Pipeline Architecture

CSSE6400

Brae Webb

March 7, 2022

So far...

Layered architectures reduce the impact of changing a layer

Why does the layer order matter?

Why does the layer order matter?

Answer

Each layer implements a different interface.

If every layer implements the same interface?

So...

Extreme layered architecture

Pipeline Architectures¹

Definition 1. Pipeline Architecture

Components connected in such a way that the output of one component is the input of another.

Can you think of a *pipeline archiecture*?

Can you think of a pipeline archiecture?

Answer

How about bash?

```
>> cat assignment.py | grep "hack" | wc -l \
| tee code-quality.txt
```

```
>> cat assignment.py | grep "hack" | wc -l \
| tee code-quality.txt
```

Notice:

• Each program performs a small well-defined task.

```
>> cat assignment.py | grep "hack" | wc -l \
tee code-quality.txt
```

Notice:

- Each program performs a small well-defined task.
- Each program implements the same interface (i.e. raw text).

wc -l

tee code-quality.txt

grep "hack"

cat assignment.py





Filters

Modular software components



Filters

Modular software components

Pipes

The flow of data between filters

Producers

Source of data.

Producers

Source of data.

Transformers

Transform data.

Producers

Source of data.

Transformers

Transform data.

Testers

Filter data.

Producers

Source of data.

Transformers

Transform data.

Testers

Filter data.

Consumers

Target for results.

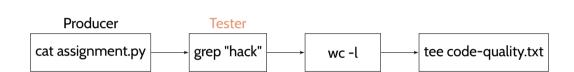
Exercise

Label the bash pipeline.

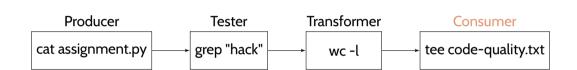


Producer

cat assignment.py grep "hack" wc -l tee code-quality.txt







Definition 2. One Direction Principle

Data should flow in one direction, this is the *down-stream*.

Definition 3. Independent Filter Principle

Filters should not rely on specific upstream or downstream components.

Corollary 1. Generic Interface

The interface between filters should be generic.

Corollary 2. Composable Filters

Filters (i.e. Transformers & Testers) can be applied in any order.

The Case Study

Bash

What does **BASH** stand for?

POSTX-2017

Shell & Utilities

Utilities

admin - create and administer SCCS files (DEVELOPMENT) ex - text editor expand - convert tabs to spaces alias - define or display aliases ar - create and maintain library archives expr - evaluate arguments as an expression asa - interpret carriage-control characters false - return false value at - execute commands at a later time fc - process the command history list awk - pattern scanning and processing language fg - run jobs in the foreground file - determine file type basename - return non-directory portion of a pathname batch - schedule commands to be executed in a batch queue find . find files fold . filter for folding lines bc - arbitrary-precision arithmetic language bg - run jobs in the background fort77 - FORTRAN compiler (FORTRAN) c99 - compile standard C programs fuser - list process IDs of all processes that have one or more files cal - print a calendar cat - concatenate and print files gencat - generate a formatted message catalog cd - change the working directory getconf - get configuration values cflow - generate a C-language flowgraph (DEVELOPHENT) get - get a version of an SCCS file (DEVELOPMENT) charp - change the file group ownership getopts - parse utility options chand - change the file modes gren - search a file for a pattern chown - change the file ownership hash - remember or report utility locations cksum . write file checksums and sizes head - copy the first part of files cmp - compare two files icony - codeset conversion command - execute a simple command id - return user identity comm - select or reject lines common to two files ipcrm - remove an XSI message queue, semaphore set, or shared memory compress - compress data segment identifier incs - report XSI interprocess communication facilities status cp - copy files jobs - display status of jobs in the current session crontab - schedule periodic background work ioin - relational database operator ceplit - enlit files based on context ctags - create a tags file (DEVELOPMENT, FORTRAN) kill - terminate or signal processes cut - cut out selected fields of each line of a file lex - generate programs for lexical tasks (DEVELOPMENT) link - call link function cycef - generate a C-language program cross-reference table In - link files (DEVELOPMENT) localedef - define locale environment date - write the date and time dd - convert and copy a file locale - get locale-specific information logger - log messages delta - make a delta (change) to an SCCS file (DEVELOPMENT) logname - return the user's login name df - report free disk space diff - compare two files lp - send files to a printer dirname - return the directory portion of a pathname ls - list directory contents du - estimate file space usage m4 - macro processor echo - write arguments to standard output mailx - process messages ed - edit text make - maintain, update, and regenerate groups of programs env - set the environment for command invocation (DEVELOPHENT)

Who has heard of *literate programming*?

The Challenge — set by Jon Bently

- 1. Read a file of text.
- 2. Determine the n most frequently used words.
- 3. Print out a sorted list of those words along with their frequencies.

Knuth's Solution

17 pages of elegant and descriptive code.

hy lon Bentley with Special Guest Ousters Don Knuth and Doug Mclirou

programming pearls

A LITERATE PROGRAM

Last monthly column introduced from Knothly stale of "Literate Programming" and his WEB system for building programs that are works of literature. This column precents a literate program by Vouth fits prining ore sketched in last month's column) and, as befits literature, a review. So without further ado, here is Knuth's progress retypeset in Communications style.

Common Words															Section								
Introduction																							
Strategic considerations																					. 1	8	
Basic input routines																					. 1	y	
Dictionary lookup																					.13	7	
The frequency counts .																					. 32	2	
Sorting a trie																							
The endgame																					.43	1	
Index																,					. 42	2	

- 1. Introduction. The purpose of this program is to solve the following problem posed by Ion Bentley:
- Given a text file and an integer k, print the k most common words in the file (and the number of their occurrences) in decreasing frequency.
- Ion intentionally left the problem community vague, but he stated that "a user should be able to find the 100 most frequent words in a twenty-page technical paper (muchly a 50K byte file) without undue emotional trauma
- Let us agree that a word is a sequence of one or more continuous letters: "Bent Lev" is a word, but "ain't" isn't. The sequence of letters should be maximal, in the cence that it cannot be lengthened without including a nonletter. Uppercase letters are considered equivalent to their lowercase counterparts, so that the words "Bentley" and *DERTLEY" and "bent lov" are exentially identical.

The given problem still isn't well defined, for the file might contain more than & words, all of the same fractioners or there might not even be as many as k words. Let's be more precise: The most common woods are to be printed in order of decreasing frequarter with words of aqual fraquency listed in alphabetic order. Printing should stop after k words have been output, if more than it words are present.

- 2. The input file is assumed to contain the given text. If it begins with a positive decimal number (preceded by optional blanks), that number will be the value of k otherwise we shall assume that k = 100. Answers will be sent to the output file.
- define default it as 100 lose this value if it isn't otherwise specified!
- 3. Bestdes solving the given problem, this program is supposed to be an example of the WEB system, for noonle who know some Parcal but who have never seen NEB before. Here is an outline of the program to be constructed:
- program common mords (input, output): type (Type declarations 17) var (Global variables 4)
- (Procedures for initialization s) (Procedures for input and output a) (Procedures for data manipulation 20) begin (The main program a):
- 4. The main idea of the WEB approach is to let the program grow in natural stages, with its parts preconted in roughly the order that they might have been written by a programmer who isn't consciolly clairvovant
- For example, each clobal variable will be introduced when we first know that it is necessary or desirable: the WEB system will take care of collecting these declarations into the proper place. We already know about one global variable, namely the number that Bantley called k Let us give it the more descriptive name may words to print.

Programming Pearls

(Global variables 4) =

may mards to print; interer

See also sections 11, 13, 18, 22, 32, and 36.

This code is used in section 3

a. As we introduce new global variables, we'll often want to give them certain starting values. This will be done by the initialize procedure, whose body will consist of various pieces of code to be specified when we think of particular kinds of initialization.

let most this many words will be printed!

(Procedures for initialization 5) = procedure initialize.

var i: integer: fall-purpose index for initializa-

begin (Set initial values 12)

This code is used in section 3

6. The WEB system, which may be thought of as a preprocessor for Pascal, includes a macro definition facility so that portable programs are easier to write For example, we have already defined 'default_k' to he 100. Here are two more examples of UPB macros they allow us to write, e.g., 'incricount[p])' as a convenient abbreviation for the statement 'coverfet -count[v] + 1'

define
$$incr(\#) = \# \leftarrow \# + 1$$
 (increment a variable)
define $dcr(\#) = \# \leftarrow \# - 1$ [decrement a variable]

7. Some of the procedures we shall be writing come to abrunt conclusions; hence it will be convenient to introduce a 'return' macro for the operation of jumping to the end of the procedure. A symbolic label 'erit' will be declared in all such procedures. and 'exite' will be placed just before the final and (No other labels or sole statements are used in the present program, but the author would find it pain. ful to eliminate these particular cons.)

define exit = 30 | (the end of a procedure) define return mente exit louisk termination!

8. Strategic considerations. What algorithms and data structures should be used for Bentley's prob lem? Clearly we need to be able to recognize different occurrences of the same word, so some sort of internal dictionary is necessary. There's no obvious way to decide that a particular word of the input cannot possibly be in the final set, until we've sotten very near the end of the file; so we might as well remember every word that appears

There should be a frequency count associated with each word, and we will eventually want to run through the words in order of decreasing frequency. But there's no need to keen these counts in order as we read through the input since the order matters only at the and

Therefore it makes sense to structure our program as follows:

(The main program s) =

initialize-

(Establish the value of our annals to exist up) (Input the text, maintaining a dictionary with

frequency counts 2(3) (Sort the dictionary by frequency 39):

(Output the results 41)

This code is used in partion 3.

9 Basic input routines | Let's switch to a hottomun approach now by writing some of the procedures.

that we know will be necessary sooner or later. Then we'll have some confidence that our program is taking shape, even though we haven't decided yet how to handle the searching or the sorting. It will be nice to get the messy details of Pascal input out of the way and off our minds Hore's a function that reads an antional positive

integer, returning zero if none is present at the beginning of the current line

```
(Procedures for input and output a) #
function read into interers
 begin n e o
 if not then
   begin while (\neg coin) \land (innut \uparrow = ' \cdot \cdot ') do
     ort(innut):
   while (innut t > '0') A (innut t < '9') do
     begin a ser litter & ordinantiti = ordifati
     ostlinnet):
     end.
   end:
 read_int - n
```

end: See also sections 15, 35, and 40. This code is used in section 2

10. We invoke reed intonly once

(Retablish the value of may words to print to) = max_words_to_print + read_int: if may mords to print = 0 then max_words_to_print - default_k

This roule is used in section 8.

11 To find words in the innut file we want a quick way to distinguish letters from nonletters. Pascal has

June 1986 Volume 29 Number 6

McIlroy's Solution

3 sort |
4 uniq -c |
5 sort -rn |
6 sed \${1}q

```
tr -cs A-Za-z '\n'
tr A-Z a-z |
```

Is literate programming bad?

Is literate programming bad?

Answer

No, the unix philosophy is just good.

The Unix Philosophy

 Write programs that do one thing and do it well.

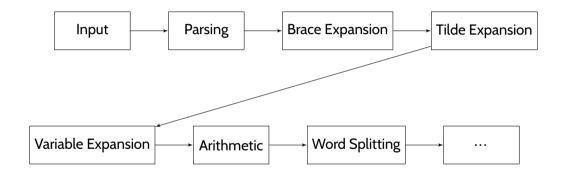
The Unix Philosophy

- Write programs that do one thing and do it well.
- Write programs to work together.

The Unix Philosophy

- Write programs that do one thing and do it well.
- Write programs to work together.
- Write programs to handle text streams, because that is a universal interface.

Bash itself is a pipeline



Case Study

Compilers

Case Study

MapReduce

What's the advantage of the map reduce pattern?

What's the advantage of the map reduce pattern?

Answer

Parallelism [1]

Using pipeline terminology, what filters do the *map* and *reduce* operators correspond to?

References

[1] Jeffrey Dean and Sanjay Ghemawat.

Mapreduce: Simplified data processing on large clusters.

In OSDI'04: Sixth Symposium on Operating System Design and Implementation, pages 137–150, San Francisco, CA, 2004.