Pipelining your music

<whoami>

Jyrki Pulliainen

Content team

Spotifier since January

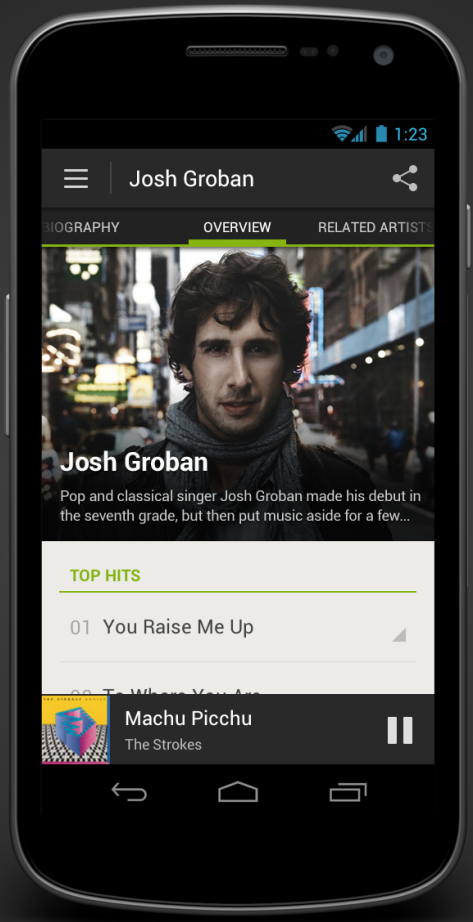
Pythonista since 2.3

@nailor

</whoami>

Spotify?


Spotify Premium - Linux Preview
File Edit View Playback Develop Help
numminen boogie
Jyrki Pulliainen
MAIN
What's New
People
Inbox
Play Queue
Devices
APPS
App Finder
Top Lists
Radio
Classify
The Guardian
Last.fm
Moodagent
Pitchfork
Songkick Concerts
Soundrop
COLLECTION
Library
Local Files
Downloads
Hide
CHRIS BROWN
LYSSNA NU
Innehåller hitsen Turn Up The Music och Don't Wake Me Up
Gästas av B.O.B., Snoop Dogg, Wiz Khalifa och Nas
DIGITALISM
DJ-KICKS
RECOMMENDED
DJ-Kicks
Digitalism
ADRIAN LUX
FEATURING DANTE
Featured Spotify Apps
1. TuneWiki
Get synced lyrics and sing along to your favorite songs!
ADD
2. Tunigo
Ett enkelt och underhållande sätt att upptäcka ny musik.
ADD
3. Digster
Playlists from your favorite artists and for all occasions!
ADDED
4. Filtr
Playlist generator with a social twist.
ADD
For Silence
Flying Saucer Attack
0:09
7:39
Aino Kopu
Anders Åstrand
All I Ever Wanted by Adria...
Jarno Tikka
The Gift Of Guilt by Gojira
Johanna Henriksson
Johannes Dahlström
Kalle Rindell
Lasse Saarinen
mharju
Holding Out For A Hero by ...
Mikko Harju
Richard Bengtsson listened to Will You Follo...
Anders Åstrand listened to All I Ever Wanted by...
Pierre Carrier listened to Never Mind by Infected...
Ville Kaisla listened to Percolator Meme (Lazer...
Mats Linander listened to Skinny Love by Bon Iver
Anders Åstrand listened to Clean Gloves, Dirty...
Mats Linander listened to Lump Sum by Bon Iver
Richard Bengtsson listened to Hollow Crown...
Ville Kaisla listened to Busy Crack by Teeth
Anders Åstrand listened to Too Close by Alex Clare
Mats Linander listened to Flume by Bon Iver
Richard Bengtsson


1:23
Josh Groban
BIOGRAPHY
OVERVIEW
RELATED ARTISTS
Josh Groban
Pop and classical singer Josh Groban made his debut in the seventh grade, but then put music aside for a few...
TOP HITS
01 You Raise Me Up
Machu Picchu
The Strokes

Spotify

Number fun

- 10M monthly active users
- 18M tracks
- 100 years of music
- 20k added **every** day

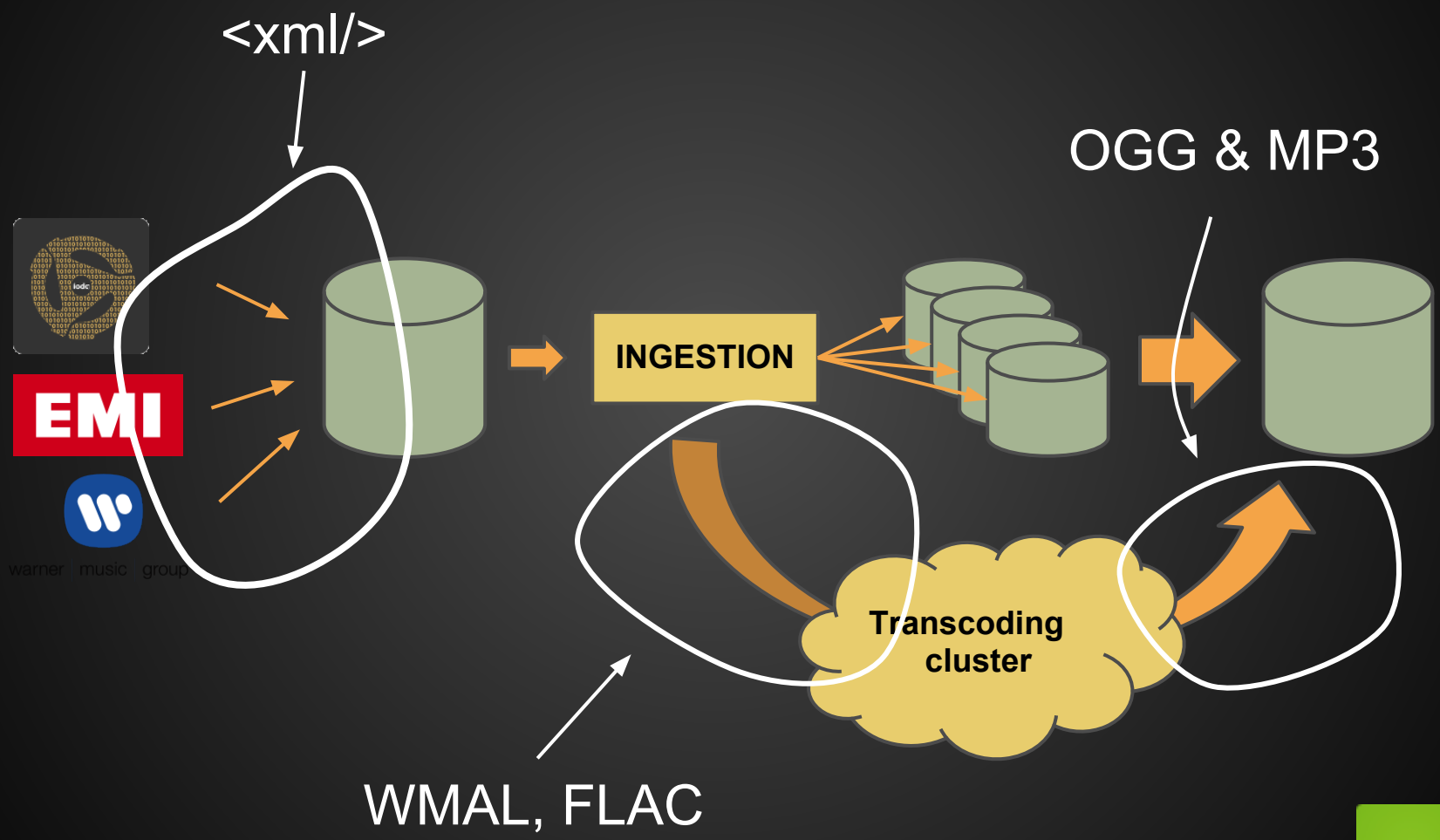
The Music Pipeline

ANTI HIPSTER

STACK

Hipster kitty by http://craigwheatart.tumblr.com

Spotify


<xml/>
OGG & MP3
EMI
warner music group
INGESTION
Transcoding cluster
WMAL, FLAC
Spotify

100s of TBs of data

Load of deliveries daily

Malformed data every day

Right tools for the right job

Spotify

Spotify

```
>>> def formerlify(_, name):
...     return 'The artist formerly known as %s' %name
>>> #Namespace stuff
>>> from lxml import etree
>>> ns = etree.FunctionNamespace('http://my.org/myfunctions')
>>> ns['hello'] = hello
>>> ns.prefix = 'f'
>>> root = etree.XML('<a><b>Prince</b></a>')
>>> print(root.xpath('f:hello(string(b))'))
... The artist formerly known as Prince
```

Fun(?) facts

- 10 different XML formats
    - Majors vs our own (indies)
    - One industry "standard"

Biggest XML 3.3M lines (350MB)

- Bible apparently fits in 3MB of XML

- lxml ftw

Spotify

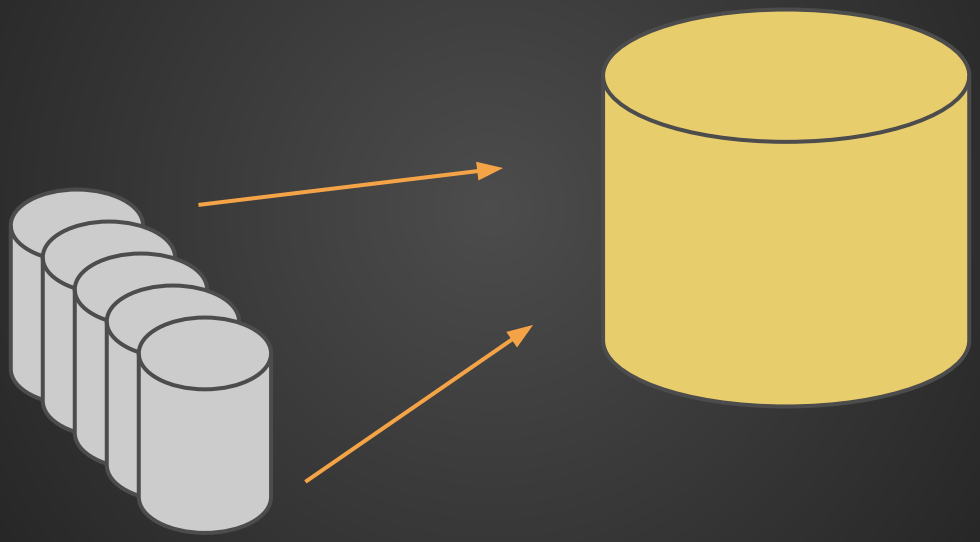
```
>>> min(timeit.repeat('etree.parse("huge.xml")', setup="from
lxml import etree", number=1, repeat=5))
2.309144973754883

>>> min(timeit.repeat('etree.parse("huge.xml")', setup="from
xml.etree import cElementTree as etree", number=1, repeat=5))
3.0681779384613037

>>> min(timeit.timeit('etree.parse("huge.xml")', setup="from
xml.etree import ElementTree as etree", repeat=5, number=1))
Killed

>>> # (with PyPy 1.9)
>>> min(timeit.repeat('etree.parse("huge.xml")', setup="from
xml.etree import ElementTree as etree, number=1, repeat=5))
23.186518907546997
```

Merging

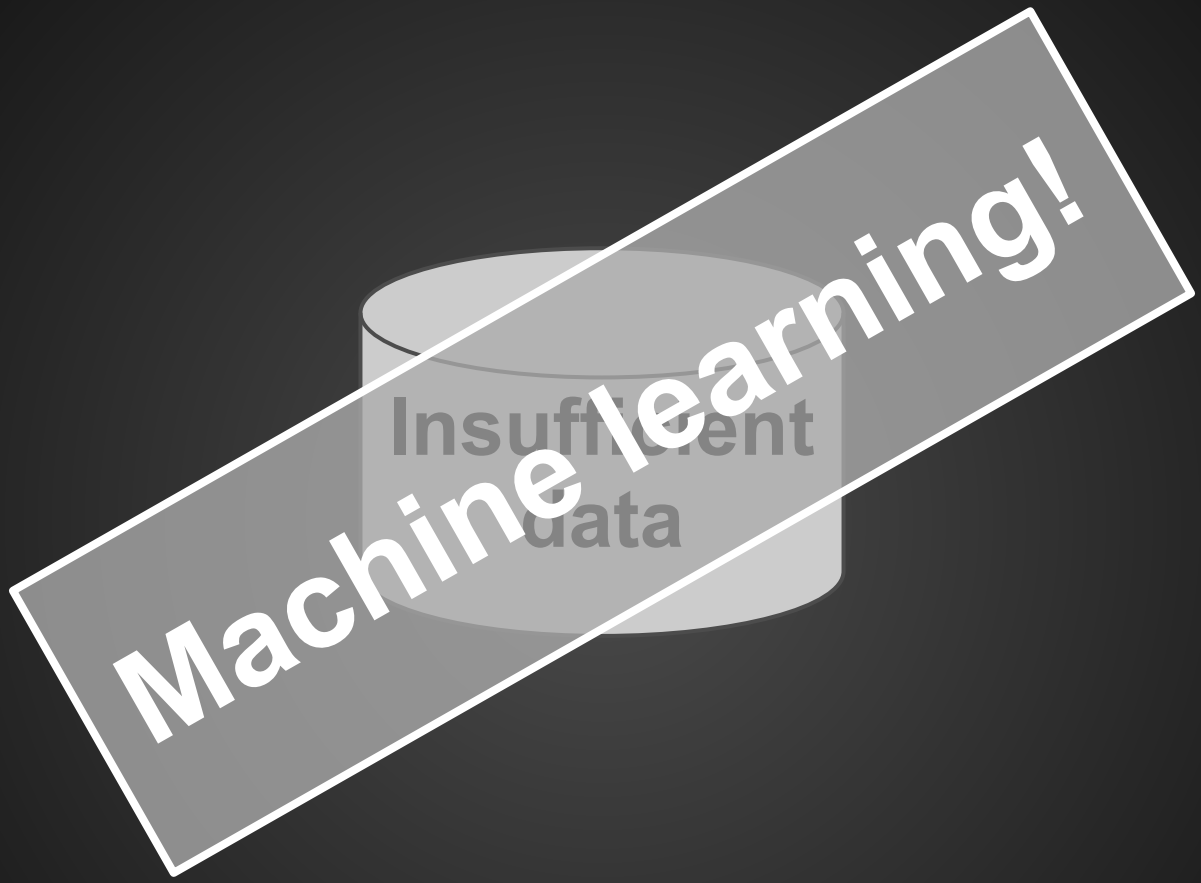


Spotify

Fun(?) facts

- Artists don't have any global or even label specific IDs
  - Multiple artists with same name
  - Even spelling differs inside a single label
- Multiple versions of the same album
- Enormous search space!
  - (18 * 10**6) ** 2 == huge number

Spotify



Spotify

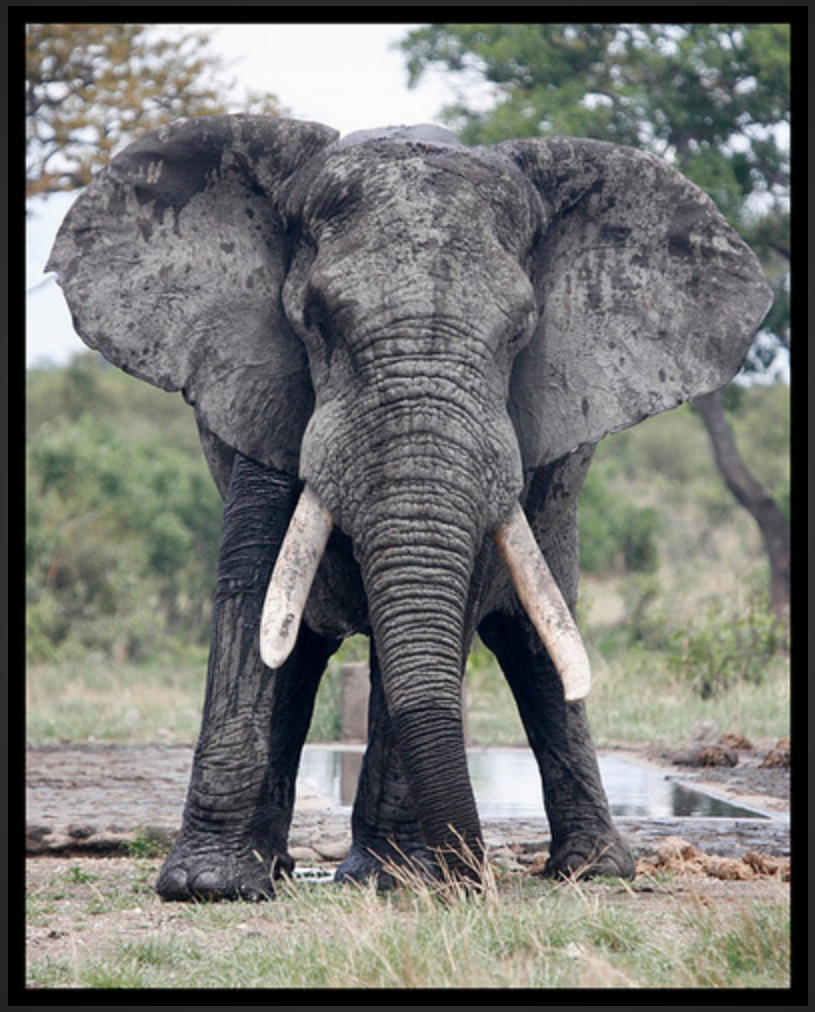
Tip: Reduce search space!

```
>>> from unicodedata import normalize
>>> key = ''.join(normalize('NFD', char)[0].lower() for char
in title)[5]
```

Side note: Levenshtein is expensive
=> use other edit distances too

(or use PyPy, 4x speed increase ftw)

Spotify



Pro-tips

- If data is relational, use relational database (duh)
- Don't over-normalize yourself, BCNF is rarely beneficial
- Weight between denormalize vs. moar indices
- Let the DB do the hard lifting, query planner is your friend!


SONY
STEREO
A
STAYIN' ALIVE!
IN
OUT
D-C60

- Asynchronous!
  - RabbitMQ + amqplib
    - One master, 49 slaves
- Isilon storage => 8Gbit/s throughput!

Spotify

Index building

....with Java

Why not Python?

- Not powerful enough for computationally intensive stuff
- We use Lucene for Search, so Java is a natural choice

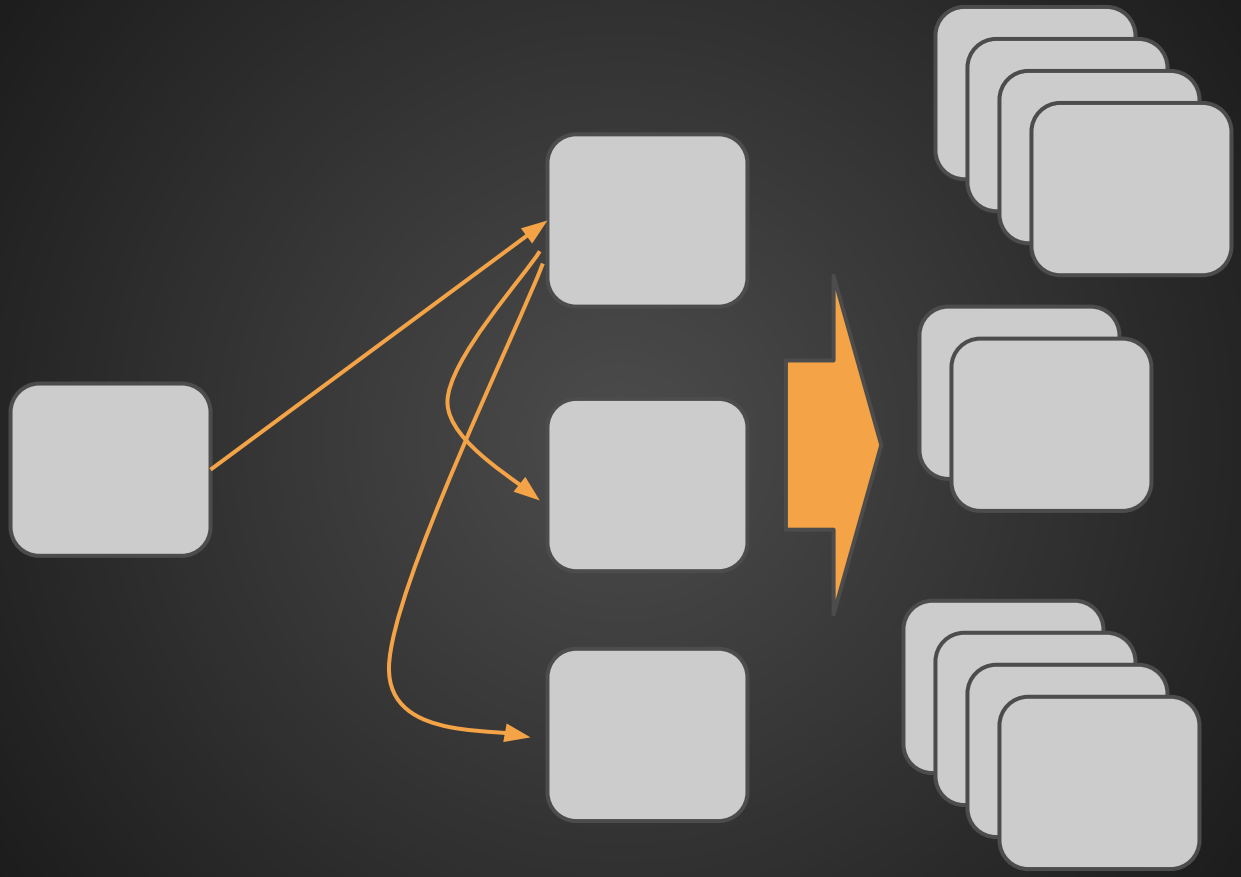
...but I'd like to try PyPy here.

The Music Distribution

TTL one day

time to live (laɪv)

Publishing an index



Spotify

SCPing around, moving hundreds of GBs daily

Future == BitTorrent

...not totally free of issues either

Index format?

Read only K/V (mostly)

Keep your eye on it

Mind the speed!

Experiment

Ditch your code

Thank you!

spoti.fi/ep_2012