



Quanticate
A Passion For Excellence

Global Solutions from the World's Largest Data-Focused CRO

Scalably crashing JVMs with Tika

Or why turning binary data to content is hard!

JVM Crashing, Tika & Scale

- How Apache Tika has crashed my JVM
- How it has crashed lots of other people's JVMs!
- ... and ways we're trying to avoid it in future
- How to tell if changes are making your data pipeline "roughly" better or worse
- Which errors to look at next – our approaches
- And what is this Apache Tika stuff anyway?

↻ Simon Willison Retweeted



Gary Bernhardt @garybernhardt · 1d

An update to twitter.com/garybernhardt/...! You can now have a x1.16xlarge EC2 instance (0.976 TiB of RAM) for \$3.836/hour (reserved) or \$6.669/hour (on demand). Your "big data" problems now fit in RAM, on-demand, at a cost of one latte per hour. Don't overengineer! Hardware is real!

Gary Bernhardt @garybernhardt

Reminder to people whose "big data" is under a terabyte: servers with 1 TB RAM can be had about \$20k. Your data set fits in RAM.

💬 15

↻ 442

❤️ 952



A slight confession

- I'm not doing any really big data stuff in my day job in production, only for testing out new ideas
- I haven't crashed my own production cluster
- But I have helped crash other people's... Sorry!
- A big thank you to everyone in the Apache Tika community who has helped with material for this talk, and shared what has / hasn't worked

Not (really) an Apache Tika talk

- You don't need to be trying to turn random binary files into useful content to benefit from the talk
- If you are, great! Plenty of tips here
- If not, learn and adapt the approaches for your own problems and systems
- If you want to know more about Apache Tika and related projects, talk to me after! Or see the video from my talk a few years ago :)

Consider your Retries

www.quanticate.com

Is it the machine or the input?

- Many Big Data systems assume that they're running on cheap unreliable hardware, with flakey networking, and problems are node related
- Task failed? Probably a bad node, try another
- Task taking ages? Speculatively try another
- Maybe it is the node / related to the node
- Maybe it's the input, and we've killed many nodes!

Not only Big Data - In IoT

- You can make the retry mistake in other world too, eg IoT (and this is a real example!)
- Perhaps you're doing vehicle tracking, around the world, and use cellular data to communicate
- If you don't get an answer, maybe there's poor / no signal, maybe vehicle is switched off
- Or maybe the message you keep retrying with is causing an OOM on the device when handled...

Consider your defaults

- What kinds of failures do you tend to get?
- What's the impact if you retry too much?
- What's the impact if you don't retry enough?
- How will you record what has failed?
- How will you know what you can retry later, and how will you re-run just those things?
- How many failures before you give up?

Detecting File Types

www.quanticate.com

Isn't detecting simple?

- Surely you just know what a file is on your computer?
- Well, probably on your computer, and maybe elsewhere?
- OK, so maybe people rename things, but it's close, no?
- Ah, the internet... But that's normally right isn't it?
- Hmm, well most web servers tell the truth, right?
- They wouldn't get it that wrong?
- A few percent of the internet, that's hardly that much?
- And people would never rename things by accident?
- Operating Systems would never “help”, would they?

Filenames

- Filenames – normally, but not always have extensions
- There aren't that many extension combinations
- There are probably more file formats than that
- No official way to reserve an extension
- So everyone just picks a “sensible” one, and hopes that don't have (too many) clashes...
- What happens if you rename a file though?
- Or have a file without one?
- *Quick, but dirty, and may not be right...*

Mime Magic

- Most file formats have a well known structure
- Most of these have a (mostly) unique pattern near the start
- These are often called Mime Magic Numbers
- In some cases, these are numbers
- More commonly, they're some sort of number / text / bit mask

- Ideally located at a fixed offset, even better, right at the start of the file
- *But not always...*

More on Magic

- PDFs (should) start with `%PDF-`
- Microsoft Office OLE2 docs start with `0xd0cf11e0a1b11ae1`
- Most Zip files start with `PK\003\004`
- AIFF starts with `FORM????AI(FF|FC)` (mask 5-8)
- PE Executables normally have `PE\000\000` at 128 or 240

- Not all of these are true constants
- Not all of these are unique - `0xffffe` can be UTF-16LE or MP3
- Container formats – Zip can be Zip, OOXML, iWorks etc

Container Formats

- Some file formats are actually containers, and can hold lots of different things in them
- For example, a **.zip** file could just be a zip of random files
- Or it could be a Microsoft OOXML file (eg **.docx**, **.pptx**)
- Or it could be an OpenDocument Format file (eg **.ods**)
- Or it could be an iWorks file (Numbers, Pages, Keynote)
- Or it could be an ePub file
- Or....
- A **.ogg** could be audio, video, text, or many!

Taking a best guess

- For the “what kind of File is this this” case, Apache Tika can help! Has detection methods
- Need to combine potentially several different approaches, and weight up likelihoods
- Sometimes we’ll get it wrong, what then?
- We may later improve, what then?
- How good is your own input data? How stable?

Build and Deploy?

www.quanticate.com

How are you deploying code?

- **TIKA-2643** – Tika call hangs indefinitely when processing a PDF on Cloudera Hadoop
- Tika has a few deployment modes, including Standalone RESTful server, Commandline App, Batch mode, OSGi bundle, and plain Java Jars
- 2 main maven artifacts **tika-core** and **tika-parsers**
- You need to get jars + dependencies out to where running, and not any older / conflicting jars!

How are you deploying code?

- **TIKA-2643** – Cloudera Hadoop was shipping a very old version of Tika, was being used first
- Frameworks can give you “bonus” jars you weren’t expecting
poi.apache.org/faq.html#faq-N10006
- Your deployment steps might be missing out jars you do need, eg missed dependencies!
wiki.apache.org/tika/Troubleshooting%20Tika

Detecting Text

www.quanticate.com

Encodings 101

- Many different ways to encode text in a file
- “A” could be **0x41** (ascii,utf8 etc), **0x0041** (utf16le), **0x4100** (utf16be), **0xC1** (ebcdic)
- 1-byte-per-character encodings historically very common
- But means that you need lots of different encodings to cope with different languages and character sets
- **0xE1** – could be: **á α c Ɔ ف ۱** (or something else too!)
- Some formats include what encoding they've used
- But many key ones, including plain text, do not!

Languages

- Different languages have different common patterns of letters, based on words, spellings and patterns
- If you see accents like á è ç then it probably isn't English
- If you see a word starting with an S, it probably isn't Spanish, but if you see lots starting “ES” it might be
- You can look for these patterns, and use those to identify what language some text might be in
- Really needs quite a bit of text to work on though, it's very hard to make meaningful guesses on just a few letters!

n-grams

- Wikipedia says “An n-gram model is a type of probabilistic language model for predicting the next item in such a sequence in the form of a $(n - 1)$ –order Markov model”
- Basically, for us, it's all the possible character combinations (including start + end markers) along with their frequency
- The “n” is the size
- Trigrams of “hello” are []he, hel, ell, llo, lo[]
- Quadgrams of “hello” are []hel, ello, llo[]
- Can be used to identify both language and encoding

False Positives, Problems



- For encoding detection to work, your tool (eg Tika) needs to recognise the file as Plain Text
- Too many control characters near the start can cause Tika and friends to decide it isn't Plain Text, so won't detect
- Some encodings are very similar, hard to tell apart
- For short runs of text, very hard to be sure what it is
- Same pattern can crop up in different languages
- Same pattern could occur between different languages when in different encodings
- What happens if we got it wrong for you?

www.quanticate.com

JVM Bugs

Uwe is the expert on this!



- Uwe Schindler gave a great talk a few years ago on JVM Bugs and Lucene, go watch the video if you haven't seen it already!
- Things really can behave differently between different JVM versions, are you prepared?
- If you try a new framework / platform version / hosting provider, how will you check / detect issues
- Pending Tika fix for an infinite loop inside the JVM, happens for Java 8+9, fine in 7, fixed again in 10

www.quanticate.com

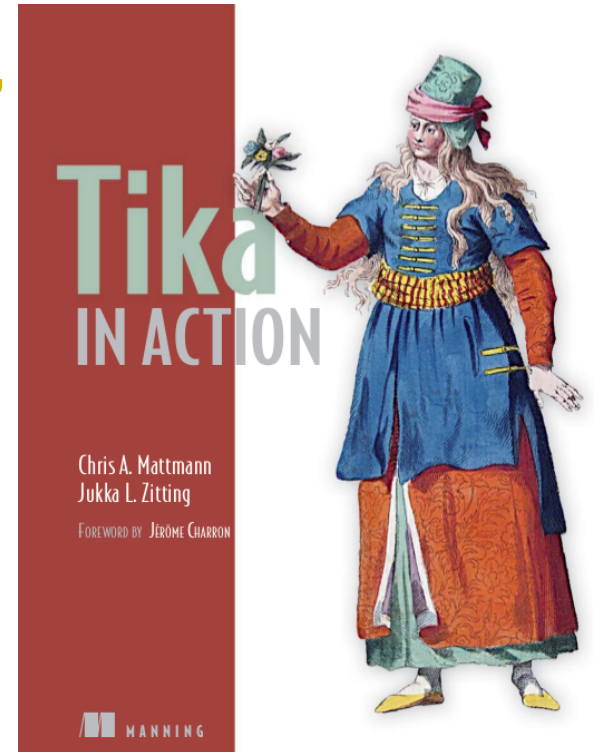
Apache Tika in a nutshell

www.quanticate.com

Apache Tika in a nutshell

“small, yellow and leech-like, and probably the oddest thing in the Universe”

- Like a Babel Fish for content!
- Helps you work out what sort of thing your content (1s & 0s) is
- Helps you extract the metadata from it, in a consistent way
- Lets you get a plain text version of your content, eg for full text indexing
- Provides a rich (XHTML) version too



(Some) Supported Formats



- Microsoft Office – Word, Excel, PowerPoint, Works, Publisher, Visio – Binary and OOXML formats
- OpenDocument (OpenOffice),
- iWorks – Keynote, Pages, Numbers
- HTML, XHTML, XML, PDF, RTF, Plain Text, CHM Help
- Compression / Archive – Zip, Tar, Ar, 7z, bz2, gz etc
- Atom, RSS, ePub Lots of Scientific formats
- Audio – MP3, MP4, Vorbis, Opus, Speex, MIDI, Wav
- Image – JPEG, TIFF, PNG, BMP, GIF, ICO

www.quanticate.com

Tika Offers

You can have any, all or none of these:

- Type Detection – what is it?
- Language + Encoding – what’s it in?
- Metadata
- Plain Text
- XHTML
- Batch mode – Try all of these
- Comparisons – Is 1.14 RC1 better or worse than 1.12?
- Top “mistakes” - Where to spend my time?

Going wrong at Scale...

www.quanticate.com

Lots of Data is Junk

- At scale, you're going to hit lots of edge cases
- At scale, you're going to come across lots of junk or corrupted documents
- 1% of a lot is still a lot...
- Bound to find files which are unusual or corrupted enough to be mis-identified
- You need to plan for failures!

Unusual Types

- If you're working on a big data scale, you're bound to come across lots of valid but unusual + unknown files
- You're never going to be able to add support for all of them!
- You should consider adding support for the more common “uncommon” unsupported types you hit
- Which means you'll need to track something about the files you couldn't understand
- If Tika knows the mimetype but has no parser available for it, just log the mimetype
- If mimetype unknown, maybe log first few bytes

Failure at Scale

- Tika will sometimes mis-identify something, so sometimes the wrong parser will run and object
- Some files will cause parsers or their underlying libraries to do something silly, such as use lots of memory or get into loops with lots to do
- Some files will cause parsers or their underlying libraries to OOM, or infinite loop, or something else bad
- Some files could (correctly or incorrectly) balloon out – how much data are you expecting back, what if too much?
- My typical may not be your typical

Log your problems!

- If you don't know where things are going wrong, how do you know where to spend your time fixing?
- Log unexpected / unsupported input (for Tika mimetypes or start of file), so you know what your most common “ignored inputs” are to work on
- Log your failures, including stack traces
- Ship the logs somewhere central
- Summarise and check from time to time!

Did you log enough?


- If your logs suggest there's a common problem occurring, and you want to investigate, have you got enough info to know which inputs are failing?
- Do you have enough details to be able to reproduce the issue, or at least have an idea of how for intermittent issues?
- Apache Lucene randomises settings when unit testing, but logs failing settings, can you copy?
- If you can't share input, can you find one you can?

Getting Better? Or Worse?

www.quanticate.com

When things go wrong

Taking a close look at the forest or open meadows reveals that there are often subtle differences in plant species across a wide landscape. Unique micro-climates, exposure to the sun, soil types, moisture availability, and a variety of other factors influence the types of plant species present in any given location. Changes in any of these factors will cause changes to



BGQOTM G IRUYK RUUQ GZ ZNK LUXKYZ UX UVKT SKGJU]Y
XK\KGRY ZNGZ ZNKXX GXK ULZKT Y[HZRK JOLKXKTIKY OT VRGTZ
YVKIOKY GIXUYY G]OJK RGTJYIGVK% CTOW[K SOIXU-
IROSGZKY\$K^VUY[XK ZU ZNK Y[T\$ YUOR Z_VKY\$ SUOYZ[XK
G\GORGHORZ_\$GTJ G \GXOKZ_ UL UZNXK LGIZUXY OTLR[KTIK ZNK
Z_VKY UL VRGTZ YVKIOKY VXKYKTZ OT GT_MO\KT RUIGZOUT%
4NGTMKY OT GT_ UL ZNKYK LGIZUXY]ORR IG[YK INGTMKY ZU

When things go wrong

You don't know what you can't find...

Statement Seasoned professional with a skilled ability to connect co-workers and clients with the information, products and services they are seeking by utilizing professional experiences, organizational and client skills both as a team and an individual.

Experience OLS: Office Liquidations Solutions May 2010 – May 2013

Statement

**OLS: Office Liquidations Solutions May
2010 – May 2013**

Experience

**Bialek Healthcare Environments June 2001
– May 2010**

Bialek Healthcare Environments June 2001 – May 2010

Design Associate, Client Services Coordinator

Furniture bid package review, quotation, response and presentation. Small office design, space planning, cost estimation, conceptual and quality for commercial interiors and food service.

What can go wrong

- Catastrophic failures
 - Out of Memory Errors
 - Infinite Hangs
 - Memory Leaks
- Exceptions: Null Pointer, etc.
- Extraction with loss of fidelity
- Missing text/metadata/attachments
- Extra text (eg placeholder, dummy, hidden, internal etc)
- Garbled text

Did this change help?

- Unit tests can tell you about major failures
- Unit tests only check the things you know about though...
- And only check a small number of files of each type
- There is a much wider variety of files out there than in even a decent-sized test suite
- File type distributions are uneven, a “minor issue” for me might be a “critical issue” for someone else!
- You can check 10 documents by eye, you can't check 10gb of files by eye, let alone 1tb!

Did this library change help?



- Large scale testing needed to spot these kinds of problems
- Needs automation of running *and* of analysis!
- Flag up key issues, look at those in detail

- Exceptions and Errors are easy (1st one per file anyway...)
- Metadata little harder, but sizes small
- Attachment counts easy, contents less so
- Finding junk text harder
- Missing text tougher still

www.quanticate.com

Did this library change help?



- Need to store output from many runs to compare
- Common terms per file can identify problems, if it goes from “differences” to “4NGTMKY” then something's wrong
- Common terms can help with missing text, but not always
- Token entropy can show garbage text, or data heavy files!

- Too many errors to fix all of them
- Need metrics to know which to focus on
- Govdocs1, Common Crawl output, more datasets needed!

www.quanticate.com

How will you re-process?



- You've analysed the difference from a change upgrade, and it looks like things got better. Great!
- You've upgraded the production system, including jars you didn't mean to have there... Great!
- But what about the files that can be helped?
- How can you re-run stuff that previously failed?
- How can you re-run things that might improve?
- Can you “restart” just bits of your pipeline? Why not?

www.quanticate.com

Why the same JVM?

www.quanticate.com

Does it need “Big Data”?



- Does this part of your problem actually need to run on your Big Data system?
- Could you run it as a pre-processing batch step elsewhere? eg Tika provides “Tika Batch” which processes a directory of files with timeouts and crash handling, could you read that output instead?
- If it’s infrequently used, could you have a pool of other services you call out to? eg Tika Server (REST)

www.quanticate.com

Does it need to be inline?



- What happens if some files take much longer than others? eg embedded images which you want to OCR + Image Recognition on? Zip of many other files?
- Can you punt these “hard” cases off to something else, then augment the original output later?
- Could this augmentation happen for other things to? eg NLP Sentiment Analysis, GROBID metadata?
- This augmentation of fails differently...

www.quanticate.com

Can you use another JVM?



- Tika provides an amazing, awesome, slightly terrifying “Forked Parser” option
- Dumps out all the jars Tika needs, but no others, to another folder, then spawns a second JVM
- Processing done in another JVM, result received back over a socket, child JVM can be killed and/or respawned if it fails
- However... 2nd JVM = 2nd bunch of memory required, and your parent system won't know about it!

www.quanticate.com

Did this change help?

www.quanticate.com

Tika Eval

- Runs Tika against a corpus of documents, or compares the results of running two different versions on those documents
 - Stack traces and exception counts
 - File id, language id
 - Attachment and metadata counts
 - Top 10 most common words
 - Content length
 - Token length statistics, Token entropy
- Gives reports on key things for a human to check

Tika Eval

- Datasets, Explanations and Run Results at <http://162.242.228.174/index.html>
- Datasets include:
 - Govdocs1
 - Common Crawl extracts
 - Fraunhofer Institute test library
 - IUST-HTMLCharDet
- Contains broken files, malware etc, take care!
- <https://issues.apache.org/jira/browse/TIKA-1302> for history

Tika Eval

	A	B
1	MIME_A_TO_MIME_B	COUNT
2	application/vnd.ms-excel -> application/vnd.ms-graph	27888
3	application/epub+zip(NEWLINE) -> application/epub+zip	855
4	text/html; charset=windows-1252 -> application/x-mobipocket-ebook	829
5	application/octet-stream -> application/vnd.wordperfect	734
6	application/octet-stream -> application/x-mobipocket-ebook	301
7	application/octet-stream -> multipart/appledouble	280
8	text/html; charset=windows-1252 -> text/plain; charset=windows-1252	239
9	image/jpeg -> multipart/appledouble	169
10	model/vnd.dwf -> model/vnd.dwf; version=6	156
11	video/mp4 -> application/mp4	102
12	text/html; charset=UTF-8 -> application/xhtml+xml; charset=UTF-8	102
13	application/x-tika-ooxml -> application/vnd.ms-excel.sheet.macroenabled.12	96
14	application/x-tika-ooxml -> application/vnd.openxmlformats-officedocument.presentationml.presentation	90
15	text/plain; charset=UTF-8 -> text/plain; charset=ISO-8859-1	76
16	image/png -> multipart/appledouble	74
17	application/octet-stream -> application/x-shapefile	66

Tika Eval

- http://162.242.228.174/mimes/mime_comparisons.html

	A	B
1	tika vs droid	cnt
2	text/plain <-> application/octet-stream	389495
3	text/plain <-> text/calendar	76511
4	application/pdf <-> application/octet-stream	57105
5	application/x-tika-ooxml <-> application/octet-stream	32150
6	application/x-tika-msoffice <-> application/x-puid-fmt-111	23981
7	image/jpeg <-> application/octet-stream	20935
8	application/gzip <-> application/x-gzip	18839
9	message/rfc822 <-> application/octet-stream	18239
10	text/html <-> application/octet-stream	16425
11	application/xhtml+xml <-> application/x-puid-fmt-471	15974
12	text/html <-> application/x-puid-fmt-471	15188
13	image/png <-> application/octet-stream	10529
14	application/xhtml+xml <-> text/html	9407
15	application/x-mspublisher <-> application/x-puid-x-fmt-257	7670
16	application/zip <-> application/octet-stream	6746
17	application/xml <-> application/octet-stream	6520
18	text/plain <-> text/x-vcalendar	5537
19	text/html <-> application/x-puid-fmt-584	4874
20	application/rss+xml <-> application/xml	4870
21	image/gif <-> application/octet-stream	4792

What's your Tika Eval equiv?



- How do you track the impact of your changes?
- How do you get a quick sense of “did this help?”
- How do you know what even changed? “Your fix last week broke everything” - but was it last week's?
- What dashboards / metrics do you have? Do you use them? Do they help?
- Are you logging the right amount of information? And the right stuff? And not doing it from scratch!

www.quanticate.com

Perils of a good quickstart

www.quanticate.com

QuickStarts are a must

- The way most people code is changing, partly as a lot more people code, which is good!
- Many people want some code that works now, if the project works they'll understand it later
- Don't want to invest lots of time until proven
- Easy quickstart is a must for a successful project, along with code snippets, stackoverflow etc
- How do you warn of “sharp edges” for production?

SOLR DIH, Tika App

- SOLR's **DataImportHandler** / **TikaEntityProcessor** lets you send binary files to SOLR, it'll have them processed by Tika and index text + metadata
- Really good for demos and getting started!
- Lets you quickly see what you have, plan for how you might integrate, what data / fields you want
- Runs in the main indexing thread! No production!
- **Tika App** – giant runnable jar, text and gui modes, great for testing and debug, **Tika Server** for prod

Think Data Pipelines

www.quanticate.com

Code steps, plan pipelines

- This stuff doesn't exist in isolation
- How will it fit with the other parts of your system?
- How will the data get in? How often?
- Where does the output go? What uses it next?
How quickly does it need to come out?
- When you improve things, what needs re-running?
- When you add another step, what to re-run?
- How will you handle versioning / changes?

Expect Failure

- Things will go wrong!
- Things will fail in ways you can't yet imagine...
- What are you logging, what are you sampling, how are you alerting, how are you investigating?
- When should something abort? When should auto-scale cancel? How to tidy up? Billing alerts?
- You have X days to tackle some technical debt, what should you focus on? What are quick wins?

Share your failures!

- Report bugs!
- Contribute fixes if you can, good bug reports if not
- Content failures normally happen on the most private of documents... For Tika share stacktrace and we'll try to find a public file with same problem
- Talk / Blog / Tweet / Present about what you've done, what worked well, what didn't
- Get your junior devs to present, not just seniors

Any Questions?

www.quanticate.com

Two related talks

- (Most of the conference is related!)
- Scalable OCR pipelines using Python, Tensorflow and Tesseract - Mark Keinhörster
Tomorrow, 14:40 to 15:00, Palais Atelier
- Calculating recommendations based on product images - Vlad Dolezal
Check for the video, you've missed it...

Nick Burch @Gagravarr

www.quanticate.com