
Spis treści

Wstęp	7
złość I Przygotowanie	11
1. Teoria	13
Agile w Big Data	13
Wielkie słowa	15
Zespoły	16
Rozpoznawanie problemów i szans	18
Adaptowanie do zmian	18
Proces wytwórczy w zwinnym Big Data	22
Programowanie w parach i przegląd kodu	24
Środowisko zwinnej pracy a produktywność	24
Przestrzeń współpracy	25
Przestrzeń prywatna	26
Przestrzeń osobista	26
Pomysły na wielkoformatowych wydrukach	26
2. Dane	29
E-mail	29
Praca z surowymi danymi	30
Surowe wiadomości e-mail	30
Dane ustrukturyzowane a dane na wpół ustrukturyzowane	31
SQL	31
NoSQL	37
Serializacja	38
Wyodrębnianie i ujawnianie cech w ewoluującym schemacie	39
Potoki danych	40

Perspektywy danych	40
Sieci	41
Szeregi czasowe	44
Język naturalny	44
Prawdopodobieństwo	45
Podsumowanie	48
3. Narzędzia zwinności	49
Skalowalność = prostota	49
Zwinne przetwarzanie w Big Data	50
Konfigurowanie wirtualnego środowiska dla języka Python	52
Serializacja zdarzeń przez Avro	52
Avro w Pythonie	53
Zbieranie danych	55
Przetwarzanie danych w Pigu	58
Instalacja	58
Publikowanie danych w MongoDB	62
Instalacja	62
Instalowanie sterownika MongoDB dla Javy	63
Instalowanie łącznika mongo-hadoop	63
Wypychanie danych z Piga do MongoDB	63
Wyszukiwarka Elasticsearch	66
Instalacja	66
ElasticSearch i Pig — Wonderdog	66
Refleksja o kształcie potoku przetwarzającego	69
Lekkie aplikacje WWW	70
Python i Flask	70
Prezentacja danych	72
Instalacja	73
Bootstrap na start	73
Wizualizacja danych: D3.js i nv3.js	78
Podsumowanie	78
4. Do chmury!	81
Wprowadzenie	81
GitHub	83
DotCloud	84
Pierwszy krok w dotCloud	85
Procesy robocze w Pythonie	87

Amazon Web Services	87
Simple Storage Service	88
Elastic MapReduce	89
MongoDB w wydaniu usługowym	94
Monitorowanie	97
Google Analytics	97
Mortar Data	98

Część II W górę piramidy 101

5. Zbieranie i wyświetlanie rekordów 105

Montaż końcowy	106
Pobieranie i serializowanie zawartości skrzynki pocztowej	107
Przetwarzanie i publikowanie wiadomości e-mail	108
Prezentowanie wiadomości w przeglądarce	110
Serwowanie wiadomości przez Flask i pymongo	110
Renderowanie strony HTML5 z szablonów Jinja2	111
Kontrola zwinności	115
Listy wiadomości	116
Generowanie list wiadomości w MongoDB	116
Anatomia prezentacji	119
Przeszukiwanie wiadomości e-mail	124
Indeksowanie wiadomości — Pig, Elasticsearch i Wonderdog	124
Wyszukiwanie wiadomości z poziomu aplikacji WWW	125
Podsumowanie	126

6. Wizualizacja danych na wykresach 129

Dobre wykresy	130
Wyodrębnianie encji: adresy e-mail	130
Wyodrębnianie adresów	131
Wizualizacja w przekroju czasowym	135
Podsumowanie	141

7. Eksplorowanie danych w raportach 143

Budowanie raportów z wieloma wykresami	144
Łączenie rekordów	147
Ekstrakcja słów z wiadomości — TF-IDF	152
Podsumowanie	158

8. Stawianie prognoz	161
Przewidywanie współczynnika odpowiedzi na wiadomości	162
Personalizacja	167
Podsumowanie	168
9. Ukierunkowywanie działań	169
Właściwości skutecznych wiadomości e-mail	170
Lepsze przewidywanie — prosty predyktor bayesowski	171
$P(\text{reply} \mid \text{from} \ \& \ \text{to})$	171
$P(\text{reply} \mid \text{token})$	171
Predykcje w czasie rzeczywistym	174
Rejestrowanie zdarzeń w aplikacji	177
Podsumowanie	179
Skorowidz	180