# 大数据技术与应用竞赛培训

## Flume

1. 简介

一个数据传输配置流程，需要指定源（source）、管道（channel）、接收器（sink）。

Agent ： （Source，Channel，Sink） 的一个配置文件。

启动Agent，即可按照配置的流程来传输数据。

1. 配置

|  |
| --- |
| # 单节点 Flume Agent 配置  # a1代表Agent的名字，并为其source，sink，channel配置名字  a1.sources = r1  a1.sinks = k1  a1.channels = c1  # 配置 source  a1.sources.r1.type = netcat  a1.sources.r1.bind = localhost  a1.sources.r1.port = 44444  # 配置 sink  a1.sinks.k1.type = logger  # 配置管道-使用 memory管道  a1.channels.c1.type = memory  a1.channels.c1.capacity = 1000  a1.channels.c1.transactionCapacity = 100  # 绑定source 和 sink 到 channel上  a1.sources.r1.channels = c1  a1.sinks.k1.channel = c1 |

1. 实例

实例1： 从一个网络端口传送数据；

使用上面的配置文件：

|  |
| --- |
| root@slave1:/usr/local/apache-flume-1.6.0-bin# bin/flume-ng agent --conf conf --conf-file /opt/flume/conf/demo.conf --name a1 -Dflume.root.logger=INFO,console |
| root@master:~# telnet slave1 44444 |

实例2：本地文件上传到HDFS；

Source查看其（目录输入源）Document：<https://flume.apache.org/releases/content/1.6.0/FlumeUserGuide.html#spooling-directory-source> 。

Sink 查看其（HDFS输出源）Document：<https://flume.apache.org/releases/content/1.6.0/FlumeUserGuide.html#hdfs-sink> 。

Channel查看其（可以设置为Memory，或File）Document：<https://flume.apache.org/releases/content/1.6.0/FlumeUserGuide.html#memory-channel> 、 <https://flume.apache.org/releases/content/1.6.0/FlumeUserGuide.html#file-channel> 。

mem.conf

|  |
| --- |
|  |

！数据传输异常？

## Hadoop

1. 命令（非官网自带命令）

|  |
| --- |
| root@master:~# cluster  Usage:cluster hdfs|yarn|spark|hive start|stop|status |

1. 简介

Hadoop是一个分布式架构：HDFS、YARN、MapReduce。

HDFS：分布式文件系统；

YARN：分布式集群资源管理框架；

MapReduce：分布式编程模型；

1. 基本命令

启动/关闭（$HADOOP\_HOME/sbin）

hdfs dfs （上传、下载、删除、查看目录/文件内容）

yarn jar

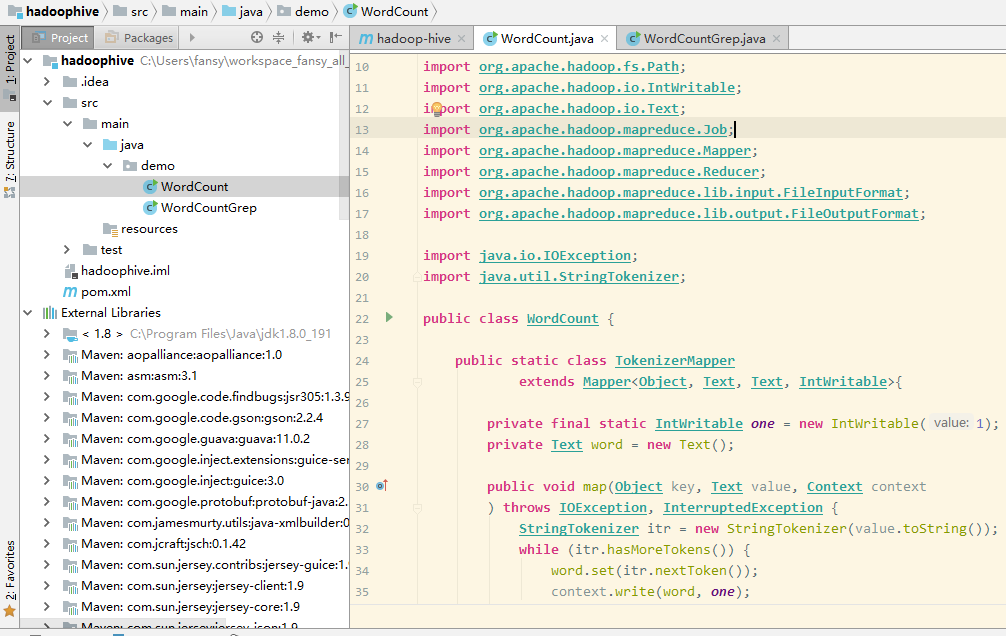
演示（分清三个文件系统）：

1. 文件上传；

b． 单词计数；

1. IDEA环境

下载示例工程（<https://github.com/fansy1990/hadoophive> ）即可配置好环境，如下：



1. MapReduce原理及编程

Map ： <K1, V1> -> <K2, V2>

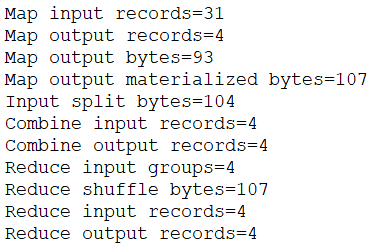
Shuffle : <K2, <V21,V22,…V 2n>>

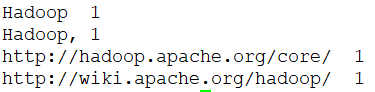
Reduce: <K2, <V21,V22,…V 2n>> -> <K3,V3>

文本文件的键值对<K1,V1>;

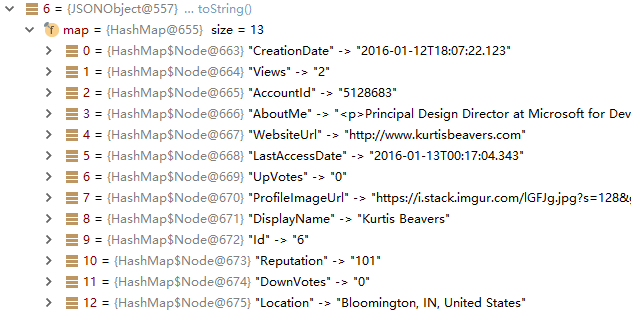
实例1：单词计数+过滤字符串

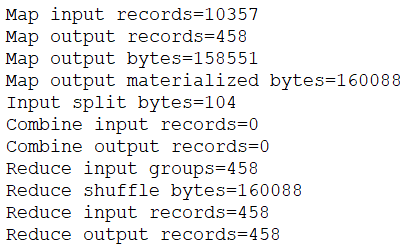
|  |
| --- |
| yarn jar hadoop-hive-1.0-SNAPSHOT.jar demo.WordCountGrep /user/root/README.txt ad /tmp/wc\_00 |





实例2： 读取json文件，并过滤其views 小于2，Reputation 小于20的数据。





## Hive

1. 启动

|  |
| --- |
| cd $HIVE\_HOME  bin/hiveserver2 &  bin/hive --service metastore & |

状态查看：jps -lm

1. 简介

大数据仓库：元数据+HDFS数据。

数据存储：HDFS；

查询引擎：MapReduce/Spark

|  |
| --- |
| root@master:~# beeline  Beeline version 1.2.1.spark2 by Apache Hive  beeline> !connect jdbc:hive2://master:10000  Connecting to jdbc:hive2://master:10000  Enter username for jdbc:hive2://master:10000: root  Enter password for jdbc:hive2://master:10000: \*\*\*\*  19/03/12 10:05:20 INFO Utils: Supplied authorities: master:10000  19/03/12 10:05:20 INFO Utils: Resolved authority: master:10000  19/03/12 10:05:20 INFO HiveConnection: Will try to open client transport with JDBC Uri: jdbc:hive2://master:10000  Connected to: Apache Hive (version 1.2.2)  Driver: Hive JDBC (version 1.2.1.spark2)  Transaction isolation: TRANSACTION\_REPEATABLE\_READ  0: jdbc:hive2://master:10000> |

1. DML、DDL

DML: <https://cwiki.apache.org/confluence/display/Hive/LanguageManual+DML>

DDL : <https://cwiki.apache.org/confluence/display/Hive/LanguageManual+DDL>

* Show/Create/Drop/Alter/Use Database
* Show/Create/Drop Table
* LOAD DATA [LOCAL] INPATH 'filepath' [OVERWRITE] INTO TABLE tablename [PARTITION (partcol1=val1, partcol2=val2 ...)]

1. HQL及实例
2. 根据Users.json 数据结构建表；
3. 使用User.json预处理后数据导入建好的表中；
4. 执行查询：总记录数、最大reputation及其对应的记录、reputation分组查询；

！最大reputation对应的记录数？