

100% Big Data
0% Hadoop
0% Java

Pavlo Baron, codecentric AG



- pavlo.baron@codecentric.de
- [@pavlobaron](https://twitter.com/pavlobaron)
- github.com/pavlobaron

So here is the short story...

СОЧИНЕНІЯ
ГРАФА
Л. Н. ТОЛСТАГО.

ЧАСТЬ ПЯТАЯ.

ВОЙНА И МИРЪ.

I

ИЗДАНИЕ ТРЕТЬЕ.

МОСКВА.

Въ Университетской типографии (Колосовъ и Б-е)
и въ Спасскомъ переулке.

1873.

sitting there, listening...



The screenshot shows the Attensity website homepage. At the top, there's a navigation bar with the Attensity logo, a language selector set to 'English', and links for Contact, Resources, Support, Blog, and a search bar. Below this is a secondary navigation bar with links for Products, Solutions, Services, Customers, Partners, and Company.

The main hero section features a large dark grey box with the headline "Your real-time window into the social web." Below the headline is a quote from Yahoo! and a green "Learn More" button. To the right of the hero section are several overlapping images of tablets and laptops displaying various social media analytics dashboards with charts and graphs.

Below the hero section, there are three tabs: "Attensity for Marketing", "Attensity for Customer Service", and "Attensity for IT". The "Attensity for Marketing" tab is active, showing a section titled "Increase the effectiveness of your social marketing strategies:" with a list of four bullet points and a "LEARN MORE" link.

To the right of the marketing section is a "Success Story" for JetBlue Airways with a "DOWNLOAD NOW" link, and a "White Paper" titled "Social Intelligence Benchmark Report" also with a "DOWNLOAD NOW" link.

Further right is an "About Attensity" section stating they are the leading provider of social analytics and engagement solutions, followed by the slogan "Listen. Analyze. Relate. Act." and a "LEARN MORE" link.

On the far right is a "Watch Video" section for the "Command Center Video" with a play button icon, and a "Request Info" section with a "REGISTER TO LEARN MORE" link and a "Newsletter" section with a "SUBSCRIBE NOW" link.

presented as Houdini magic...



so you telling me...



it's smoke and mirrors?

Smells like a bunch of queues,
pipes and filters...



Looks like some NLP...



Sounds like some math...

$$\left[\frac{1}{2} Z_2(\alpha, \alpha) + \frac{1}{24} Z_4(\alpha, \alpha, \alpha, \alpha) + \dots \right]$$

$$(S(V^*) \otimes \wedge^*(V))$$

$$V = \bigoplus_{i \geq 0} V_i$$

$$d_{i+1} V_i \subset V_{i+1}$$

$$V^* = \bigoplus_{i \geq 0} V_i^*$$

$$\mathcal{L}_{\text{fin}}^2(V) \xrightarrow{\sim} \mathcal{L}_{\text{fin}}(V)$$

$$\mathcal{L}_{\text{fin}}^2(V) \xrightarrow{\text{quas}} \mathcal{L}_{\text{fin}}(V)$$

$$\mathcal{L}_{\text{fin}}^2(V) \xrightarrow{\text{Hoch}} \mathcal{L}_{\text{fin}}(V)$$

Seems like some basic ML...



methinks: I can tinker that.
I have 2 nights in the hotel...



Fire!



Know the use cases...



Consume a feed where people
say what they think before
they think what they say...



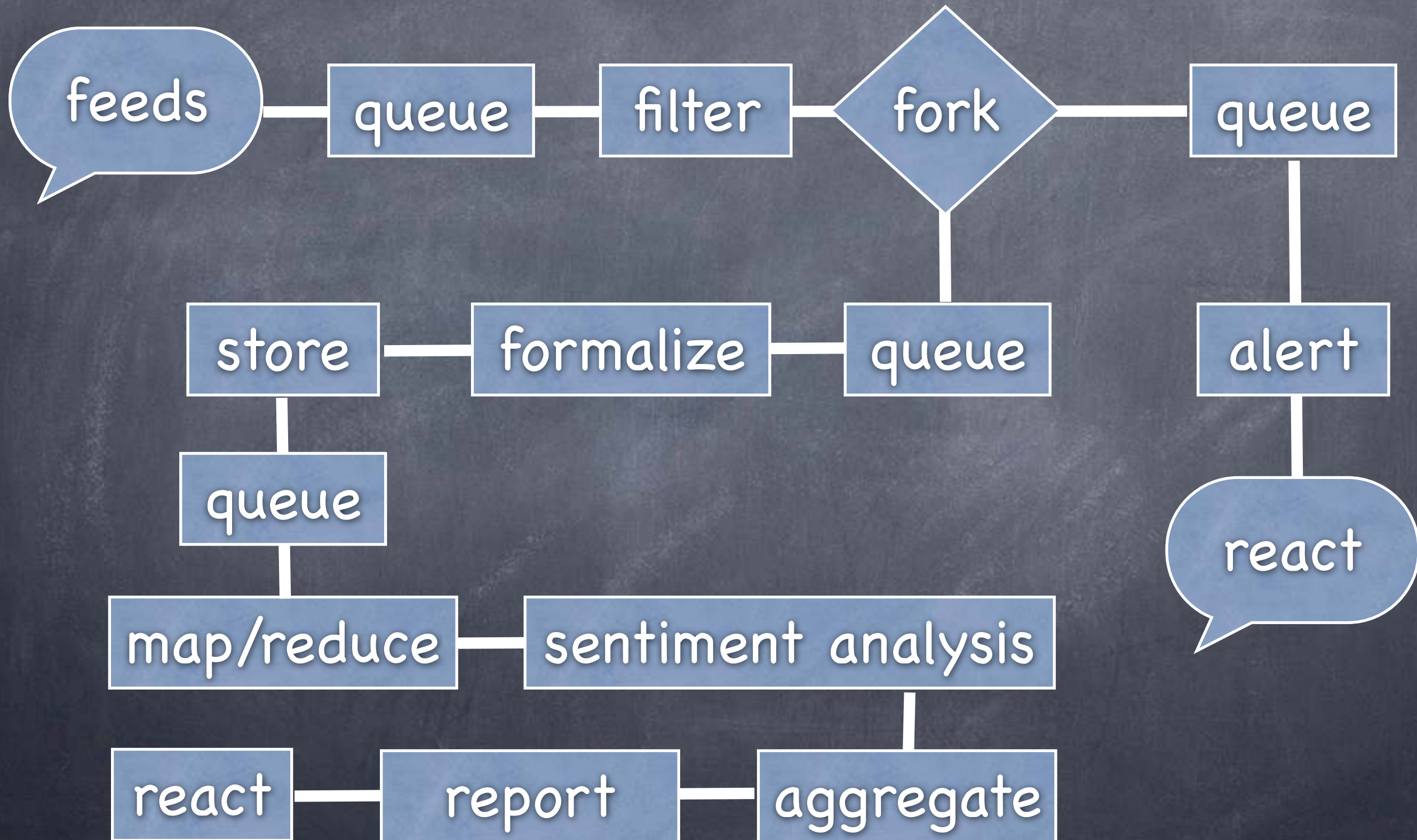
Drink Big Data warm, straight from the fire hose...



Then fork for immediate
notification and batch
analytics...



Some bubbles



Some tech

- Languages: Python, Erlang
- Feeds: Tweepy, crawlers, feed readers
- Queueing: RabbitMQ through Pika
- Store: Riak through protobufs
- Map/reduce: modified Disco to run workers on Riak-nodes data-locally

Some math

- Analytics: NLP with NLTK
- Algo training: nltk-trainer with pickle=true
- Algos: naive Bayes, decision tree, binary classification based on trigram frequencies
- simple name and antiword filtering based on public and own corpora

Some numbers...



... 'cause numbers are sexy

When numbers
become too
sexy for your
[hat|car|cat],
they mutate
into
numbers pr0n



Some numbers, revisited

- I'm not into numbers pr0n
- numbers need to be just good enough for what you're trying to solve

But it's still the
easiest way to
impress,
especially
without
solving a
concrete
problem



So, finally, some numbers (on my MBA)

- Feed: ~10000 chaotic msg/min
- Store: ~8000 formalized msg/min, N=3, quorum, 3 nodes
- Analytics: ~7000 msg/min (filtered, pos/neg aggregation, location based aggregation)
- Demo: ~1500000 tweets, pos/neg aggregation, stream processing in ~7min, map/red in ~15sec

Some lessons learned



The Beliebers...

- 

Maria @IGoWildForBiebs 14m
Retweet If your Twitter is about **Justin Bieber** ♥
Expand
- 

BassCannonKaplan @Avi_Kaplan 19m
"@scotthoying: **Justin Bieber**'s mom's ringtone is @avi_kaplan speaking; not kidding" TRUE. #dying
Expand
- 

Ellen DeGeneres @EllenDeGenares 23m
Who wants tickets to see 1D and **Justin Bieber** on my show? If you do, follow @MenHumor and retweet this!
Expand
- 

Scott Hoying @scotthoying 25m
Justin Bieber's mom's ringtone is @avi_kaplan speaking; not kidding
Expand
- 

Mitch Grassi @mitchgrassi 26m
Justin Bieber's mom is very sweet and made Avi record a voice memo of him talking like Barry White
Expand
- 

FUTURES @futuresband 45m
Cheers @justinbieber. Get yours merch.wearefutures.co.uk
twitpic.com/az4z8d
View photo
- 

← **BELIEVE IN HIM** ∞ @JustinFirstKiss 47m
OMG = OMB. Girl = Shawty. Lets Go = Leggo. Peace = Payce. Style = Swag. Fan = Belieber. Inspiration = **Justin Bieber** :)

More than 60% of the
Twitter sample stream is
useless garbage...



Real names...



DHH 

@dhh

Creator of Ruby on Rails, Partner at 37signals, Co-author of NYT Best-Seller Rework, and racing driver in ALMS.

Chicago, USA · <http://david.heinemeierhansson.com>

Absurd profile bios...



Location...



Zvi

@nivertech

FOLLOWS YOU

Chief Founder Monkey @ ZADATA

Israel - <http://www.ZADATA.com>



Pavlo Baron

@pavlobaron

Just me. And some sushi.

Senior Rubber Duck - <http://www.pbit.org>

Language...



For trigrams in NLTK, use Spanish as “anti-class” to tell English/German from the rest

Disco workers on Riak nodes...

- PITA and a lot of tinkering, but necessary for data locality
- Extending Disco is relatively easy, but changing it is hard...
- Flooding, asynchronous, separate key/value listing in low-level Riak goes very well with Erlang port based Python/Erlang message exchange in Disco. Not
- Extended Disco to use RabbitMQ between the worlds (h/t Dan North for the idea)

Mixing Python and Erlang in one project...

- Forgetting punctuation in Erlang code all the time when quickly switching from Python
- Terribly missing pattern matching in Python
- Considering to embed Python in Erlang, but it might become a double PITA then

Sentiment analysis...



Well, actually, strong
sentiment analysis...



Very unreliable given the
human nature...



nicolette @nikkyelago

8m

Omg it's cold as **fuck**. I love it

Expand

In addition to the NLTK's movie reviews corpus, use these for "neg" classification

Linus Torvalds goes off on Linux and Git

September 25th, 2012 | Programming, Satire

I was in a coffee shop in Portland, Oregon and happened to spot Linus Torvalds sitting alone at a window table. I asked the creator of the Linux operating system and the Git source code about his feelings on him. Over the next fifteen minutes we talked about programming.

Mr. Lr

Programmers Need To Learn Statistics Or I Will Kill Them All

FAQ



Q: Why the heck are you
doing this?



A

- Because I can
- Because I want
- Because I want to learn
- Because I want to go deep on low-level
- Because it's very interesting to combine computer science with math

Q: Why not just use Hadoop?



A

- Because I didn't want to run this on the JVM
- Because I have 2 use cases, and only one of them is suitable for batch map/reduce

Q: Why didn't you want to
run this on the JVM?



A: well, technically seen,
Big Data area is growing
on the JVM

- Hadoop
- Pig
- Storm, Kafka, Esper
- Mahout
- OpenNLP

A: but I didn't want this
Big Data on my drive

~/m2

A: and I am evaluating some alternatives to the ecosystem



Q: Why are you queueing at all? Others do gazillions of msg/sec without queues



A

- I could, if instead of filters and batch analytics of chaotic text, it would be just about building trivial sums
- with growable numbers like this, you want to protect any sort of reliable data store from getting flooded by writes, RDBMS or NoSQL store
- Because I need to do some pipes and filters
- Because I'm mixing and crossing borders of data sources and technologies
- Because almost all frameworks that you might consider also do some queueing or buffering

Q: Why did you use Erlang and Python?



A

- Because reliability and distribution are built into the Erlang VM and I don't need separate coordinators or to reinvent the wheel
- Because both, Python and Erlang, are "functional" enough for what I need day-by-day
- Because Python has been for many years the platform of choice for scientists, thus there are available clever and mature math libraries
- Because Disco is on Python and Erlang, Riak and RabbitMQ are on Erlang

Q: isn't Python slow like hell?



A

- it's not operating at the speed of light
- yes, it is slower at some points
- I'm also testing PyPy to improve performance for the case I should need it, 'cause right now it works just fast enough without explicit bottle-necks in the given architecture, even on one single MBA

Q: MBA is boring. Can you make it real web scale?



A

- well, to be precise, I'm operating on web data
- I can scale queues with RabbitMQ
- I can scale storage with Riak
- I can scale the map/reduce supported analytics with Disco/Riak
- I can scale data sources/feeds, machines, hardware, networks, infrastructure, logins etc. You name it

Q: what's in the future?



A

- I don't have my crystal ball with me
- I've started to implement Pig Latin engine in Python called "Sau" (German for pig), to offer data scientists a comfortable interface and to allow them to run existing Pig scripts on this stack
- I'm going to add more data sources, improve throughput where necessary and work on some low level Disco modifications to change the way it utilizes Erlang in my case
- We will integrate my Disco extensions with Disco 0.5

Q: what do we learn about
Big Data here?



A

- Big Data is about the “what”, followed by the “how” and enabled by the “what with”

A

- It's about gathering data, analyzing it, gaining useful information out of it, finding new ways to gather and use information and deriving steps for business improvements, strategy planning, doing soft intelligence aka enterprise level stalking or, even more important, helping make the world a better place – it's up to you

A

- It's not about building SkyNet – even if this will be built one day, it will be pretty boring. It's about building recommender and decision support systems, thus letting machines do stupid, repeated jobs fast and human beings make high quality decisions

A

- It's not about plain numbers. It's about numbers that are good enough to carry the solution. Not less, but also not more than that

A

- It's a huge field for geeks with aspiration to learn new things, dig into math and computer science, play with different platforms and tools and pick the right tool chain

Oh, and did the demo run?



Thank you!



Most images originate from
[istockphoto.com](https://www.istockphoto.com)

except few ones taken
from Wikipedia or Flickr (CC)
and product pages
or generated through public
online generators