빅 데이터 분석 및 데이터 처리
웹 관리 도구

어니컴 / 전수현
shjeon@onycom.com

ONYCOM

Agenda

- Motivation
- Mission
- Introduction
- Development Environment
- Main Committer
- Feature
- Milestone
- Sites
- Demo

Motivation

- Hadoop 중심의 Big Data EcoSystem의 출현으로 기존 솔루션 활용의 어려움
  - ETL, Scheduler, Import & Export, Mining Algorithm 등 모든 것이 Hadoop을 중심으로 통합되어야 하는 특징
- 데이터 프로세싱을 다루는 사람들은 기본적으로 UI를 기반으로 업무를 수행 하나 적절한 도구는 존재하지 않음
- 현장 요구사항
- 데이터를 다루기 위해서 필요한 순쉬운 개발 도구의 필요성

Mission

- MapReduce, Pig, Hive 등 다양한 오픈소스가 잘 통합될 수 있는 UI를 제공한다.
- **개발**부터 **운영**까지 모두 적용할 수 있어야 한다.
- 복잡하지 않은 **단순한** UI 및 현장 요건을 충분히 수렴한 UI를 제공해야 한다.
- MapReduce 모듈을 **자유롭게** 추가하여 운영할 수 있어야 한다.
- 기존 상용 솔루션과 **잘 연계**하도록 한다. (예; 타 솔루션)

Introduction

- 과제명
  - 빅 데이터 분석 및 처리를 위한 웹 관리 도구 (Open Flamingo Hadoop Manager)
    - NIPA 커뮤니티 지원 사업 프로젝트로 진행 중

- 주요 목표
  - 누구나 손쉽게 빅 데이터 기술을 활용하여 데이터를 가공할 수 있는 개발 및 운영 환경을 제공하는 것

Introduction

- 지원 기능
  - HDFS Browser (0.2)
  - MapReduce Job Monitoring (Dashboard) (0.2)
  - Workflow Engine & Web based Designer for Big Data Processing (0.2)
  - Hadoop EcoSystem & Commercial Solution Integration (0.3)
  - File & Hadoop Cluster Management (0.3)
  - MapReduce based ETL & Mining Algorithm (0.3)
  - Web based MapReduce, Pig, Hive Editor (0.4)
- 라이센스
  - GPLv3 (Web User Interface)
  - Apache License 2.0 (Server + CLI)

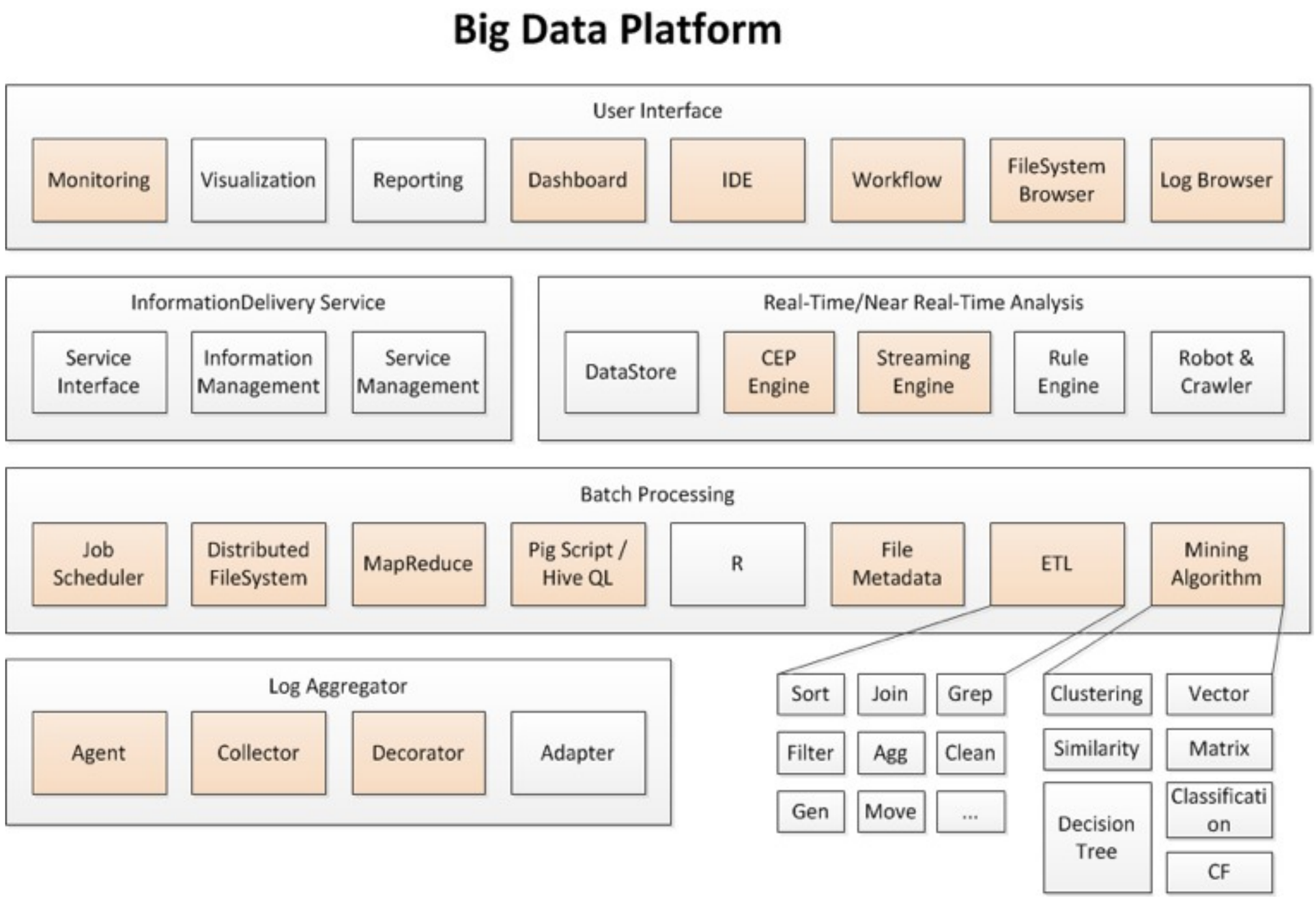
Development Environment

- Used Framework
  - Spring Framework
  - Quartz Job Scheduler, MySQL
  - Sencha ExtJS
  - OpenGraph
- Project Management : Apache Maven
- Build Server : TeamCity
- Issue Tracker : Atlassian JIRA
- WIKI : Atlassian Confluence

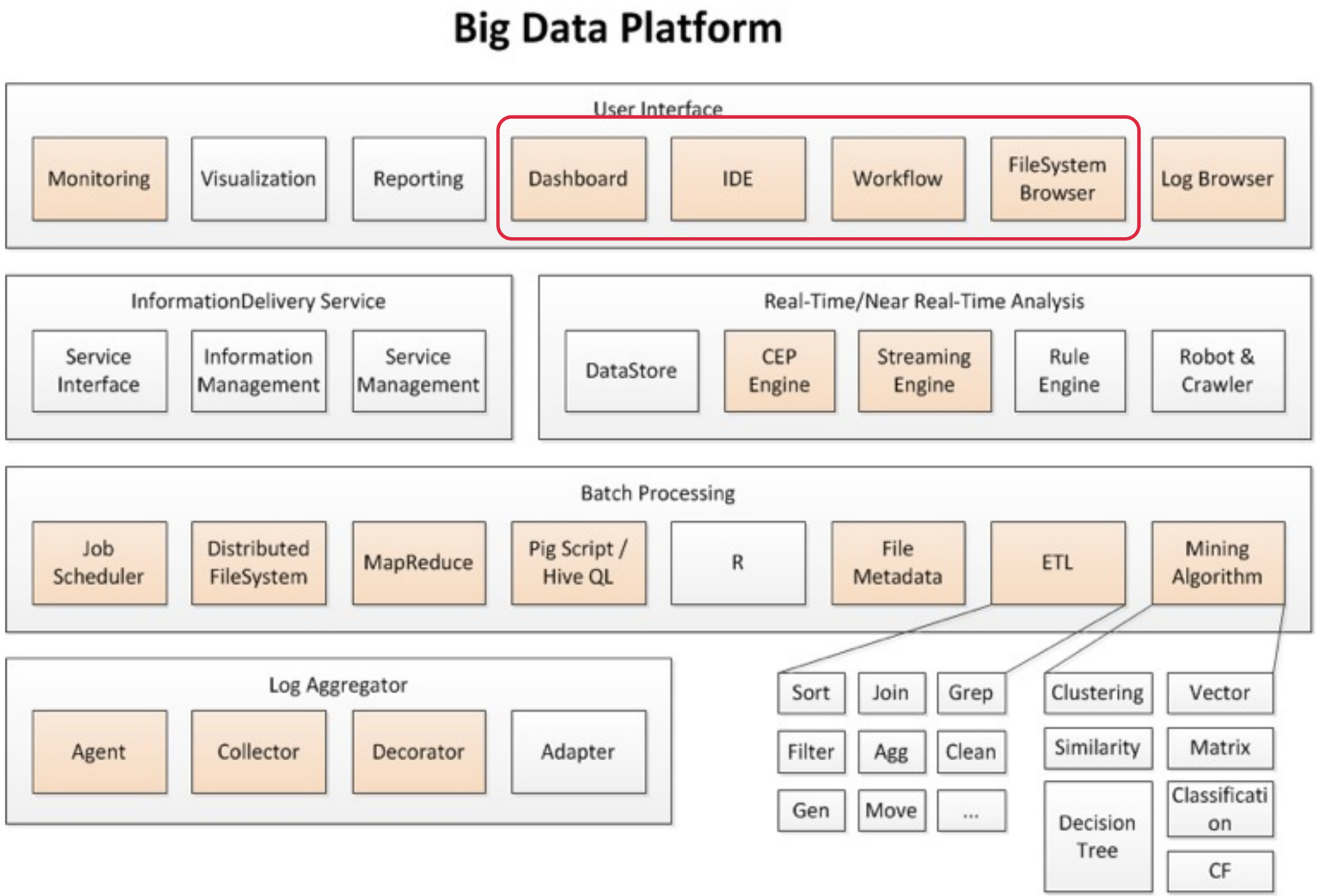
Main Committer

- **김병곤**
  - JBoss User Group 운영자
  - 지경부/SW 마에스트로 멘토
  - Flamingo Project 첫시작
- **전수현**
  - 여자개발자모임터 운영자
  - 한국공개소프트웨어(KOSSA) 멘토
- **이승백**
  - 오픈소스 Gant Chart & Open Graph 개발자
  - JavaScript 전문가

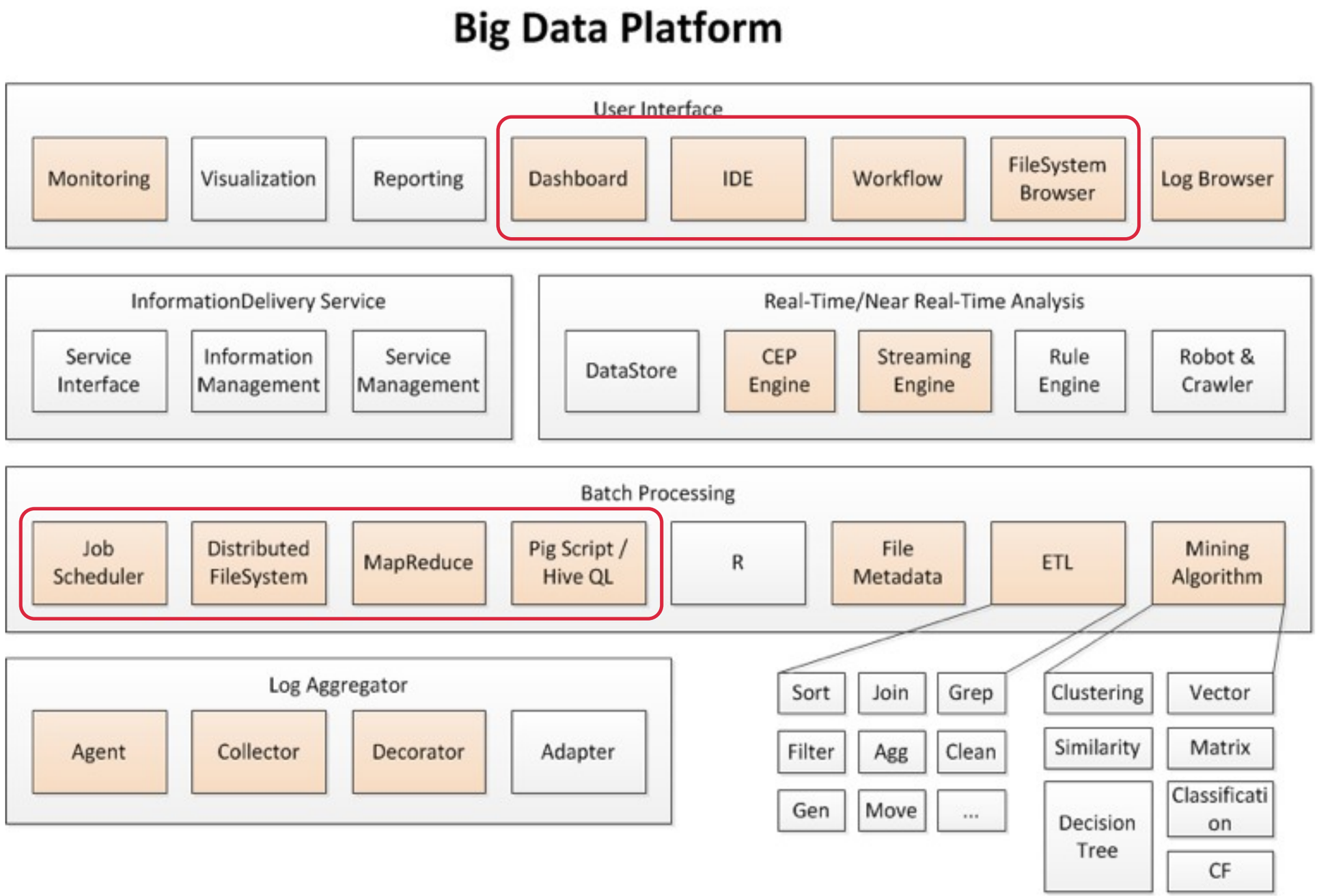
Our Big Data Platform View



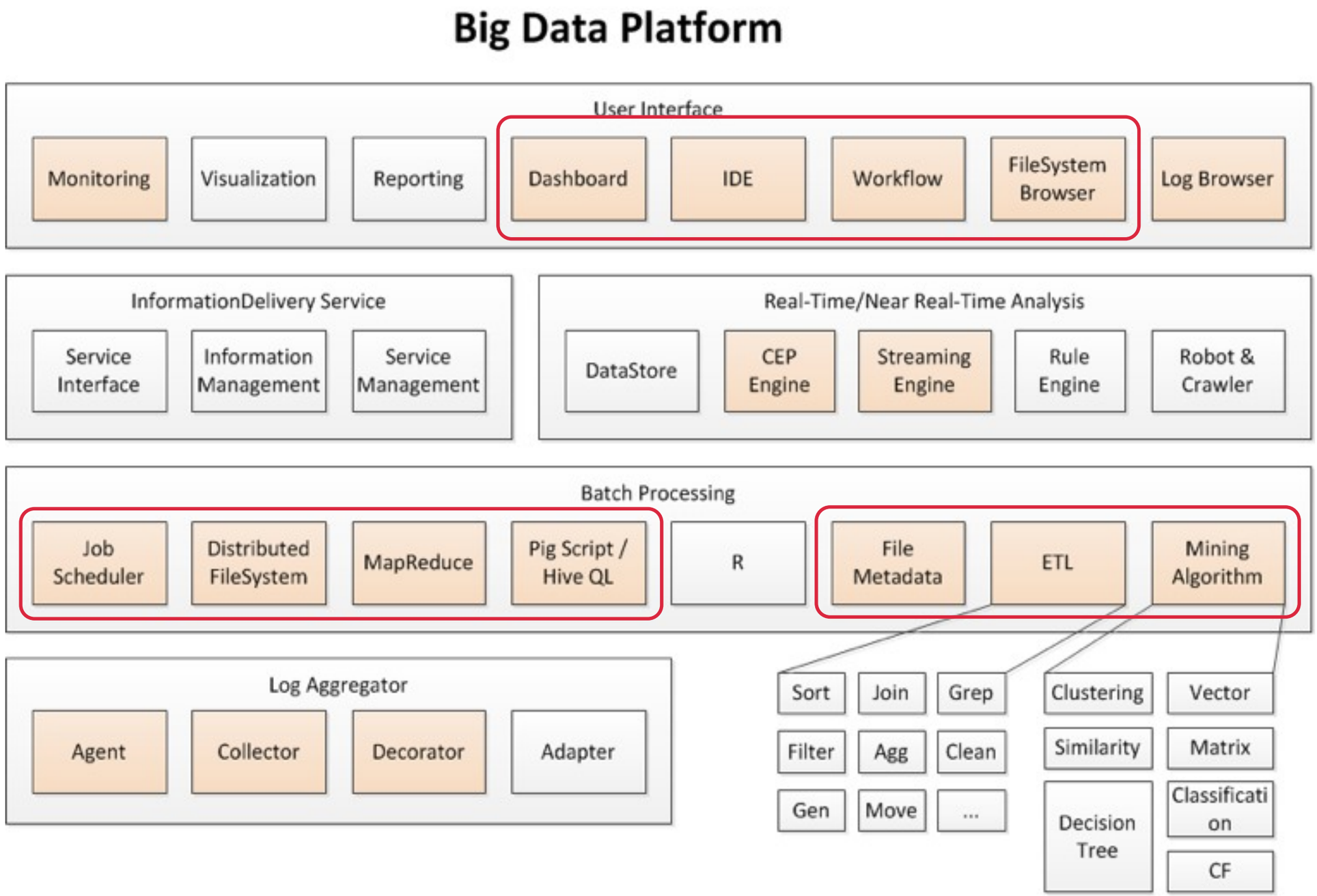
Our Big Data Platform View



Our Big Data Platform View

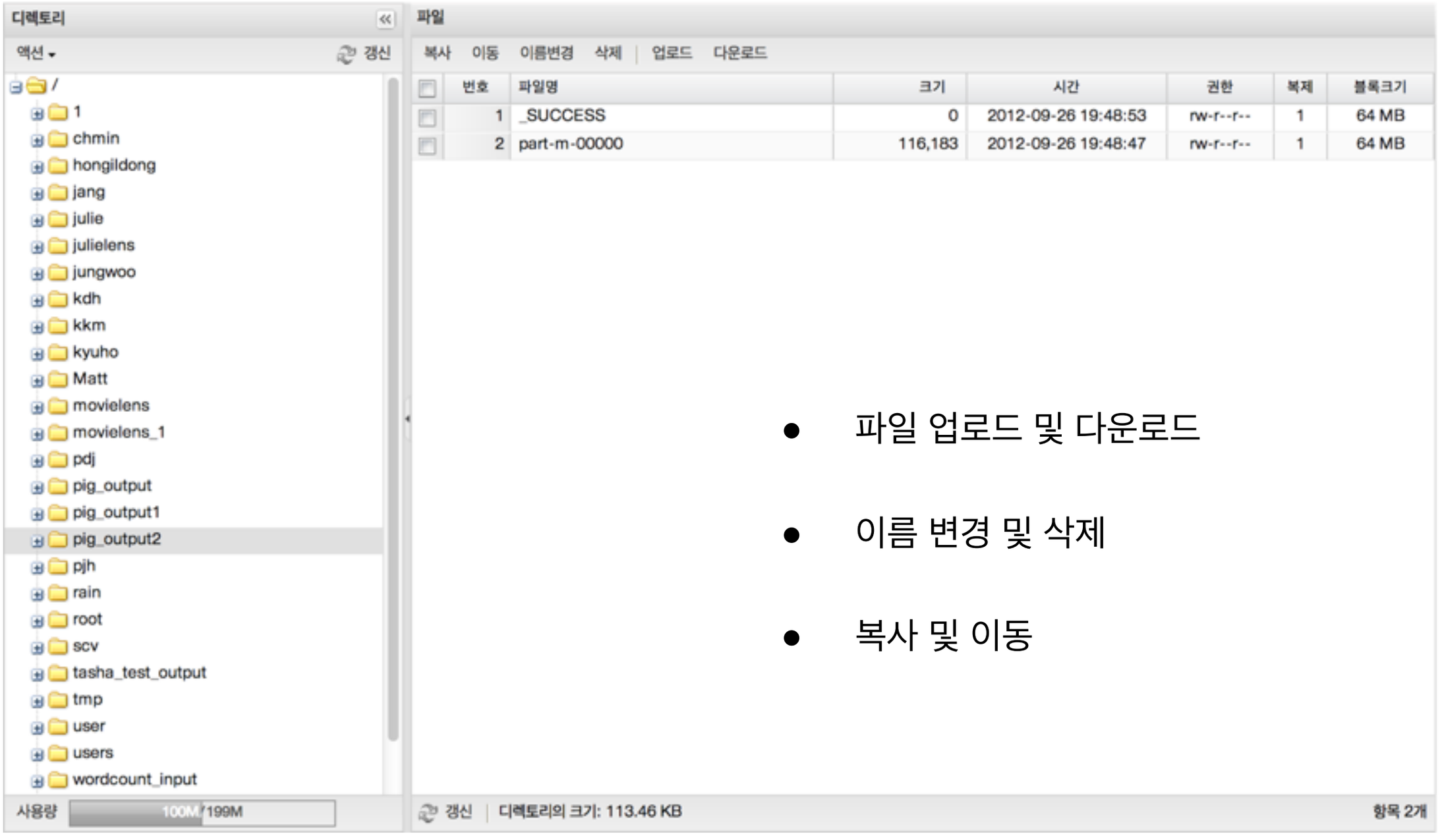


Our Big Data Platform View



Feature - HDFS Browser

- 개발 및 관리 시 가장 많이 사용하는 기능
- 파일 보기 기능 및 다량의 파일이 디렉토리에 있는 경우 지원해야 함



- 파일 업로드 및 다운로드
- 이름 변경 및 삭제
- 복사 및 이동

Feature - HDFS Browser (예정)

- 멀티 Hadoop Cluster 선택 기능 (0.2)
- 텍스트 파일 뷰어 (0.3)
- 드래그 앤 드롭 형태의 디렉토리 및 파일 관리 (0.3)
- 시퀀스 파일 변환 기능 (0.4)
- 대용량 멀티 파일 업로드 기능 (0.4)

Feature - Command Line Interface

```
#flamingo
Flamingo Client에서 사용할수 있는 커맨드 목록입니다.
사용방법: flamingo 커맨드
'커맨드'자리에는 다음의 커맨드를 사용할 수 있습니다.
job <커맨드> :  지정한 배치 작업 제어 관련 커맨드를 실행합니다.
        -list:  현재 등록되어 있는 Job 및 워크플로우 목록을 표시합니다.
        -regist <워크플로우 XML 파일> <.properties>:    지정한 워크플로우 파일을 등록합니다.
        -get <워크플로우 Instance ID>:  등록되어 있는 워크플로우를 파일로 저장한다.
        -run <워크플로우 Instance ID>:  지정한 워크플로우를 실행합니다.
        -delete <워크플로우 Instance ID 또는 배치 작업 ID>:     지정한 워크플로우 또는 배치 작업을 삭제합니다.
        이미 스케줄링 되어 있는 워크플로우는 삭제할 수 없습니다.
        삭제하려면 스케줄링 되어 있는 배치 작업을 중지하고 삭제한 이후에 워크플로우를 삭제해야 합니다.
        -stop <배치 작업 ID>:   실행중인 배치 작업의 스케줄링을 중지합니다. 중지한 이후에는 더이상 동작하지 않습니다.
        -pause <배치 작업 ID>:  실행중인 배치 작업의 스케줄링을 일시 중지합니다.
        -resume <배치 작업 ID>:         일시 중지중인 배치 작업의 스케줄링을 다시 시작합니다.
        -schedule "<Cron Expression>" <워크플로우 ID> <배치 작업명>:    지정한 워크플로우를 스케줄링합니다.
... 생략

#flamingo job -regist /home/flamingo/workflows/recommendation.xml
#flamingo job -schedule “0 30 * * * * ?” WF_2011090909_1 “영화 평가 점수 기반 추천 프로세스”
```

Feature - Workflow Engine

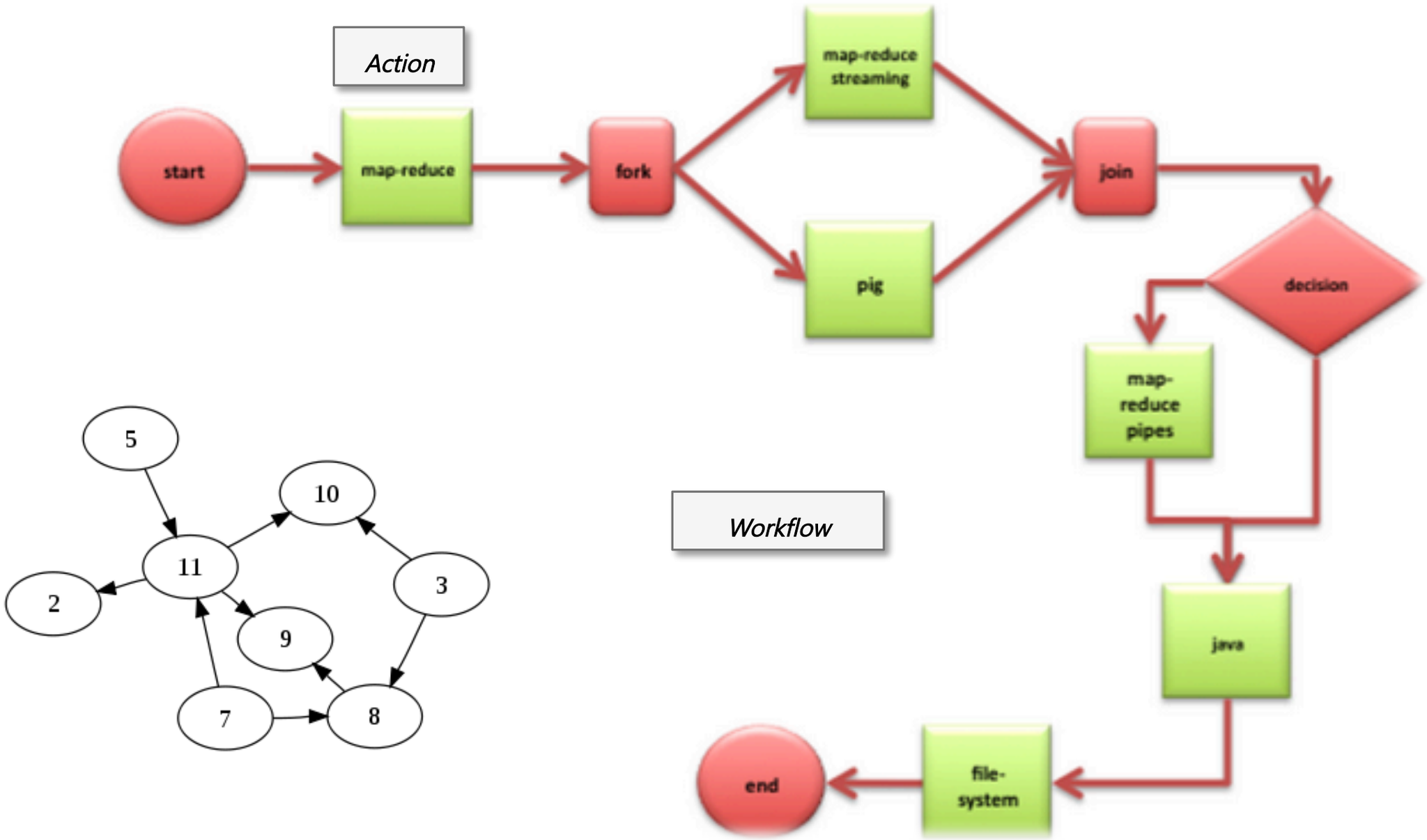
- Apache Oozie의 문제점
  - 한국 환경을 제대로 지원하지 않는 **Timezone** 문제
  - Timezone 문제로 인한 Coordinator Service (Scheduler)의 기능 오류
  - 매우 **형편 없는** 스케줄링 기능(Interval 기반)
  - **복잡한** 사용법
  - UI Integration의 **어려움**
    - 그래프를 그리는 방법에 따라서 실패와 성공이 결정됨
    - 그래프의 각 노드는 모두 완전한 단독으로만 동작하도록 되어 있음

Feature - Workflow Engine

- 복잡한 UI를 구성하더라도 최대한 손쉽게 그릴 수 있는 구조
- Direct Acyclic Graph (DAG) 기반 엔진
- XML 기반 워크플로우
- Expression Language 지원 (`${timestamp('yyyyMMdd')}`)
- Cron Trigger 기반 Job Scheduler
- Command Line Interface 지원
- MapReduce, Shell, Pig, Hive 지원
    - (예정) Sqoop, FTP

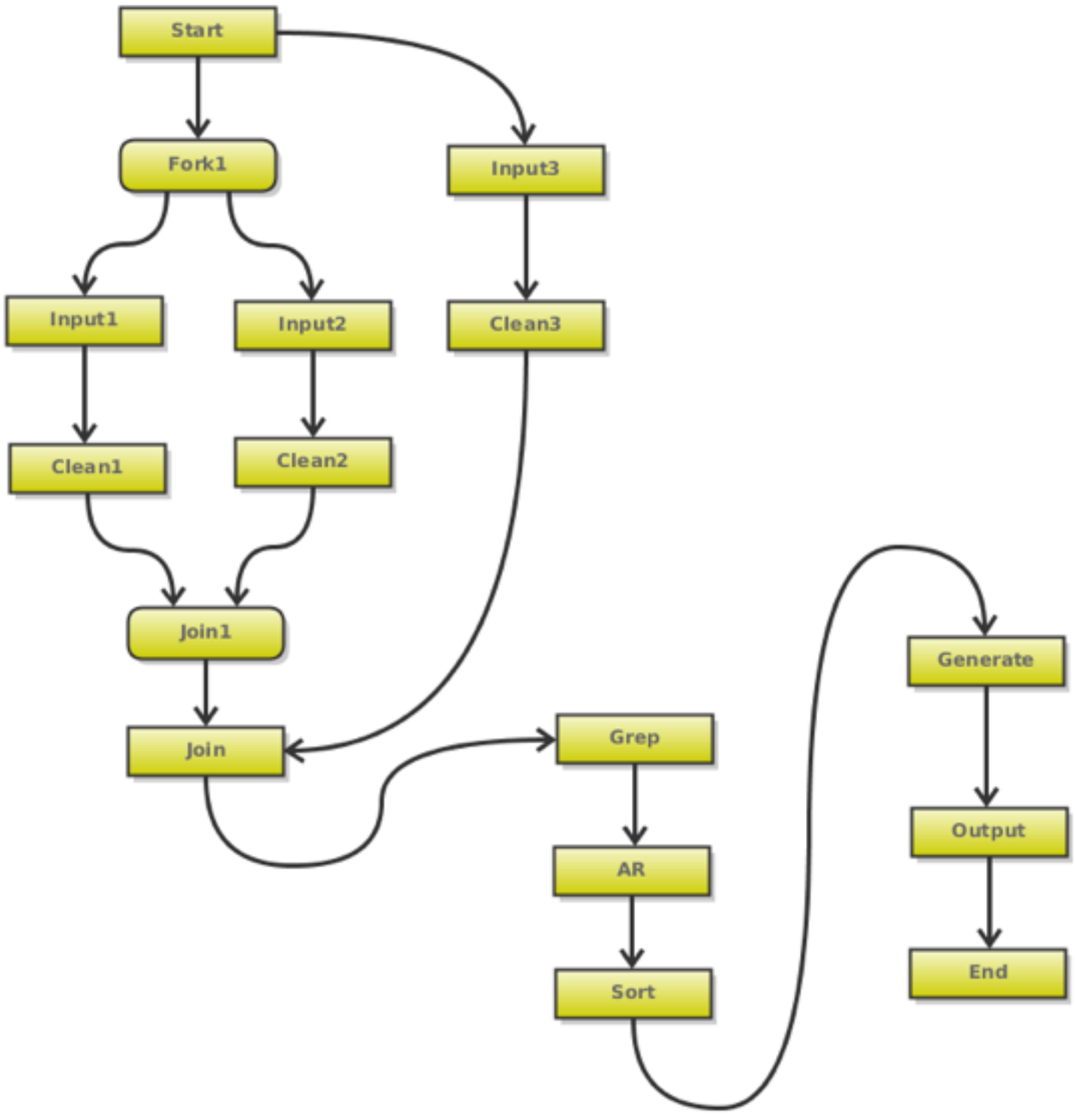
Feature - Workflow Engine - DAG Engine

- Direct Acyclic Graph (DAG) :: 순환하지 않는 그래프



Feature - Workflow Engine - DAG Engine

- Apache Oozie는 어떻게 그래프를 그리느냐에 따라서 동작이 실패할 수 있지만 Flamingo의 그래프는 제한이 없는 그래프를 가짐



Feature - Workflow Engine - EL

```
<action name="s1" to="endStep" description="">
  <shell>
    <workingDirectory>${user.dir}</workingDirectory>
    <program>${user.dir}/src/test/resources/helloworld.sh</program>
    <args>
      <variable value="${workflowId}"/>
      <variable value="${uniqueId}"/>
      <variable value="${date}"/>
      <variable value="${actionId}"/>
      <variable value="hello"/>
      <variable value="world"/>
      <variable value="${user.dir}"/>
      <variable value="${user.dir},${user.home}"/>
      <variable value="${GB}-${MB}"/>
      <variable value="${user.dir}${concat('a','b')}"/>
    </args>
    <envs>
      <variable name="A" value="1"/>
      <variable name="B" value="2"/>
    </envs>
  </shell>
</action>
```

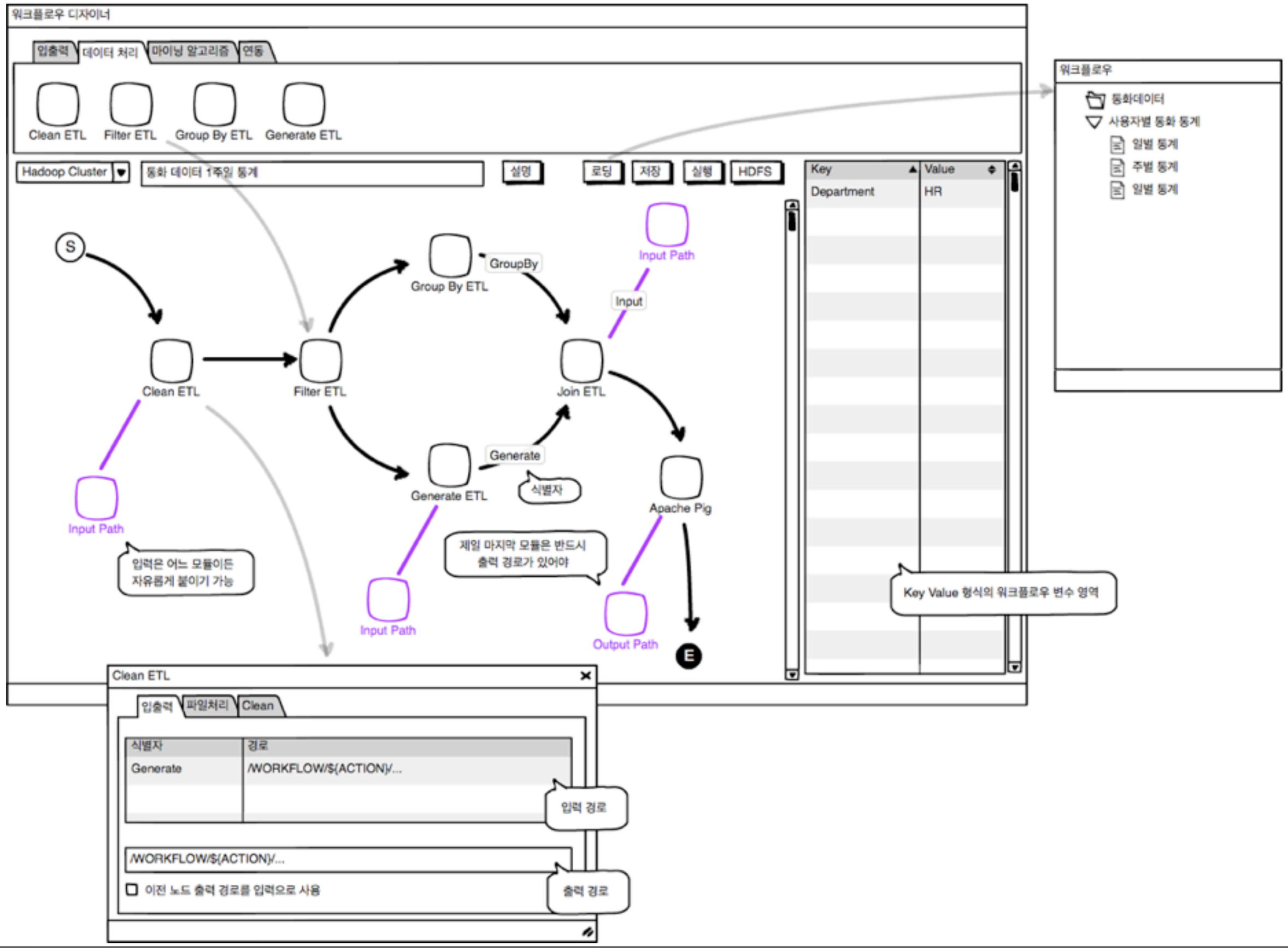
Feature - Workflow Engine - MapReduce

```
<action name="cleanETL2" to="endStep" description="Prepare, Output을 명시적으로 지정한 Clean ETL MR">
  <mapreduce>
    <prepare>
      <delete path="/movielens_output"/>
    </prepare>
    <clusterName>dev</clusterName>
    <jar>/home/hadoop-manager/trunk/src/test/resources/flamingo-mapreduce-0.1.jar</jar>
    <className>clean</className>
    <variables>
      <variable name="input" value="${inputPath}"/>
      <variable name="output" value="${outputPath}"/>
      <variable name="inputDelimiter" value="${delimiter}"/>
      <variable name="outputDelimiter" value="${delimiter}"/>
      <variable name="columnsToClean" value="1"/>
      <variable name="columnSize" value="5"/>
    </variables>
    <input>
      <inputPaths>
        <inputPath>/movielens/users.dat</inputPath>
      </inputPaths>
    </input>
    <output>
      <outputPath>/movielens_output</outputPath>
    </output>
  </mapreduce>
</action>
```

Feature - Workflow Engine - Apache Pig

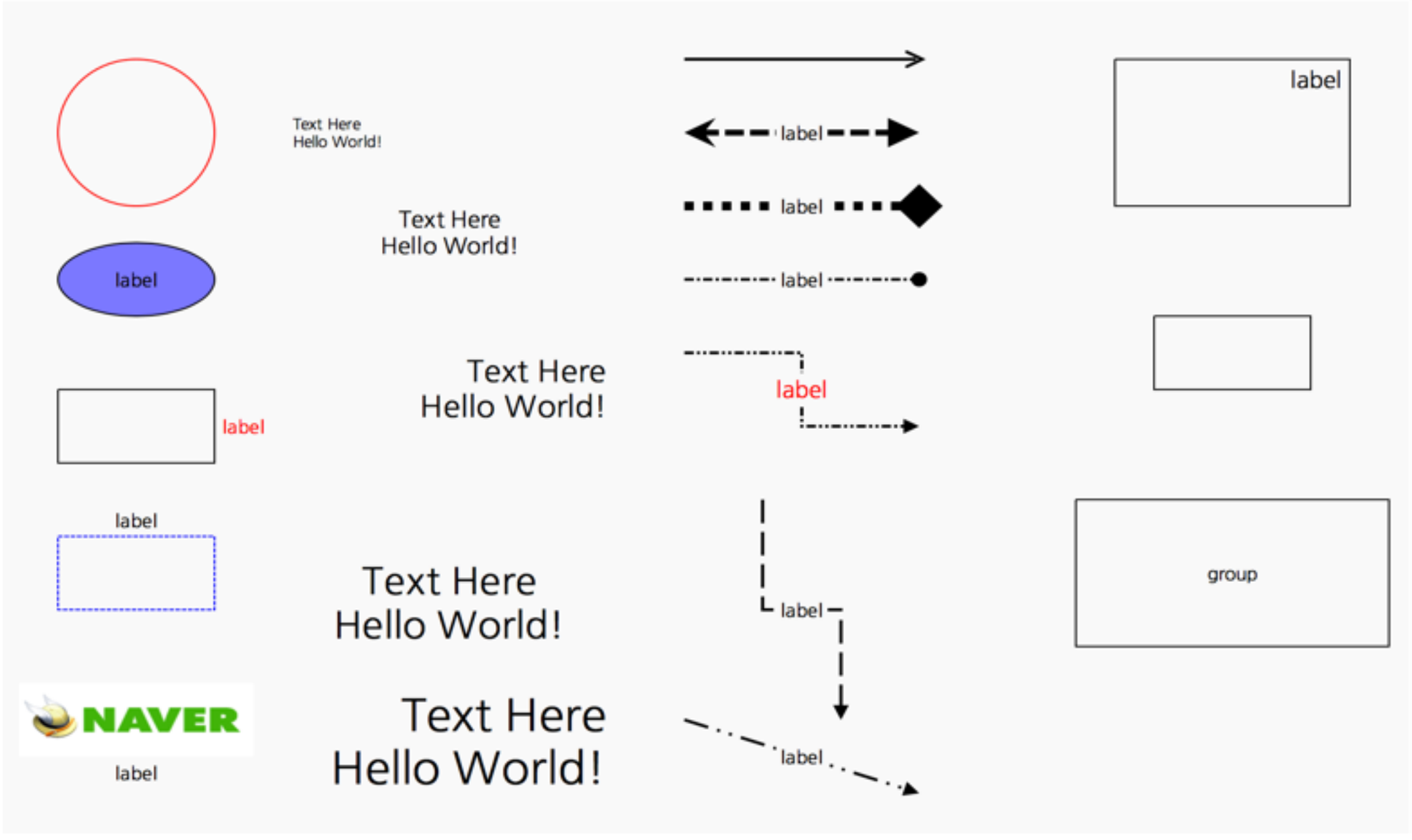
```
<action name="Pig Job" to="endStep" description="Order By Pig Job">
  <pig>
    <clusterName>dev</clusterName>
    <variables>
      <variable name="INPUT" value="/movielens"/>
      <variable name="delimiter" value="^"/>
    </variables>
    <script>
      <![CDATA[
        A = LOAD '${INPUT}' Using PigStorage('${delimiter}');
        B = ORDER A BY $0 DESC;
        STORE B INTO '/movielens_output' USING PigStorage('${delimiter}');
      ]]>
    </script>
  </pig>
</action>
```

Feature - Workflow Designer

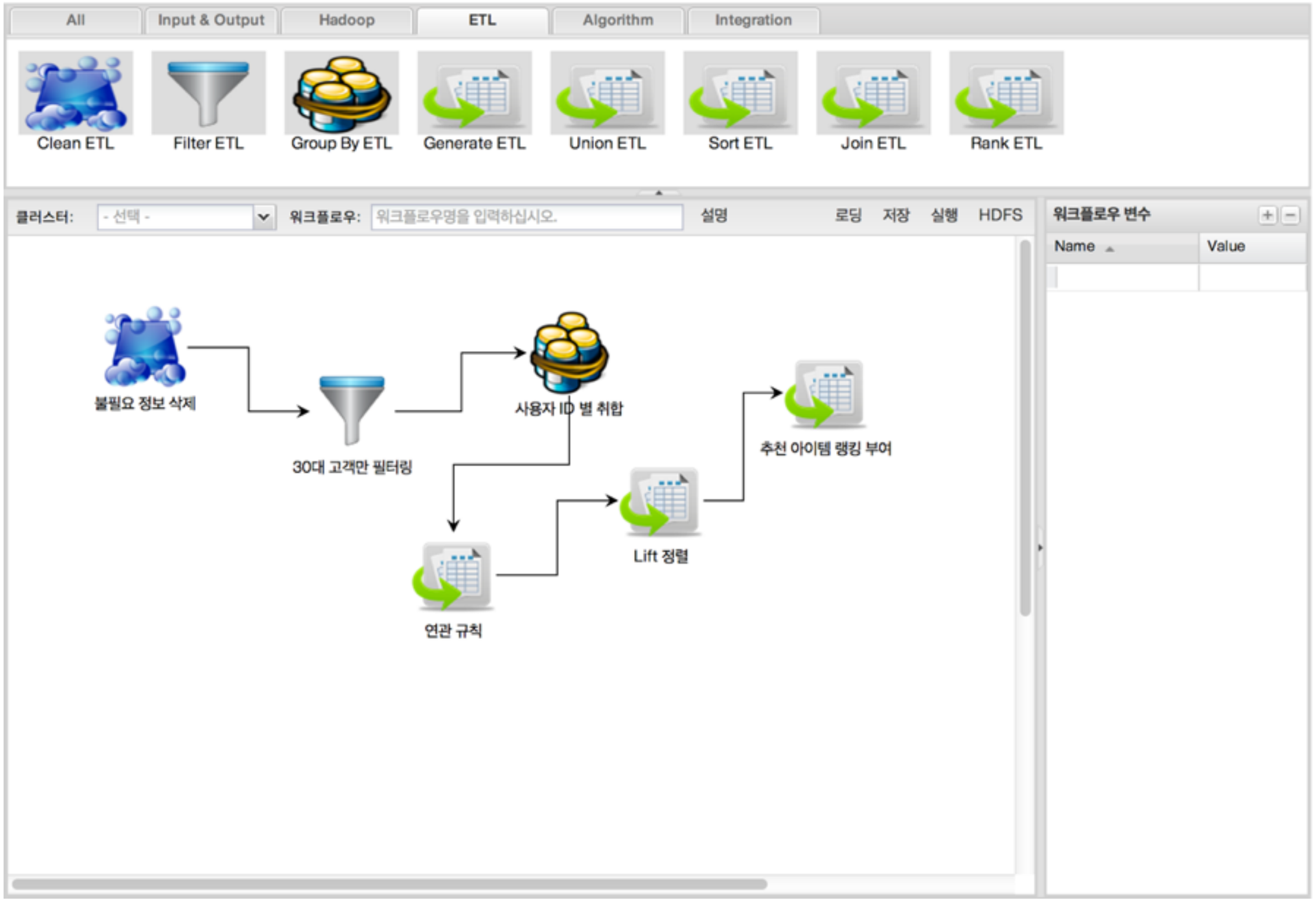


Web based Drawing Framework

MxGraph와 같은 javascript 그래프 엔진은 웹 기반 워크플로우를 그려내기 위한 핵심 기술

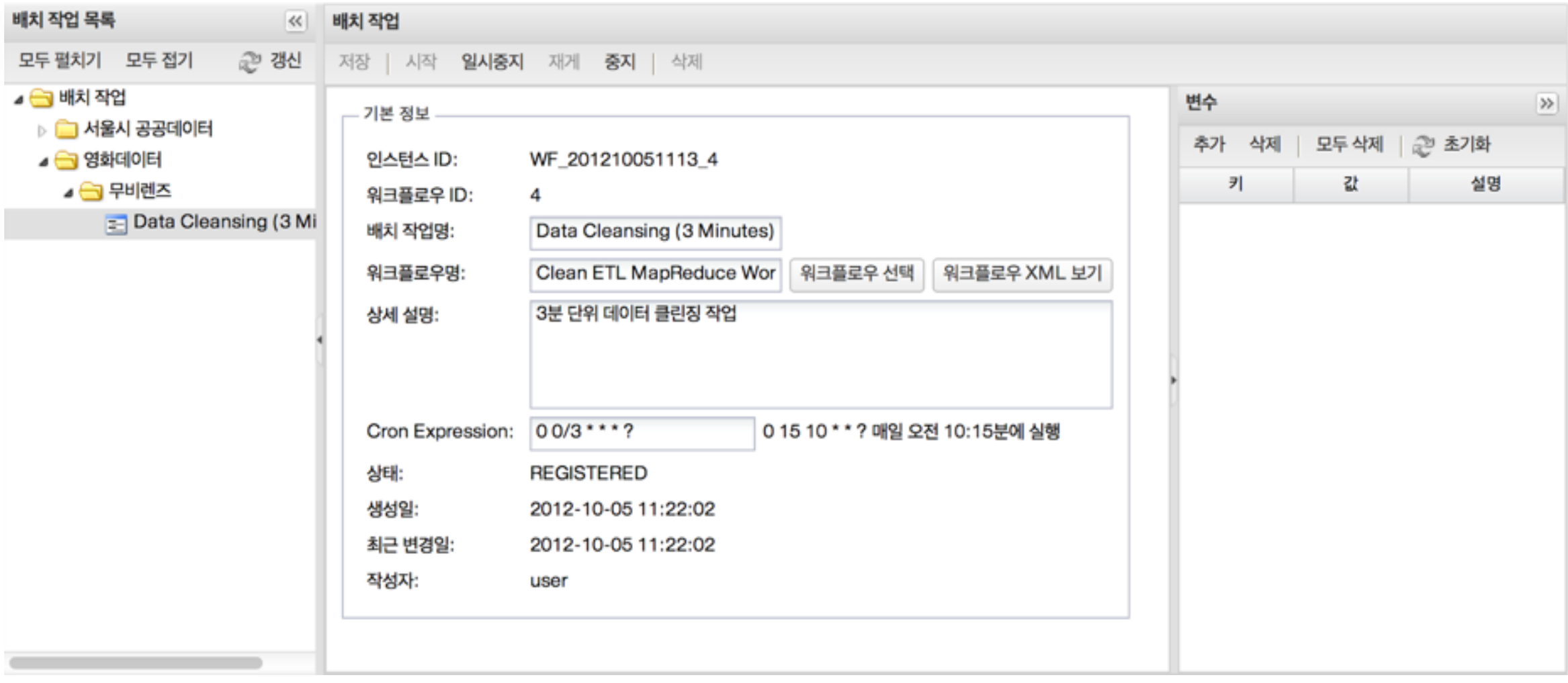


Feature - Workflow Designer

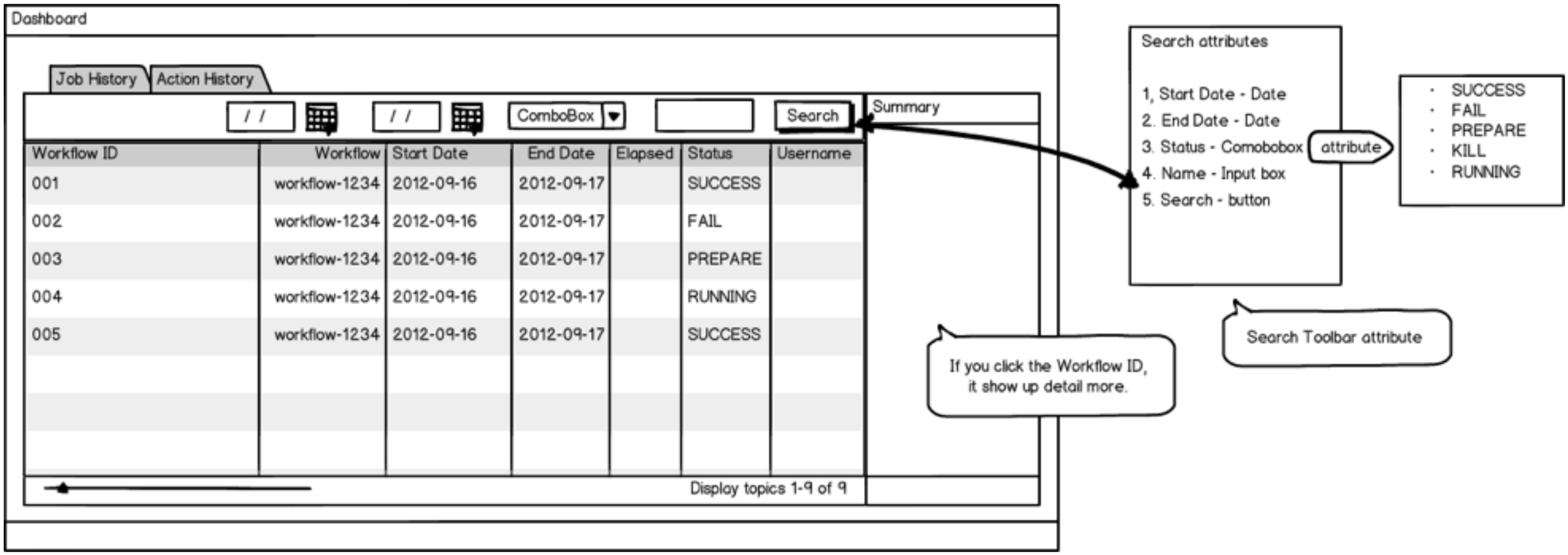


Feature - Job Scheduler

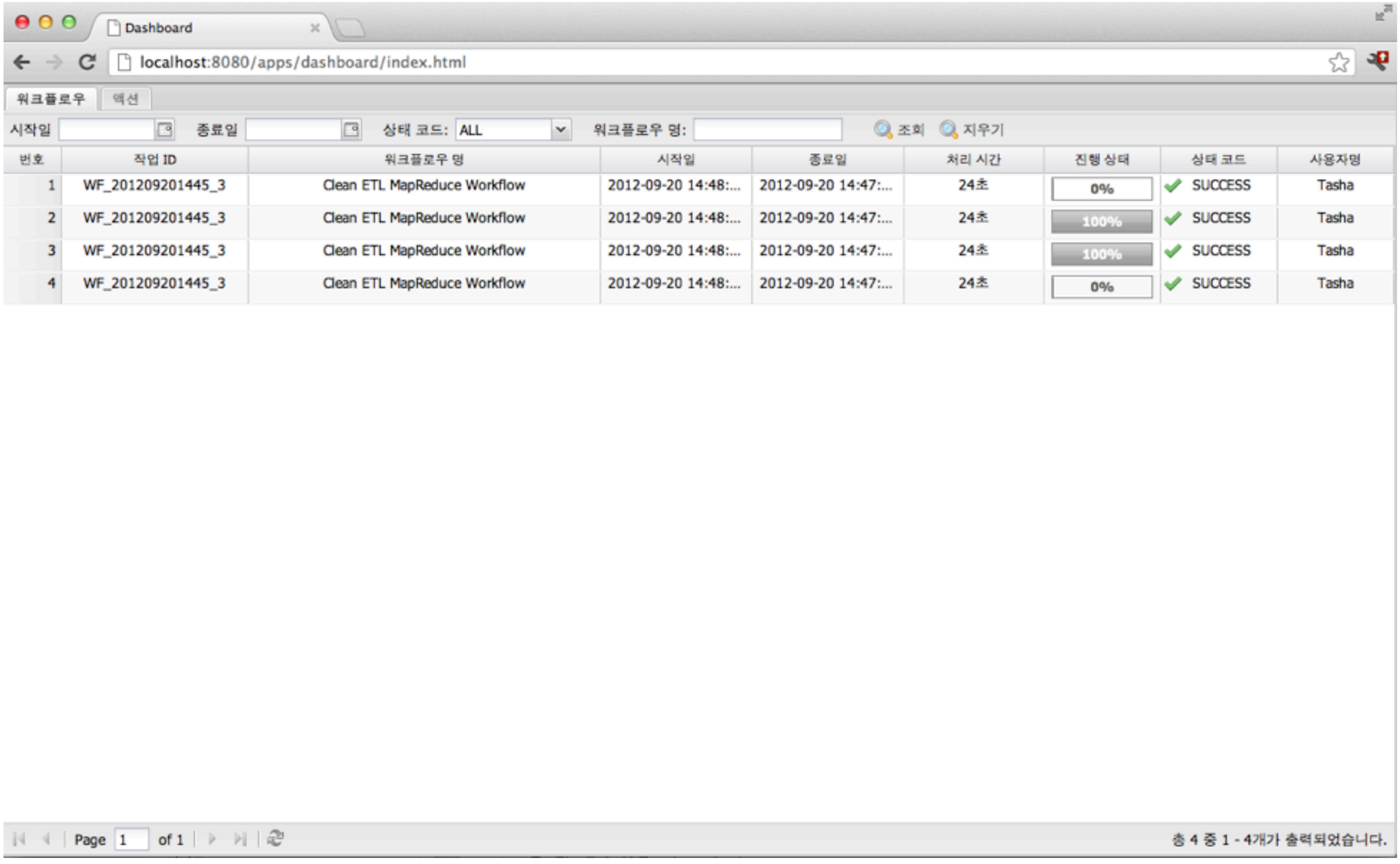
- 작성한 워크플로우를 등록하고 스케쥴링 하는 기능
- Cron Trigger 방식, Cron Trigger 등록은 문자열 등록이 아닌 UI 기반으로 변경 중
- 워크플로우에 적용되어 있는 변수를 적용하는 기능 제공



Feature - Dashboard



Feature - Dashboard



Milestone

Version	Date	Features
0.2	2012.11.30	Workflow Job Dashboard HDFS Browser Command Line Interface Workflow Engine Workflow Designer Flamingo ETL Integration
0.3	2013.1.30	HDFS Browser - Drag And Drop, 텍스트 파일 뷰어 Apache Mahout Integration 타 솔루션 Integration Sqoop, FTP Integration Flamingo ETL Integration
0.4	2013.3.30	File Metadata Management Pig & Hive IDE 대용량 멀티 파일 업로드 시퀀스 파일 변환 기능

Sites

프로젝트 커뮤니티	Open Flamingo facebook page
소스코드 브라우저	fisheye.openflamingo.org
위키	confluence.openflamingo.org/display/DESIGN/Hadoop+Manager
이슈 트래커	jira.openflamingo.org