# Hadoop 集群实操实验 (20 分钟)

**什么是 Hadoop 集群?**

Hadoop 集群是由计算机组成的集合，称为节点，这些节点通过网络连接在一起，以对大数据集进行并行计算。名称节点是 Hadoop 分布式文件系统（HDFS）的主节点。它在 RAM 中维护文件的元数据，以便快速访问。实际的 Hadoop 集群设置涉及大量资源，这超出了本实验的范围。在本实验中，您将使用 Docker 化的 Hadoop 创建一个 Hadoop 集群，具体包括：

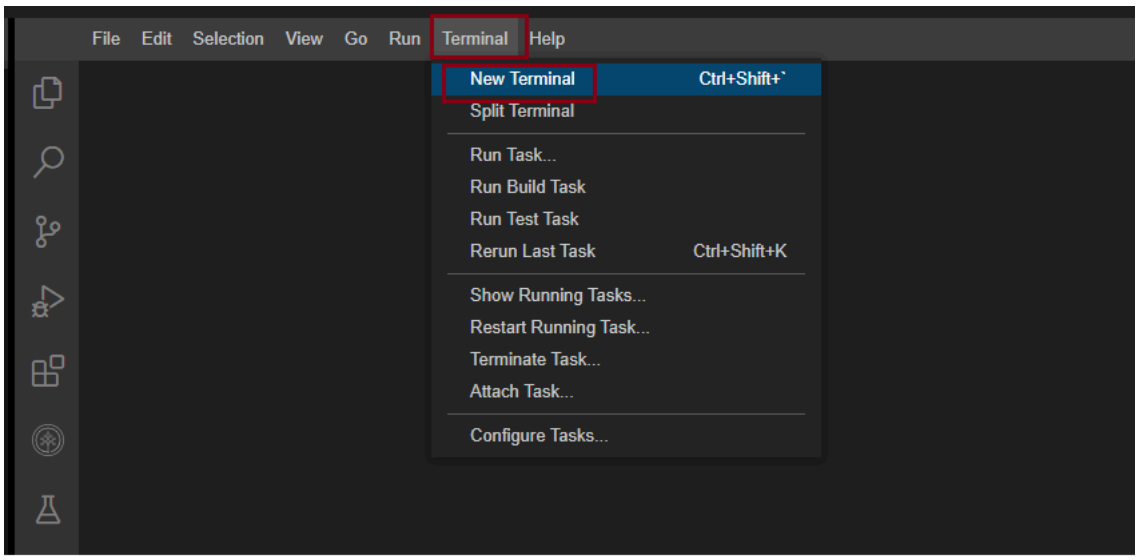
1. Namenode
2. Datanode
3. Node Manager
4. Resource manager
5. Hadoop 历史服务器

## 目标

- 运行一个 Docker 化的 Hadoop 集群实例
- 在 HDFS 中创建一个文件并在 GUI 上查看它

### 设置集群节点 Docker 化 Hadoop

1. 启动一个新的终端




2. 将仓库克隆到你的 theia 环境中。

```
git clone https://github.com/ibm-developer-skills-network/ooxwv-docker_hadoop.git
```

3. 导航到 docker-hadoop 目录以进行构建。

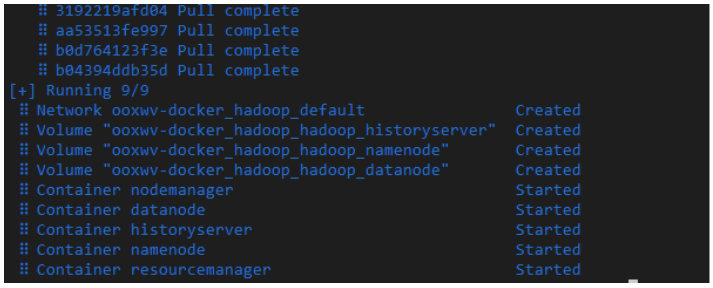
```
cd ooxwv-docker_hadoop
```

4. 编写 Docker 应用程序。

```
docker-compose up -d
```

**Compose** 是一个用于定义和运行多容器 Docker 应用程序的工具。它使用 YAML 文件来配置服务，使我们能够仅通过一个配置文件创建和启动所有服务。

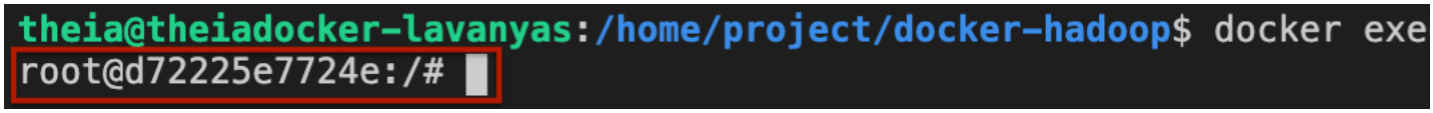
您将看到所有五个容器已创建并启动。




5. 在 bash 上以挂载驱动器的方式运行 namenode。

```
docker exec -it namenode /bin/bash
```

6. 您将观察到提示会发生变化，如下所示。




# 探索 Hadoop 环境

正如你在本课程的视频和阅读中所学到的，Hadoop 环境是通过编辑一组配置文件来配置的：

- **hadoop-env.sh** 作为主文件，用于配置 YARN、HDFS、MapReduce 和与 Hadoop 相关的项目设置。
- **core-site.xml** 定义 HDFS 和 Hadoop 核心属性。
- **hdfs-site.xml** 管理存储节点元数据、fsimage 文件和日志文件的位置。
- **mapred-site-xml** 列出 MapReduce 配置的参数。
- **yarn-site.xml** 定义与 YARN 相关的设置。它包含 Node Manager、Resource Manager、Containers 和 Application Master 的配置。

对于 Docker 镜像，这些 xml 文件已经配置好。你可以通过运行以下命令在目录 **/opt/hadoop-3.2.1/etc/hadoop/** 中查看这些文件。

```
ls /opt/hadoop-3.2.1/etc/hadoop/*.xml
```

### 在 HDFS 中创建文件

1. 在 HDFS 中，创建一个名为 `user/root/input` 的目录结构。

```
hdfs dfs -mkdir -p /user/root/input
```

2. 将所有的 hadoop 配置 xml 文件复制到输入目录中。

```
hdfs dfs -put $HADOOP_HOME/etc/hadoop/*.xml /user/root/input
```

3. 在当前目录中创建一个 `data.txt` 文件。

```
curl https://raw.githubusercontent.com/ibm-developer-skills-network/ooxwv-docker_hadoop/master/SampleMapReduce.txt --output data.txt
```

4. 将 `data.txt` 文件复制到 `/user/root`。

```
hdfs dfs -put data.txt /user/root/
```

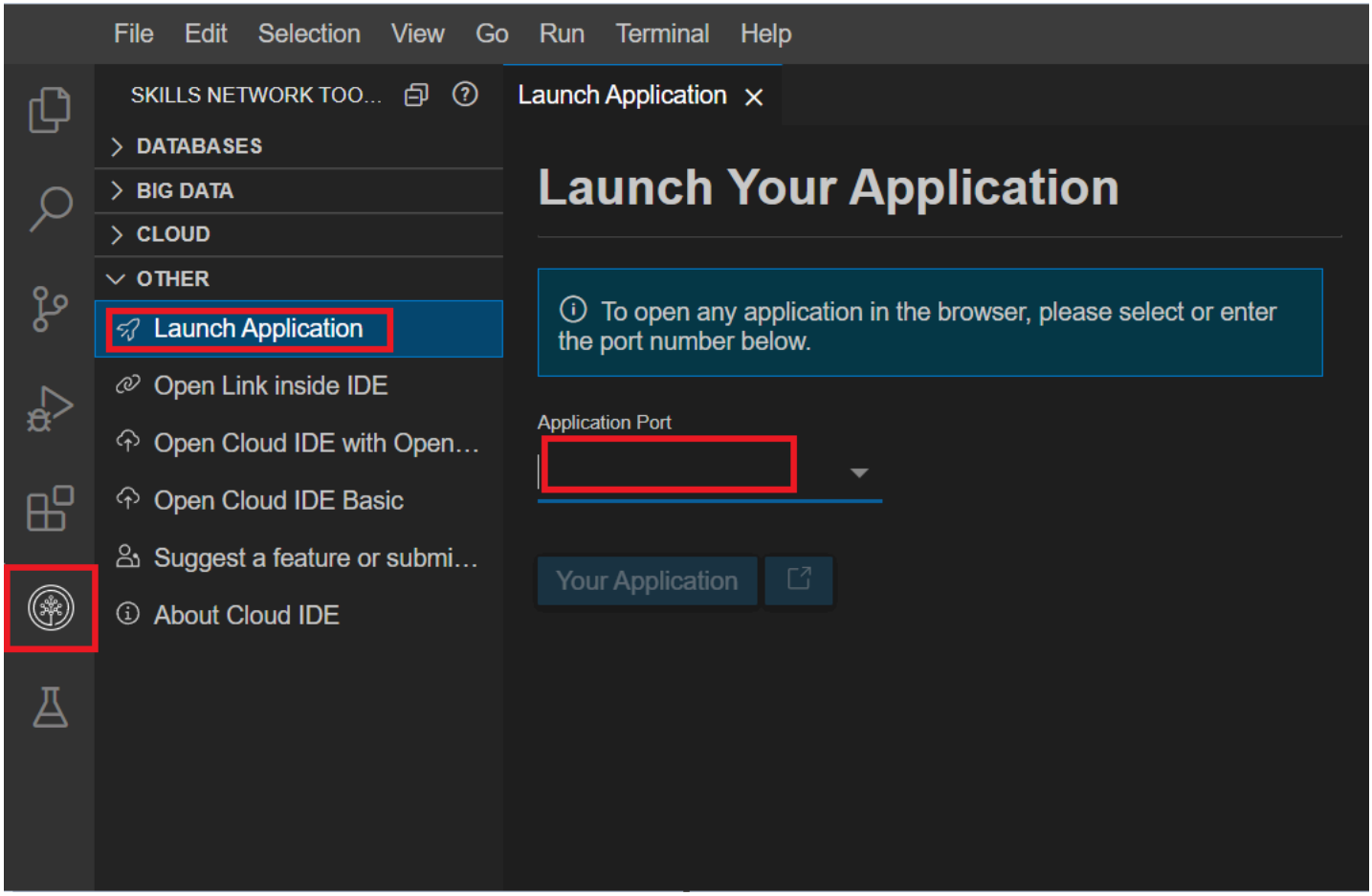
5. 通过查看文件内容检查文件是否已被复制到 HDFS。

```
hdfs dfs -cat /user/root/data.txt
```

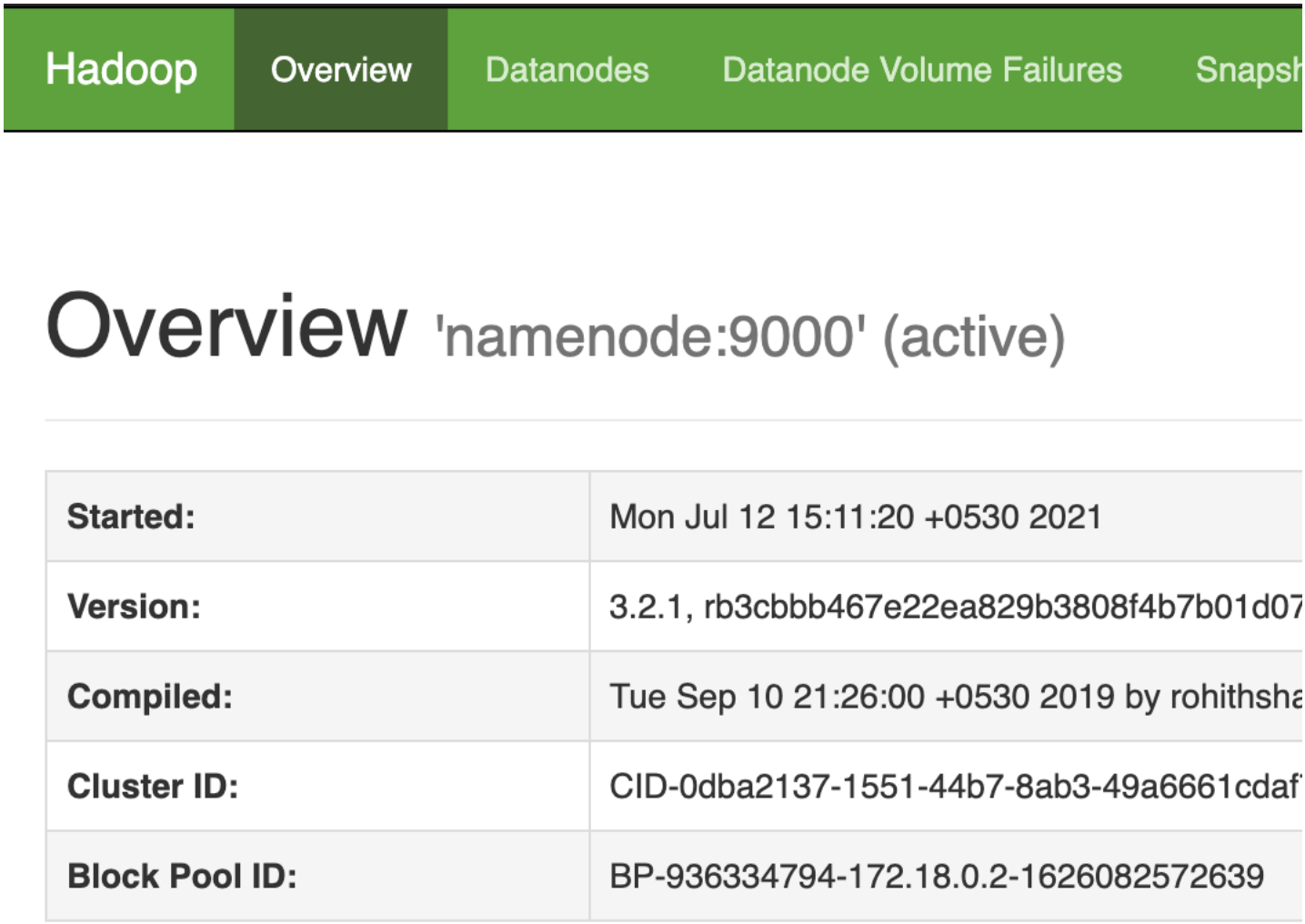
## 查看 HDFS

1. 点击下面的按钮或点击左侧的技能网络按钮，这将打开“技能网络工具箱”。然后点击其他选项，再点击启动应用程序。从那里你应该能够输入端口号 `9870` 并启动。

View HDFS




2. 这将打开 Hadoop 节点的图形用户界面 (GUI)。点击 `Utilities` -> `浏览文件系统` 来浏览文件。




3. 通过点击 `user` 然后 `root` 来查看你刚刚创建的目录中的文件。

# Browse Directory

/

Show 25 entries

| | Permission | Owner | Group | Size | Last Modified | Replic |
|---|---|---|---|---|---|---|
| | drwxr-xr-x | root | supergroup | 0 B | Nov 10 16:00 | 0 |
| | drwxr-xr-x | root | supergroup | 0 B | Nov 10 16:01 | 0 |

Showing 1 to 2 of 2 entries

Hadoop, 2019.

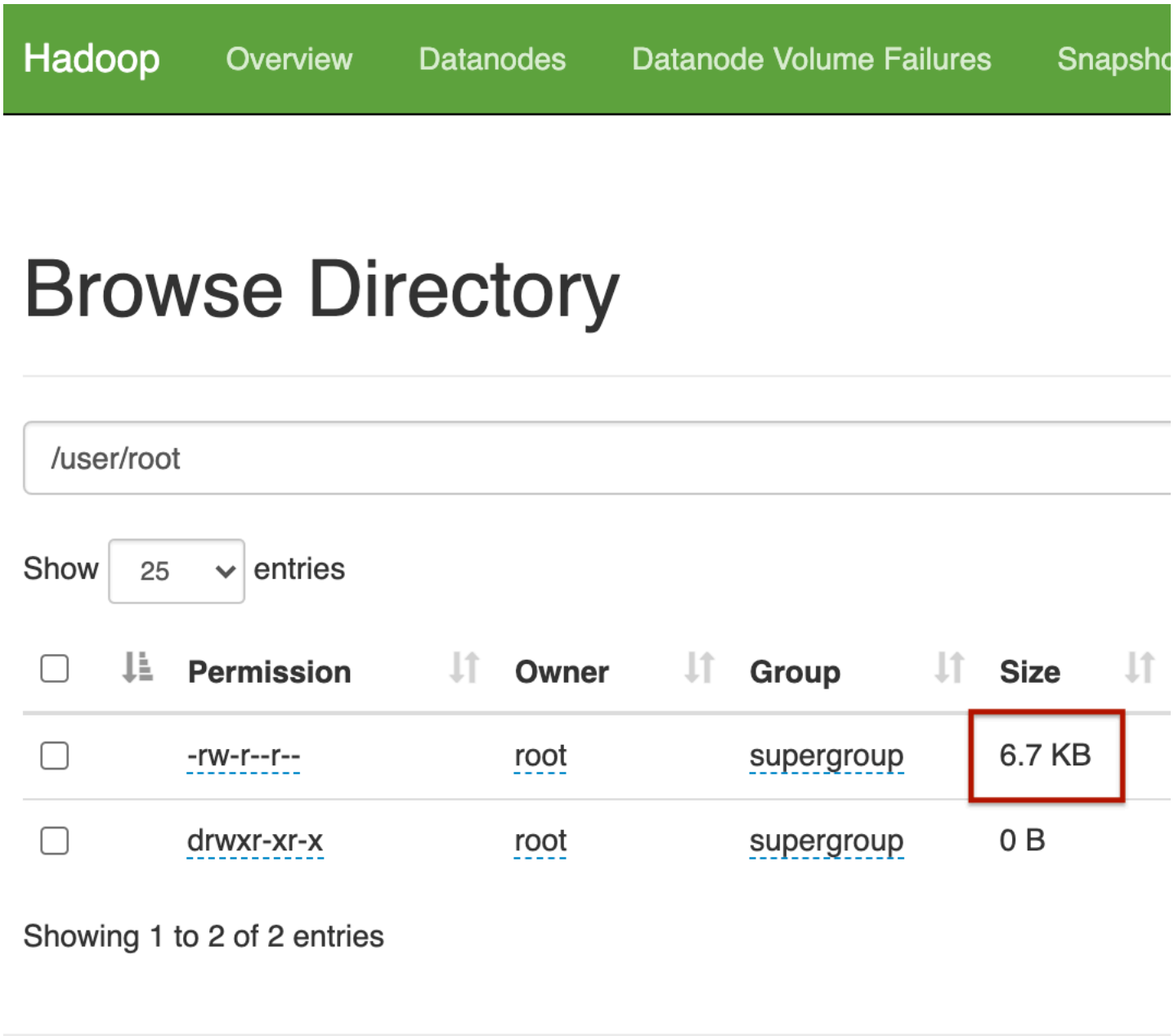
# Browse Directory

/user

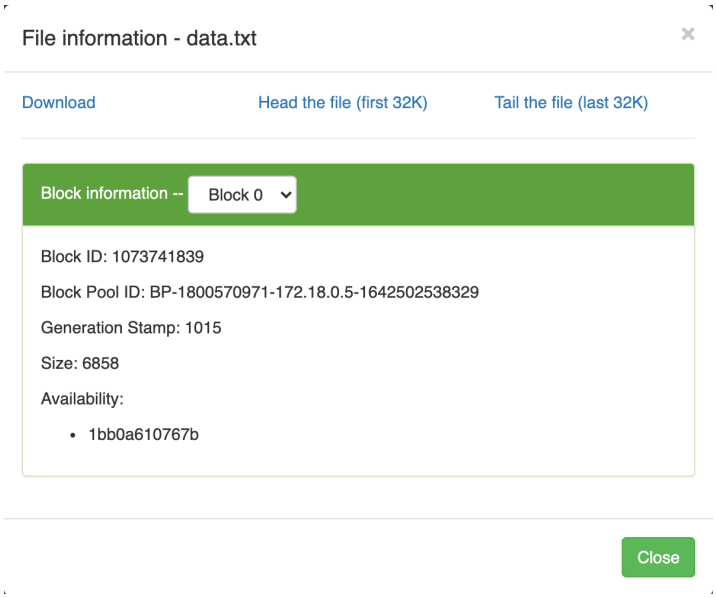
Show 25 entries

| | Permission | Owner | Group | Size | Last Modified | Replicatio |
|---|---|---|---|---|---|---|
| | drwxr-xr-x | root | supergroup | 0 B | Nov 10 16:01 | 0 |

Showing 1 to 1 of 1 entries




4. 注意到块大小为 128 MB，尽管文件大小实际上要小得多。这是因为 HDFS 使用的默认块大小为 128 MB。
5. 你可以点击文件查看文件信息。它会给你提供关于文件的字节数、块 ID 等信息。




**恭喜你！你已经：**

- 使用 Docker 部署了 Hadoop

- 在 HDFS 中创建了数据并在 GUI 上查看了它

发推特分享你的成就！

## Author(s)

Lavanya T S

**© IBM Corporation. 保留所有权利。**