

Lessons

Hadoop comes with several MapReduce applications. In the Cloudera VM, these applications are in `/usr/jars/hadoop-examples.jar`. You can see a list of all the applications by running `hadoop jar /usr/jars/hadoop-examples.jar`.

```
[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 input files.
  aggregatewordhist: An Aggregate based map/reduce program that computes the histogram of the 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 counts 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 files.
  wordmedian: A map/reduce program that counts the median length of the words in the input files.
```

Each of these MapReduce applications can be run in the terminal. To see how to run a specific application, append the application name to the end of the command line. For example, to see how to run `wordcount`, run `hadoop jar /usr/jars/hadoop-examples.jar wordcount`.

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

The output tells you how to run `wordcount`:

```
1  Usage: wordcount <in> [<in>...] <out>
```

The `<in>` and `<out>` denote the names of the input and output, respectively. The square brackets around the second `<in>` mean that the second input is optional, and the `...` means that more than one input can be used.

This usage says that wordcount is run with one or more inputs and one output, the input(s) are specified first, and the output is specified last.

✓ Complete

