Map Reduce and Design Patterns
Lecture 6

Fang Yu

Software Security Lab.
Department of Management Information Systems
College of Commerce, National Chengchi University
http://soslab.nccu.edu.tw

Cloud Computation, April 21, 2015

Meta-Patterns

Patterns that deal with patterns

- job chaining: piecing together several patterns to solve complex, multistage problems
- job merging: optimization for performing several analytics in the same MapReduce job, effectively killing multiple birds with one stone

Job Chaining

With the driver: have a master driver that simply fires off multiple job-specific drivers

- Take the driver for each MapReduce job and call them in the sequence they should run
- Be sure that the output path of the first job is the input path of the second
- The temporary directories should be cleaned up
- You can also fire off multiple jobs in parallel by using `Job.submit()` instead of `Job.waitForCompletion` . The submit method returns immediately to the current thread and runs the job in the background.

Job Chaining

Do it in a shell script: each job in the chain is fired off separately in the way you would run it from the command line from inside of a shell script

- The master driver is a scripting language instead of Java
- Upside: Use jobs that have already been productionalized to work through a command- line interface. The shell script can interact with services, systems, and tools that are not Java centric
- Downside: it may be harder to implement more complicated job flows in which jobs are running in parallel. You can run jobs in the background and then test for success, but it may not be as clean as in Java.

Implementation

- *Oozie*, an open source Apache project, has functionality for building workflows and coordinating job running
- `JobControl` and `ControlledJob` classes

Chain Folding

Folding the chain to combine map phases to mapper or dredger

- Each record can be submitted to multiple mappers, or to a reducer and then a mapper
- Push the processes that decrease the amount of data into the previous reducer, while keeping the processes that increase the amount of data where they are
- Save a lot of time reading files and transmitting data
- `ChainMapper` and `ChainReducer` are special mapper and reducer classes that allow you to run multiple map phases in the mapper and multiple map phases after the reducer.

Job Merging

A process that allows two unrelated jobs that are loading the same data to share the MapReduce pipeline

- The data needs to be loaded and parsed only once.
- Tag, if-statement, multiple outputs.
- The more the merrier?! When the jobs are merged, theyll have to run together and the source code will have to be kept together