关于tsx-随手笔记

作者：杨鹏飞

时间：2015-07-03

目 录

[1 序言 1](#_Toc423528822)

[2 归并模块 2](#_Toc423528823)

[3 更新时间 3](#_Toc423528824)

# 序言

笔记。

# tsx-streaming

## 简介

最近在做利用Spark streaming和Kafka进行数据分析的研究， 整理一些相应的开发文档， 做了一些代码实践。 本文特意将这些资料记录下来。

本文最后列出了一些参考的文档，实际调研中参考了很多的资料，并没有完全将它们记录下来， 只列出了主要的一些参考资料。 当前的版本：

Spark: 1.2.0

Kafka: 0.8.1.1

Spark Streaming属于Spark的核心api，它支持高吞吐量、支持容错的实时流数据处理。 有以下特点:

易于使用 提供了和批处理一致的高级操作API，可以进行map, reduce, join, window。

容错 Spark Streaming可以恢复你计算的状态， 包括lost work和operator state (比如 sliding windows)。 支持worker节点和driver 节点恢复。

## Spark集成优点

可以结合批处理流和交互式查询。

可以重用批处理的代码。

可以直接使用内置的机器学习算法、图算法包来处理数据。

它可以接受来自文件系统, Akka actors, rsKafka, Flume, Twitter, ZeroMQ和TCP Socket的数据源或者你自己定义的输入源。



它的工作流程像下面的图所示一样，接受到实时数据后，给数据分批次，然后传给Spark Engine处理最后生成该批次的结果流。



Spark Streaming提供了一个高级的抽象模型，叫做discretized stream或者叫做DStream,它代表了一个持续的数据流。DStream既可以从Kafka、Flume 和 Kinesis中产生, 或者在其它DStream上应用高级操作得到。 内部实现上一个DStream代表一个RDD序列。

# 更新时间

2015-06-30 周二，随手笔记，入职记录；

2015-07-01 周二，随手笔记，常用名称记录；