

Introduction

1. Introduction	11
2. Buts et objectifs de l'ouvrage	12
3. À qui s'adresse cet ouvrage ?	15
4. Comment lire cet ouvrage ?	15
5. Ce que ce livre n'est pas	16
6. Précisions de l'auteur	17
7. Remerciements	17
8. Dédicace	18

Contexte de création d'Hadoop

1. Introduction	19
2. Contexte d'Hadoop	20
3. Approche conceptuelle d'Hadoop	21

4. Qui utilise Hadoop ?	24
4.1 Effervescence des projets impliquant l'utilisation d'Hadoop en Europe	24
4.2 Cas d'utilisation d'Hadoop	25
5. Conclusion	26

Architecture infrastructurelle d'Hadoop

1. Introduction	27
2. Types d'architectures distribuées	28
2.1 Architectures délocalisées ou client/serveur	28
2.1.1 Architecture client/serveur à deux niveaux (two-tiers)	28
2.1.2 Architecture client/serveur à trois niveaux (three-tiers)	30
2.2 Architectures distribuées	31
2.2.1 Cluster computing ou grappe de calcul	33
2.2.2 Grid computing ou grille de calcul	36
2.3 Caractéristiques du cluster computing	37
2.3.1 Scalabilité horizontale : le facteur clé d'adoption du cluster	37
2.3.2 Tolérance aux pannes	40
2.3.3 Haute disponibilité	41

2.3.4 Mesure de la disponibilité d'un système	43
3. Modes de partage des ressources dans le cluster	45
3.1 Shared-nothing	45
3.2 Shared-memory	47
3.3 Shared-disk	48
4. Modes de communication des nœuds dans le cluster	49
4.1 Modèle maître/esclave	50
4.2 Modèle peer-to-peer	51
5. Modes de traitement de données dans le cluster	52
5.1 Traitement sur disque ou batch processing	53
5.2 Traitement en mémoire ou in-memory processing	55
6. Types de parallélismes des calculs dans un cluster	57
6.1 Parallélisme asynchrone simultané	58
6.2 Parallélisme pipeline	60
6.3 Parallélisme indépendant	61
7. Choix architecturaux d'un cluster Hadoop	63
8. Conclusion	65

9. Guide d'étude du chapitre	66
9.1 Présentation des guides d'étude	66
9.2 Comment utiliser le guide d'étude ?	66
10. À retenir	71
MapReduce	
1. Introduction	73
2. Mapreduce : un nouveau paradigme	74
3. Détails conceptuels des phases du MapReduce	75
3.1 Phase Map	75
3.2 Phase Shuffle	77
3.3 Phase Reduce	79
4. Détails techniques de l'exécution du MapReduce dans un cluster	80
5. Exemples d'application du MapReduce	82
5.1 Calcul d'un index inversé	85
5.2 Jointure de deux tables relationnelles	85

5.3 Exemple de patron de conception du Map/Reduce en Python	87
6. Modèles alternatifs au MapReduce	88
6.1 Tez : le moteur d'optimisation du MapReduce	90
6.2 Spark : le moteur in-memory distribué	92
6.3 Architectures ? : la couche Streaming du MapReduce	93
7. Conclusion	97
8. Guide d'étude du chapitre	97
9. À retenir	101
Hadoop	
1. Introduction	103
2. Spécificités d'un cluster Hadoop	104
2.1 « Conscience des casiers » des nœuds du cluster (rack awareness)	106
2.2 Localisation des données dans le cluster (data locality)	107
3. Détails d'exécution du MapReduce dans un cluster Hadoop	108
4. Gestion des défaillances dans un cluster Hadoop	114

4.1 Gestion de la défaillance du nœud de référence	114
4.2 Gestion de la défaillance des nœuds de données	115
4.3 Gestion des « retardataires » (stragglers)	115
5. Hadoop Streaming	116
6. Conclusion	118
7. Guide d'étude du chapitre	119
8. À retenir	122
HDFS	
1. Introduction	123
2. Pourquoi a-t-on besoin d'un système de fichiers ?	124
2.1 Principes de stockage des données sur le disque dur	125
2.2 Principes de stockage des données dans un cluster	129
2.2.1 Principes de stockage des données dans un cluster shared-disk	129
2.2.2 Principes de stockage des données dans un cluster shared-nothing	133
3. Définition du HDFS dans le cluster Hadoop	136

3.1 Définition et rôle du HDFS dans le cluster	136
3.2 Processus de maintien de la haute disponibilité du cluster	139
3.3 Interactivité avec le HDFS	141
4. Conclusion	142
5. Guide d'étude du chapitre	143
6. À retenir	146
Futur d'Hadoop : limites d'Hadoop et YARN	
1. Introduction	147
2. Limites d'Hadoop	148
2.1 Modèle de calcul d'Hadoop	148
2.2 HDFS	148
2.3 Haute disponibilité du cluster	149
2.4 Sécurité du cluster	150
3. YARN et développements en cours sur Hadoop	151
3.1 Définition du YARN	151
3.2 Fonctionnement du YARN	154

3.3 Fédération HDFS	156
4. Conclusion	157
5. Guide d'étude du chapitre	158
6. À retenir	162
SQL dans Hadoop	
1. Introduction	163
2. Étude de l'écosystème Hadoop	164
3. Langages d'abstraction	167
3.1 Hive	169
3.1.1 Infrastructure technique de Hive	169
3.1.2 Écriture des requêtes HiveQL	171
3.2 Pig	174
4. Moteurs natifs SQL sur Hadoop	178
4.1 Fonctionnement des bases de données parallèles (MPP DB)	179
4.1.1 Architecture des bases de données parallèles	180
4.1.2 Exécution des requêtes SQL dans les bases de données parallèles	

4.2 Fonctionnement des moteurs natifs SQL sur Hadoop	185
4.3 Impala : le moteur SQL sur Hadoop de Cloudera	189
	191
5. Conclusion	194
6. Guide d'étude du chapitre	194
7. À retenir	202
Streaming	
1. Introduction	205
2. Domaine temporel	206
3. Approches de traitement streaming	209
3.1 Approche batch du traitement streaming	210
3.1.1 Batch par fenêtrage	210
3.1.2 Batch par sessions	211
3.2 Approche continue du traitement streaming	211
3.2.1 Fenêtres	212
3.2.2 Techniques de traitement événement par événement	214
3.2.3 Techniques de traitement agnostiques au temps (time-agnostic)	214

3.2.4 Techniques d'approximation	214
3.2.5 Techniques de fenêtrage par temps de traitement	215
3.2.6 Techniques de fenêtrage par temps d'événement	216
4. Idempotence	219
4.1 Nature du traitement	220
4.1.1 Traitements de nature déterministe	220
4.1.2 Traitements de nature aléatoire	221
4.2 État	222
4.2.1 Définition de la notion d'état	222
4.2.2 Utilisation ou non de l'état	223
4.2.3 Mécanisme de sauvegarde de l'état	225
5. Disponibilité d'un système streaming	226
6. Conclusion	227
7. Guide d'étude du chapitre	228
8. À retenir	235

Apache Storm

1. Introduction	237
2. Définition de Storm	238
3. Fonctionnement de Storm	240
4. Topologies	242
4.1 Philosophie et fonctionnement des topologies	242
4.2 Topologies DRPC	245
5. Utilisation de Storm	246
6. Storm et Hadoop	249
6.1 Storm-YARN	249
6.2 Storm et architecture ?	251
7. Conclusion	255
8. Guide d'étude du chapitre	255
9. À retenir	258

Adoption d'Hadoop

1. Introduction	261
2. Distributions Hadoop	263
3. Distribution Cloudera d'Hadoop	264
4. Distribution Hortonworks d'Hadoop	265
5. Distribution MapR d'Hadoop	266
6. Tableau récapitulatif des outils proposés	268
7. Guide de sélection d'une distribution Hadoop	270
8. Conclusion	276
9. Guide d'étude du chapitre	277
10. À retenir	278

Transition numérique

1. Introduction	281
2. Changement	282
2.1 Principes qui régissent le changement	

2.1.1 Principe 1 : le changement est un processus, pas un événement	283
2.1.2 Principe 2 : le changement annonce son arrivée par des signes	284
2.1.3 Principe 3 : le changement contient une opportunité qui lui est inhérente	284
2.1.4 Principe 4 : le changement est un processus inéluctable	285
2.2 Clés qui donnent accès aux opportunités du changement	286
2.2.1 Clé 1 : évitez le comportement de la grenouille - le changement est processus	287
2.2.2 Clé 2 : évitez le comportement du crabe - le changement est inéluctable	287
2.2.3 Clé 3 : évitez la présomption - tout change	290
2.2.4 Clé 4 : changez votre perception - le changement est normal	291
2.2.5 Clé 5 : soyez intentionnel - le changement contient une opportunité	292
3. Transition vers le Numérique et Hadoop	294
3.1 Caractéristiques de l'ère numérique	298
3.1.1 Âge de l'information	299
3.1.2 Âge de la communication	299
3.1.3 Âge de la globalisation	300
3.2 Pourquoi apprendre Hadoop ?	302
3.2.1 Raison 1 : apprendre Hadoop vous positionne en pionnier	304
3.2.2 Raison 2 : apprendre Hadoop hausse votre valeur professionnelle	305
3.2.3 Raison 3 : apprendre Hadoop vous permet de couvrir la majorité des problématiques de traitement de données	306

3.3 Quelques conseils	307
3.3.1 Profils métier Hadoop	307
3.3.2 Certifications éditeurs	311
3.3.3 Masters spécialisés	313
3.3.4 Kaggle et meetup Hadoop	315
4. Conclusion	316
5. Guide d'étude du chapitre	316
6. À retenir	319
Liens et références utiles	
1. Liens utiles	323
2. Bibliographie	327
Réponses des guides d'étude	
1. Guide d'étude du chapitre Architecture infrastructurelle d'Hadoop	329
2. Guide d'étude du chapitre MapReduce	332

3. Guide d'étude du chapitre Hadoop	336
4. Guide d'étude du chapitre HDFS	338
5. Guide d'étude du chapitre Futur d'Hadoop : limites d'Hadoop et Yarn	340
6. Guide d'étude du chapitre SQL dans Hadoop	344
7. Guide d'étude du chapitre Streaming 101	351
8. Guide d'étude du chapitre Apache Storm	357
9. Guide d'étude du chapitre Adoption d'Hadoop	359
10. Guide d'étude du chapitre Transition numérique	361
Index	365