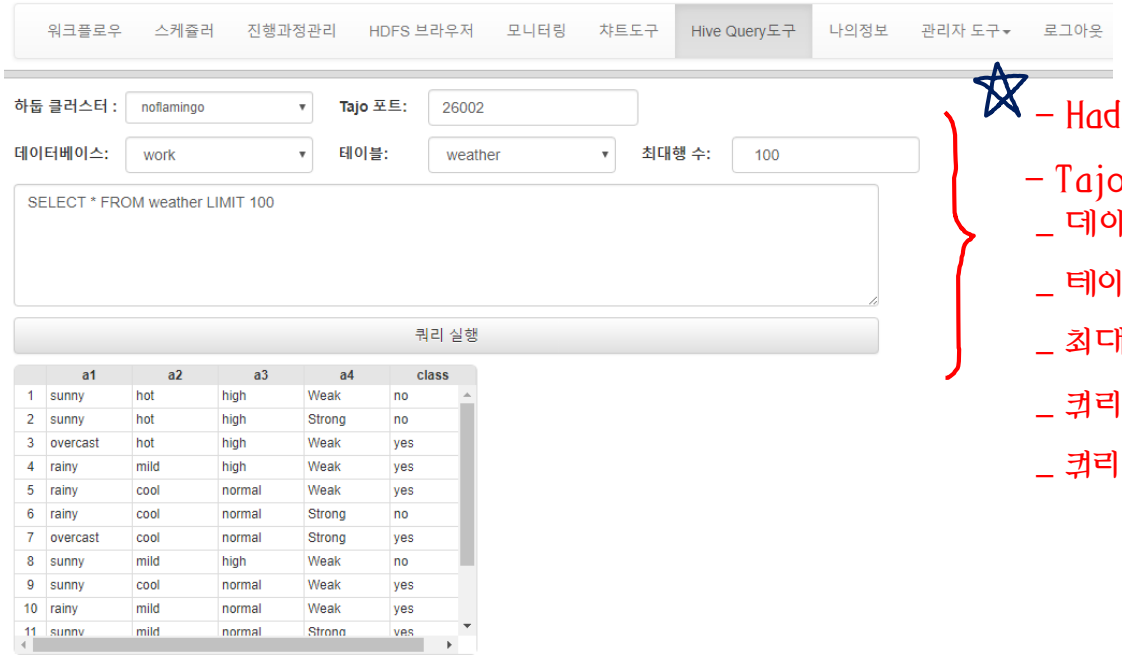
차트도구 화면



5. 기능 소개

5.9 Hive Query 도구

- Hive Query 도구를 통하여 Hadoop 클러스터에 데이터를 조회 가능
- Hive Query 도구에서는 Tajo SQL을 작성하여 적용 가능
- Query 의 결과는 하단 테이블 뷰를 통하여 바로 보기 가능



- Hadoop 클러스터 : Engine 선택
- Tajo 포트 : Tajo Port 지정
- 데이터베이스 : 생성된 DB 선택
- 테이블 : 생성된 Table 선택
- 최대행 수: query 결과 출력 라인 수
- 쿼리: list를 보여주는 query command
- 쿼리 실행: query 실행

6. 활용 예제

세부 목차

1. 데이터 전처리(ETL)
2. 통계분석 - Numeric Statistics
3. 시각화 - 차트도구

6. 활용 예제

6.1 데이터 전처리 (ETL)(1/9)

- **데이터 전처리 (ETL; Extraction, Transformation, Load)**

  - 사용 목적 : 원하는 속성 변수만 추출하거나 데이터 형태의 변형을 주고 싶은 경우
    군집, 분류 등 다른 알고리즘을 구동하기 위해 데이터 전처리가 필요한 경우
  - 기본 개념 : 입력데이터의 추출, 가공 등의 작업을 통한 전처리 수행
  - 입력 형식 : 수치 데이터 , 기호 데이터
  - 출력 형식 : 지정한 방법에 의해 전처리 된 데이터

6. 활용 예제

6.1 데이터 전처리 (ETL)(2/9)

- **샘플 데이터 다운로드**
  - iris.csv : 꽃잎의 모양 데이터가 담긴 파일
  - 데이터 형식 0: 꽃받침 길이, 1: 꽃받침 너비, 2: 꽃잎 길이, 3: 꽃잎 너비, 4: 꽃 이름
  - 다운로드 위치 : https://github.com/onycom-ankus/ankus_library/blob/master/test_data/iris.csv

onycom-ankus / ankus_library

Watch 7 | Star 8 | Fork 13

Code | Issues 0 | Pull requests 0 | Projects 0 | Wiki | Insights | Settings

Branch: master | ankus_library / test_data / iris.csv | Find file | Copy path

prismdata 데이터 셋입니다. d56ec31 on 29 Sep 2017

1 contributor

151 lines (150 sloc) | 3.57 KB | Raw | Blame | History

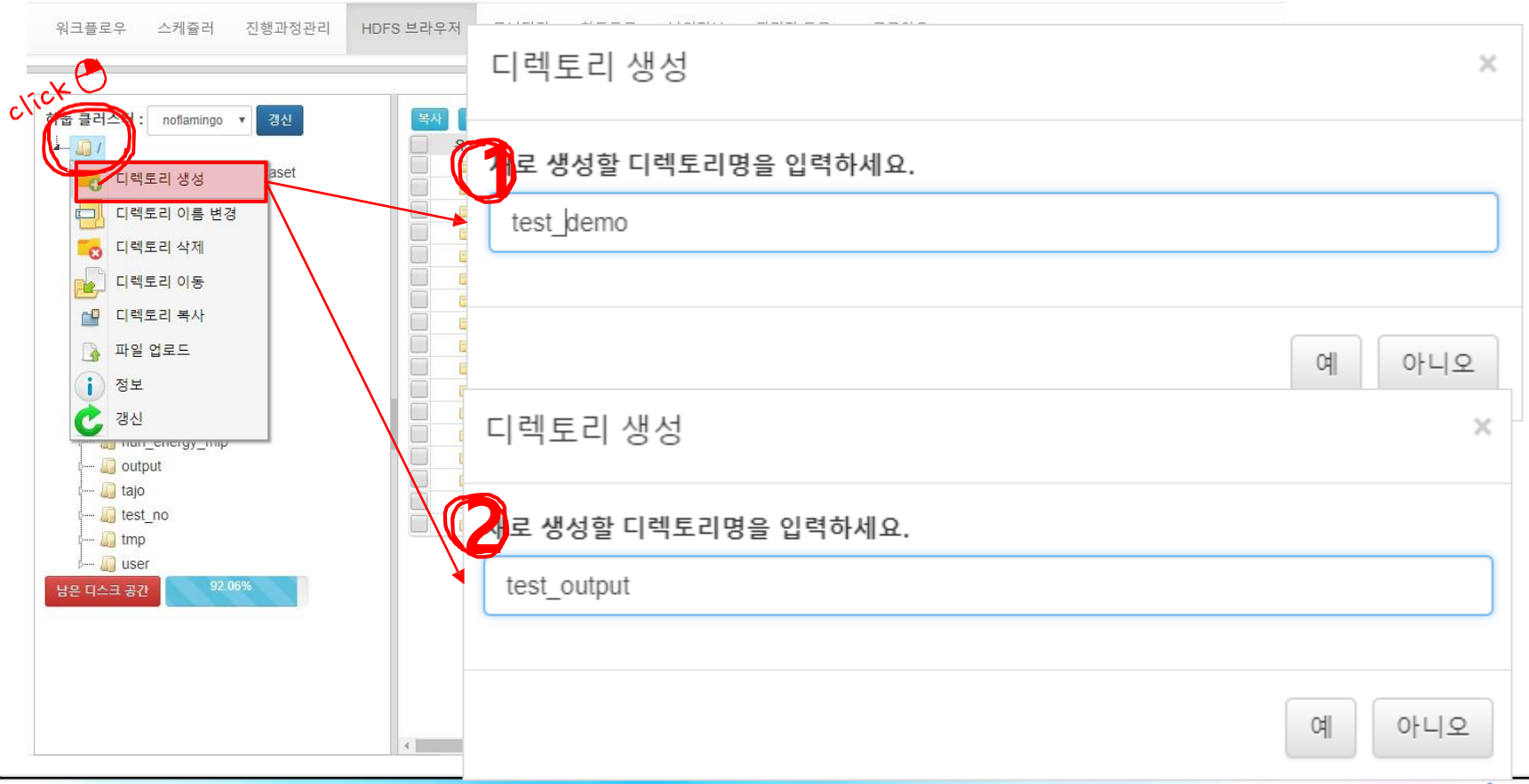
Search this file...

1	5.1	3.5	1.4	0.2	setosa
2	4.9	3	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa

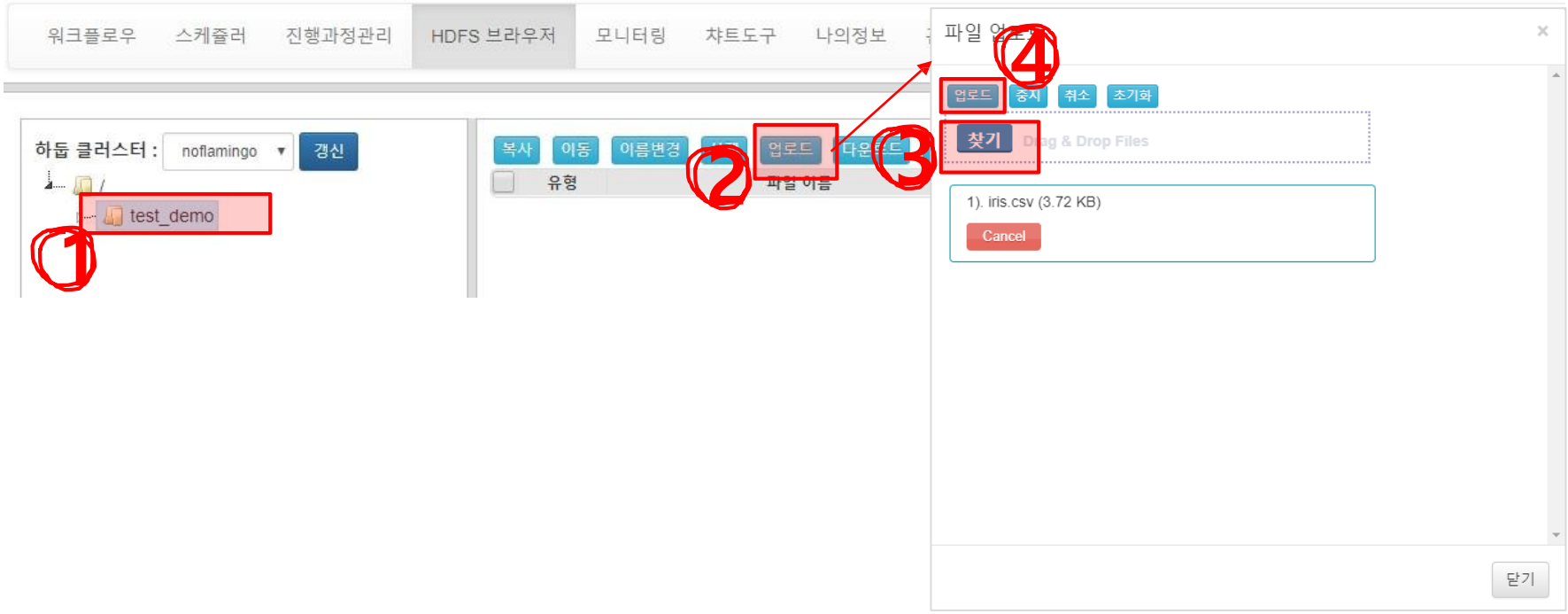
6. 활용 예제

6.1 데이터 전처리 (ETL)(3/9)

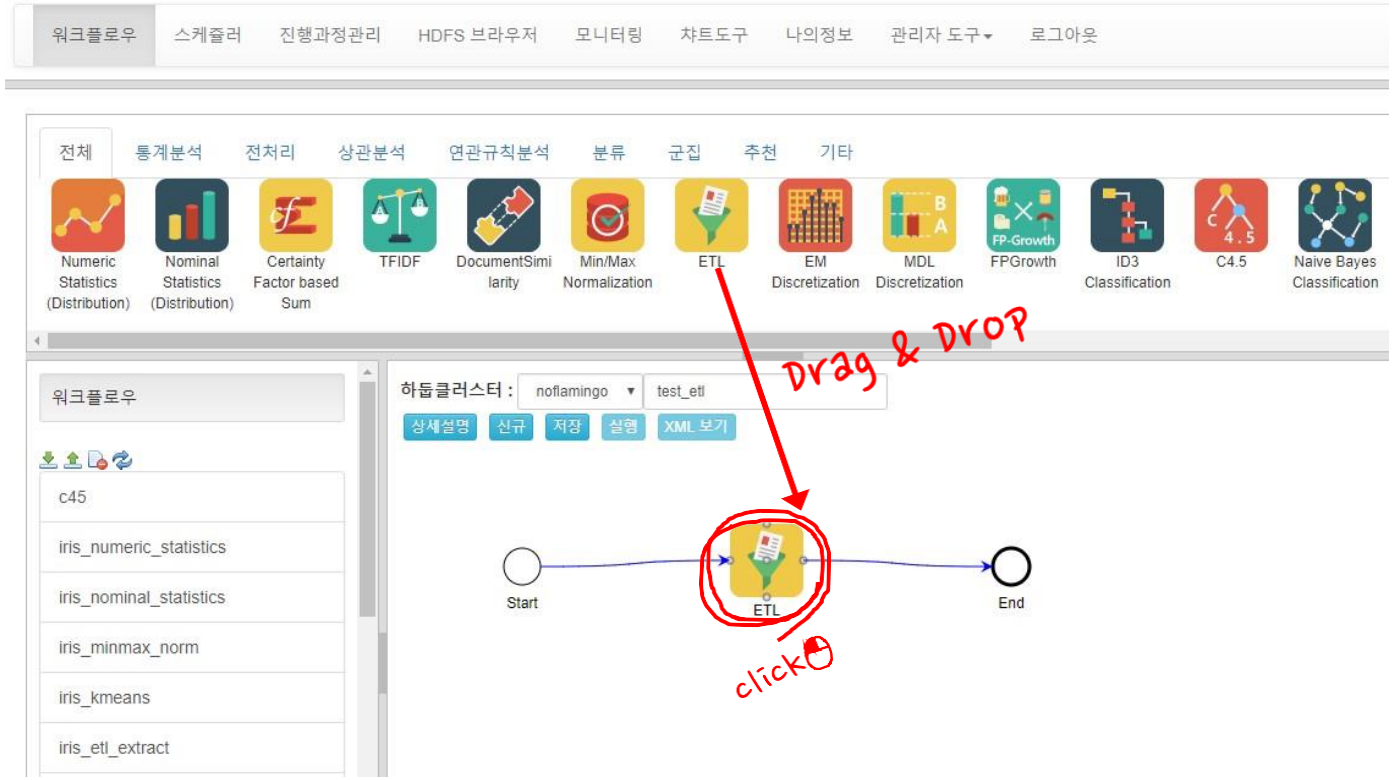
- **샘플 파일 업로드 디렉토리와 결과 출력 디렉토리 생성**
  - HDFS 브라우저 메뉴로 이동
  - 좌측 루트(/) 디렉토리에서 마우스 오른쪽 클릭 후 디렉토리 생성 선택
  - 각각 "test_demo" 와 "test_output" 디렉토리 생성



- **iris.csv 파일 업로드**
  - test_demo 폴더로 이동
  - 업로드 버튼을 눌러 파일 업로드창 띄움
  - 찾기 버튼을 눌러 다운로드한 "iris.csv" 파일을 선택
  - 업로드 버튼을 눌러 파일 업로드 실행



- **ETL 워크플로우 생성**
  - 워크플로우 메뉴로 이동
  - ETL 아이콘을 디자인 공간으로 드래그 하여 이동
  - Start 와 end 에 연결
  - ELT 아이콘을 더블 클릭하여 설정 메뉴 띄움



- **ETL 세부 설정**
  - 입력 경로에서 찾기를 눌러 업로드한 iris.csv 파일 선택
  - 컬럼 구분자를 콤마로 변경
  - 출력 경로를 처음에 생성한 "test_ouput" 디렉토리로 선택
  - 파라미터 옵션

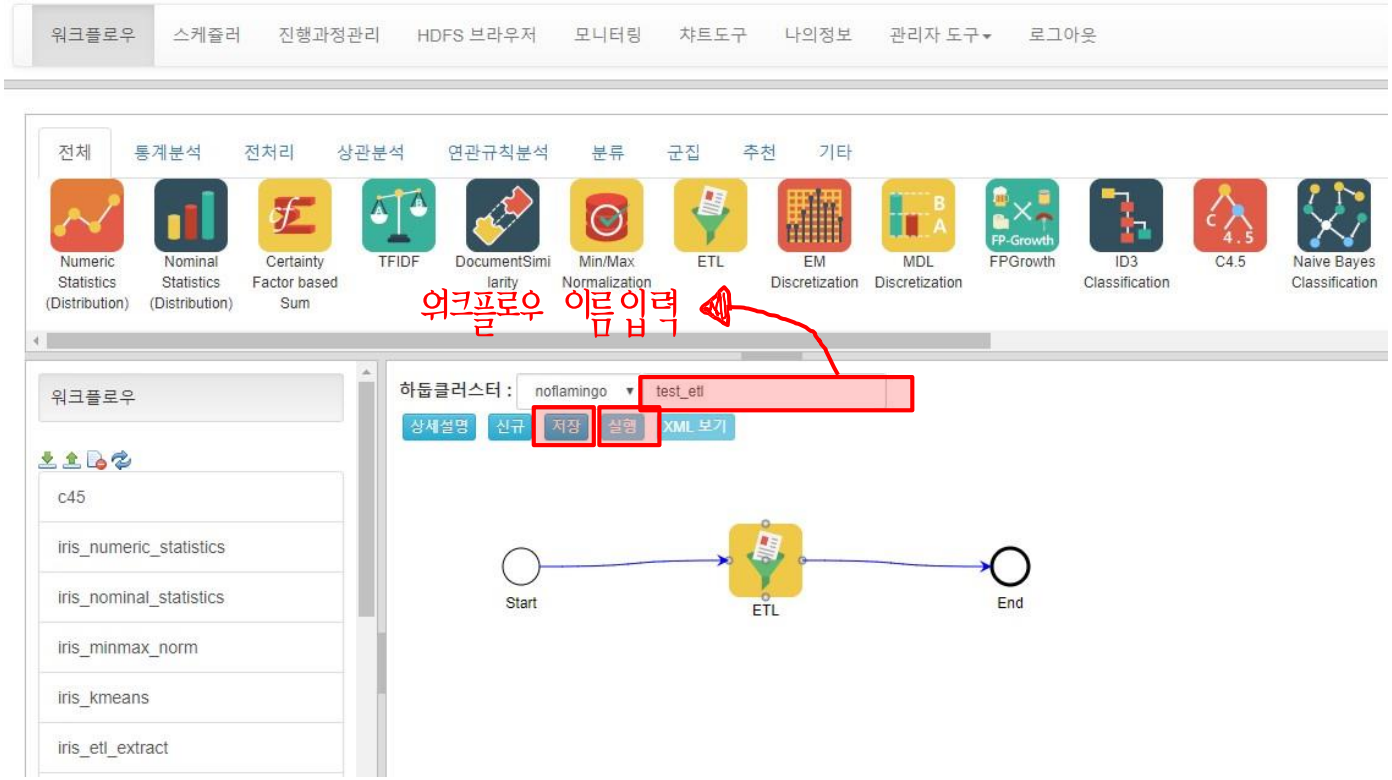
    Method: Filter include (룰에 맞는 값만 추출)

    Rule : '0, 5.1|4, setosa'

    (0 인덱스의 값이 5.1 이거나,

    4 인덱스의 값이 setosa 인 것만 추출)

- **워크플로우 실행**
  - 워크플로우의 이름을 “test_etl” 로 입력
  - 저장 후 실행 버튼 클릭



6. 활용 예제

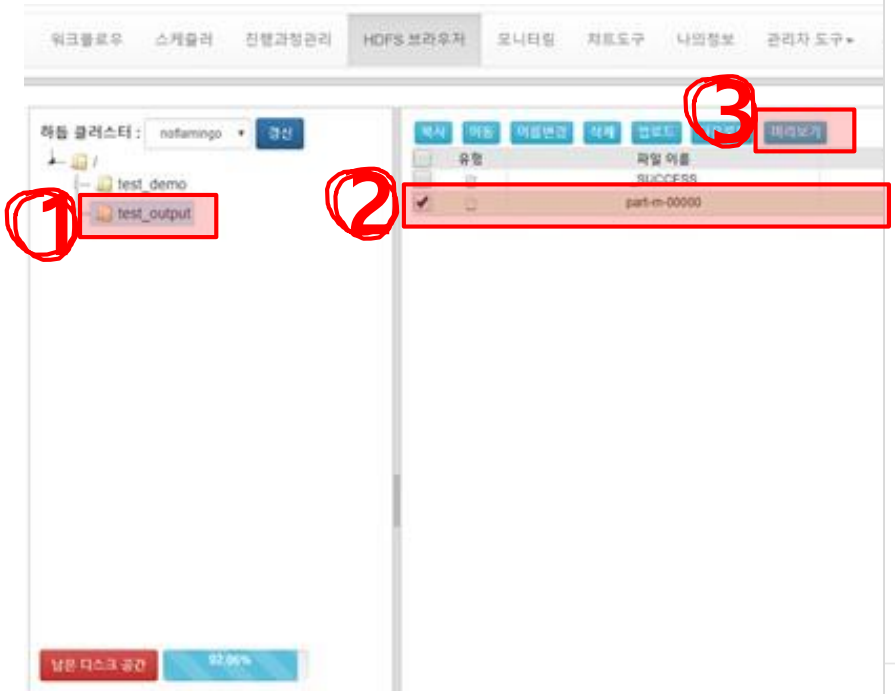
6.1 데이터 전처리 (ETL)(8/9)

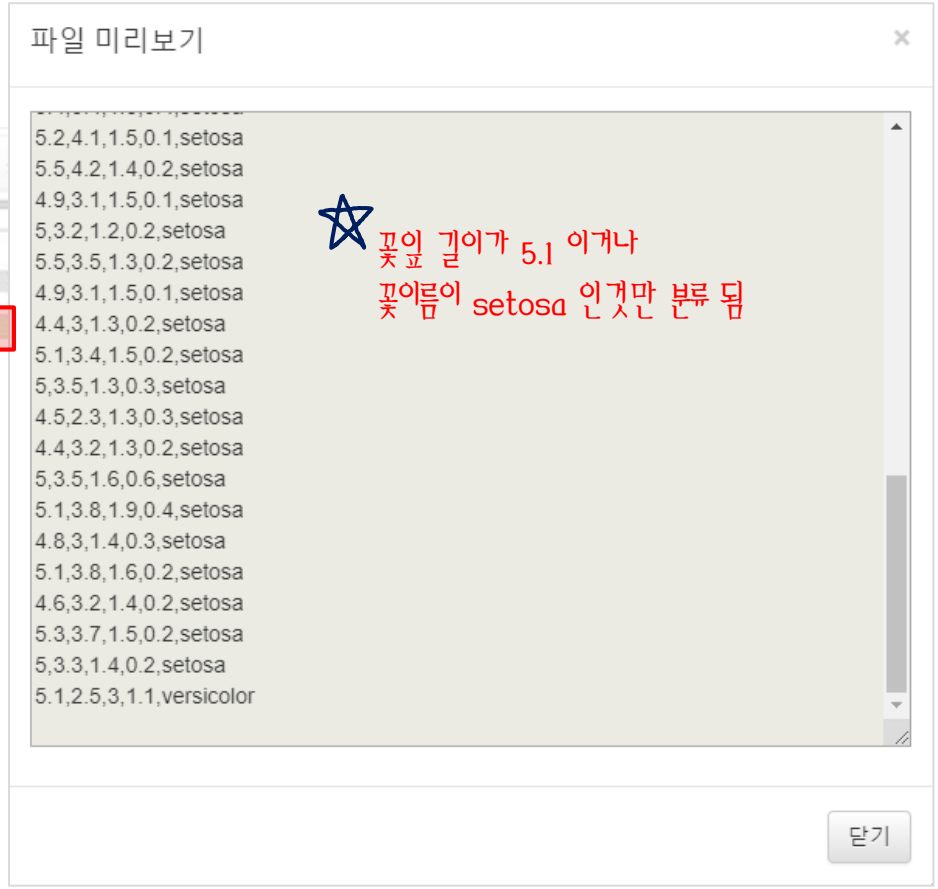
- **워크플로우 처리 확인**
  - 진행과정관리 메뉴로 이동
  - 현재 실행중인 워크플로우 진행 결과 확인

6. 활용 예제

6.1 데이터 전처리 (ETL)(9/9)

- **데이터 전처리 결과 확인**
  - HDFS 브라우저 메뉴로 이동
  - test_output 디렉토리로 이동
  - part-m-00000 파일 선택 후 미리보기 클릭
  - 결과 확인





- **통계분석 – Numeric Statistics**

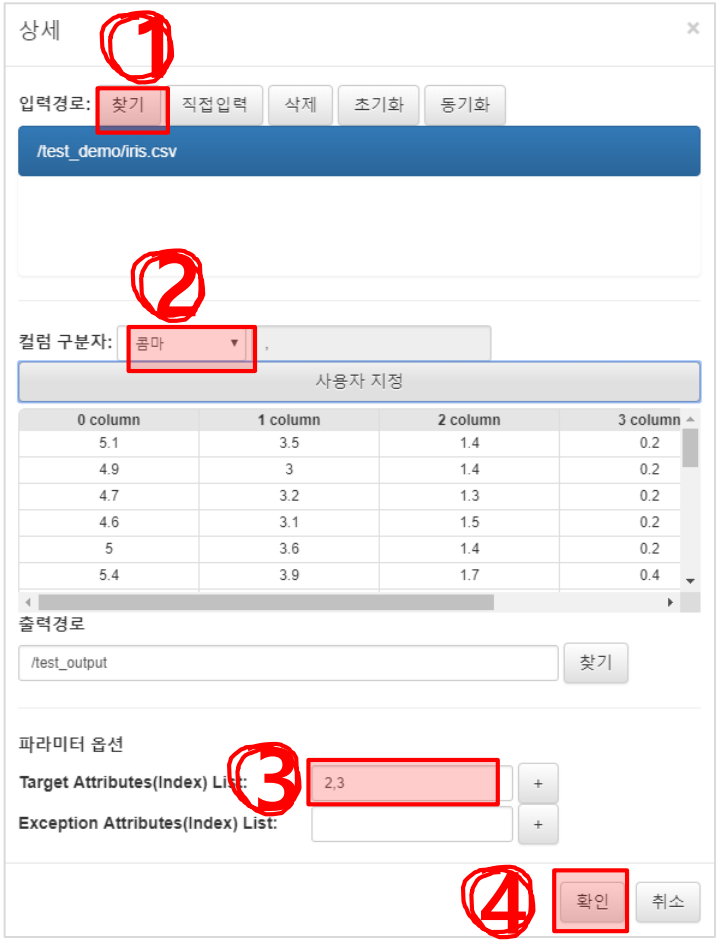
- 사용 목적 : 기초통계 분석
- 기본 개념 : 수치 속성에 대하여 기초 통계를 산출
- 입력 형식 : 수치 데이터
- 출력 형식 : 수치 데이터

(합계(sum) / 평균(average) / 조화평균(harmonic average) /
기하평균(geographic average) / 분산(variance) / 표준편차(standard deviation) /
최대값(max) / 최소값(min) / 데이터 개수(data cnt.)/
1사분위수(1st Quartile) / 2사분위수(2nd Quartile) / 3사분위수(3rd Quartile))

- **Numeric Statistics 워크플로우 생성**
  - 워크플로우 메뉴로 이동
  - Numeric Statistics 아이콘을 디자인 공간으로 드래그 하여 이동
  - Start 와 end 에 연결
  - ELT 아이콘을 더블 클릭하여 설정 메뉴 띄움

6.2 통계분석 – Numeric Statistics(3/6)

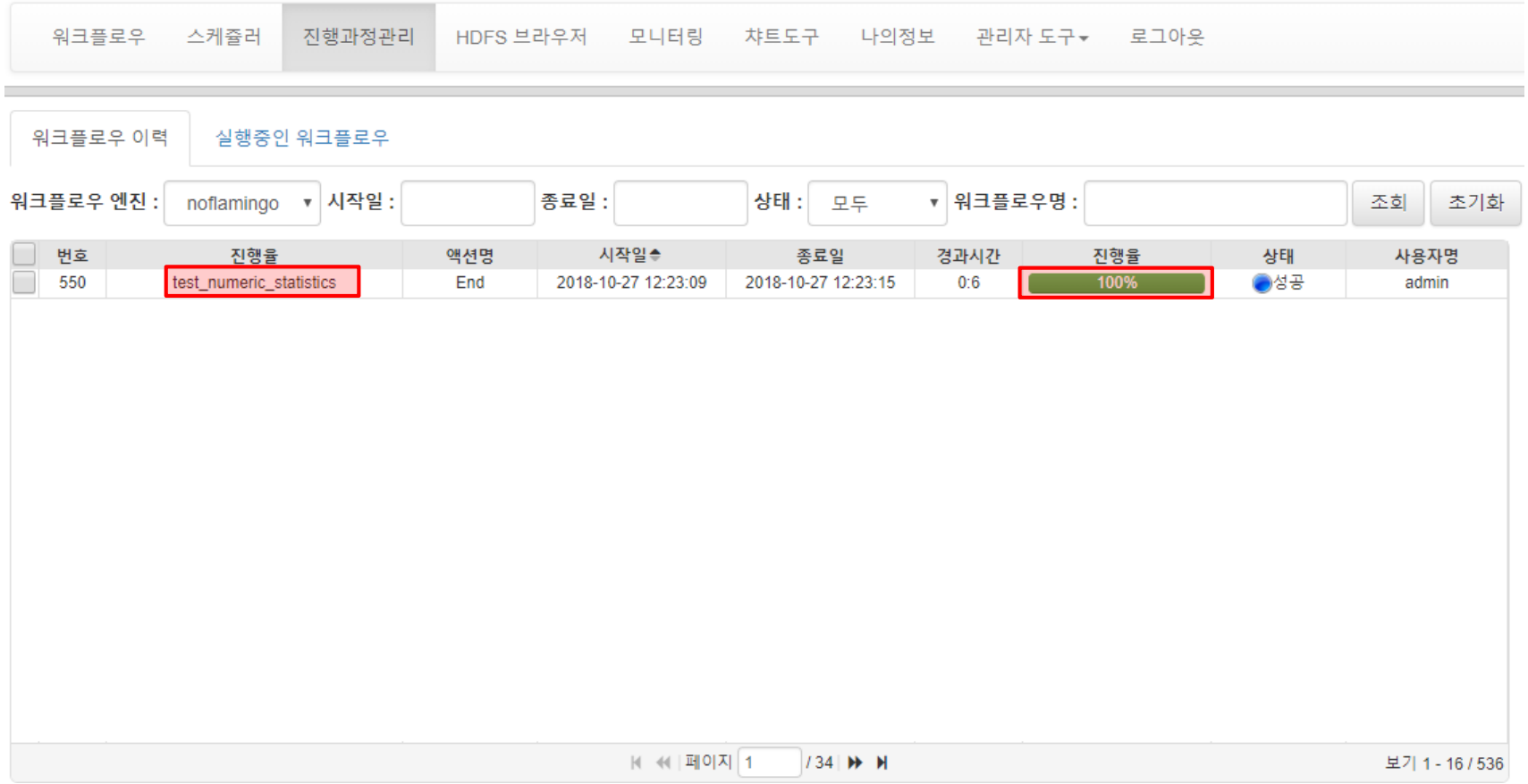
- **Numeric Statistics 세부 설정**
  - 입력 경로에서 찾기를 눌러 업로드한 iris.csv 파일 선택
  - 컬럼 구분자를 콤마로 변경
  - 출력 경로를 처음에 생성한 "test_ouput" 디렉토리로 선택
  - 파라미터 옵션
    * Target Attributes(Index) List : 0,1,2,3 선택



6. 활용 예제

6.2 통계분석 – Numeric Statistics(4/6)

- **워크플로우 처리 확인**
  - 진행과정관리 메뉴로 이동
  - 현재 실행중인 워크플로우 진행 결과 확인



6. 활용 예제

6.2 통계분석 – Numeric Statistics(5/6)

- **분석 결과 확인**
  - HDFS 브라우저 메뉴로 이동
  - test_output 디렉토리로 이동
  - reuslt 파일 선택 후 다운로드

6.2 통계분석 – Numeric Statistics(6/6)

- **분석 결과 확인**

  - 다운받은 result 파일을 엑셀로 열어 확인

Numeric Statistics (Distribution)

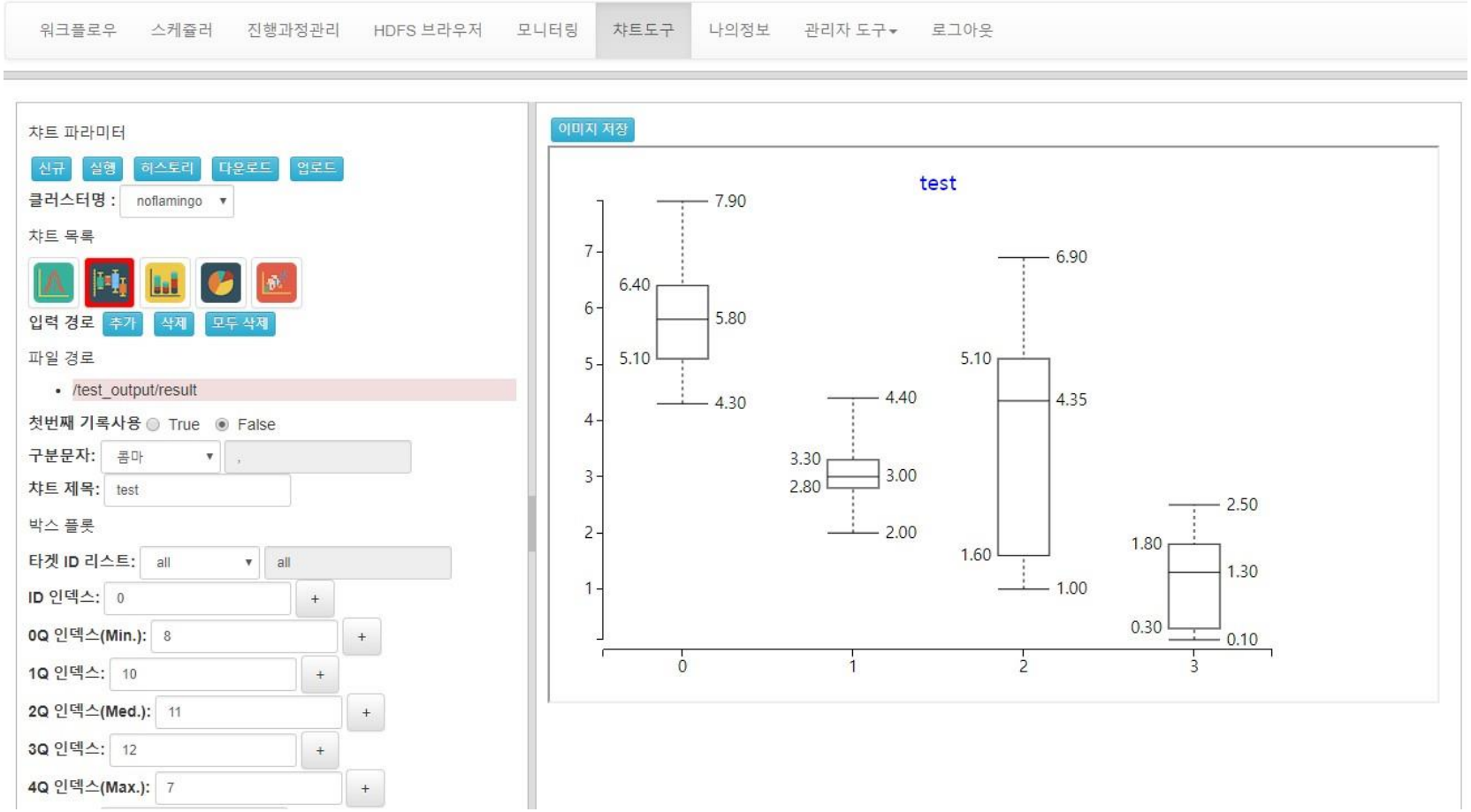
AttrIndex	Sum	Average	Harmonic/	Geographi	Variance	StandardD	Max	Min	DataCnt	1stQuartile	2ndQuartile	3rdQuartile
0	876.5	5.843333	5.728905	5.78572	0.681122	0.825301	7.9	4.3	150	5.1	5.8	6.4
1	458.1	3.054	2.993137	3.023582	0.186751	0.432147	4.4	2	150	2.8	3	3.3
2	563.8	3.758667	2.696472	3.239757	3.092425	1.758529	6.9	1	150	1.6	4.35	5.1
3	179.8	1.198667	0.486646	0.837827	0.578532	0.760613	2.5	0.1	150	0.3	1.3	1.8

0: 꽃받침 길이, 1: 꽃받침 너비, 2: 꽃잎 길이, 3: 꽃잎 너비에 대한 기초 통계

(합계(sum) / 평균(average) / 조화평균(harmonic average) / 기하평균(geographic average) / 분산(variance) / 표준편차(standard deviation) / 최대값(max) / 최소값(min) / 데이터 개수(data cnt.)/

1사분위수(1st Quartile) / 2사분위수(2nd Quartile) / 3사분위수(3rd Quartile))

- **통계분석 결과 차트로 확인**
  - 차트도구 메뉴로 이동

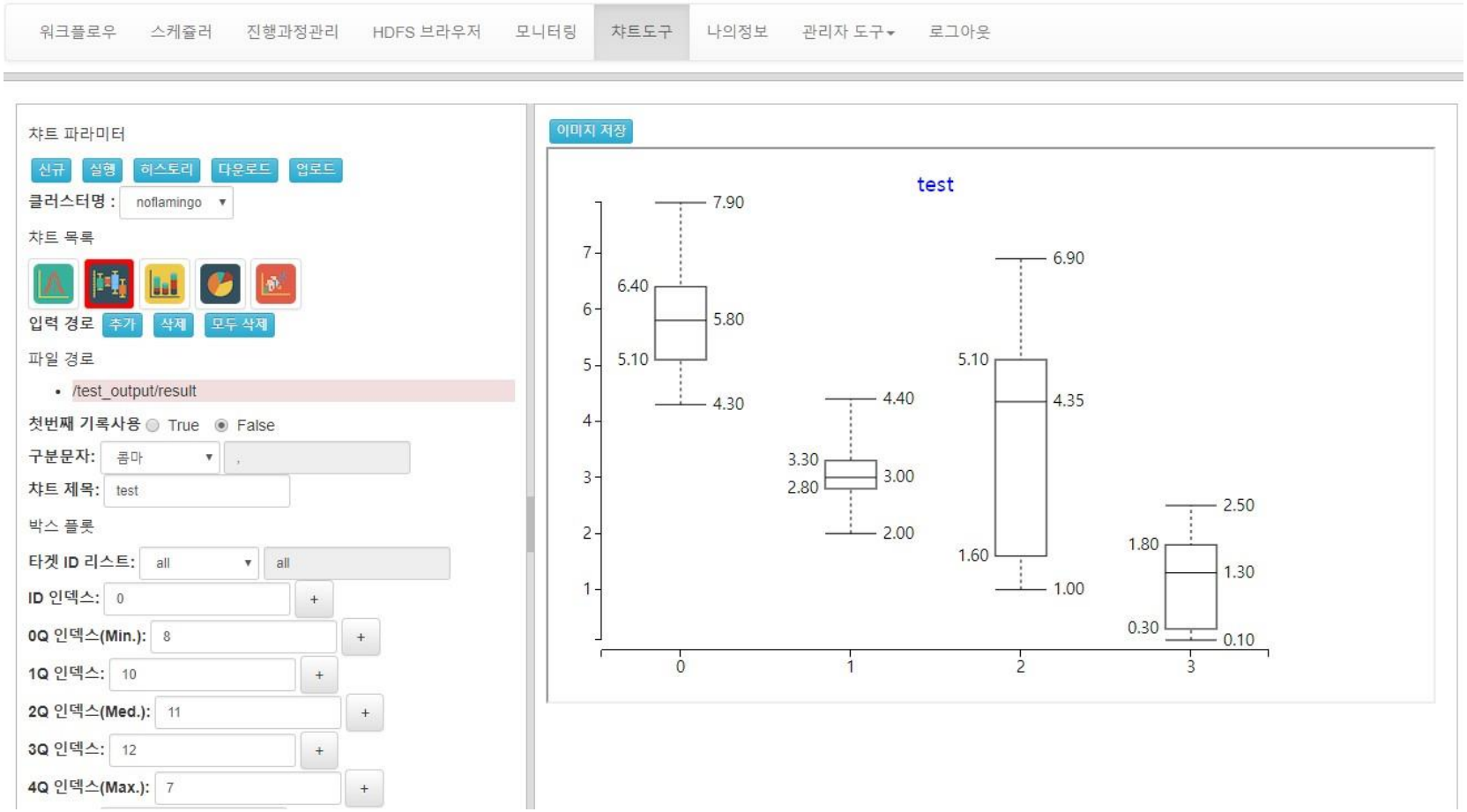


- **차트 상세 설정**
  - 박스플롯 차트 설정
  - 아래 설정 모두 입력 후 실행 버튼 클릭

설정 완료 후 click

1. 박스 플롯 차트 선택
2. 결과 파일 선택 /test_output/result
3. 입력 데이터 구분문자 콤마 선택
4. Chart 이름 입력
5. ID 인덱스 : 0 입력
6. 0Q 인덱스 : 8 입력
7. 1Q 인덱스 : 10 입력
8. 2Q 인덱스 : 11 입력
9. 3Q 인덱스 : 12 입력
10. 4Q 인덱스 : 7 입력

- **결과 확인**


워크플로우
스케쥴러
진행과정관리
HDFS 브라우저
모니터링
챠트도구
나의정보
관리자 도구
로그아웃
챠트 파라미터
신규
실행
히스토리
다운로드
업로드
클러스터명 :
noflamingo
챠트 목록
입력 경로
추가
삭제
모두 삭제
파일 경로
/test_output/result
첫번째 기록사용
True
False
구분문자:
콤마
,
챠트 제목:
test
박스 플롯
타겟 ID 리스트:
all
all
ID 인덱스:
0
0Q 인덱스(Min.):
8
1Q 인덱스:
10
2Q 인덱스(Med.):
11
3Q 인덱스:
12
4Q 인덱스(Max.):
7
이미지 저장
test
7.90
6.40
5.80
5.10
4.30
4.40
3.30
3.00
2.80
2.00
6.90
5.10
4.35
1.60
1.00
2.50
1.80
1.30
0.30
0.10
0
1
2
3

7. FAQ

Q ankus Analyzer G 는 실시간 분석, 배치 분석 등이 모두 가능한가요?

A ankus Analyzer G 는 실시간 분석(일부), 배치 분석 등이 모두 가능합니다. ankus Analyzer G 는 Hadoop 기반 환경에서 수행되는 분석 도구 입니다. 기존의 분석 모델을 수집되는 데이터에 바로 적용하는 필터 형태로 쓰는 경우에는 현재도 ankus Analyzer G 는 적용 가능합니다.

Q ankus Analyzer G 를 사용시 라이선스는 어떻게 되나요?

A ankus Analyzer G 의 라이선스(GPL v3)는 무료이며, 소스코드를 사용자 제품에 포함시키거나 혹은 그와 함께 배포하는 경우에는 해당 소스에 설치정보를 함께 제공해야 합니다.다만, 기술지원 서비스와 더 다양한 분석 알고리즘이 필요한 사용자는 ankus의 컨설팅을 제공하는 업체를 통해 기술 지원을 받을 수 있습니다.

8. 용어정리

용어	설명
Apache Hadoop	Apache Hadoop 소프트웨어 라이브러리는 간단한 프로그래밍 모델을 사용하여 여러대의 컴퓨터 클러스터에서 대규모 데이터 세트를 분산 처리 할 수있게 해주는 프레임워크. 단일 서버에서 수천대의 머신으로 확장 할 수 있도록 설계됨
데이터 마이닝 (Data mining)	데이터 마이닝은 대규모로 저장된 데이터 안에서 체계적이고 자동적으로 통계적 규칙이나 패턴을 찾아 내는 것임, 다른 말로는 KDD(데이터베이스 속의 지식 발견, knowledge-discovery in databases)라고도 일컫음
기계 학습 (Machine Leaning)	인공 지능의 한 분야로, 컴퓨터가 학습할 수 있도록 하는 알고리즘과 기술을 개발하는 분야를 말함, 가령, 기계 학습을 통해서 수신한 이메일이 스팸인지 아닌지를 구분할 수 있도록 훈련
Data Warehouse (DW)	데이터 웨어하우스란 사용자의 의사 결정에 도움을 주기 위하여, 기간시스템의 데이터베이스에 축적된 데이터를 공통의 형식으로 변환해서 관리하는 데이터베이스, 줄여서 DW로도 불림
컴퓨터 클러스터 (Computer Cluster)	여러 대의 컴퓨터들이 연결되어 하나의 시스템처럼 동작하는 컴퓨터들의 집합, 클러스터의 구성 요소들은 일반적으로 고속의 근거리 통신망으로 연결, 서버로 사용되는 노드에는각각의운영체제가실행

Open Source Software Installation & Application Guide

이 저작물은 크리에이티브 커먼즈 [저작자표시-비영리-동일조건 변경허락 2.0 대한민국 라이선스]에 따라 이용하실 수 있습니다.