

Extracted from:

Data Crunching

Solve Everyday Problems Using Java, Python, and More

This PDF file contains pages extracted from Data Crunching, one of the Pragmatic Starter Kit series of books for project teams. For more information, visit http://www.pragmaticprogrammer.com/starter_kit.

Note: This extract contains some colored text (particularly in code listing). This is available only in online versions of the books. The printed versions are black and white. Pagination might vary between the online and printer versions; the content is otherwise identical.

Copyright © 2005 The Pragmatic Programmers, LLC.

All rights reserved.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form, or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior consent of the publisher.

Contents

1	Introduction	1
1.1	Name That Molecule	1
1.2	There's One in Every Crowd...	2
1.3	And the Moral Is...	3
1.4	Questions About Data Crunching	4
1.5	Road Map	6
2	Text	8
2.1	Reversing a File	8
2.2	Reformatting Data	10
2.3	Handling Multiline Records	16
2.4	Checking for Collisions	22
2.5	Including One File in Another	27
2.6	The Unix Shell	30
2.7	Very Large Data Sets	37
2.8	Summary	38
3	Regular Expressions	39
3.1	The Shell	40
3.2	Basic Patterns	41
3.3	Extracting Matched Values	48
3.4	Practical Applications	57
3.5	Speaking in Tongues	67
3.6	Other Systems	69
3.7	Summary	72
4	XML	74
4.1	A Quick Introduction	74
4.2	SAX	79
4.3	DOM	90
4.4	XPath	99
4.5	XSLT	104
4.6	Summary	112

5	Binary Data	114
5.1	Numbers	115
5.2	Input and Output	119
5.3	Strings	123
5.4	Summary	133
6	Relational Databases	134
6.1	Simple Queries	135
6.2	Nesting and Negation	144
6.3	Aggregation and Views	148
6.4	Creating, Updating, and Deleting	152
6.5	Using SQL in Programs	159
6.6	Summary	163
7	Horseshoe Nails	164
7.1	Unit Testing	164
7.2	Encoding and Decoding	173
7.3	Floating-Point Arithmetic	175
7.4	Dates and Times	178
7.5	Summary	182
A	Resources	183
A.1	Bibliography	183