Storm Applied: Strategies For Real-time Event Processing
Synopsis

Summary Storm Applied is a practical guide to using Apache Storm for the real-world tasks associated with processing and analyzing real-time data streams. This immediately useful book starts by building a solid foundation of Storm essentials so that you learn how to think about designing Storm solutions the right way from day one. But it quickly dives into real-world case studies that will bring the novice up to speed with productionizing Storm. Purchase of the print book includes a free eBook in PDF, Kindle, and ePub formats from Manning Publications. Summary Storm Applied is a practical guide to using Apache Storm for the real-world tasks associated with processing and analyzing real-time data streams. This immediately useful book starts by building a solid foundation of Storm essentials so that you learn how to think about designing Storm solutions the right way from day one. But it quickly dives into real-world case studies that will bring the novice up to speed with productionizing Storm. About the Technology It’s hard to make sense out of data when it’s coming at you fast. Like Hadoop, Storm processes large amounts of data but it does it reliably and in real time, guaranteeing that every message will be processed. Storm allows you to scale with your data as it grows, making it an excellent platform to solve your big data problems. About the Book Storm Applied is an example-driven guide to processing and analyzing real-time data streams. This immediately useful book starts by teaching you how to design Storm solutions the right way. Then, it quickly dives into real-world case studies that show you how to scale a high-throughput stream processor, ensure smooth operation within a production cluster, and more. Along the way, you’ll learn to use Trident for stateful stream processing, along with other tools from the Storm ecosystem. This book moves through the basics quickly. While prior experience with Storm is not assumed, some experience with big data and real-time systems is helpful. What’s InsideMapping real problems to Storm componentsPerformance tuning and scalingPractical troubleshooting and debuggingExactly-once processing with TridentAbout the AuthorsSean Allen, Matthew Jankowski, and Peter Pathirana lead the development team for a high-volume, search-intensive commercial web application at TheLadders.

Table of Contents
Introducing Storm
Core Storm concepts
Topology design
Creating robust topologies
Moving from local to remote topologies
Tuning in Storm
Resource contention
Storm internals
Trident

Book Information

Paperback: 280 pages
Publisher: Manning Publications; 1 edition (April 12, 2015)
Language: English
First some context as to why I wanted to read this book. I have extensive knowledge in the messaging, BPM and Java area having worked on a number of software products in this area. I have been interested in understanding how to use Hadoop and this in turn led me to Storm. I am a beginner on Storm and was looking for a book that explained the fundamentals as the concepts seemed to be similar to those I have used in broker software. Onto the book. The things I thought were good: 1. It is extremely easy to read, very well written and with a number of very good examples on differing Storm topologies. The authors do an excellent job of stepping through the examples in a lot of detail which helps the reader understand the concepts behind the product. Throughout the book it is obvious that the authors have used Storm extensively. 2. The flow through the chapters is good, starting with core Storm concepts in Chapter 2. Chapter 3 is also good for beginners as it explains some things to take into consideration when building Storm topologies. There are also some good chapters following on how build and deploy a Storm cluster in production and discusses issues that can arise, although some of these e.g. JVM and O/S tuning are the sort of things that companies would already have a handle on, in my opinion. Nevertheless, some useful information applied to Storm processing. The things I felt were missing or lacking: 1. Some real-world performance figures. This is all about "real-time" processing and consequently I would have expected some examples of what sort of throughput a user could expect for certain hardware configurations. The authors throw in a statement on page 9, "In fact it does not get any speedier than this". I realize this would be a whole book by itself but some numbers would be useful. The authors say they have deployed it in a production, so I would have expected to see some numbers. 2. The authors state this book is for beginners so I would have expected to see a section at the beginning on how to actually install Storm in local mode, so that the reader could test out the
code examples. Even a pointer would have been helpful but nothing appears like this until page 109.3. Having spent a lot of time with Enterprise-ready guaranteed message processing (once and only once), I know how complicated this really is to achieve. Consequently I feel that chapter 4 spent a long time discussing this and the conclusion I got reading the chapter is that it is not really guaranteed for all messages in all environments (which is what I expected). From the start of the chapter, it seemed to imply that it was possible with Storm but really there is more to it especially when interacting with external services. 4. No mention was made of the overhead in providing the various levels of "guaranteed message processing". Usually there is some loss in throughput as it becomes more guaranteed. 5. The tuning section was good but there was no real explanation of how the authors arrived at figures of 32, 32 and 8 for bolt tasks in one of the examples. The bottleneck was identified and then these numbers seemed to be pulled from nowhere. At least, I did not see a scientific explanation. 6. I did not find the chapter on Trident at all useful for me personally and I am not sure it added a lot to the book.

I picked the book up after reading Marz and Warren's "Big Data", seeing this one as the logical next step (Marz is also the creator of Storm). I once previously tried getting acquainted with Storm through the online documentation, but found the terminology to be rather complicated: there's topologies, spouts, bolts, tuple trees, executors and tasks, to name a few. This is where the book really shines for me. The first four chapters delved deeply (yet gently) into the Storm terminology. They are extremely well written, and the book flows almost like a novel. In addition, the examples give (counting git commits, grouping social network check-ins, and a credit card authorization system) are chosen perfectly to illustrate the points, and are something every reader (even one from a non-STEM background, such as myself) will easily relate to. The rest of the book is dedicated to detailed explanations of setting up Storm, tuning it for performance, internals of the platform, as well as a very good cookbook-style chapter on dealing with different performance issues the authors have seen and dealt with. An additional example here, of a flash-sale based e-commerce website, ties in very well. Overall, the book is a excellent, and I'm happy that I have it. The style of writing and exposition is something all technical book authors should definitely strive for.

This is probably the best of (not so many) available books on Storm. It goes in much detail explaining how Storm works and its structure. A little bit underestimates the reader (too many simple details), but maybe it is targeted at people who are not Java developers. Unfortunately, Trident is explained very poorly, but same is true for all other books on Storm. Looks like nobody of those
writers really understands Trident.

Download to continue reading...


Dmca