



Infini.e: An Open Analytic Platform  
Driven by MongoDB & Hadoop

# Agenda

---

- Who we are
- What Infnit.e is
- Architecture
  - Use of Open Source
  - Elasticsearch / MongoDB / Hadoop combo
  - Focus on MongoDB
  - Focus on Hadoop
- Demo
- Questions



# Who we are

---

## IKANOW (ikanow.com)

- Our vision is to enable agile intelligence through open analytics
- Our engineering vision is to use the best OSS technologies to build a document analysis platform that will enable this and then Open Source it back to the community
  - <https://github.com/IKANOW/Infinit.e>
  - <http://bit.ly/ikanow-oss>



# What Informat.e is

---

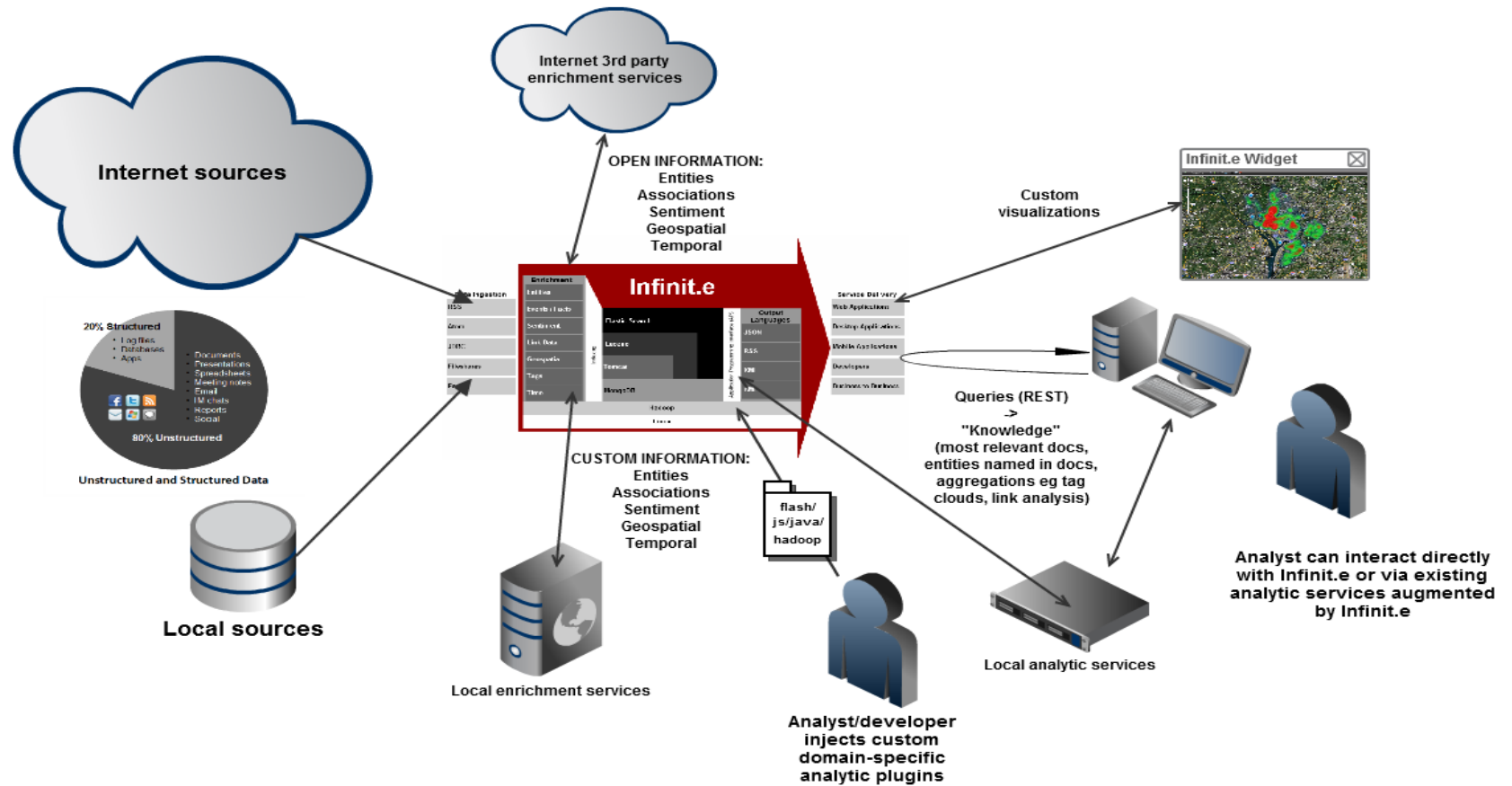
Informat.e is a scalable framework for:

- Collecting,
- Storing,
- Enriching,
- Retrieving,
- Analyzing, and
- Visualizing

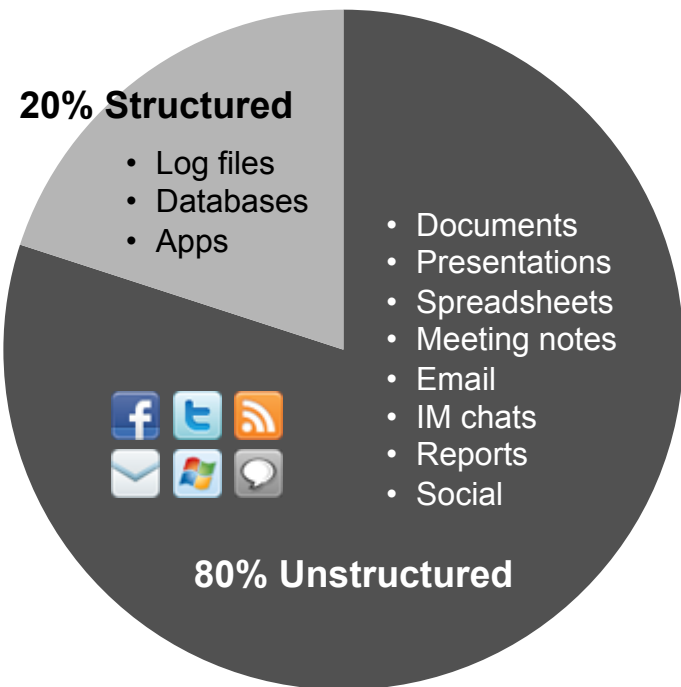
Unstructured documents and structured records



# What Infinitt.e is - Overview



# What Informat.e is - Documents



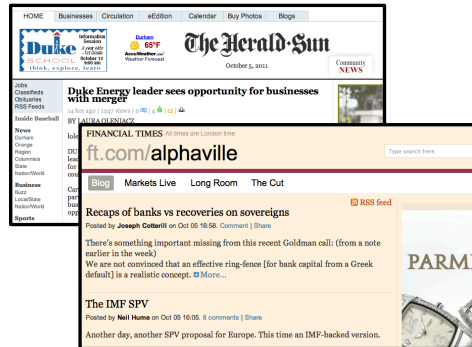
**Unstructured and Structured Data**



- Entities
- Events
- Facts
- Sentiment
- Geospatial
- Temporal
- Themes



# What Informat.e is - Documents



Duke and Progress announced **merger plans in January 2012**

Bernanke, 57 said in his testimony **price increases** "have begun to moderate" after a jump in **oil costs earlier this year**

**Who**  
people, organizations, facilities, company



Tablet ownership levels hit 18% in **China**, the **UK** and **US** versus 3% in **November 2010**

**What**  
events, summaries, facts, themes

**When**  
past, present, future dates

ORACLE

<?xml?>



Microsoft  
SQL Server

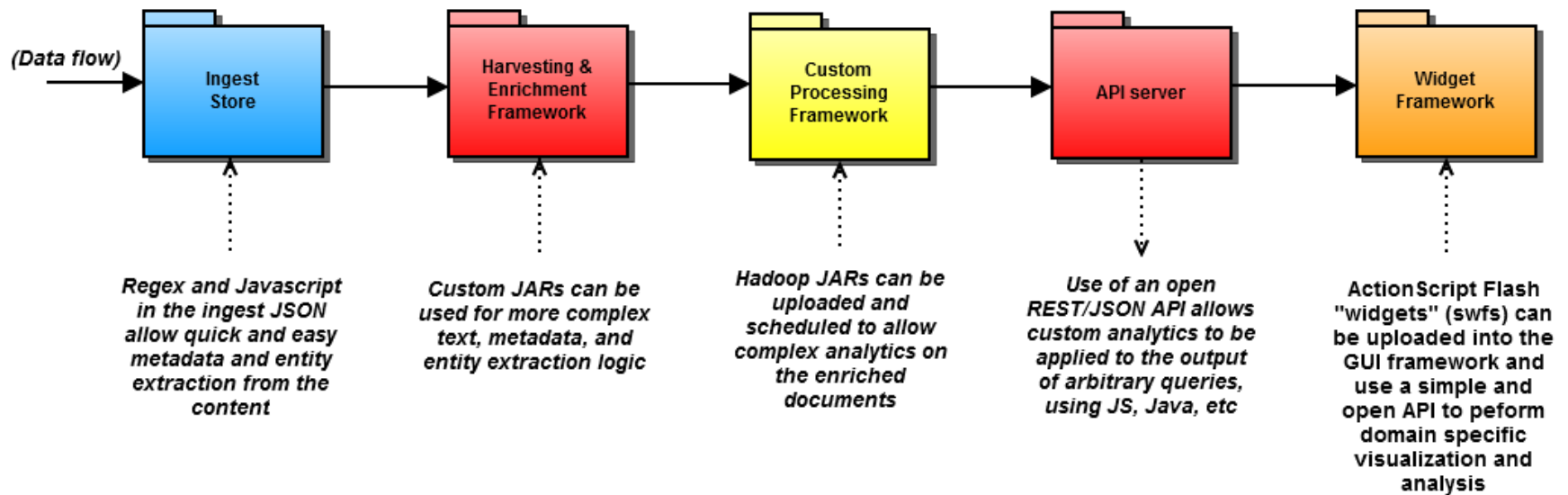
IKANOW

```
<Incident>
  <uid>20101043423</uid>
  <subject>1 person killed in armed attack by
  suspected Boko Haram in Maiduguri, Borno,
Nigeria</subject>
  <multipleDays>No</multipleDays>
  <eventDate>06/04/2011</eventDate>
</Incident>
```

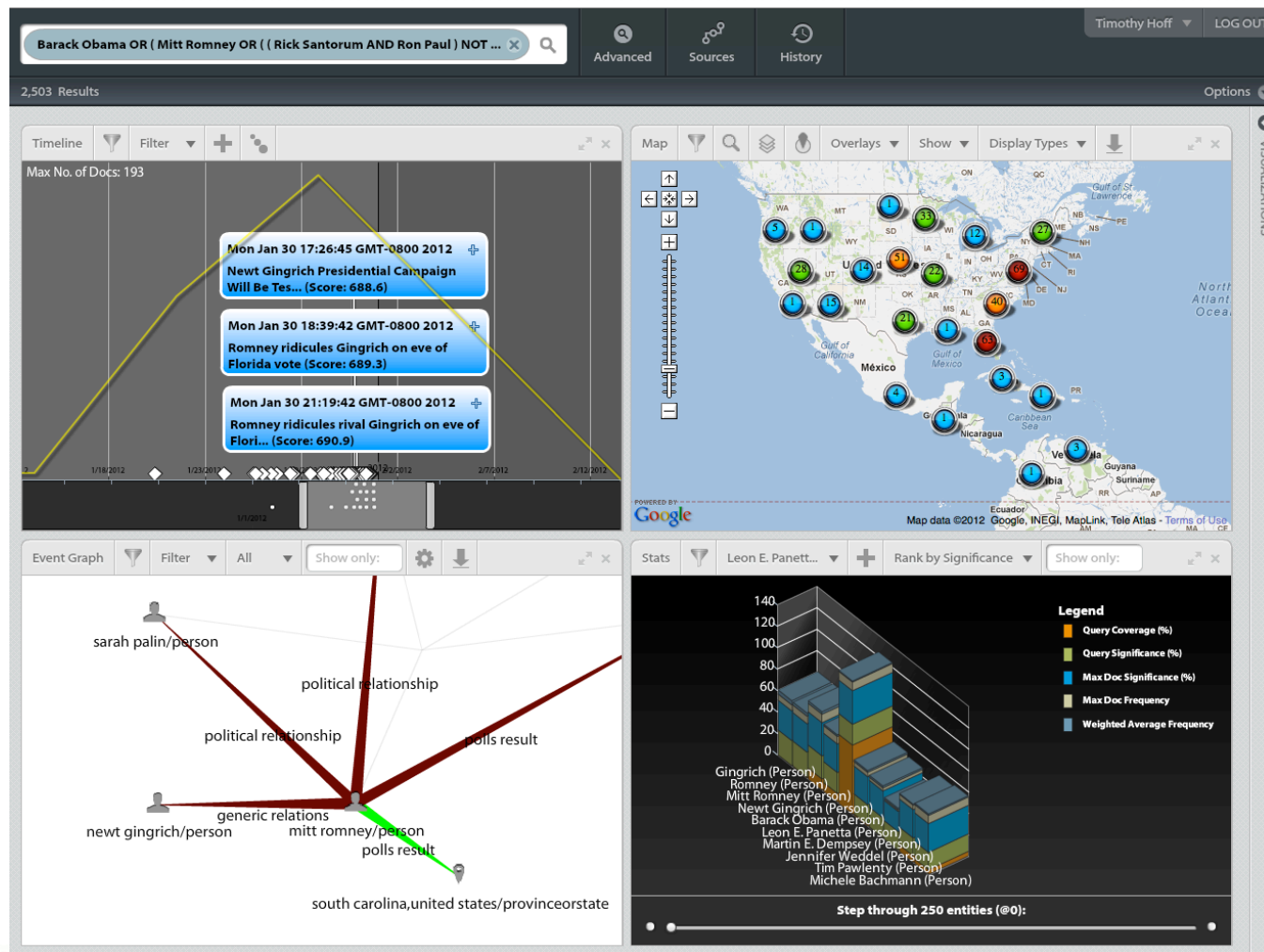
**Where**  
city, state, country, coordinate

# What Informat.e is - Framework

---

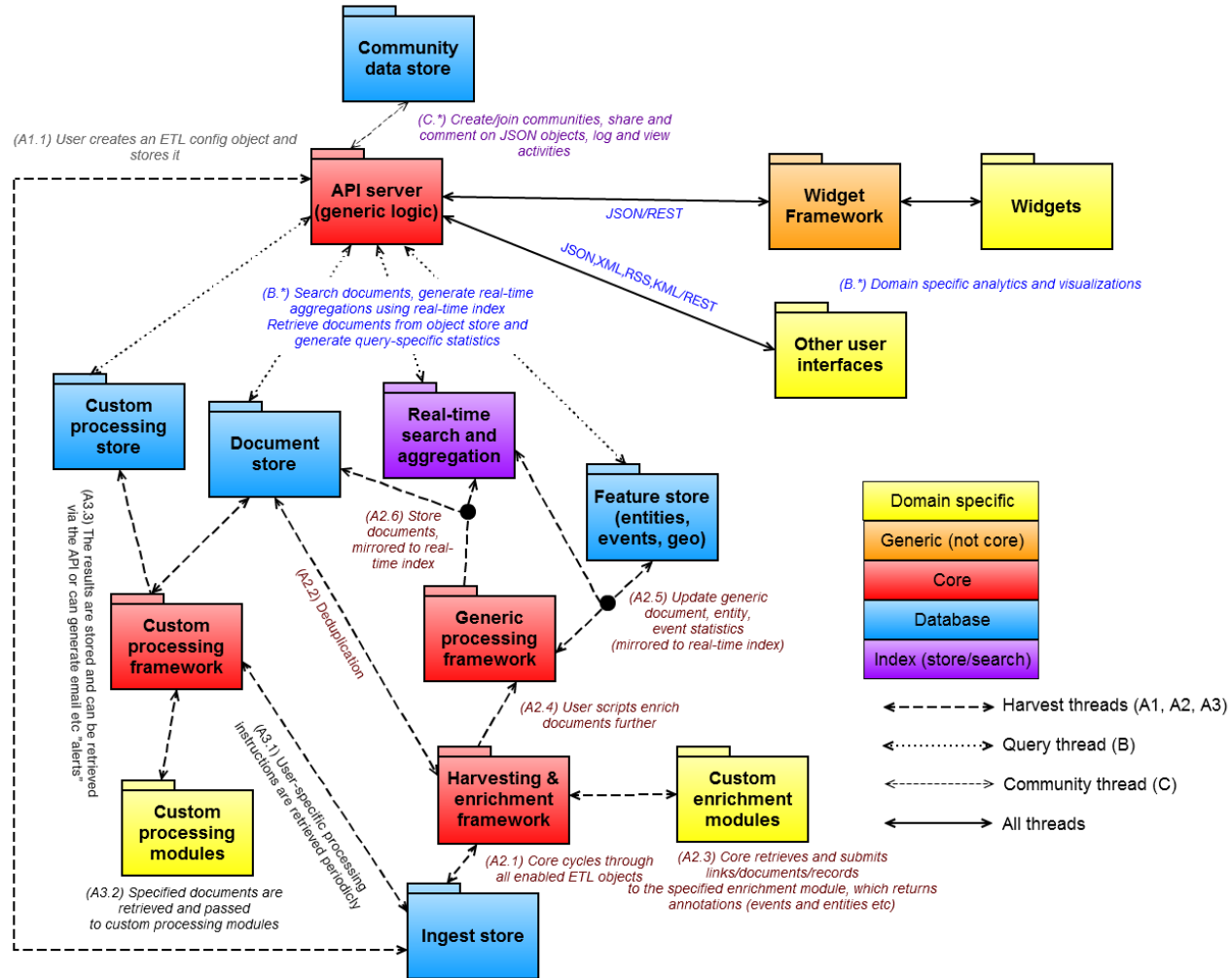


# What Informat.e is - Visualization



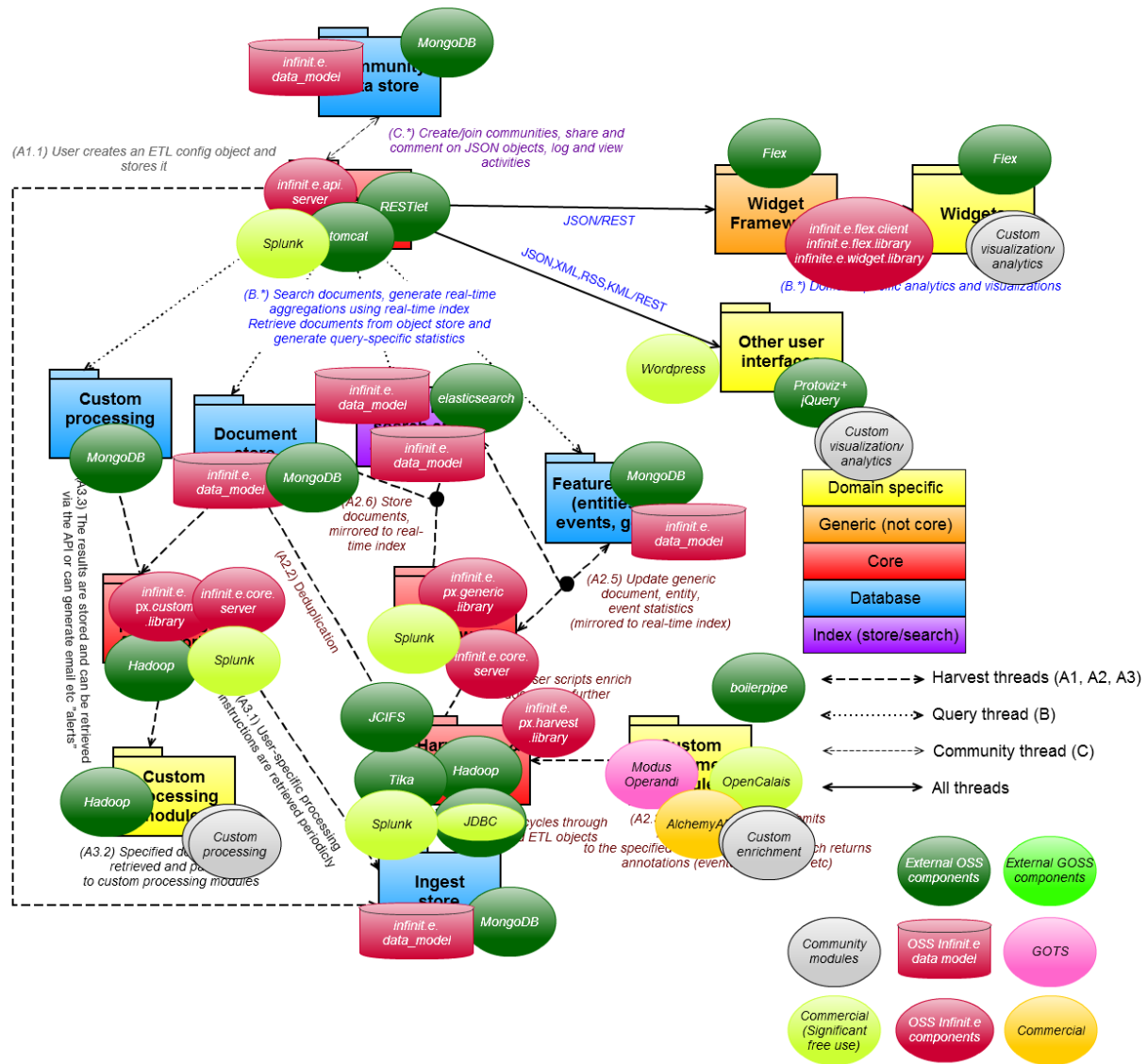
# Architecture

## Use of Open Source



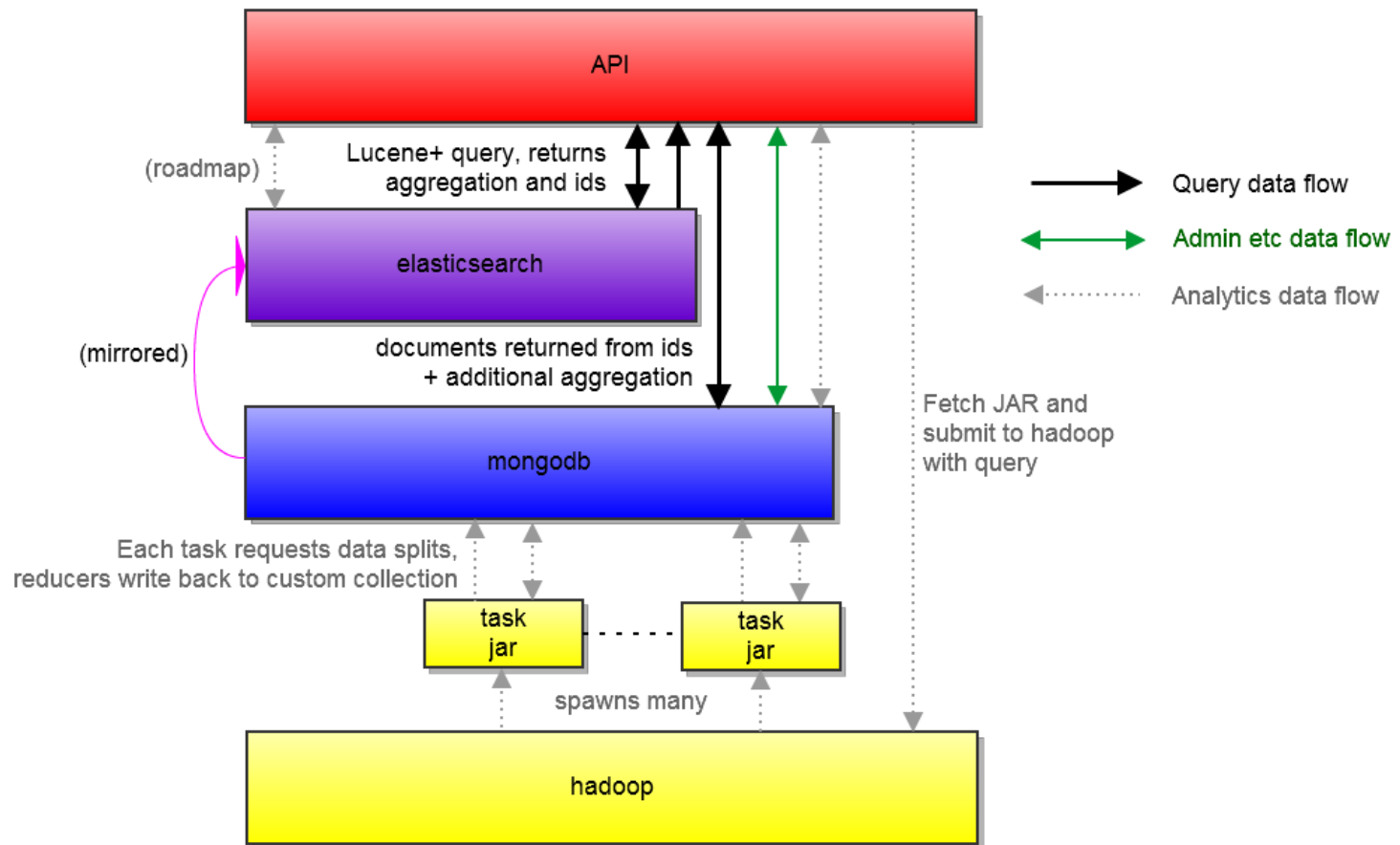
# Architecture

## Use of Open Source



# Architecture

## The 3 Key Elements



# Architecture

Focus on MongoDB

---

3 key areas of benefit:

- Development
- Integration
- Deployment

# MongoDB Development

---

Document analysis – lots of complex generic logic written in Java

- The “records” are all complex objects
  - BSON/JSON is a perfect representation
- Usually code maintainability is most important
  - BSON → “Plain Old Java Object”
    - (we use GSON, probably JACKSON is better; though GSON extensions for MongoDB types like dates and ObjectIds worked nicely)
- Sometimes performance is most important
  - Option to stay in BSON



# MongoDB – Dev Examples

---

- Converting to “POJO”

```
DocumentPojo docIn = new DocumentPojo();
docIn.setId(new ObjectId(idStr));
DocumentPojo docOut = DocumentPojo.fromDb(
    DbManager.getDocument().getMetadata().findOne(docIn.toDb()));
```

- Hybrid

```
BasicDBObject query = new BasicDBObject(DocumentPojo.communityId_,
    new BasicDBObject(MongoDbManager.in_, communityIdList));
// (then as above)
```

- Working in BSON only

```
BasicDBList l = (BasicDBList)(f.get(DocumentPojo.entities_));
for(Iterator<?> e0 = l.iterator(); e0.hasNext();){
    BasicDBObject e = (BasicDBObject)e0.next();
```



# MongoDB

## Changing Data Model

---

Standard requirement, particularly for an evolving project based on whatever functionality can be derived from the latest technologies...

- Example
  - We have [sentiment](#) as a property of [entity](#) (person/place/organization)
  - [association](#) links 2 [entity](#) objects via a [verb](#)
  - *New capability: NLP engine can now provide directed sentiment from one entity to another!*
- Often requires no extra dev effort at all...
  - Adding fields, eg just add [sentiment](#) to [association](#) above
- Otherwise, built in JSON format makes data model migrations easy
  - Have performed 2 major data model changes in 18 months, both via simple map/reduce scripts, with backwards compatibility



# MongoDB Integration

---

**Infini.e is based on NoSQL and web 2.0 technologies**

- ElasticSearch – JSON engine
- Javascript/Actionscript – JSON a key component
- NLP SaaS engines – JSON-based

**A key component of the custom ingest/enrichment is the ability to tag arbitrary source-specific metadata onto documents**

- Allows custom search / analytics / visualization
- “Best of both worlds” in conjunction with generic data model
- Schema-less storage is essential



# MongoDB Deployment

---

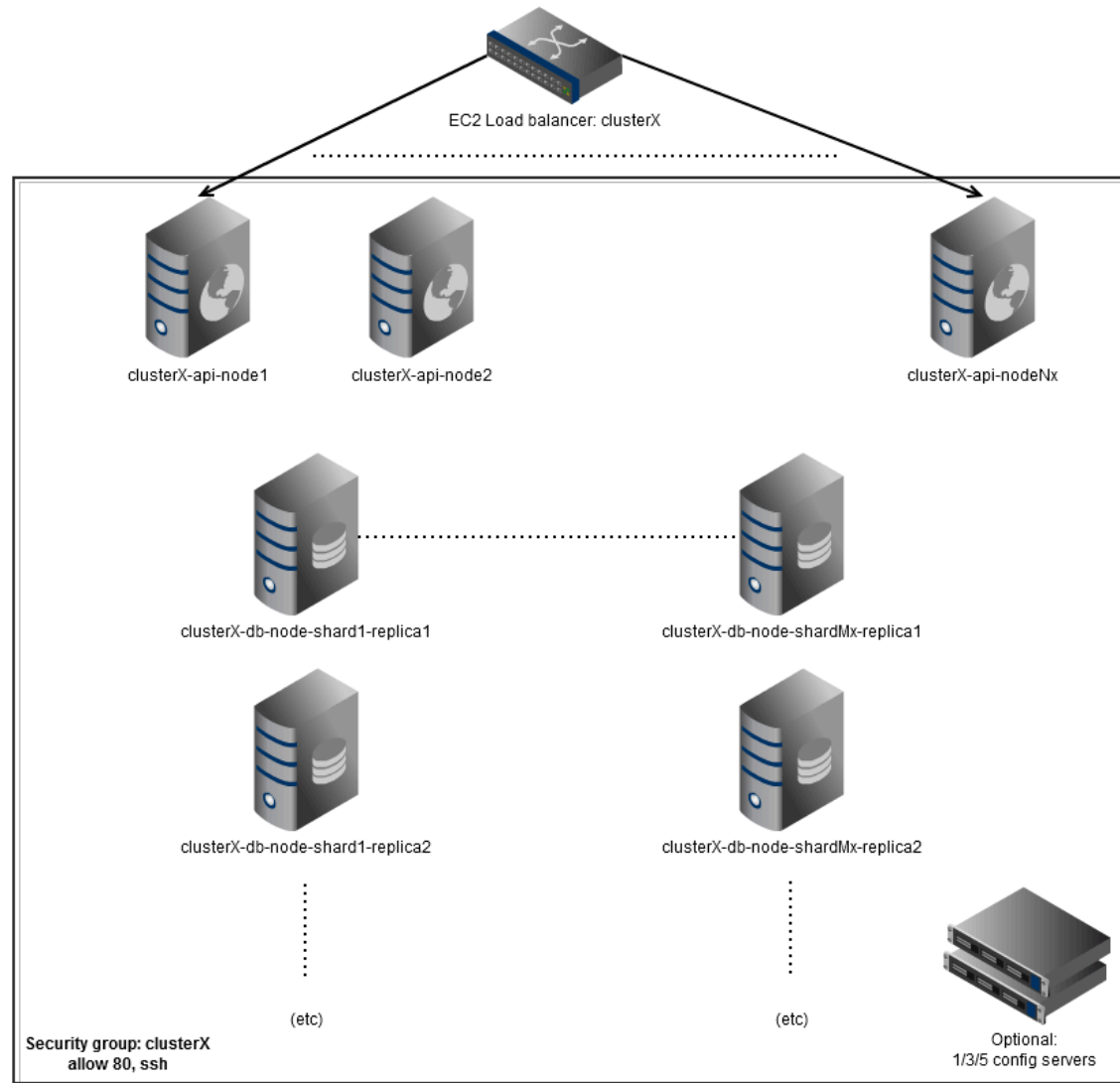
Need to scale in many directions:

- Writes due to new documents
- Reads for queries
- The ability to scale execution of domain specific logic
  - On ingest
  - Batch analytics

Infinitt.e is designed to use platforms like EC2 to scale



# MongoDB Deployment



# MongoDB Deployment

---

## MongoDB scalability

- Works!
  - Scales to arbitrary sizes in both read/write dimensions
- Sophisticated sharding keys provide powerful/flexible balancing
- Downsides:
  - Building an initial cluster is quite complex
  - Managing cluster changes is quite fiddly
- For Infit.e we used CloudFormation templates and (RPM-based) install scripts to manage the cluster
  - Works OK, a graphical tool and some more robustness would be nice
    - (on our roadmap, but not very close!)



# MongoDB Deployment

---

## MongoDB/EC2 integration

- m1.xlarge works best for our needs (m1.large is fine for ~0.5M docs)
  - 4 cores, 15GB
  - 4 500GB ephemeral disks that we RAID-0 together
    - (without that performance dropped off a cliff at >1M docs)



# Architecture

## Focus on Hadoop

---

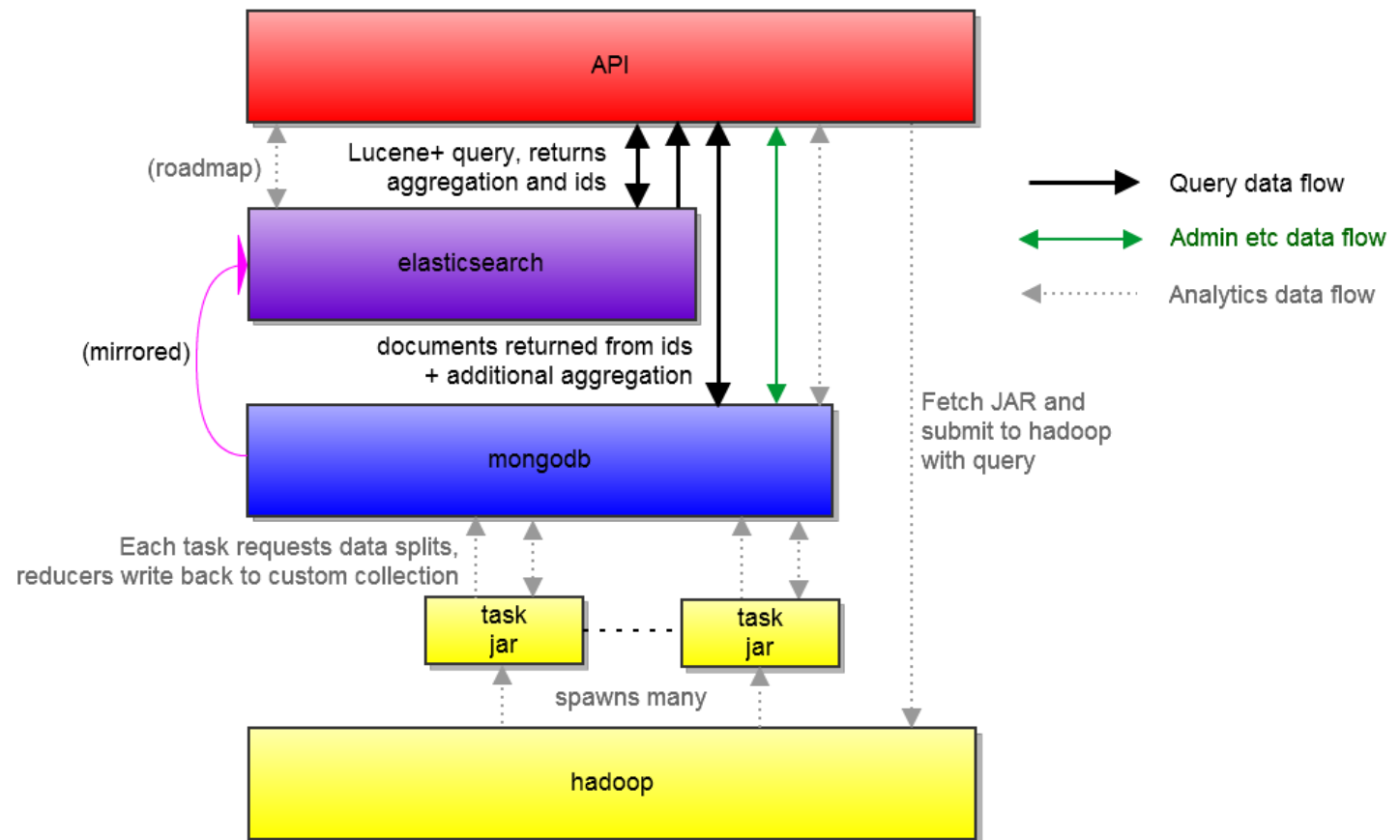
### Why Hadoop?

- Queries/aggregation/visualization is an excellent first step for document analysis, and is often all that's required
  - More complex analytics requires
    - Access to all of the data, not pre-aggregated or selected
    - A high level programming language, mature libraries etc
- Hadoop is becoming the de-facto standard for data analytics
  - Open Source, very customizable
  - Proven scalability
  - Java libraries
  - Mahout project (machine learning libraries for Hadoop)
  - Amazon elastic cloud



# Architecture

## MongoDB / Hadoop



# Infinite Demonstration

The screenshot displays the Infinite application interface. At the top, a search bar contains the word "Twitter", which is circled in red. Below the search bar, the interface is divided into several sections:

- Map:** A world map showing global sentiment distribution with red and green heatmaps.
- Network Graph:** A central graph showing relationships between various Twitter handles, such as "epn/twitterhandle" and "ozkacinh0/twitterhandle".
- Sentiment Chart:** A 3D bar chart showing sentiment scores for various keywords. A legend indicates: Positive Sentiment (green), Aggregate Sentiment (blue), and Negative Sentiment (red). The chart is titled "Step through 2224 entities (@0):".
- Entity Significance Chart:** A 3D bar chart showing sentiment scores for various hashtags, such as "#PrimarySchoolMemories (HashTag)" and "#oofm (HashTag)". It is titled "Step through <=2224 entities (@0):".
- US Map:** A smaller map of the United States showing sentiment distribution, also circled in red.

Additional elements include a "Filter Visualizations" sidebar on the right, a "Doc Browser" section, and a "Zestimate Chart" provided by Zillow at the bottom right.

# Infinite Demonstration

The screenshot displays the Infinite File Upload Tool interface. At the top, a browser window shows the URL `infinite.rr.ikanow.com`. The main interface is divided into several sections:

- Twitter Search:** Shows 85,351 results for the query `veet[tweets_t]`. Navigation options include Map, Overlays, Show, Display Types, Event Graph, Filter, and All.
- World Map:** A map of the world with red and green heatmaps indicating sentiment distribution across continents like North America, Europe, and Africa.
- Network Graph:** A central graph showing relationships between various Twitter handles, such as `epn/twitterhandle`, `ozkajinh0/twitterhandle`, and `retweet`.
- File Uploader:** A central modal window titled "File Uploader" with the following details:
  - Filter On: `application/java-archive`
  - Edit: `GenericSentimentPlugin` (with a Delete button)
  - Title: `GenericSentimentPlugin`
  - Description: `GenericSentimentPlugin: Aggregates document level sentiment by geographic location`
  - Communities: `Counter Terrorism`, `Law Enforcement Fusion Center`, `Nigerian Terrorism Events in 2010`, `Sentiment Analysis`
  - File: `Choose File` (No file chosen)
  - Share URL: `http://infinite.rr.ikanow.com/api/share/get/4f8c7abde4b09c`
  - Owner: `cvitter@ikanow.com`
  - Buttons: `Submit`, `Log Out`
- Data Visualizations:** Two 3D bar charts at the bottom. The left chart is titled "Step through 2224 entities (@0):" and lists keywords like `tomorrow (Keyword)`, `photo (Keyword)`, `Dont (Keyword)`, `Hate (Keyword)`, `school (Keyword)`, `Girl (Keyword)`, `Video (Keyword)`, `damn (Keyword)`, `happy (Keyword)`, and `Skip (Keyword)`. The right chart is titled "Step through <=2224 entities (@0):" and lists hashtags like `#PrimarySchoolMemories (HashTag)`, `#oofm (HashTag)`, `#Nf (HashTag)`, `#itanic (HashTag)`, `#FUCK (HashTag)`, `#twitter (HashTag)`, `#NowPlaying (HashTag)`, `#PretextosParaNoDebatir (HashTag)`, `#Syria (HashTag)`, and `#fui (HashTag)`.
- Right Sidebar:** Contains various widgets including "Filter Visualizations", "REIT Explorer", "Zestimate Chart" (provided by Zillow), and "Twitter Sentiment".

# Infinite Demonstration

The screenshot displays the Infinite File Upload Tool interface. The main window shows a Hadoop job execution details for 'Hadoop job\_201204161654\_0013 on ip-10-12-105-125'. The job status is 'Succeeded' and it finished in 3 minutes and 10 seconds. A table below shows the job's progress:

| Kind   | % Complete | Num Tasks | Pending | Running | Complete | Killed | Failed/Killed Task Attempts |
|--------|------------|-----------|---------|---------|----------|--------|-----------------------------|
| map    | 100.00%    | 554       | 0       | 0       | 554      | 0      | 0 / 2                       |
| reduce | 100.00%    | 1         | 0       | 0       | 1        | 0      | 0 / 0                       |

Below the table is a 'Job Counters' section with a table showing various counters for Map and Reduce tasks, including 'SLOTS\_MILLIS\_MAPS', 'Launched reduce tasks', and 'Total time spent by all reduces waiting after reserving slots (ms)'. The 'SLOTS\_MILLIS\_REDUCES' counter is highlighted with a red circle.

The interface also features a Twitter sentiment visualization on the left, showing a map of North America with red and green heatmaps indicating sentiment levels. A 3D bar chart below the map shows sentiment scores for various keywords like 'tomorrow', 'photo', 'Dont', 'Hate', 'school', 'Girl', 'Video', 'damn', 'happy', and 'Skip'. On the right, there are social media widgets including a REIT Explorer chart and a Zillow Zestimate Chart.

# Infinite Demonstration



# Thank You!!!

---

Alex Piggott  
Director of Product Engineering  
[apiggott@ikanow.com](mailto:apiggott@ikanow.com)

