

大数据技术体系详解：原理、架构与实践



[大数据技术体系详解：原理、架构与实践 下载链接1](#)

著者:董西成

出版者:机械工业出版社

出版时间:2018-4

装帧:平装

isbn:9787111590729

作者介绍:

董西成，资深大数据技术实践者和研究者，对大数据基础架构有非常深刻的认识和理解，有着丰富的实践经验。熟悉常见的开源大数据解决方案，包括Hadoop和spark生态系统等，擅长底层分布式系统的优化和开发。撰写了大量Hadoop和spark等大数据相关的技术文章并分享在自己的博客上，由于文章技术含量高，所以非常受欢迎。出版有大数据领域负有盛名的专著：《Hadoop技术内幕：深入解析MapReduce架构设计与实现原理》和《Hadoop技术内幕：深入解析YARN架构设计与实现原理》。

目录: 前言

第一部分 概述篇

第1章 企业级大数据技术体系概述 2

1.1 大数据系统产生背景及应用场景 2

1.1.1 产生背景 2

1.1.2 常见大数据应用场景 3

1.2 企业级大数据技术框架 5

1.2.1 数据收集层 6

1.2.2 数据存储层 7

1.2.3 资源管理与服务协调层 7

1.2.4 计算引擎层 8

1.2.5 数据分析层 9

1.2.6 数据可视化层 9

1.3 企业级大数据技术实现方案 9

1.3.1 Google大数据技术栈 10

1.3.2 Hadoop与Spark开源大数据技术栈 12

1.4 大数据架构: Lambda Architecture 15

1.5 Hadoop与Spark版本选择及安装部署 16

1.5.1 Hadoop与Spark版本选择 16

1.5.2 Hadoop与Spark安装部署 17

1.6 小结 18

1.7 本章问题 18

第二部分 数据收集篇

第2章 关系型数据的收集 20

2.1 Sqoop概述 20

2.1.1 设计动机 20

2.1.2 Sqoop基本思想及特点 21

2.2 Sqoop基本架构 21

2.2.1 Sqoop1基本架构 22

2.2.2 Sqoop2基本架构 23

2.2.3 Sqoop1与Sqoop2对比 24

2.3 Sqoop使用方式 25

2.3.1 Sqoop1使用方式 25

2.3.2 Sqoop2使用方式 28

2.4 数据增量收集CDC 31

2.4.1 CDC动机与应用场景 31

2.4.2 CDC开源实现Canal 32

2.4.3 多机房数据同步系统Otter 33

2.5 小结 35

2.6 本章问题 35

第3章 非关系型数据的收集 36

3.1 概述 36

3.1.1 Flume设计动机 36

3.1.2 Flume基本思想及特点 37

3.2 Flume NG基本架构 38

3.2.1 Flume NG基本架构 38

3.2.2 Flume NG高级组件 41

3.3 Flume NG数据流拓扑构建方法 42

3.3.1 如何构建数据流拓扑 42

3.3.2 数据流拓扑实例剖析 46

3.4 小结 50

3.5 本章问题 50

第4章 分布式消息队列Kafka 51

4.1 概述 51

4.1.1 Kafka设计动机 51

4.1.2 Kafka特点 53

4.2 Kafka设计架构 53

4.2.1 Kafka基本架构 54

4.2.2 Kafka各组件详解 54

4.2.3 Kafka关键技术点 58

4.3 Kafka程序设计 60

4.3.1 Producer程序设计 61

4.3.2 Consumer程序设计 63

4.3.3 开源Producer与Consumer实现 65

4.4 Kafka典型应用场景 65

4.5 小结 67

4.6 本章问题 67

第三部分 数据存储篇

第5章 数据序列化与文件存储格式 70

5.1 数据序列化的意义 70

5.2 数据序列化方案 72

5.2.1 序列化框架Thrift 72

5.2.2 序列化框架Protobuf 74

5.2.3 序列化框架Avro 76

5.2.4 序列化框架对比 78

5.3 文件存储格式剖析 79

5.3.1 行存储与列存储 79

5.3.2 行式存储格式 80

5.3.3 列式存储格式ORC、Parquet与CarbonData 82

5.4 小结 88

5.5 本章问题 89

第6章 分布式文件系统 90

6.1 背景 90

6.2 文件级别和块级别的分布式文件系统 91

6.2.1 文件级别的分布式系统 91

6.2.2 块级别的分布式系统 92

6.3 HDFS基本架构 93

6.4 HDFS关键技术 94

6.4.1 容错性设计 95

6.4.2 副本放置策略 95

6.4.3 异构存储介质 96

6.4.4 集中式缓存管理 97

6.5 HDFS访问方式 98

6.5.1 HDFS shell 98

6.5.2 HDFS API 100

6.5.3 数据收集组件 101

6.5.4 计算引擎 102

6.6 小结 102

6.7 本章问题 103

第7章 分布式结构化存储系统 104

7.1 背景 104

7.2 HBase数据模型 105

7.2.1 逻辑数据模型 105

7.2.2 物理数据存储 107

7.3 HBase基本架构 108

7.3.1 HBase基本架构 108

7.3.2 HBase内部原理	110
7.4 HBase访问方式	114
7.4.1 HBase shell	114
7.4.2 HBase API	116
7.4.3 数据收集组件	118
7.4.4 计算引擎	119
7.4.5 Apache Phoenix	119
7.5 HBase应用案例	120
7.5.1 社交关系数据存储	120
7.5.2 时间序列数据库OpenTSDB	122
7.6 分布式列式存储系统Kudu	125
7.6.1 Kudu基本特点	125
7.6.2 Kudu数据模型与架构	126
7.6.3 HBase与Kudu对比	126
7.7 小结	127
7.8 本章问题	127
第四部分 分布式协调与资源管理篇	
第8章 分布式协调服务ZooKeeper	130
8.1 分布式协调服务的存在意义	130
8.1.1 leader选举	130
8.1.2 负载均衡	131
8.2 ZooKeeper数据模型	132
8.3 ZooKeeper基本架构	133
8.4 ZooKeeper程序设计	134
8.4.1 ZooKeeper API	135
8.4.2 Apache Curator	139
8.5 ZooKeeper应用案例	142
8.5.1 leader选举	142
8.5.2 分布式队列	143
8.5.3 负载均衡	143
8.6 小结	144
8.7 本章问题	145
第9章 资源管理与调度系统YARN	146
9.1 YARN产生背景	146
9.1.1 MRv1局限性	146
9.1.2 YARN设计动机	147
9.2 YARN设计思想	148
9.3 YARN的基本架构与原理	149
9.3.1 YARN基本架构	149
9.3.2 YARN高可用	152
9.3.3 YARN工作流程	153
9.4 YARN资源调度器	155
9.4.1 层级队列管理机制	155
9.4.2 多租户资源调度器产生背景	156
9.4.3 Capacity/Fair Scheduler	157
9.4.4 基于节点标签的调度	160
9.4.5 资源抢占模型	163
9.5 YARN资源隔离	164
9.6 以YARN为核心的生态系统	165
9.7 资源管理系统Mesos	167
9.7.1 Mesos基本架构	167
9.7.2 Mesos资源分配策略	169
9.7.3 Mesos与YARN对比	170
9.8 资源管理系统架构演化	170

- 9.8.1 集中式架构 171
- 9.8.2 双层调度架构 171
- 9.8.3 共享状态架构 172
- 9.9 小结 173
- 9.10 本章问题 173
- 第五部分 大数据计算引擎篇
- 第10章 批处理引擎MapReduce 176
 - 10.1 概述 176
 - 10.1.1 MapReduce产生背景 176
 - 10.1.2 MapReduce设计目标 177
 - 10.2 MapReduce编程模型 178
 - 10.2.1 编程思想 178
 - 10.2.2 MapReduce编程组件 179
 - 10.3 MapReduce程序设计 187
 - 10.3.1 MapReduce程序设计基础 187
 - 10.3.2 MapReduce程序设计进阶 194
 - 10.3.3 Hadoop Streaming 198
 - 10.4 MapReduce内部原理 204
 - 10.4.1 MapReduce作业生命周期 204
 - 10.4.2 MapTask与ReduceTask 206
 - 10.4.3 MapReduce关键技术 209
 - 10.5 MapReduce应用实例 211
 - 10.6 小结 213
 - 10.7 本章问题 213
- 第11章 DAG计算引擎Spark 215
 - 11.1 概述 215
 - 11.1.1 Spark产生背景 215
 - 11.1.2 Spark主要特点 217
 - 11.2 Spark编程模型 218
 - 11.2.1 Spark核心概念 218
 - 11.2.2 Spark程序基本框架 220
 - 11.2.3 Spark编程接口 221
 - 11.3 Spark运行模式 227
 - 11.3.1 Standalone模式 229
 - 11.3.2 YARN模式 230
 - 11.3.3 Spark Shell 232
 - 11.4 Spark程序设计实例 232
 - 11.4.1 构建倒排索引 232
 - 11.4.2 SQL GroupBy实现 234
 - 11.4.3 应用程序提交 235
 - 11.5 Spark内部原理 236
 - 11.5.1 Spark作业生命周期 237
 - 11.5.2 Spark Shuffle 241
 - 11.6 DataFrame、Dataset与SQL 247
 - 11.6.1 DataFrame/Dataset与SQL的关系 248
 - 11.6.2 DataFrame/Dataset程序设计 249
 - 11.6.3 DataFrame/Dataset程序实例 254
 - 11.7 Spark生态系统 257
 - 11.8 小结 257
 - 11.9 本章问题 258
- 第12章 交互式计算引擎 261
 - 12.1 概述 261
 - 12.1.1 产生背景 261
 - 12.1.2 交互式查询引擎分类 262

- 12.1.3 常见的开源实现 263
- 12.2 ROLAP 263
 - 12.2.1 Impala 263
 - 12.2.2 Presto 267
 - 12.2.3 Impala与Presto对比 271
- 12.3 MOLAP 271
 - 12.3.1 Druid简介 271
 - 12.3.2 Kylin简介 272
 - 12.3.3 Druid与Kylin对比 274
- 12.4 小结 274
- 12.5 本章问题 274
- 第13章 流式实时计算引擎 276
 - 13.1 概述 276
 - 13.1.1 产生背景 276
 - 13.1.2 常见的开源实现 278
 - 13.2 Storm基础与实战 278
 - 13.2.1 Storm概念与架构 279
 - 13.2.2 Storm程序设计实例 282
 - 13.2.3 Storm内部原理 285
 - 13.3 Spark Streaming基础与实战 290
 - 13.3.1 概念与架构 290
 - 13.3.2 程序设计基础 291
 - 13.3.3 编程实例详解 298
 - 13.3.4 容错性讨论 300
 - 13.4 流式计算引擎对比 303
 - 13.5 小结 304
 - 13.6 本章问题 304
- 第六部分 数据分析篇
- 第14章 数据分析语言HQL与SQL 308
 - 14.1 概述 308
 - 14.1.1 背景 308
 - 14.1.2 SQL On Hadoop 309
 - 14.2 Hive架构 309
 - 14.2.1 Hive基本架构 310
 - 14.2.2 Hive查询引擎 311
 - 14.3 Spark SQL架构 312
 - 14.3.1 Spark SQL基本架构 312
 - 14.3.2 Spark SQL与Hive对比 313
 - 14.4 HQL 314
 - 14.4.1 HQL基本语法 314
 - 14.4.2 HQL应用实例 320
 - 14.5 小结 322
 - 14.6 本章问题 322
- 第15章 大数据统一编程模型 325
 - 15.1 产生背景 325
 - 15.2 Apache Beam基本构成 327
 - 15.2.1 Beam SDK 327
 - 15.2.2 Beam Runner 328
 - 15.3 Apache Beam编程模型 329
 - 15.3.1 构建Pipeline 330
 - 15.3.2 创建PCollection 331
 - 15.3.3 使用Transform 334
 - 15.3.4 side input与side output 340
 - 15.4 Apache Beam流式计算模型 341

15.4.1 window简述 342
15.4.2 watermark、trigger与accumulation 344
15.5 Apache Beam编程实例 346
15.5.1 WordCount 346
15.5.2 移动游戏用户行为分析 348
15.6 小结 350
15.7 本章问题 350
第16章 大数据机器学习库 351
16.1 机器学习库简介 351
16.2 MLlib 机器学习库 354
16.2.1 Pipeline 355
16.2.2 特征工程 357
16.2.3 机器学习算法 360
16.3 小结 361
16.4 本章问题 361
• • • • • ([收起](#))

[大数据技术体系详解：原理、架构与实践 下载链接1](#)

标签

大数据

hadoop

分布式

图书馆

架构

技术

数据

图书馆k

评论

太基础了。。

全书16章 作者应该是技术出身 全书代码 技术架构等 较为专业

入门挺好，深入还得继续看其他的书

适合快速入门了解

适合入门，了解整个大数据技术体系，介绍得比较全面，易懂

体系比较完善，没有太多废话，不是堆代码，适合大数据从业人员中级以下水平的读物

对了解大数据的总体架构和主要组建有帮助

比较基础，适合入门对大数据框架各个层次进行了介绍。

教科书

[大数据技术体系详解：原理、架构与实践_下载链接1_](#)

[大数据技术体系详解：原理、架构与实践_下载链接1](#)