

## 2.5 빅데이터 클러스터 구성

### 기본 소프트웨어 설치

하둡, 주키퍼 등 기본구성

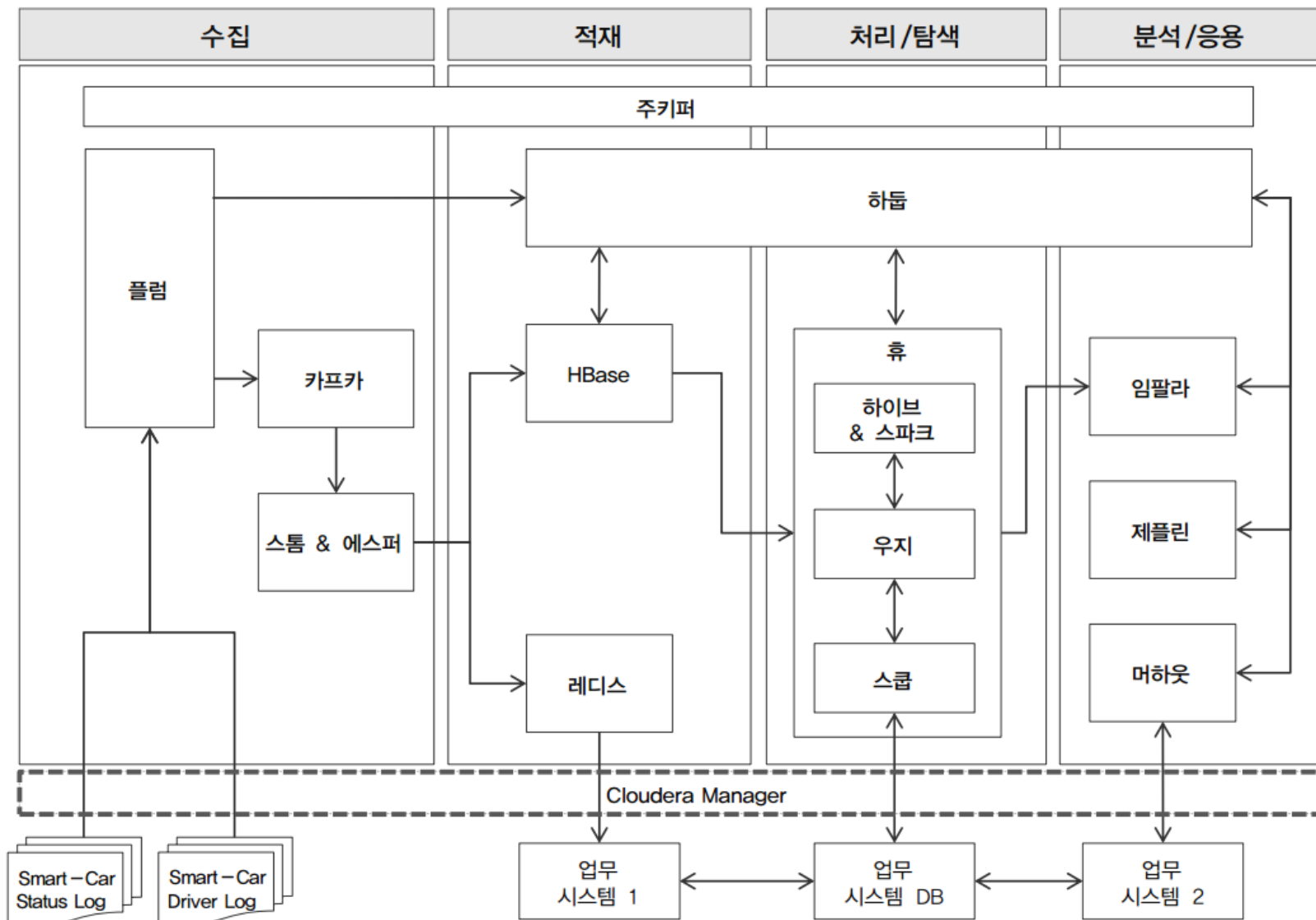


그림 2.65 빅데이터 파일럿 프로젝트 아키텍처에서 CM 영역

## 클러스터 설치

✓ 시작

✓ Cluster Basics

● Specify Hosts

○ 리포지토리 선택

## Specify Hosts

1 **i** In Cloudera Express you can install CDH 6.0 or higher on up to 100 hosts.

새 호스트 현재 관리되는 호스트 (2)

이 호스트들은 소속된 클러스터가 없습니다. 일부를 선택하여 클러스터를 형성하십시오.

2

<input checked="" type="checkbox"/>	호스트 이름 ↑	IP 주소 ↕	랙 ↕	CDH 버전 ↕	코어 ↕
<input checked="" type="checkbox"/>	server01.hadoop.com	192.168.56.101	/default	없음	1
<input checked="" type="checkbox"/>	server02.hadoop.com	192.168.56.102	/default	없음	1

Displaying 1 - 2 of 2

뒤로

3 계속

방식 선택

☐ 패키지 사용 ?

☒ Parcel 사용 (권장됨) ?

추가 옵션

Proxy 설정

CDH 버전

이 Cloudera Manager 버전(6.3.1)보다 최신 버전인 CDH는 표시되지 않습니다.

☒ CDH-6.3.2-1.cdh6.3.2.p0.1605554

☐ CDH-5.16.2-1.cdh5.16.2.p0.8

1

추가 Parcel

☐ ACCUMULO-1.9.2-1.ACCUMULO6.1.0.p0.908695

☐ ACCUMULO-1.7.2-5.5.0.ACCUMULO5.5.0.p0.8

☒ 없음

☐ KAFKA-4.1.0-1.4.1.0.p0.4

☒ 없음

☐ SPARK-0.9.0-1.cdh4.6.0.p0.98

☒ 없음

☐ SQOOP\_NETEZZA\_CONNECTOR-1.5.1c6

☐ SQOOP\_NETEZZA\_CONNECTOR-1.5.1c5

☒ 없음

☐ SQOOP\_TERADATA\_CONNECTOR-1.7c5

☒ 없음

☐ mkl-2020.0.166

☒ 없음

2

3

뒤로

계속

그림 2.73 설치할 CDH 버전 및 추가 Parcel 선택

## JDK 설치 옵션

Oracle Binary Code License Agreement for the Java SE Platform Products and JavaFX

ORACLE AMERICA, INC. ("ORACLE"), FOR AND ON BEHALF OF ITSELF AND ITS SUBSIDIARIES AND AFFILIATES UNDER COMMON CONTROL, IS WILLING TO LICENSE THE SOFTWARE TO YOU ONLY UPON THE CONDITION THAT YOU ACCEPT ALL OF THE TERMS CONTAINED IN THIS BINARY CODE LICENSE AGREEMENT AND SUPPLEMENTAL LICENSE TERMS (COLLECTIVELY "AGREEMENT"). PLEASE READ THE AGREEMENT CAREFULLY. BY SELECTING THE "ACCEPT LICENSE AGREEMENT" (OR THE EQUIVALENT) BUTTON AND/OR BY USING THE SOFTWARE YOU ACKNOWLEDGE THAT YOU HAVE READ THE TERMS AND AGREE TO THEM. IF YOU ARE AGREEING TO THESE TERMS ON BEHALF OF A COMPANY OR OTHER LEGAL ENTITY, YOU REPRESENT THAT YOU HAVE THE LEGAL AUTHORITY TO BIND THE LEGAL ENTITY TO THESE TERMS. IF YOU DO NOT HAVE SUCH AUTHORITY, OR IF YOU DO NOT WISH TO BE BOUND BY THE

☐ Oracle Java SE Development Kit(JDK) 설치

Oracle 바이너리 코드 사용권 계약에 동의하고 JDK를 설치하려면 이 상자를 선택하십시오. 현재 설치된 JDK를 사용하려면 선택 취소 상태로 두십시오.

그림 2.74 JDK 사용권 계약 동의

※ 이과정은 Skip 됩니다.

WARNING: This Cloudera offering includes Oracle's Unlimited Strength Java(TM) Cryptography Extension (JCE) Policy Files for the Java(TM) Platform, Standard Edition (Java SE) Runtime Environment. Due to import restrictions of some countries, the version of the JCE Policy Files that are bundled in the Java Runtime Environment, or JRE(TM), allow "strong" but limited cryptography to be used. The Unlimited Strength JCE Policy Files included in this Cloudera offering, however, provides "unlimited strength" policy files which contain no restrictions on cryptographic strengths. Please note that some countries may legally prohibit the import of unlimited encryption strength policy files. You are responsible for determining whether you are subject to legal restrictions on cryptographic strength, and if so, you should not download

☐ Java Unlimited Strength 암호화 정책 파일 설치

1

지역법이 Unlimited Strength 암호화 배포를 허용하고 보안 클러스터를 실행하고 있을 경우 이 확인란을 선택하십시오.

뒤로

계속

2

그림 2.75 Java 암호화 정책 설치 동의

※ 이과정은 Skip 됩니다.

SSH 로그인 정보를 제공합니다.

Cloudera 패키지를 설치하려면 호스트에 대한 루트 액세스가 필요합니다. 이 설치 관리자는 SSH를 통해 호스트에 연결하고 루트로 직접 로그인하거나 암호 없이 `sudo/pbrun` 권한을 가진 다른 사용자로 로그인하여 루트가 됩니다.

모든 호스트를 다음으로 로그인: ☒ root  
☐ 다른 사용자

위에서 선택한 사용자에 대한 암호 또는 공용 키 인증을 통해 연결할 수 있습니다.

인증 방법: ☒ 모든 호스트가 동일한 암호 허용  
☐ 모든 호스트가 동일한 개인 키 허용

암호 입력:	<input type="text"/>
암호 확인:	<input type="text"/>

1

SSH 포트:

동시에 진행하는 설치 수:

(많은 설치를 한꺼번에 실행하면 많은 양의 네트워크 대역폭 및 다른 시스템 리소스가 소모됩니다.)

뒤로

계속

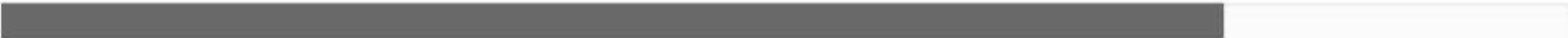
2

그림 2.76 클러스터 내 SSH 접속 정보 입력

※ 이과정은 Skip 됩니다.

# Install Agents

설치를 진행 중입니다.



3개 중 1개의 호스트가 완료되었습니다. 설치 중단

호스트 이름	IP 주소	진행률	상태	
server01.hadoop.com	192.168.56.101	<div></div>	✓ 설치가 완료되었습니다.	세부 정보
server02.hadoop.com	192.168.56.102	<div></div>	○ oracle-j2sdk1.8 패키지를 설치하는 중...	세부 정보
server03.hadoop.com	192.168.56.103	<div></div>	○ oracle-j2sdk1.8 패키지를 설치하는 중...	세부 정보

그림 2.77 CM 에이전트 및 기본 패키지 설치



# Install Parcels

선택한 Parcel을 다운로드하여 클러스터의 모든 호스트에 설치하는 중입니다.

1

▼ CDH 6.3.2-1.cdh6.3.2.p0.1605554	다운로드됨: 100%	배포됨: 3/3 (31.8 GiB/s)	압축 해제됨: 3/3	활성화됨: 3/3
<div><div></div><div></div><div></div></div>				

뒤로

계속

2

그림 2.78 CM Parcel 설치 및 완료

## Inspect Cluster

**i** You have created a new empty cluster. Cloudera recommends that you run the following inspections. For accurate measurements, Cloudera recommends that they are performed sequentially.

### ✔ Inspect Network Performance

#### > 고급 옵션

상태 ✔ 마지막 실행 a few seconds ago 소요 시간 21.86s

검사기 결과 표시 ↗

다시 실행

더 있음 ▼

### ! Inspect Hosts

Warning(s) were detected, review the inspector results to determine if any of the warnings need to be addressed.

상태 ✔ 마지막 실행 a few seconds ago 소요 시간 26.3s

검사기 결과 표시 ↗

다시 실행

더 있음 ▼

- ☐ Fix the issues and run the inspection tools again.
- ☐ Quit the wizard and Cloudera Manager will delete the temporarily created cluster.
- ☒ I understand the risks, let me continue with cluster setup. **1**

뒤로

계속

**2**

그림 2.79 Inspect Cluster 작업

Select Services

설치할 서비스 조합을 선택하십시오.

Essentials

Management and support for Cloudera's distribution including Hadoop.  
서비스: HDFS, YARN(MapReduce 2 포함), ZooKeeper, Oozie, Hive 및 Hue

Data Engineering

Process, develop, and serve predictive models.  
서비스: HDFS, YARN(MapReduce 2 포함), ZooKeeper, Oozie, Hive, Hue 및 Spark

Data Warehouse

The modern data warehouse for today, tomorrow, and beyond.  
서비스: HDFS, YARN(MapReduce 2 포함), ZooKeeper, Oozie, Hive, Hue 및 Impala

Operational Database

Real-time insights for modern data-driven business.  
서비스: HDFS, YARN(MapReduce 2 포함), ZooKeeper, Oozie, Hive, Hue 및 HBase

모든 서비스

Everything you need to become information-driven, with complete use of the Cloudera ecosystem.  
서비스: HDFS, YARN(MapReduce 2 포함), ZooKeeper, Oozie, Hive, Hue, HBase, Impala, Kudu, Solr, Spark, YARN, ZooKeeper

사용자 지정 서비스

보유한 서비스를 선택하십시오. 선택한 서비스에 필요한 서비스가 자동으로

서비스 유형	설명
<input type="checkbox"/> HBase	Apache HBase는 대규모 데이터 세트에 임의의 실시간 읽기/쓰기 액세스를 제공합니다(HDFS와 ZooKeeper 필요).
<input checked="" type="checkbox"/> HDFS	Apache HDFS(Hadoop Distributed File System)는 Hadoop 애플리케이션이 사용하는 기본 스토리지 시스템입니다. HDFS는 데이터 블록에 대한 여러 개의 복제본을 생성하고 이를 클러스터 전반에 걸쳐 컴퓨팅 호스트에 배포하여 안정적이고 매우 빠른 계산을 지원합니다.
<input type="checkbox"/> Hive	Hive는 SQL과 유사한 언어인 HiveQL을 제공하는 데이터 웨어하우스 시스템입니다.
<input type="checkbox"/> Hue	Hue는 CDH(Cloudera Distribution Including Apache Hadoop)에서 작동하는 GUI(그래픽 사용자 인터페이스)입니다(HDFS, MapReduce, Hive 필요).
<input type="checkbox"/> Impala	Impala에서는 HDFS 및 HBase에 저장된 데이터에 대해 실시간 SQL 쿼리 인터페이스를 제공합니다. Impala에는 Hive 서비스가 필요하며 Hue와 Hive Metastore를 공유합니다.
<input type="checkbox"/> Isilon	EMC Isilon is a distributed filesystem.
<input type="checkbox"/> Kafka	Apache Kafka is publish-subscribe messaging rethought as a distributed commit log.
<input type="checkbox"/> Key-Value Store Indexer	Key-Value Store Indexer는 HBase에 포함된 테이블 안의 데이터의 변경 사항을 수신 대기하고 Solr을 사용하여 인덱싱합니다.
<input type="checkbox"/> Kudu	Kudu is a true column store for the Hadoop ecosystem.
<input type="checkbox"/> Oozie	Oozie는 클러스터의 데이터 처리 작업을 관리하는 워크플로우 조정 서비스입니다.
<input type="checkbox"/> Solr	Solr은 HDFS에 저장된 데이터를 인덱싱 및 검색하는 배포 서비스입니다.
<input type="checkbox"/> Spark	Apache Spark is an open source cluster computing system. This service runs Spark as an application on YARN.
<input checked="" type="checkbox"/> YARN (MR2 Included)	YARN이라고도 하는 MRv2(Apache Hadoop MapReduce 2.0)는 MapReduce 애플리케이션을 지원하는 데이터 계산 프레임워크입니다(HDFS 필요).
<input checked="" type="checkbox"/> ZooKeeper	Apache ZooKeeper는 구성 데이터를 유지하고 동기화하는 중앙 집중식 서비스입니다.

저사양 파일럿 환경: DataNode 선택 시 Server02 하나만 선택한다.

#### HDFS

<b>NameNode</b> × 1 새로 만들기 server01.hadoop.com	<b>SecondaryNameNode</b> server01.hadoop.com	<b>Balancer</b> server01.hadoop.com	<b>HttpFS</b> 호스트 선택
<b>NFS Gateway</b> 호스트 선택	<b>DataNode</b> × 2 새로 만들기 server[02-03].hadoop.com ▼		

그림 2.81 CM을 이용한 소프트웨어 설치 – HDFS

- NameNode: Server01 선택
- SecondaryNameNode: Server01 선택
- Balancer: Server01 선택
- HttpFS: 미설치
- NFS Gateway: 미설치
- DataNode: Server02, Server03 선택 (※ 저사양 환경은 Server02만 선택)

저사양 파일럿 환경: Cloudera Management Service의 설치 위치를 모두 Server01로 선택한다.

#### C Cloudera Management Service

<b>C</b> Service Monitor × ... server03.hadoop.com ▼	<b>C</b> Activity Monitor 호스트 선택	<b>C</b> Host Monitor × 1 새로 만들기 server03.hadoop.com ▼	<b>C</b> Reports Manager × ... server03.hadoop.com ▼
<b>C</b> Event Server × 1 새로 만들기 server03.hadoop.com ▼	<b>C</b> Alert Publisher × 1 새로 만들기 server03.hadoop.com ▼	<b>C</b> Telemetry Publisher 호스트 선택	

그림 2.82 CM을 이용한 소프트웨어 설치 - Cloudera Management Service

- Service Monitor: Server03 선택
- Active Monitor: 미설치
- Host Monitor: Server03 선택
- Report Manager: Server03 선택
- Event Server: Server03 선택
- Alert Publisher: Server03 선택
- Telemetry Publisher: 미설치

(※ 저사양 환경은 모두 Server01로 선택)

## YARN (MR2 Included)

ResourceManager x ...

server01.hadoop.com

JobHistory Server x ...

server01.hadoop.com

NodeManager x 2 새로 만들기

DataNode(으)로 저장 ▼

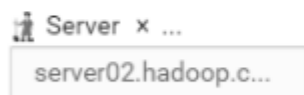
그림 2.83 CM을 이용한 소프트웨어 설치 - YARN(MR2 Included)

- ResourceManager: Server01 선택
- JobHistory Server: Server01 선택
- NodeManager: “DataNode로 저장” 선택

## 역할 할당 사용자 지정

여기에서 새 서비스에 대한 역할 할당을 사용자 지정할 수 있지만 단일 호스트에 너무 많은 수의 역할을 할당하는 등 올바르지 않게 할당할 경우, 성능이 저하될 수 있습니다.

역할 할당을 호스트별로 볼 수도 있습니다. [호스트별로 보기](#)




## 그림 2.84 CM을 이용한 소프트웨어 설치 - ZooKeeper

- Server: Server02 선택

## 데이터베이스 설정

데이터베이스 연결을 구성 및 테스트할 수 있습니다. 설치 가이드 [🔗](#)의 **Installing and Configuring an External Database** 섹션에 설명된 대로 데이터베이스를 먼저 생성하십시오.

- ☐ 사용자 지정 데이터베이스 사용    ☒ **내장된 데이터베이스 사용** **1**

 The embedded PostgreSQL database is not supported for use in production environments. 내장된 데이터베이스를 사용할 경우 암호가 자동으로 생성됩니다. 복사해 두십시오.

### Reports Manager

현재 **server03.hadoop.com**에서 실행하도록 할당되었습니다.

유형	호스트 이름 *	데이터베이스 이름 *	사용자 이름 *
PostgreSQL ▼	server01.hadoop.com:7432	rman	rman

암호 \*

ZyrkTlm57w

**2**

테스트 연결

그림 2.85 CM을 이용한 소프트웨어 설치 – 데이터베이스 설정

※ 이과정은 Skip 됩니다.



## 첫 번째 실행 명령

상태 완료됨 Mar 8, 9:55:28 PM 5.1m

Finished First Run of the following services successfully: ZooKeeper, HDFS, YARN (MR2 Included), Cloudera Management Service.

### ▼ 1/1단계가 완료되었습니다.

☒ Show All Steps ☐ Show Only Failed Steps ☐ Show Only Running Steps

▼  Run a set of services for the first time 4단계가 성공적으로 완료되었습니다.	Mar 8, 9:55:28 PM	5.1m
▼  6 단계 순차 실행 4단계가 성공적으로 완료되었습니다.	Mar 8, 9:55:28 PM	5.1m
▶  Ensuring that the expected software releases are installed on hosts.	Mar 8, 9:55:28 PM	4.83s
▶  4 단계 병렬 실행	Mar 8, 9:55:33 PM	36.71s
▶  2 단계 병렬 실행	Mar 8, 9:56:10 PM	94.49s
▶  2 단계 병렬 실행	Mar 8, 9:57:44 PM	2.5m
▶  YARN (MR2 Included) 시작  YARN (MR2 Included)	Mar 8, 10:00:14 PM	9.15s
▶  Verifying successful startup of services	Mar 8, 10:00:23 PM	12.66s

뒤로

계속

그림 2.86 CM을 이용한 소프트웨어 설치 – 설치 실행

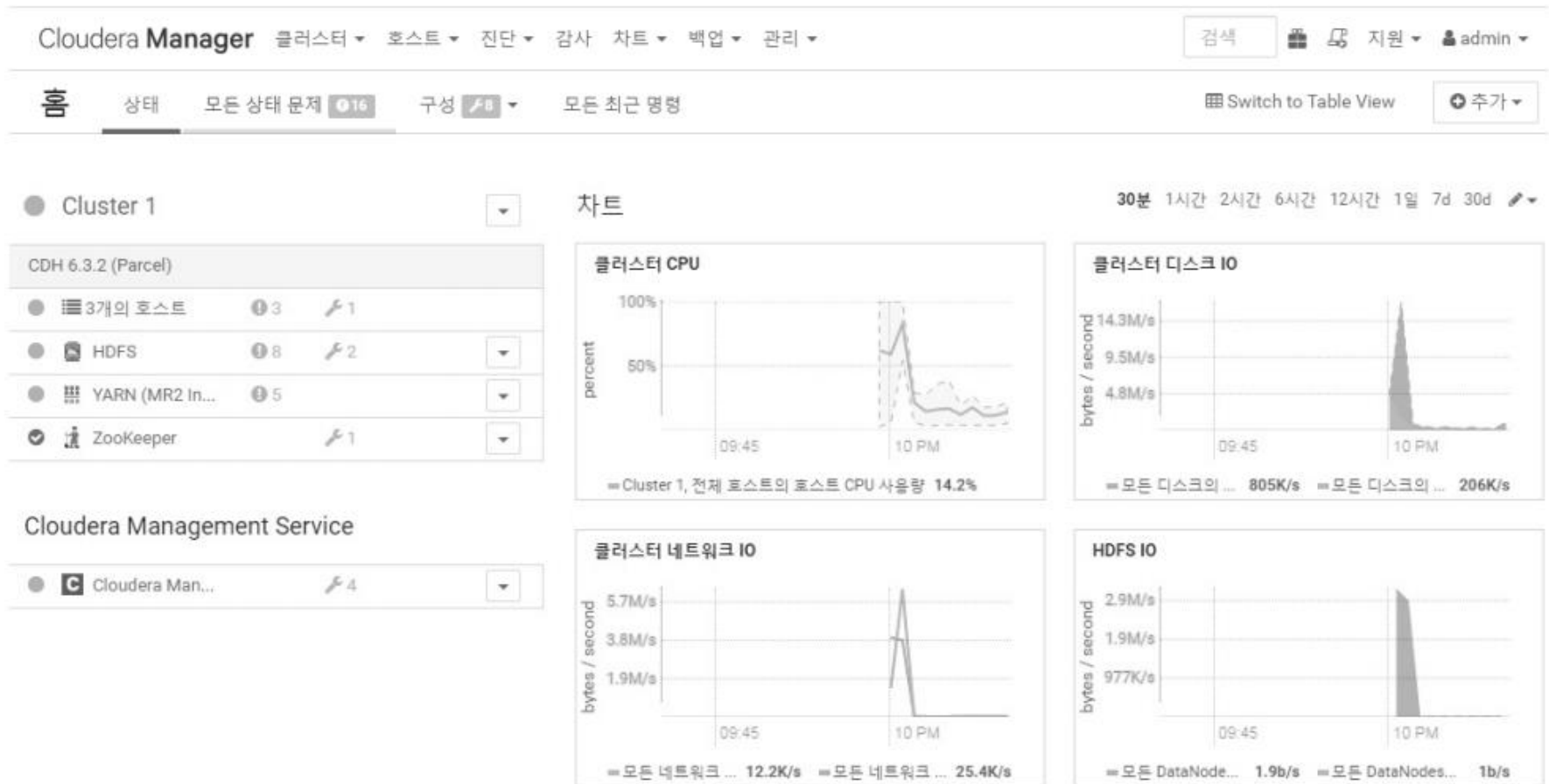


그림 2.88 CM을 이용한 소프트웨어 설치 – Cluster1의 홈 화면

## 2.5 빅데이터 클러스터 구성

### 기본 소프트웨어 설치

하둡, 주키퍼 등 기본구성  
(실습)

## Tip \_ CM 리소스 모니터링

---

CM에서는 각 서버의 리소스(CPU, 메모리, 디스크, I/O 등)와 설치된 소프트웨어(하둡, 주키퍼 등)를 모니터링하면서 현재 상태값을 보여준다. 색깔에 따라 양호(초록), 주의(노랑), 불량(빨강)으로 분류되며, CM의 에이전트(Agent)가 지속적으로 체크하면서 상태값을 업데이트한다. 불량으로 표시되도 해당 서버 또는 소프트웨어가 정지 상태가 아니라면 파일럿 프로젝트를 진행하는 데 문제는 없다.

추가적으로 개발 PC를 리부팅했거나, 오라클 버추얼 박스를 종료해서 CM이 강제 종료될 경우 클러스터(Cluster)와 클라우데라 관리 서비스(Cloudera Management Service)가 불안정한 상태로 시작되어 모니터링 상태도 알 수 없음으로 표시된다. 이럴 땐 CM 홈에서 Cluster1 우측의 콤보박스를 선택해서 중지시킨 다음 Cluster1을 다시 시작한다. 클라우데라 관리 서비스 또한 같은 방법으로 재시작한다. 다소 불편한 작업이지만 365일 가동될 수 없는 개인의 파일럿 환경임을 감안하자.

**저사양 파일럿 환경:** Cloudera Management Service 기능을 모두 정지한다.

- 앞으로 프로젝트를 진행하면서 저사양 PC 환경에서는 리소스 부족 현상이 자주 발생한다. 원활한 파일럿 프로젝트 진행을 위해 Cloudera Management Service의 모니터링 기능은 그림 2.88까지만 확인하고 정지시킨다. 고사양 PC 환경에서도 리소스 부족 현상이 발생하면 Cloudera Manager Service를 정지한다.
- HBase 서비스: Cloudera Management 서비스: CM의 홈 → [Cloudera Management Service] → [정지]