Cloudera 安装和应用指南

(一) Hadoop 预装映像

Hadoop 预装映像将 Hadoop 生态系统的相关软件预先组装在一起，发布成预装的软件映像产品。对于初学者来说，从头开始组装自己的软件堆栈可能会很混乱，需要做很多工作。设置整个堆栈的任务可能会消耗大量的项目时间和人力，增加了开发的时间。获取预构建的映像类似于购买预组装的家具，你可以获得一个现成的软件堆栈，其中包含预安装的操作系统、所需的库和应用程序软件。它使你避免了将不同部分以正确的方式放在一起的麻烦，可以立即开始使用。这些预构建的软件映像一般需要由使用虚拟化软件的虚拟机启用。

虚拟化软件的好处之一是它允许你在几分钟内就可以运行现成的软件堆栈，而无需过多细节。软件堆栈映像是一个大文件，虚拟化软件提供了一个可以运行软件堆栈映像的平台。

许多公司提供了 Hadoop 平台的软件堆栈映像，各自选择包括了许多 Hadoop 工具。Hortonworks 是为 Mac 和 Windows 平台提供预构建软件堆栈的公司之一。Cloudera 是另一家提供预安装和组装软件堆栈映像的公司。许多其他公司也提供类似的映像。此外，供应商网站上提供了许多面向初学者的在线教程，用于用户对使用这些映像及其包含的开源工具进行自我培训。

选择供应商后，你可以查看他们的网站以获取有关如何快速入门的教程，网上也有很多资源可以帮助你。可以选择云来部署预构建的映像，这将进一步加快应用程序部署的过程。在部署过程中，最好评估哪种方法对你的业务模型和组织最具成本效益。Cloudera，Hortonworks 等公司提供了有关如何在云上设置预构建映像的分步指南。总而言之，使用预构建的软件包具有许多好处，可以显着加速你的大数据项目。即使是小团队也可以快速制作原型，部署和验证他们的项目创意。开发的分析解决方案可以在几个小时内扩展到更大的规模并提高数据处理速度。这些公司还为大型成熟应用程序提供企业级解决方案。

因为有很多公司提供现成的 Hadoop 解决方案，这意味着你有很多备选项来从中选出最适合你的项目。

(二) 基于 Cloudera 的大数据应用实践

本节介绍如何在计算机上使用 Cloudera 映像进行 Hadoop 软件的部署，如何在 HDFS 中存储和复制数据，以及如何在 Hadoop 上运行你的第一个应用程序 WordCount 进行实践编程。

1. 下载和安装 Cloudera 虚拟机映像

本节以 Windows 操作系统为例来下载和安装 Cloudera 虚拟机映像，Mac 操作系统类似。请按照以下说明来下载并安装虚拟机 VirutalBox，并在此基础上下载和安装 Cloudera，来快速开启你的 Hadoop 编程之旅。后面章节中将指导如何使用 Cloudera 虚拟机环境来进

行 Hadoop 的实践操作。

学习目标

- 下载并安装 VirtualBox。
- 下载并安装 Cloudera 虚拟机（VM）映像。
- 启动 Cloudera 虚拟机。

硬件要求：

（A）四核处理器（推荐 VT-x 或 AMD-V），64 位;

（B）8GB 内存;

（C）20 GB 可用磁盘。

如何查找硬件信息：单击“开始”按钮打开“系统”，右键单击“计算机”，然后单击“属性”可查看硬件信息。最近几年购买的大多数计算机都支持 8GB 内存，可满足最低要求。计算机还需要具有高速互联网连接，因为需要下载多达 4 GB 大小的文件。

说明：

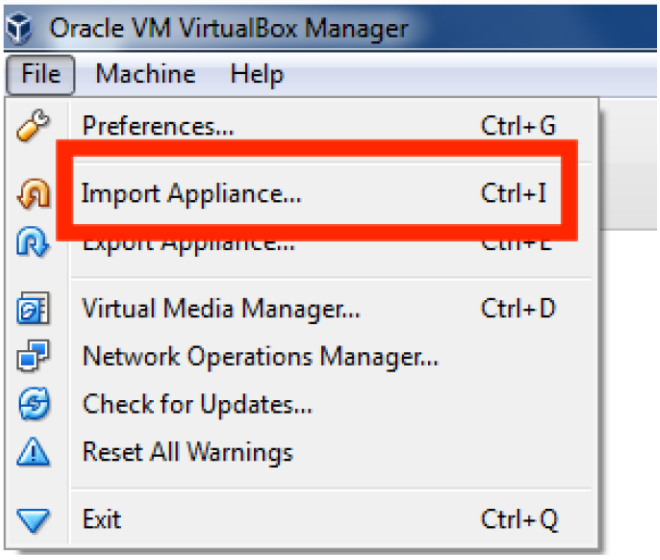
1.安装 VirtualBox。到网址 https://www.virtualbox.org/wiki/Downloads 中下载并安装 VirtualBox。本书使用 Virtualbox5.1.X 版本，因此，建议单击该网址页面上的“VirtualBox 5.1 builds”的相关链接，并下载旧版本的软件包以便于与本说明中的屏幕截图保持一致。然而，如果你选择使用新版本的 VirtualBox，应该不会有太大的不同。对于 Windows 操作系统，请选择“VirtualBox 5.1.X for Windows hosts x86/amd64”链接进行下载，其中“X”是代表最新版本的数字。

2. 下载 Cloudera VM。到网址 https://downloads.cloudera.com/demo_vm/virtualbox/cloudera-quickstart-vm-5.4.2-0-virtualbox.zip 中下载 Cloudera VM。虚拟机压缩文件超过 4GB，因此需要一些时间进行下载。

3. 解压 Cloudera VM。右键单击文件“cloudera-quickstart-vm-5.4.2-0-virtualbox.zip”进行解压。

4. 启动 VirtualBox。

5. 导入映像文件。通过 File -> Import Appliance 导入 Cloudera VM。



1　在 VirtualBox 中导入 Cloudera VM

6. 点击文件夹图标。

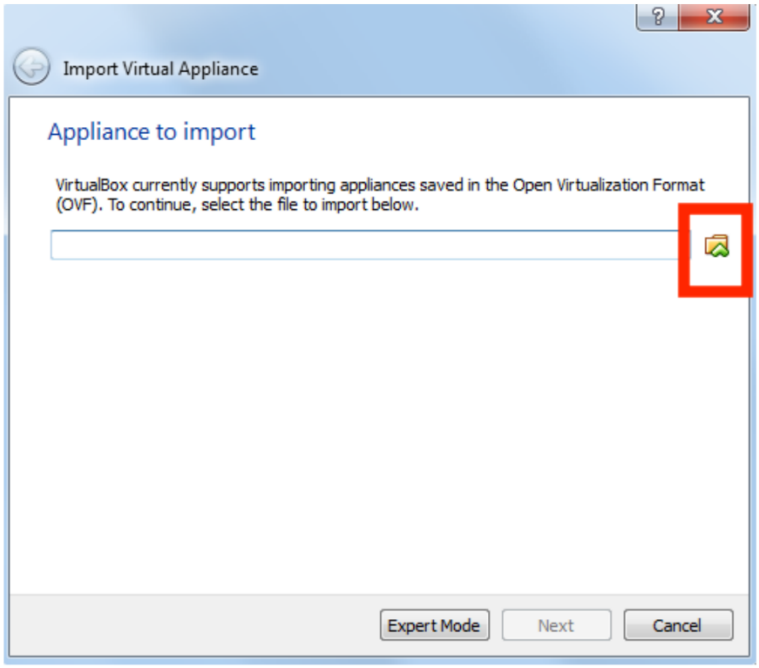


图 2　点击文件夹图标

7. 从你解压 VirtualBox VM 的文件夹中选择“cloudera-quickstart-vm-5.4.2-0-virtualbox.ovf”，并打开它。

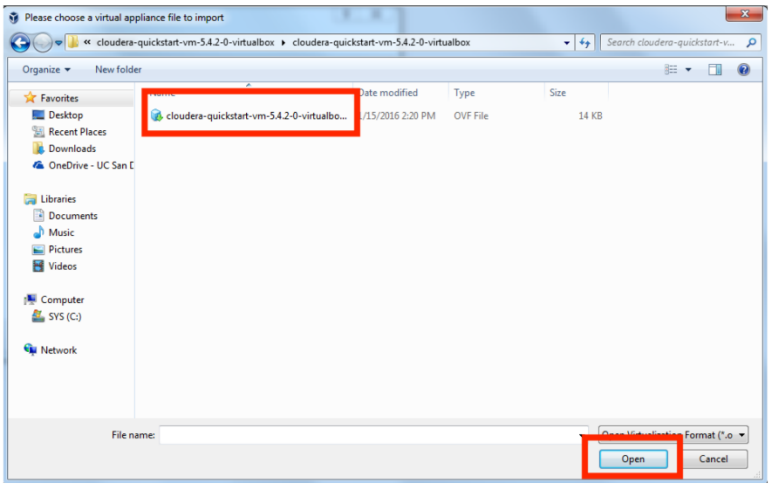


图 3　打开 ovf 文件

8. 点击“Next”。

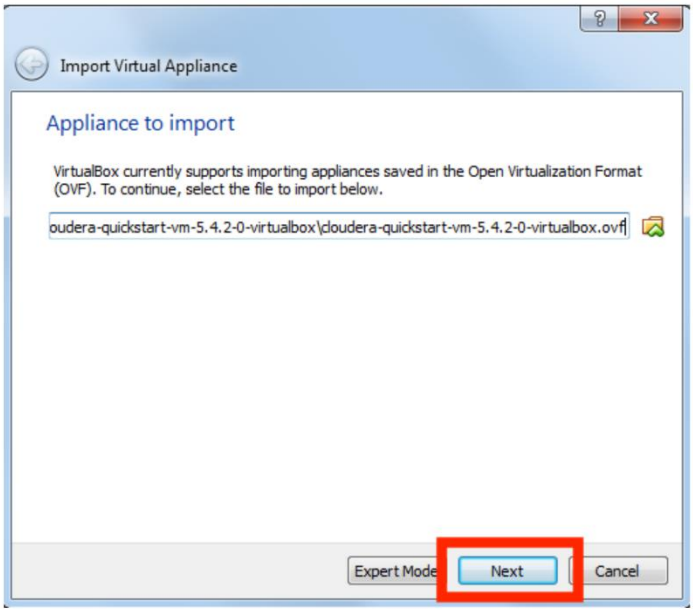


图 4　点击 Next

9. 点击 Import。

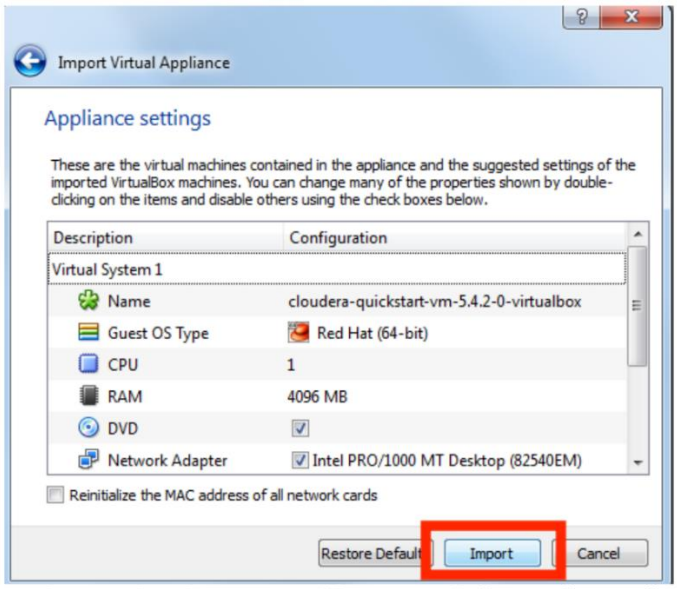


图 5 点击 Import

10. 将导入虚拟机映像，可能需要几分钟。

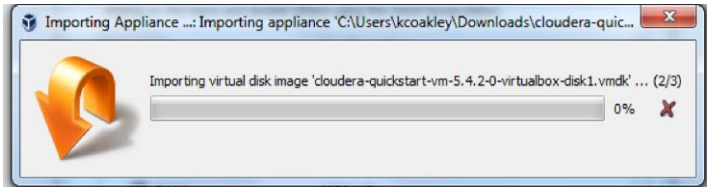


图 6　导入虚拟机映像

11.启动 Cloudera VM。导入完成后，quickstart-vm-5.4.2-0 虚拟机将出现在 VirtualBox 窗口的左侧。选择它并单击“Start”按钮启动虚拟机。

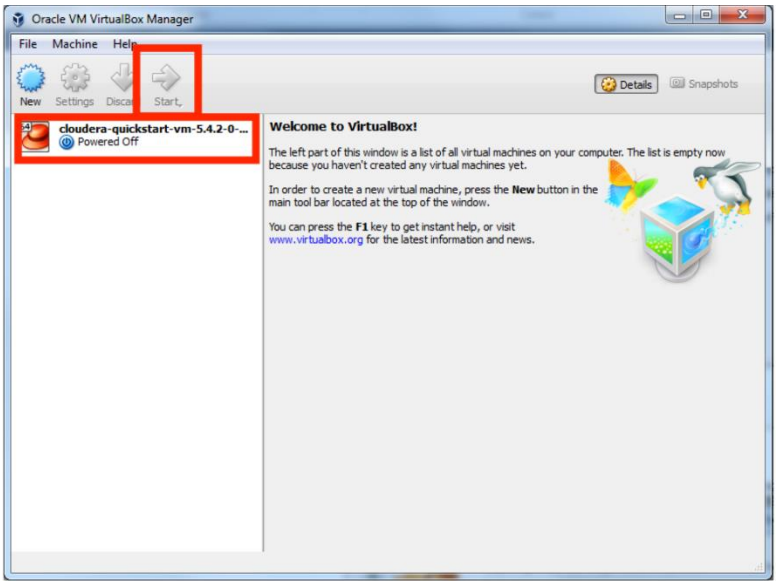


图 7　启动 Cloudera VM（1）

因为许多 Hadoop 工具需要启动，虚拟机将需要较长时间才能启动。

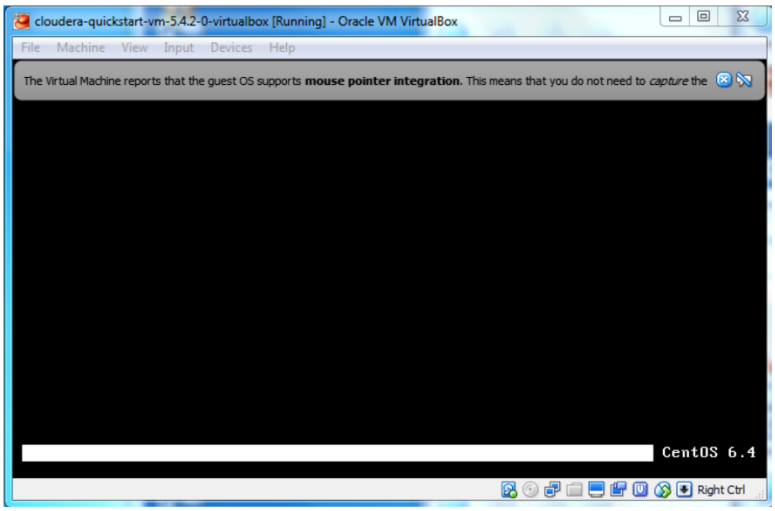


图 8　启动 Cloudera VM（2）

12. Cloudera VM 桌面。启动成功后，Cloudera VM 桌面将显示一个浏览器。

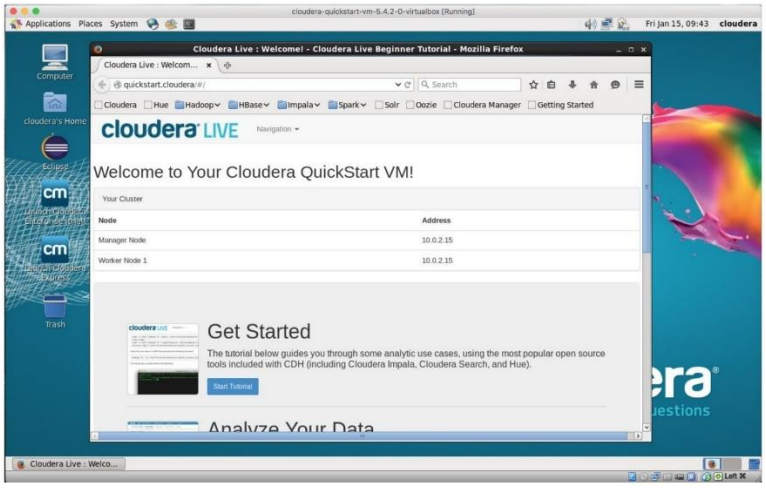


图 9　Cloudera VM 桌面

Cloudera VM 成功安装后，就可以基于此进行 Hadoop 的应用和编程实践了。

2. 将数据复制到 Hadoop 分布式文件系统中

学习目标

- 使用命令行应用程序与 Hadoop 交互。
- 将文件复制到 Hadoop 分布式文件系统（HDFS）中或从中移出。

说明：

1. 打开浏览器。

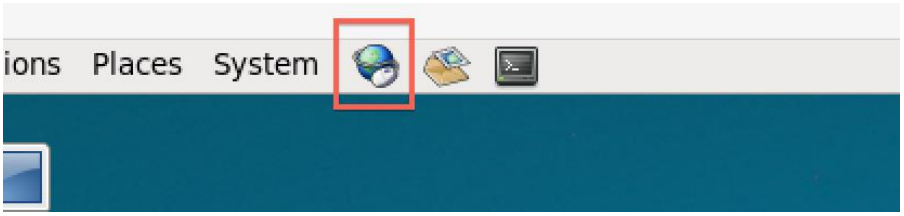


图 10 打开浏览器

2. 下载“t8.shakespeare.txt”（莎士比亚全集），作为练习文本文件，以复制到 HDFS 中。在浏览器中输入以下链接：

http://ocw.mit.edu/ans7870/6/6.006/s08/lecturenotes/files/t8.shakespeare.txt

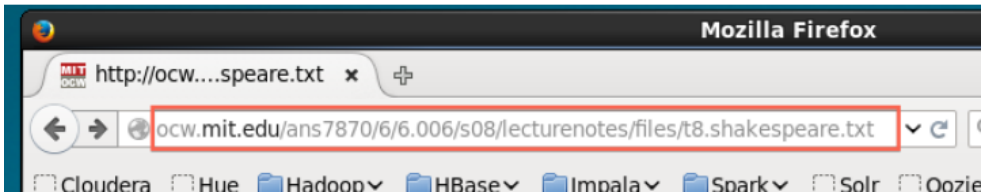


图 11 下载文本文件

加载页面后，单击"打开"菜单按钮。

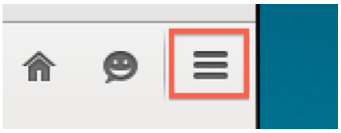


图 12 打开文本文件

点击“保存页面”。

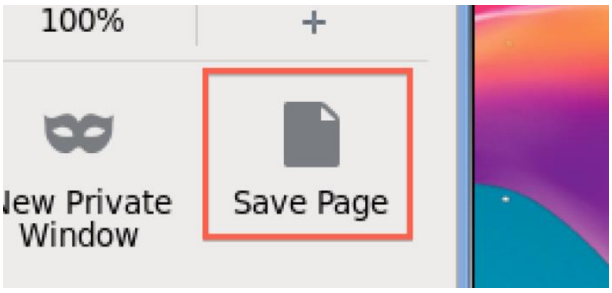


图 13 保存文本文件

将输出文件名更改为“words.txt”，然后单击保存。

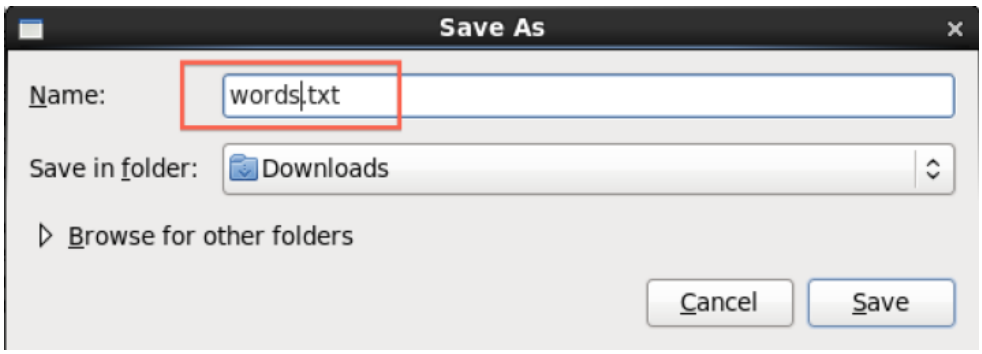


图 14　文件重命名

3. 打开终端 Shell。

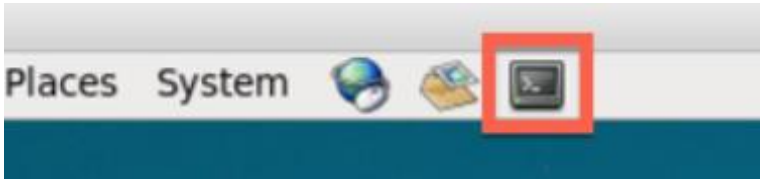


图 15　打开终端 Shell

在 Shell 中执行命令：*cd Downloads*，从当前目录切换为下载目录。

```
[cloudera@quickstart ~]$ cd Downloads/
[cloudera@quickstart Downloads]$
```

运行 *ls* 以查看 words.txt 是否已保存。

```
[cloudera@quickstart Downloads]$ ls
words.txt
```

4. 将文件从本地文件系统复制到 HDFS 中。运行 hadoop fs –copyFromLocal words.txt 将文本文件从本地文件系统复制到 HDFS 文件系统中。

```
[cloudera@quickstart Downloads]$ hadoop fs -copyFromLocal words.txt
[cloudera@quickstart Downloads]$
```

5. 验证文件是否已复制到 HDFS。运行 hadoop fs –ls 以验证文件是否已复制到 HDFS。

```
[cloudera@quickstart Downloads]$ hadoop fs -ls
Found 1 items
-rw-r--r--   1 cloudera cloudera    5458199 2016-02-12 15:14 words.txt
[cloudera@quickstart Downloads]$
```

6. 在 HDFS 中复制文件。可以在 HDFS 中创建文件的副本。运行 *hadoop fs -cp words.txt words2.txt* 来创建 words.txt 的副本并命名为 words2.txt ：

```
[cloudera@quickstart Downloads]$ hadoop fs -cp words.txt words2.txt
[cloudera@quickstart Downloads]$
```

可以通过运行 *hadoop fs -ls* 来查看复制的新文件

```
[cloudera@quickstart Downloads]$ hadoop fs -ls
Found 2 items
-rw-r--r--   1 cloudera cloudera    5458199 2016-02-12 15:14 words.txt
-rw-r--r--   1 cloudera cloudera    5458199 2016-02-12 15:15 words2.txt
[cloudera@quickstart Downloads]$
```

7. 从 HDFS 复制文件到本地文件系统。我们还可以将文件从 HDFS 复制到本地文件系统。运行 hadoop fs -copyToLocal words2.txt，将 words2.txt 复制到本地目录。

```
[cloudera@quickstart Downloads]$ hadoop fs -copyToLocal words2.txt
[cloudera@quickstart Downloads]$
```

然后，运行 *ls* 以查看 words2.txt 是否存在，以查看文件是否已成功复制到本地，。

```
[cloudera@quickstart Downloads]$ ls
words2.txt  words.txt
[cloudera@quickstart Downloads]$
```

8. 删除 HDFS 中的文件。运行 *hadoop fs -rm words2.txt* 在 HDFS 中删除 words2.txt。

```
[cloudera@quickstart Downloads]$ hadoop fs -rm words2.txt
16/02/12 15:17:01 INFO fs.TrashPolicyDefault: Namenode trash configuration: Dele
tion interval = 0 minutes, Emptier interval = 0 minutes.
Deleted words2.txt
[cloudera@quickstart Downloads]$
```

运行 hadoop fs -ls 以查看文件是否消失。

```
[cloudera@quickstart Downloads]$ hadoop fs -ls
Found 1 items
-rw-r--r--   1 cloudera cloudera    5458199 2016-02-12 15:14 words.txt
[cloudera@quickstart Downloads]$
```

3. 运行 WordCount 程序

学习目标

- 执行 WordCount 应用程序。
- 将 WordCount 的结果从 HDFS 中复制出来。

说明：

1.打开终端 Shell。在 VirtualBox 中启动 ClouderaVM（如果尚未运行），然后打开终端 Shell。有关这些步骤的详细说明，请参阅前面的章节。

2.运行系统自带的 MapReduce 示例程序，并查看用法。可以通过运行 *hadoop jar /usr/jars/hadoop-examples.jar* 来查看 Hadoop 示例程序的列表。我们可以看到将要运行的 WordCount。

```
[cloudera@quickstart ~]$ hadoop jar /usr/jars/hadoop-examples.jar
An example program must be given as the first argument.
Valid program names are:
  aggregatewordcount: An Aggregate based map/reduce program that counts the words in the in
put files.
  aggregatewordhist: An Aggregate based map/reduce program that computes the histogram of t
he words in the input files.
  bbp: A map/reduce program that uses Bailey-Borwein-Plouffe to compute exact digits of Pi.
  dbcount: An example job that count the pageview counts from a database.
  distbbp: A map/reduce program that uses a BBP-type formula to compute exact bits of Pi.
  grep: A map/reduce program that counts the matches of a regex in the input.
  join: A job that effects a join over sorted, equally partitioned datasets
  multifilewc: A job that counts words from several files.
  pentomino: A map/reduce tile laying program to find solutions to pentomino problems.
  pi: A map/reduce program that estimates Pi using a quasi-Monte Carlo method.
  randomtextwriter: A map/reduce program that writes 10GB of random textual data per node.
  randomwriter: A map/reduce program that writes 10GB of random data per node.
  secondarysort: An example defining a secondary sort to the reduce.
  sort: A map/reduce program that sorts the data written by the random writer.
  sudoku: A sudoku solver.
  teragen: Generate data for the terasort
  terasort: Run the terasort
  teravalidate: Checking results of terasort
  wordcount: A map/reduce program that counts the words in the input files.
  wordmean: A map/reduce program that counts the average length of the words in the input f
iles.
  wordmedian: A map/reduce program that counts the median length of the words in the input
files.
```

图 16　Hadoop 示例程序的列表

3.查看 WordCount 命令行参数。每个示例应用程序都可以在终端中运行。要查看如何运行特定的应用程序，可以将应用程序名称附加到上述命令后面。例如，要查看如何运行 WordCount，运行 *hadoop jar /usr/jars/hadoop-examples.jar wordcount* 即可。

```
[cloudera@quickstart Downloads]$ hadoop jar /usr/jars/hadoop-examples.jar wordcount
Usage: wordcount <in> [<in>...] <out>
[cloudera@quickstart Downloads]$
```

输出 WordCount 的使用方法中，<in>和<out>分别表示输入和输出的文件或目录名称。方括号表示第二个<in>或更多输入是可选的。该信息显示 WordCount 采用一个或多个输入文件的名称以及一个输出目录的名称。请注意，这些文件位于 HDFS 文件系统 中，而不是本地文件系统中。

4. 验证输入文件是否在 HDFS 中存在。之前我们下载了莎士比亚的完整作品，并命名为“words.txt”，并将它复制到了 HDFS 中。运行 *hadoop fs -ls,* 来确保此文件仍在 HDFS 中，以便我们可以在其上运行 WordCount。

```
[cloudera@quickstart Downloads]$ hadoop fs -ls
Found 1 items
-rw-r--r--   1 cloudera cloudera    5458199 2016-02-12 15:14 words.txt
[cloudera@quickstart Downloads]$
```

5. 运行 WordCount。运行 WordCount，查找 words.txt 中所有单词的出现次数: hadoop jar /usr/jars/hadoop-examples.jar wordcount words.txt out

```
[cloudera@quickstart Downloads]$ hadoop jar /usr/jars/hadoop-examples.jar wordcount words.txt out
16/02/12 15:27:34 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
16/02/12 15:27:35 INFO input.FileInputFormat: Total input paths to process : 1
16/02/12 15:27:35 INFO mapreduce.JobSubmitter: number of splits:1
```

当 WordCount 执行时，Hadoop 会打印 Map 和 Reduce 的执行进度。当字数计数完成时，两边都会显示 100%。

```
16/02/12 15:27:46 INFO mapreduce.Job:  map 0% reduce 0%
16/02/12 15:27:54 INFO mapreduce.Job:  map 100% reduce 0%
16/02/12 15:28:02 INFO mapreduce.Job:  map 100% reduce 100%
16/02/12 15:28:02 INFO mapreduce.Job: Job job_1455318527581_0001 completed successfully
```

6. 浏览 WordCount 输出目录。WordCount 完成后，验证是否已创建输出。首先，运行 *hadoop fs –ls*，查看输出目录 out 已经在 HDFS 中创建。

```
[cloudera@quickstart Downloads]$ hadoop fs -ls
Found 2 items
drwxr-xr-x   - cloudera cloudera          0 2016-02-12 15:28 out
-rw-r--r--   1 cloudera cloudera    5458199 2016-02-12 15:14 words.txt
[cloudera@quickstart Downloads]$ _
```

我们可以看到 HDFS 中现在有两个项目：words.txt 是我们之前创建的文本文件，out 是 WordCount 创建的目录。

7. 查看输出目录内部。WordCount 创建的 out 目录包含多个文件。通过运行 *hadoop fs -ls out* 查看目录内部：

```
[cloudera@quickstart Downloads]$ hadoop fs -ls out
Found 2 items
-rw-r--r--   1 cloudera cloudera          0 2016-02-12 15:28 out/_SUCCESS
-rw-r--r--   1 cloudera cloudera     717768 2016-02-12 15:28 out/part-r-00000
[cloudera@quickstart Downloads]$ _
```

其中，文件 part-r-00000 包含来自 WordCount 的统计结果。文件_SUCCESS 表示 WordCount 已成功执行。

8. 将 WordCount 结果从 HDFS 复制到本地文件系统。通过运行 *hadoop fs –copyToLocal out/part-r-00000 local.txt* 将 part-r-00000 复制到本地文件系统，并重命名为“local.txt”。

```
[cloudera@quickstart Downloads]$ hadoop fs -copyToLocal out/part-r-00000 local.txt
[cloudera@quickstart Downloads]$
```

9. 查看 WordCount 结果。查看结果的内容：*more local.txt*

```
[cloudera@quickstart Downloads]$ more local.txt
```

结果文件的每一行都显示输入文件中某个单词的出现次数。例如，" Accuse "在输入中出现四次，但" Accusing "仅出现一次:

```
Accost- 1
Account 1
Accountant      1
Accounted       1
Accoutred       1
Accurs'd        2
Accurs'd,       1
Accursed        4
Accusativo,     2
Accuse  4
Accusing        1
Acheron 2
Acheron,        1
Aches   1
Achiev'd        1
```