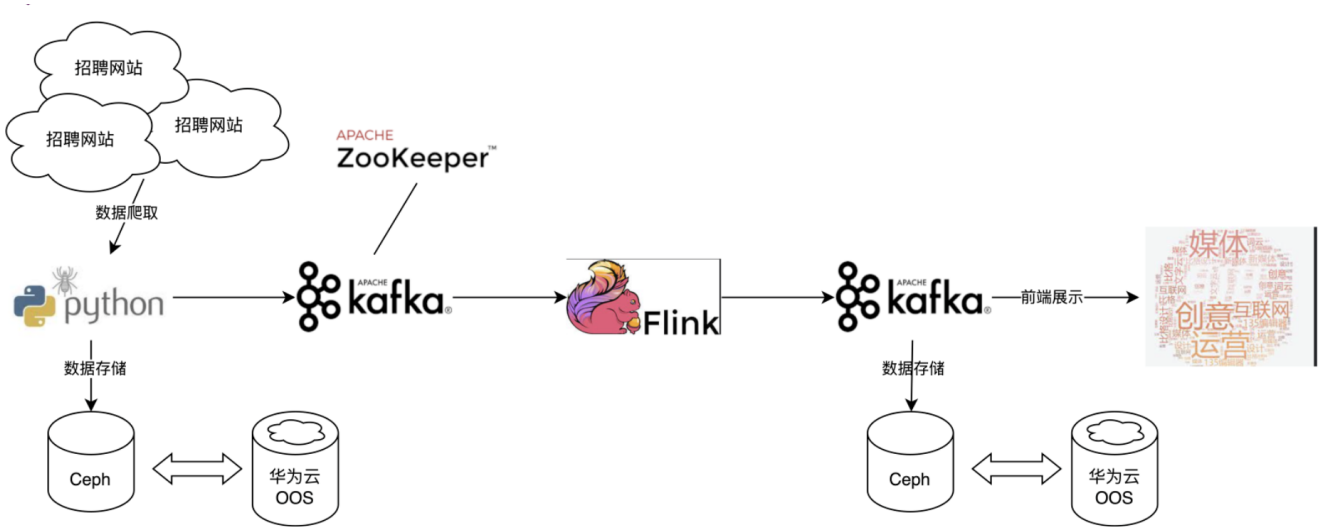
# 人才招聘洞察分析

朱民笑 张浩东 郭臣 杨沛然 张铭益 杨襄龙 蒋一杰 戴书睿

架构




# Ceph部分

1、关闭防火墙和selinux

```
sed -i  "s/SELINUX=enforcing/SELINUX=permissive/g" /etc/selinux/config
setenforce 0
systemctl stop firewalld
systemctl disable firewalld
```

2、配置hosts文件

保证集群内主机名与ip解析正常（每个节点都需要配置）

```
[root@ceph-node1 ~]# cat /etc/hosts
127.0.0.1   localhost localhost.localdomain localhost4
localhost4.localdomain4
::1         localhost localhost.localdomain localhost6
localhost6.localdomain6
192.168.56.125  ceph-node1
192.168.56.126  ceph-node2
192.168.56.127  ceph-node3
[root@ceph-node1 ~]# ping ceph-node2
PING ceph-node2 (192.168.56.126) 56(84) bytes of data.
64 bytes from ceph-node2 (192.168.56.126): icmp_seq=1 ttl=64 time=0.616 ms
……
```

3、创建部署用户及配置sudo权限（所有节点都执行）

a.考虑到使用root用户的安全性问题，所以这里创建一个 ceph-admin 普通用户做为部署及运维使用 b.再加上ceph-deploy会在节点安装软件包，所以创建的用户需要无密码 sudo 权限

```
[root@ceph-node1 ~]# useradd ceph-admin
[root@ceph-node1 ~]# echo "123456" | passwd --stdin ceph-admin
Changing password for user ceph-admin.
passwd: all authentication tokens updated successfully.

[root@ceph-node1 ~]# echo "ceph-admin ALL = NOPASSWD:ALL" | tee
/etc/sudoers.d/ceph-admin
ceph-admin ALL = NOPASSWD:ALL
[root@ceph-node1 ~]# chmod 0440 /etc/sudoers.d/ceph-admin
[root@ceph-node1 ~]# ll /etc/sudoers.d/ceph-admin
-r--r-----. 1 root root 30 Oct 19 16:06 /etc/sudoers.d/ceph-admin
```

测试

```
[root@ceph-node1 ~]# su - ceph-admin
Last login: Mon Oct 19 16:11:51 CST 2020 on pts/0
[ceph-admin@ceph-node1 ~]$ sudo su -
Last login: Mon Oct 19 16:12:04 CST 2020 on pts/0
[root@ceph-node1 ~]# exit
logout
[ceph-admin@ceph-node1 ~]$ exit
logout
```

4、配置ssh无密码访问（在主节点node1上执行）

```
[root@ceph-node1 ~]# su - ceph-admin
[ceph-admin@ceph-node1 ~]$ ssh-keygen            （每一步都按回车，口令密码留空）
[ceph-admin@ceph-node1 ~]$ ssh-copy-id ceph-admin@ceph-node1
[ceph-admin@ceph-node1 ~]$ ssh-copy-id ceph-admin@ceph-node2
[ceph-admin@ceph-node1 ~]$ ssh-copy-id ceph-admin@ceph-node3
```

5、配置ntp时间同步

配置时间同步目的：因在时间一致的情况下，才可保证集群正常运行 配置时间同步方式：node1连接网络上的ntp服务器同步时间，node2,3连接node1同步时间（即node1既为ntp服务端，也为客户端） 注：ntpd启动后需要等待几分钟去同步

```
yum -y intall ntp（安装ntp，全部节点都需要执行）
```

```
node1节点操作：
vim /etc/ntp.conf
注释掉默认的配置项：
    #server 0.centos.pool.ntp.org iburst
    #server 1.centos.pool.ntp.org iburst
    #server 2.centos.pool.ntp.org iburst
    #server 3.centos.pool.ntp.org iburst
添加配置项：
server  ntp1.aliyun.com     #阿里云ntp服务器
server 127.127.1.0     #本地ntp服务器，配置此项是为了在外网ntp连接异常的情况下还能保证ntp正常，维护集群稳定

node2/node3节点操作：
vim /etc/ntp.conf
同样注释掉默认的server配置项：
添加配置项：
server 192.168.56.125     #node1-ntp服务器

全部节点都执行：
systemctl restart ntpd
systemctl enable ntpd

查看ntp连接情况和状态
[root@ceph-node1 ~]# ntpq -p
     remote           refid      st t when poll reach   delay   offset
jitter
==============================================================================
====
*120.25.115.20   10.137.53.7      2 u   41  128  377   30.382   -1.019
1.001
 LOCAL(0)        .LOCL.           5 l  806   64    0    0.000    0.000
0.000

 [root@ceph-node2 ~]# ntpq -p
     remote           refid      st t when poll reach   delay   offset
jitter
==============================================================================
====
*ceph-node1      120.25.115.20    3 u   20   64  377    2.143   33.254
10.350

[root@ceph-node1 ~]# ntpstat
synchronised to NTP server (120.25.115.20) at stratum 3
   time correct to within 27 ms
   polling server every 128 s
```

二、开始部署Ceph集群

1、添加阿里云的base源和epel源（所有节点都执行）

```
备份系统原本的源
[root@ceph-node1 ~]# mkdir /mnt/repo_bak
```

```
[root@ceph-node1 ~]# mv /etc/yum.repos.d/* /mnt/repo_bak
添加新源
[root@ceph-node1 ~]# wget -O /etc/yum.repos.d/CentOS-Base.repo
http://mirrors.aliyun.com/repo/Centos-7.repo
[root@ceph-node1 ~]# wget -O /etc/yum.repos.d/epel.repo
http://mirrors.aliyun.com/repo/epel-7.repo
```

2、添加ceph的yum源（所有节点都执行）

注意事项： 这里的yum源是确定了ceph的版本，在源中的baseurl项中rpm-nautilus即代表着是ceph的nautilus版本的rpm包（nautilus是ceph的14.x版本）如果需要安装其他版本，还需要替换为其他版本号，12.x版本是luminous，13.x版本是rpm-mimic。 详情可以去ceph官方源中查看：download.ceph.com/

```
vim /etc/yum.repos.d/ceph.repo
[Ceph]
name=Ceph
baseurl=http://download.ceph.com/rpm-nautilus/el7/x86_64
enabled=1
gpgcheck=1
type=rpm-md
gpgkey=https://download.ceph.com/keys/release.asc
priority=1

[Ceph-noarch]
name=Ceph noarch packages
baseurl=http://download.ceph.com/rpm-nautilus/el7/noarch
enabled=1
gpgcheck=1
type=rpm-md
gpgkey=https://download.ceph.com/keys/release.asc
priority=1

[ceph-source]
name=Ceph source packages
baseurl=http://download.ceph.com/rpm-nautilus/el7/SRPMS
enabled=1
gpgcheck=1
type=rpm-md
gpgkey=https://download.ceph.com/keys/release.asc
priority=1
```

更新yum缓存及系统软件

```
yum makecache
yum -y update
```

可查看ceph版本，判断yum是否配置正确

```
[root@ceph-node1 yum.repos.d]# yum list ceph --showduplicates |sort -r
 * updates: mirrors.cn99.com
Loading mirror speeds from cached hostfile
Loaded plugins: fastestmirror
 * extras: mirrors.163.com
ceph.x86_64                          2:14.2.9-0.el7
Ceph
ceph.x86_64                          2:14.2.8-0.el7
Ceph
ceph.x86_64                          2:14.2.7-0.el7
Ceph
ceph.x86_64                          2:14.2.6-0.el7
Ceph
ceph.x86_64                          2:14.2.5-0.el7
Ceph
ceph.x86_64                          2:14.2.4-0.el7
Ceph
ceph.x86_64                          2:14.2.3-0.el7
Ceph
ceph.x86_64                          2:14.2.2-0.el7
Ceph
ceph.x86_64                          2:14.2.11-0.el7
Ceph
ceph.x86_64                          2:14.2.1-0.el7
Ceph
ceph.x86_64                          2:14.2.10-0.el7
Ceph
ceph.x86_64                          2:14.2.0-0.el7
Ceph
ceph.x86_64                          2:14.1.1-0.el7
Ceph
ceph.x86_64                          2:14.1.0-0.el7
Ceph
 * base: mirrors.163.com
Available Packages

[root@ceph-node1 yum.repos.d]# yum list ceph-deploy --showduplicates |sort
-r
 * updates: mirrors.cn99.com
Loading mirror speeds from cached hostfile
Loaded plugins: fastestmirror
 * extras: mirrors.163.com
ceph-deploy.noarch                   2.0.1-0                         Ceph-
noarch
ceph-deploy.noarch                   2.0.0-0                         Ceph-
noarch
ceph-deploy.noarch                   1.5.39-0                        Ceph-
noarch
ceph-deploy.noarch                   1.5.38-0                        Ceph-
noarch
ceph-deploy.noarch                   1.5.37-0                        Ceph-
noarch
ceph-deploy.noarch                   1.5.36-0                        Ceph-
```

```
noarch
ceph-deploy.noarch                      1.5.35-0                        Ceph-
noarch
ceph-deploy.noarch                      1.5.34-0                        Ceph-
noarch
ceph-deploy.noarch                      1.5.33-0                        Ceph-
noarch
ceph-deploy.noarch                      1.5.32-0                        Ceph-
noarch
ceph-deploy.noarch                      1.5.31-0                        Ceph-
noarch
ceph-deploy.noarch                      1.5.30-0                        Ceph-
noarch
ceph-deploy.noarch                      1.5.29-0                        Ceph-
noarch
 * base: mirrors.163.com
Available Packages
```

3、安装ceph-deploy（在主节点node1上执行）

```
[root@ceph-node1 ~]# su - ceph-admin
[ceph-admin@ceph-node1 ~]$ sudo yum -y install python-setuptools   #安装
ceph依赖包
[ceph-admin@ceph-node1 ~]$ sudo yum install ceph-deploy   （默认会选择安装2.0最
新版本）

查看ceph-deploy安装版本
[root@ceph-node1 ~]# ceph-deploy --version
2.0.1
```

4、初始化集群（在主节点node1上执行） 创建集群安装目录（ceph-deploy部署程序会将文件输出到当前目录）

```
[ceph-admin@ceph-node1 ~]$ mkdir cluster
[ceph-admin@ceph-node1 ~]$ cd cluster/

创建集群（后边是指定哪些节点做为mon监视器使用，所以选择规划中部署mon的节点-node1）
[ceph-admin@ceph-node1 cluster]$ ceph-deploy new ceph-node1
[ceph_deploy.conf][DEBUG ] found configuration file at: /home/ceph-
admin/.cephdeploy.conf
[ceph_deploy.cli][INFO  ] Invoked (2.0.1): /bin/ceph-deploy new ceph-node1
[ceph_deploy.cli][INFO  ] ceph-deploy options:
[ceph_deploy.cli][INFO  ]  username                      : None
[ceph_deploy.cli][INFO  ]  func                          : <function new
at 0x7f14c44c9de8>
[ceph_deploy.cli][INFO  ]  verbose                       : False
[ceph_deploy.cli][INFO  ]  overwrite_conf                : False
[ceph_deploy.cli][INFO  ]  quiet                         : False
[ceph_deploy.cli][INFO  ]  cd_conf                       :
```

```
<ceph_deploy.conf.cephdeploy.Conf instance at 0x7f14c3c424d0>
[ceph_deploy.cli][INFO  ]  cluster                       : ceph
[ceph_deploy.cli][INFO  ]  ssh_copykey                   : True
[ceph_deploy.cli][INFO  ]  mon                           : ['ceph-node1']
[ceph_deploy.cli][INFO  ]  public_network                : None
[ceph_deploy.cli][INFO  ]  ceph_conf                     : None
[ceph_deploy.cli][INFO  ]  cluster_network               : None
[ceph_deploy.cli][INFO  ]  default_release               : False
[ceph_deploy.cli][INFO  ]  fsid                          : None
[ceph_deploy.new][DEBUG ] Creating new cluster named ceph
[ceph_deploy.new][INFO  ] making sure passwordless SSH succeeds
[ceph-node1][DEBUG ] connection detected need for sudo
[ceph-node1][DEBUG ] connected to host: ceph-node1
[ceph-node1][DEBUG ] detect platform information from remote host
[ceph-node1][DEBUG ] detect machine type
[ceph-node1][DEBUG ] find the location of an executable
[ceph-node1][INFO  ] Running command: sudo /usr/sbin/ip link show
[ceph-node1][INFO  ] Running command: sudo /usr/sbin/ip addr show
[ceph-node1][DEBUG ] IP addresses found: [u'192.168.56.125']
[ceph_deploy.new][DEBUG ] Resolving host ceph-node1
[ceph_deploy.new][DEBUG ] Monitor ceph-node1 at 192.168.56.125
[ceph_deploy.new][DEBUG ] Monitor initial members are ['ceph-node1']
[ceph_deploy.new][DEBUG ] Monitor addrs are ['192.168.56.125']
[ceph_deploy.new][DEBUG ] Creating a random mon key...
[ceph_deploy.new][DEBUG ] Writing monitor keyring to ceph.mon.keyring...
[ceph_deploy.new][DEBUG ] Writing initial config to ceph.conf...

[ceph-admin@ceph-node1 cluster]$ ls
ceph.conf  ceph-deploy-ceph.log  ceph.mon.keyring

在当前目录下的ceph.conf中添加以下两行内容
public_network = 192.168.56.0/24
cluster_network = 192.168.56.0/24

安装Ceph包至其他节点
（其中 --no-adjust-repos 参数含义：使用本地配置的源，不更改源。以防出现问题）
[ceph-admin@ceph-node1 cluster]$ ceph-deploy install --no-adjust-repos
ceph-node1 ceph-node2 ceph-node3
```

如果出现“RuntimeError: Failed to execute command: ceph --version”报错，是因为服务器网络问题导致，下载ceph安装包速度太慢，达到5分钟导致超时，可以重复执行，或者单独在所有节点执行yum -y install ceph即可

初始化mon节点

在2.0.1版本的ceph-deploy中在该初始化的时候就会做收集密钥的动作，无需再执行 ceph-deploy gatherkeys {monitor-host} 这个命令

[ceph-admin@ceph-node1 cluster]$ ceph-deploy mon create-initial

5、添加OSD

如果是里边有数据的磁盘，还需先清除数据：（详细可查看 ceph-depoy disk zap --help）

```
列出所有节点上所有可用的磁盘
[ceph-admin@ceph-node1 cluster]$ ceph-deploy disk list ceph-node1 ceph-
node2 ceph-node3
清除数据
sudo ceph-deploy disk zap {osd-server-name} {disk-name}
    eg: sudo ceph-deploy disk zap ceph-node2 /dev/sdb

如果是干净的磁盘，可忽略上边清除数据的操作，直接添加OSD即可
（我这里是新添加的/dev/sdb磁盘）
[ceph-admin@ceph-node1 cluster]$ ceph-deploy osd create --data /dev/sdb
ceph-node1
[ceph-admin@ceph-node1 cluster]$ ceph-deploy osd create --data /dev/sdb
ceph-node2
[ceph-admin@ceph-node1 cluster]$ ceph-deploy osd create --data /dev/sdb
ceph-node3

可以看到cpeh将新增OSD创建为LVM格式加入ceph集群中
[ceph-admin@ceph-node1 cluster]$ sudo pvs
  PV         VG                                        Fmt  Attr PSize
PFree
  /dev/sdb   ceph-ab1b8533-018e-4924-8520-fdbefbb7d184 lvm2 a--  <10.00g
0
```

6、允许主机以管理员权限执行 Ceph 命令 将ceph-deploy命令将配置文件和 admin key复制到各个ceph节点，其他节点主机也能管理ceph集群 [ceph-admin@ceph-node1 cluster]$ ceph-deploy admin ceph-node1 ceph-node2 ceph-node3

7、部署MGR用于获取集群信息 [ceph-admin@ceph-node1 cluster]$ ceph-deploy mgr create ceph-node1

查看集群状态

```
[ceph-admin@ceph-node1 cluster]$ sudo ceph health detail
HEALTH_OK
[ceph-admin@ceph-node1 cluster]$ sudo ceph -s
  cluster:
    id:     e9290965-40d4-4c65-93ed-e534ae389b9c
    health: HEALTH_OK

  services:
    mon: 1 daemons, quorum ceph-node1 (age 62m)
    mgr: ceph-node1(active, since 5m)
    osd: 3 osds: 3 up (since 12m), 3 in (since 12m)

  data:
    pools:   0 pools, 0 pgs
    objects: 0 objects, 0 B
    usage:   3.0 GiB used, 27 GiB / 30 GiB avail
    pgs:
```

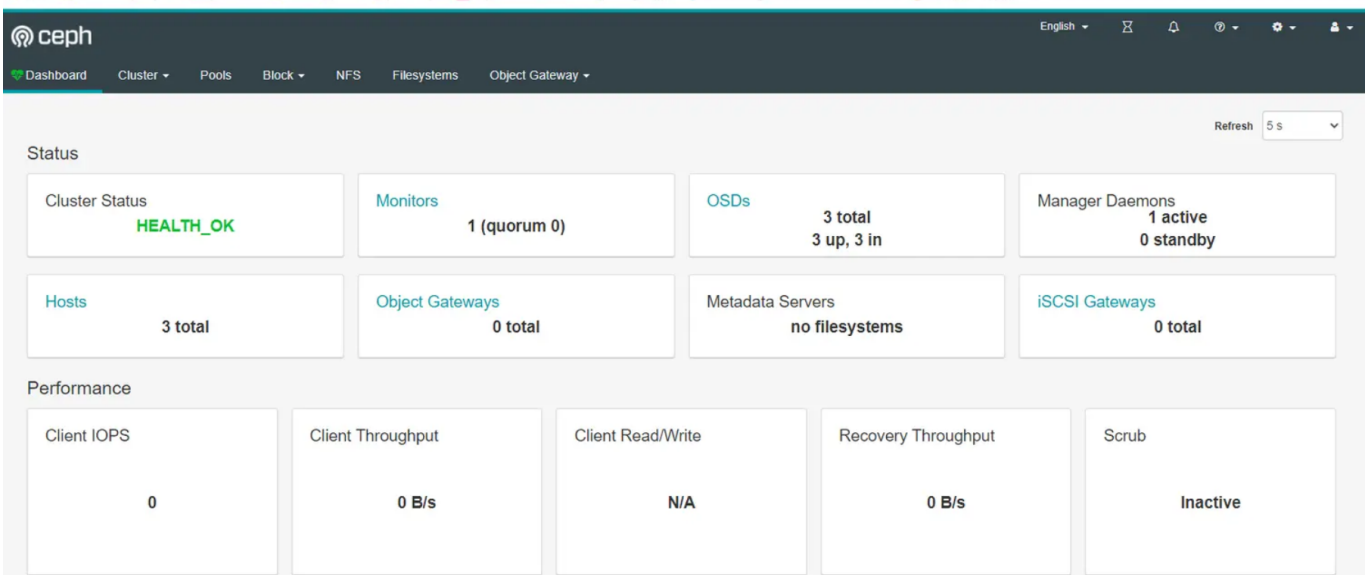
```
如果查看集群状态为“HEALTH_WARN mon is allowing insecure global_id reclaim”，是
因为开启了不安全的模式，将之禁用掉即可：

 [ceph-admin@ceph-node1 cluster]$ sudo ceph config set mon
auth_allow_insecure_global_id_reclaim false

因/etc/ceph/下key文件普通用户没有读权限，所以普通用户无权直接执行ceph命令
如果需要ceph-admin普通用户也可直接调用集群，增加对ceph配置文件的读权限即可
（想要每个节点普通用户都可以执行ceph相关命令，那就所有节点都修改权限）
[ceph-admin@ceph-node1 ~]$ ll /etc/ceph/
total 12
-rw-------. 1 root root 151 Oct 21 17:33 ceph.client.admin.keyring
-rw-r--r--. 1 root root 268 Oct 21 17:35 ceph.conf
-rw-r--r--. 1 root root  92 Oct 20 04:48 rbdmap
-rw-------. 1 root root   0 Oct 21 17:30 tmpcmU035
[ceph-admin@ceph-node1 ~]$ sudo chmod +r
/etc/ceph/ceph.client.admin.keyring
[ceph-admin@ceph-node1 ~]$ ll /etc/ceph/
total 12
-rw-r--r--. 1 root root 151 Oct 21 17:33 ceph.client.admin.keyring
-rw-r--r--. 1 root root 268 Oct 21 17:35 ceph.conf
-rw-r--r--. 1 root root  92 Oct 20 04:48 rbdmap
-rw-------. 1 root root   0 Oct 21 17:30 tmpcmU035
[ceph-admin@ceph-node1 ~]$ ceph -s
  cluster:
    id:     130b5ac0-938a-4fd2-ba6f-3d37e1a4e908
    health: HEALTH_OK

  services:
    mon: 1 daemons, quorum ceph-node1 (age 20h)
    mgr: ceph-node1(active, since 20h)
    osd: 3 osds: 3 up (since 20h), 3 in (since 20h)

  data:
    pools:   0 pools, 0 pgs
    objects: 0 objects, 0 B
    usage:   3.0 GiB used, 27 GiB / 30 GiB avail
    pgs:
```

三、配置Mgr-Dashboard模块 开启dashboard模块

```
[ceph-admin@ceph-node1 ~]$ sudo ceph mgr module enable dashboard

如果报错如下：
Error ENOENT: all mgr daemons do not support module 'dashboard', pass --
force to force enablement

那是因为没有安装ceph-mgr-dashboard，在mgr节点上安装即可
[ceph-admin@ceph-node1 ~]$ sudo yum -y install ceph-mgr-dashboard

默认情况下，仪表板的所有HTTP连接均使用SSL/TLS进行保护。
要快速启动并运行仪表板，可以使用以下命令生成并安装自签名证书
```

```
[ceph-admin@ceph-node1 ~]$ sudo ceph dashboard create-self-signed-cert
Self-signed certificate created

创建具有管理员角色的用户:
[ceph-admin@ceph-node1 ~]$ sudo ceph dashboard set-login-credentials admin
admin
******************************************************************
***          WARNING: this command is deprecated.              ***
*** Please use the ac-user-* related commands to manage users. ***
******************************************************************
Username and password updated


之前用的“admin admin”，现在好像不能直接这样写了，需要将密码写在一个文件中读取，不然会报
错
“dashboard set-login-credentials <username> : Set the login credentials.
Password read from -i <file>”

那就加上-i参数来创建也是一样
[ceph-admin@ceph-node1 cluster]$ echo admin > userpass
[ceph-admin@ceph-node1 cluster]$ sudo ceph dashboard set-login-credentials
admin -i userpass
******************************************************************
***          WARNING: this command is deprecated.              ***
*** Please use the ac-user-* related commands to manage users. ***
******************************************************************
Username and password updated

查看ceph-mgr服务:
[ceph-admin@ceph-node1 ~]$ sudo ceph mgr services
{
    "dashboard": "https://ceph-node1:8443/"
}
```

浏览器访问测试:



# Flink部分

所有jar包可在对应module的target目录中获取。

## FilerWatcher

jar包，用于监听是否有新数据被爬取，将新数据输入Kafka的某个topic。 arg0:kafka topic名 arg1:监听路径

```
java -jar FileWatcher-1.0-SNAPSHOT.jar arg0 arg1
```

## FilerWriter

jar包，用于监听Kafka的某个topic是否有新消息，并将新消息写入文件。 arg0:kafka topic名 arg1:写入文件全路径名

```
java -jar FileWriter-1.0-SNAPSHOT.jar arg0 arg1
```

## Flink

依赖：

zookeeper3.4.13 配置在2181端口
kafka2.1.1 配置在9092端口

Flink集群配置了三个节点master，worker1，worker2，每个节点中有一个slot。在启动集群后，浏览器打开master:8081 进入flink dashboard提交任务。

任务jar包：
ciyun-1.0-SNAPSHOT.jar
任务入口：

```
# 计算北京地区Python岗位的描述关键词词频
com.zmy.CiyunJob
```

flink_python-1.0-SNAPSHOT.jar
任务入口：

```
# 计算北京各地区Python岗位的数量
com.zmy.PythonAreaJob
# 计算北京地区Python岗位的学历要求
com.zmy.PythonDegreeJob
```

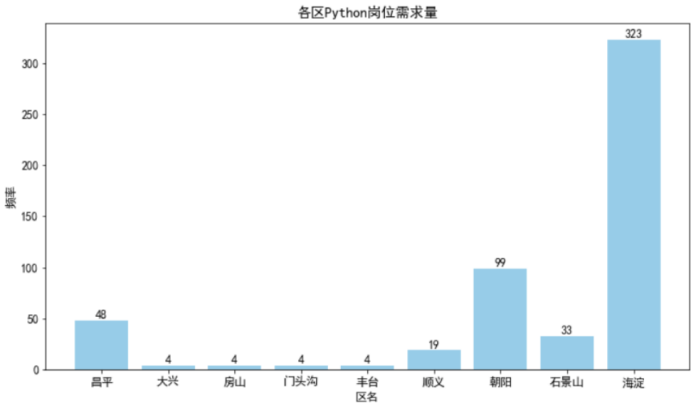
salary-1.0-SNAPSHOT.jar
任务入口：

```
# 计算北京地区Python岗位按地区分组计算平均工资
com.zmy.SalaryJob
```

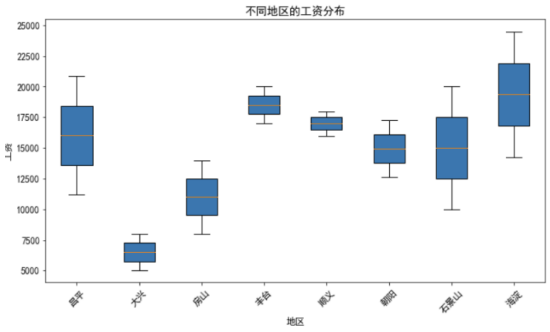
# 结果

北京地区Python岗位描述中关键词词云


北京各区Python岗位需求量




北京地区Python岗位工资分布




北京Python岗位学历要求分布