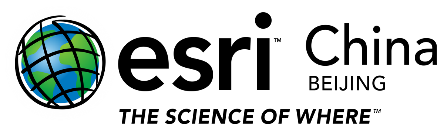
Esri大数据平台搭建指南之HDFS



易智瑞中国信息技术有限公司

2017 年 3 月

**——制定及修订记录——**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **版本** | **完成日期** | **编写/修订纪要** | **编写者** | **备注** |
| **V1.0** | **2017-3-8** | **初稿** | **胡源** |  |
| **V1.1** | **2017-3-20** | **修订** | **胡源** | **内容完善** |
| **V1.2** | **2017-3-31** | **修订** | **胡源** | **截图、链接更新** |
|  |  |  |  |  |

目录

[Esri大数据平台搭建指南之HDFS 1](#_Toc479169251)

[1. HDFS简介 4](#_Toc479169252)

[2. 搭建环境概览 4](#_Toc479169253)

[2.1硬件环境 4](#_Toc479169254)

[2.2操作系统 4](#_Toc479169255)

[2.3软件介质 5](#_Toc479169256)

[2.4测试数据 5](#_Toc479169257)

[2.5可选软件 5](#_Toc479169258)

[3. HDFS搭建 5](#_Toc479169259)

[3.1基础环境准备 5](#_Toc479169260)

[3.2配置SSH免密登陆 8](#_Toc479169261)

[3.3启动hadoop集群 11](#_Toc479169262)

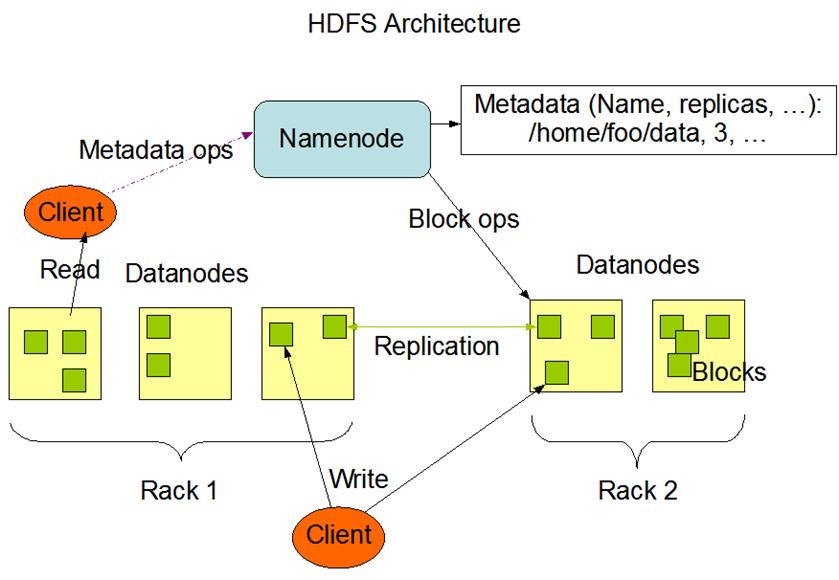
[4. 注册数据 12](#_Toc479169263)

[5. 执行分析 15](#_Toc479169264)

1. **HDFS简介**

HDFS（Hadoop Distributed File System），是一个分布式文件系统。它有一定高度的容错性，而且提供了高吞吐量的数据访问，非常适合大规模数据集上的应用。HDFS 提供了一个高度容错性和高吞吐量的海量数据存储解决方案。

HDFS架构图如下：



1. **搭建环境概览**

2.1硬件环境

四台centos服务器：

CPU：4核/结点

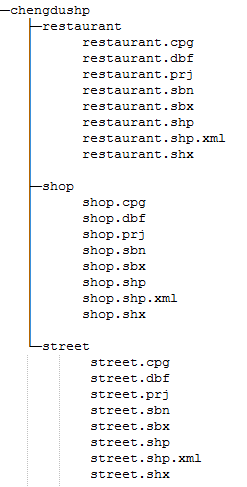
内存：4G /结点

磁盘：40G/结点

2.2操作系统

CentOS版本: [CentOS-6.8-x86\_64-minimal](http://centos.ustc.edu.cn/centos/6.8/isos/x86_64/CentOS-6.8-x86_64-minimal.iso)（点击下载）

***提示****：其他Linux版本也可参考此配置教程，但命令会有些差别。*

2.3软件介质

JDK：jdk-7u25-linux-x64.tar.gz

Hadoop：hadoop-2.7.3.tar.gz

（下载地址：<http://pan.baidu.com/s/1jHCTZca>）

2.4测试数据

成都市路网和餐饮点分布矢量数据，数据目录结构如右图所示。

2.5可选软件

Xshell、Xftp

***备注****：Xshell可用于在windows本机和hadoop集群中的机器之间执行远程命令，Xftp可用于在windows本机和hadoop集群中的机器之间传输文件。*

1. **HDFS搭建**

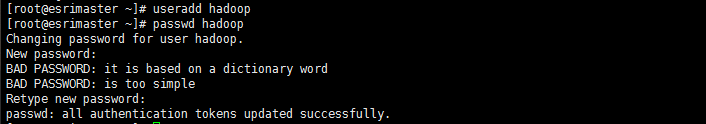
3.1基础环境准备

***备注****：四台机器该准备步骤完全一致，故这里以esrimaster机器为例。（如在虚拟机环境下，可克隆虚拟机并修改克隆机的MAC地址和IP、主机名以减少工作量）。*

1. 安装centos并启动。
2. 添加并配置hadoop账户。

useradd hadoop

passwd hadoop



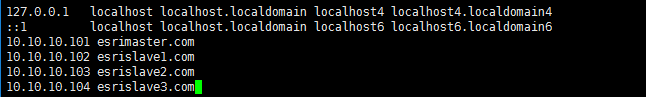
1. 关闭防火墙并设置为开机不自动启动。

service iptables stop

chkconfig iptables off

1. 配置IP与主机名的映射。

vi /etc/hosts



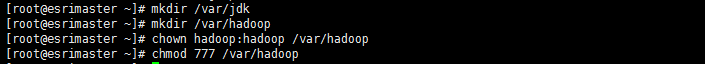
1. 创建jdk、hadoop安装路径并赋予权限。

mkdir /var/jdk

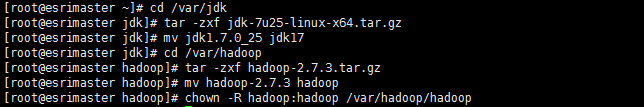
mkdir /var/hadoop

chown hadoop:hadoop /var/hadoop

chmod 777 /var/hadoop



1. 通过xftp软件向虚拟机传输hadoop、jdk包到上一步指定目录，解压jdk、hadoop并重命名为jdk17、hadoop，并更改hadoop安装目录的所有权。



1. 配置java和hadoop环境变量并使其生效。

vi /etc/profile



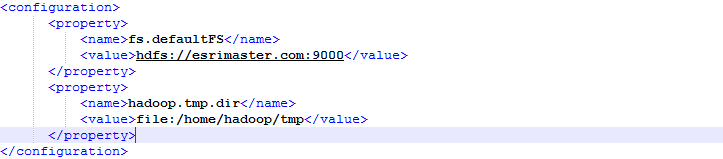
source /etc/profile

1. 测试java安装。

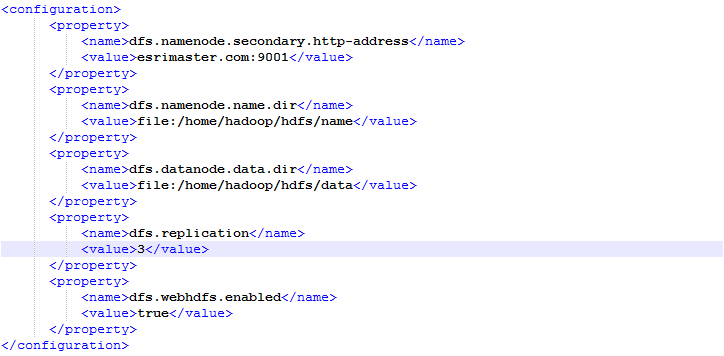


1. 修改$HADOOP\_HOME/etc/hadoop中配置文件。

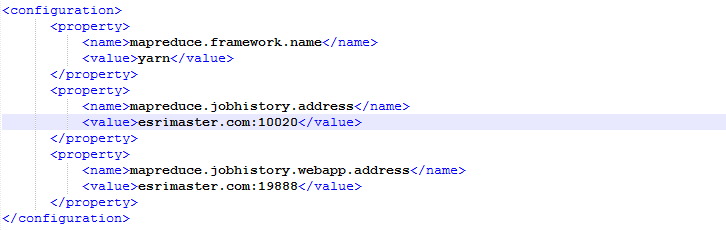
* core-site.xml



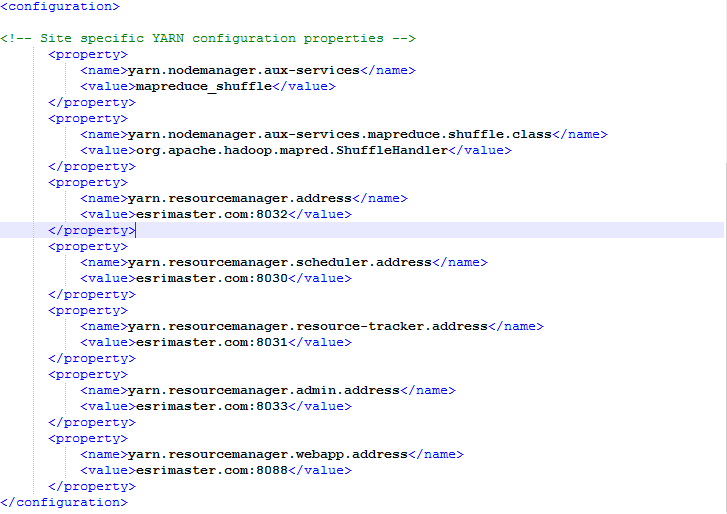
* hdfs-site.xml



* mapred-site.xml



* yarn-site.xml



* slaves



* hadoop-env.sh



* yarn-env.sh



3.2配置SSH免密登陆

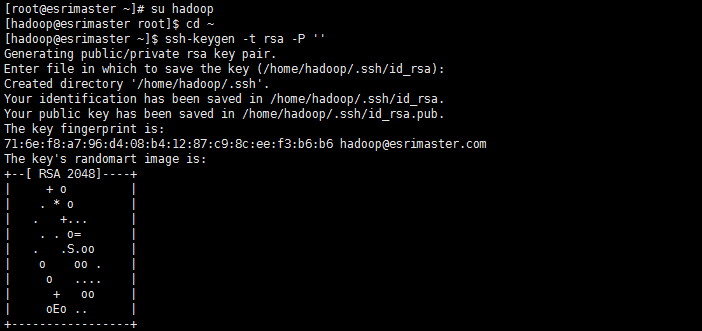
***备注****：以esrimaster机器为例部署，其他机器部署方式与此相同。*

1. 切换到hadoop用户并创建密钥。

su hadoop

cd ~

ssh-keygen -t rsa -P ‘’

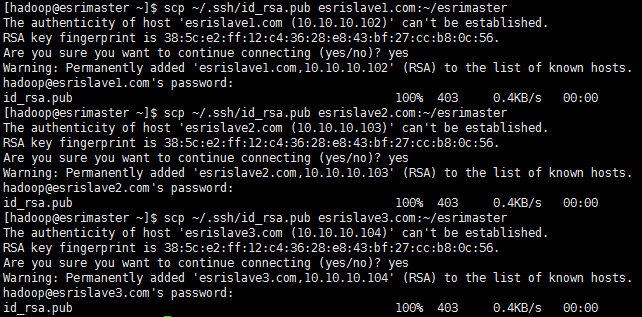


1. 创建文件夹来存储其他机器的公钥。

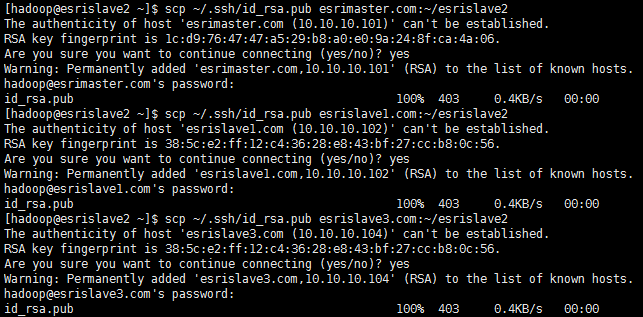


1. 在其他机器上重复2）步骤， 分别在每台机器上创建3个文件夹（除本机外的其他三台机器数量为需要创建的文件夹数）。
2. 将每台机器创建的公钥发送到其他机器创建的对应文件夹，不同机器在发送的时候又有不同，下面以机器esrimaster与esrislave2为例说明。

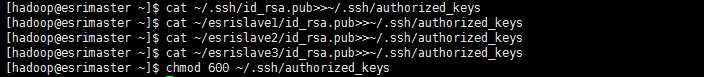
esrimaster机器：



esrislave2机器：

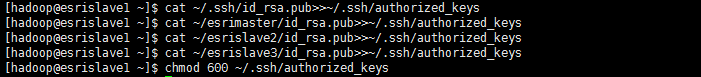


1. 将所有公钥写入authorized\_keys文件中并修改该文件的权限。



1. 在其他机器上重复5）步骤。

例如，在esrislave1机器上为：



1. 每台机器上，切换到root用户，修改/etc/ssh/sshd\_config文件并重启SSH服务。

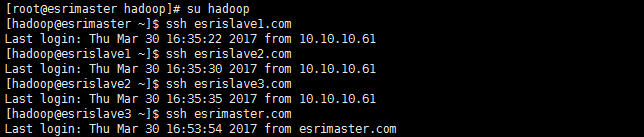
su

vi /etc/ssh/sshd\_config



service sshd restart

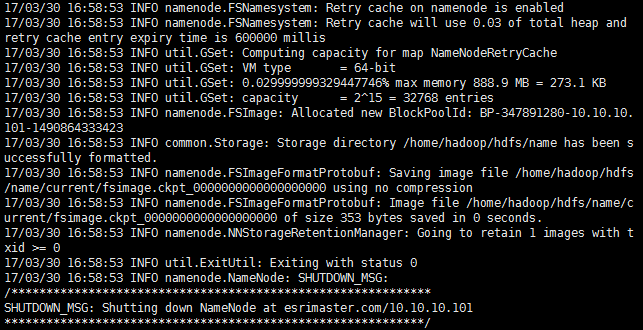
1. 切换回hadoop用户，在每一台机器上测试各台机器之间的免密登陆，无需输入密码即配置成功。



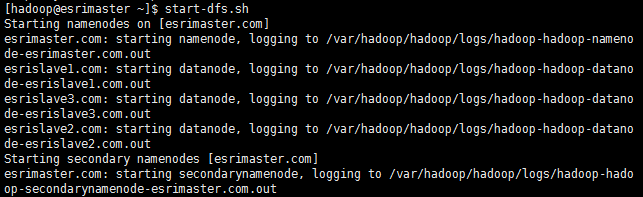
3.3启动hadoop集群

1. 格式化HDFS。

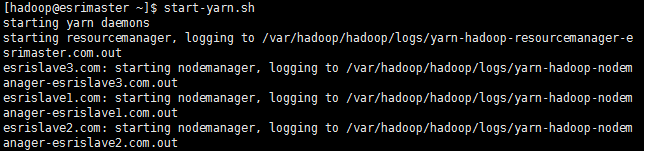
hdfs namenode -format



1. 启动HDFS。

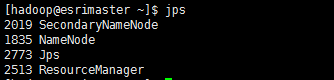


1. 启动Yarn。



1. 查看hadoop集群各机器的进程。

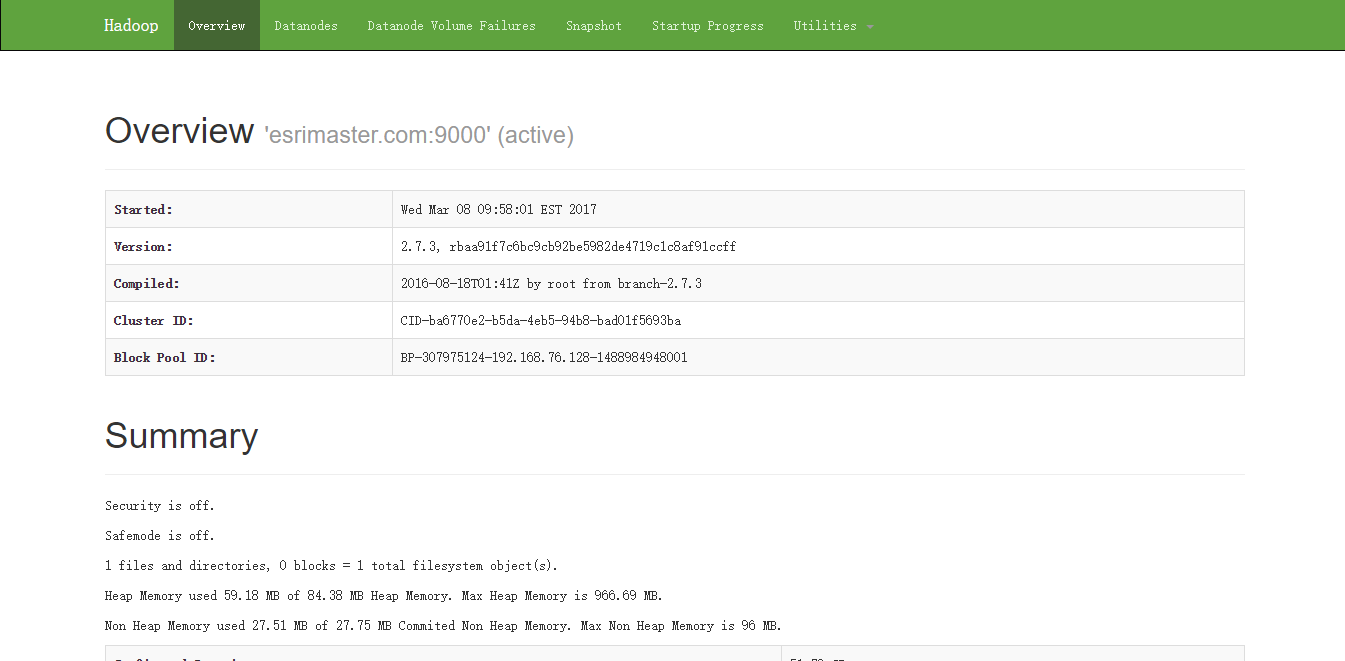
esrimaster机器：



esrislave各机器：



1. 在浏览器中通过esrimaster.com:50070来查看HDFS集群信息。

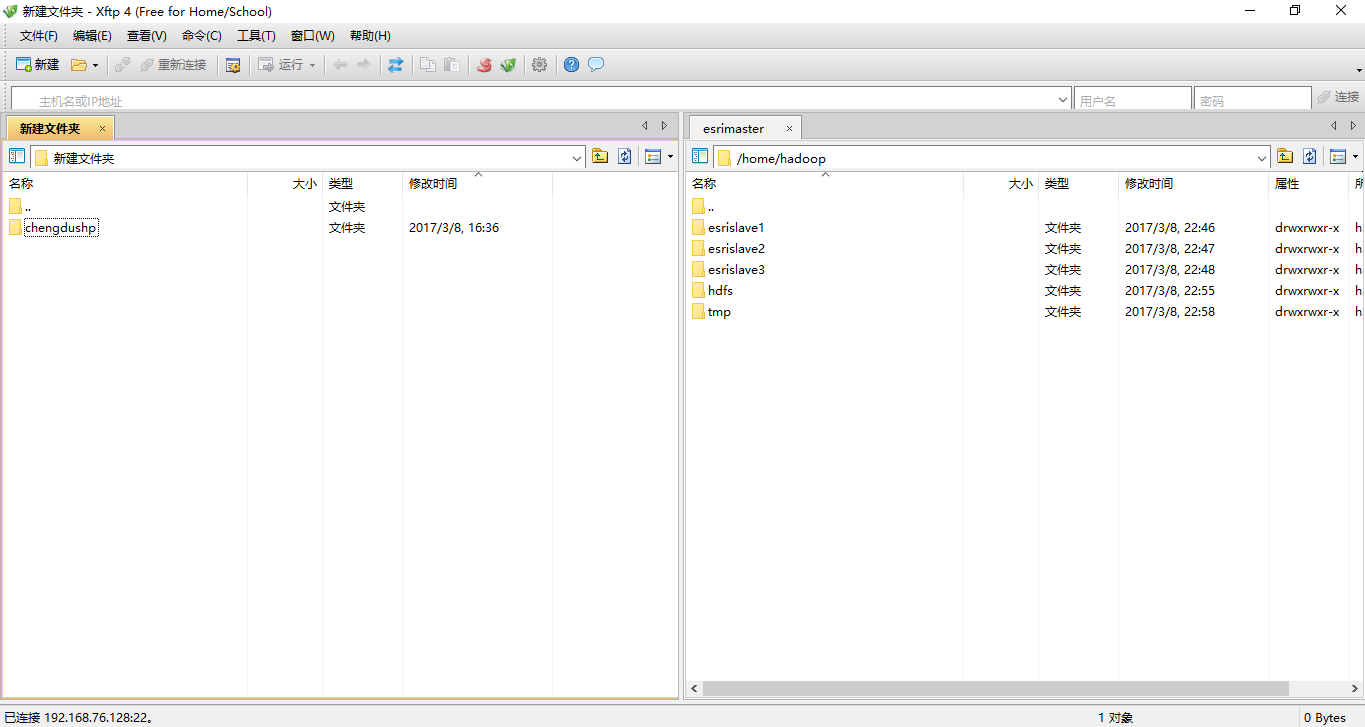


1. **注册数据**

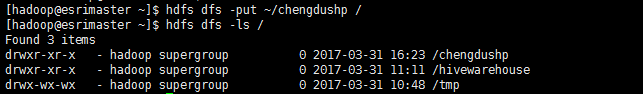
***备注：****将数据导入到HDFS上，可以通过Xftp软件将本机数据上传到hadoop集群中的任意一台集群中，再通过HDFS命令上传到HDFS。也可以通过geoprocessing-tools-for-hadoop中的Copy To HDFS工具直接将本地数据传输至HDFS。*

1. 使用Xftp软件将chengdushp文件夹上传至esrimaster节点。

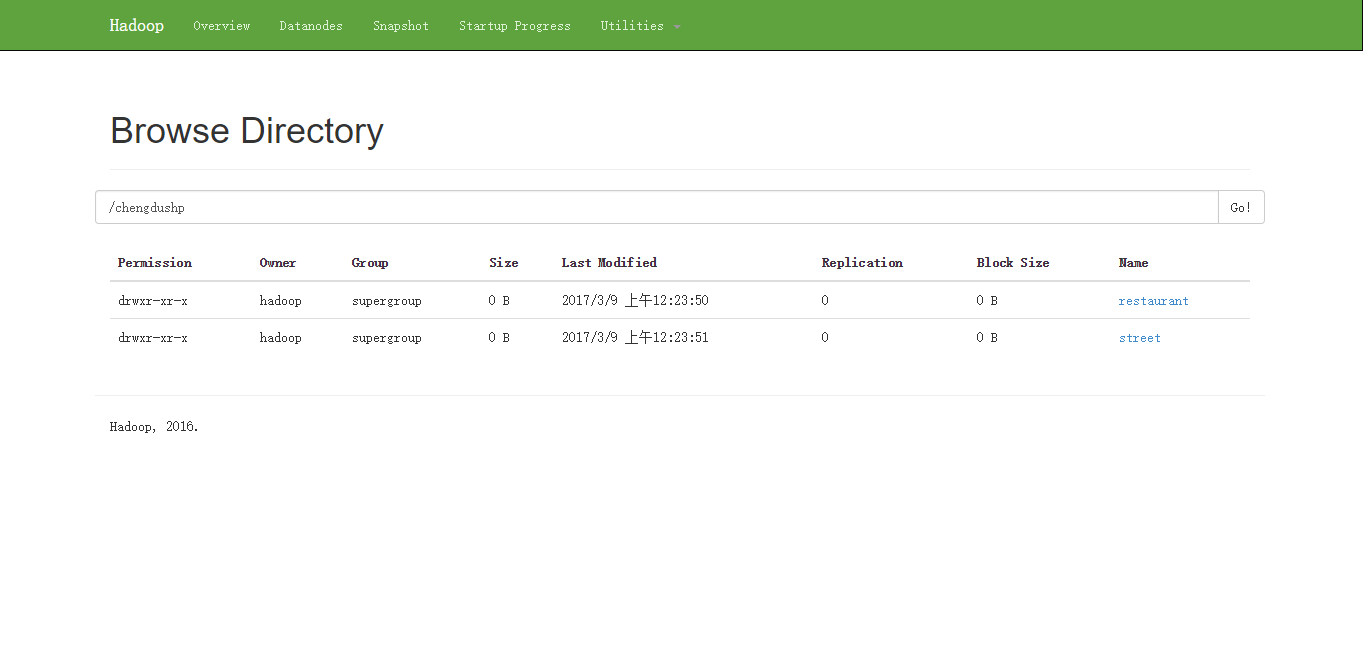
***备注****：上传至其他节点上传也可以*



1. 将chengdushp文件夹导入至HDFS。



1. 在浏览器中查看HDFS中的数据（其中，street文件夹中存放的是成都市路网数据，restaurant文件夹中存放的是餐饮点矢量数据）。



1. 打开GeoAnalytics Server集群中主节点manager地址，进入 **/站点/Data Stores**，点击**注册**，选择**大数据文件共享**，弹出框中选择**HDFS**，填写大数据文件共享**名称**和数据文件夹在HDFS中的**路径**。



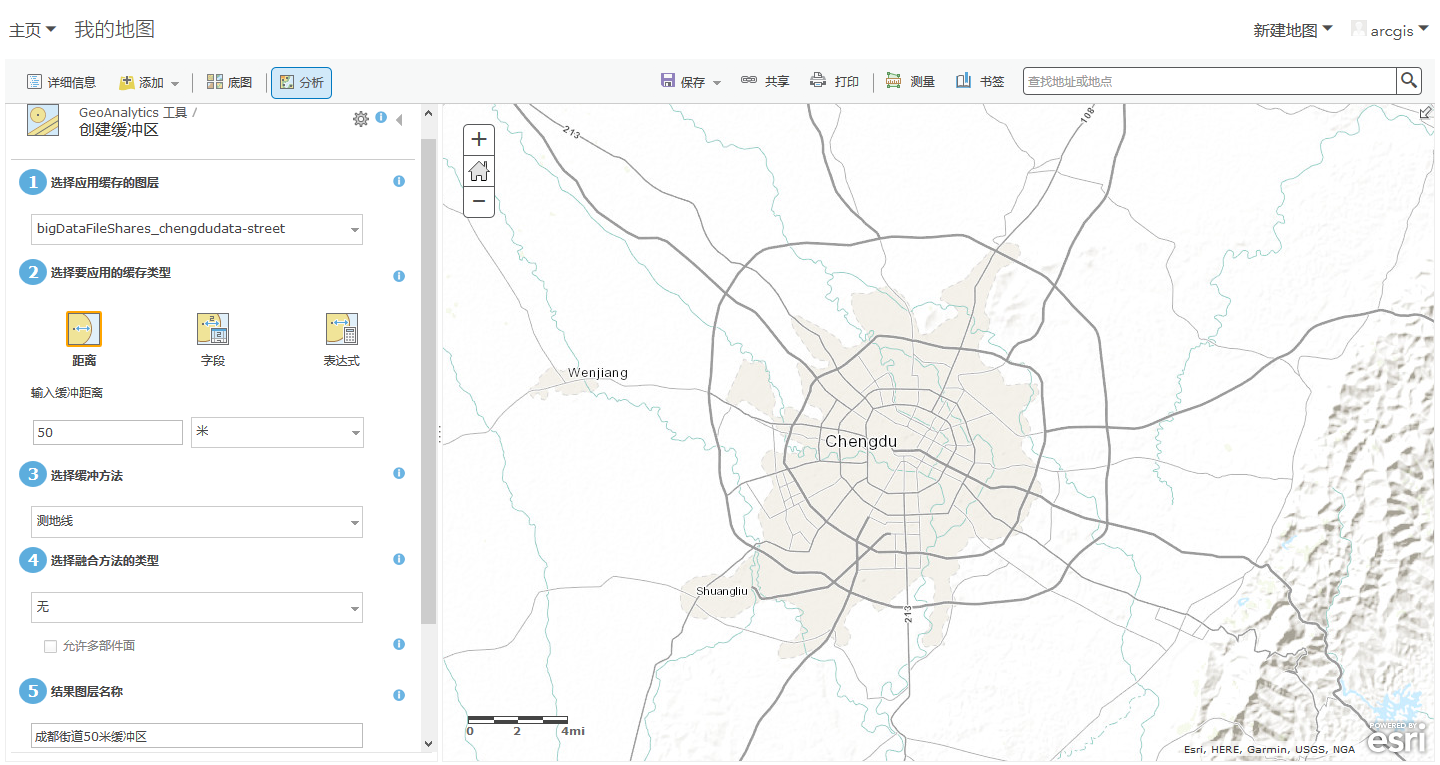
1. 点击创建，稍等片刻后即可看到数据注册成功，点击右侧铅笔按钮，可以查看并修改数据注册信息。



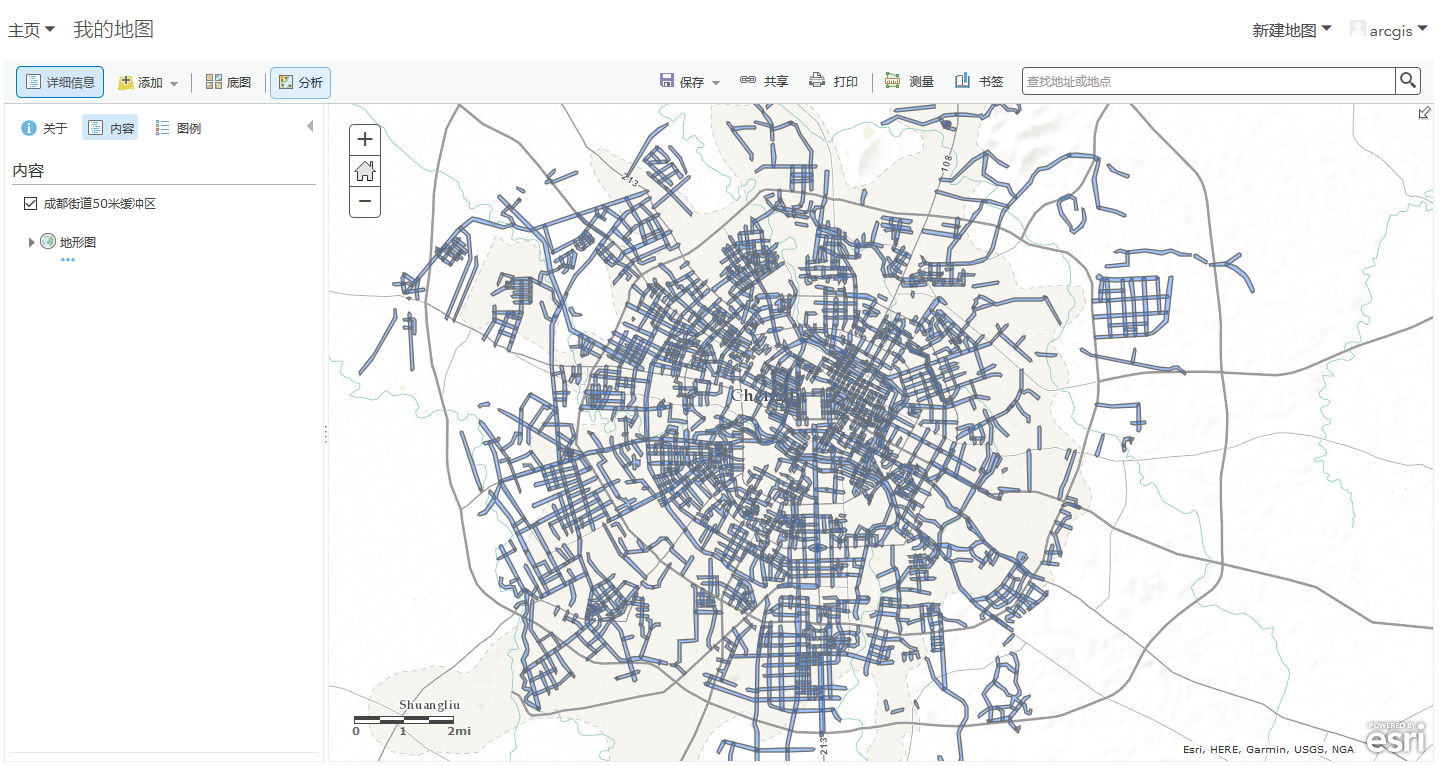
1. **执行分析**

***说明：****通过成都路网数据构建50米的缓冲区，然后分析出每条道路缓冲区中的餐饮数，来反映出成都市各路段的餐饮密集程度。*

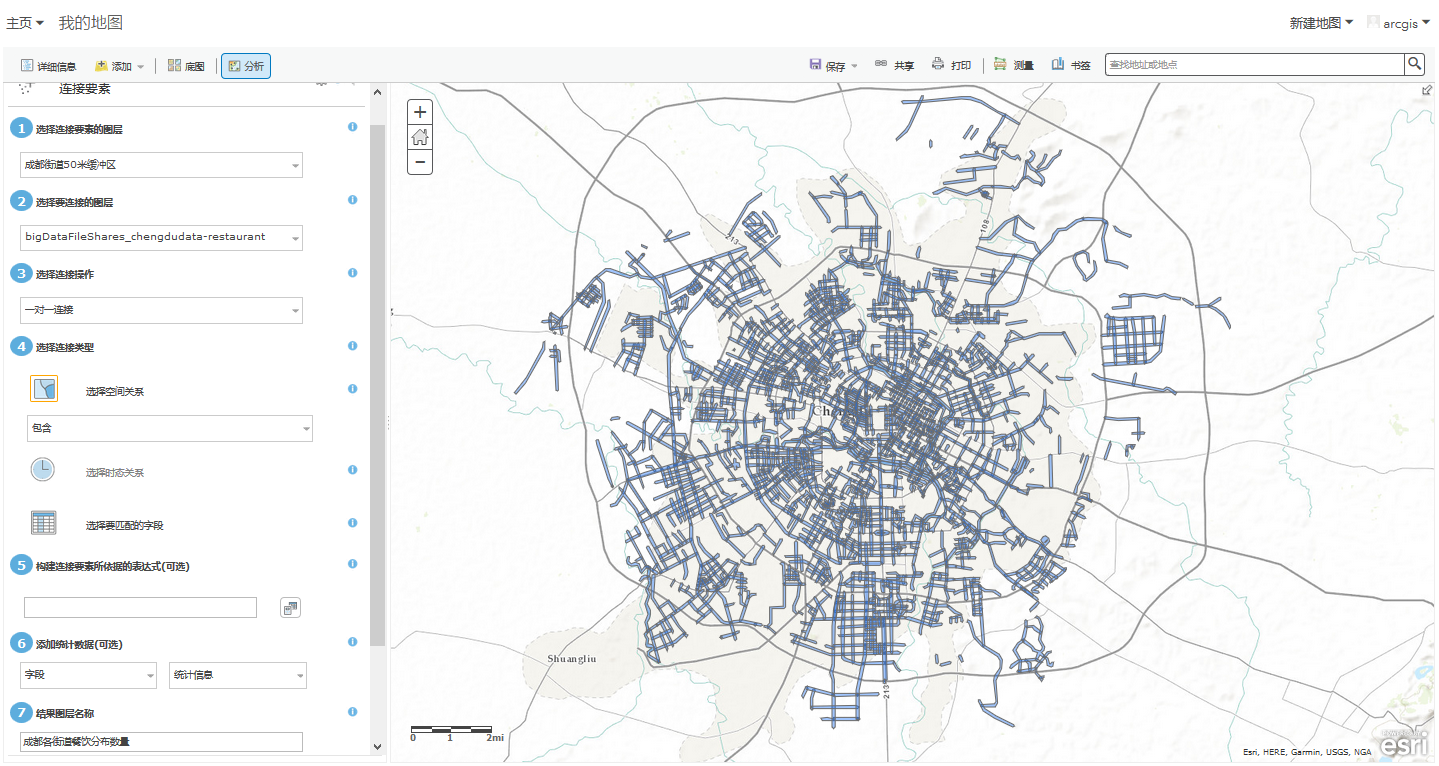
1. 进入Portal for ArcGIS，在地图中点击**分析🡪GeoAnalytics工具🡪临近分析🡪创建缓冲区**，在分析环境中将处理坐标系设置为102100，使用路网数据生成五十米的缓冲区。分析参数设置如下：



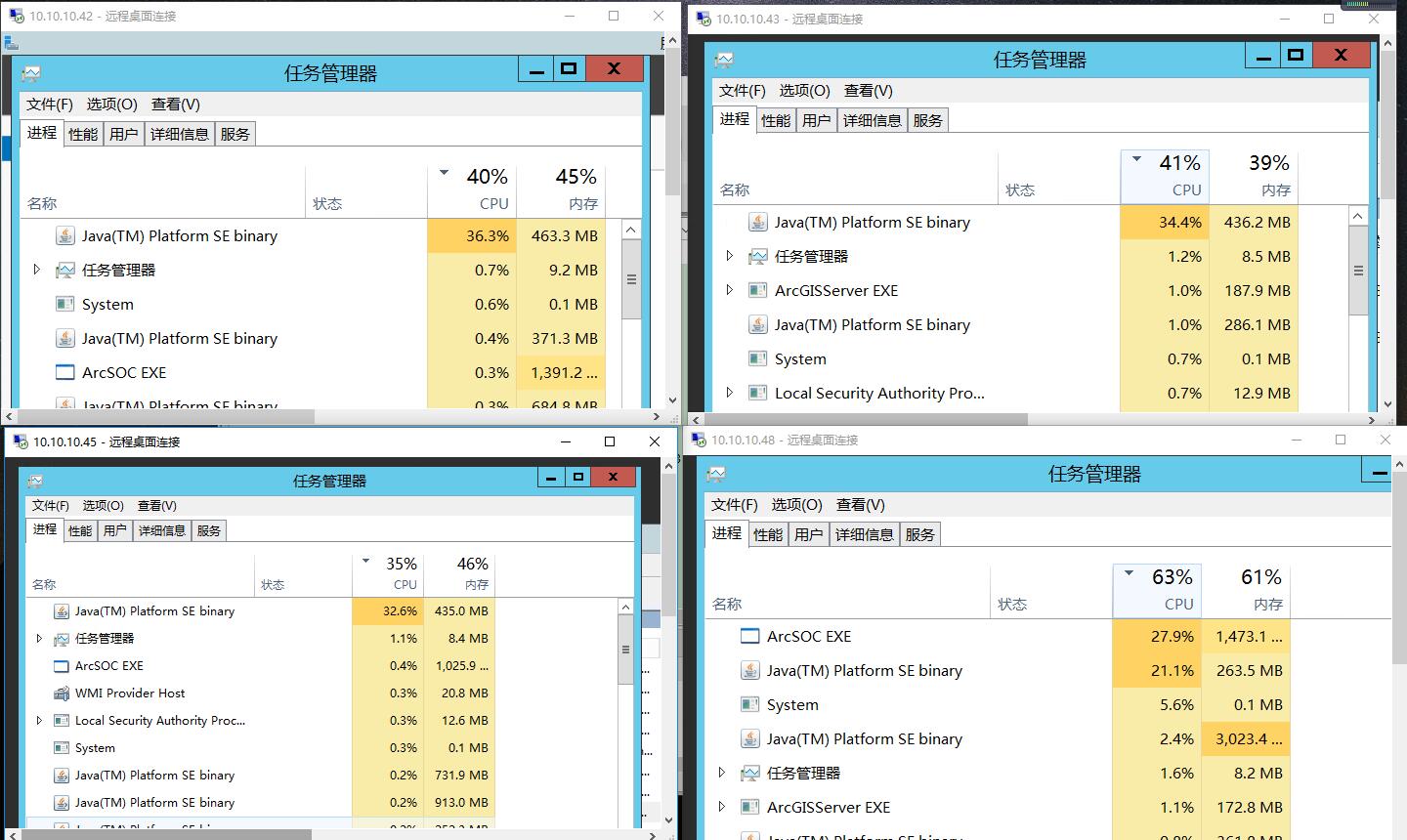
1. 点击运行分析，分析结果如下所示：



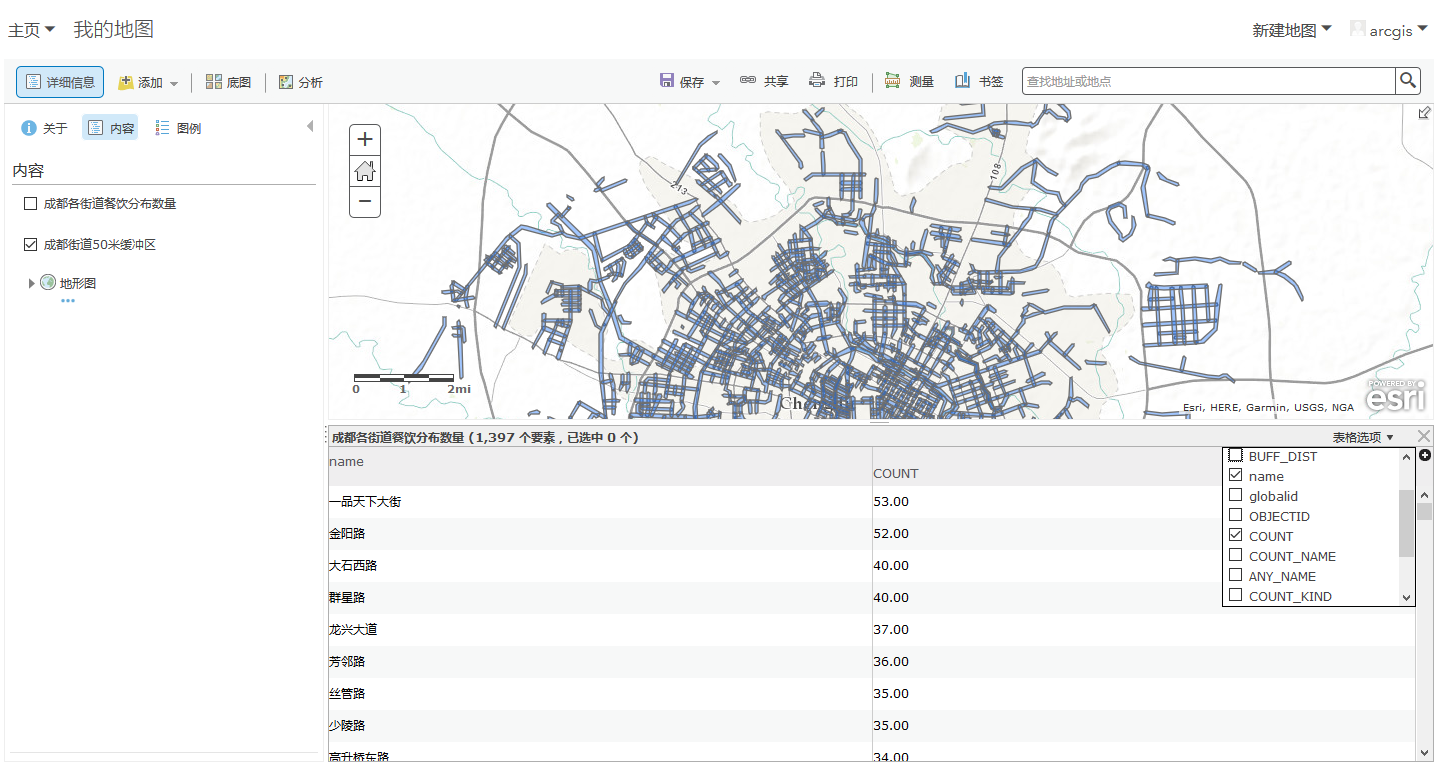
1. 通过**连接要素工具**分析出道路上的餐饮数，参数如下所示，结果图层名称设置为:“成都各街道餐饮分布数量”，并在分析环境中将处理坐标系设置为102100，然后开始执行。



1. 分析过程中GeoAnalystics Server的CPU和内存的监控。可以看到当开始分析时，CPU和内存占用迅速上升，分析结束后，CPU和内存又恢复到稳定状态。



1. 查看分析结果。



1. 分析结果可视化。

