大数据聚集框架部署文档

[**一、 引言 1**](#_Toc24645)

[1.1 编写目的](#_Toc10302)

[**二、 概述 1**](#_Toc22793)

[2.1 框架简介](#_Toc10302)

[**三、 术语 1**](#_Toc22793)

[3.1 名词解释](#_Toc10302)

[**四、 模块安装 3**](#_Toc21676)

[4.1 安装条件](#_Toc10302)

[4.2 安装步骤](#_Toc14304)

[**五、 其它软件的安装 3**](#_Toc21676)

[5.1 spark2的部署](#_Toc10302)

大数据聚集框架部署

http://bbs.zwlbs.com/forum.php?mod=viewthread&tid=5943

(出处: F3平台论坛-中位科技)

安装顺序：1、**yarn+MapReduce 2.**  **spark的部署 3.**  gather的部署

# 引言

**1.1 编写目的**

本文档是聚集框架部署文档，其提供了分布式安装的方法和步骤。本文档适用于聚集框架0.2.0版本。

# 概述

**2.1 框架简介**

聚集框架每天凌晨分布式的跑报表任务，并将聚集结果回写到hbase。

# 术语

**3.1 名词解释**

1 HADOOP： Apache开源的分布式框架

2 HDFS: Hadoop的分布式文件系统

3 YARN: Hadoop的资源管理系统

4 SPARK： Apache开源的分布式计算引擎

5 HBASE： Apache开源的分布式数据仓库

5 PHOENIX： 为Hbase提供sql功能的框架

# 模块安装

**4.1.1 安装条件**

**4.1.1 安装条件**

|  |  |  |
| --- | --- | --- |
| 序号 | 名称 | 版本 |
| 1 | Java | 1.8 |
| 2 | Scala | 2.11.8 |
| 3 | Spark | 2.0.0 |
| 4 | Hbase | 1.1.2 |
| 5 | Zookeeper | 3.4.6 |
| 6 | Hadoop | 2.7.3 |
| 7 | Phoenix | 4.8.1-hbase-1.1 |

Spark的部署详见：**5.1 spark2的部署**

**4.1.2 安装步骤（一台）.**

**集群需先安装yarn+MapReduce**

#切换到聚集安装目录，如果不存在则新建目录。

ls /cluster。若报无此目录，则新建目录，mkdir /cluster；

cd /cluster

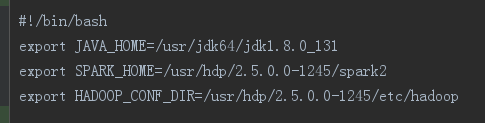
#将聚集框架安装包解压到/cluster目录下。

tar -zvxf gather-0.1.0.tar.gz –C /cluster

#修改环境变量配置文件

vim /cluster/gather-0.2.0/conf/gather-env.sh

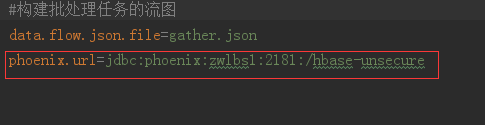
分别修改JAVA\_HOME, HADOOP\_HOME,SPARK\_HOME的环境变量



#修改gather.properties配置文件

vim /cluster/gather-0.2.0/conf/gather.properties

增加phoenix.url配置参数

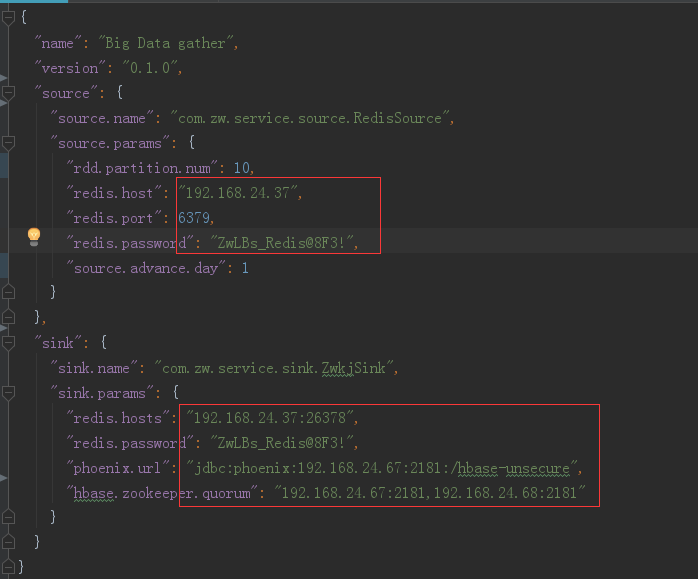


#修改聚集框架配置文件

vim /cluster/gather-0.2.0/conf/gather.json

修改source.params的redis.host; redis.port; redis.password参数

修改sink.params的redis.hosts; redis.password; phoenix.url; hbase.zookeeper.quorum参数



#将建表，聚集任务加到定时

配置修改

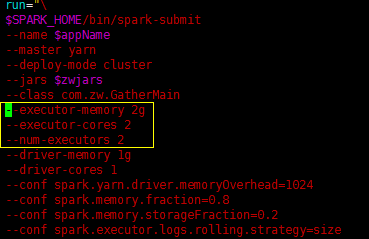
vim /cluster/gather-0.2.0/conf/gather.json



vim /cluster/gather-0.2.0/bin/start.sh

memory代表占用内存

num-executors代表核心，executor代表2个 及占用4核。上图中的rdd.partition.num代表4核。



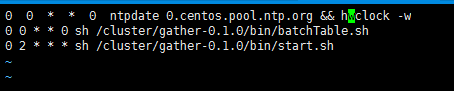
crontab -e

建表语句每周运行一次，建本月和下月的表（如果不存在）；

聚集框架凌晨2点运行：

0 0 \* \* 0 sh /cluster/gather-0.2.0/bin/batchTable.sh

0 2 \* \* \* sh /cluster/gather-0.2.0/bin/start.sh



#初始化表，不存在则建

sh /cluster/gather-0.2.0/bin/batchTable.sh

#安装完毕

# 其它软件的安装

**5.1 spark的部署（两台）**

#切换到spark安装目录。

cd /opt

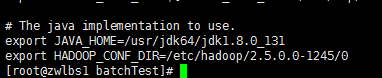
#将spark安装包解压到/opt目录下。

unzip spark.zip

#修改spark运行配置。

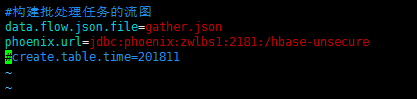
vi /opt/spark2/conf/spark-env.sh

在最后面增加export HADOOP\_CONF\_DIR=/etc/hadoop/2.5.0.0-1245/0（其为实际Hadoop conf目录）



单独创建上个月的表：

vim gather.properties



将最后一条注释取消，修改月份后

sh /cluster/gather-0.2.0/bin/batchTable.sh