

机器学习平台多租户部署

第一部分：基础软件安装

1.1 软件准备

集群已部署 spark2.2 以上，jdk1.8.0_111，版本并不是越新越好，会受版本兼容性影响。
由于是公共软件，不能安装在某个用户目录下。

下载 py3.5

wget https://repo.continuum.io/archive/Anaconda3-4.2.0-Linux-x86_64.sh

下载 pyarrow

wget

https://pypi.python.org/packages/0c/19/805aa541740279bc8a198eeeb57509de5551f55f0cbc6371fa897bfc3245/pyarrow-0.8.0-cp35-cp35m-manylinux1_x86_64.whl

下载 jupyterhub（管理 jupyter notebook 多租户）

wget

<https://pypi.python.org/packages/bd/36/2c98cae181c50d955a9f7157ee0a1db80b234fd8b8c11b76b0a37efb695a/jupyterhub-0.8.1-py3-none-any.whl>

下载 npm

wget <https://npm.taobao.org/mirrors/node/latest-v8.x/node-v8.8.1-linux-x64.tar.gz>

1.2 python3.5 安装

sh Anaconda3-4.2.0-Linux-x86_64.sh

安装目录: /data/anaconda3

配置环境变量 /etc/profile

```
export ANACONDA_HOME=/data/anaconda3
export PATH=$ANACONDA_HOME/bin:$PATH
```

1.3 npm 安装

cd /data

tar -zxvf node-v8.8.1-linux-x64.tar.gz

```
ln -s node-v8.8.1-linux-x64 node
```

```
export NODE_HOME= /work/zhongls/node-v8.8.1-linux-x64/  
export PATH=$NODE_HOME/bin:$PATH
```

1.4 环境生效

```
source /etc/profile
```

版本查看

```
[root@mvx16226 data]# python -V  
Python 3.5.2 :: Anaconda 4.2.0 (64-bit)  
[root@mvx16226 data]# npm -version  
5.4.2
```

1.5 jupyterhub 安装

```
pip install jupyterhub-0.8.1-py3-none-any.whl
```

1.6 configurable-http-proxy 安装

```
unzip configurable-http-proxy-master.zip  
mv configurable-http-proxy-master  
/data/node/lib/node_modules/configurable-http-proxy  
cd /data/node/lib/node_modules/configurable-http-proxy  
npm install -g configurable-http-proxy
```

第二部分: jupyterhub 多租户配置

2.1 jupyterhub_config 配置

创建配置文件

```
cd /data/jupyterhub_conf  
jupyterhub --generate-config  
sudo openssl req -x509 -nodes -days 200 -newkey rsa:1024 -keyout mykey.key -out  
mycert.pem
```

编辑配置

```
vi jupyterhub_config.py
```

```
c = get_config()
```

```
c.JupyterHub.ssl_cert = '/data/jupyterhub_conf/mycert.pem'
c.JupyterHub.ssl_key = '/data/jupyterhub_conf/mykey.key'
c.JupyterHub.port = 8011
```

2.2 ipython 配置

jupyter notebook 启动 python/pyspark 时的环境配置

```
mkdir -p /var/lib/hive/.ipython/profile_default/startup
```

```
cd /var/lib/hive/.ipython/profile_default/startup
```

```
vi 00-first.py
```

```
import os
import sys
#os.environ["JAVA_HOME"] = "/var/lib/hive/app/jdk1.8.0_111/jre"
os.environ["SPARK_HOME"] = "/data/spark/python"
os.environ["PYLIB"] = os.environ["SPARK_HOME"] + "/python/lib"
os.environ["PYSARK_PYTHON"] = "/data/anaconda3/bin/python"
sys.path.insert(0, os.environ["PYLIB"] + "/py4j-0.10.6-src.zip")
sys.path.insert(0, os.environ["PYLIB"] + "/pyspark.zip")
```

2.3 创建新用户

```
sudo groupadd ml_dev
```

```
sudo useradd tancz1 -g ml_dev -d /data/tancz1
```

```
sudo passwd tancz1
```

```
mkdir -p /data/tancz1/.ipython/profile_default/startup
```

```
cp /var/lib/hive/.ipython/profile_default/startup/00-first.py
```

```
/data/tancz1/.ipython/profile_default/startup/00-first.py
```

```
chown -R tancz1:ml_dev /data/tancz1/.ipython
```

第三部分: jupyterhub 多租户管理

使用 root 启动服务

```
jupyterhub --config=/root/jupyterhub_config.py
```

```
jupyterhub --no-ssl
```

登录个人账号，密码跟 linux 服务器一致



Sign in

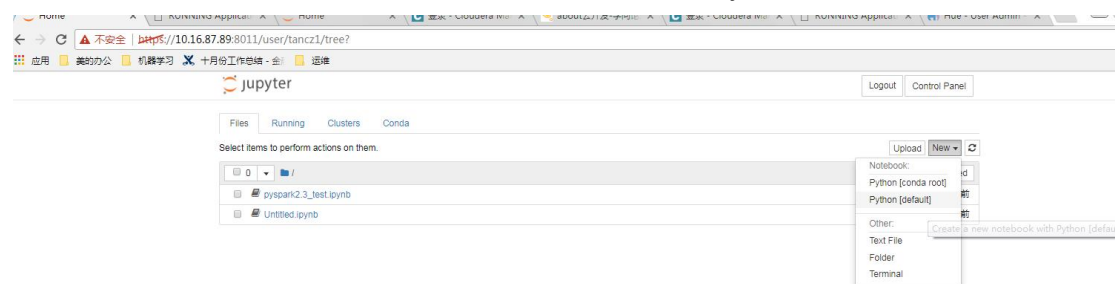
Username:

tancz1

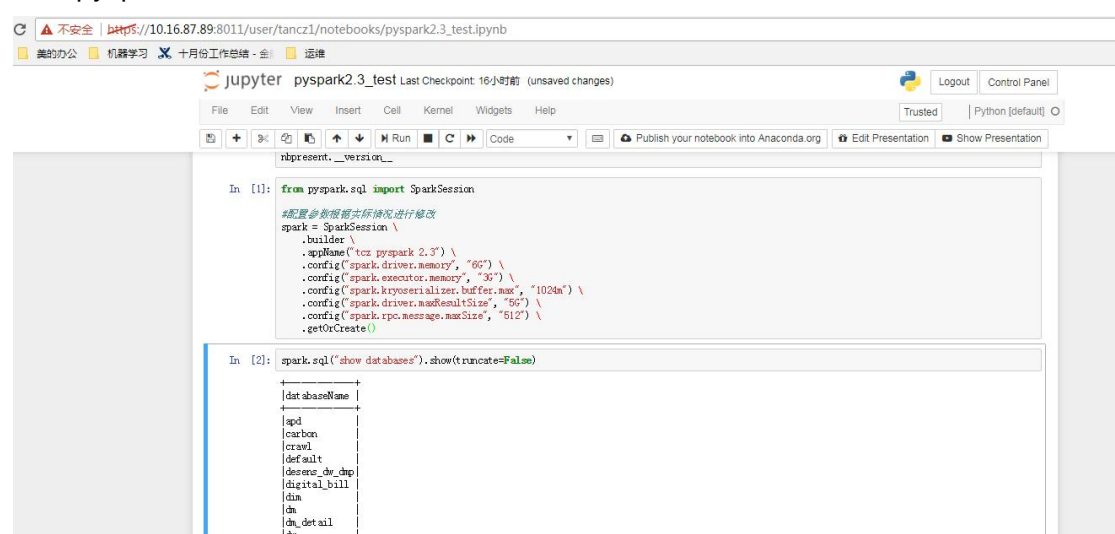
Password:

Sign In

按个人账号维护脚本，可针对用户进行 hive 表访问权限和 yarn 资源控制




运行 pyspark 脚本



对应 yarn 进程

← → ↻ 10.16.87.89:8088/cluster/apps/RUNNING

应用 美知办公 机器学习 十月份工作总结 - 金 运维



RUNNING Applications

Logged in as: dr.

Cluster

About

Nodes

Applications

NEW

NEW_SAVING

SUBMITTED

ACCEPTED

RUNNING

FINISHED

FAILED

KILLED

Scheduler

Tools

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	VCores Used	VCores Total	VCores Reserve
4	0	1	3	1	1 GB	25.03 GB	0 B	1	48	0

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	Rebooted Nodes
3	0	0	0	0	0

User Metrics for dr,who

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Containers Pending	Containers Reserved	Memory Used	Memory Pending	Memory Reserved	VCores Used	VCores Pending	VCores Reserved
0	0	0	0	0	0	0 B	0 B	0 B	0 B	0	0	0

Show 20 entries

Search:

ID	User	Name	Application Type	Queue	StartTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCoers	Allocated Memory MB	Reserved CPU VCoers	Reserved Memory MB	Progress	Tracking UI
application_1521026821710_0004	tancz1	tcz pyspark 2.3	SPARK	root.users.tancz1	Thu Mar 15 11:50:35 +0800 2018	N/A	RUNNING	UNDEFINED	1	1	1024	0	0		ApplicationMa