和其他 Shell 工具不一样的是，在其他 Shell 工具中，你只能使用单机的硬盘和内存来操作数据，而 spark-shell 可用来与分布存储在多台机器上的内存或者硬盘上的数据进行交互，并且处理过程的分发由 Spark 自动控制完成，不需要用户参与。

4.2.1 spark-shell 命令

在 Linux 终端中运行 spark-shell 命令，就可以启动进入 spark-shell 交互式执行环境。spark-shell 命令及其常用的参数如下：

```
$ ./bin/spark-shell  --master  <master-url>
```

Spark 的运行模式取决于传递给 SparkContext 的<master-url>的值。<master-url>可以是表 4-1 中的任一种形式。

表 4-1　　spark-shell 命令中的<master-url>参数及其含义

<master-url>	含义
local	使用一个 Worker 线程本地化运行 Spark（完全不并行）
local[*]	使用与逻辑 CPU 个数相同数量的线程来本地化运行 Spark（“逻辑 CPU 个数”等于“物理 CPU 个数”乘以“每个物理 CPU 包含的 CPU 核数”）
local[K]	使用 K 个 Worker 线程本地化运行 Spark（理想情况下，K 应该根据运行机器的 CPU 核数来确定）
spark://HOST:PORT	Spark 采用独立（Standalone）集群模式，连接到指定的 Spark 集群，默认端口是 7077
yarn-client	Spark 采用 YARN 集群模式，以客户端模式连接 YARN 集群，集群的位置可以在 HADOOP_CONF_DIR 环境变量中找到；当用户提交了作业之后，不能关掉 Client，Driver Program 驻留在 Client 中，负责调度作业的执行；该模式适合运行交互类型的作业，常用于开发测试阶段
yarn-cluster	Spark 采用 YARN 集群模式，以集群模式连接 YARN 集群，集群的位置可以在 HADOOP_CONF_DIR 环境变量中找到；当用户提交了作业之后，就可以关掉 Client，作业会继续在 YARN 上运行；该模式不适合运行交互类型的作业，常用于企业生产环境
mesos://HOST:PORT	Spark 采用 Mesos 集群模式，连接到指定的 Mesos 集群，默认接口是 5050

在 Spark 中采用 Local 模式启动 spark-shell 的命令主要包含以下参数：

- master：这个参数表示当前的 spark-shell 要连接到哪个 Master，如果是 local[*]，就是使用 Local 模式（单机模式）启动 spark-shell，其中，中括号内的星号表示需要使用几个 CPU 核心（Core），也就是启动几个线程模拟 Spark 集群；
- jars：这个参数用于把相关的 JAR 包添加到 CLASSPATH 中；如果有多个 jar 包，可以使用逗号分隔符连接它们。

比如，要采用 Local 模式，在 4 个 CPU 核心（Core）上运行 spark-shell，命令如下：

```
$ cd  /usr/local/spark
$ ./bin/spark-shell  --master local[4]
```

或者，可以在 CLASSPATH 中添加 code.jar，命令如下：

```
$ cd  /usr/local/spark
$ ./bin/spark-shell  --master  local[4]  --jars  code.jar
```

可以执行“spark-shell --help”命令，获取完整的选项列表，具体如下：

```
$ cd /usr/local/spark
$ ./bin/spark-shell  --help
```