《Spark编程基础（Python版）》

教材官网： http://dblab.xmu.edu.cn/post/spark-python/

温馨提示：编辑幻灯片母版，可以修改每页PPT的厦大校徽和底部文字

第6章 Spark Streaming

（PPT版本号：2020年1月版）

林子雨

厦门大学计算机科学系

E-mail: ziyulin@xmu.edu.cn

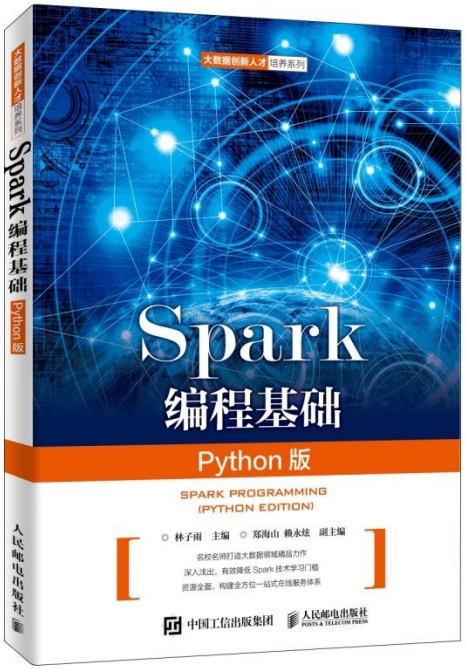
主页： **http://dblab.xmu.edu.cn/post/linziyu**

扫一扫访问教材官网

课程教材

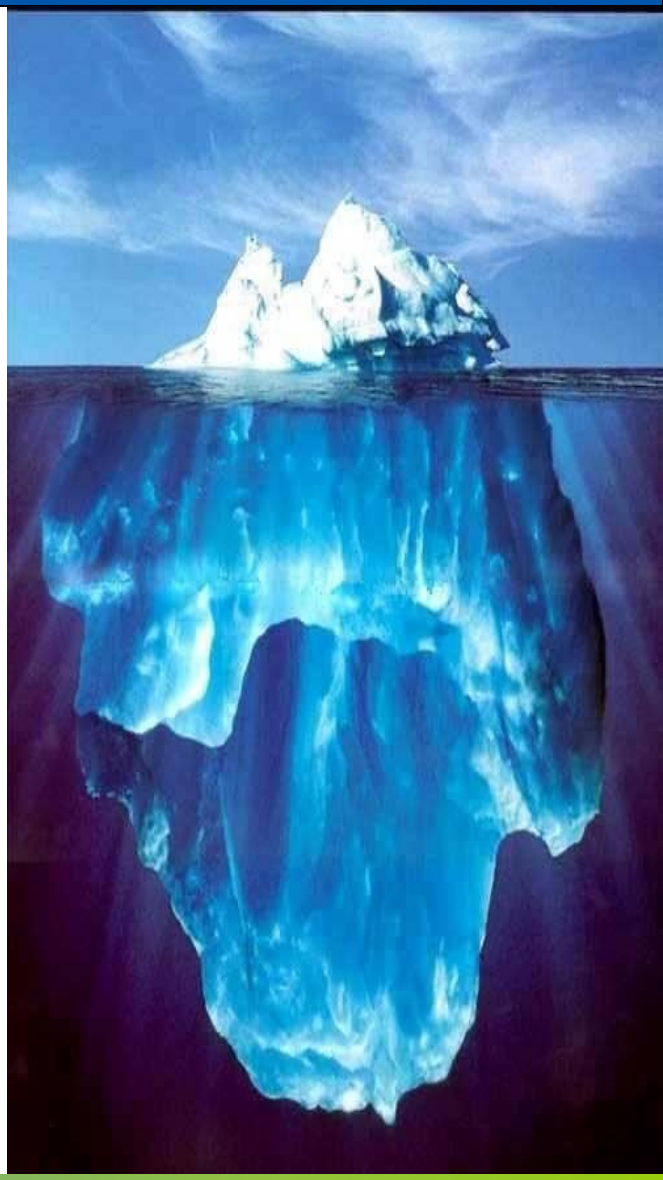
林子雨，郑海山，赖永炫 编著《Spark编程基础（Python版）》

教材官网：http://dblab.xmu.edu.cn/post/spark-python/
ISBN:978-7-115-52439-3 人民邮电出版社



本书以Python作为开发Spark应用程序的编程语言，系统介绍了Spark编程的基础知识。全书共8章，内容包括大数据技术概述、Spark的设计与运行原理、Spark环境搭建和使用方法、RDD编程、Spark SQL、Spark Streaming、Structured Streaming、Spark MLlib等。本书每个章节都安排了入门级的编程实践操作，以便读者更好地学习和掌握Spark编程方法。本书官网免费提供了全套的在线教学资源，包括讲义PPT、习题、源代码、软件、数据集、授课视频、上机实验指南等。

提纲

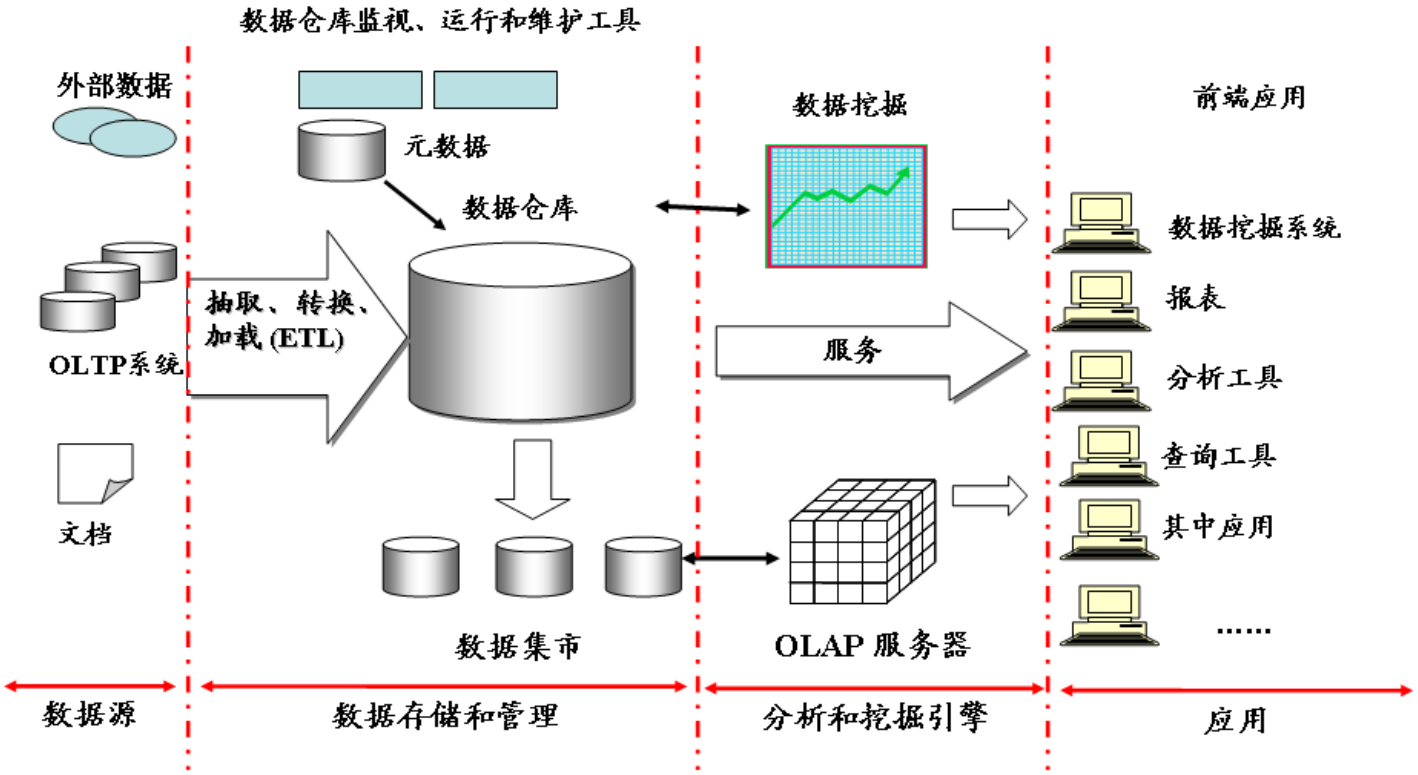


6.1 流计算概述

- 6.1.1 静态数据和流数据
- 6.1.2 批量计算和实时计算
- 6.1.3 流计算概念
- 6.1.4 流计算框架
- 6.1.5 流计算处理流程

6.1.1 静态数据和流数据

- 很多企业为了支持决策分析而构建的数据仓库系统，其中存放的大量历史数据就是静态数据。技术人员可以利用数据挖掘和OLAP（On-Line Analytical Processing）分析工具从静态数据中找到对企业有价值的信息



6.1.1 静态数据和流数据

- 近年来，在Web应用、网络监控、传感监测等领域，兴起了一种新的数据密集型应用——流数据，即数据以大量、快速、时变的流形式持续到达
- 实例：PM2.5检测、电子商务网站用户点击流

流数据具有如下特征：

- 数据快速持续到达，潜在大小也许是无穷无尽的
- 数据来源众多，格式复杂
- 数据量大，但是不十分关注存储，一旦经过处理，要么被丢弃，要么被归档存储
- 注重数据的整体价值，不过分关注个别数据
- 数据顺序颠倒，或者不完整，系统无法控制将要处理的新到达的数据元素的顺序

6.1.2 批量计算和实时计算

- 对静态数据和流数据的处理，对应着两种截然不同的计算模式：批量计算和实时计算

•批量计算：充裕时间处理静态数据，如Hadoop

•流数据不适合采用批量计算，因为流数据不适合用传统的关系模型建模

•流数据必须采用实时计算，响应时间为秒级

•数据量少时，不是问题，但是，在大数据时代，数据格式复杂、来源众多、数据量巨大，对实时计算提出了很大的挑战。因此，针对流数据的实时计算——流计算，应运而生

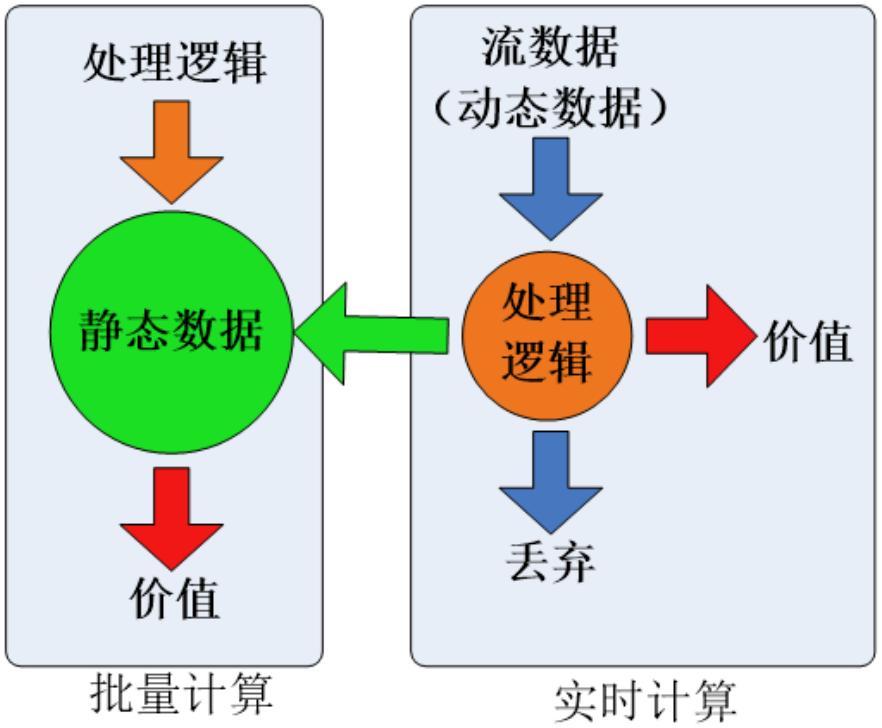


图 数据的两种处理模型

6.1.3 流计算概念

- 流计算：实时获取来自不同数据源的海量数据，经过实时分析处理，获得有价值的信息

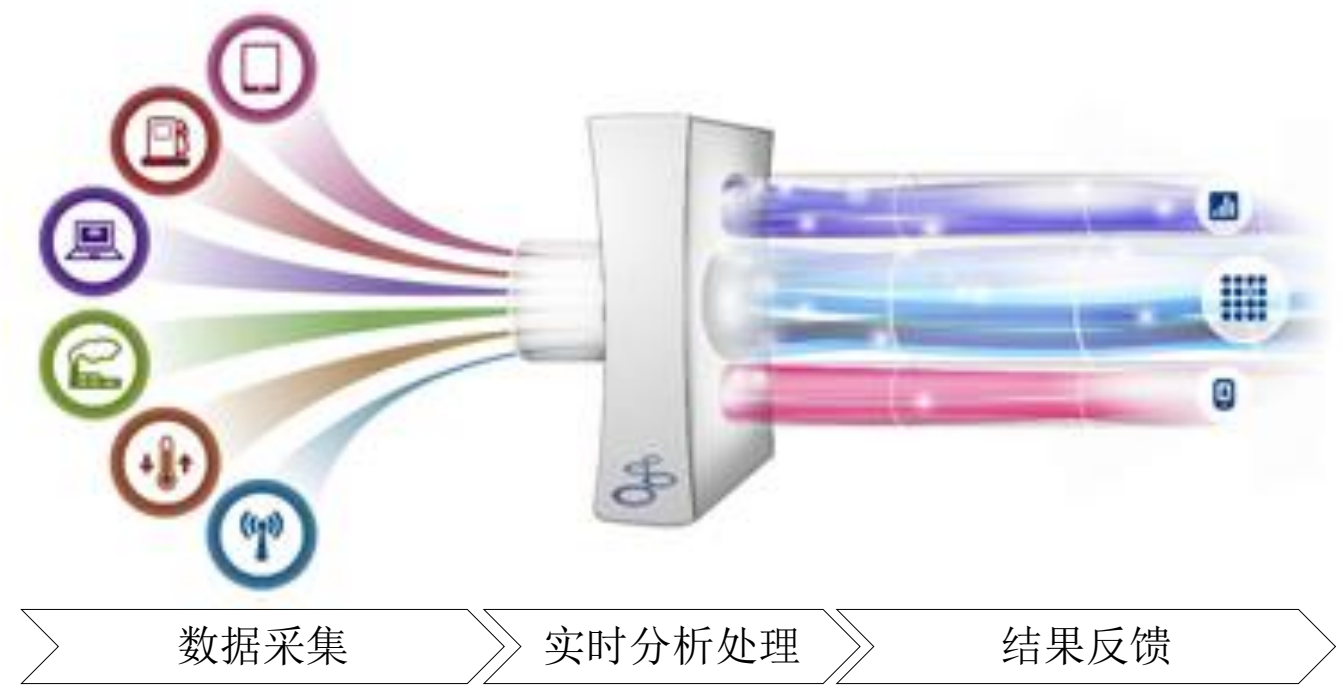


图 流计算示意图

6.1.3 流计算概念

- 流计算秉承一个基本理念，即**数据的价值随着时间的流逝而降低**，如用户点击流。因此，当事件出现时就应该立即进行处理，而不是缓存起来进行批量处理。为了及时处理流数据，就需要一个低延迟、可扩展、高可靠的处理引擎

对于一个流计算系统来说，它应达到如下需求：

- 高性能：处理大数据的基本要求，如每秒处理几十万条数据
- 海量式：支持TB级甚至是PB级的数据规模
- 实时性：保证较低的延迟时间，达到秒级别，甚至是毫秒级别
- 分布式：支持大数据的基本架构，必须能够平滑扩展
- 易用性：能够快速进行开发和部署
- 可靠性：能可靠地处理流数据

6.1.4 流计算框架

- 当前业界诞生了许多专门的流数据实时计算系统来满足各自需求
- 目前有三类常见的流计算框架和平台：商业级的流计算平台、开源流计算框架、公司为支持自身业务开发的流计算框架
- 商业级：IBM InfoSphere Streams和IBM StreamBase
- 较为常见的是开源流计算框架，代表如下：
  - Twitter Storm：免费、开源的分布式实时计算系统，可简单、高效、可靠地处理大量的流数据
  - Yahoo! S4（Simple Scalable Streaming System）：开源流计算平台，是通用的、分布式的、可扩展的、分区容错的、可插拔的流式系统
- 公司为支持自身业务开发的流计算框架：
  - Facebook Puma
  - Dstream（百度）
  - 银河流数据处理平台（淘宝）

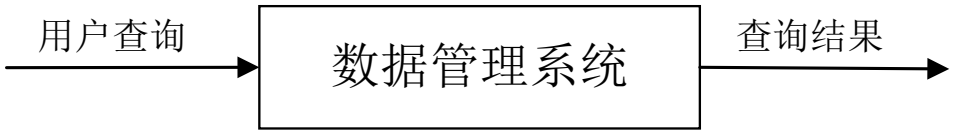
6.1.5 流计算处理流程

- 1. 概述
- 2. 数据实时采集
- 3. 数据实时计算
- 4. 实时查询服务

6.1.5 流计算处理流程

1. 概述

- 传统的数据处理流程，需要先采集数据并存储在关系数据库等数据管理系统中，之后由用户通过查询操作和数据管理系统进行交互

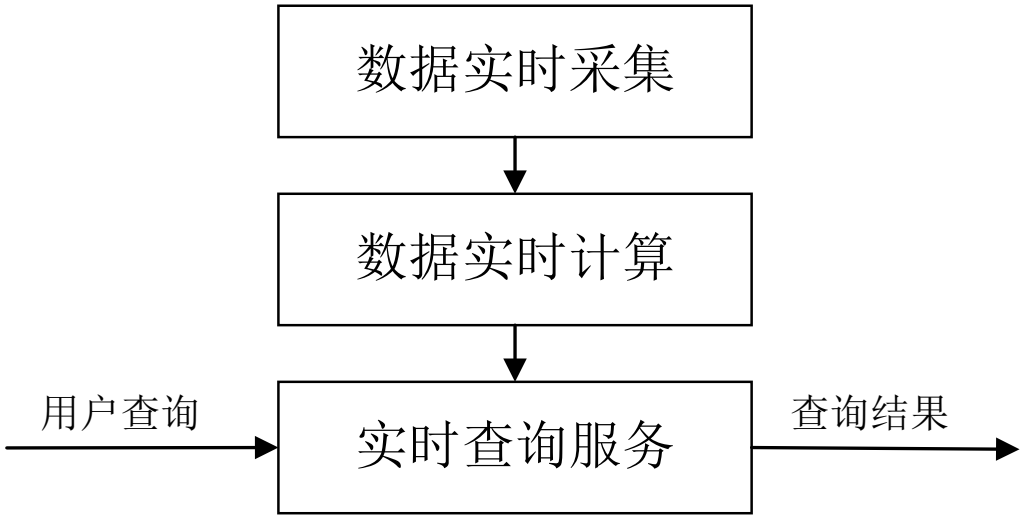


传统的数据处理流程示意图

- 传统的数据处理流程隐含了两个前提：
  - **存储的数据是旧的**。存储的静态数据是过去某一时刻的快照，这些数据在查询时可能已不具备时效性了
  - **需要用户主动发出查询来获取结果**

6.1.5 流计算处理流程

- 流计算的处理流程一般包含三个阶段：数据实时采集、数据实时计算、实时查询服务

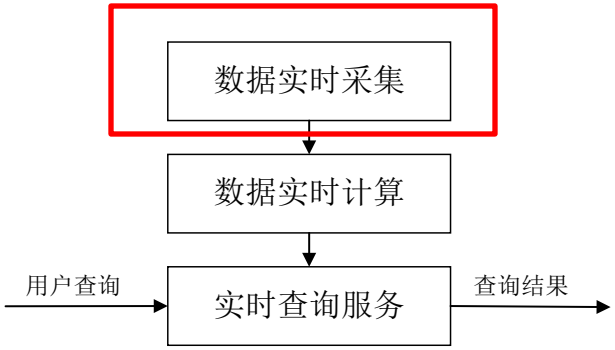


流计算处理流程示意图

6.1.5 流计算处理流程

2. 数据实时采集

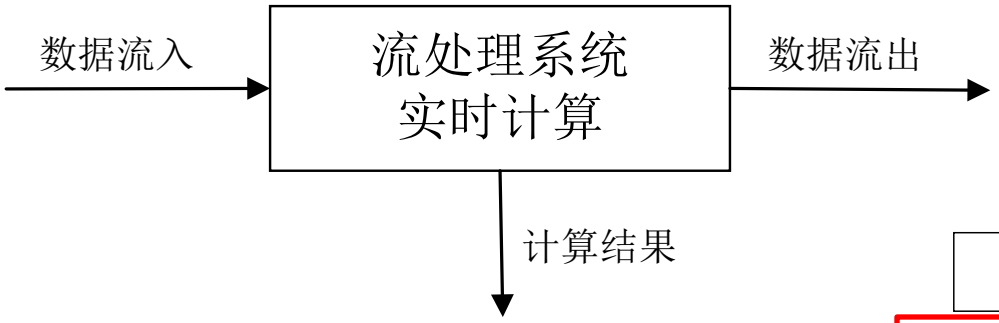
- 数据实时采集阶段通常采集多个数据源的海量数据，需要保证实时性、低延迟与稳定可靠
- 以日志数据为例，由于分布式集群的广泛应用，数据分散存储在不同的机器上，因此需要实时汇总来自不同机器上的日志数据
- 目前有许多互联网公司发布的开源分布式日志采集系统均可满足每秒数百MB的数据采集和传输需求，如：
  - Facebook的Scribe
  - LinkedIn的Kafka
  - 淘宝的Time Tunnel
  - 基于Hadoop的Chukwa和Flume



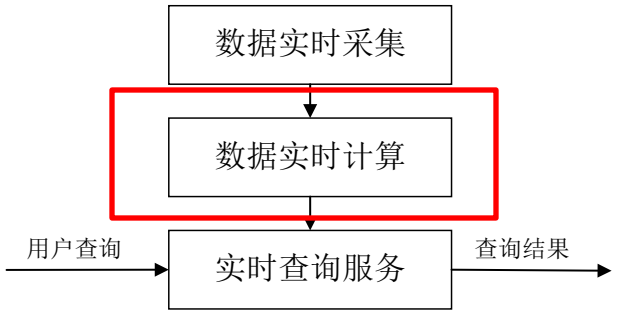
6.1.5 流计算处理流程

3. 数据实时计算

- 数据实时计算阶段对采集的数据进行实时的分析和计算，并反馈实时结果
- 经流处理系统处理后的数据，可视情况进行存储，以便之后再进行分析计算。在时效性要求较高的场景中，处理之后的数据也可以直接丢弃



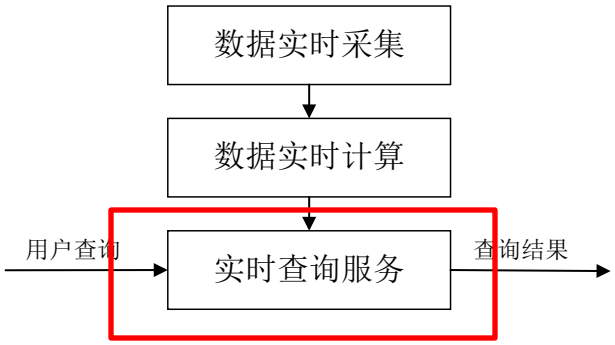
数据实时计算流程



6.1.5 流计算处理流程

4.实时查询服务

- 实时查询服务：经由流计算框架得出的结果可供用户进行实时查询、展示或储存
- 传统的数据处理流程，用户需要主动发出查询才能获得想要的结果。而在流处理流程中，实时查询服务可以不断更新结果，并将用户所需的结果实时推送给用户
- 虽然通过对传统的数据处理系统进行**定时**查询，也可以实现不断地更新结果和结果推送，但通过这样的方式获取的结果，仍然是根据过去某一时刻的数据得到的结果，与实时结果有着本质的区别



6.1.5 流计算处理流程

- 可见，流处理系统与传统的数据处理系统有如下不同：
  - 流处理系统处理的是实时的数据，而传统的数据处理系统处理的是预先存储好的静态数据
  - 用户通过流处理系统获取的是实时结果，而通过传统的数据处理系统，获取的是过去某一时刻的结果
  - 流处理系统无需用户主动发出查询，实时查询服务可以主动将实时结果推送给用户

6.2 Spark Streaming

6.2.1 Spark Streaming设计
6.2.2 Spark Streaming与Storm的对比
6.2.3 从“Hadoop+Storm”架构转向Spark架构

6.2.1 Spark Streaming设计

•Spark Streaming可整合多种输入数据源，如Kafka、Flume、HDFS，甚至是普通的TCP套接字。经处理后的数据可存储至文件系统、数据库，或显示在仪表盘里

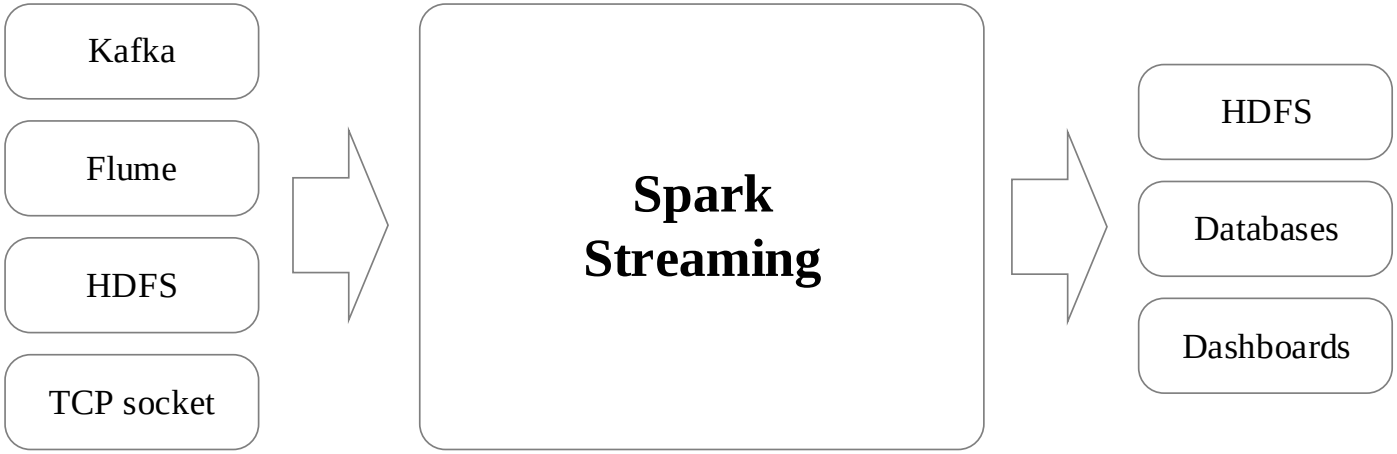


图 Spark Streaming支持的输入、输出数据源

6.2.1 Spark Streaming设计

Spark Streaming的基本原理是将实时输入数据流以时间片（秒级）为单位进行拆分，然后经Spark引擎以类似批处理的方式处理每个时间片数据

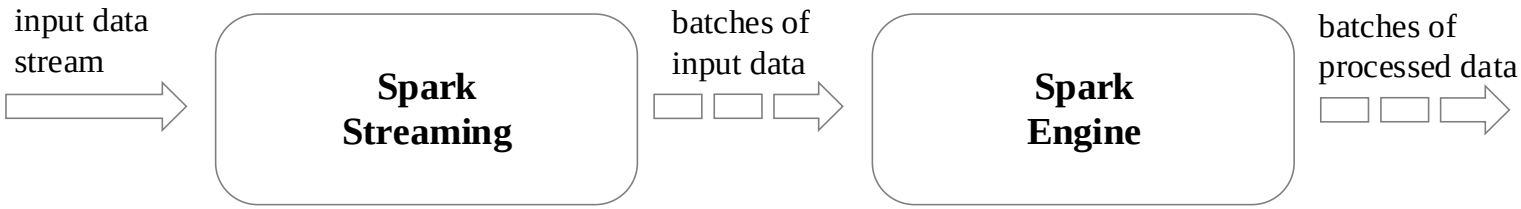


图 Spark Streaming执行流程

6.2.1 Spark Streaming设计

Spark Streaming最主要的抽象是DStream（Discretized Stream，离散化数据流），表示连续不断的数据流。在内部实现上，Spark Streaming的输入数据按照时间片（如1秒）分成一段一段，每一段数据转换为Spark中的RDD，这些分段就是Dstream，并且对DStream的操作都最终转变为对相应的RDD的操作

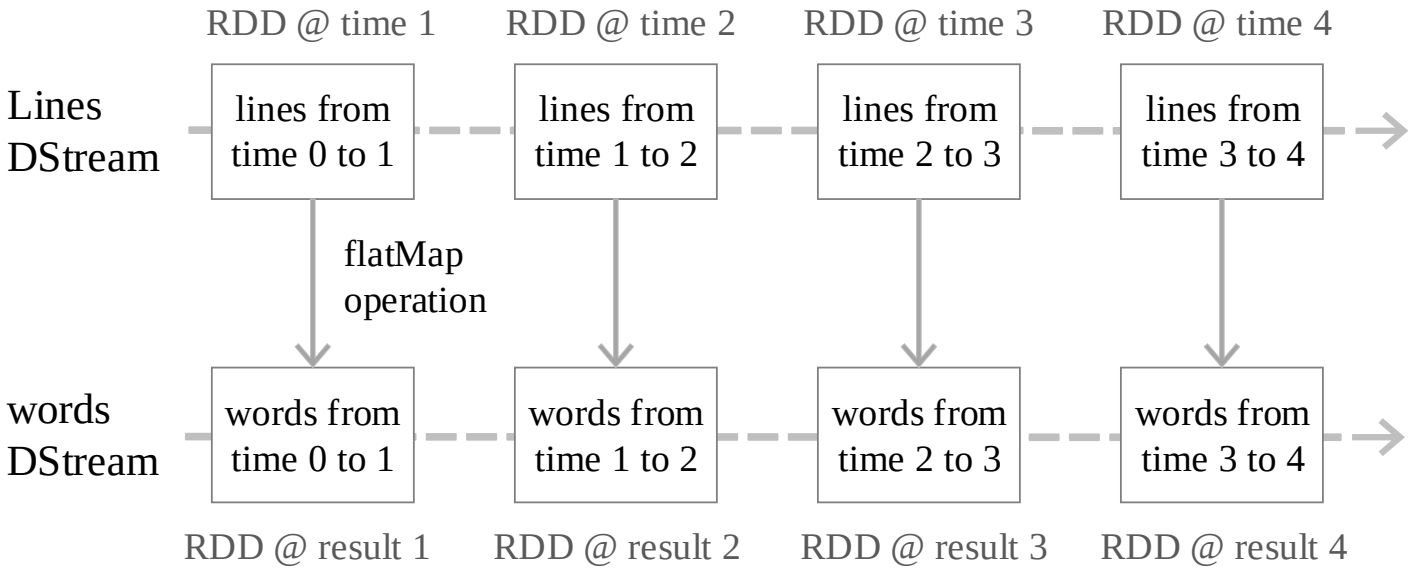


图 DStream操作示意图

6.2.2 Spark Streaming与Storm的对比

•Spark Streaming和Storm最大的区别在于，Spark Streaming无法实现毫秒级的流计算，而Storm可以实现毫秒级响应

•Spark Streaming构建在Spark上，一方面是因为Spark的低延迟执行引擎（100ms+）可以用于实时计算，另一方面，相比于Storm，RDD数据集更容易做高效的容错处理

•Spark Streaming采用的小批量处理的方式使得它可以同时兼容批量和实时数据处理的逻辑和算法，因此，方便了一些需要历史数据和实时数据联合分析的特定应用场合

6.2.3 从“Hadoop+Storm”架构转向Spark架构

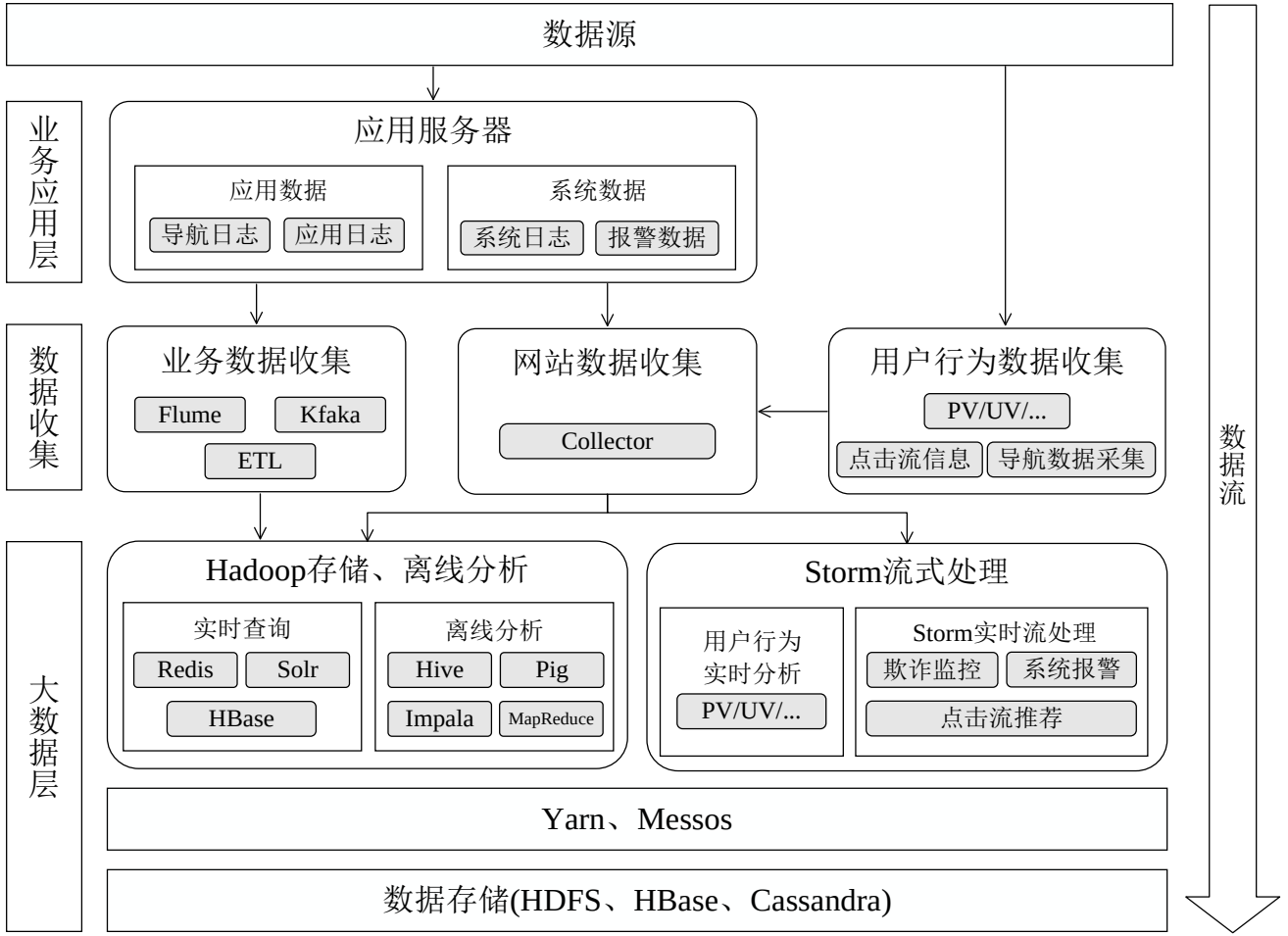


图 采用Hadoop+Storm部署方式的一个案例

6.2.3 从“Hadoop+Storm”架构转向Spark架构

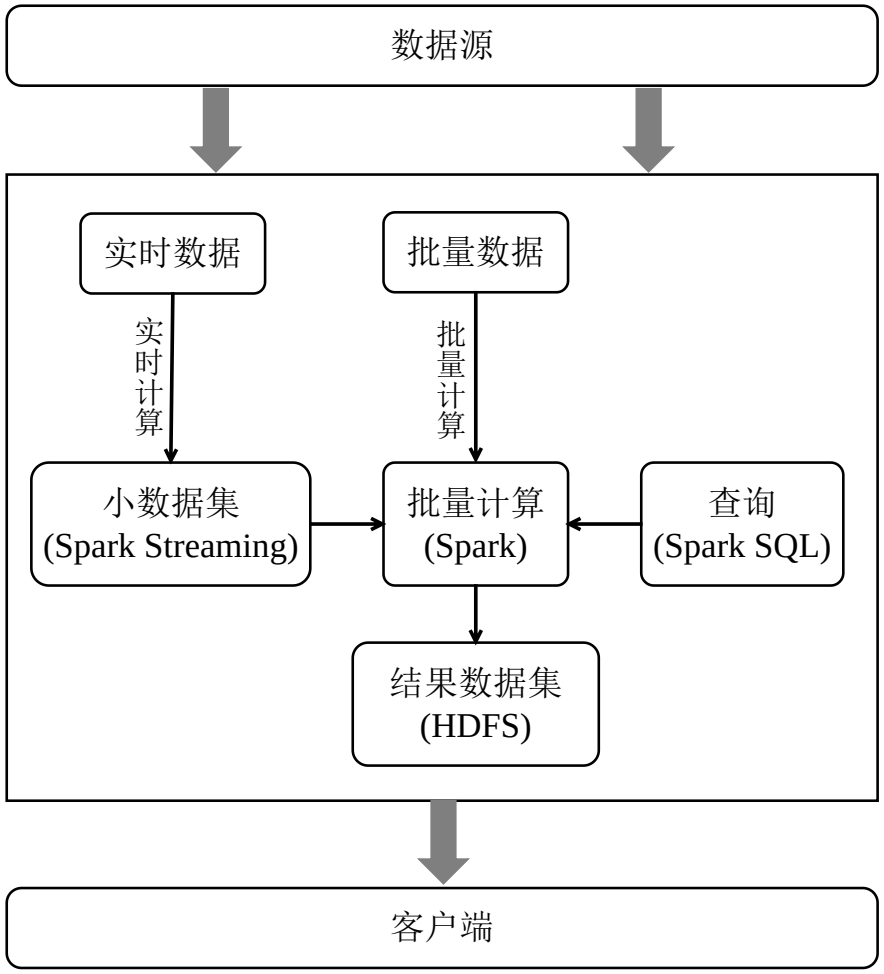


图　用Spark架构满足批处理和流处理需求

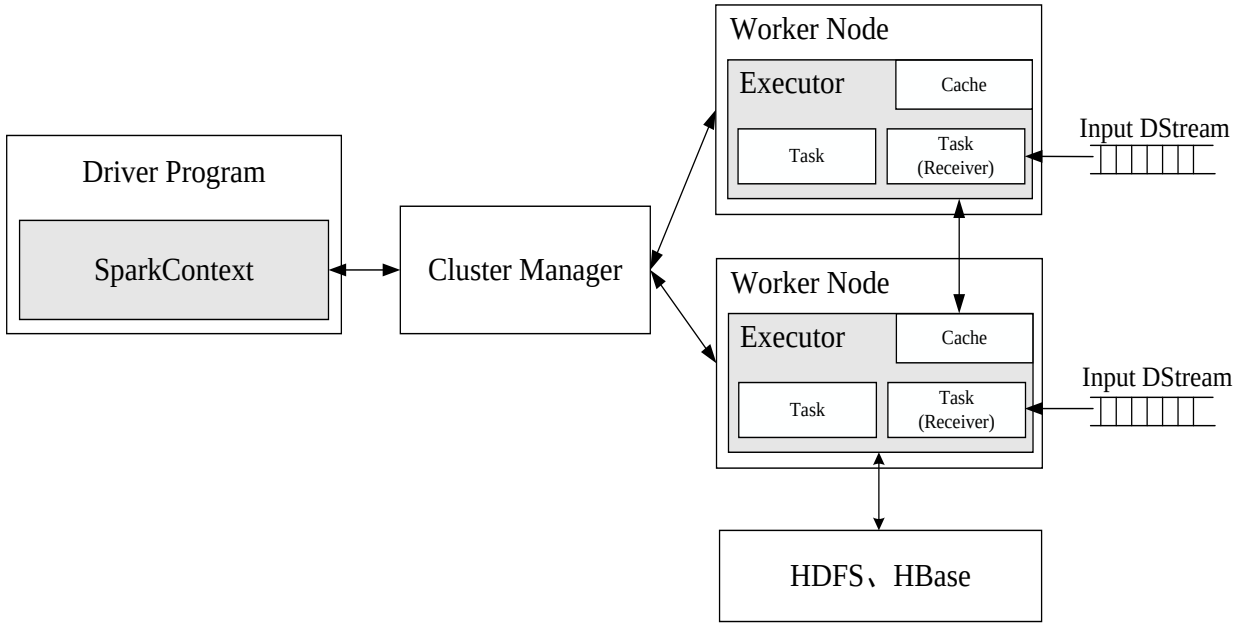
采用Spark架构具有如下优点：

- 实现一键式安装和配置、线程级别的任务监控和告警；
- 降低硬件集群构建、软件维护、任务监控和应用开发的难度；
- 便于做成统一的硬件、计算平台资源池。

6.3 DStream操作概述

6.3.1 Spark Streaming工作机制
6.3.2 Spark Streaming程序的基本步骤
6.3.3 创建StreamingContext对象

6.3.1 Spark Streaming工作机制



- 在Spark Streaming中，会有一个组件Receiver，作为一个长期运行的task跑在一个Executor上
- 每个Receiver都会负责一个input DStream（比如从文件中读取数据的文件流，比如套接字流，或者从Kafka中读取的一个输入流等等）
- Spark Streaming通过input DStream与外部数据源进行连接，读取相关数据

6.3.2 Spark Streaming程序的基本步骤

编写Spark Streaming程序的基本步骤是：

1.通过创建输入DStream来定义输入源

2.通过对DStream应用转换操作和输出操作来定义流计算

3.用streamingContext.start()来开始接收数据和处理流程

4.通过streamingContext.awaitTermination()方法来等待处理结束（手动结束或因为错误而结束）

5.可以通过streamingContext.stop()来手动结束流计算进程

6.3.3 创建StreamingContext对象

•如果要运行一个Spark Streaming程序，就需要首先生成一个StreamingContext对象，它是Spark Streaming程序的主入口

•可以从一个SparkConf对象创建一个StreamingContext对象。
•在pyspark中的创建方法：进入pyspark以后，就已经获得了一个默认的SparkConext对象，也就是sc。因此，可以采用如下方式来创建StreamingContext对象：

```
>>> from pyspark.streaming import StreamingContext
>>> ssc = StreamingContext(sc, 1)
```

6.3.3 创建StreamingContext对象

如果是编写一个独立的Spark Streaming程序，而不是在pyspark中运行，则需要通过如下方式创建StreamingContext对象：

```
from pyspark import SparkContext, SparkConf
from pyspark.streaming import StreamingContext
conf = SparkConf()
conf.setAppName('TestDStream')
conf.setMaster('local[2]')
sc = SparkContext(conf = conf)
ssc = StreamingContext(sc, 1)
```

6.4 基本输入源

6.4.1 文件流
6.4.2 套接字流
6.4.3 RDD队列流

6.4.1 文件流

1.在pyspark中创建文件流

```
$ cd /usr/local/spark/mycode
$ mkdir streaming
$ cd streaming
$ mkdir logfile
$ cd logfile
```

6.4.1 文件流

进入pyspark创建文件流。请另外打开一个终端窗口，启动进入pyspark

```
>>> from pyspark import SparkContext
>>> from pyspark.streaming import StreamingContext
>>> ssc = StreamingContext(sc, 10)
>>> lines = ssc. \
... textFileStream('file:///usr/local/spark/mycode/streaming/logfile')
>>> words = lines.flatMap(lambda line: line.split(' '))
>>> wordCounts = words.map(lambda x :
(x,1)).reduceByKey(lambda a,b:a+b)
>>> wordCounts.pprint()
>>> ssc.start()
>>> ssc.awaitTermination()
```

6.4.1 文件流

上面在pyspark中执行的程序，一旦你输入ssc.start()以后，程序就开始自动进入循环监听状态，屏幕上会显示一堆的信息，如下：

```
-------------------------------------------
Time: 2018-12-30 15:35:30
-------------------------------------------


-------------------------------------------
Time: 2018-12-30 15:35:40
-------------------------------------------


-------------------------------------------
Time: 2018-12-30 15:35:50
-------------------------------------------
```

在“/usr/local/spark/mycode/streaming/logfile”目录下新建一个log.txt文件，就可以在监听窗口中显示词频统计结果

6.4.1 文件流

2. 采用独立应用程序方式创建文件流

```
$ cd /usr/local/spark/mycode
$ cd streaming
$ cd logfile
$ vim FileStreaming.py
```

6.4.1 文件流

用vim编辑器新建一个FileStreaming.py代码文件，请在里面输入以下代码：

```
#!/usr/bin/env python3

from pyspark import SparkContext, SparkConf
from pyspark.streaming import StreamingContext

conf = SparkConf()
conf.setAppName('TestDStream')
conf.setMaster('local[2]')
sc = SparkContext(conf = conf)
ssc = StreamingContext(sc, 10)
lines = ssc.textFileStream('file:///usr/local/spark/mycode/streaming/logfile')
words = lines.flatMap(lambda line: line.split(' '))
wordCounts = words.map(lambda x : (x,1)).reduceByKey(lambda a,b:a+b)
wordCounts.pprint()
ssc.start()
ssc.awaitTermination()
```

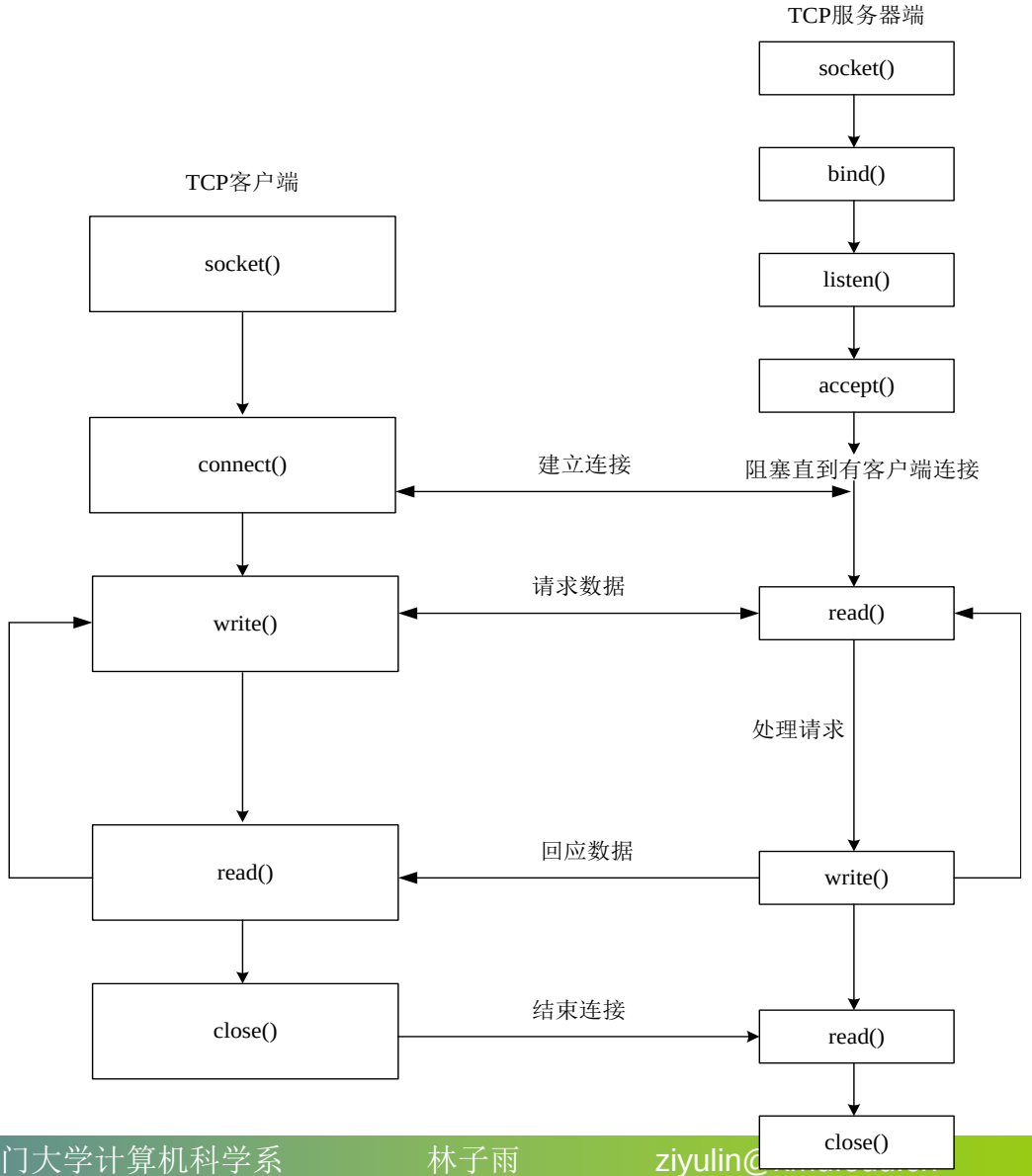
6.4.1 文件流

```
$ cd /usr/local/spark/mycode/streaming/logfile/
$ /usr/local/spark/bin/spark-submit FileStreaming.py
```

6.4.2 套接字流

•Spark Streaming可以通过Socket端口监听并接收数据，然后进行相应处理

1.Socket工作原理



6.4.2 套接字流

2.使用套接字流作为数据源

```
$ cd /usr/local/spark/mycode
$ mkdir streaming #如果已经存在该目录，则不用创建
$ cd streaming
$ mkdir socket
$ cd socket
$ vim NetworkWordCount.py
```

6.4.2 套接字流

请在NetworkWordCount.py文件中输入如下内容：

```
#!/usr/bin/env python3

from __future__ import print_function
import sys
from pyspark import SparkContext
from pyspark.streaming import StreamingContext

if __name__ == "__main__":
    if len(sys.argv) != 3:
        print("Usage: NetworkWordCount.py <hostname> <port>", file=sys.stderr)
        exit(-1)
    sc = SparkContext(appName="PythonStreamingNetworkWordCount")
    ssc = StreamingContext(sc, 1)
    lines = ssc.socketTextStream(sys.argv[1], int(sys.argv[2]))
    counts = lines.flatMap(lambda line: line.split(" ")) \
                  .map(lambda word: (word, 1)) \
                  .reduceByKey(lambda a, b: a+b)
    counts.pprint()
    ssc.start()
    ssc.awaitTermination()
```

6.4.2 套接字流

新打开一个窗口作为nc窗口，启动nc程序：

```
$ nc -lk 9999
```

再新建一个终端（记作“流计算终端”），执行如下代码启动流计算：

```
$ cd /usr/local/spark/mycode/streaming/socket
$ /usr/local/spark/bin/spark-submit NetworkWordCount.py localhost 9999
```

•可以在nc窗口中随意输入一些单词，监听窗口就会自动获得单词数据流信息，在监听窗口每隔1秒就会打印出词频统计信息，大概会在屏幕上出现类似如下的结果：

```
-------------------------------------------
Time: 2018-12-24 11:30:26
-------------------------------------------
('Spark', 1)
('love', 1)
('I', 1)
(spark,1)
```

6.4.2 套接字流

3. 使用Socket编程实现自定义数据源

- 下面我们再前进一步，把数据源头的产生方式修改一下，不要使用nc程序，而是采用自己编写的程序产生Socket数据源

```
$ cd /usr/local/spark/mycode/streaming/socket
$ vim DataSourceSocket.py
```

6.4.2 套接字流

```
#!/usr/bin/env python3
import socket
# 生成socket对象
server = socket.socket()
# 绑定ip和端口
server.bind(('localhost', 9999))
# 监听绑定的端口
server.listen(1)
while 1:
    # 为了方便识别，打印一个“我在等待”
    print("I'm waiting the connect...")
    # 这里用两个值接受，因为连接上之后使用的是客户端发来请求的这个实例
    # 所以下面的传输要使用conn实例操作
    conn,addr = server.accept()
    # 打印连接成功
    print("Connect success! Connection is from %s " % addr[0])
    # 打印正在发送数据
    print('Sending data...')
    conn.send('I love hadoop I love spark hadoop is good spark is fast'.encode())
    conn.close()
    print('Connection is broken.‘)
```

6.4.2 套接字流

执行如下命令启动Socket服务端：

```
$ cd  /usr/local/spark/mycode/streaming/socket
$ /usr/local/spark/bin/spark-submit DataSourceSocket.py
```

启动客户端，即NetworkWordCount程序。新建一个终端（记作“流计算终端”），输入以下命令启动NetworkWordCount程序：

```
$ cd  /usr/local/spark/mycode/streaming/socket
$ /usr/local/spark/bin/spark-submit NetworkWordCount.py localhost 9999
```

```
-------------------------------------------
Time: 2018-12-30 15:16:17
-------------------------------------------
('good', 1)
('hadoop', 2)
('is', 2)
('love', 2)
('spark', 2)
('I', 2)
('fast', 1)
```

6.4.3 RDD队列流

- 在调试Spark Streaming应用程序的时候，我们可以使用streamingContext.queueStream(queueOfRDD)创建基于RDD队列的DStream
- 新建一个RDDQueueStream.py代码文件，功能是：每隔1秒创建一个RDD，Streaming每隔2秒就对数据进行处理

6.4.3 RDD队列流

```
#!/usr/bin/env python3

import time
from pyspark import SparkContext
from pyspark.streaming import StreamingContext

if __name__ == "__main__":
    sc = SparkContext(appName="PythonStreamingQueueStream")
    ssc = StreamingContext(sc, 2)
    #创建一个队列，通过该队列可以把RDD推给一个RDD队列流
    rddQueue = []
    for i in range(5):
        rddQueue += [ssc.sparkContext.parallelize([j for j in range(1, 1001)], 10)]
        time.sleep(1)
    #创建一个RDD队列流
    inputStream = ssc.queueStream(rddQueue)
    mappedStream = inputStream.map(lambda x: (x % 10, 1))
    reducedStream = mappedStream.reduceByKey(lambda a, b: a + b)
    reducedStream.pprint()
    ssc.start()
    ssc.stop(stopSparkContext=True, stopGraceFully=True)
```

6.4.3 RDD队列流

下面执行如下命令运行该程序：

```
$ cd  /usr/local/spark/mycode/streaming/rddqueue
$ /usr/local/spark/bin/spark-submit RDDQueueStream.py
```

```
-------------------------------------------
Time: 2018-12-31 15:42:15
-------------------------------------------
(0, 100)
(8, 100)
(2, 100)
(4, 100)
(6, 100)
(1, 100)
(3, 100)
(9, 100)
(5, 100)
(7, 100)
```

6.5 高级数据源

6.5.1 Kafka简介
6.5.2 Kafka准备工作
6.5.3 Spark准备工作
6.5.4 编写Spark Streaming程序使用Kafka数据源

6.5.1 Kafka简介

- Kafka是一种高吞吐量的分布式发布订阅消息系统，用户通过Kafka系统可以发布大量的消息，同时也能实时订阅消费消息
- Kafka可以同时满足在线实时处理和批量离线处理

- 在公司的大数据生态系统中，可以把Kafka作为数据交换枢纽，不同类型的分布式系统（关系数据库、NoSQL数据库、流处理系统、批处理系统等），可以统一接入到Kafka，实现和Hadoop各个组件之间的不同类型数据的实时高效交换

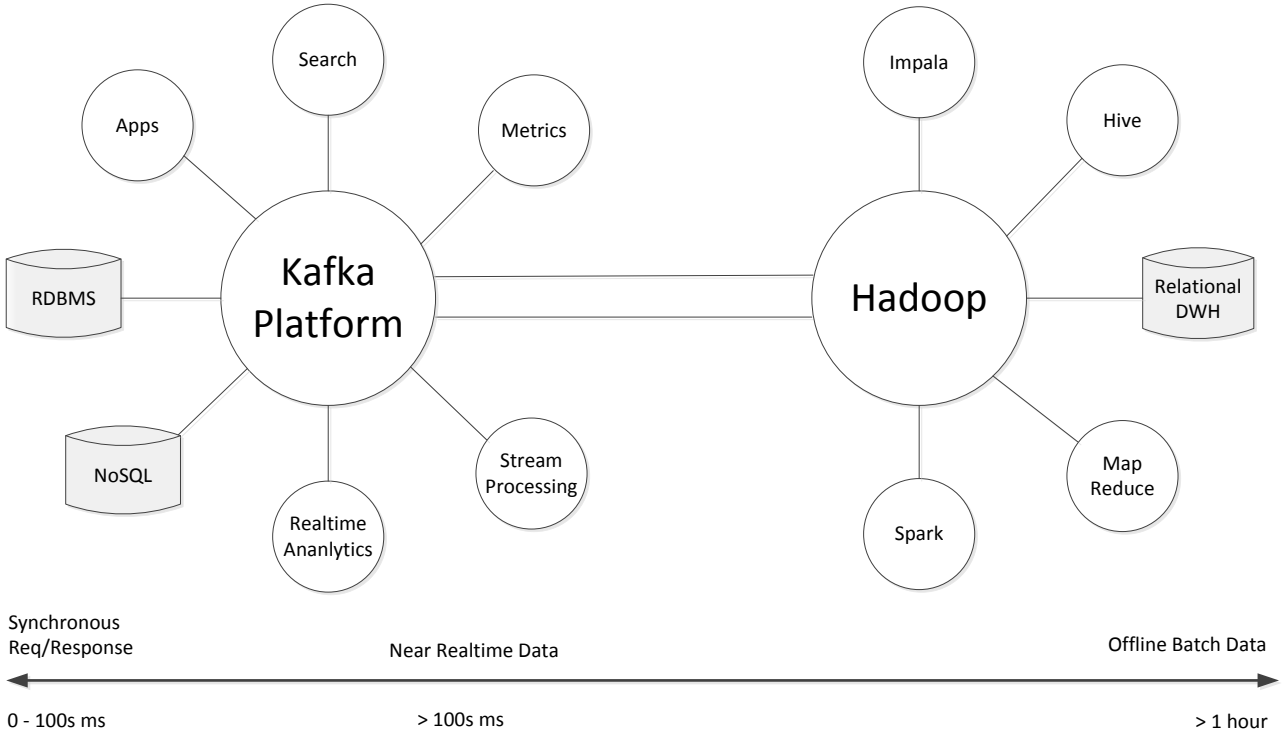


图 Kafka作为数据交换枢纽

6.5.1 Kafka简介

•Broker
Kafka集群包含一个或多个服务器，这种服务器被称为broker

•Topic
每条发布到Kafka集群的消息都有一个类别，这个类别被称为Topic。（物理上不同Topic的消息分开存储，逻辑上一个Topic的消息虽然保存于一个或多个broker上，但用户只需指定消息的Topic即可生产或消费数据而不必关心数据存于何处）

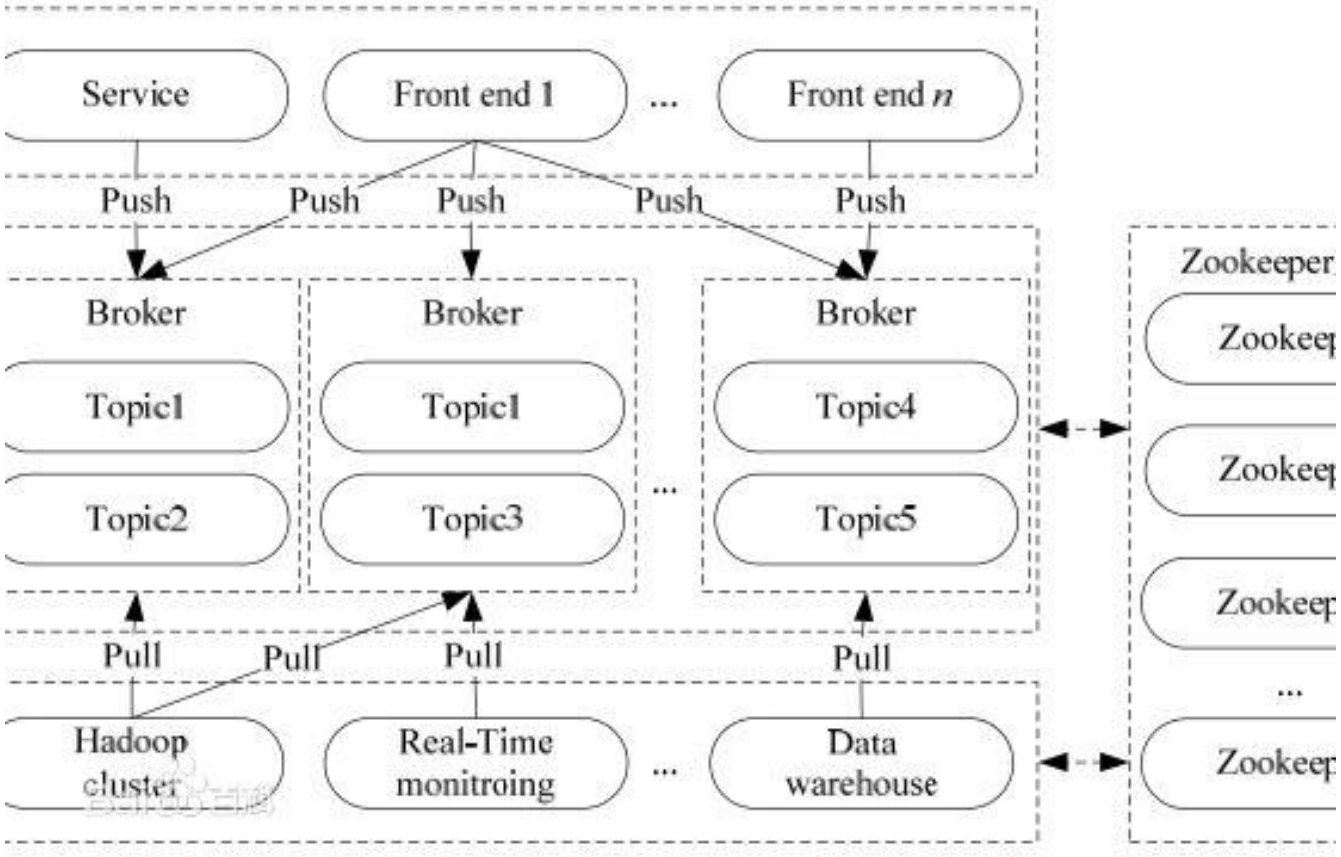
•Partition
Partition是物理上的概念，每个Topic包含一个或多个Partition.

•Producer
负责发布消息到Kafka broker

•Consumer
消息消费者，向Kafka broker读取消息的客户端。

•Consumer Group
每个Consumer属于一个特定的Consumer Group（可为每个Consumer指定group name，若不指定group name则属于默认的group）

6.5.1 Kafka简介



Kafka的运行依赖于Zookeeper。Topic、Consumer、Patition、Broker等注册信息都存储在ZooKeeper中。

6.5.2 Kafka准备工作

1.安装Kafka
2.启动Kafka
3.测试Kafka是否正常工作

6.5.2 Kafka准备工作

1.安装Kafka

- 关于Kafka的安装方法，请参考厦门大学数据库实验室建设的高校大数据课程公共服务平台的技术博客文章《Kafka的安装和简单实例测试》
- http://dblab.xmu.edu.cn/blog/1096-2/
- 这里假设已经成功安装Kafka到“/usr/local/kafka”目录下

高校大数据课程

公 共 服 务 平 台

平台每年访问量超过100万次

6.5.2 Kafka准备工作

1.安装Kafka

说明：本课程下载的安装文件为Kafka_2.11-0.8.2.2.tgz，前面的2.11就是该Kafka所支持的Scala版本号，后面的0.8.2.2是Kafka自身的版本号

6.5.2 Kafka准备工作

2.启动Kafka

打开一个终端，输入下面命令启动Zookeeper服务：

```
$ cd /usr/local/kafka
$ ./bin/zookeeper-server-start.sh config/zookeeper.properties
```

千万不要关闭这个终端窗口，一旦关闭，Zookeeper服务就停止了

6.5.2 Kafka准备工作

打开第二个终端，然后输入下面命令启动Kafka服务：

```
$ cd /usr/local/kafka
$ bin/kafka-server-start.sh config/server.properties
```

千万不要关闭这个终端窗口，一旦关闭，Kafka服务就停止了

6.5.2 Kafka准备工作

3.测试Kafka是否正常工作

再打开第三个终端，然后输入下面命令创建一个自定义名称为“wordsendertest”的Topic：

```
$ cd  /usr/local/kafka
$ ./bin/kafka-topics.sh  --create  --zookeeper  localhost:2181  \
>--replication-factor  1  --partitions  1  --topic  wordsendertest
#可以用list列出所有创建的Topic，验证是否创建成功
$ ./bin/kafka-topics.sh  --list  --zookeeper  localhost:2181
```

replication-factor：每个partition的副本个数

6.5.2 Kafka准备工作

下面用生产者（Producer）来产生一些数据，请在当前终端（记作“数据源终端”）内继续输入下面命令：

```
$ ./bin/kafka-console-producer.sh  --broker-list  localhost:9092 \
>  --topic  wordsendertest
```

上面命令执行后，就可以在当前终端内用键盘输入一些英文单词，比如可以输入：

hello hadoop
hello spark

6.5.2 Kafka准备工作

现在可以启动一个消费者，来查看刚才生产者产生的数据。请另外打开第四个终端，输入下面命令：

```
$ cd /usr/local/kafka
$./bin/kafka-console-consumer.sh  --zookeeper  localhost:2181  \
> --topic  wordsendertest  --from-beginning
```

可以看到，屏幕上会显示出如下结果，也就是刚才在另外一个终端里面输入的内容：

hello hadoop
hello spark

6.5.3 Spark准备工作

1.添加相关jar包

Kafka和Flume等高级输入源，需要依赖独立的库（jar文件）

对于Spark2.4.0版本，如果要使用Kafka，则需要下载spark-streaming-kafka-0-8_2.11相关jar包

spark-streaming-kafka-0-8_2.11-2.4.0.jar

下载地址：
http://mvnrepository.com/artifact/org.apache.spark/spark-streaming-kafka-0-8_2.11/2.4.0

6.5.3 Spark准备工作

把jar文件复制到Spark目录的jars目录下

```
$ cd /usr/local/spark/jars
$ mkdir kafka
$ cd ~
$ cd 下载
$ cp ./spark-streaming-kafka-0-8_2.11-2.4.0.jar
/usr/local/spark/jars/kafka
```

继续把Kafka安装目录的libs目录下的所有jar文件复制到“/usr/local/spark/jars/kafka”目录下，请在终端中执行下面命令：

```
$ cd /usr/local/kafka/libs
$ cp ./* /usr/local/spark/jars/kafka
```

6.5.3 Spark准备工作

然后，修改Spark配置文件，命令如下：

```
$ cd  /usr/local/spark/conf
$ vim spark-env.sh
```

把Kafka相关jar包的路径信息增加到spark-env.sh，修改后的spark-env.sh类似如下：

```
export SPARK_DIST_CLASSPATH=$(/usr/local/hadoop/bin/hadoop classpath):$(/usr/local/hbase/bin/hbase classpath):/usr/local/spark/jars/hbase/*:/usr/local/spark/examples/jars/*:/usr/local/spark/jars/kafka/*:/usr/local/kafka/libs/*
```

6.5.4 编写Spark Streaming程序使用Kafka数据源

KafkaWordCount.py

```
#!/usr/bin/env python3

from __future__ import print_function
import sys
from pyspark import SparkContext
from pyspark.streaming import StreamingContext
from pyspark.streaming.kafka import KafkaUtils

if __name__ == "__main__":
    if len(sys.argv) != 3:
        print("Usage: KafkaWordCount.py <zk> <topic>", file=sys.stderr)
        exit(-1)
    sc = SparkContext(appName="PythonStreamingKafkaWordCount")
    ssc = StreamingContext(sc, 1)
    zkQuorum, topic = sys.argv[1:]
    kvs = KafkaUtils. \
        createStream(ssc, zkQuorum, "spark-streaming-consumer", {topic: 1})
    lines = kvs.map(lambda x: x[1])
    counts = lines.flatMap(lambda line: line.split(" ")) \
        .map(lambda word: (word, 1)) \
        .reduceByKey(lambda a, b: a+b)
    counts.pprint()
    ssc.start()
    ssc.awaitTermination()
```

6.5.4 编写Spark Streaming程序使用Kafka数据源

然后，新建一个终端（记作“流计算终端”），执行KafkaWordCount.py，命令如下：

```
$ cd  /usr/local/spark/mycode/streaming/kafka/
$ /usr/local/spark/bin/spark-submit  \
> ./KafkaWordCount.py localhost:2181 wordsendertest
```

这时再切换到之前已经打开的“数据源终端”，用键盘手动敲入一些英文单词

在流计算终端内就可以看到类似如下的词频统计动态结果

```
-------------------------------------------
Time: 2018-12-31 10:40:42
-------------------------------------------
('hadoop', 1)

-------------------------------------------
Time: 2018-12-31 10:40:43
-------------------------------------------
('spark', 1)
```

6.6 转换操作

6.6.1 DStream无状态转换操作
6.6.2 DStream有状态转换操作

6.6.1 DStream无状态转换操作

•map(func) ：对源DStream的每个元素，采用func函数进行转换，得到一个新的Dstream

•flatMap(func)： 与map相似，但是每个输入项可用被映射为0个或者多个输出项

•filter(func)： 返回一个新的DStream，仅包含源DStream中满足函数func的项

•repartition(numPartitions)： 通过创建更多或者更少的分区改变DStream的并行程度

•reduce(func)：利用函数func聚集源DStream中每个RDD的元素，返回一个包含单元素RDDs的新DStream

•count()：统计源DStream中每个RDD的元素数量

•union(otherStream)： 返回一个新的DStream，包含源DStream和其他DStream的元素

6.6.1 DStream无状态转换操作

•countByValue()：应用于元素类型为K的DStream上，返回一个（K，V）键值对类型的新DStream，每个键的值是在原DStream的每个RDD中的出现次数

•reduceByKey(func, [numTasks])：当在一个由(K,V)键值对组成的DStream上执行该操作时，返回一个新的由(K,V)键值对组成的DStream，每一个key的值均由给定的recuce函数（func）聚集起来

•join(otherStream, [numTasks])：当应用于两个DStream（一个包含（K,V）键值对,一个包含(K,W)键值对），返回一个包含(K, (V, W))键值对的新Dstream

•cogroup(otherStream, [numTasks])：当应用于两个DStream（一个包含（K,V）键值对,一个包含(K,W)键值对），返回一个包含(K, Seq[V], Seq[W])的元组

•transform(func)：通过对源DStream的每个RDD应用RDD-to-RDD函数，创建一个新的DStream。支持在新的DStream中做任何RDD操作

6.6.1 DStream无状态转换操作

无状态转换操作实例：
之前“套接字流”部分介绍的词频统计，就是采用无状态转换，每次统计，都是只统计当前批次到达的单词的词频，和之前批次无关，不会进行累计

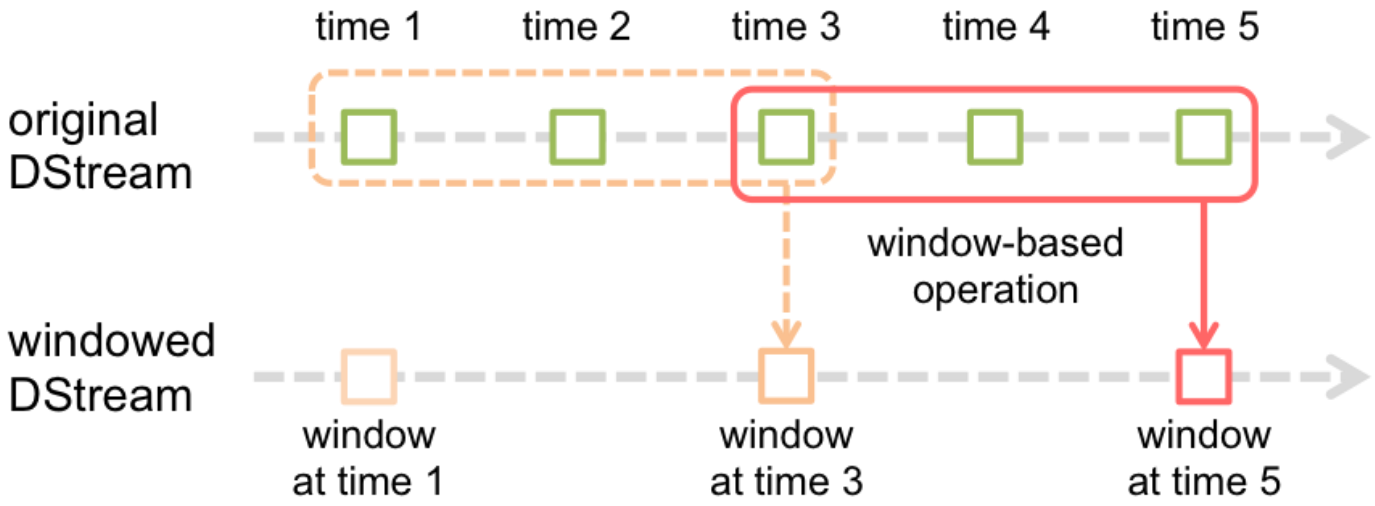
6.6.2 DStream有状态转换操作

1. 滑动窗口转换操作
2. updateStateByKey操作

6.6.2 DStream有状态转换操作

1. 滑动窗口转换操作

•事先设定一个滑动窗口的长度（也就是窗口的持续时间）
•设定滑动窗口的时间间隔（每隔多长时间执行一次计算），让窗口按照指定时间间隔在源DStream上滑动
•每次窗口停放的位置上，都会有一部分Dstream（或者一部分RDD）被框入窗口内，形成一个小段的Dstream
•可以启动对这个小段DStream的计算



6.6.2 DStream有状态转换操作

一些窗口转换操作的含义：

•window(windowLength, slideInterval) 基于源DStream产生的窗口化的批数据，计算得到一个新的Dstream

•countByWindow(windowLength, slideInterval) 返回流中元素的一个滑动窗口数

•reduceByWindow(func, windowLength, slideInterval) 返回一个单元素流。利用函数func聚集滑动时间间隔的流的元素创建这个单元素流。函数func必须满足结合律，从而可以支持并行计算

•countByValueAndWindow(windowLength, slideInterval, [numTasks]) 当应用到一个(K,V)键值对组成的DStream上，返回一个由(K,V)键值对组成的新的DStream。每个key的值都是它们在滑动窗口中出现的频率

6.6.2 DStream有状态转换操作

一些窗口转换操作的含义：

•reduceByKeyAndWindow(func, windowLength, slideInterval, [numTasks]) 应用到一个(K,V)键值对组成的DStream上时，会返回一个由(K,V)键值对组成的新的DStream。每一个key的值均由给定的reduce函数(func函数)进行聚合计算。注意：在默认情况下，这个算子利用了Spark默认的并发任务数去分组。可以通过numTasks参数的设置来指定不同的任务数

•reduceByKeyAndWindow(func, invFunc, windowLength, slideInterval, [numTasks]) 更加高效的reduceByKeyAndWindow，每个窗口的reduce值，是基于先前窗口的reduce值进行增量计算得到的；它会对进入滑动窗口的新数据进行reduce操作，并对离开窗口的老数据进行“逆向reduce”操作。但是，只能用于“可逆reduce函数”，即那些reduce函数都有一个对应的“逆向reduce函数”（以InvFunc参数传入）

6.6.2 DStream有状态转换操作

WindowedNetworkWordCount.py

```
#!/usr/bin/env python3
from __future__ import print_function
import sys
from pyspark import SparkContext
from pyspark.streaming import StreamingContext
if __name__ == "__main__":
    if len(sys.argv) != 3:
        print("Usage: WindowedNetworkWordCount.py <hostname> <port>", file=sys.stderr)
        exit(-1)
    sc = SparkContext(appName="PythonStreamingWindowedNetworkWordCount")
    ssc = StreamingContext(sc, 10)
    ssc.checkpoint("file:///usr/local/spark/mycode/streaming/socket/checkpoint")
    lines = ssc.socketTextStream(sys.argv[1], int(sys.argv[2]))
    counts = lines.flatMap(lambda line: line.split(" "))\
                .map(lambda word: (word, 1))\
                . reduceByKeyAndWindow(lambda x, y: x + y, lambda x, y: x - y, 30, 10)
    counts.pprint()
    ssc.start()
    ssc.awaitTermination()
```

6.6.2 DStream有状态转换操作

reduceByKeyAndWindow(lambda x, y: x + y, lambda x, y: x - y, 30, 10)

实现增量计算

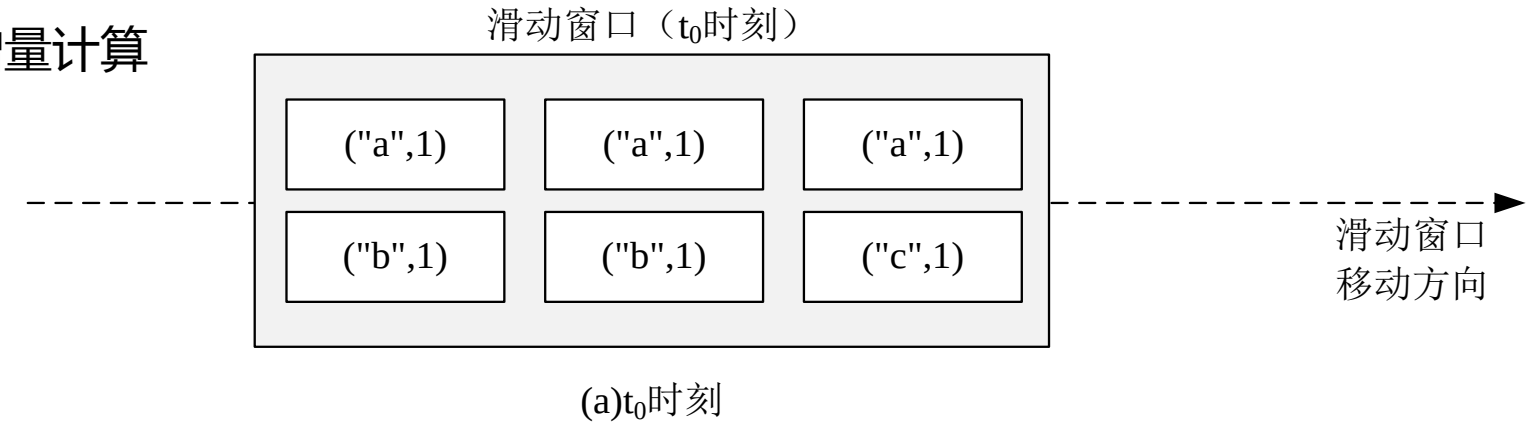


图 reduceByKeyAndWindow操作过程示意图

6.6.2 DStream有状态转换操作

为了测试程序的运行效果，首先新建一个终端（记作“数据源终端”），执行如下命令运行nc程序：

```
$ cd /usr/local/spark/mycode/streaming/socket/
$ nc -lk 9999
```

再新建一个终端（记作“流计算终端”），运行客户端程序WindowedNetworkWordCount.py，命令如下：

```
$ cd /usr/local/spark/mycode/streaming/socket/
$ /usr/local/spark/bin/spark-submit  \
> WindowedNetworkWordCount.py localhost 9999
```

在数据源终端内，用键盘连续敲入10个“hadoop”，每个hadoop单独占一行（即输入一个hadoop就回车），再用键盘连续敲入10个“spark”，每个spark单独占一行。这时，可以查看流计算终端内显示的词频动态统计结果，可以看到，随着时间的流逝，词频统计结果会发生动态变化。

6.6.2 DStream有状态转换操作

2. updateStateByKey操作

需要在跨批次之间维护状态时，就必须使用updateStateByKey操作

词频统计实例：

对于有状态转换操作而言，本批次的词频统计，会在之前批次的词频统计结果的基础上进行不断累加，所以，最终统计得到的词频，是所有批次的单词的总的词频统计结果

6.6.2 DStream有状态转换操作

NetworkWordCountStateful.py

```
#!/usr/bin/env python3
from __future__ import print_function
import sys
from pyspark import SparkContext
from pyspark.streaming import StreamingContext
if __name__ == "__main__":
    if len(sys.argv) != 3:
        print("Usage: NetworkWordCountStateful.py <hostname> <port>", file=sys.stderr)
        exit(-1)
    sc = SparkContext(appName="PythonStreamingStatefulNetworkWordCount")
    ssc = StreamingContext(sc, 1)
    ssc.checkpoint("file:///usr/local/spark/mycode/streaming/stateful/")
    # RDD with initial state (key, value) pairs
    initialStateRDD = sc.parallelize([(u'hello', 1), (u'world', 1)])

    def updateFunc(new_values, last_sum):
        return sum(new_values) + (last_sum or 0)

    lines = ssc.socketTextStream(sys.argv[1], int(sys.argv[2]))
    running_counts = lines.flatMap(lambda line: line.split(" "))\
                          .map(lambda word: (word, 1))\
                          .updateStateByKey(updateFunc, initialRDD=initialStateRDD)
    running_counts.pprint()
    ssc.start()
    ssc.awaitTermination()
```

6.6.2 DStream有状态转换操作

新建一个终端（记作“数据源终端”），执行如下命令启动nc程序：

```
$ nc  -lk  9999
```

新建一个Linux终端（记作“流计算终端”），执行如下命令提交运行程序：

```
$ cd /usr/local/spark/mycode/streaming/stateful
$ /usr/local/spark/bin/spark-submit \
> NetworkWordCountStateful.py localhost 9999
```

在数据源终端内手动输入一些单词并回车，再切换到流计算终端，可以看到已经输出了类似如下的词频统计信息：

```
-------------------------------------------
Time: 2018-12-30 20:53:02
-------------------------------------------
('hadoop', 1)
('world', 1)
('hello', 1)
('spark', 1)
-------------------------------------------
```

6.7 输出操作

在Spark应用中，外部系统经常需要使用到Spark DStream处理后的数据，因此，需要采用输出操作把DStream的数据输出到数据库或者文件系统中

6.7.1 把DStream输出到文本文件中
6.7.2 把DStream写入到MySQL数据库中

6.7.1 把DStream输出到文本文件中

请在NetworkWordCountStatefulText.py代码文件中输入以下内容：

```
#!/usr/bin/env python3

from __future__ import print_function
import sys
from pyspark import SparkContext
from pyspark.streaming import StreamingContext
```

6.7.1 把DStream输出到文本文件中

```
if __name__ == "__main__":
    if len(sys.argv) != 3:
        print("Usage: NetworkWordCountStateful.py <hostname> <port>", file=sys.stderr)
        exit(-1)
    sc = SparkContext(appName="PythonStreamingStatefulNetworkWordCount")
    ssc = StreamingContext(sc, 1)
    ssc.checkpoint("file:///usr/local/spark/mycode/streaming/stateful/")
    # RDD with initial state (key, value) pairs
    initialStateRDD = sc.parallelize([(u'hello', 1), (u'world', 1)])
    def updateFunc(new_values, last_sum):
        return sum(new_values) + (last_sum or 0)
    lines = ssc.socketTextStream(sys.argv[1], int(sys.argv[2]))
    running_counts = lines.flatMap(lambda line: line.split(" "))\
                          .map(lambda word: (word, 1))\
                          .updateStateByKey(updateFunc, initialRDD=initialStateRDD)
    running_counts.saveAsTextFiles("file:///usr/local/spark/mycode/streaming/stateful/outpu
    running_counts.pprint()
    ssc.start()
    ssc.awaitTermination()
```

6.7.2 把DStream写入到MySQL数据库中

启动MySQL数据库，并完成数据库和表的创建：

```
$ service mysql start
$ mysql -u root -p
$ #屏幕会提示你输入密码
```

在此前已经创建好的“spark”数据库中创建一个名称为“wordcount”的表：

```
mysql> use spark
mysql> create table wordcount (word char(20), count int(4));
```

6.7.2 把DStream写入到MySQL数据库中

由于需要让Python连接数据库MySQL，所以，需要首先安装Python连接MySQL的模块PyMySQL，请在Linux终端中执行如下命令：

```
$ sudo apt-get update
$ sudo apt-get install python3-pip
$ pip3 -V
$ sudo pip3 install PyMySQL
```

6.7.2 把DStream写入到MySQL数据库中

NetworkWordCountStatefulDB.py

```
#!/usr/bin/env python3

from __future__ import print_function
import sys
import pymysql
from pyspark import SparkContext
from pyspark.streaming import StreamingContext

if __name__ == "__main__":
    if len(sys.argv) != 3:
        print("Usage: NetworkWordCountStateful <hostname> <port>", file=sys.stderr)
        exit(-1)
    sc = SparkContext(appName="PythonStreamingStatefulNetworkWordCount")
    ssc = StreamingContext(sc, 1)
    ssc.checkpoint("file:///usr/local/spark/mycode/streaming/stateful")
    # RDD with initial state (key, value) pairs
    initialStateRDD = sc.parallelize([(u'hello', 1), (u'world', 1)])
```

6.7.2 把DStream写入到MySQL数据库中

```
def updateFunc(new_values, last_sum):
    return sum(new_values) + (last_sum or 0)

lines = ssc.socketTextStream(sys.argv[1], int(sys.argv[2]))
running_counts = lines.flatMap(lambda line: line.split(" "))\
                      .map(lambda word: (word, 1))\
                      .updateStateByKey(updateFunc, initialRDD=initialStateRDD)
running_counts.pprint()
```

6.7.2 把DStream写入到MySQL数据库中

```
def dbfunc(records):
        db = pymysql.connect("localhost","root","123456","spark")
        cursor = db.cursor()
        def doinsert(p):
            sql = "insert into wordcount(word,count) values ('%s', '%s')" % (str(p[0]),
str(p[1]))
            try:
                cursor.execute(sql)
                db.commit()
            except:
                db.rollback()
        for item in records:
            doinsert(item)
```

6.7.2 把DStream写入到MySQL数据库中

```
def func(rdd):
        repartitionedRDD = rdd.repartition(3)
        repartitionedRDD.foreachPartition(dbfunc)

    running_counts.foreachRDD(func)
    ssc.start()
    ssc.awaitTermination()
```

附录A：主讲教师林子雨简介

主讲教师：林子雨

单位：厦门大学计算机科学系
E-mail: ziyulin@xmu.edu.cn
个人网页：http://dblab.xmu.edu.cn/post/linziyu
数据库实验室网站：http://dblab.xmu.edu.cn

扫一扫访问个人主页

林子雨，男，1978年出生，博士（毕业于北京大学），现为厦门大学计算机科学系助理教授（讲师），曾任厦门大学信息科学与技术学院院长助理、晋江市发展和改革局副局长。中国计算机学会数据库专业委员会委员，中国计算机学会信息系统专业委员会委员。国内高校首个“数字教师”提出者和建设者，厦门大学数据库实验室负责人，厦门大学云计算与大数据研究中心主要建设者和骨干成员，2013年度和2017年度厦门大学教学类奖教金获得者，荣获2017年福建省精品在线开放课程、2018年厦门大学高等教育成果特等奖、2018年福建省高等教育教学成果二等奖、2018年国家精品在线开放课程。主要研究方向为数据库、数据仓库、数据挖掘、大数据、云计算和物联网，并以第一作者身份在《软件学报》《计算机学报》和《计算机研究与发展》等国家重点期刊以及国际学术会议上发表多篇学术论文。作为项目负责人主持的科研项目包括1项国家自然科学青年基金项目(No.61303004)、1项福建省自然科学青年基金项目(No.2013J05099)和1项中央高校基本科研业务费项目(No.2011121049)，主持的教改课题包括1项2016年福建省教改课题和1项2016年教育部产学协作育人项目，同时，作为课题负责人完成了国家发改委城市信息化重大课题、国家物联网重大应用示范工程区域试点泉州市工作方案、2015泉州市互联网经济调研等课题。中国高校首个“数字教师”提出者和建设者，2009年至今，“数字教师”大平台累计向网络免费发布超过500万字高价值的研究和教学资料，累计网络访问量超过500万次。打造了中国高校大数据教学知名品牌，编著出版了中国高校第一本系统介绍大数据知识的专业教材《大数据技术原理与应用》，并成为京东、当当网等网店畅销书籍；建设了国内高校首个大数据课程公共服务平台，为教师教学和学生学习大数据课程提供全方位、一站式服务，年访问量超过200万次。

附录B：大数据学习路线图

大数据学习路线图访问地址：http://dblab.xmu.edu.cn/post/10164/

附录C：《大数据技术原理与应用》教材

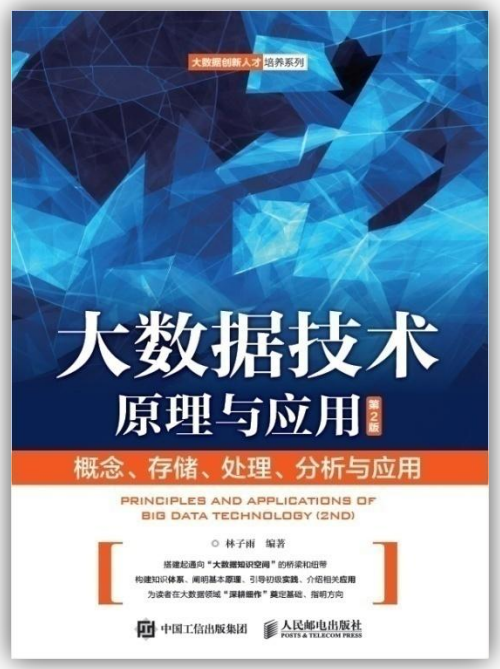
《大数据技术原理与应用——概念、存储、处理、分析与应用（第2版）》，由厦门大学计算机科学系林子雨博士编著，是国内高校第一本系统介绍大数据知识的专业教材。人民邮电出版社 ISBN:978-7-115-44330-4 定价：49.80元

全书共有15章，系统地论述了大数据的基本概念、大数据处理架构Hadoop、分布式文件系统HDFS、分布式数据 库HBase、NoSQL数据库、云数据库、分布式并行编程模型MapReduce、Spark、流计算、图计算、数据可视化以及大数据在互联网、生物医学和物流等各个领域的应用。在Hadoop、HDFS、HBase和MapReduce等重要章节，安排了入门级的实践操作，让读者更好地学习和掌握大数据关键技术。

本书可以作为高等院校计算机专业、信息管理等相关专业的大数据课程教材，也可供相关技术人员参考、学习、培训之用。

欢迎访问《大数据技术原理与应用——概念、存储、处理、分析与应用》教材官方网站：
http://dblab.xmu.edu.cn/post/bigdata

扫一扫访问教材官网



附录D：《大数据基础编程、实验和案例教程》

本书是与《大数据技术原理与应用（第2版）》教材配套的唯一指定实验指导书

大数据教材

1+1黄金组合

厦门大学林子雨编著

配套实验指导书

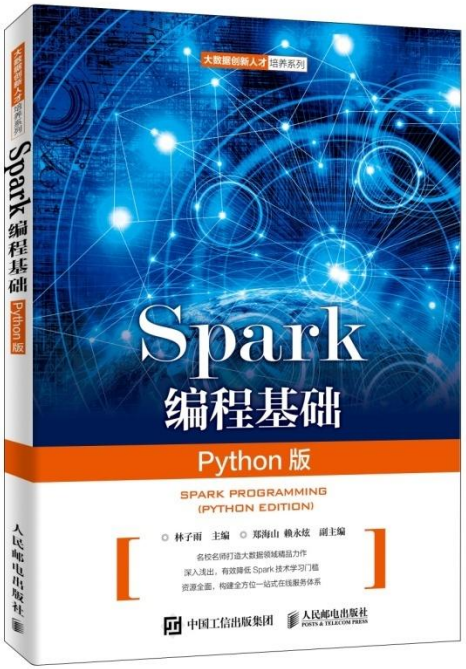
- 步步引导，循序渐进，详尽的安装指南为顺利搭建大数据实验环境铺平道路
- 深入浅出，去粗取精，丰富的代码实例帮助快速掌握大数据基础编程方法
- 精心设计，巧妙融合，五套大数据实验题目促进理论与编程知识的消化和吸收
- 结合理论，联系实际，大数据课程综合实验案例精彩呈现大数据分析全流程

清华大学出版社 ISBN:978-7-302-47209-4 定价：59元

附录E：《Spark编程基础（Python版）》

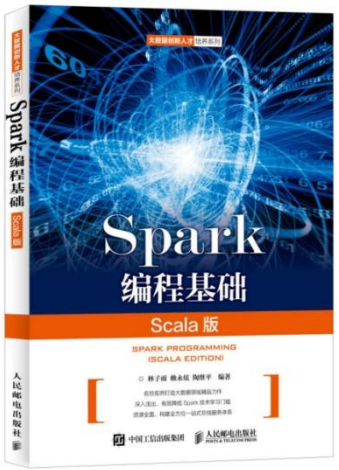
林子雨，郑海山，赖永炫 编著《Spark编程基础（Python版）》

教材官网：http://dblab.xmu.edu.cn/post/spark-python/
ISBN:978-7-115-52439-3 人民邮电出版社



本书以Python作为开发Spark应用程序的编程语言，系统介绍了Spark编程的基础知识。全书共8章，内容包括大数据技术概述、Spark的设计与运行原理、Spark环境搭建和使用方法、RDD编程、Spark SQL、Spark Streaming、Structured Streaming、Spark MLlib等。本书每个章节都安排了入门级的编程实践操作，以便读者更好地学习和掌握Spark编程方法。本书官网免费提供了全套的在线教学资源，包括讲义PPT、习题、源代码、软件、数据集、授课视频、上机实验指南等。

附录F：《Spark编程基础（Scala版）》



《Spark编程基础（Scala版）》

厦门大学 林子雨，赖永炫，陶继平 编著

披荆斩棘，在大数据丛林中开辟学习捷径
填沟削坎，为快速学习Spark技术铺平道路
深入浅出，有效降低Spark技术学习门槛
资源全面，构建全方位一站式在线服务体系

人民邮电出版社出版发行，ISBN:978-7-115-48816-9
教材官网：http://dblab.xmu.edu.cn/post/spark/

本书以Scala作为开发Spark应用程序的编程语言，系统介绍了Spark编程的基础知识。全书共8章，内容包括大数据技术概述、Scala语言基础、Spark的设计与运行原理、Spark环境搭建和使用方法、RDD编程、Spark SQL、Spark Streaming、Spark MLlib等。本书每个章节都安排了入门级的编程实践操作，以便读者更好地学习和掌握Spark编程方法。本书官网免费提供了全套的在线教学资源，包括讲义PPT、习题、源代码、软件、数据集、授课视频、上机实验指南等。

附录G：高校大数据课程公共服务平台

高校大数据课程

公 共 服 务 平 台

http://dblab.xmu.edu.cn/post/bigdata-teaching-platform/

扫一扫访问平台主页

扫一扫观看3分钟FLASH动画宣传片


Thank You!