ECE4721J - Homework 1

Methods and Tools for Big Data

Kexuan Huang 518370910126

May 30, 2022

Ex. 1 - Processes, cgroups, and namespaces

1. Write a short summary describing what cgroups are.

Control groups, usually referred to as cgroups, are a Linux kernel feature which allow processes to be organized into hierarchical groups whose usage of various types of resources can then be limited and monitored. The kernel's cgroup interface is provided through a pseudo-filesystem called cgroupfs. Grouping is implemented in the core cgroup kernel code, while resource tracking and limits are implemented in a set of per-resource-type subsystems (memory, CPU, and so on). [a]

[a] Linux manual page

2. Explain the differences and similarities between cgroups and processes in Linux.

Similarity: cgroup is a collection of processes.
Difference: cgroups limits, accounts for, and isolates the resource usage of a collection of processes.[a]

[a] Wikipedia

3. How does kernel namespace increase the security of the OS?

Namespaces are a feature of the Linux kernel that partitions kernel resources such that one set of processes sees one set of resources while another set of processes sees a different set of resources. Examples of such resources are process IDs, hostnames, user IDs, file names, and some names associated with network access, and interprocess communication.[a]

The key feature of namespaces is that they isolate processes from each other. On a server where you are running many different services, isolating each service and its associated processes from other services means that there is a smaller blast radius for changes, as well as a smaller footprint for security-related concerns.[b]

[a] Wikipedia
[b] Nginx

Ex. 2 — Increasingly large dataset

Please refer to `hw_1_code/ex2/README.md`

1. Basic hardware profile.

a) What CPU does your computer have?

2.7 GHz Quad-Core Intel Core i7

b) How much RAM does your computer have?

16 GB 2133 MHz LPDDR3

c) Explain how you will monitor the RAM and CPU usage in the following questions.

Command `top` and `htop`

2. Determine the following information:

a) Which carrier is most commonly late?

```
Most commonly late carrier: DL
Late count: 8064705 times
```

b) Which are the three most commonly late origins, due to bad weather?

```
DFW delays 72276 times
ATL delays 58137 times
ORD delays 57754 times
```

c) What is the longest delay experienced for each carrier?

```
US: 1646
WN: 883
NW: 2601
PA (1): 1070
TW: 1086
UA: 1437
DL: 1439
HP: 1309
ML (1): 472
AA: 1521
AS: 1140
CO: 1187
OH: 1242
OO: 996
XE: 927
TZ: 1173
EV: 1200
FL: 1345
HA: 1317
MQ: 1710
B6: 1048
DH: 1050
PI: 1418
PS: 569
EA: 1380
F9: 899
YV: 715
9E: 1956
AQ: 1021
```

3. Can you discover any pattern explaining departure delays?

Solution:

For `2008.csv.bz2`, the multiple linear regression gives:

$$\begin{aligned}\mathbf{DepDelay} = &\ 0.25038515 * \mathbf{DayOfWeek} \\ &+ 0.08558365 * \mathbf{DepTime} \\ &- 0.07303154 * \mathbf{CRSDepTime} \\ &- 0.01676935 * \mathbf{ArrTime} \\ &+ 0.01328631 * \mathbf{CRSArrTime}\end{aligned}$$

Ex. 3 — Very basic Java

1. Given a text file where each line is composed of three fields, first-name, name, and email, write a very short and simple Java program generating a text file where (i) the order of the lines is random and (ii) each line is composed of the previous fields in the following order: name, first-name, and email.

> Please refer to `hw_1_code/ex3_1/README.md`

2. Use inheritance an polymorphism to define various types of vehicles owned by a company. The definition of the actual objects is left to your imagination. Write a short program to demonstrate your work.

> Please refer to `hw_1_code/ex3_2/README.md`