# LESSON 1

## 하둡 에코시스템

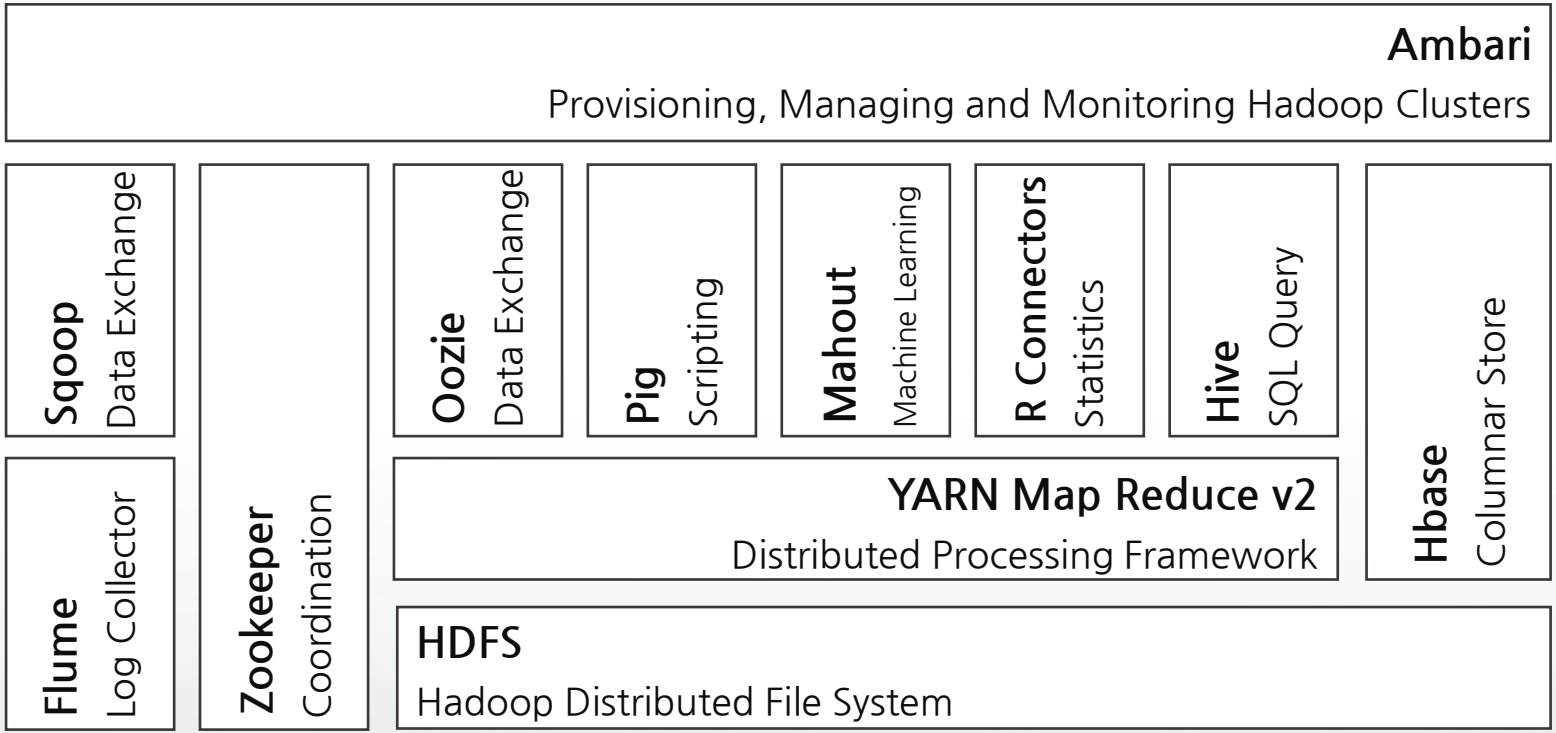
# 하둡 에코시스템

## 하둡 에코(eco)시스템은

**하둡에서 제공하지 않는 기능을 보완/개선하는 패키지들을 모두 말하는 것**

- 하둡을 SQL과 같은 쿼리언어로 사용하게 해주는 하이브(Hive),
- 하둡의 입력을 파일에서 RDBMS를 사용하게 해주는 스쿱(SQOOP),
- 하둡의 입력을 웹서버 로그파일로 해주는 Flume(***),
- 하둡의 분산파일 시스템인 HDFS를 NoSQL 저장소로 만들어 주는 HBase 등이 있다
- R은?

# 하둡 에코시스템


Apache Hadoop Ecosystem
Ambari
Provisioning, Managing and Monitoring Hadoop Clusters
Sqoop
Data Exchange
Flume
Log Collector
Zookeeper
Coordination
Oozie
Data Exchange
Pig
Scripting
Mahout
Machine Learning
R Connectors
Statistics
Hive
SQL Query
Hbase
Columnar Store
YARN Map Reduce v2
Distributed Processing Framework
HDFS
Hadoop Distributed File System

# LESSON 2

# Apache Sqoop

# Apache Sqoop

- RDBMS데이터를 하둡에서  접근할 수 있게 만들어줌


- 1.4대에서 2.0대로 변경되면서
  - 클라이언트/서버 형태로 변경
  - 실습에서는 1.4사용

# Apache Sqoop

## Sqoop 설치

- 다운로드(하둡 버전 확인 필요)

  $ wget http://apache.tt.co.kr/sqoop/1.4.6/sqoop-1.4.6.bin__hadoop-2.0.4-alpha.tar.gz

- 압축해제

  $ tar xvfz sqoop-1.4.6.bin__hadoop-2.0.4-alpha.tar.gz

## Sqoop 요구사항

- MySQL이 설치되어 있어야 함

MySQL설치 동영상

# Apache Sqoop

## Sqoop 세팅

환경변수 설정

```
export SQOOP_HOME=/sqoop-1.4.6.bin__hadoop-2.0.4-alpha

export PATH=$PATH:$SQOOP_HOME/bin

export CLASSPATH=$CLASSPATH:$SQOOP_HOME/lib/*
```

MySQL JDBC 드라이버 설치

```
$ apt-get install libmysql-java
```

드라이버,하둡파일 복사/

```
$ cp /usr/share/java/mysql-connector-java.jar $SQOOP_HOME/lib
$ cp -r $HADOOP_HOME/share/hadoop/mapreduce/* $SQOOP_HOME/lib
```

# Apache Sqoop

## Sqoop 세팅

스쿱 설정파일 수정

```
$ cp conf/sqoop-env-template.sh conf/sqoop-env.sh
```

파일 내용 중 값 지정

```
HADOOP_COMMON_HOME=/hadoop-2.7.2(하둡 홈디렉토리)
HADOOP_MAPRED_HOME=/hadoop-2.7.2/share/hadoop/mapreduce
```

동작확인(MySQL 테스트 DB world가 있어야 함)

```
$ sqoop import --connect jdbc:mysql://mysql_db/sample?useSSL=false ₩
  --username root --table city -P
```

스쿱 실습 동영상

# LESSON 3

# Apache Flume

# Apache Flume

- http://flume.apache.org
- 웹 서버 로그파일을 하둡의 HDFS로 업로드 해줌

# Apache Flume

## Apache 웹서버 설치

### Flume 설치

다운로드

$ wget http://apache.mirror.cdnetworks.com/flume/1.6.0/apache-flume-1.6.0-bin.tar.gz

환경변수

export FLUME_HOME=/apache-flume-1.6.0-bin

export FLUME_CONF_DIR=$FLUME_HOME/conf

export CLASSPATH=$JAVA_HOME/lib/*:$SQOOP_HOME/lib/*:.

export FLUME_CLASSPATH=$FLUME_CONF_DIR

export PATH=$PATH:$FLUME_HOME/bin

# Apache Flume

## Apache 웹서버 설치

### Flume 설치

설정파일

$FLUME_HOME/conf/flume-env.sh

$ cp conf/flume-env.sh.template flume-env.sh

Flume-env.sh 수정

$JAVA_OPTS="-Xms500m -Xmx1000m -Dcom.sun.management.jmxremote"

export JAVA_HOME=/usr/lib/jvm/java-7-openjdk-amd64(JAVA_HOME 설정)

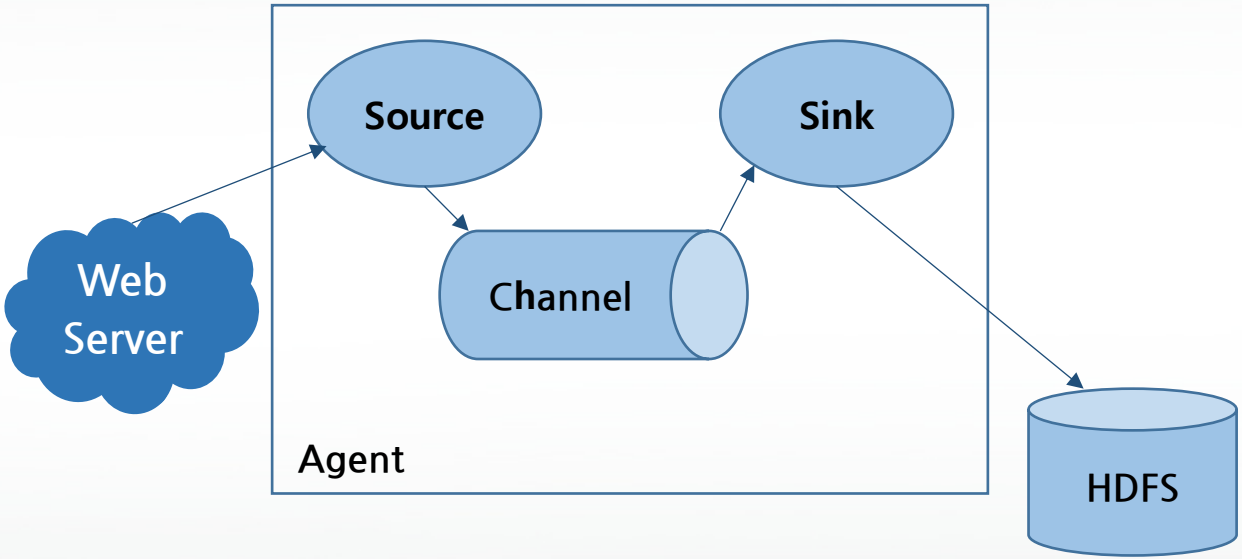
Flume 수행

$ flume-ng --help

$ flume-ng agent -conf-file ~/flume.conf --name agent

# Apache Flume

## Apache 웹서버 설치



# Apache Flume

## flume.conf(아파치 웹서버로그->하둡 hdfs)

```
agent.sources = seqGenSrc
agent.channels = memoryChannel
agent.sinks = hdfsSink

# For each one of the sources, the type is defined
agent.sources.seqGenSrc.type = exec
agent.sources.seqGenSrc.command = tail -F /var/log/apache2/access.log

# The channel can be defined as follows.
agent.sources.seqGenSrc.channels = memoryChannel

# Each sink's type must be defined
agent.sinks.hdfsSink.type = hdfs
agent.sinks.hdfsSink.hdfs.path = hdfs://localhost:9000/flume/data
agent.sinks.hdfsSink.rollInterval = 30
agent.sinks.hdfsSink.sink.batchSize = 100

#Specify the channel the sink should use
agent.sinks.hdfsSink.channel = memoryChannel

# Each channel's type is defined.
agent.channels.memoryChannel.type = memory

# Other config values specific to each type of channel(sink or source)
# can be defined as well
# In this case, it specifies the capacity of the memory channel
agent.channels.memoryChannel.capacity = 100000
agent.channels.memoryChannel.transactionCapacity = 10000
```

# Apache Flume

## flume.conf(아파치 웹서버로그->하둡 hdfs)

```
$ flume-ng agent -conf-file ~/flume.conf --name agent &
```

```
$ hadoop fs -mkdir /flume
```

```
$ hadoop fs -mkdir /data
```

FLUME 실습 동영상