Processing Text and Binary Files

Dr. Chris Brown

In This Lesson

Reading binary files

Logging

Processing log files

Modifying Configuration Files

Reading Binary Files

Reading Binary Files – The Basics

```
file = open("datafile", "rb")

data = file.read(20)
```

Open for binary reading

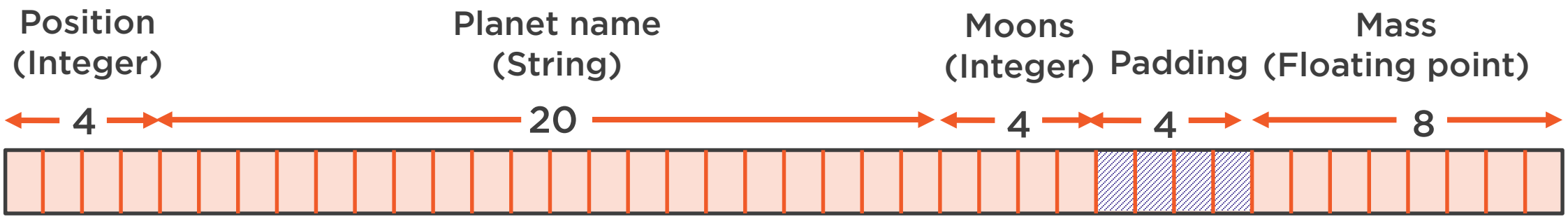
Returns a
bytes object

Read 20 bytes
Default: read to end of file

Planetary Record Format

"C" structure

```
struct planet {
    int position;
    char name[20];
    int moons;
    double mass;
};
```



Writing the Data (in "C")

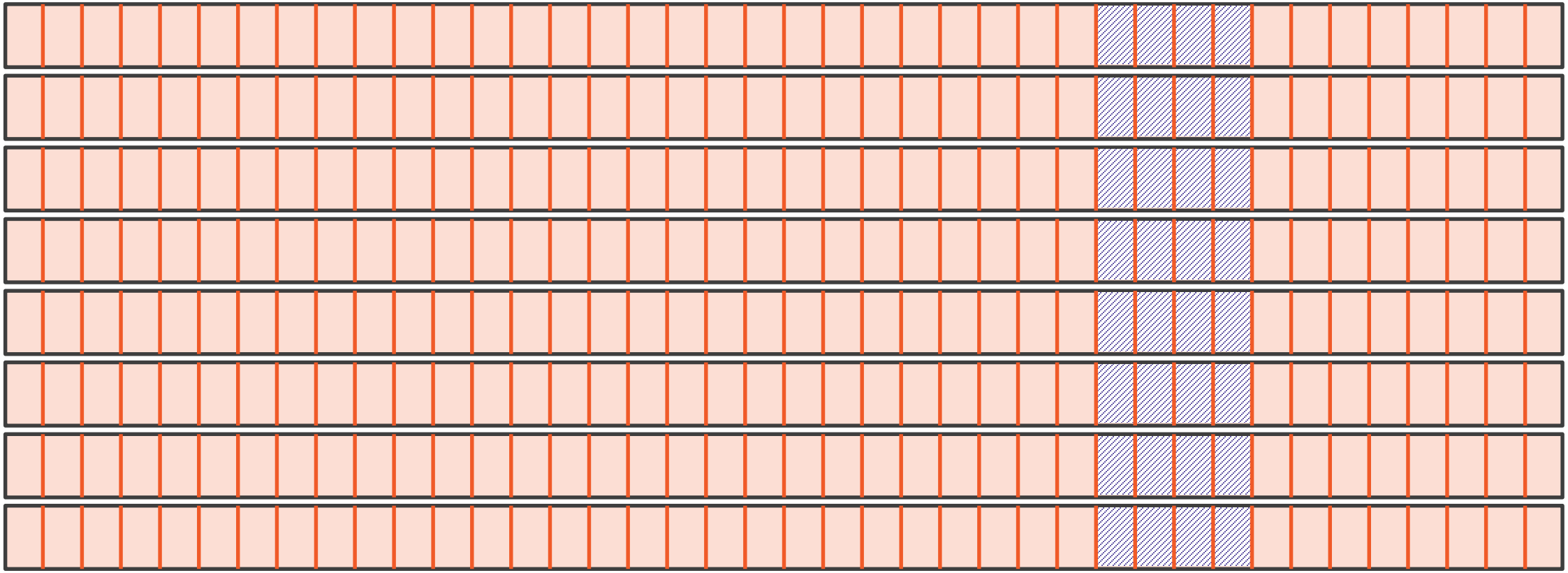
```
int main()
{
    struct planet solar_system[] = {
        { 1, "Mercury", 0,  0.06 },
        { 2, "Venus",   0,  0.82 },
        { 3, "Earth",   1,  1.00 },
        { 4, "Mars",    2,  0.11 },
        { 5, "Jupiter", 67, 317.8 },
        { 6, "Saturn",  62, 95.0 },
        { 7, "Uranus",  27, 14.5 },
        { 8, "Neptune", 14, 17.0 }
    };

    FILE * fd;
    int nplanets = sizeof solar_system / sizeof (struct planet);

    fd = fopen("planets.dat", "wb");

    fwrite(solar_system, sizeof(struct planet), nplanets, fd);
    fclose(fd);
    return 0;
}
```

The Whole File



File size = 8 x 40 = 320 bytes

Unpacking the Structure

```
import struct

tuple = struct.unpack(format, bytes)
```

Returns a tuple of unpacked values

Format string specifying the structure of the data

The byte sequence to be unpacked

Unpack Format Codes

Format	C Type	Python type	Standard size
`i`	`int`	integer	4
`I`	`unsigned int`	integer	4
`l`	`long`	integer	4
`L`	`unsigned long`	integer	4
`q`	`long long`	integer	8
`Q`	`unsigned long long`	integer	8
`n`	`ssize_t`	integer	
`N`	`size_t`	integer	
`e`	(7)	float	2
`f`	`float`	float	4
`d`	`double`	float	8
`s`	`char[]`	bytes	

Unpacking the `planet` Record

```
file = open("planets.dat", "rb")
content = file.read(40)
pos, name, moons, mass = struct.unpack("@i20sid", content)
```

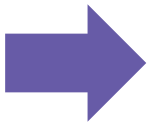
Logging

When to `print()` and When to Log

Short-running interactive script

`print()` **to** `stdout` **or** `stderr`

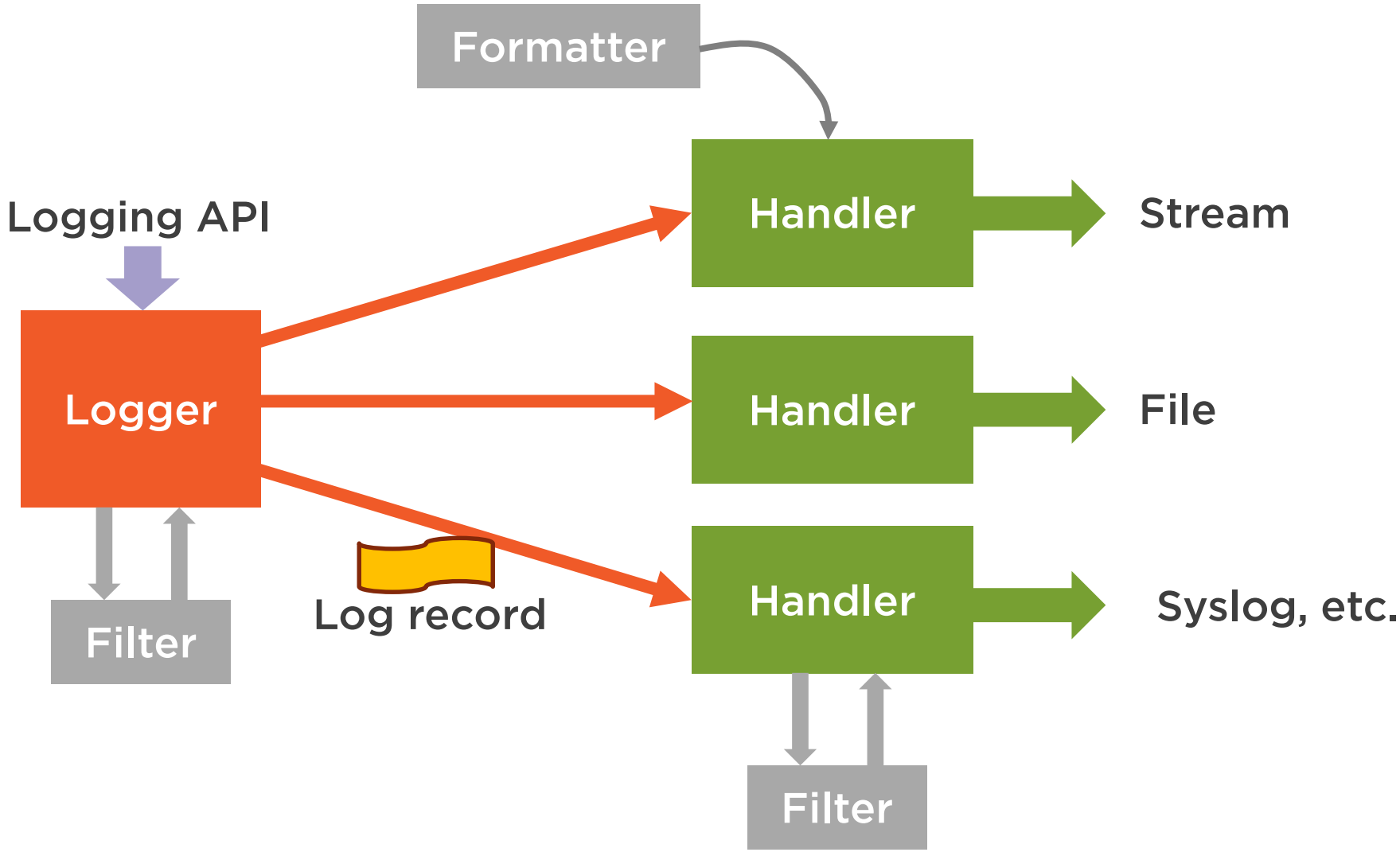
Long-running background service



`log events to:`

- `file`
- `syslog`
- `systemd journal`

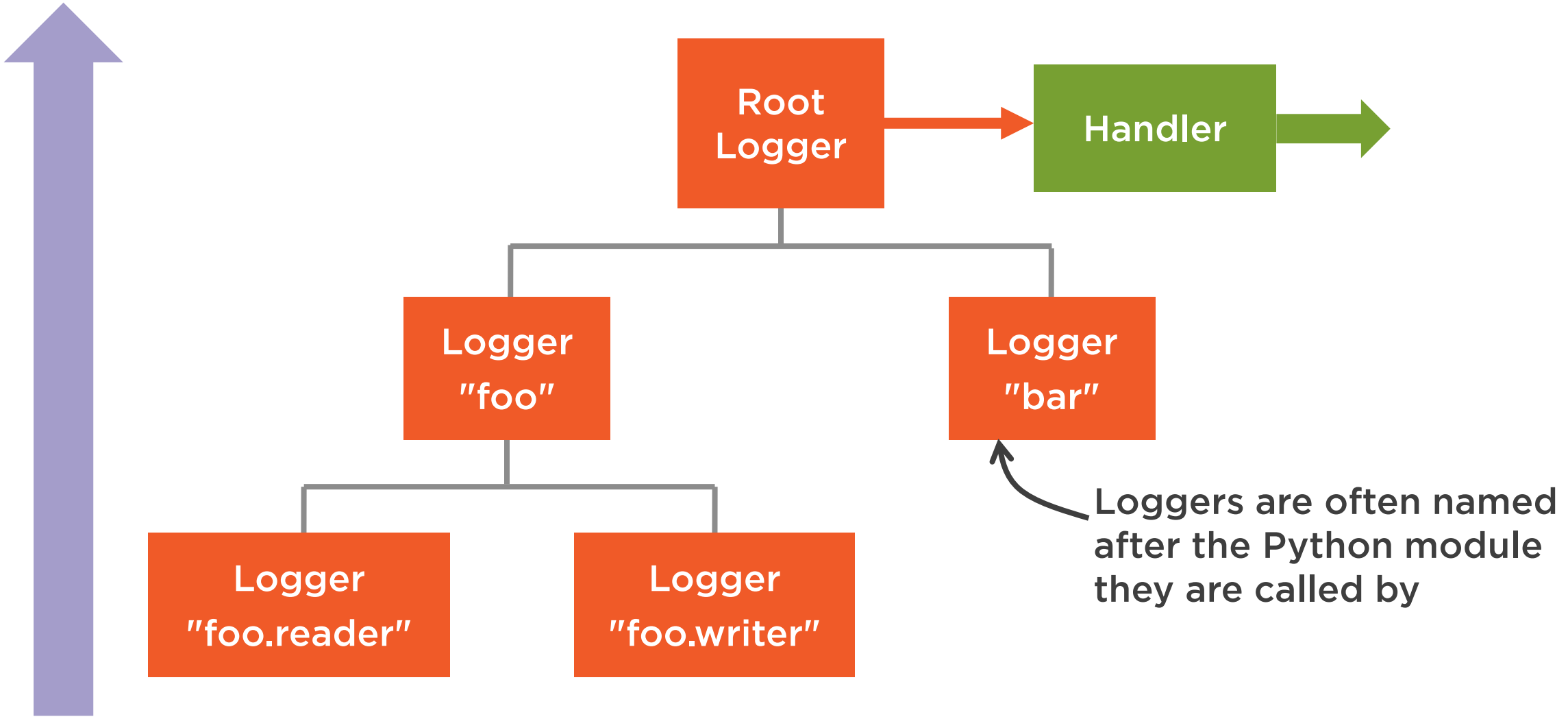
Logging Architecture



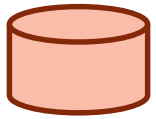
Log Record Attributes

Attribute name	Format
args	You shouldn't need to format this yourself.
asctime	`%(asctime)s`
created	`%(created)f`
exc_info	You shouldn't need to format this yourself.
filename	`%(filename)s`
funcName	`%(funcName)s`
levelname	`%(levelname)s`
levelno	`%(levelno)s`
lineno	`%(lineno)d`
module	`%(module)s`
msecs	`%(msecs)d`
message	`%(message)s`

Logger Hierarchy



Logging Handlers

StreamHandler	Logs to a stream e.g. `stdout`, `stderr`
FileHandler	Logs to a file 
WatchedFileHandler	Useful for files that are periodically "rotated" by an external program (e.g. `logrotate`)
SysLogHandler	Logs to a local or remote `syslog` daemon
SMTPHandler	Logs to an email address 
SocketHandler	Connects to a specified TCP host/port
JournalHandler	Logs to the `systemd` journal

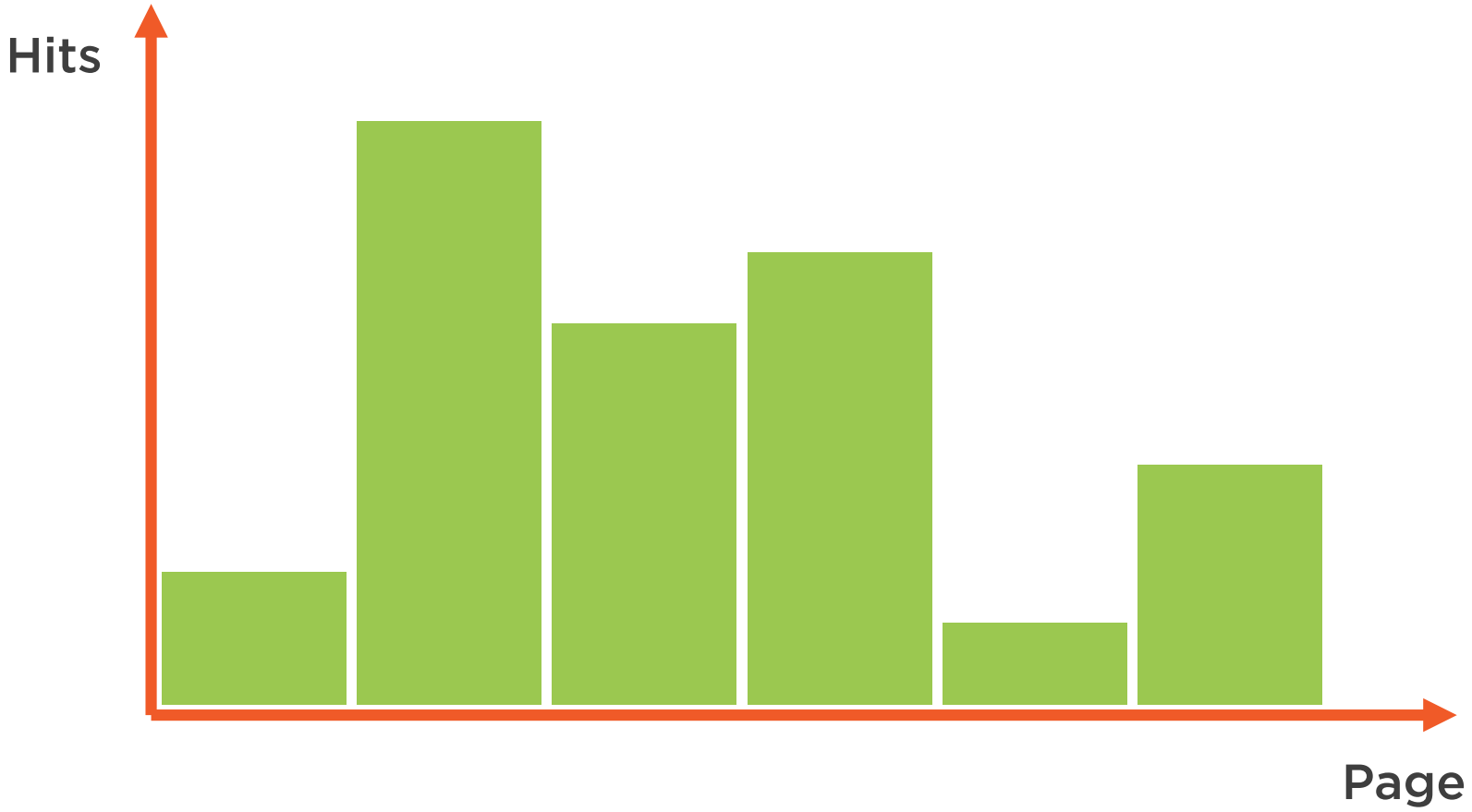
Logging Levels

Level	Usage
NOTSET	This logger's level is unspecified
DEBUG	Detailed output intended to verify correct operation of the code. Not for production use
INFO	A significant but normal event; e.g. service started or stopped, new client connected.
WARNING	A potential problem; e.g. low disk space, insecure configuration choice
ERROR	Something definitely wrong, but not terminal
CRITICAL	A serious problem that may force the program to abort

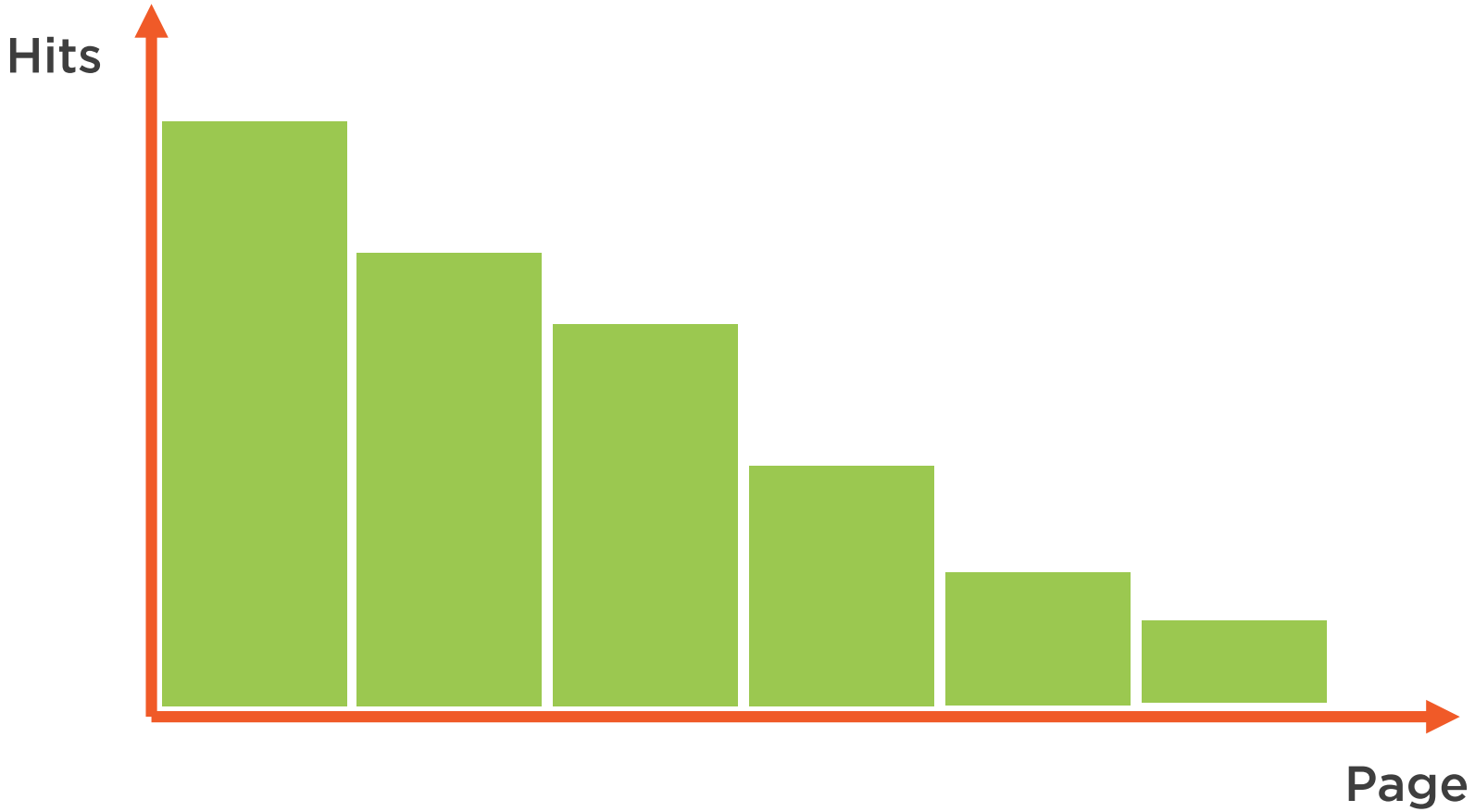
Increasing Severity

Processing an Apache Access Log

Building the Histogram



Sorting the Histogram



Apache Server Access Log

```
51.255.65.16 - - [09/Oct/2016:04:38:34 +0100] "GET /ccp51/cgi-bin/cp-
app.cgi?rnd=3012635&rrc=N&affl=&cip=163.172.66.109&act=&aff=&pg=track
HTTP/1.1" 200 21764 "-" "Mozilla/5.0 (compatible; AhrefsBot/5.1;
+http://ahrefs.com/robot/)"
180.76.15.144 - - [09/Oct/2016:04:38:42 +0100] "GET /ccp51/cgi-bin/cp-
app.cgi?usr=51F139517&rrc=N&affl=&cip=&act=&aff=&catstr=HOME:tin_can_tools
&pg=ste_cat&ref=tin_can_tools HTTP/1.1" 200 25992 "-" "Mozilla/5.0
(compatible; Baiduspider/2.0; +http://www.baidu.com/search/spider.html)"
51.255.65.93 - - [09/Oct/2016:04:38:48 +0100] "GET /ccp51/cgi-bin/cp-
app.cgi?rnd=3554092&rrc=N&affl=&cip=164.132.161.21&act=&aff=&pg=privacy
HTTP/1.1" 200 23525 "-" "Mozilla/5.0 (compatible; AhrefsBot/5.1;
+http://ahrefs.com/robot/)"
54.86.105.131 - - [09/Oct/2016:04:39:38 +0100] "GET
/linux_kernel_internals_and_device_driver_programming.php HTTP/1.1" 404
332 "-" "User-Agent: Mozilla/5.0 (Windows NT 6.1; WOW64)
AppleWebKit/537.31 (KHTML, like Gecko) Chrome/26.0.1410.64 Safari/537.31"
```

Dissecting the Access Log

```
86.4.46.202 - - [09/Oct/2016:04:42:08 +0100] "GET /contactus.php HTTP/1.1" 200 7779
```

```
splitline = line.split()
url = splitline[6]
```

```
url = line.split()[6]
```

Removing the Query String

```
163.172.66.161 - - [09/Oct/2016:04:40:26 +0100] "GET /cgi-bin/cp-app.cgi?rnd=84427
```

```
url = line.split()[6].split("?")[0]
```

Building the Histogram

```
hist = dict()
for line in open("access_log"):
    url = line.split()[6].split('?')[0]
    hist[url] += 1
```

```
hist = dict()
for line in open("access_log"):
    url = line.split()[6].split('?')[0]
    if url in hist:
        hist[url] += 1
    else:
        hist[url] = 1
```

Asking Forgiveness not Permission

```
hist = dict()
for line in open("access_log"):
    url = line.split()[6].split('?')[0]
    try:
        hist[url] += 1
    except KeyError:
        hist[url] = 1
```

Using the `defaultdict` Class

```
import collections

hist = collections.defaultdict(int)
for line in open("access_log"):
    url = line.split()[6].split('?')[0]
    hist[url] += 1
```

Sorting the Histogram

```
sorted(hist, key=hist.get, reverse=True)
```

```
for url in sorted(hist, key=hist.get, reverse=True):
    print (url, hist[url])
```

Automatic Updates to `/etc/fstab`

Old and New Styles

```
# <file system> <mount point>   <type>  <options>       <dump>  <pass>
/dev/sda1 /               ext4    errors=remount-ro 0       1
/dev/sda5 none            swap    sw                0       0
```

```
# <file system> <mount point>   <type>  <options>       <dump>  <pass>
UUID=781ffbee-4471-45f1-8086-a01160786143 /         ext4    errors=remount-ro 0       1
UUID=97eaef25-16c3-4efb-9a57-33767723a93f none      swap    sw                0       0
```

Pseudo-Code

```
for each line in fstab:
    if the line contains a partition name:
        map the partition name to a UUID using lsblk
        write the modified line to the output file
    else
        copy the line to the output file unchanged
```

Dissecting the File

```
# <file system> <mount point>   <type>  <options>       <dump>  <pass>
/dev/sda1 /               ext4    errors=remount-ro 0       1
/dev/sda5 none            swap    sw                0       0
```

```
r"(/dev/sd[ab][0-9])(.*)"
```

Lesson Summary

Reading binary files

- Unpacking structures

Logging

- Loggers, Handlers, Formatters
- Log levels

Processing an apache access log

- Dissecting the file,
- building and sorting a histogram

Modifying /etc/fstab

- More regular expressions!
- Running an external command

Course Summary

Why use Python?

- Comparison with bash

Development environments

- The REPL, iPython, IDLE, PyCharm

Managing the file system

- The os module
- Collections (lists, tuples, dictionaries)

Interacting with Linux

- Accessing the command line
- Streams and filters
- Handling signals

Course Summary (Continued)

Combining Python with other tools

- Running a program with subprocess
- Writing a network client (mail)
- Using a common file format (tar)

Strings

- String literals, formatting, testing
- Dates and times, regular expressions

File I/O

- binary files, logging,
- processing apache access log
- modifying /etc/fstab

Going Further

"Python Fundamentals"

- Austin Bingham and Robert Smallshire
- A more systematic view of the language
- Defining your own classes, etc.

"Python – Beyond the Basics"

- Organizing larger programs
- Closures, decorators, iterables
- Inheritance and Polymorphism

And that endless supply of batteries!

- Keep reading at `docs.python.org`