Hadoop Map-Reduce 实操实验 (20 分钟)

目标

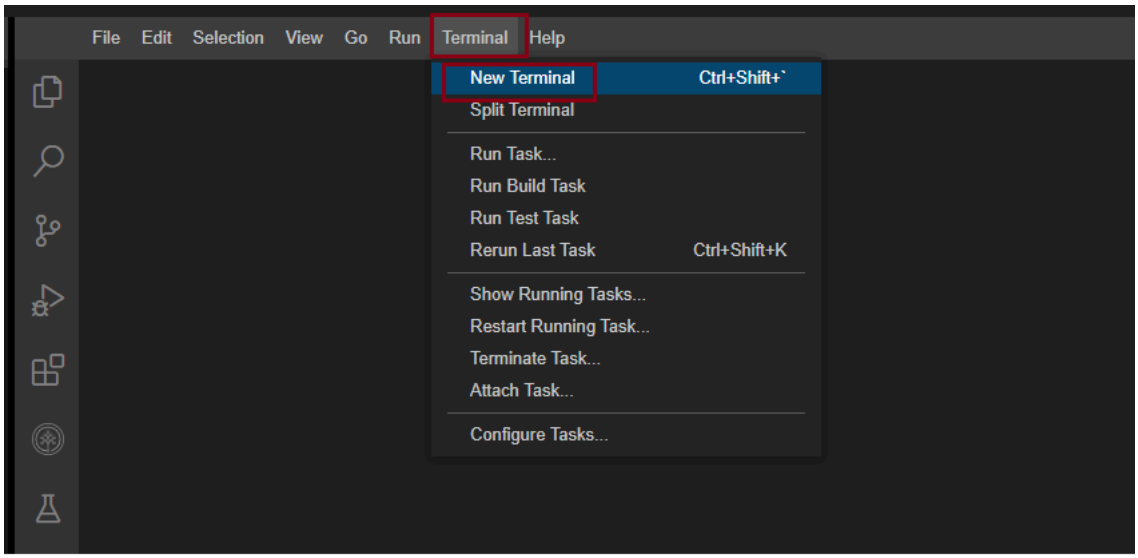
- 运行单节点 Hadoop 实例
- 使用 Hadoop **Map Reduce** 进行单词计数。

设置单节点 Hadoop

本实验中概述的步骤使用单节点 Hadoop 版本 3.3.6 **Hadoop** 在大型网络服务器集群上以完全分布式模式部署时最为有用，共享大量数据。然而，为了基本理解，我们将在单个节点上配置 Hadoop。

在本实验中，我们将运行 WordCount 示例，使用输入文本并查看输入文件的内容如何被 WordCount 处理。

1. 启动一个新终端



2. 通过运行以下命令将 hadoop-3.2.3.tar.gz 下载到你的 theia 环境中。

```
curl https://dlcdn.apache.org/hadoop/common/hadoop-3.3.6/hadoop-3.3.6.tar.gz --output hadoop-3.3.6.tar.gz
```

3. 在当前目录中提取 tar 文件。

```
tar -xvf hadoop-3.3.6.tar.gz
```

4. 导航到 hadoop-3.3.6 目录。

```
cd hadoop-3.3.6
```

5. 检查 hadoop 命令以确认其是否已设置。这将显示 hadoop 脚本的使用文档。

```
bin/hadoop
```

6. 运行以下命令将 data.txt 下载到您当前的目录。

```
curl https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-BD0225EN-SkillsNetwork/labs/data/data.txt --output data.
```

7. 在 data.txt 上运行 Map reduce 应用程序进行单词计数，并将输出存储在 **/user/root/output** 中。

```
bin/hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-examples-3.3.6.jar wordcount data.txt output
```

这可能需要一些时间。

8. 一旦字数统计成功运行，您可以运行以下命令查看生成的输出文件。

```
ls output
```

您应该看到 **part-r-00000** 和 **_SUCCESS**，这表明字数统计已完成。

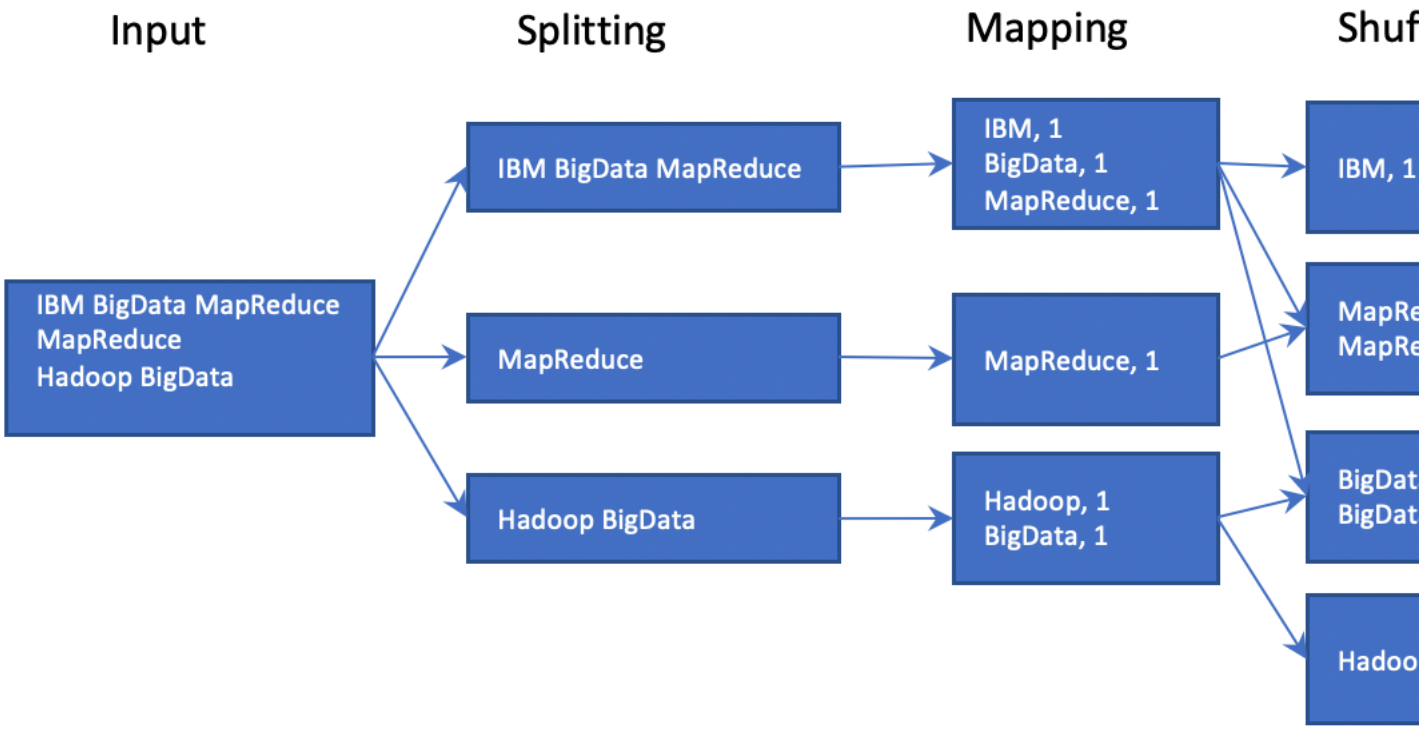
在处理过程中，您可能只会在输出目录中看到 *'_temporary'*。请等待几分钟，然后再次运行命令，直到您看到如上所示的输出。

9. 运行以下命令以查看字数统计输出。

```
cat  output/part-r-00000
```

```
theia@theiadocker-lavanyas:/home/project/hadoop-3.2.2$ c
BigData 2
Hadoop  1
IBM     1
MapReduce       2
```

下图展示了MapReduce字数统计的过程。



练习实验室

1. 对以下内容的文件进行字数统计。

```
Italy Venice
Italy Pizza
Pizza Pasta Gelato
```

▼ 点击这里获取关于如何开始的提示
– 删除 data.txt 文件和输出文件夹

```
rm data.txt
```

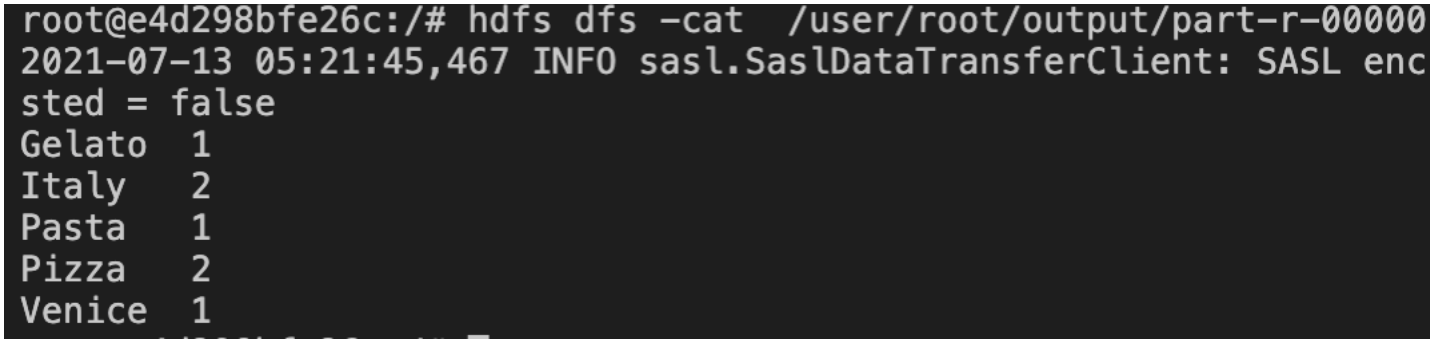
```
rm -rf output
```

▼ 点击这里获取关于如何创建文件以进行字数统计的提示
创建一个包含所需内容的 data.txt 文件。您可以使用文件编辑器。
▼ 点击这里获取关于如何对文件进行字数统计的解决方案
运行以下命令

```
bin/hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-examples-3.3.6.jar wordcount data.txt output
```

▼ 点击这里查看示例输出

输出将如下所示。



恭喜！您已经完成了：

- 使用 Docker 部署了 Hadoop
- 将数据复制到 HDFS
- 使用 MapReduce 进行字数统计

发推并分享您的成就！

作者

Lavanya T S

贡献者

Aije Egwaikhide

© IBM Corporation. 保留所有权利。