CS425, Distributed Systems: Fall 2015
Machine Programming 1 – Distributed Log Querier

Yunsheng Wei (wei29) Neha Chaube (nchaub2)

The Distributed Log Querier consists of four main components:-

Class	Description
Catalog	Stores system information including host information (host name, IP address, port number), encoding, etc., and can be extended in future projects.
RemoteGrepClient	Sends `grep' command to servers, and output matched lines received from servers to standard output.
Server	Server is responsible for accepting query from client. For this project, it only supports `grep' command, but can be easily extended for future projects.
Grep	Main workhorse that Implements `grep' functionality along with some of its options, it also supports wildcard in file pattern.

We design the project in a Client-Server model. The RemoteGrepClient will spawn several threads, one for sending `grep' command to each server, and once accepts the socket, the server will spawn a new thread to execute the `grep' command, and sends the results back to client.

The RemoteGrepClient and Server are both deployed in all the machines. The Client program can run the grep command from any of the machines and query the logs present on all machines and display the results on its terminal.

Apache Commons CLI library is used for parsing command line options passed to `grep'. Json-simple library is used to encode and decode command line options of `grep' before sending to and receiving from socket, in case the pattern string has quote or space.
The whole MP is implemented in Java.

Test

DistributedGrepTest.java contains test cases that test whether the distributed grep result is the same as local grep result. It orders the result of distributed grep result and strips corresponding headers, and then compare it with local grep result line by line.

Performance

170 MB log files split across 7 servers

Pattern	Frequent pattern (“HTTP”)	Somewhat frequent (“edu|com”)	Rare pattern (“tia1\.eskimo\.com”)
Avg latency	18s	8s	1.5s

Catalog

Catalog stores information about the whole system. Its inner class Host stores information about all the machines present in the distributed system including the host name, IP address and port number. It also stores the encoding of the while system(“UTF-8”).
The details of the machines are stored in “conf/host_list” file.

RemoteGrepClient

RemoteGrepClient is a client program which sends `grep' command to servers, and outputs matched lines received from servers to standard output. The usage of RemoteGrepClient is exactly the same as grep, except that at least one file should be given (i.e. cannot read from standard input, which does not make sense in distributed settings). Any invalid grep command will be detected here, so it will never send invalid grep commands to servers.

Server

This is the main class for Server. Every machine in system runs an instance of Server class. Whenever a request is received by the server, the Server accepts the connection and spawns a new thread (an instance of ServerThread) to handle the request. This worker thread is responsible for handling the request from the client. The result of the grep operation is sent back to the client over the socket.

Grep

Grep is a utility which is like UNIX command: grep. This is the main class which implements the grep functionality. It can be used both as an API and a command line tool.
The command line syntax for performing grep is: **java Grep [-options] [pattern] [file ...]**
As standard grep, the [file] can have wildcard in it.