Tracking Down The Bad Guys

Tom Barber - NASA JPL
Big Data Conference - Vilnius
Nov 2017

Who am I?

Tom Barber

Data Nerd

Open Source Business Intelligence Developer

Director of Meteorite BI and Spicule LTD

Software Developer at NASA JPL in the Computer Science for Data Intensive Applications Group

Some stuff I’ve worked on

Saiku Analytics - OLAP Browser

Apache OODT - Data Processing Toolkit originally by NASA

Pentaho Business Intelligence Suite - Business Intelligence Suite

Apache Tika - Metadata Extraction

Apache DRAT - Release Auditing

Sparkler - Spark Crawler

How I ended up working at NASA

It's an easy 6 step process!!

1. Want to be an astronaut
2. Learn about data
3. Volunteer some time on open source code and meet new people
4. Decide to change job direction
5. Ask random people for new job
6. Get offered a job at NASA

See.... easy!

Work we do at NASA JPL

We build applications for high volume data processing and visualisation.

- Genomics Search and data extraction
- Polar data discovery
- Data processing for a number of satellite programs
- Storage and archive of planetary data
- Track down criminals!!!

What is DARPA?

Defense Advanced Research Projects Agency

Founded in response to Sputnik

Invests in technology for (US) National Security

Work with external partners to bring knowledge to programs

The Challenge

Memex seeks to develop software that advances online search capabilities far beyond the current state of the art.

The goal is to invent better methods for interacting with and sharing information, so users can quickly and thoroughly organize and search subsets of information relevant to their individual interests.

The technologies developed in the program would provide the mechanisms for improved content discovery, information extraction, information retrieval, user collaboration and other key search functions.

(https://www.darpa.mil/program/memex)

The Challenge

Benefits of the program include:

- Development of next-generation of search technologies to revolutionize the discovery, organization and presentation of domain-specific content
- Creation of a new domain-specific search paradigm to discover relevant content and organize it in ways that are more immediately useful to specific tasks
- Extension of current search capabilities to the deep web and nontraditional content
- Improved interfaces for military, government and commercial enterprises to find and organize publically available information on the Internet

The Challenge

Initially, DARPA plans to develop Memex to address a key Defense Department mission: fighting human trafficking.

Human trafficking is a factor in many types of military, law enforcement and intelligence investigations and has a significant web presence to attract customers.

The use of forums, chats, advertisements, job postings, hidden services, etc., continues to enable a growing industry of modern slavery.

An index curated for the counter-trafficking domain, along with configurable interfaces for search and analysis, would enable new opportunities to uncover and defeat trafficking enterprises.

Who's On The Team?

DR. CHRIS MATTMANN
Principal Investigator

PAUL RAMIREZ
Co-Investigator

WAYNE BURKE
Data Scientist

DR. LEWIS MCGIBBNEY
Engineering Applications Software Engineer

ASITANG MISHRA
Software Engineer

SUJEN SHAH
Software Engineer

KYLE HUNDMAN
Data Scientist

LAUREN WONG
UX Designer

ROB TAPELLA
UX Designer

KARANJEET SINGH
Data Scientist Intern (USC)

And me!

Domain Search

Human Trafficking

Weapons Smuggling

Child Exploitation

Drug Smuggling

Financial Fraud

What Needed Building

Web Crawlers and Scrapers

Content Extraction Pipelines

Indexers

Visual Analytics

Who’s been involved?

DARPA

NASA JPL

Continuum Analytics

USC

Harvard

Kitware

Students and coders who like a challenge!

Technologies We’ve Developed

Sparkler

About The Spark Crawler

- Open Source Web Crawler
- Inspired By Apache Nutch
- Horizontally Scaling
- Runs On Top Of Apache Spark

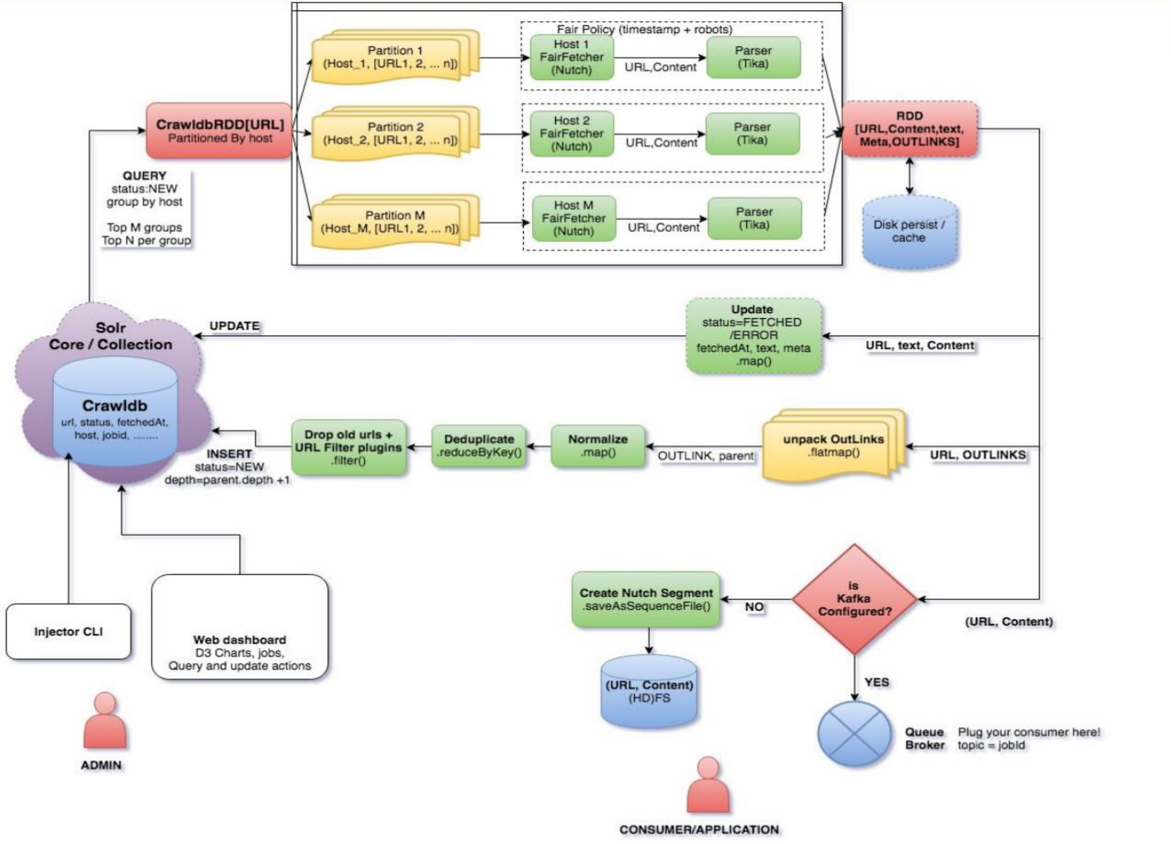
Important Design Choices

- Progress Reports - No more black box crawling
- Easy to deploy and use
- Newer libraries and technologies
- High Performance
- Fault Tolerant

Sparkler Technologies

- Batch Crawling
- Apache Solr
- Maven Build
- OSGI plugins
- Apache Kafka
- Apache Tika
- Data Visualisation with Banana

Sparkler Internals



Sparkler Power

Apache Lucene/Solr powered database

- Required an indexed database for real time analytics
- Exposure of the data structure via REST API
- Easy to build web applications over the top
- Scalable deployment options with Solr/Solr Cloud
- Spark and Solr interaction using the CrawldbRDD

Sparkler Partitioning

- Politeness
  - Don’t DDOS servers when running in distributed mode
- First Version
  - Group by: hostname
  - Sort by: depth, score
- Customisation
  - Definable Solr queries
  - Use Solr boosting to alter results
- Lazy Evaluations
- Delay Between Requests

Sparkler Plugins

- Interfaces inspired by Apache Nutch
- Using standard OSGI technology
- Migration of Nutch plugins to Sparkler

Sparkler Javascript Rendering

- First class support of Javascript Execution
- Can deploy on Spark Cluster without hacks
- Preserves cookies and sessions for future use

Sparkler & Kafka

- Output stream from crawler into 3rd party application
- Integration made easy via Queues
- Scaleable
- Distributable
- Fault Tolerant
- Can stream larger objects, videos etc

Sparkler & Tika

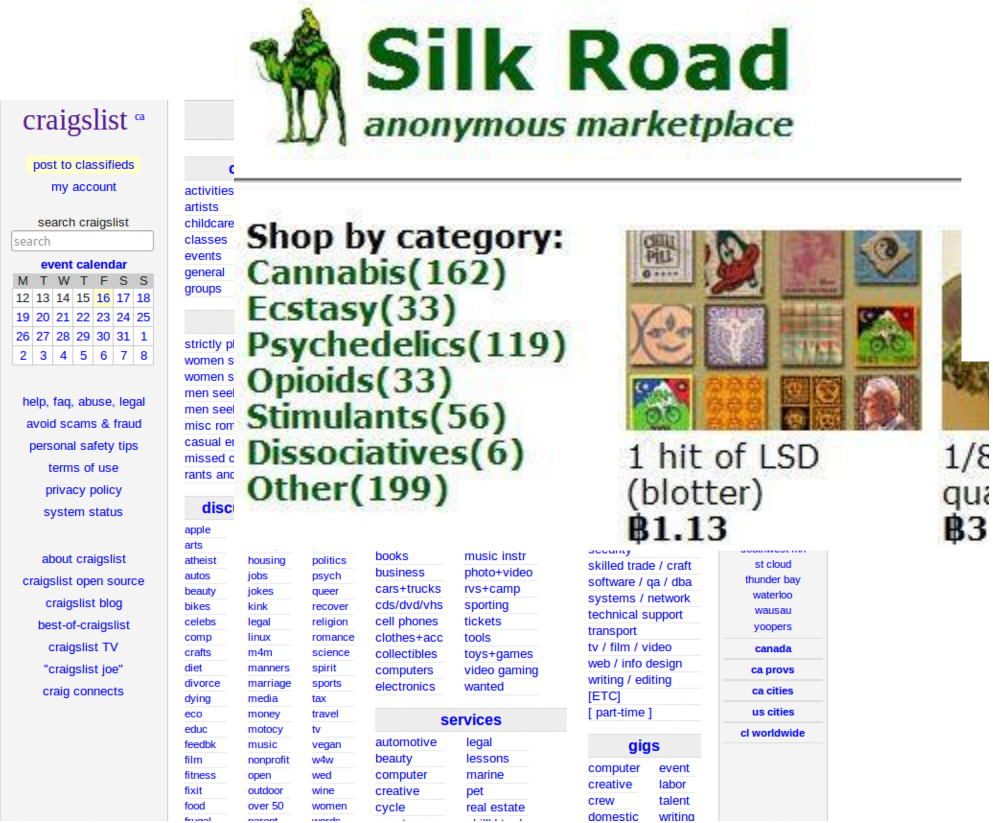
- Tika - A pluggable toolkit of parsers
- Detects and extracts metadata, text, urls, images and other content
- Can parse over 1000 content types
- Used in Sparkler to discover outgoing links

Sparkler Analytics

- Charts and Graphs - Show real time summary of crawl
- Distribution of URL’s across hosts/domains
- Temporal Activities
- Status Reports
- Easy to update
- Based upon Banana from Lucidworks (sadly discontinued)

Apache Tika

Data The Criminals Like



craigslist ca

post to classifieds

my account

search craigslist

search

event calendar

M	T	W	T	F	S	S
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	1
2	3	4	5	6	7	8

help, faq, abuse, legal

avoid scams & fraud

personal safety tips

terms of use

privacy policy

system status

about craigslist

craigslist open source

craigslist blog

best-of-craigslist

craigslist TV

"craigslist joe"

craig connects

c

activities
artists
childcare
classes
events
general
groups

strictly pl
women s
women s
men seel
men seel
misc rom
casual er
missed c
rants and

disc

apple
arts
atheist
autos
beauty
bikes
celebs
comp
crafts
diet
divorce
dying
eco
educ
feedbk
film
fitness
fixit
food
frugal

housing
jobs
jokes
kink
legal
linux
m4m
manners
marriage
media
money
motocy
music
nonprofit
open
outdoor
over 50
parent

politics
psych
queer
recover
religion
romance
science
spirit
sports
tax
travel
tv
vegan
w4w
wed
wine
women
words

books
business
cars+trucks
cds/dvd/vhs
cell phones
clothes+acc
collectibles
computers
electronics

music instr
photo+video
rvs+camp
sporting
tickets
tools
toys+games
video gaming
wanted

services

automotive
beauty
computer
creative
cycle

legal
lessons
marine
pet
real estate

security
skilled trade / craft
software / qa / dba
systems / network
technical support
transport
tv / film / video
web / info design
writing / editing
[ETC]
[part-time]

gigs

computer
creative
crew
domestic

event
labor
talent
writing

southwest mn
st cloud
thunder bay
waterloo
wausau
yoopers

canada

ca provs

ca cities

us cities

cl worldwide

Silk Road

anonymous marketplace

Shop by category:

Cannabis(162)

Ecstasy(33)

Psychedelics(119)

Opioids(33)

Stimulants(56)

Dissociatives(6)

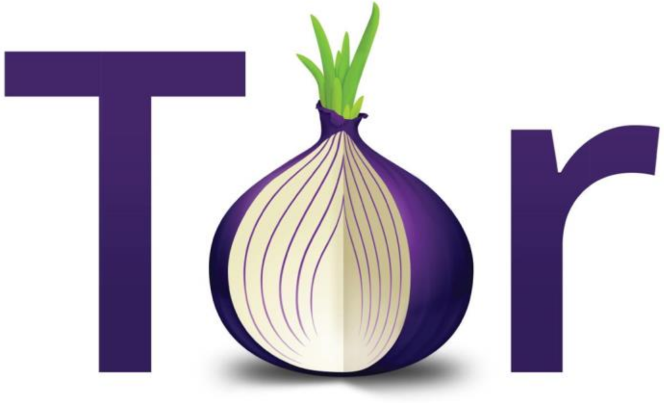
Other(199)

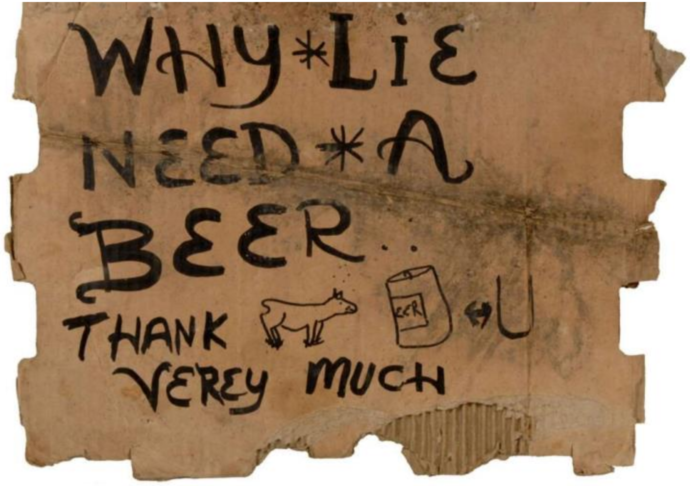
1 hit of LSD (blotter)

฿1.13

1/8 qua

฿3





How Many Content Types?!!!

- 16,000 to 51,000 content types and climbing
- They all share some things in common:
  - They need parsing
  - They will contain text and structure
  - They have metadata

Why Content Types Are Important

- They make computing convenient
- They help the internet function
- They allow us to search!

Third Party Parsers

- Most of the custom applications come with software libraries and tools to read/write these files
- Rather than re-invent the wheel, figure out a way to take advantage of them
- Parsing text and structure is a difficult problem
- Not all libraries parse text in equivalent manners
- Some are faster than others
- Some are more reliable than others

Metadata Extraction

- Important to follow common Metadata models
    - Dublin Core
    - Word Metadata
    - XMP
    - EXIF
    - Lots of standards and models out there
        - The use and extraction of common models allows for content intercomparison
    - All standardizes mechanisms for searching
    - You always know for X file type that field Y is there and of type String or Int or Date

Language Identification

- Hard to parse out text and metadata from different languages
    - French document: J’aime la classe de CS 572!
    - Metadata:
        - Publisher: L’Universitaire de Californie en Etas-Unis de Sud
        - English document: I love the CS 572 class!
    - Metadata:
        - Publisher: University of Southern California
- How to compare these 2 extracted texts and sets of metadata when they are in different languages?
- How to translate them?

We need to find out if this:

Is a weapons smuggler

Or, just a guy with a gun!

Or both...

A Bit About Tika

- A content analysis and detection toolkit
- A set of Java APIs providing MIME type detection, language identification, integration of various parsing libraries
- A rich Metadata API for representing different Metadata models
- A command line interface to the underlying Java code
- A GUI interface to the Java code
- Translation API
- REST server
- Easy to extend
- Under heavy development

Improvements made to Tika

- Tika doesn’t natively handle images and video even though it’s used in crawling the web
- Improve two specific areas
  - Optical Character Recognition (OCR)
  - EXIF metadata extraction

OCR and EXIF

- Many dark web images include text as part of the image caption
  - Sometimes the text in the image is all we have to search for since an accompanying description is not provided
  - Image text can relate previously unlinkable images with features
  - Some challenges: Imagine running this at the scale of 40+Million images
- EXIF metadata
  - Allows feature relationships to be made between e.g., camera properties (model number; make; date/time; geo location; RGB space, etc.)

Tesseract

- Originally by HP
- Developed by Google
- Server side OCR engine
- Apache 2.0 Licensed
- Able to decode both LTR and RTL texts

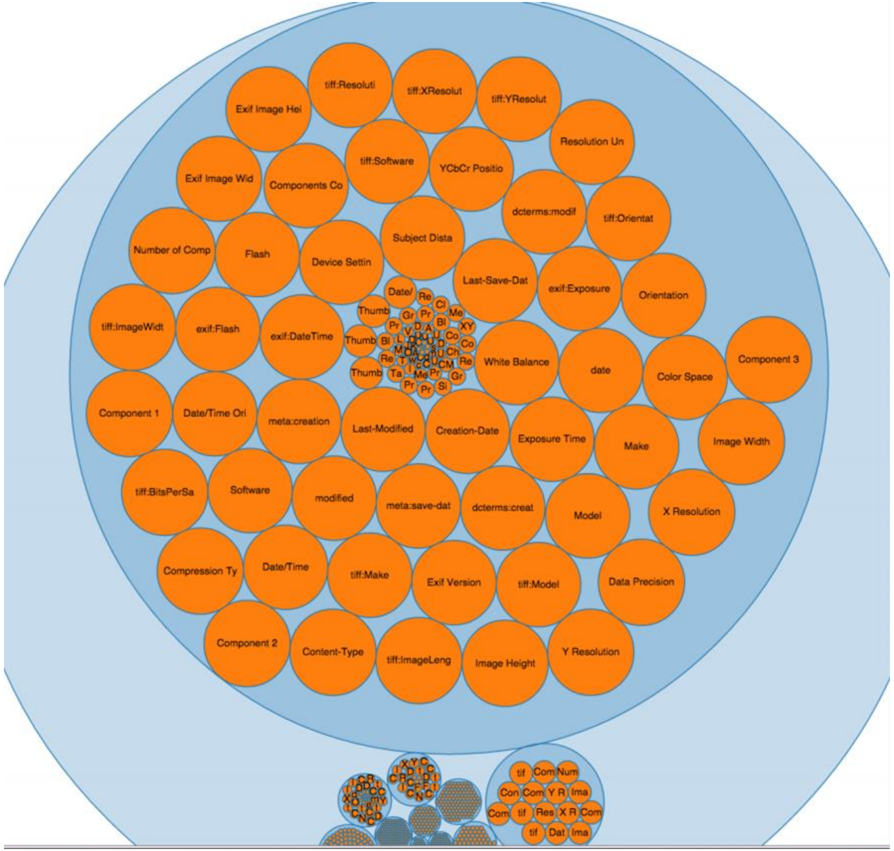
Criminals Are Pretty Dumb

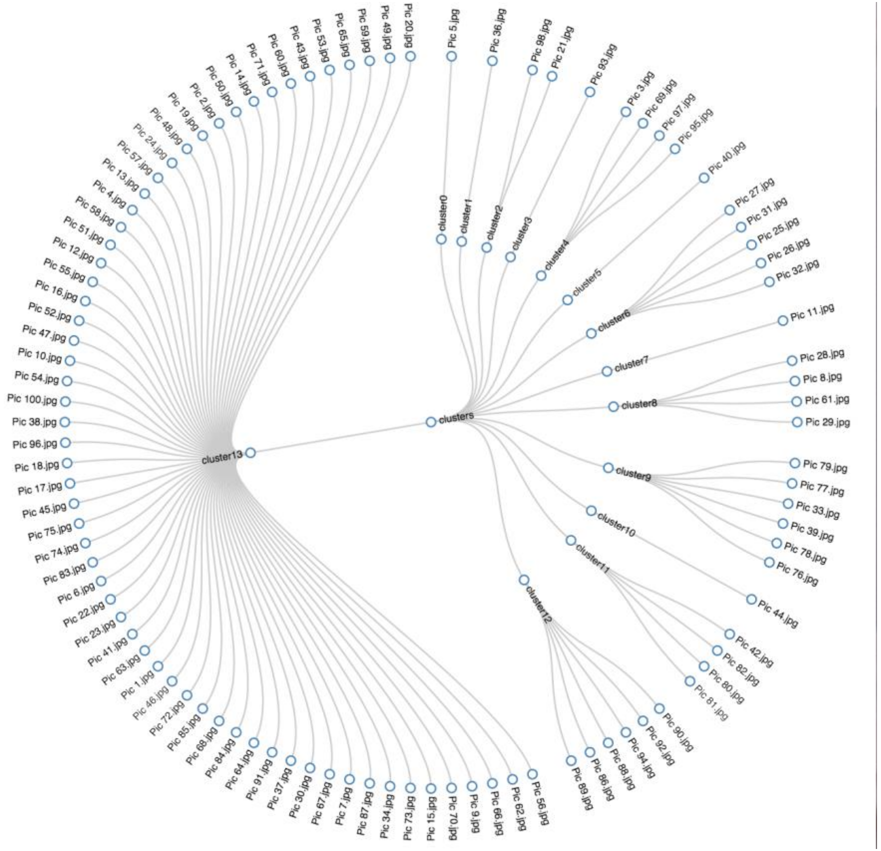
- Example EXIF metadata
    - Camera Settings; Scene Capture Type; White Balance Mode; Flash; Fnumber (Fstop); File Source; Exposure Mode; Xresolution; Yresolution; Recommended EXIF interoperability Rules, Thumbnail compression; Image Height; Image Width; Flash Output; AF Area Height; Model; Model Serial Number; Shooting Mode; Exposure Compensation.. AND MANY MORE
- These represent a “feature space” that can be used to relate images even without looking directly at the image
- Believe it or not, EXIF data remains in a lot of the images we find on the web

Tika Image Similarity

- First pass it a directory e.g., of Images
  - For each file (image) in the directory
    - Run Tika, extract EXIF features
    - Add all unique features to “golden feature set”
- Loop again
  - Use extracted EXIF metadata for file, compute size of feature set, and names, compute containment which is a “distance” of each document to the golden feature set
- Set a threshold on distance, then you have clusters

Image Similarity Visualisation





ImageCat

- OCR and EXIF metadata around images
- Can handle similarity measures
- Can allow for search of features in images based on text
- Can relate images based on EXIF properties (all taken with flash on; all taken in same geographic region, etc.)
- How do you do this at the scale of the Internet
- “Deep Web” as defined by DARPA in domain of e.g., human trafficking ~60M web pages, 40M images

ImageCat Innards

- Apache OODT – ETL, Map Reduce over long lists of files
  - Partition files into 50k chunks
  - Ingest into Solr Extracting RequestHandler
- Apache Solr / Extracting RequestHandler
- Augmented Tika + Tesseract OCR
- Tika + EXIF metadata

Other Parts To The Puzzle

- Domain Specific machine learning exercises
- Regular meetings and hackathons to drive research
- Buy in by law enforcement

How Open Source Helps!

Criminals Evolve, But Are Still Dumb

- Exif data gets left in images
- Text gets added to descriptions
- Alt Text
- They post in the same places
- The dark web might be dark but it’s still accessible

Part Of A Larger Community

- Many organisations involved in Memex
- Many volunteers spent unpaid hours working on integrations, extensions and brand new code
- Apache Foundation is core to a lot of what we do at NASA JPL

Core to everything we do….

- Open source data platforms
    - Cost effective
    - Extendable
    - Community Driven

You’ll Find Most Of Memex On The Web

- https://github.com/memex-explorer
- https://github.com/USCDataScience/sparkler
- https://github.com/apache/tika
- https://github.com/darpamemex/program-index

Thanks To

- Chris Mattmann
- Karanjeet Singh
- Thamme Gowda

Original Slides can be found here:

http://schd.ws/hosted_files/apachebigdataeu2016/e0/Sparkler%20-%20ApacheCon%20EU.pdf

http://events.linuxfoundation.org/sites/events/files/slides/ACNA15_Mattmann_Tika_Video2.pdf