

Seminar Databasesystems

By
Samuel Kurath

University of Applied Science, HSR

Autumn 2016

©2015 – ALL RIGHTS RESERVED.

Advisor:
Prof. Keller Stefan

Seminar Databasesystems

ABSTRACT

This paper is splitted in three parts, it begins with an overview of streams and their difficulties. Followed by a part about Apache Storm a distributed stream processing framework. And finally a concrete implementation based on a given problem with Storm. The goal of the implementation is to do some queries and analysis on the minutely updated Augmented Diffs of OpenStreetMap.

Contents

0	INTRODUCTION	2
1	STREAM PROCESSING	3
1.1	Initial example	4
1.2	Eight Rules For Stream Procesing	7
2	APACHE STORM	10
2.1	Topology	11
2.2	Spout	13
2.3	Bolt	14
3	IMPLEMENTATION	16
3.1	Test Setup	17
3.2	Queries	17
3.3	Topology of the queries	18
3.4	Benchmark	19
4	CONCLUSION	21
	REFERENCES	22



Introduction

During the master of science in engineering the students have to absolve two seminars. The goal of these is to elaborate a theme on your own, discuss the result in group and write a paper about the topic.

The Databasesystems Seminar does a focus on streams and their processing. Stream processing is a strong growing subject in reference to the huge amount of data we are exposed and produce nowadays.

A big force in generating this data is the rapidly increasing amount of Internet of Things *IoT* sensors, the willingness of the people to populate a lot of personal information on social media platforms and also the expanding interest in data collection of companies.

With this amount of data new problems in collecting, processing, storing etc. appear and thus new solutions and technical tools to solve them appear too.

*The man who is swimming against the stream
knows the strength of it.*

Woodrow Wilson

1

Stream Processing

Streams are older than computers so it is not a big surprise that streams and the processing of them isn't an absolutely new topic in the computer science world. Historically there are techniques like **logging** or in the domain driven development area there is **event sourcing**, which are very similar to streaming processing. But they have not always been such an important deal like they are today with the unbelievable amount of data. Antecedent to handle this streams was an event driven way and do analytics after storing the data.

INITIAL EXAMPLE

In this section I'd like to give you a small example of how streams and stream processing could be more and more relevant. And I would like to do this by the example of a factory producing delicious frosted yoghurt.

FROZEN YOGURT FACTORY

Let me explain the *old* event driven way with a small example.

Imagine a factory which produces frozen yogurt in different flavors. They weigh and register every cup of yogurt at the end of the assembly line.



Figure 1.1: frozen yogurt assembly line

COMMON ARCHITECTURE

The factory has sensors which weight the cups and this weights are sent to a server. The server handles the request and stores the frozen yogurt with his weight in the database. for analytic tasks there is a web application. If you are now interested in the total amount of produced cups. You can simply open the browser go to the analytic page. This starts a request to the server and the server will call the database with a query like "select count(*) from frozen_yogurt". After the query is executed you will get the result from the server and have the aggregated number on your screen.

(This is more or less a default example of a Three-tier architecture)

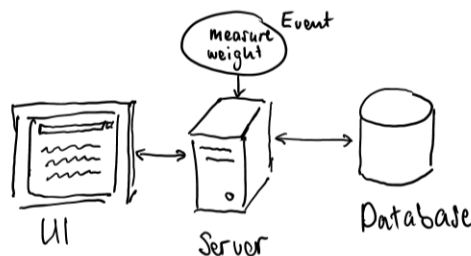


Figure 1.2: Three-tier architecture

Now the owner of the factory does a very good business and they are able to expand the production. And with this expansion they also add a lot more of sensors to the assembly line like, a temperature sensor, optical recognition to check if the cups are always full and a lot more.

The requirements of the system are also updated the owner want to have statistic about the production all the time and want's immediately notifications if for example the temperature is to high.

These new requirements leads to new challenges in the architecture of the system and are hard to implement with the current state and the enornus data produced by the all sensors.

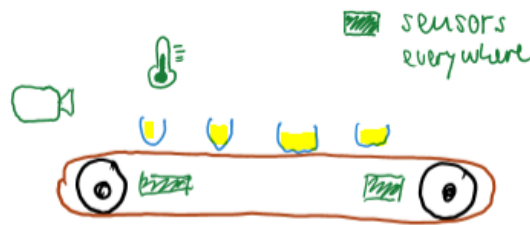


Figure 1.3: sensors everywhere

STREAMING ARCHITECTURE

A good way to handle this new requirements is to continuous aggregate and filter the stream of data before it is stored in the database. For this purpose there has grown up a lot of new techniques and frameworks during the last few years.

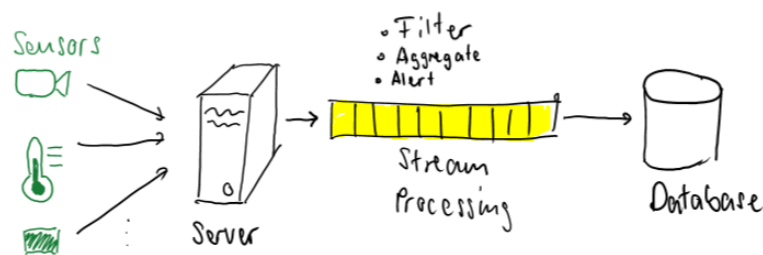


Figure 1.4: streaming architecture

CONCLUSION

The new streaming architecture of the frozen yogurt factory has got a lot of advantages but also has to handle some big difficulties. Now it is possible to get very fast access to production statistics, deal with the huge amount of data and throw alerts in real time. It's also helpfull in the developers perspective, because you can add several consumer to the stream. Normaly this stream is immutable if you use a technologie like kafka, which has the advantage that there aren't concurrency problems anymore. Thus the requirements in such real time stream processing systems are ambitious and we will face them in the next section.

EIGHT RULES FOR STREAM PROCESING

In the paper "The 8 Requirements of Real-Time Stream Processing" Michael Stonebraker, Uur etintemel and Stan Zdonik bring their thoughts about real-time stream processing together and write it down. It faces the fundamental ideas of stream processing and sums them up in eight rules which every stream processing engine should require. This allows us to measure and compare different technologies. Thus leads me to take a closer look at these rules and add my reflection to them.

RULE 1: KEEP THE DATA MOVING

The first requirement for a real-time stream processing system is to process messages in-stream, without any requirement to store them to perform any operation or sequence of operations. Ideally the system should also use an active (i.e., non-polling) processing model.

This is also for me a very important point. The stream processing should not slow down the "in-stream". How can this target be reached? The paper says be aware of expensive storage operation. But I could imagine that this question depends most on the architecture of the processing engine. So handle the streams as immutable and decouple the processing from the stream sink.

RULE 2: QUERY USING SQL ON STREAMS (STREAMSQL)

The second requirement is to support a high-level StreamSQL language with built-in extensible stream- oriented primitives and operators.

Use something like "StreamSQL" is not a must have requirement in my eyes. I would even go further and say that a higher-level query language could not always fit the fast changing streams. So a flexible and easy to expand query language would be a better choice for me. Even in spite of the heavy maintenance and intelligibility.

RULE 3: HANDLE STREAM IMPERFECTIONS (DELAYED, MISSING AND OUT-OF-ORDER DATA)

The third requirement is to have built-in mechanisms to provide resiliency against stream imperfections, including missing and out-of-order data, which are commonly

present in real-world data streams.

If the data of the stream does not appear like it used to be the stream processing engine should not crash and handle the problem. For me an other very fundamental requirement in the point of view that such streams often come from distributed systems. Which leads to a lot of hard to control error sources.

RULE 4: GENERATE PREDICTABLE OUTCOMES

The fourth requirement is that a stream processing engine must guarantee predictable and repeatable outcomes.

The same input should always result in the same output. Easy said and easy to understand rule but in processing huge amount of data this could be in some cases very difficult. But you still want to be able to process data and not very entry in the stream is absolutely important. Thus it could be a good strategy to more or less randomly more or less randomly throw some messages away. Nevertheless the decisions you do and the processing on it should be reproducible.

RULE 5: INTEGRATE STORED AND STREAMING DATA

The fifth requirement is to have the capability to efficiently store, access, and modify state information, and combine it with live streaming data. For seamless integration, the system should use a uniform language when dealing with either type of data.

I agree with this rule and past data and also current data should possible haven an influent on the results of the processing. So it is a good idea to use a uniform language. But I also think that the streaming engine needs to have the ability and flexibility to deal with different and unexpected data.

RULE 6: GUARANTEE DATA SAFETY AND AVAILABILITY

The sixth requirement is to ensure that the applications are up and available, and the integrity of the data maintained at all times, despite failures.

Of course streaming engines need to be high-availability. But this is a very hard deal in the perspective of high performance, because high-availability also leads to

some inevitable overheads. Thus i would say that it should be possible to handle guarantee integrity of mission-critical information and also support less critical information to be processed in a super performant but less integrity way.

RULE 7: PARTITION AND SCALE APPLICATIONS AUTOMATICALLY

The seventh requirement is to have the capability to distribute processing across multiple processors and machines to achieve incremental scalability. Ideally, the distribution should be automatic and transparent.

In the case of huge amount of data the ability to scale vertically is an absolute must have feature of every stream processing engine. For me this is one of the most important rules.

RULE 8: PROCESS AND RESPOND INSTANTANEOUSLY

The eighth requirement is that a stream processing system must have a highly-optimized, minimal-overhead execution engine to deliver real-time response for high-volume applications.

Like at the rule 7 sad, this is a hard tradeoff between availability and integrity. Thus a stream processing engine has to deal with this two requirements as good as possible and find a best fitting solution. Because the possible throughput can also have a decisive impact on the choice of a technology.

*There are some things you learn best in calm,
and some in storm.*

Willa Cather

2

Apache Storm

Apache Storm is a reliable, distributed and fault-tolerant system for stream processing. The beginnings of the project were at Backtype (later bought by Twitter) and created by Nathan Marz. He open sourced Storm on September the 19th in 2011. The project rapidly got a big development community and on September the 18, 2013 Nathan moved Storm to Apache Incubator.

Storm works with different types of components which are responsible for clear defined task. These components are bundled and managed in a so called **Topology**. The entrypoint and the stream input is handled by a **Spout**, the spout passes to **Bolts**. Bolts are responsible for the main data processing and persist the data. They can be chained or parallelised in a way that fits best for your current problem.



Figure 2.1: Storm

TOPOLOGY

In general Storm passes tuples between the different components. To organize this tuples there are topologies.

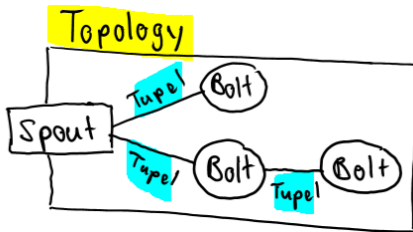


Figure 2.2: Topology example

GROUPING

The Topology defines the grouping, this means how streams are consumed by the bolts and how they consume them. There are four kinds of grouping *Shuffle Grouping*, *Fields Grouping*, *All Grouping* and *Custom Grouping*.

SHUFFLE GROUPING

Shuffle Grouping takes a single entry from the source and sends each tuple to a randomly chosen bolt, which is listening to this kind of tuples. It also guarantees that each consumer gets the same number of tuples.

FIELDS GROUPING

With Fields Grouping it is possible to send tuples to specified bolts based on the fields of the tuple. It takes care that the same combination of fields is always sent to the same bolt.

ALL GROUPING

All Grouping sends every tuple to all the bolts. This is helpful for tasks like signals or use different filters for alerting systems.

CUSTOM GROUPING

It is possible to build your own grouping based on the stream. This is very helpful if you have to make fine granular decisions about which bolt has to handle which tuple.

SPOUT

The entry point of each topology are the so called spouts. Their main task is to procure the data from the input source. After the access of the source the spout emits a list of defined fields, which are consumed by the bolts. The spouts are also responsible for reliability, thus you have to take care to implement them in a fault-tolerant way. This means that a spout must have the ability to deal with unprocessable and unpredictable messages from the stream.

To handle the reliability at the spout you can add an ID to every message. If a message is correctly processed by all the target bolts the *ack* method of the spout is called. But if there were troubles or the timeout is reached the *fail* method will be triggered.

EXAMPLE

As an example you can imagine a spout which consumes data from a message queue and emits the data to the interested bolts. All of this data stream is defined in storm topology.

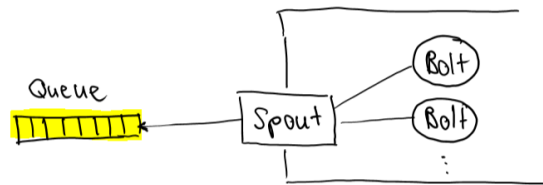


Figure 2.3: Spout example

CODE

And in the code this could look like the following.

```
TopologyBuilder builder = new TopologyBuilder();
MessageQueueSpout spout = new MessageQueueSpout();
PrinterBolt bolt = new PrinterBolt();

builder.setSpout("spout", spout);
builder.setBolt("bolt", bolt).shuffleGrouping("spout");
```


BOLT

Bolts are the heartpieces, the engine of storm, they are responsible for the hard work of processing the data. A bolt takes tuples as input, does some operations on it and finally produces tuples as output. Normally a bolt implements the *IRichBolt* interface, which has the following methods.

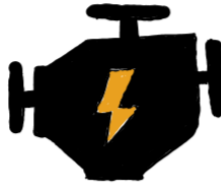


Figure 2.4: Bolt engine

```
PUBLIC VOID PREPARE(MAP MAP, TOPOLOGYCONTEXT TOPOLOGYCONTEXT)
```

This method is always called before the bolts starts to process tuples (before `execute(...)`). A possible task could be to reset a counter.

```
PUBLIC VOID EXECUTE(TUPLE TUPLE, BASICOUTPUTCOLLECTOR BASICOUTPUTCOLLECTOR)
```

The execute methode processes a singel tuple of input, so it is the core function of every bold.

```
PUBLIC VOID DECLAREOUTPUTFIELDS(OUTPUTFIELDSDECLARER OUTPUTFIELDSDECLARER)
```

The prepare method is responsible for defining the output schema of the bolt. If bolts are concatenated the output is a tuple ,but if the current bolt is may be the last bolt in the chain, and does a task like ,print to the standard output, you don't have to set an output type.

```
PUBLIC VOID CLEANUP()
```

Is called when a bolt is going to shutdown. So you should do some tasks like a destructor does. For example disconnect an active database connection.

A good idea is about ten percent and implementation and hard work, and luck is 90 percent.

Guy Kawasaki

3

Implementation

This chapter is about the implementation with Apache Storm to solve the defined queries and to do the benchmarks. The idea of the test scenario is to use "OSM Augmented Diffs" as a stream. This diffs extend the ordinary minutely diffs of OSM with more information. The result contains all the nodes, ways and relations changed during the time periode.

TEST SETUP

For the implementation there is a given test setup. The Open Street Map Augmented diffs are minutely collected by a script executed as a cron job. The script acts as a producer for Kafka and the data is now accessible from these message queue. Furthermore the data is ready to be consumed by the stream processing engine and in this case ready for Apache Storm.

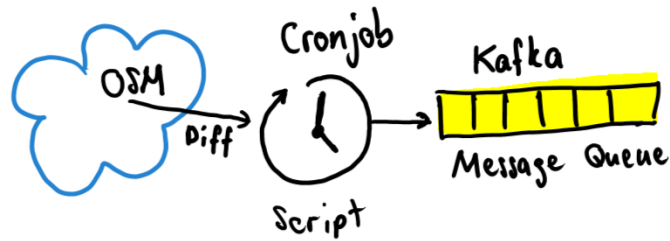


Figure 3.1: Test setup

QUERIES

The following list is about the queries which has to be done on the Augmented Diff stream.

- A) Leader board of top 10 OSM active users
- B) Leader board of top 10 OSM objects added
- C) Node objects with suspicious keys and values
- D) Way objects with only user tag "area=yes" without other user tags

TOPOLOGY OF THE QUERIES

With the queries we have to solve in mind. I have to design a fitting architecture for Apache Storm. So the first step is to get the messages from kafka, then convert the message string to JSON, do the queries on the JSON stream and finally make the result available.

This leads me to the following Apache Storm topology:

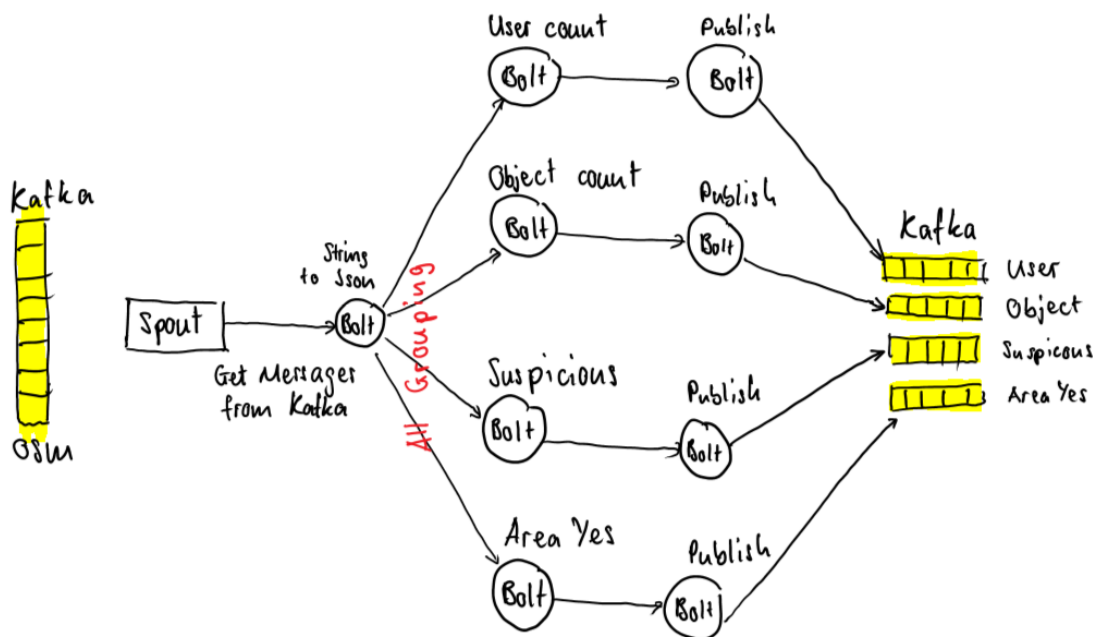


Figure 3.2: Topic for queries

Let's have a closer look to some implementation details of the image above. To read the messages from the kafka osm topic I used the *KafkaSpout* class from `org.apache.storm.kafka`. This class handles the communication with kafka and acts like a consumer, furthermore it makes the string messages available for the bolts. The consumed messages are in text format and needs to be casted to JSON for this purpose I implemented a *StringToJsonBolt* class. After this conversion I used an all grouping approach, which emits the JSON data to all connected bolts. Each following bolt is responsible for defined query, thus leads to the classes *UserCountBolt*, *ObjectCountbolt*, *SuspiciousBolt*

and *AreaYesBolt*. When the bolts have done their queries on the JSON data there is always a connected *KafkaBolt* class from `org.apache.storm.kafka`, which publishes the results to the fitting kafka topic.

BENCHMARK

Benchmarking is a very difficult topic and strongly depends on various parameters like the underlying hardware. Thus we decide to make this as hardware independent as possible. The idea is to produce a lot of small kafka message, like IoT does, and to count them. Then repeat this process and relate the number of processed messages with the time spent.

For the message generation we had a small python script called "benchmark.py" which produces as much messages as you want and stores them in to kafka. The messages have got the following structure:

packetId	TopicName	qos	retainFlag	payload	dupFlag
0	weather	5	False	83	True
1	computer	2	True	82	False

Table 3.1: MQTT Messages

To realize the benchmark with apache storm I tried two different topics. The first with only one Bolt, which is the non parallelism way. This leads to the blue curve on the diagramm.

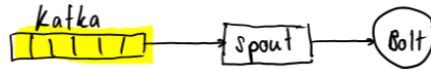


Figure 3.3: Benchmark 1 Bolt

For the second approach I use four Bolts for parallelism. The decision for four Bolts depends on the four cores of the notebook which I used for the benchmarks. The result is the red curve on the diagramm.

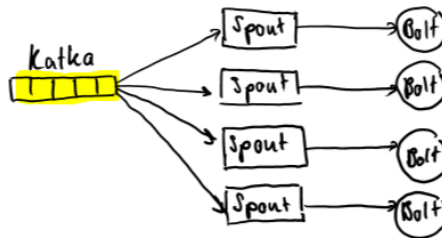


Figure 3.4: Benchmark 4 Bolts

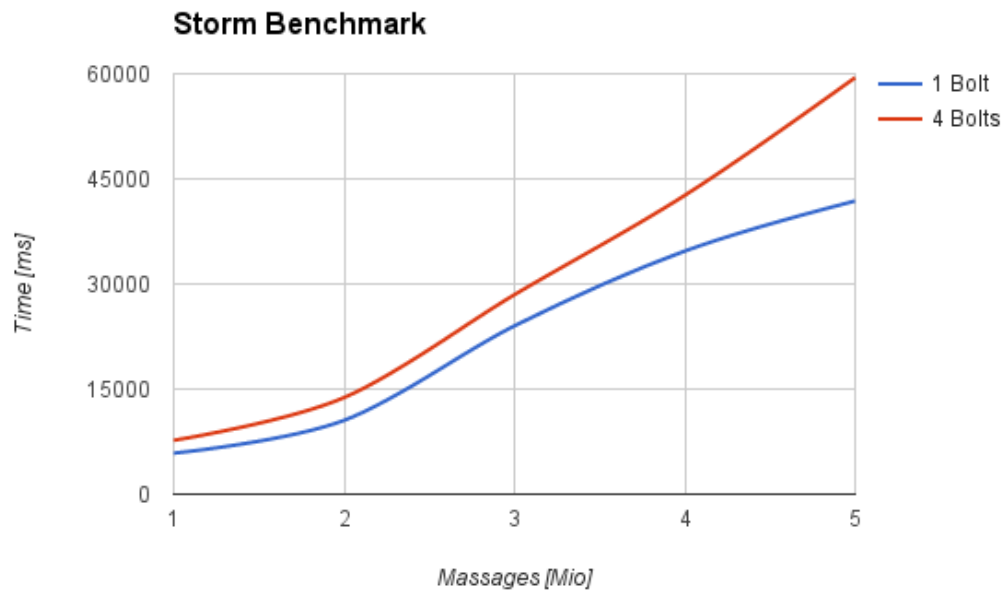


Figure 3.5: Benchmark Diagramm

CONCLUSION

To take a good conclusion out of this benchmark is a hard task. On the one hand side it's no problem for storm to process millions of messages. But on the other hand side I couldn't perform better with more messages. It was more or less a linear behavior. A bit unsuspected was that the parallelism topic was slower than to work with only one Bolt. Maybe this behavior would change with bigger messages. I think the bottleneck in the construction was the communication with kafka. And so the throughput was limited by the networking.

4

Conclusion

References

- [1] Akidau, T., Balikov, A., Bekiroğlu, K., Chernyak, S., Haberman, J., Lax, R., McVeety, S., Mills, D., Nordstrom, P., & Whittle, S. (2013). Millwheel: fault-tolerant stream processing at internet scale. *Proceedings of the VLDB Endowment*, 6(11), 1033–1044.
- [2] Neumeyer, L., Robbins, B., Nair, A., & Kesari, A. (2010). S4: Distributed stream computing platform. In *2010 IEEE International Conference on Data Mining Workshops* (pp. 170–177).: IEEE.
- [3] Stonebraker, M., Çetintemel, U., & Zdonik, S. (2005). The 8 requirements of real-time stream processing. *ACM SIGMOD Record*, 34(4), 42–47.