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

Agenda

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

IKANOW

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

IKANOW

What Infinit.e is

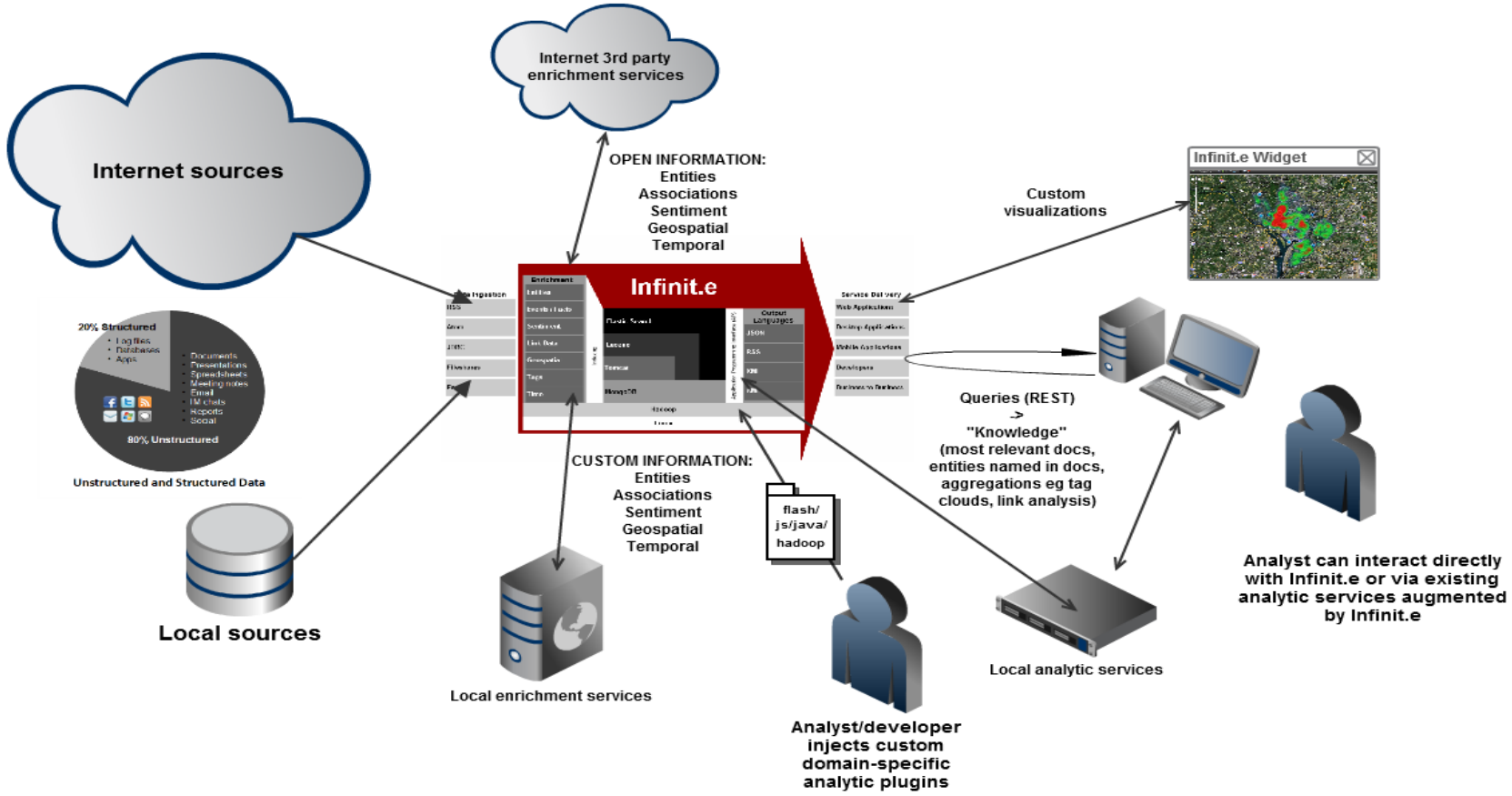
Infinit.e is a scalable framework for:

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

Unstructured documents and structured records

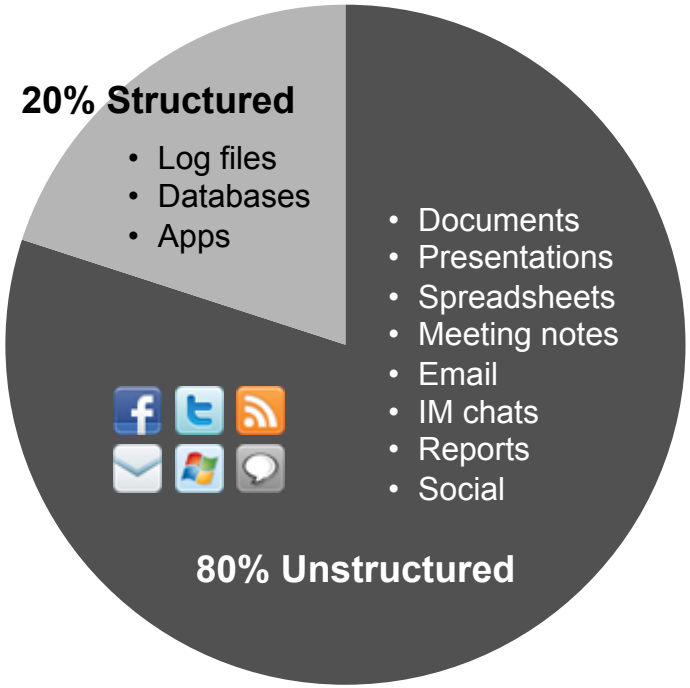
IKANOW

What Infinit.e is - Overview

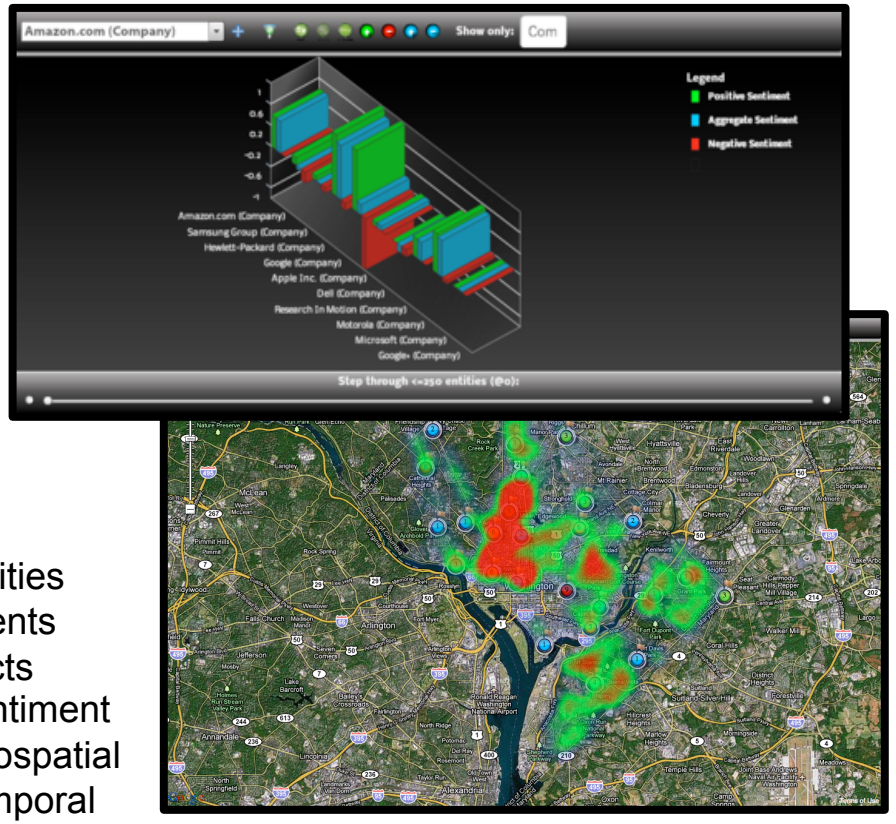


IKANOW

What Infinit.e is - Documents



Unstructured and Structured Data



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

IKANOW

What Infinit.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

What
events, summaries, facts, themes

When
past, present, future dates

Where
city, state, country, coordinate

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

ORACLE

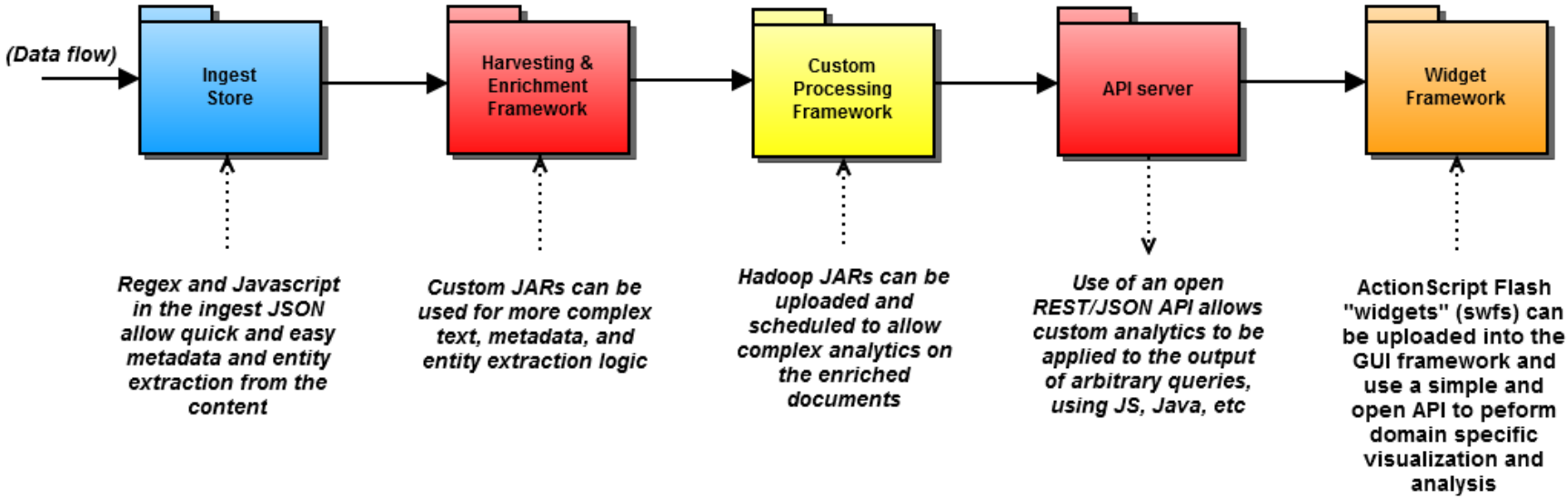
<?xml?>

Microsoft SQL Server

IKANOW

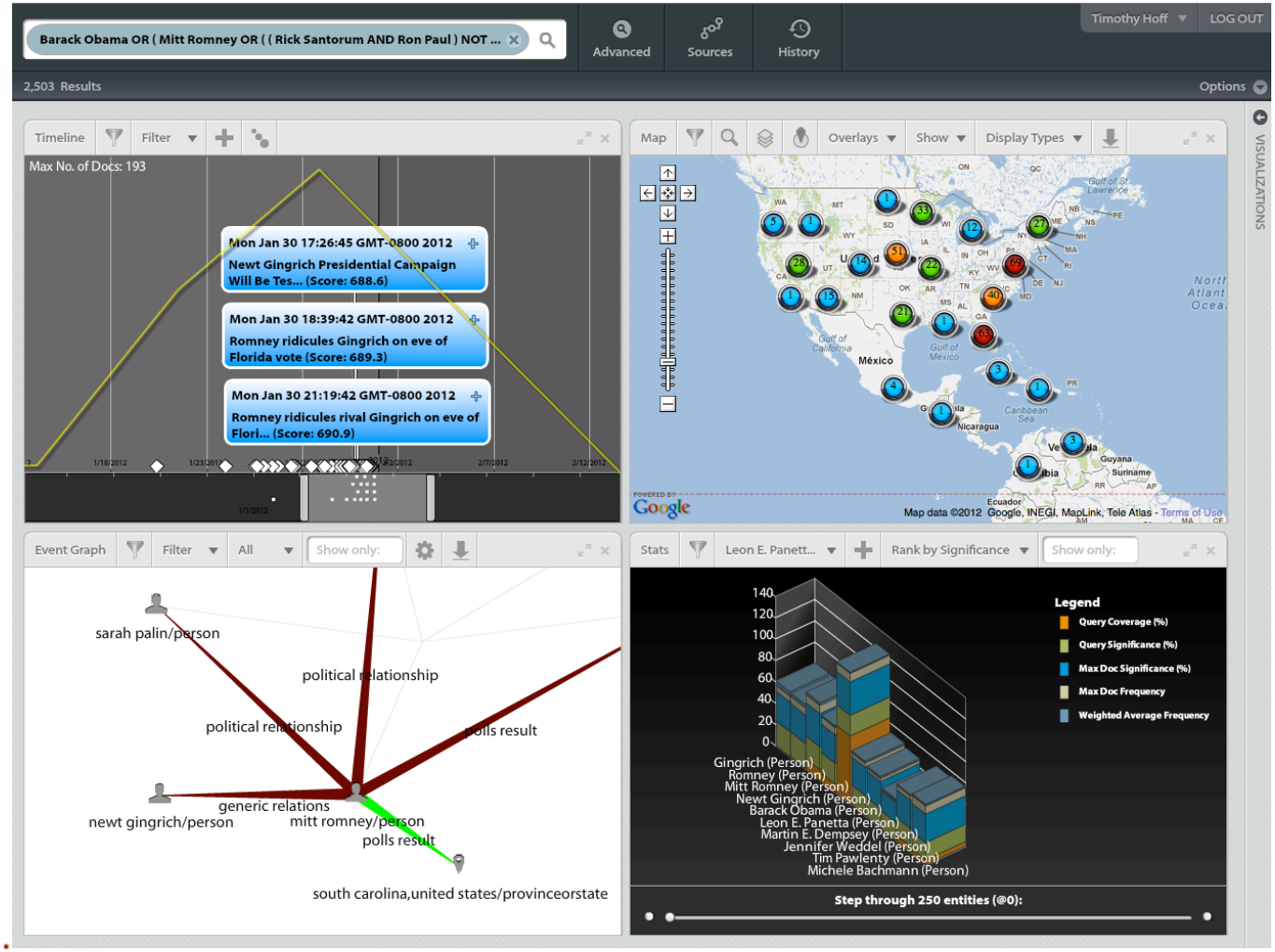
```
<Incident>
  <uid>20101043423</uid>
  <subject>1 person killed in armed attack by
 suspected Boko Haram in Maiduguri, Borno,
 Nigeria</subject>
  <multipleDays>No</multipleDays>
  <eventDate>06/04/2011</eventDate>
</Incident>
```

What Infinit.e is - Framework



IKANOW

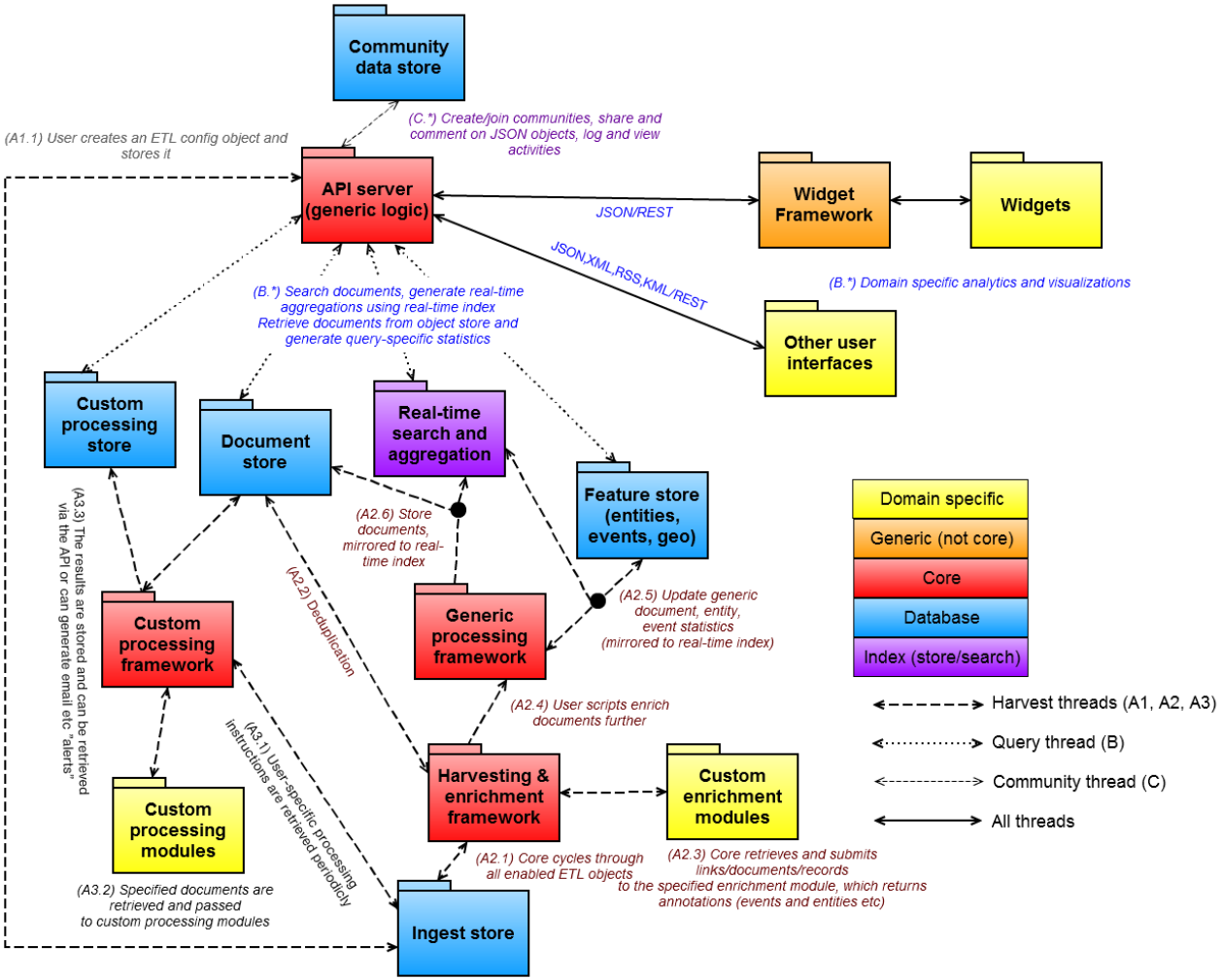
What Infinit.e is - Visualization



IKANOW

Architecture

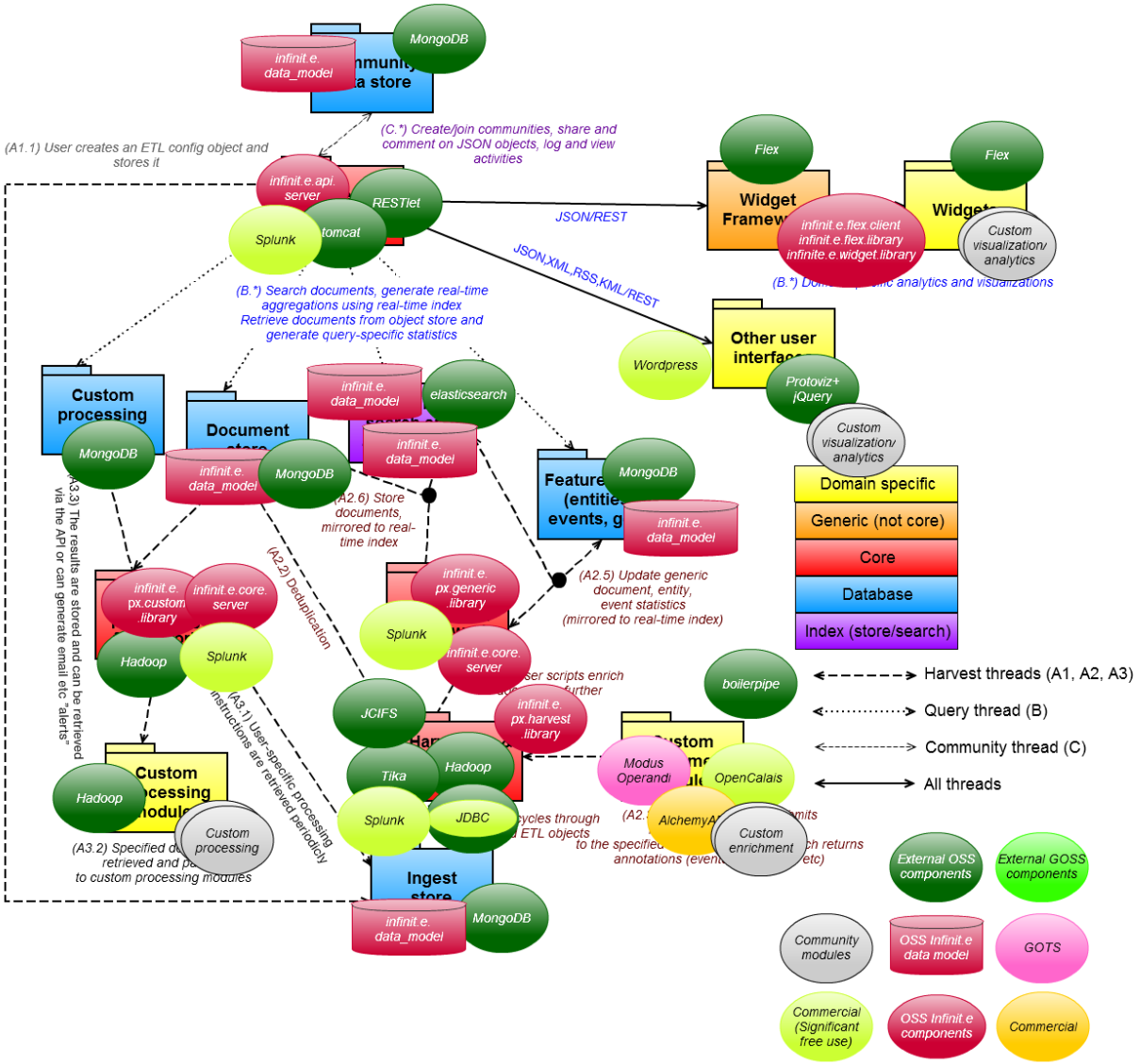
Use of Open Source



IKANOW

Architecture

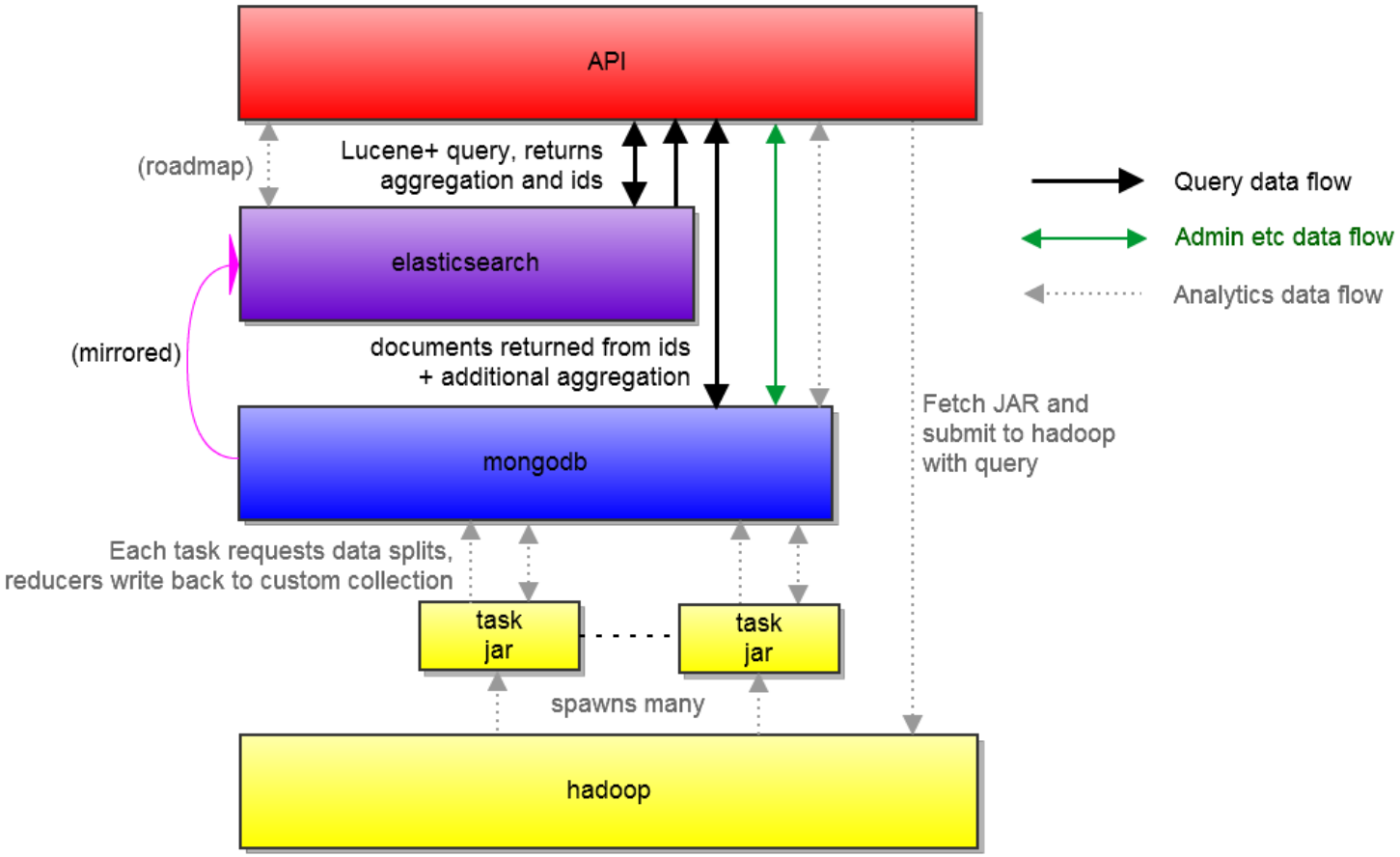
Use of Open Source



IKANOW

Architecture

The 3 Key Elements



IKANOW

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

IKANOW

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();
```

IKANOW

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

IKANOW

MongoDB Integration

Infinit.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

IKANOW

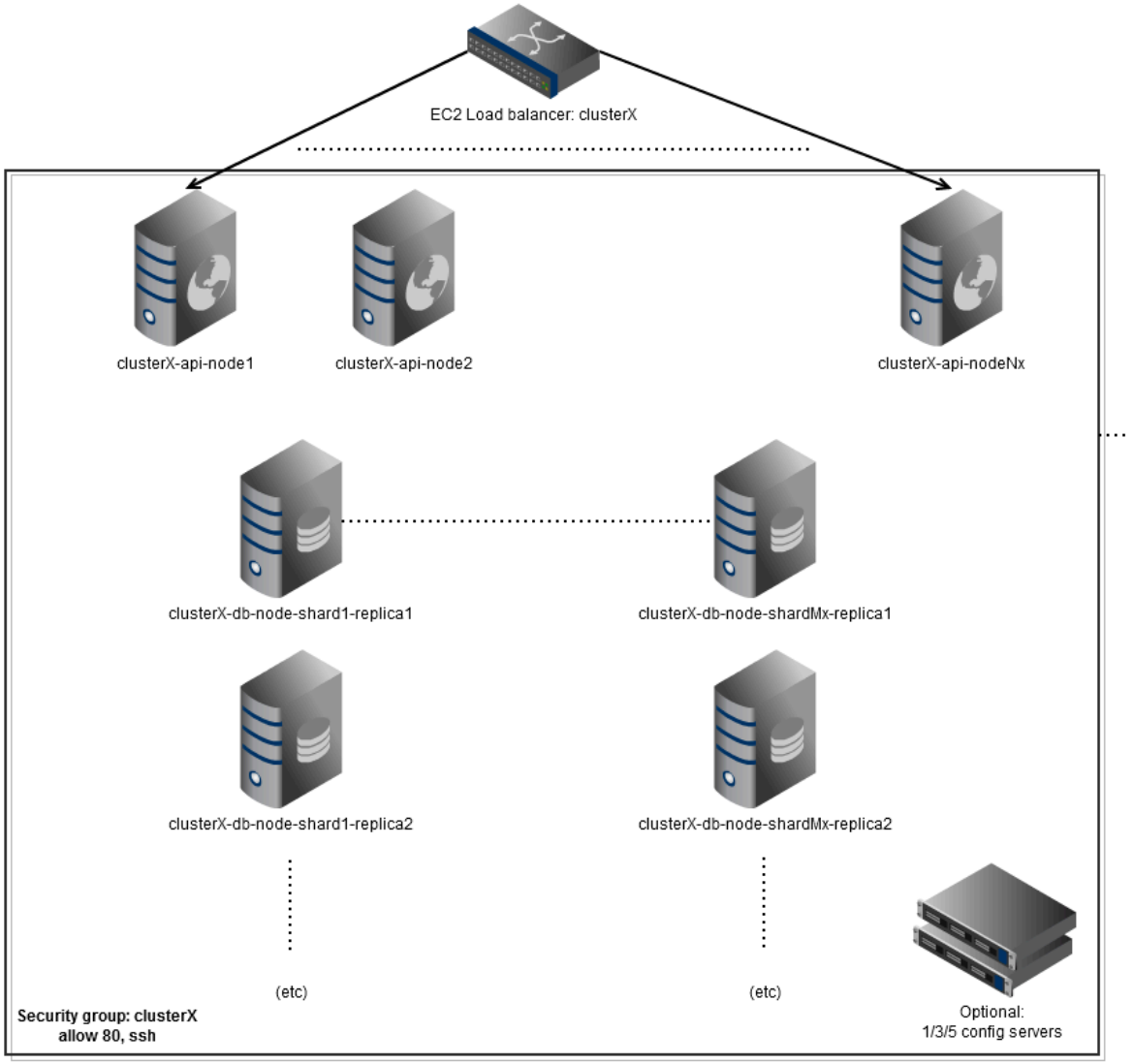
MongoDB Deployment

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

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

MongoDB Deployment



IKANOW

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 Infinit.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!)

IKANOW

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

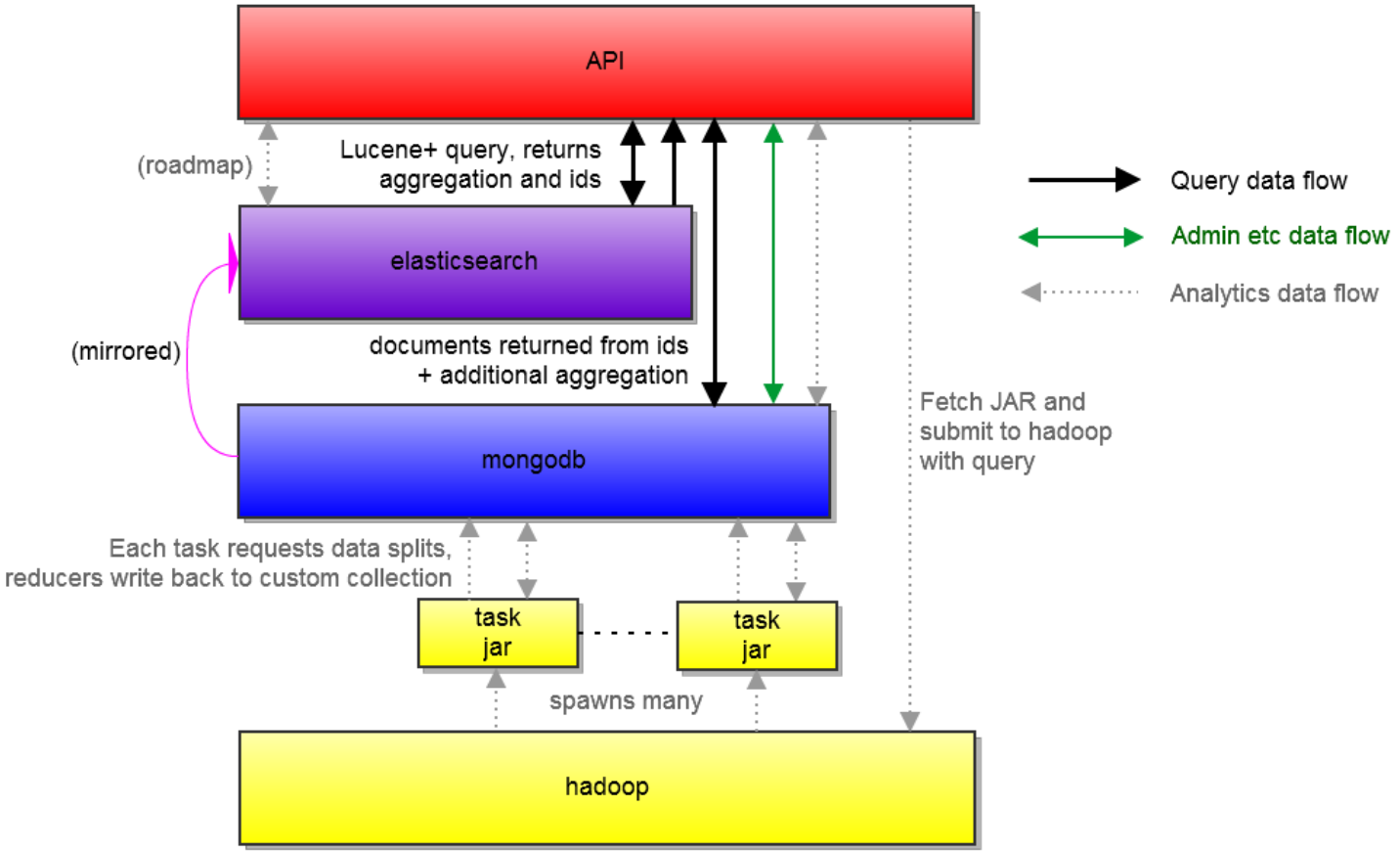
Why Hadoop?

- 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

IKANOW

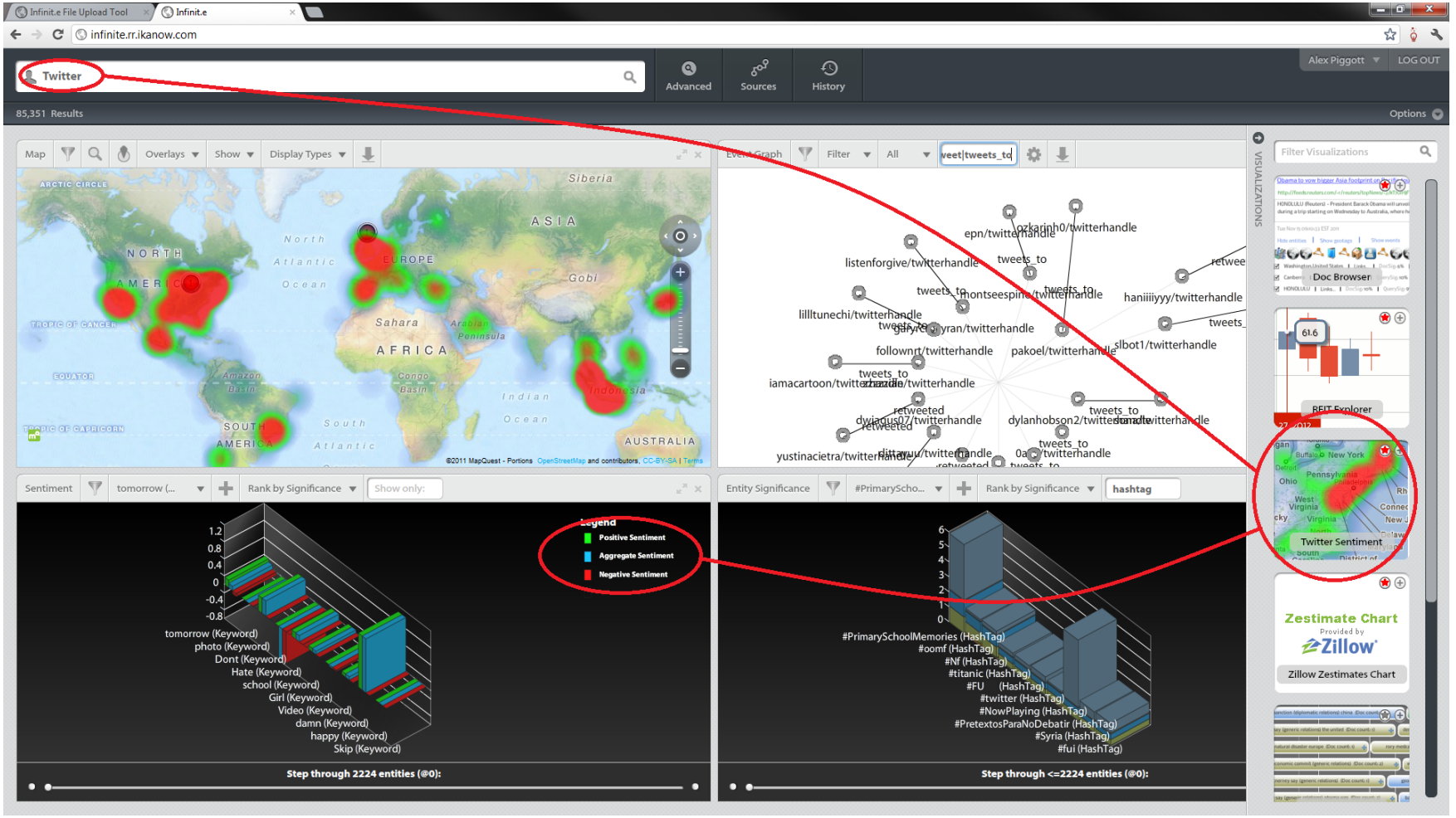
Architecture

MongoDB / Hadoop



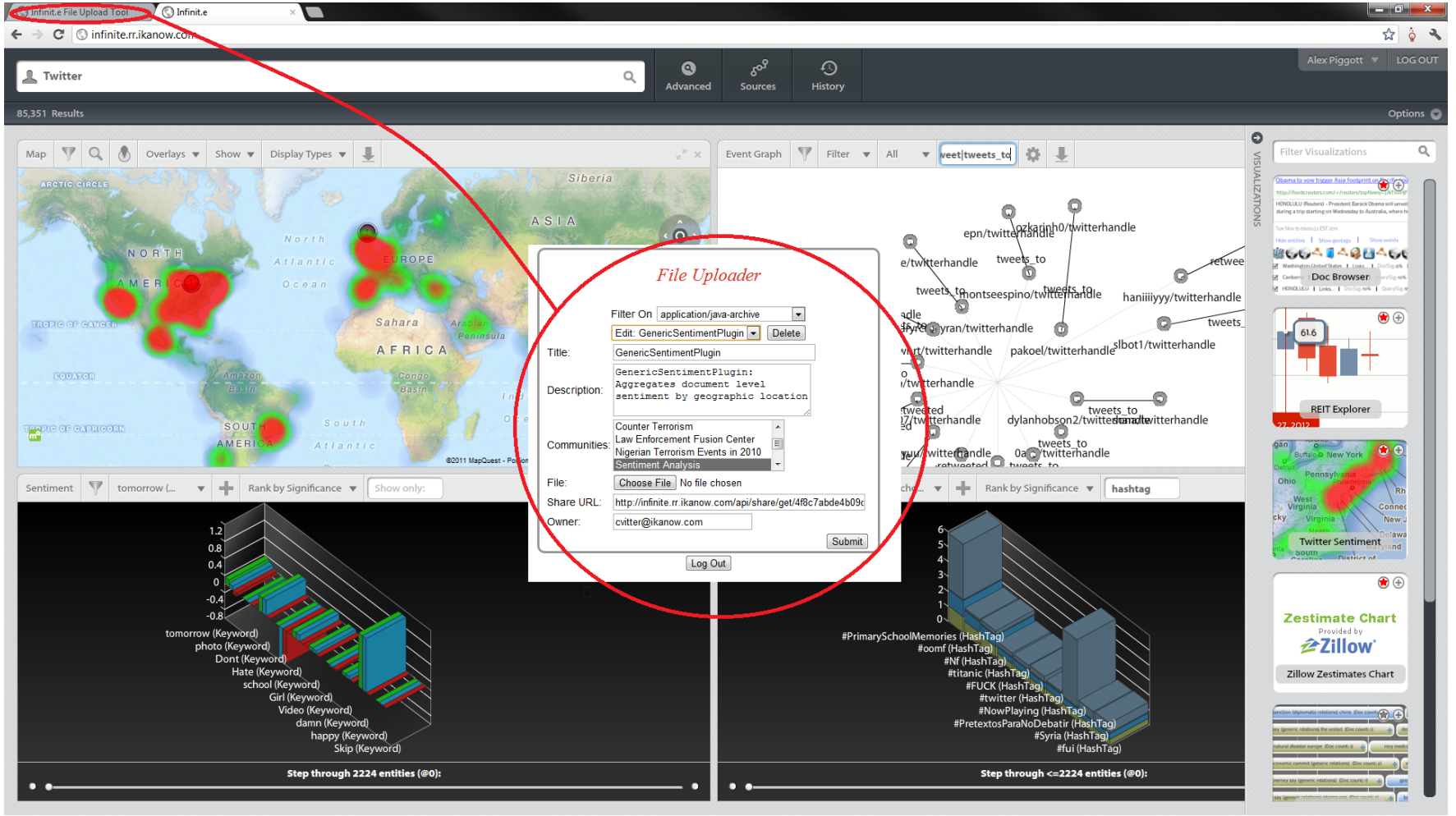
IKANOW

Infinit.e Demonstration



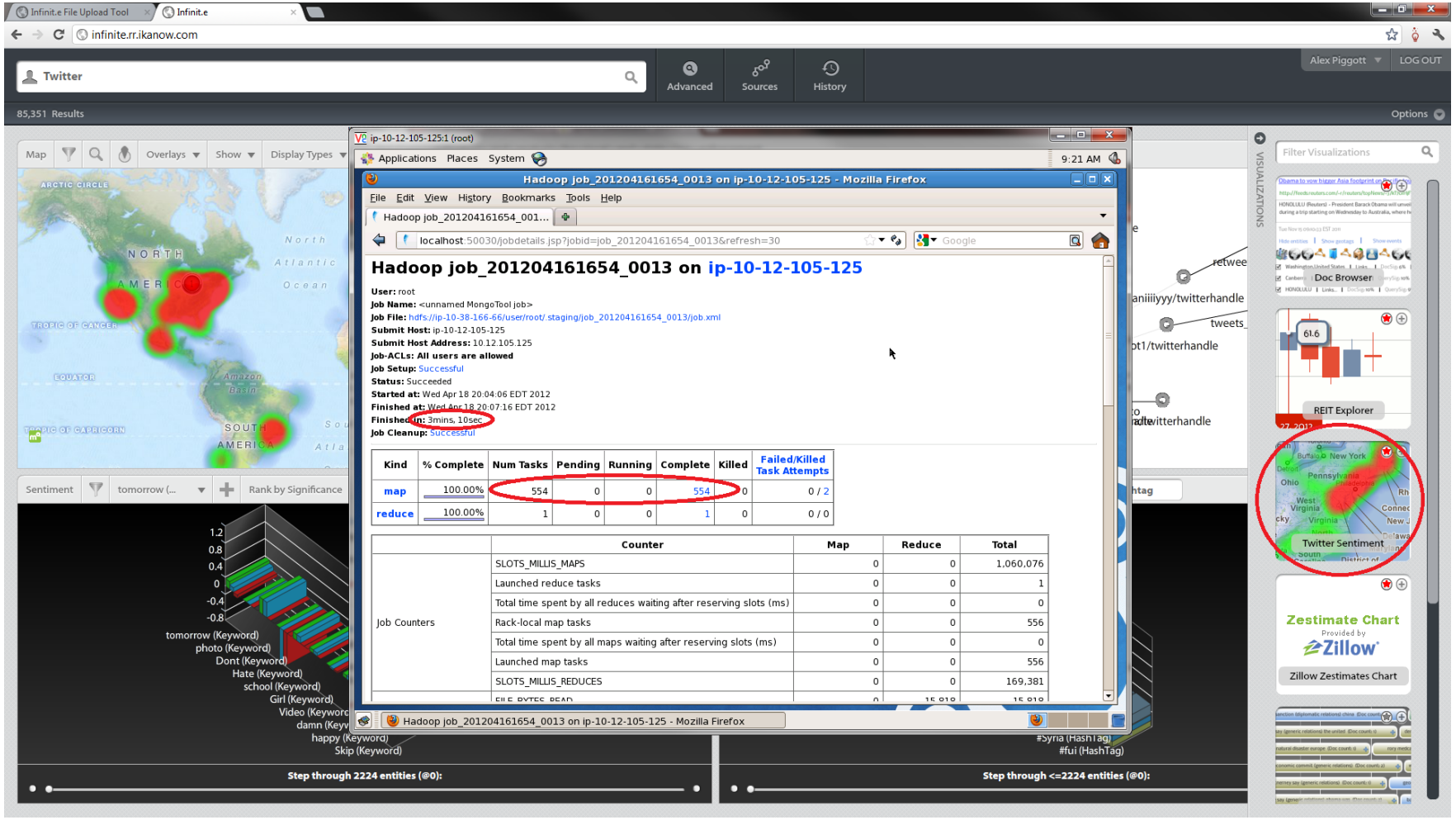
IKANOW

Infinit.e Demonstration



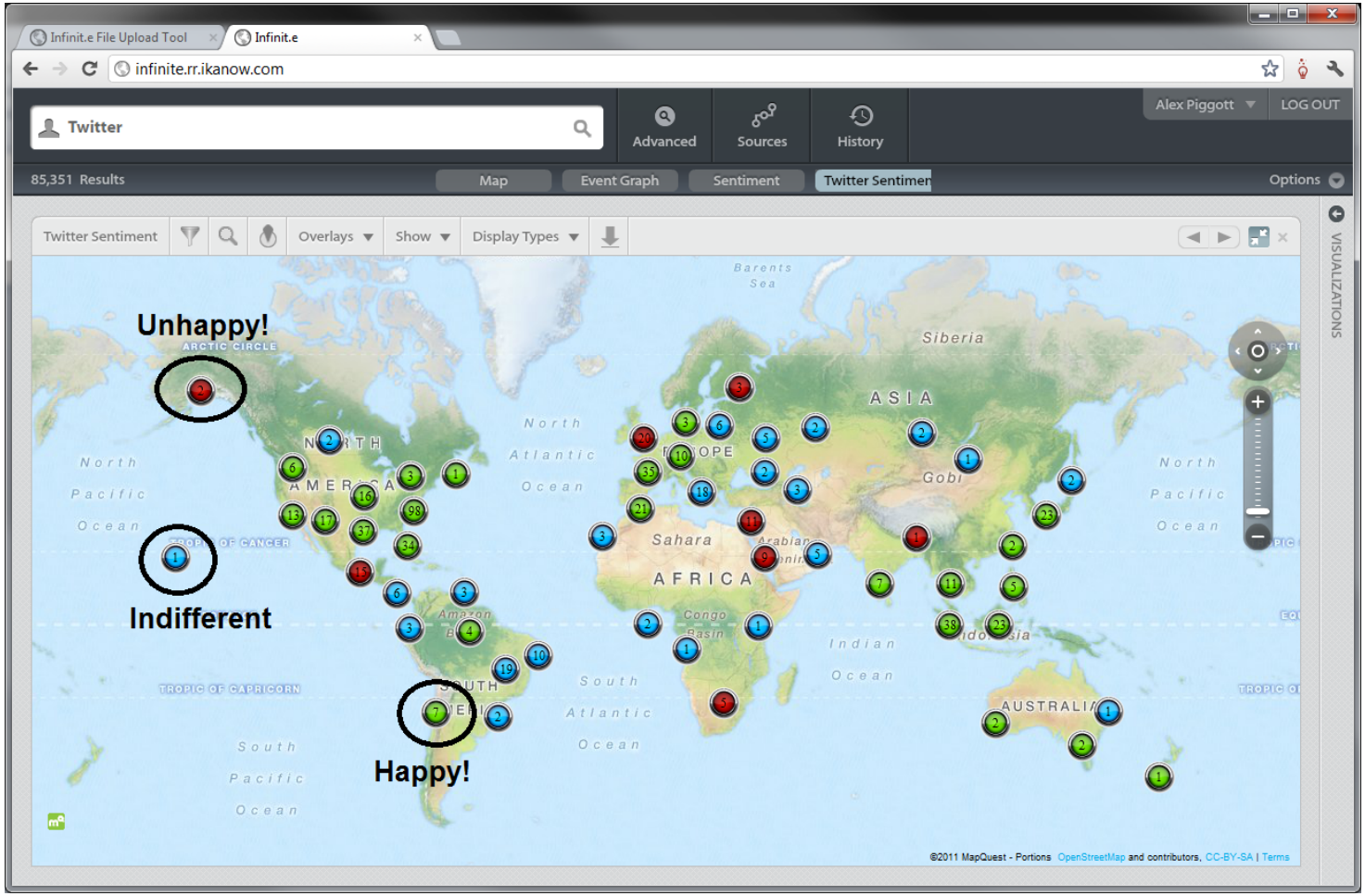
IKANOW

Infinit.e Demonstration



IKANOW

Infinit.e Demonstration



IKANOW

Thank You!!!

Alex Piggott
Director of Product Engineering
apiggott@ikanow.com

IKANOW