

The Future of Search

Fishing The Streams of Big Data

Charlie Hull - Managing Director
9th June 2015
IKO Conference, Singapore

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



www.flax.co.uk



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



www.flax.co.uk



Where are we now?

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



www.flax.co.uk



Where are we now?

- ♦ What sort of enterprise projects do we see at Flax?
 - Migrations
 - Greenfield
 - Speculative
- ♦ Unsurprisingly most are using open source



www.flax.co.uk



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



www.flax.co.uk



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”?



www.flax.co.uk



What's happening with open source?

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

Solr 

 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)

Solr 

 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

Solr 

 elasticsearch.



FLAX

www.flax.co.uk



Analytics based on search

- Log files,
messages,
transactions

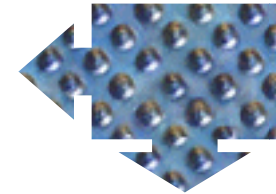


www.flax.co.uk

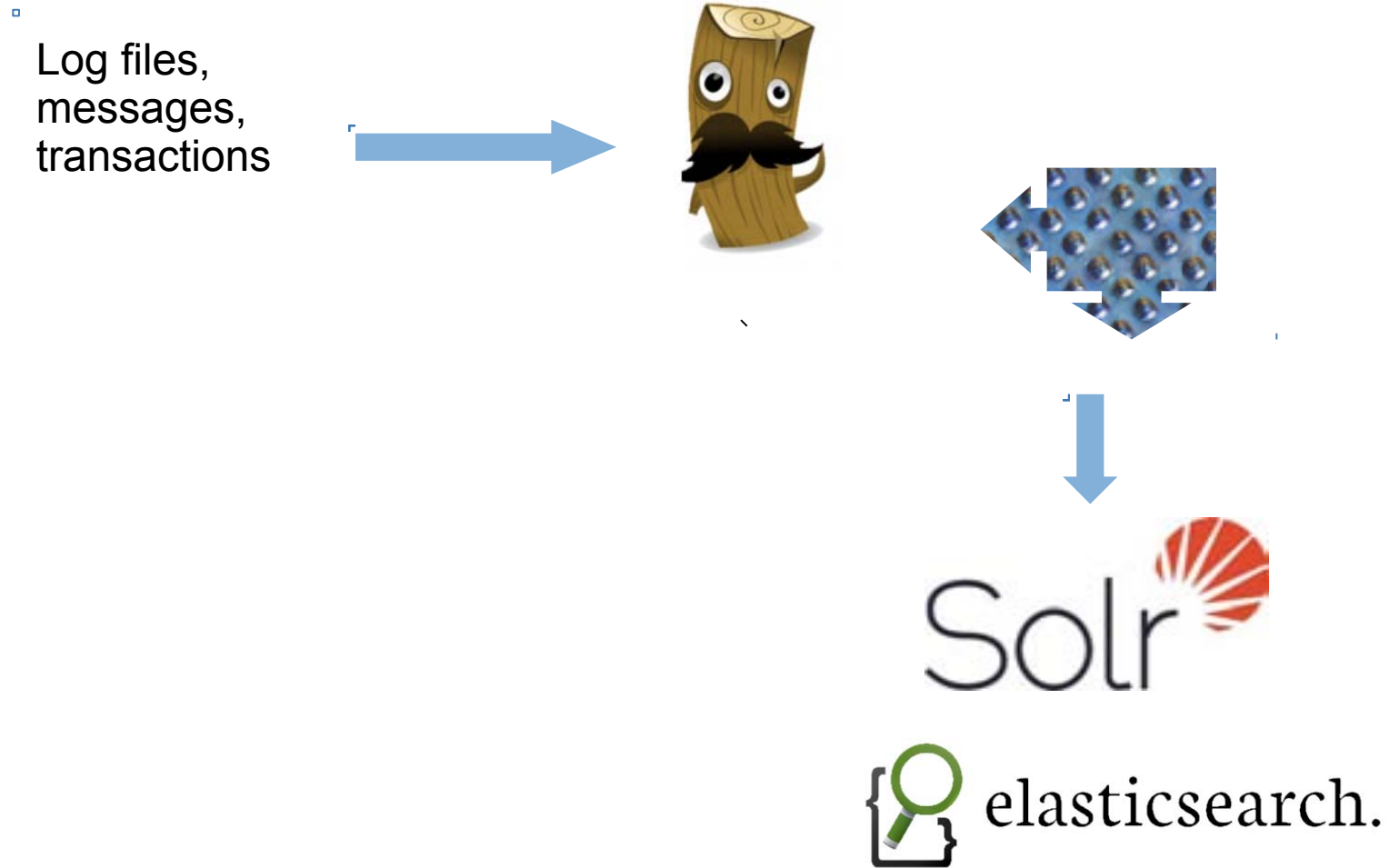


Analytics based on search

- Log files,
messages,
transactions

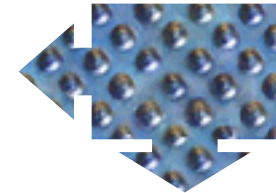


Analytics based on search



Analytics based on search

Log files,
messages,
transactions



elasticsearch.

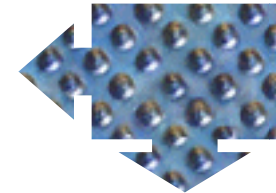


www.flax.co.uk



Analytics based on search

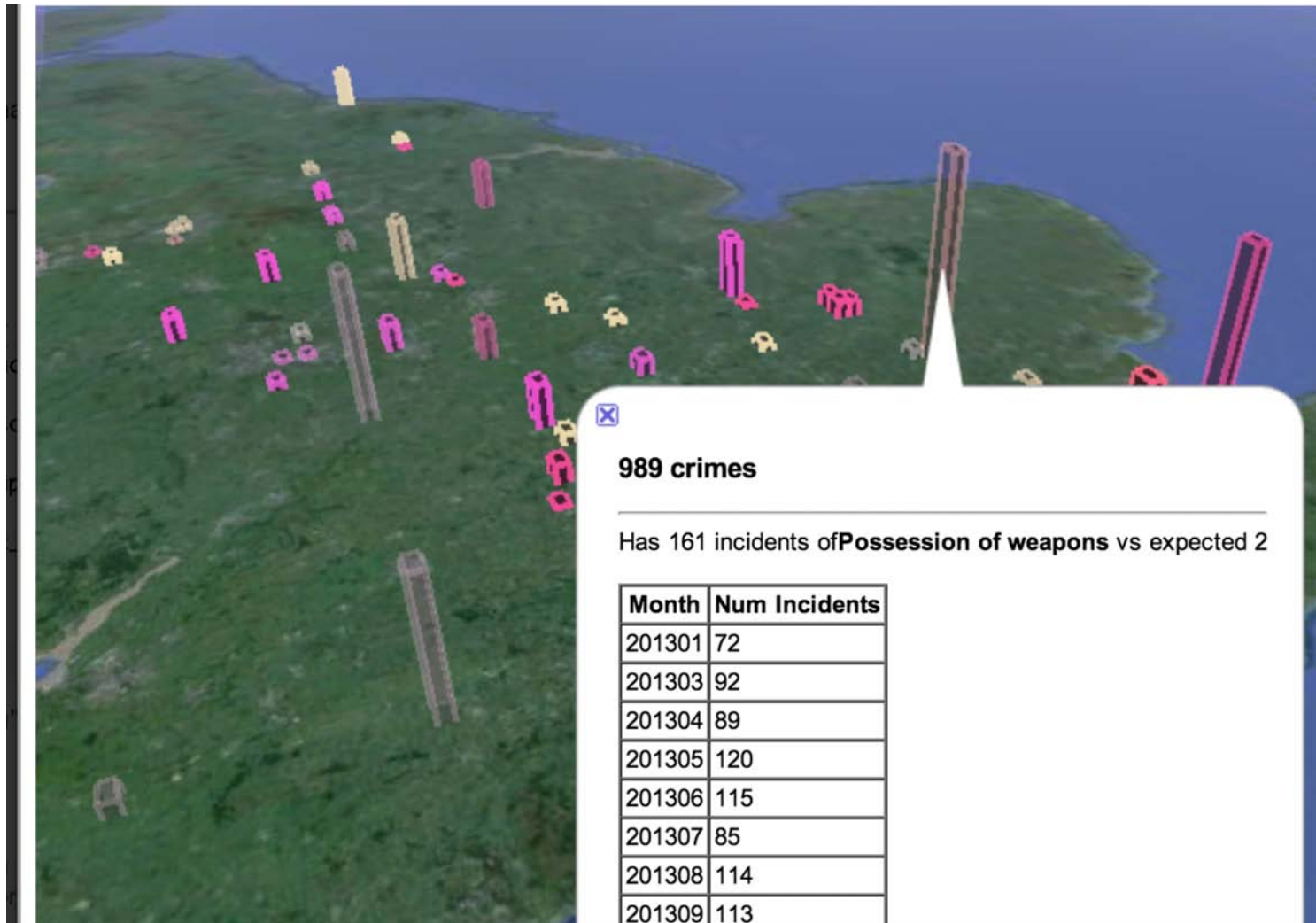
Log files,
messages,
transactions



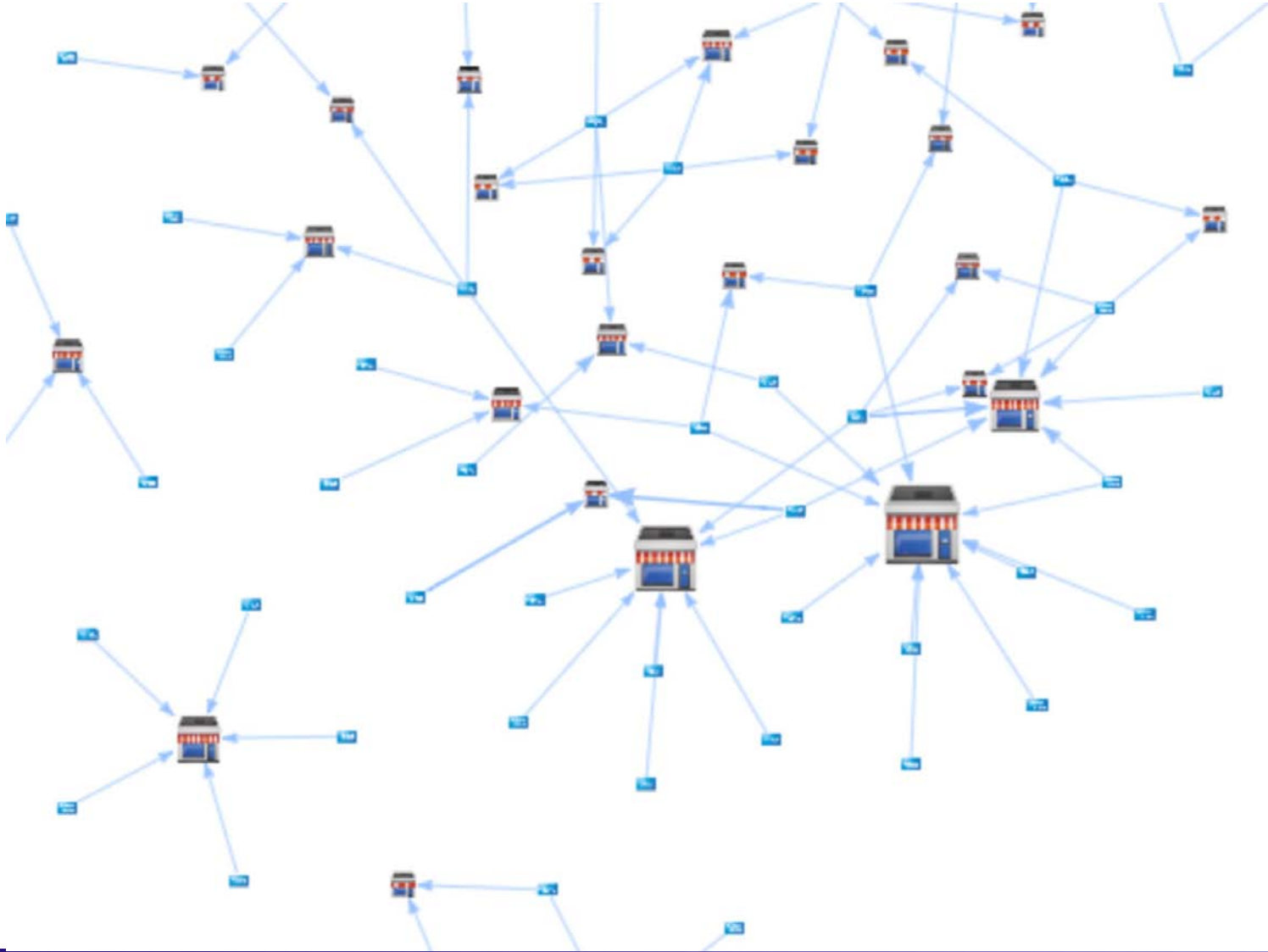
www.flax.co.uk



Analytics & visualisations



Analytics & visualisations



www.flax.co.uk



Trends

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



www.flax.co.uk



Trends

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



www.flax.co.uk



Trends

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



www.flax.co.uk



Charlie's All Purpose Big Data Graph

,

Data

.

Time



www.flax.co.uk



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



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

“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” *Gartner*²



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

“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” *Gartner*²

“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.”
*Forrester*³



www.flax.co.uk



Some predictions

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



www.flax.co.uk



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)



www.flax.co.uk



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



www.flax.co.uk



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**

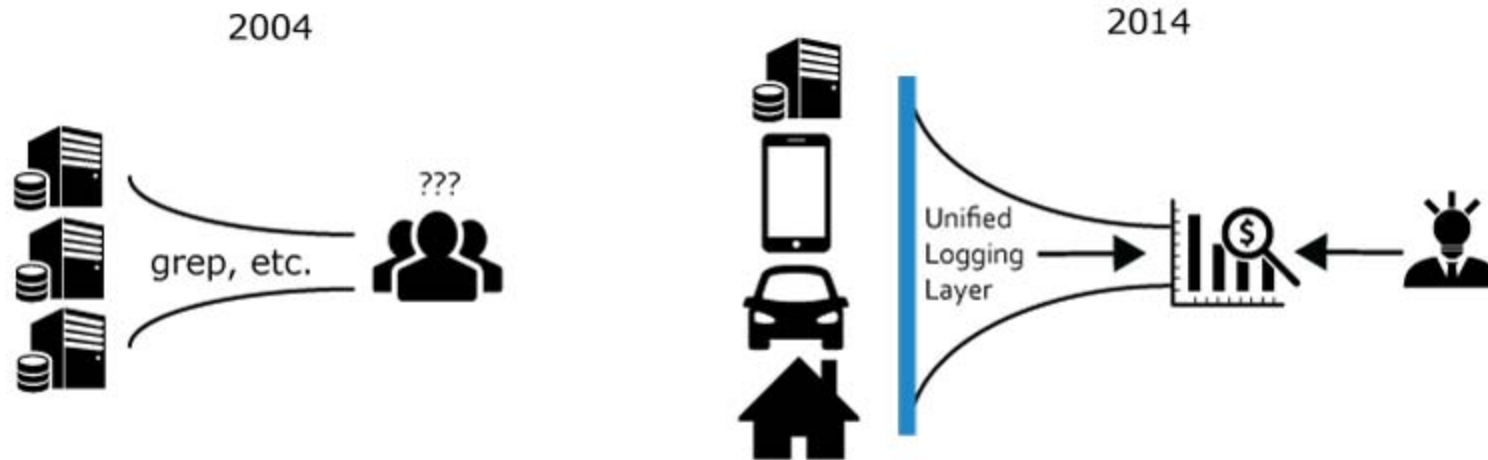


www.flax.co.uk



New ideas

- ◆ Unified Log

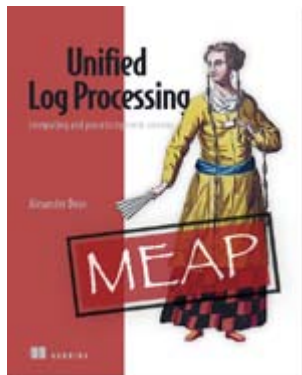
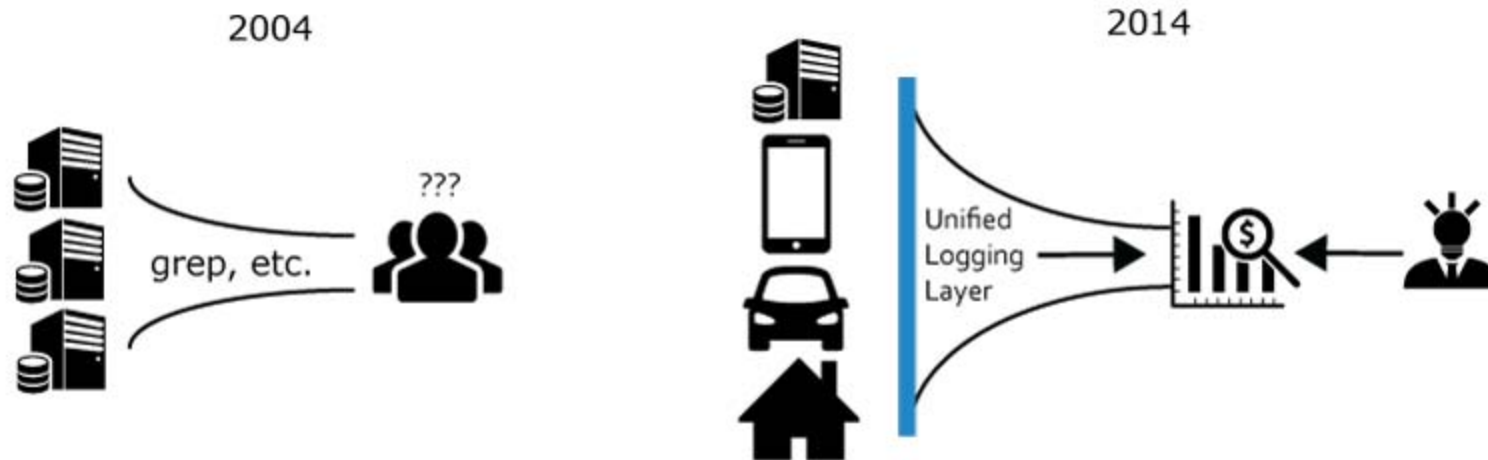


4



New ideas

◆ Unified Log



“The Log: What every software engineer should know about real-time data's unifying abstraction” – *Jay Kreps, LinkedIn (now Confluent)*⁶



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* ⁷

8



www.flax.co.uk



New technologies

- ◆ Streaming data platforms



Amazon Kinesis



New technologies

- ◆ Streaming data platforms



Amazon Kinesis



New technologies

- ♦ Streaming data platforms



Amazon Kinesis



New technologies

- ▶ Streaming data platforms



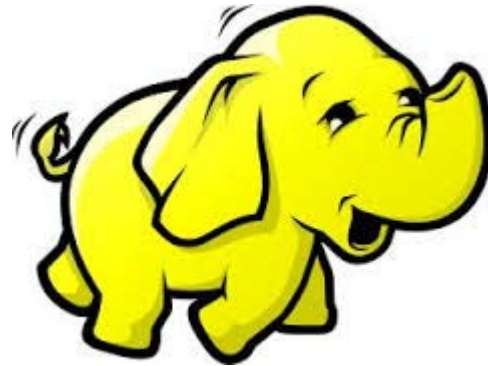
Amazon Kinesis



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



www.flax.co.uk



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



www.flax.co.uk



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



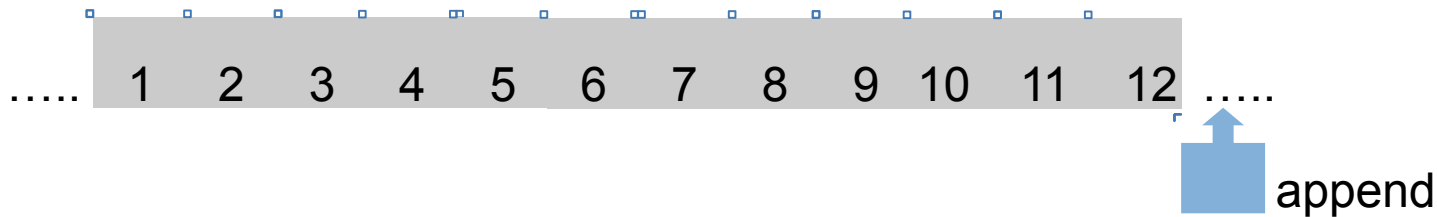
www.flax.co.uk



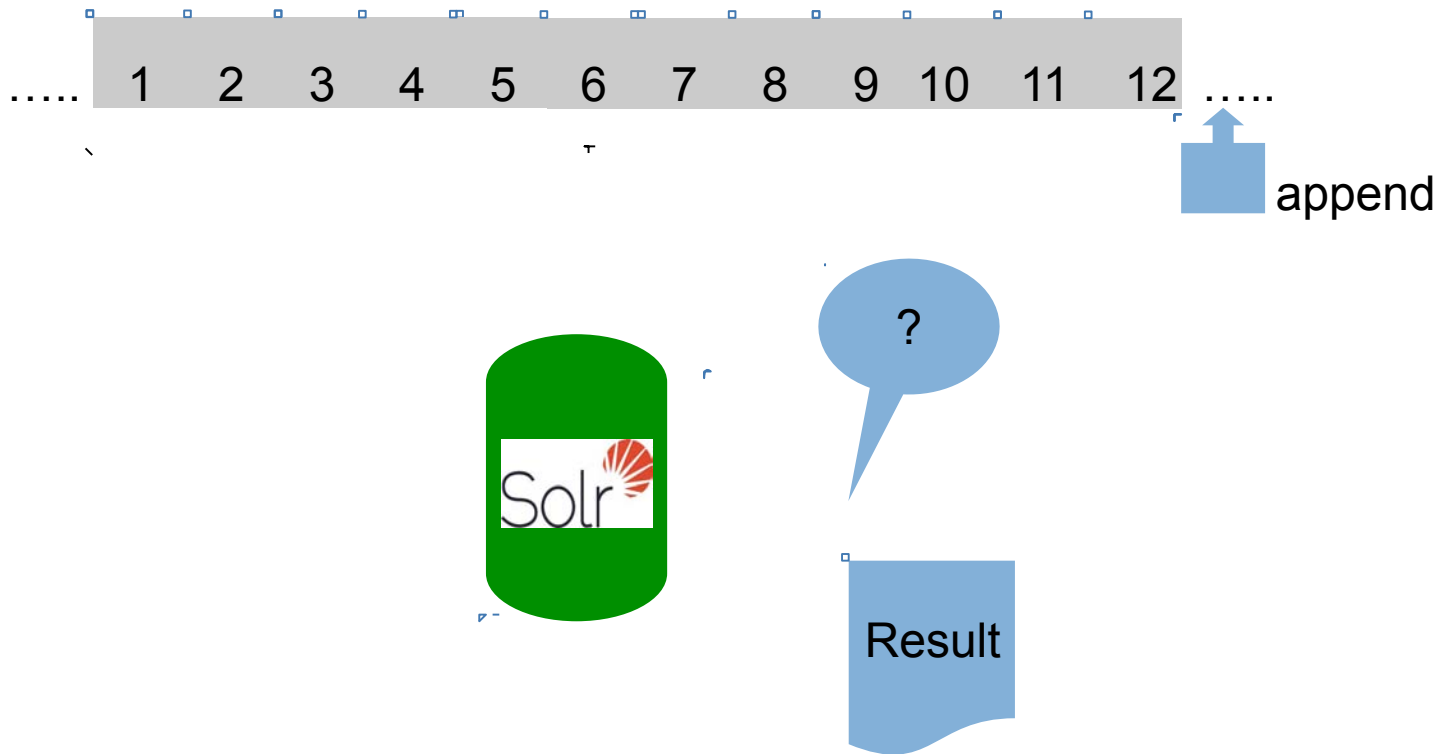
What's a stream?

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

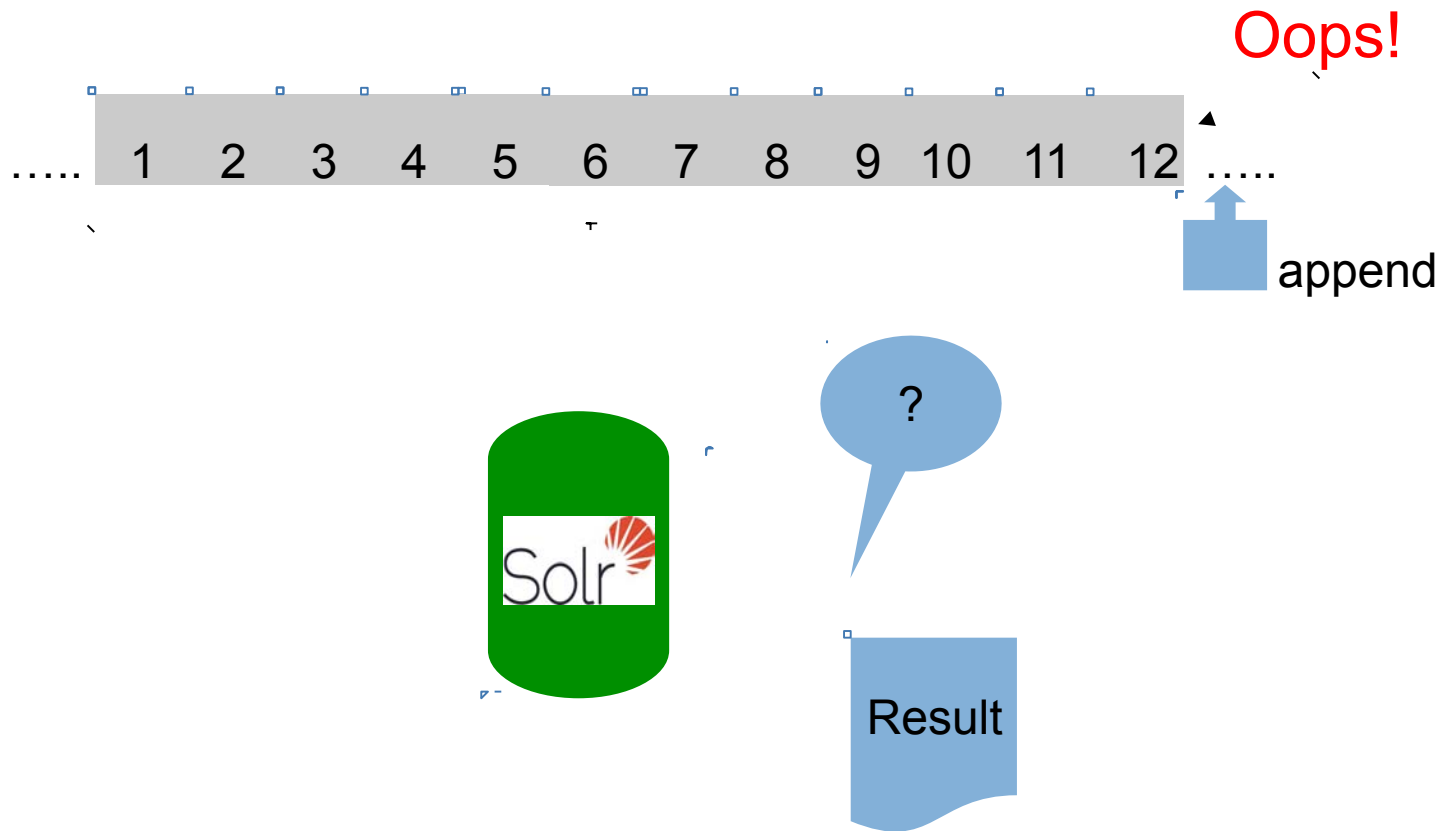
Martin Kleppman, LinkedIn ⁹



How do we search it?



How do we search it?



Here's something we're doing already

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



www.flax.co.uk



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



www.flax.co.uk



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*



www.flax.co.uk



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 ";!CELLPHONE*"; OR ";!
TELECOM*"; OR ";!LANDLINE*"; OR ";!TELEPHONE*"; OR ";PHONE*"; OR ";!TELEKOM*"; OR ";TELCO*"; OR ";VODAFONE"; OR
";T-MOBILE"; OR ";TMOBILE"; OR ";!TELEFONICA"; OR ";BT"; OR ";!MOBILE USER*"; OR ";TEXT MESSAG*"; OR
";SMARTPHONE*"; OR ";!VIRGIN !MEDIA*"; 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 ";!ENTERPRIS*"; OR ";!BUSINESS*"; OR ";PRICE*"; OR ";JOINT*"; OR ";NEW VENTURE*"; OR
";PRICING"; OR ";COST*"; OR ";CHAIRM?N"; OR ";APPOINT*"; OR ";!EXECUTIVE"; OR ";SALE*"; OR ";SELL*"; 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 ";COST CUT*"; OR ";CUT* COST*";
OR ";FIBRE OPTIC*"; OR ";TAX"; OR ";TAXES"; OR ";TAXED"; OR ";EXPAND*"; OR ";!EXPANSION"; OR ";EMPLOY*"; OR
";STAFF"; OR ";WORKER*"; OR ";SPOKESM?N"; OR ";DEBUT"; OR ";BRAND*"; OR ";DIRECTOR*";) OR ((";FAIR"; OR
";UNFAIR"; OR ";%UNSCRUPULOUS"; OR ";NOT FAIR"; OR ";UNJUST*"; 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*";)))))
```

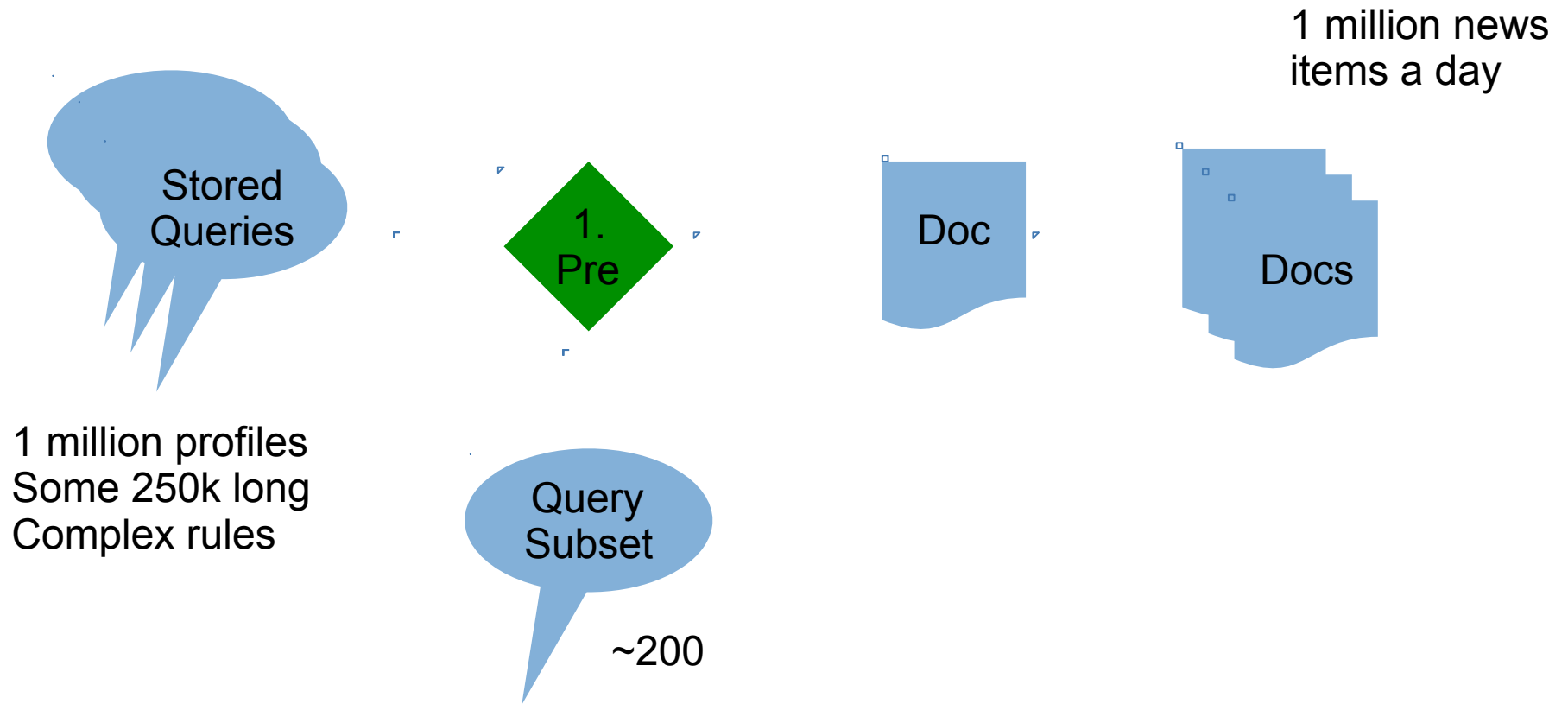
...and that's an easy one!



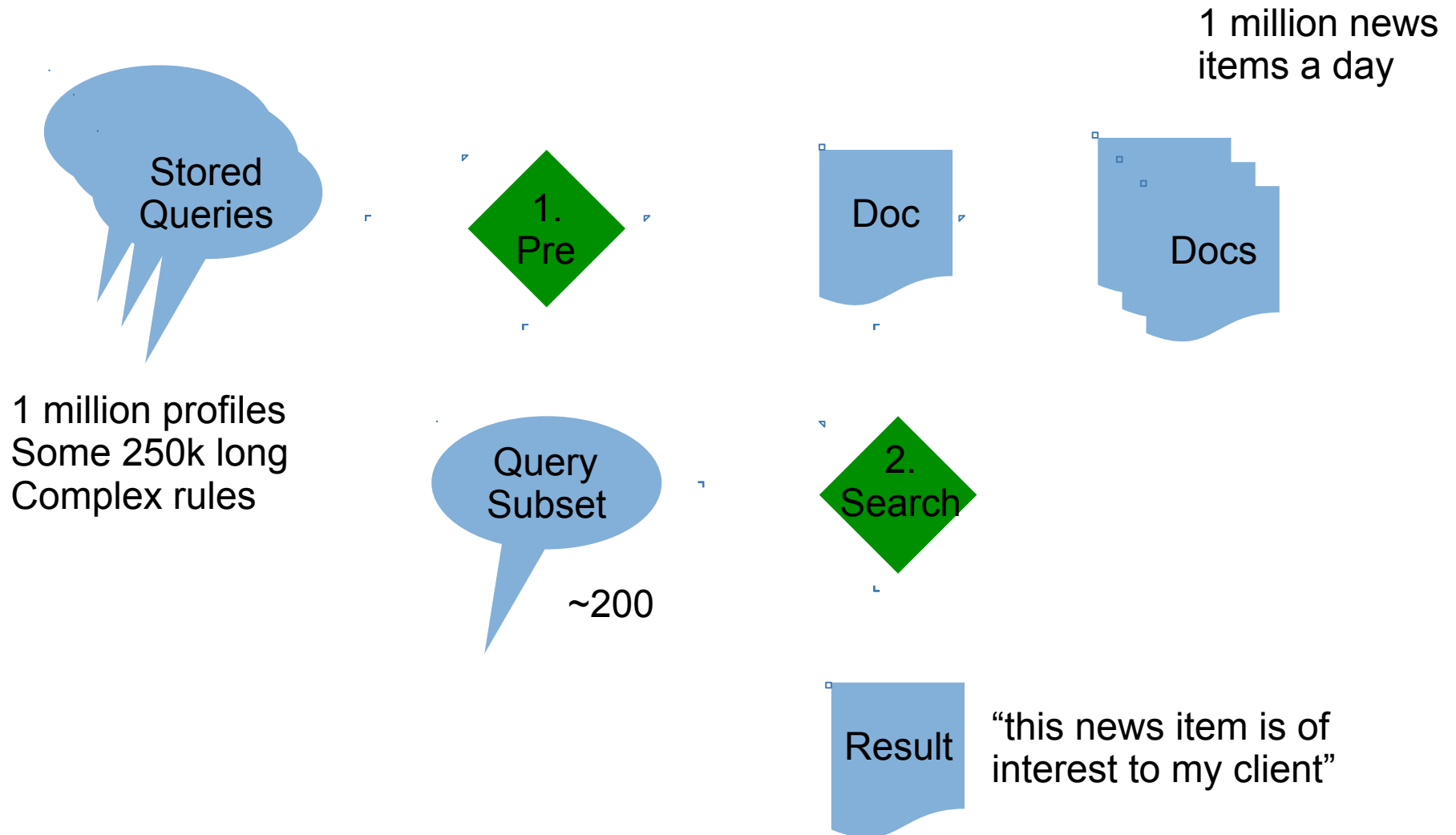
www.flax.co.uk



Turning search upside down..

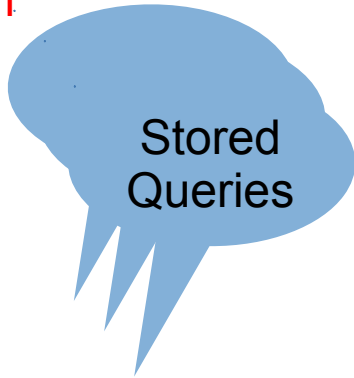


Turning search upside down..

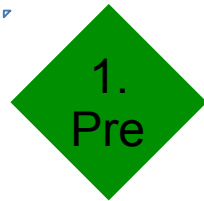


Turning search upside down..

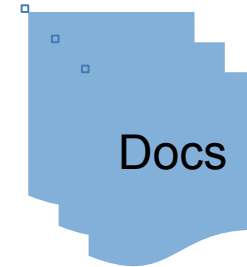
Stream



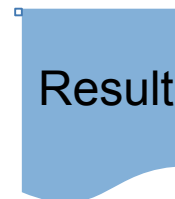
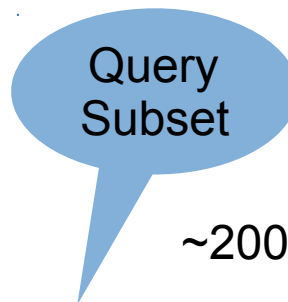
1 million profiles
Some 250k long
Complex rules



1 million news
items a day



Stream



Stream



www.flax.co.uk



Searching streaming data at scale

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



www.flax.co.uk



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



www.flax.co.uk



Searching streaming data at scale

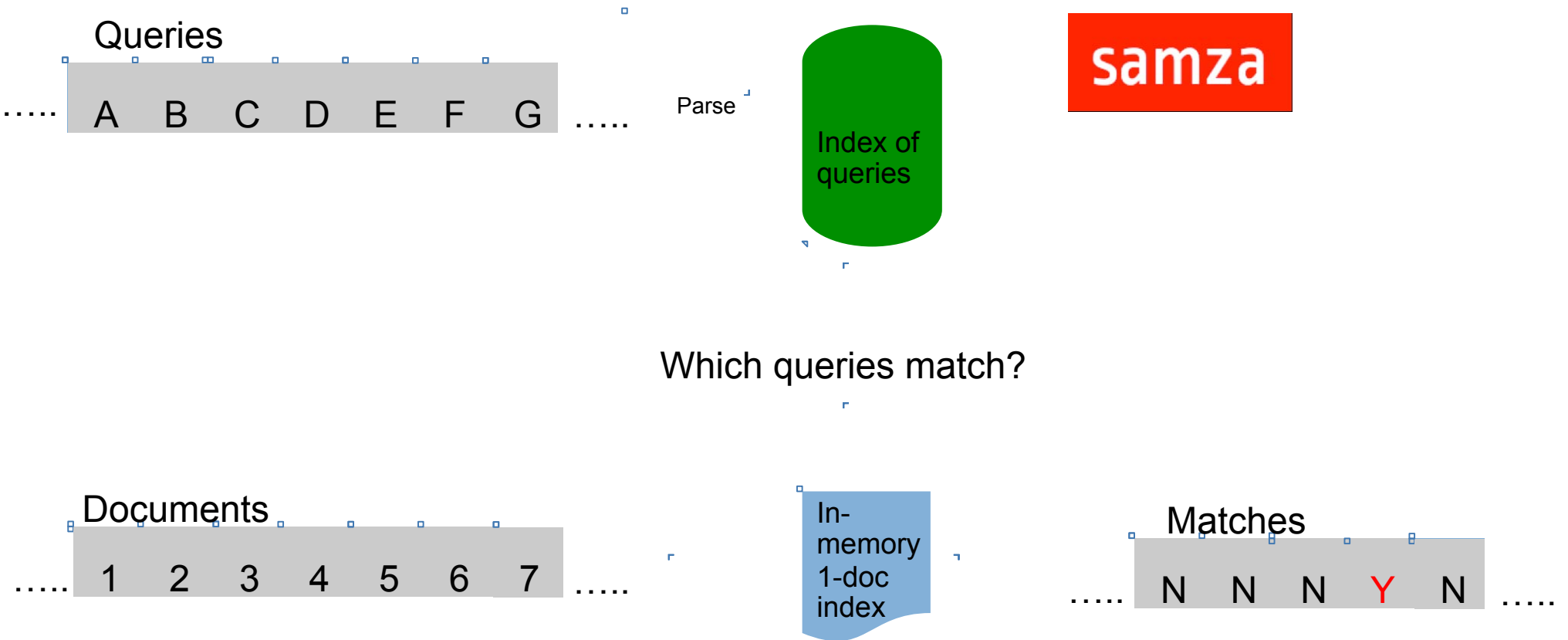
- ◆ Flax solution scales to 1m queries over 1m items/day
- ◆ We need to be faster:
 - Network monitoring – up to 1m items/second
 - Restaurant reservations/reviews – 840m messages/day
 - IoT
- ◆ Here's a prototype:
 - <https://github.com/romseygeek/samza-luwak>
 - “... you’ll be able to perform full-text search on streams at arbitrary scale, simply by adding new partitions and adding more machines to the cluster.” *Martin Kleppman, LinkedIn* ⁹



www.flax.co.uk



Real-time scalable search for streams



What could you do with this?

- ◆ Build queries using a standard search box, facets etc.



www.flax.co.uk



What could you do with this?

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



www.flax.co.uk



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



www.flax.co.uk



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



www.flax.co.uk



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



www.flax.co.uk



In conclusion....

- ◆ Big Data....is about to get a lot bigger



www.flax.co.uk



In conclusion....

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



www.flax.co.uk



Thankyou!

Any questions?

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



www.flax.co.uk



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>
3. Big Data Forrester Wave 2014
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



www.flax.co.uk

