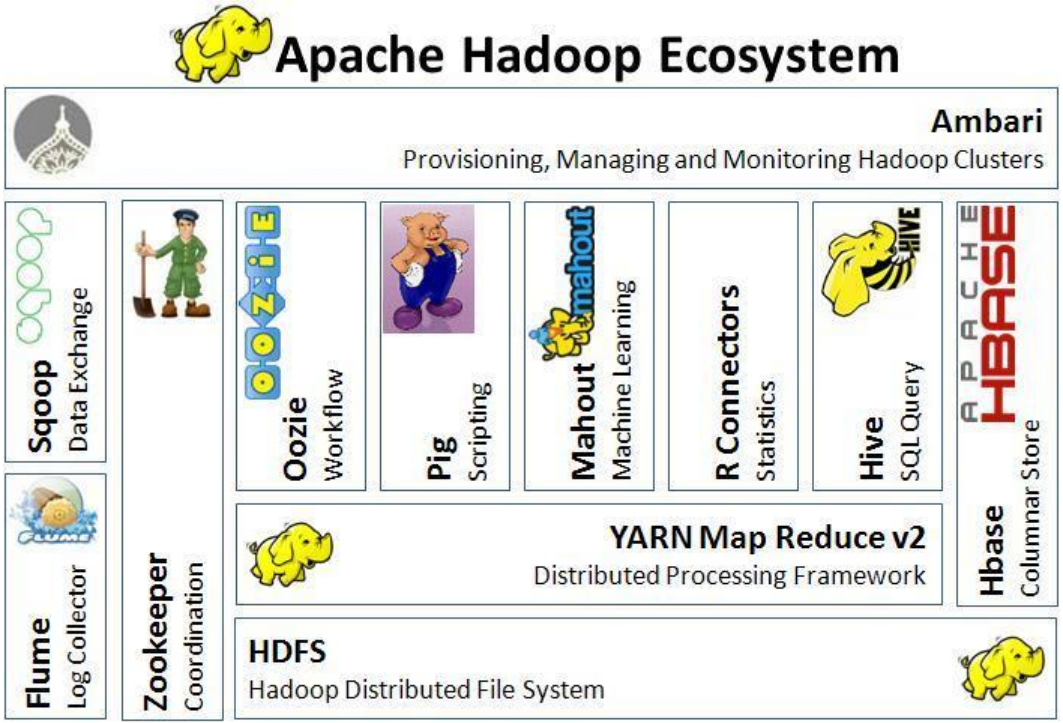
빅데이터 마스터과정
(DAM)

하석재
CEO, 2HCUBE
sjha72@gmail.com

하둡 에코시스템


Apache Hadoop Ecosystem
Ambari
Provisioning, Managing and Monitoring Hadoop Clusters
Sqoop
Data Exchange
Flume
Log Collector
Zookeeper
Coordination
Oozie
Workflow
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

Apache Sqoop

- RDBMS데이터를 하둡에서 접근할 수 있게 만들어 줌
  - RDBMS->HDFS -> RDBMS
- 1.4대에서 2.0대로 변경되면서
  - 클라이언트/서버 형태로 변경
  - 실습에서는 1.4사용

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

스쿱 요구사항

- MySQL이 설치되어 있어야 함

스쿱 세팅

환경변수 설정

export SQOOP_HOME=**/home/vagrant**/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

스쿱 세팅

스쿱 설정파일 수정

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

파일 내용 중 값 지정

HADOOP_COMMON_HOME=/hadoop-2.7.3(하둡 홈디렉토리)

HADOOP_MAPRED_HOME=/hadoop-2.7.3/share/hadoop/mapreduce

동작확인(MySQL 테스트 DB world가 있어야 함)

```
$ sqoop import --connect jdbc:mysql://192.168.0.123/world?useSSL=false \
  --username root --table city -P
```

Apache Flume

- http://flume.apache.org
- 웹서버 로그파일을 하둡의 HDFS로 업로드해 줌

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=/home/vagrant/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

Flume 설치

- 설정파일

  $FLUME_HOME/conf/flume-env.sh

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

  Flume-env.sh 수정

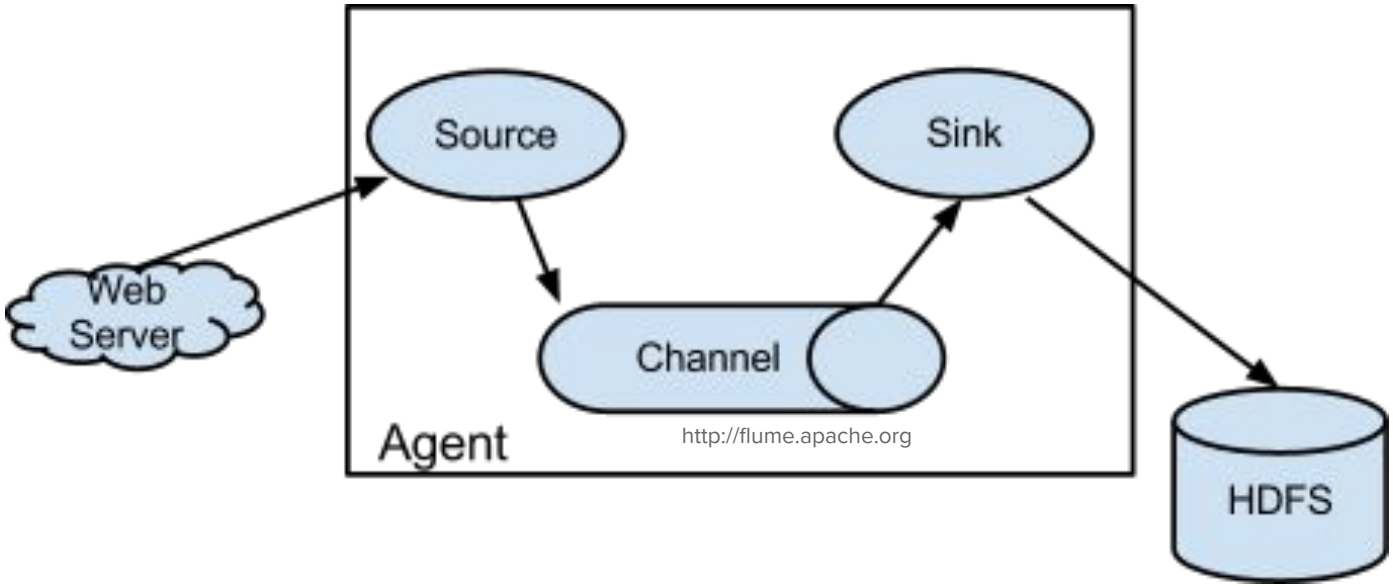
  $JAVA_OPTS="-Xms500m -Xmx1000m -Dcom.sun.management.jmxremote"
  export JAVA_HOME=/usr/lib/jvm/java-7-openjdk-amd64(JAVA_HOME 설정)

  Flume 수행

  $ flume-ng --help

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

Flume 아키텍처


Source
Sink
Web Server
Channel
Agent
http://flume.apache.org
HDFS

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
```

flume 실행

```
$ hdfs dfs -mkdir /flume
$ hdfs dfs -mkdir /flume/data

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