The Future of Search
Fishing The Streams of Big Data
Who are Flax?

- We design, build and support open source powered search applications

- Based in Cambridge U.K., technology agnostic & independent – but open source exponents

- UK Authorized Partner of Lucidworks

- Customers in recruitment, government, e-commerce, news & media, bioinformatics, consulting, law...

[Logos of Flax and its partners]
Where are we now?

- What sort of enterprise projects do we see at Flax?
  - Migrations
  - Greenfield
  - Speculative
Where are we now?

- What sort of enterprise projects do we see at Flax?
  - Migrations
  - Greenfield
  - Speculative

- Unsurprisingly most are using open source
Where are we now?

- What sort of enterprise projects do we see at Flax?
  - Migrations
  - Greenfield
  - Speculative

- Unsurprisingly most are using open source

- ....unless they're Sharepoint
Where are we now?

- What sort of enterprise projects do we see at Flax?
  - Migrations
  - Greenfield
  - Speculative

- Unsurprisingly most are using open source

- ....unless they're Sharepoint

- So has open source “won”?
What's happening with open source?

Two main stacks
- Apache Lucene/Solr
  - Lucidworks (Fusion, SiLK)
  - Elasticsearch
  - Elastic (ELK)
What's happening with open source?

- Two main stacks
  - Apache Lucene/Solr
    - Lucidworks (Fusion, SiLK)
  - Elasticsearch
    - Elastic (ELK)

- Lots of other players (less for Elasticsearch)
What's happening with open source?

- Two main stacks
  - Apache Lucene/Solr
    - Lucidworks (Fusion, SiLK)
  - Elasticsearch
    - Elastic (ELK)

- Lots of other players (less for Elasticsearch)

- Focus on:
  - Log file analysis
  - Stability & scalability
  - Visualisation
Analytics based on search

- Log files,
  messages,
  transactions
Analytics based on search

Log files, messages, transactions
Analytics based on search

Log files, messages, transactions

Solr

elasticsearch.

www.flax.co.uk
Analytics based on search

Log files, messages, transactions

Solr

elasticsearch.

flax

www.flax.co.uk
Analytics based on search

Log files, messages, transactions

Solr

elasticsearch.

flax www.flax.co.uk
Analytics & visualisations

989 crimes

Has 161 incidents of *Possession of weapons* vs expected 2

<table>
<thead>
<tr>
<th>Month</th>
<th>Num Incidents</th>
</tr>
</thead>
<tbody>
<tr>
<td>201301</td>
<td>72</td>
</tr>
<tr>
<td>201303</td>
<td>92</td>
</tr>
<tr>
<td>201304</td>
<td>89</td>
</tr>
<tr>
<td>201305</td>
<td>120</td>
</tr>
<tr>
<td>201306</td>
<td>115</td>
</tr>
<tr>
<td>201307</td>
<td>85</td>
</tr>
<tr>
<td>201308</td>
<td>114</td>
</tr>
<tr>
<td>201309</td>
<td>113</td>
</tr>
</tbody>
</table>
Analytics & visualisations
Trends

- Big Data
- Internet of Things
- Cloud
- Semantic Search
- Mobile / Wearables
Trends

- Big Data
- Internet of Things
- Cloud
- Semantic Search
- Mobile / Wearables
Trends

- Big Data
- Internet of Things
- Cloud
- Semantic Search
- Mobile / Wearables
Charlie's All Purpose Big Data Graph

Data

Time
Charlie's All Purpose Big Data Graph

Data

Time

www.flax.co.uk
Analyst quotes

“The IoT has the potential to connect 10X as many (28 billion) “things” to the Internet by 2020, ranging from bracelets to cars.”

*Goldman Sachs*¹
Analyst quotes

“The IoT has the potential to connect 10X as many (28 billion) “things” to the Internet by 2020, ranging from bracelets to cars.” Goldman Sachs¹

“IoT threatens to generate massive amounts of input data from sources that are globally distributed. Transferring the entirety of that data to a single location for processing will not be technically and economically viable” Gartner²
“The IoT has the potential to connect 10X as many (28 billion) “things” to the Internet by 2020, ranging from bracelets to cars.”

Goldman Sachs¹

“IoT threatens to generate massive amounts of input data from sources that are globally distributed. Transferring the entirety of that data to a single location for processing will not be technically and economically viable”

Gartner²

“Streaming analytics is anything but a sleepy, rearview mirror analysis of data. No, it is about knowing and acting on what’s happening in your business at this very moment — now.”

Forrester³
Some predictions

- It will become increasingly difficult, if not impossible, to store data for later processing
Some predictions

- It will become increasingly difficult, **if not impossible**, to store data for later processing

- It must therefore be processed as it appears, in **real-time** (Real Time Analytics)
Some predictions

- It will become increasingly difficult, **if not impossible**, to store data for later processing.

- It must therefore be processed as it appears, in **real-time** (Real Time Analytics).

- Much of this data will be **unstructured**, noisy and badly formatted.
Some predictions

- It will become increasingly difficult, if not impossible, to store data for later processing.
- It must therefore be processed as it appears, in real-time (Real Time Analytics).
- Much of this data will be unstructured, noisy and badly formatted.
- There are many exciting applications - if this is done right.
New ideas

- Unified Log

2004

grep, etc.

2014

Unified Logging Layer
New ideas

Unified Log

“The Log: What every software engineer should know about real-time data's unifying abstraction” – Jay Kreps, LinkedIn (now Confluent)
More new ideas

• Everything is a stream

“...think of a database as an always-growing collection of immutable facts. You can query it at some point in time — but that’s still old, imperative style thinking. A more fruitful approach is to take the streams of facts as they come in, and functionally process them in real-time”.- Martin Kleppman, LinkedIn
New technologies

- Streaming data platforms

Amazon Kinesis
New technologies

- Streaming data platforms

Amazon Kinesis

Spark Streaming
New technologies

- Streaming data platforms

Amazon Kinesis

Spark Streaming

Apache Storm

Distributed • Resilient • Real-time

www.flax.co.uk
New technologies

- Streaming data platforms

![Amazon Kinesis](image1)

![Spark Streaming](image2)

![Apache Storm](image3)

![Apache Kafka](image4)

A high-throughput distributed messaging system.

![samza](image5)

![flax](image6)

[www.flax.co.uk](http://www.flax.co.uk)
No elephant in the room?

- Hadoop is for batch processing, not stream processing
- ...although Storm etc. can run on HDFS
So what's this got to do with Search?

How can you currently process data in a stream?
- SQL like
- Regular Expressions
- Machine Learning
So what's this got to do with Search?

- How can you currently process data in a stream?
  - SQL like
  - Regular Expressions
  - Machine Learning

- Why not use full-text search?
  - Easier to create queries
  - Great with unstructured data
  - Handles noisy data
So what's this got to do with Search?

- How can you currently process data in a stream?
  - SQL like
  - Regular Expressions
  - Machine Learning

- Why not use full-text search?
  - Easier to create queries
  - Great with unstructured data
  - Handles noisy data

- But you can do search already!
  - But only near real time
What's a stream?

- “..an append-only, totally ordered sequence of records (also called events or messages).”

Martin Kleppman, LinkedIn ⁹
How do we search it?
How do we search it?

Oops!

Result

append

Solr
Here's something we're doing already

- Search for Media Monitoring
  - Many complex stored profiles (searches)
  - High volume of news stories every day
Here's something we're doing already

- Search for Media Monitoring
  - Many complex stored profiles (searches)
  - High volume of news stories every day

- The solution:
  - Build an index of the stored profiles
  - Turn each news story into a query
  - Search the stored profiles to find which might match
  - Use these candidates to search the news story
Here's something we're doing already

- Search for Media Monitoring
  - Many complex stored profiles (searches)
  - High volume of news stories every day

- The solution:
  - Build an index of the stored profiles
  - Turn each news story into a query
  - Search the stored profiles to find which *might* match
  - Use these candidates to search the news story

- We call it 'search turned upside down'
  - But it's effectively *searching a stream*
Like this...

(((";MOBILE PHONE*"; OR ";PHONE MAST*"; OR ";HANDSET*"; OR ";CELL* PHONE*"; OR ";3G"; OR ";GPRS"; OR ";G.P.R.S"; OR ";GENERAL !RADIO PACKET SERVICE*"; OR ";GSM"; OR ";G.S.M"; OR ";GLOBAL SYSTEM FOR !MOBILE COMM*"; OR ";HSDPA"; OR ";H.S.D.P.A"; OR ";HIGH SPEED DOWNLINK !PACKET ACCESS"; OR ";HSUPA"; OR ";H.S.U.P.A"; OR ";HIGH SPEED !UPLINK !PACKET ACCESS"; OR ";UMTS"; OR ";U.M.T.S"; OR ";MVNO"; OR ";M.V.N.O"; OR ";SMS"; OR ";SHORT MESSAGE !SERVICE*"; OR ";MMS"; OR ";MULTIMEDIA MESSAGE !SERVICE*"; OR ";MOBILES"; OR ";MOBILE"; OR ";CELLPHONE*"; OR ";TELECOM"; OR ";LANDLINE*"; OR ";PHONE*"; OR ";!TELEPHONE*"; OR ";TELEFONICA"; OR ";BT"; OR ";MOBILE USER*"; OR ";TEXT MESSAG*"; OR ";SMARTPHONE"; OR ";MOBILE USER*"; OR ";CABLE & WIRELESS"; OR ";CABLE AND !WIRELESS";) W/48 (((";PROFIT*"; OR ";LOSS*"; OR ";BAN"; OR ";BANNED"; OR ";PREMIUM RATE*"; OR ";FINANC*"; OR ";REFINANC*"; OR ";OFFICE OF FAIR TRADING"; OR ";MERGER*"; OR ";ACQUISIT*"; OR ";ACQUIR*"; OR ";TAKEOVER*"; OR ";BUYOUT*"; OR ";BUY-OUT*"; OR ";NEW PRODUCT*"; OR ";INVEST*"; OR ";SHARES"; OR ";MARKET*"; OR ";ACCOUNT*"; OR ";!MONEY"; OR ";CASH*"; OR ";SECURIT*"; OR ";ENTERPRISE*"; OR ";BUSINESS*"; OR ";PRICE*"; OR ";JOINT*"; OR ";NEW VENTURE*"; OR ";FULL YEAR*"; OR ";REGULAT*"; OR ";DIRECTIVE*"; OR ";LAW*"; OR ";LAWS"; OR ";LEGISLAT*"; OR ";GREEN PAPER"; OR ";WHITE PAPER*"; OR ";MEDIAWATCH"; OR ";MORAL*"; OR ";ETHIC*"; OR ";ADVERT*"; OR ";AD"; OR ";ADS"; OR ";MARKETING"; OR ";COMPLAIN*"; OR ";MIS-SOLD*"; OR ";MIS-SELL*"; OR ";SPONSOR"; OR ";COSTCUT*"; OR ";CUT* COST*"; OR ";FIBRE OPTIC*"; OR ";TAX"; OR ";TAXES"; OR ";EXPAND*"; OR ";EXPANSION"; OR ";EMPLOY*"; OR ";STAFF"; OR ";WORKER*"; OR ";SPOKESM?N*"; OR ";DEBUT"; OR ";BRAND*"; OR ";DIRECTOR*";) OR (";FAIR"; OR ";UNFAIR"; OR ";UNFAIR"; OR ";!PENALISE*"; W/12 ";!CHARG*"; OR ";!TARIFF*"; OR ";PRICE PLAN*"; OR ";GLOBAL*"))) AND NOT (";EXPRESS OFFER"; OR ";TIMES OFFER"; OR ";READER OFFER"; OR ";CALLS COST*";) W/6 (";FROM A LANDLINE"; OR ";FROM LANDLINE*"; OR ";BT LANDLINE*";)))

…and that's an easy one!
Turning search upside down..

- Stored Queries
- 1 million profiles
- Some 250k long
- Complex rules

- 1. Pre
- Query Subset
- ~200

- Doc
- Docs

- 1 million news items a day
Turning search upside down..

1 million profiles
Some 250k long
Complex rules

1 million news items a day

1. Pre

Query Subset

~200

2. Search

“this news item is of interest to my client”
Turning search upside down..

1 million profiles
Some 250k long
Complex rules

1. Pre

Query Subset

~200

1 million news
items a day

Docs

Doc

Result

Stream

Stream
Searching streaming data at scale

- Flax solution scales to 1m queries over 1m items/day
Searching streaming data at scale

- Flax solution scales to 1m queries over 1m items/day

- We need to be faster:
  - Network monitoring – up to 1m items/second
  - Restaurant reservations/reviews – 840m messages/day
  - IoT
Searching streaming data at scale

- Flax solution scales to 1m queries over 1m items/day

- We need to be faster:
  - Network monitoring – up to 1m items/second
  - Restaurant reservations/reviews – 840m messages/day
  - IoT

- Here's a prototype:
  - https://github.com/romseygeek/samza-luwak
  - “... you’ll be able to perform full-text search on streams at arbitrary scale, simply by adding new partitions and adding more machines to the cluster.” Martin Kleppman, LinkedIn

flax www.flax.co.uk
Real-time scalable search for streams

Which queries match?

Documents

Queries

In-memory 1-doc index

Index of queries

Parse

Matches

A B C D E F G
What could you do with this?

- Build queries using a standard search box, facets etc.
What could you do with this?

- Build queries using a standard search box, facets etc.
- Set them to run on a stream of data
What could you do with this?

- Build queries using a standard search box, facets etc.
- Set them to run on a stream of data
- Matches appear as another stream
What could you do with this?

- Build queries using a standard search box, facets etc.
- Set them to run on a stream of data
- Matches appear as another stream
- Use cases
  - Monitor a network
  - Watch social media
  - Check IoT for faults, errors, trends
  - Look for patterns in customer interactions
What could you do with this?

- Build queries using a standard search box, facets etc.
- Set them to run on a stream of data
- Matches appear as another stream

Use cases
- Monitor a network
- Watch social media
- Check IoT for faults, errors, trends
- Look for patterns in customer interactions

- Combine with analytics & visualisations
In conclusion....

- Big Data....is about to get a lot bigger
In conclusion....

- Big Data....is about to get a lot bigger
- Search can help!
Thankyou!

Any questions?

charlie@flax.co.uk
www.flax.co.uk/blog
+44 (0) 8700 118334
Twitter: @FlaxSearch
References

1. Internet of Things - HorizonWatch 2015 Trend Report, Bill Chamberlin, IBM
2. Gartner Says the Internet of Things Will Transform the Data Center http://www.gartner.com/newsroom/id/2684915
4. Unified Logging Layer: Turning Data into Action, Kiyoto Tamura, Fluentd
5.Unified Log Processing, Alexander Dean
6. The Log: What every software engineer should know about real-time data's unifying abstraction, Jay Kreps, LinkedIn/Confluent
7. Turning the database inside-out with Apache Samza, Martin Kleppman
8. Designing Data-Intensive Applications, Martin Kleppmann
9. Realtime Full-text Search with Luwak and Samza, Martin Kleppmann