Indice generale

Introduzione xiii

Struttura del libro xiii
Cosa serve per questo libro xiv
Lo scopo del libro xiv
Convenzioni xv
Codice degli esempi xv

Gli autori xvii

I revisori xix

Capitolo 1 Per iniziare 1

Una nota sulle versioni 1
Panoramica su Hadoop 2
Componenti di Hadoop 3
Componenti comuni 4
Storage 4
Calcolo 5
Meglio se insieme 5
Hadoop 2: dov'è l'affare? 6
Storage in Hadoop 2 6
Calcolo in Hadoop 2 7
Distribuzioni di Apache Hadoop 9
Un doppio approccio 10
AWS: infrastruttura on demand di Amazon 10
Simple Storage Service (S3) 10
Elastic MapReduce (EMR) 10
Come iniziare 11
Cloudera QuickStart VM 11
Amazon EMR 11

Utilizzare Elastic MapReduce 12
Rendere Hadoop operativo 12
L'interfaccia AWS a riga di comando 14
Eseguire gli esempi 15
Elaborazione dei dati con Hadoop 16
Perché Twitter? 16
Creare il primo dataset 16
Accesso programmato con Python 19
Riepilogo 21

Capitolo 2 Storage 23

Funzionamento interno di HDFS 23
Avvio del cluster 24
Replica dei blocchi 25
Accedere al file system HDFS tramite riga di comando 25
Esplorare il file system HDFS 26
Proteggere i metadati del file system 27
Il Secondary NameNode non ci salva 28
NameNode HA di Hadoop 2 28
Configurazione del client 29
Come funziona un failover 29
Apache ZooKeeper: un file system diverso 30
Implementare un lock distribuito con ZNode sequenziali 31
Implementare l'adesione a un gruppo e l'elezione di un leader usando ZNode effimeri 32
API Java 33
Componenti 33
Per saperne di più 33
Failover automatico dei NameNode 33
Snapshot HDFS 34
File system di Hadoop 36
Interfacce di Hadoop 37
Gestire e serializzare i dati 37
L'interfaccia Writable 37
Le classi wrapper 38
Classi wrapper per gli array 38
Le interfacce Comparable e WritableComparable 39
Storage dei dati 39
Serializzazione e contenitori 39
Compressione 39
Formati di file general purpose 40
Formati di dati orientati alle colonne 41
Riepilogo 45

Capitolo 3 Elaborazione: MapReduce e oltre.....47

MapReduce.....47
API Java per MapReduce.....48
La classe Mapper.....49
La classe Reducer.....50
La classe Driver.....51
Combiner.....52
Partizionamento dei file.....53
Implementazione dei mapper e dei reducer forniti da Hadoop.....54
Condividere i dati di riferimento.....54
Scrivere programmi MapReduce.....55
Come iniziare.....55
Eseguire gli esempi.....55
WordCount, l'Hello World di MapReduce.....56
Co-occorrenze di parole.....59
Trending topic.....60
Sentiment degli hashtag.....64
Pulizia del testo con ChainMapper.....67
Panoramica sull'esecuzione di un job di MapReduce.....70
Avvio.....70
Suddividere l'input.....71
Assegnazione delle attività.....71
Avvio dell'attività.....71
Monitorare il lavoro del JobTracker.....72
Input della classe Mapper.....72
Esecuzione del mapper.....72
Output del mapper e input del reducer.....73
Input del reducer.....73
Esecuzione del reducer.....73
Output del reducer.....73
Chiusura.....73
Input/Output.....74
InputFormat e RecordReader.....74
La classe InputFormat fornita da Hadoop.....75
La classe RecordReader fornita da Hadoop.....75
OutputFormat e RecordWriter.....75
La classe OutputFormat fornita da Hadoop.....76
SequenceFile.....76
YARN.....76
L'architettura di YARN.....77
Ciclo di vita di un'applicazione YARN.....78
Pensare a livelli.....80
Modelli di esecuzione.....80

YARN nel mondo reale: il calcolo oltre MapReduce 81
Il problema con MapReduce 81
Tez 81
Apache Spark 84
Apache Samza 85
YARN oggi e oltre 86
Riepilogo 86

Capitolo 4 Computazione in tempo reale con Samza 89

Elaborazione degli stream con Samza 89
Come funziona Samza 90
L'architettura ad alto livello di Samza 90
Il miglior amico di Samza: Apache Kafka 91
Integrazione con YARN 92
Un modello indipendente 93
Hello Samza! 93
Creare un job di parsing di un tweet 94
Il file di configurazione 95
Portare i dati di Twitter in Kafka 97
Eseguire un job di Samza 98
Samza e HDFS 99
Windowing 99
Flussi di lavoro con più job 101
Sentiment analysis dei tweet 102
Attività stateful 107
Riepilogo 110

Capitolo 5 Computazione iterativa con Spark 111

Apache Spark 112
Computazione dei cluster con i working set 112
Distribuzione 114
Iniziare con Spark 115
Scrivere ed eseguire applicazioni standalone 116
L'ecosistema di Spark 119
Spark Streaming 119
GraphX 119
MLlib 119
Spark SQL 120
Elaborare i dati con Apache Spark 120
Costruire ed eseguire gli esempi 120
Elaborazione dei dati sugli stream 123
Analisi dei dati con Spark SQL 125
Spark e Samza Streaming a confronto 127
Riepilogo 128

Capitolo 6 Analisi dei dati con Apache Pig.....129

Panoramica su Pig.....130
Per iniziare.....131
Eseguire Pig.....131
Grunt, la shell interattiva di Pig.....132
Fondamenti di Apache Pig.....133
Programmare Pig.....135
Tipi di dati di Pig.....135
Funzioni di Pig.....136
Lavorare con i dati.....138
Estendere Pig (UDF).....141
Repository di UDF.....141
Analizzare lo stream di Twitter.....142
Prerequisiti.....142
Esplorazione del dataset.....142
Metadati dei tweet.....143
Preparazione dei dati.....143
Statistiche top n.....144
Manipolazione di datetime.....146
Catturare le interazioni dell'utente.....147
Analisi dei link.....149
Utenti influenti.....150
Riepilogo.....153

Capitolo 7 Hadoop e SQL.....155

Perché SQL su Hadoop.....156
Altre soluzioni SQL su Hadoop.....156
Prerequisiti.....156
Panoramica su Hive.....158
La natura delle tabelle di Hive.....160
L'architettura di Hive.....160
Tipi di dati.....161
Istruzioni DDL.....161
Formati di file e storage.....163
Query.....167
Strutturare le tabelle di Hive per i vari carichi di lavoro.....169
Partizionare una tabella.....169
Scrivere degli script.....175
Hive e Amazon Web Services.....175
Hive e S3.....175
Hive su Elastic MapReduce.....177
Estendere HiveQL.....177
Interfacce programmatiche.....179
JDBC.....179
Thrift.....181

L'iniziativa Stinger....182
Impala....183
L'architettura di Impala....184
Coesistenza con Hive....184
Una filosofia diversa....185
Drill, Tajo e oltre....186
Riepilogo....186

Capitolo 8 Gestione del ciclo di vita dei dati....189

Cos'è la gestione del ciclo di vita dei dati....189
Importanza della gestione del ciclo di vita dei dati....190
Strumenti di supporto....190
Costruire la capacità per l'analisi dei tweet....190
Ottenere i dati dei tweet....191
Oozie....191
Produrre dati derivati....204
Le sfide dei dati esterni....209
Validazione dei dati....210
Gestire le modifiche al formato....211
Gestire l'evoluzione dello schema con Avro....211
Raccogliere dati supplementari....216
Programmare i workflow....216
Altri trigger di Oozie....218
Assemblare il tutto....219
Altri strumenti di supporto....219
Riepilogo....220

Capitolo 9 Facilitare il lavoro di sviluppo....221

Scegliere un framework....221
Hadoop Streaming....222
Conteggio delle parole in streaming in Python....223
Differenze tra i job quando si usa lo streaming....225
Trovare parole importanti nel testo....225
Kite Data....231
Data Core....231
Data HCatalog....232
Data Hive....233
Data MapReduce....233
Data Spark....233
Data Crunch....233
Apache Crunch....234
Per iniziare....235
Concetti....235
Serializzazione dei dati....236
Pattern di elaborazione dei dati....237

Implementazione ed esecuzione delle pipeline 238
Esempi di Crunch 239
Kite Morphlines 243
Riepilogo 250

Capitolo 10 Eseguire un cluster Hadoop 251

Sono uno sviluppatore, le operations non mi interessano! 251
Best practice per Hadoop e DevOps 252
Cloudera Manager 252
Pagare o non pagare? 253
Gestione dei cluster con Cloudera Manager 253
Monitorare con Cloudera Manager 254
API Cloudera Manager 255
Il lock-in di Cloudera Manager 255
Ambari, l'alternativa open source 256
Le operations nel mondo di Hadoop 2 256
Condividere le risorse 258
Costruire un cluster fisico 258
Layout fisico 259
Costruire un cluster su EMR 261
Considerazioni sui file system 262
Ottenere i dati in EMR 262
Istanze di EC2 e raffinamento 263
Raffinamento dei cluster 263
Considerazioni sulla JVM 263
Ottimizzazione di map e reduce 264
Sicurezza 264
Evoluzione del modello di sicurezza di Hadoop 265
Oltre l'autorizzazione di base 265
Il futuro della sicurezza di Hadoop 266
Conseguenze dell'uso di un cluster protetto 266
Monitorare 267
Hadoop, dove i fallimenti non contano 267
Monitoraggio integrato 267
Metriche a livello di applicazione 268
Risoluzione dei problemi 268
Livelli di log 269
Accedere ai logfile 270
ResourceManager, NodeManager e Application Manager 272
NameNode e DataNode 279
Riepilogo 279

Capitolo 11 Come proseguire 283

Distribuzioni alternative 284
Distribuzione di Cloudera per Hadoop 284
Hortonworks Data Platform 285

MapR 285
E il resto… 286
Scegliere una distribuzione 286
Altri framework di calcolo 286
Apache Storm 286
Apache Giraph 287
Apache HAMA 287
Altri progetti interessanti 287
HBase 287
Sqoop 288
Whirr 288
Mahout 289
Hue 289
Altre astrazioni di programmazione 291
Cascading 291
Risorse per gli AWS 292
SimpleDB e DynamoDB 292
Kinesis 292
Data Pipeline 293
Fonti di informazione 293
Codice sorgente 293
Mailing list e forum 293
Gruppi di LinkedIn 294
HUG 294
Conferenze 294
Riepilogo 294

Indice analitico 295