Hadoop-based Open Source eDiscovery: FreeEed

(Easy as popcorn)

Hello!

- Sujee Maniyam & Mark Kerzner
- Founders @ Elephant Scale
- consulting and training around Hadoop, Big Data technologies
- Enterprise support and customization for FreeEed
- Oh, we wrote a free, open-source book on Hadoop http://hadoopilluminated.com/
- Contact : info@elephantscale.com | www.elephantscale.com

Why You May Find This Talk Interesting..

- Hadoop is all the rage these days
- 'log processing' is the most common use case for Hadoop
  - Click stream logs
  - Application logs….
  - …etc…etc…
- This is a slightly different use case for Hadoop
  - Analyzing emails & documents at scale…
  - And searching them interactively
  - Using 'Hadoop-On-Demand' in the cloud
- We think it is an interesting application
- Hope you do too… ☺

What is FreeEed?

Open source

eDiscovery tool,

Built on Hadoop

FreeEed : Free Electronic Discovery

- FreeEed is an open source product for eDiscovery
- Benefits for legal
  - 300+ file formats indexed
  - Deduping, chain-of-custody, OCR, other standard features
  - Browser-based review
- Technical benefits for legal techs and IT
  - Big Data technologies, scalable, open, extensible
- Open source project stats
  - 2+ years in development
  - 100s of downloads, ~1000 commits, 3 committers

Open Source !

- Is it really open source?

YES !!

- Apache 2 License
  - Very business friendly license
  - Same license as : Hadoop, Solr ….
- http://freeeed.org/
- https://github.com/markkerzner/FreeEed

Business use case and background

- What is eDiscovery
- History of legal discovery
- eDiscovery statistics
- Challenges

Business (legal) use case

- Duty to disclose information – rule FRCP 26
  - Preserve relevant information
  - Produce information on request
  - Keep the information for X years
  - Sanctions for obstruction
  - Sanctions for non-compliance

Before the thirties

- Court room was full of surprises

Civil discovery changes this

Discovery basics

- Obligations of the parties
  - At the start of a lawsuit or litigation possibility, preserve relevant data
  - Produce data at request, within timelines
  - Review the data before production
  - Can request eDiscovery from opponents
  - Store and archive

Interesting facts about eDiscovery

- Most of these are proprietary or under NDA
  - Representative case size: 5GB to 500GB
  - Cost per GB of processing: $5-200, ~$100
  - Takes 25-50% of litigation budget
  - Days to process and months to review
  - Preservation: 3-7 years
  - 500 providers, with 10 majors

Challenges of eDiscovery

- Data sizes in the TB
- Seasonal loads, tight deadlines
- Hundreds of file formats
- Heavy read/write load in review
- Text analytics is of paramount importance
- Huge price tickets obstruct justice

Meet FreeEed

- Open source eDiscovery tool
  - built on open technologies
  - And FREE!
- Built to handle large volumes of data
  - 100s GB → Tera bytes
- Easy to use GUI dashboard
- Scales with work-load
- Comes in Virtual Machine for easy adoption / preliminary analysis

Meet FreeEed

- Built on Hadoop
  - Scalable as Hadoop
- Cloud integration
  - Run Hadoop in Cloud at scale
  - No need to maintain a Hadoop cluster in-house
- Ultra fast search
- Preview / Tagging

Design goals

- Built on open source components
- Big Data scalable
- Preservation, chain of custody, archiving
- Scalable technically and business-ly
- Stable (don't laugh, people get different results on different runs)
- Extensible

Packaging architecture

- Comes as VM's
- Grab as few or as many as you want
- No mixing of matters
- No ethical problems
- Preserve for as many years as you want
- 1 VM = 1 corn, FreeEed = free popcorn

FreeEed makes lawyers happy

Next:

- eDiscovery background
- FreeEed features
- Technical Overview

FreeEed Technology Stack

- Challenge 1) How do we handle large volumes of data?

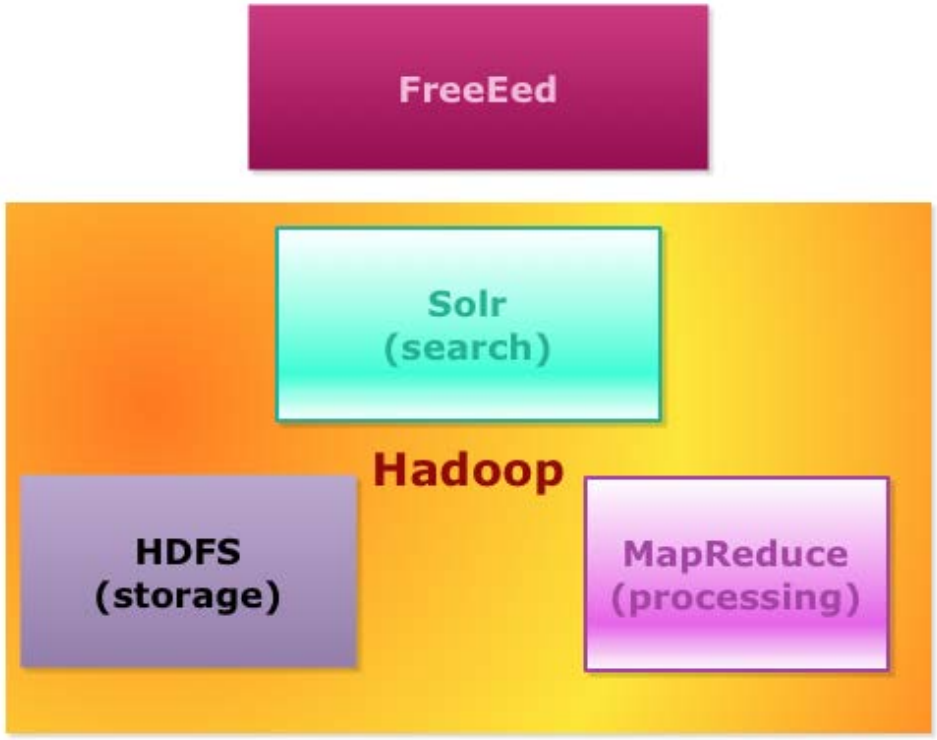
Chronicle / Frederic Larson

Hadoop Quick Overview

- Hadoop provide distributed computing / processing
  - 'OS' for distributed systems
- Built for scale
  - Can handle Petabytes of data
- Open source (Apache project)
- Thriving community
  - Vendors
  - Tools
  - …

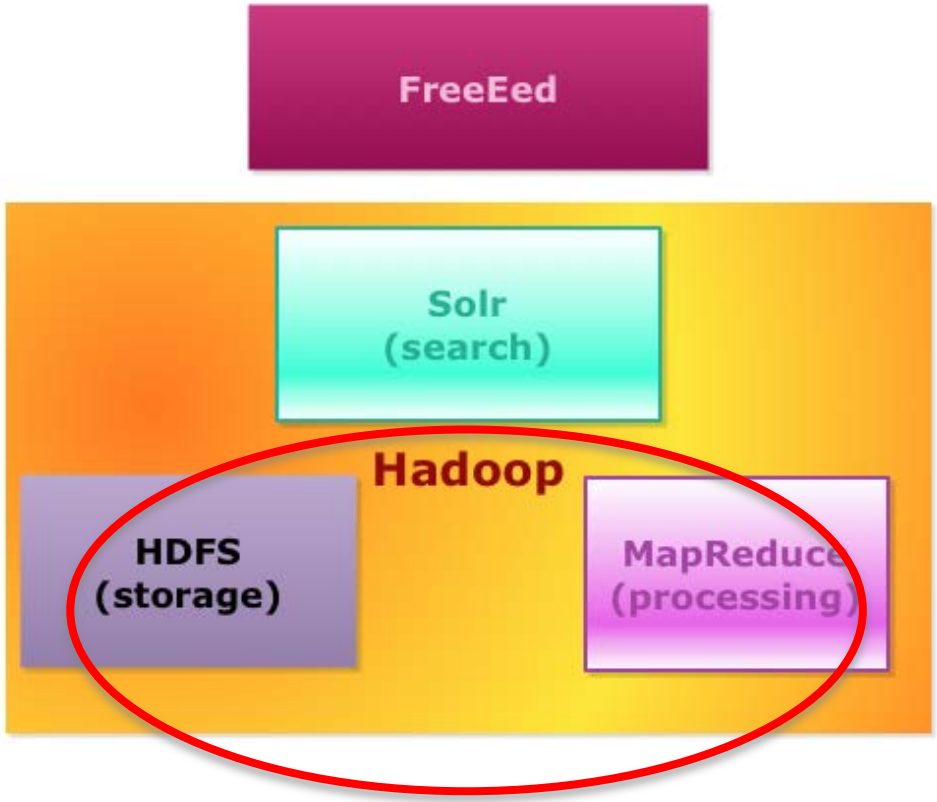
How do we use Hadoop?

- Hadoop is our base platform

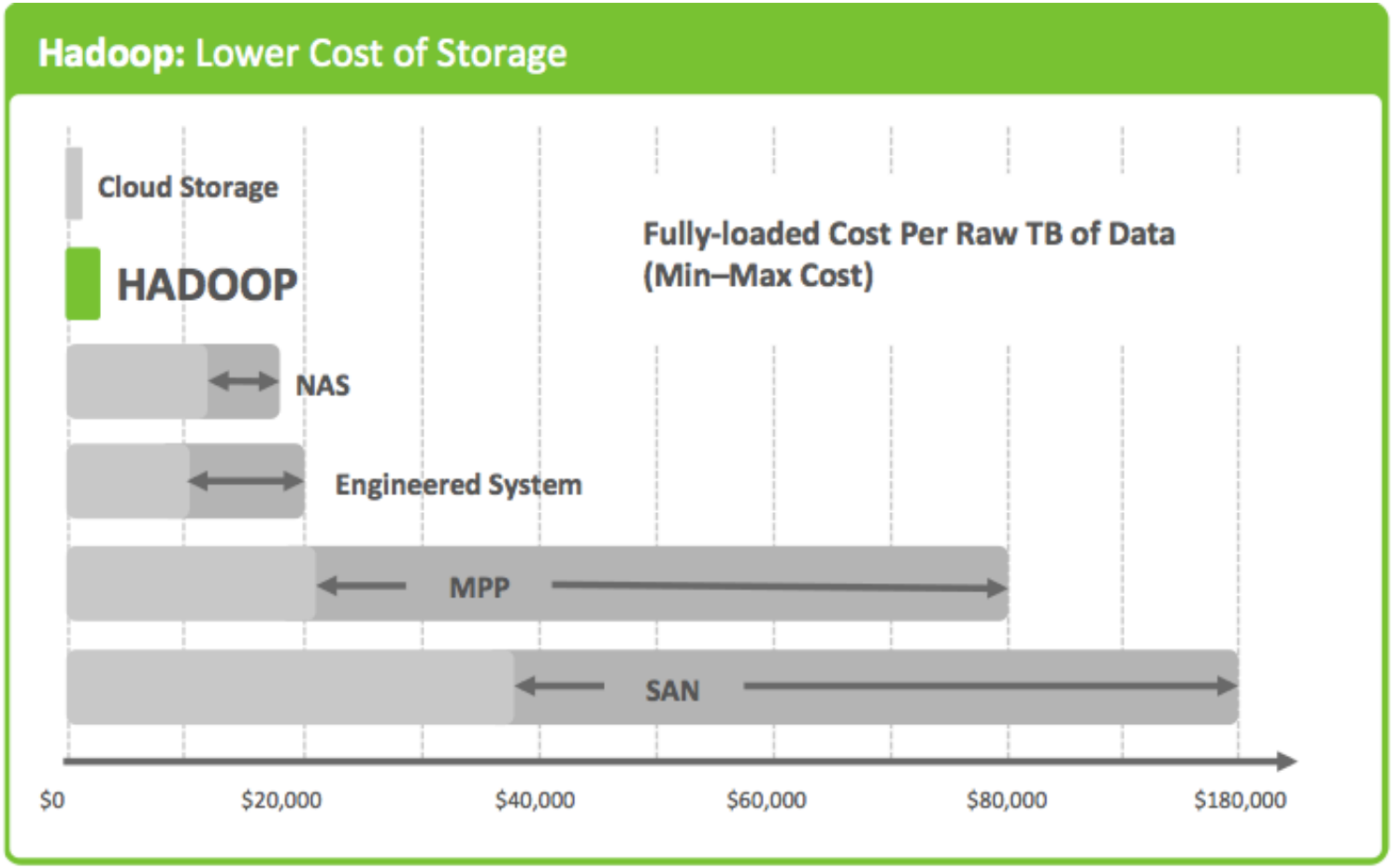


How do we use Hadoop?

- Hadoop is our base platform



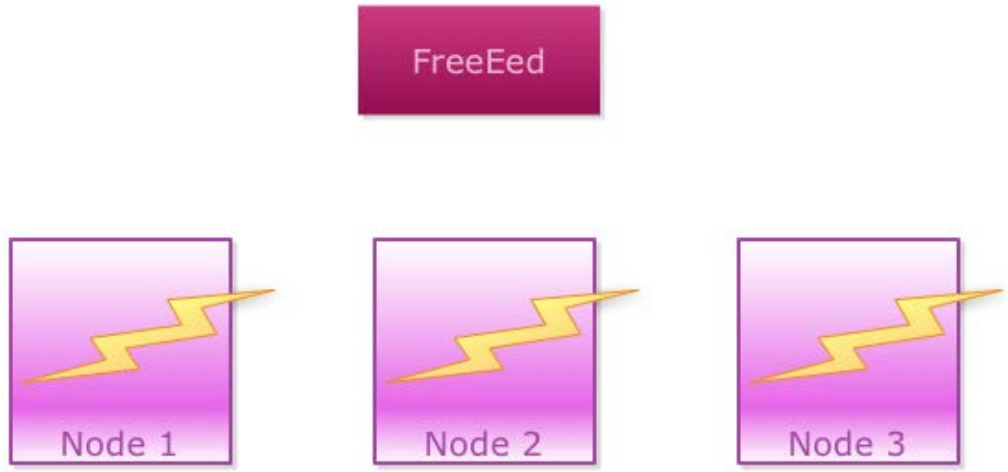
Hadoop : Cheap Storage



Source : HortonWorks

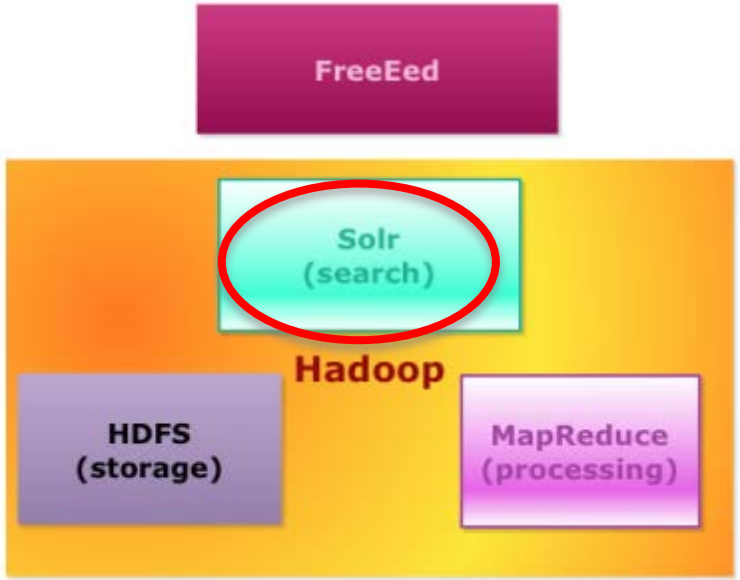
Hadoop : Distributed Processing

- Process large data in distributed fashion
- Use MapReduce to process data in parallel
- E.g : Enron dataset
  - Size ~100G
  - Amazon cloud, 50 nodes
  - Processing time : 1 hour

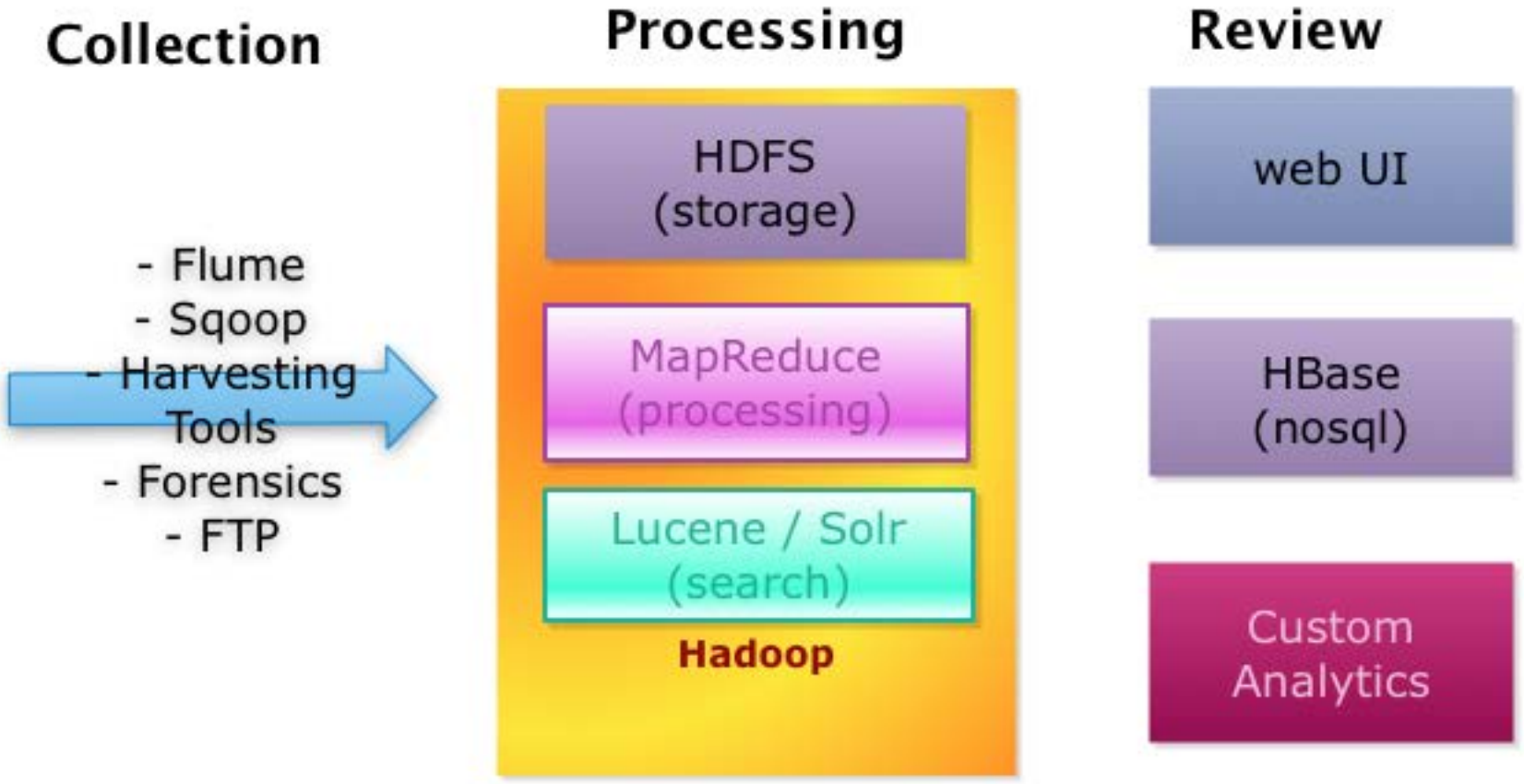


Text indexing / search

- Lucene / Solr is our index / search platform
  - Lightning fast !
  - Open source, proven
- MapReduce jobs create Lucene indexes
- Index files are fed to Solr



Overall Technology Stack



Next : Analyzing Emails

Email : Semi-Structured data

Structure

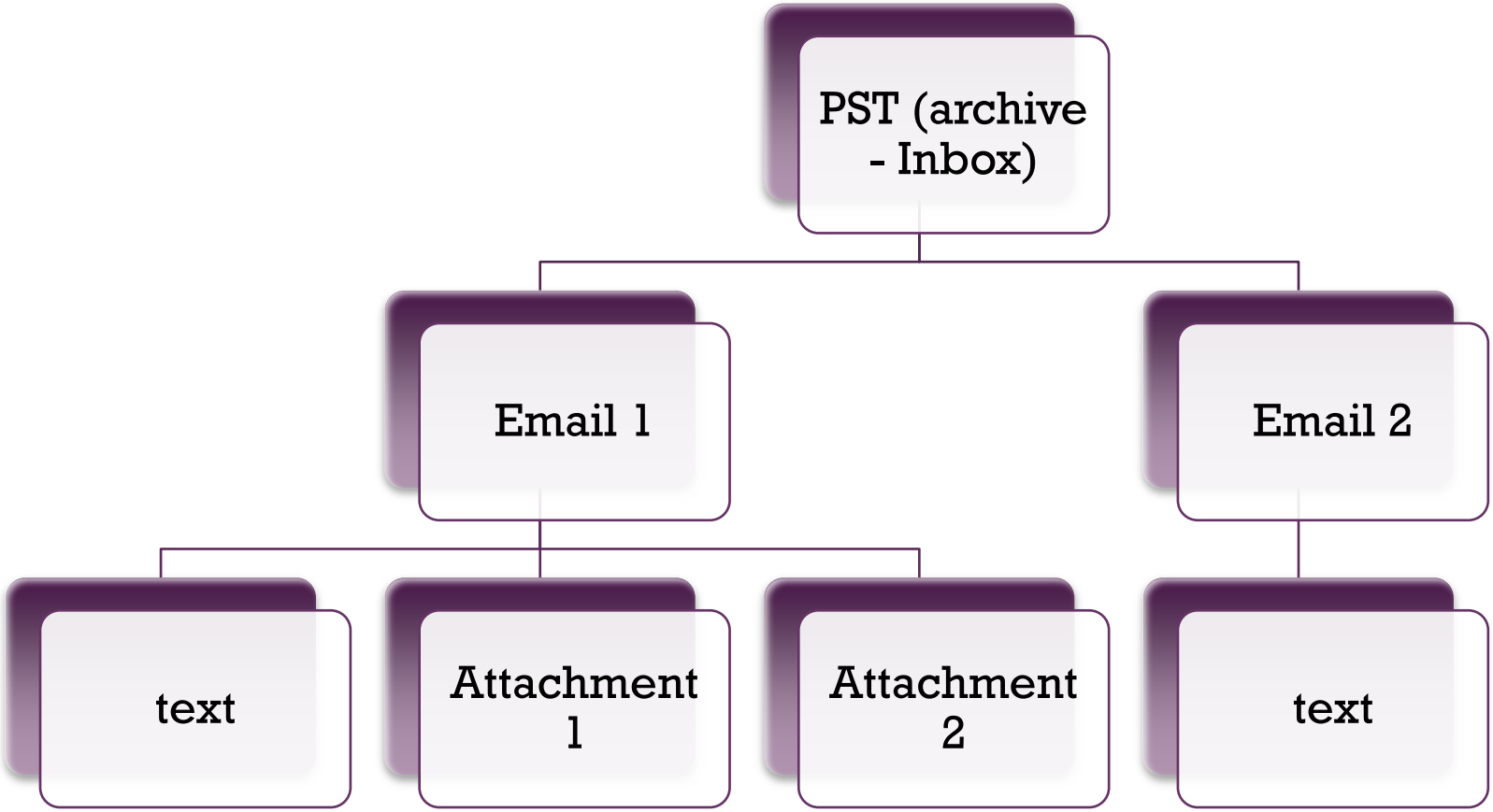
Message-ID: 123.bob@elephantscale.com
Date: Thu, 17 Apr 2014 14:02:01 -08:00 (PDT)
From : bob@elephantscale.com
To : sujee@elephantscale.com, mark@elephantscale.com
Subject: Project proposal – v2
…
X-tags will follow
….

Text

Hello guys,
see attached file for project
…..

Binary

Email Structure



Email Processing

- Each email content is analyzed
- Email text is extracted
- Attachments are extracted
  - Drilled into (zip file containing word docs)
- Text is extracted from each attachment
- Emails have to de-duplicated

Email Deduplication

- Deduplicate: find documents that are the same
- Use hash signature of emails (MD5 or SHA-1)
- Take advantage of Hadoop MapReduce sorting ☺
- Reduce the number of documents for lawyers' review
- Configurable email hash signature definition
- Preserve the "master" document, mark the position of all "duplicates" in context - may be very important to answers questions like "who knew it?" or "who done it?"

Analyzing Email : Tools

- Microsoft Outlook (PST, OST) unpacking
  - readpst (very stable Linux util)
  - JPST – proprietary driver
- LotusNotus (NSF) files
  - Separate server with LotusNotes client installed
- JavaMail API
- Tika to parse attachments

Next : Cloud Integration

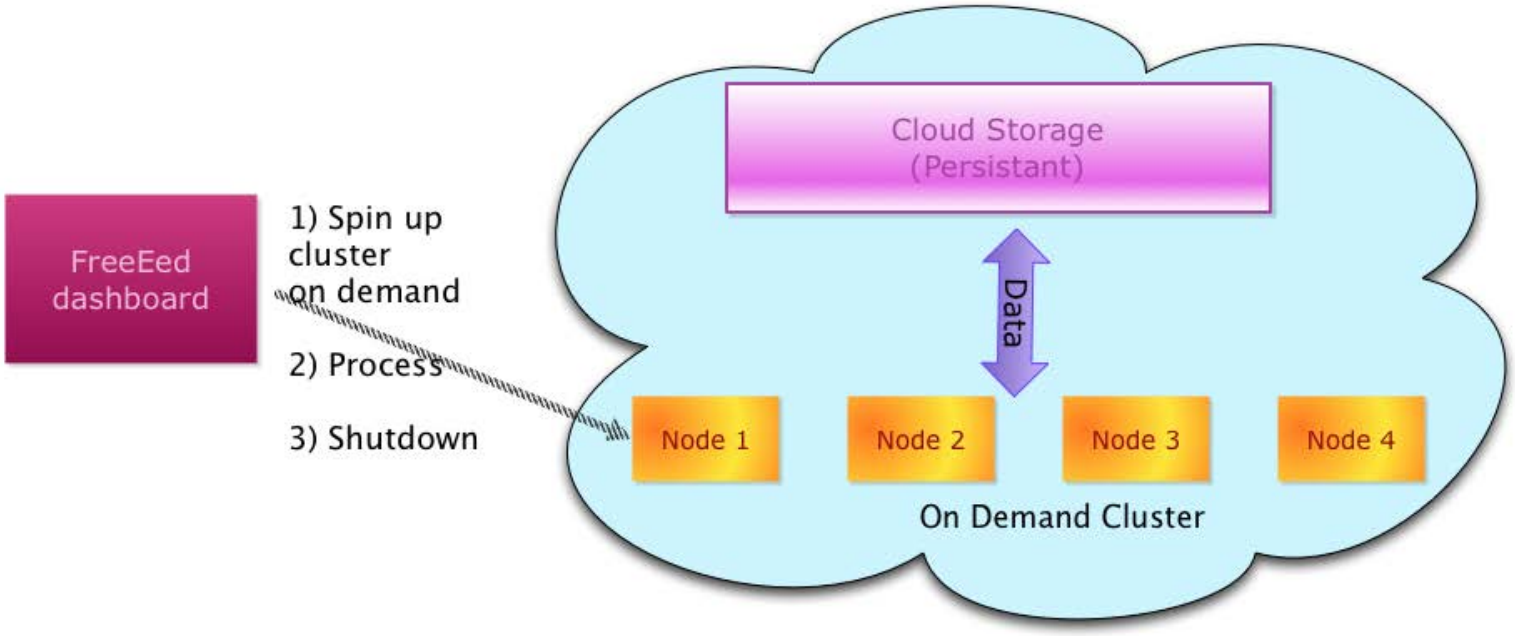
Cloud Integration

- Cloud has leveled the playing field for big data processing
- 'Infrastructure on Demand'
- No need to invest in costly data centers
- Major Cloud vendors
  - Amazon (first!), Rackspace, Google Cloud ….
- Competition is good!
  - Prices for compute and storage keep going down
- Most cloud vendors offer 'on demand Hadoop'
- Perfect fit for FreeEed !

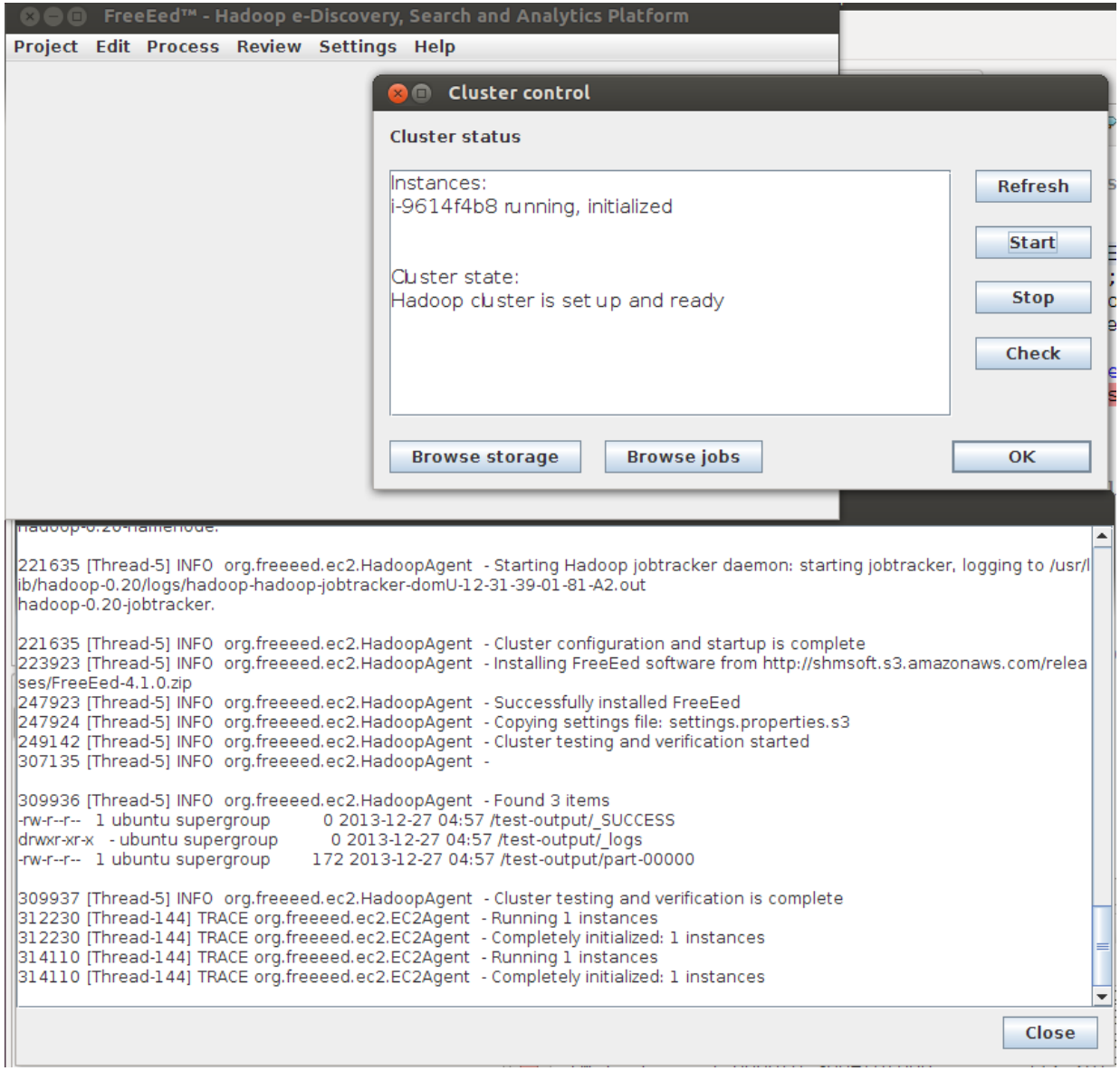
Cloud Cost Model

- Example storage cost on Amazon S3
  - First terabyte : 3c / GB
  - → $30 / TB / month
- Example of Compute Cost
  - Hadoop class machine : 50c / hr
- Crunching Enron email archive on EC2
  - 50 instances
  - 1 hr run time
  - Total expense : $30 !!!

Cloud integration



Cluster start-up on EC2



Cloud Vendor Support

- Native integration with Amazon Cloud (AWS)
- Single click launch / control of Hadoop on Amazon
- Next : Azure (Microsoft)

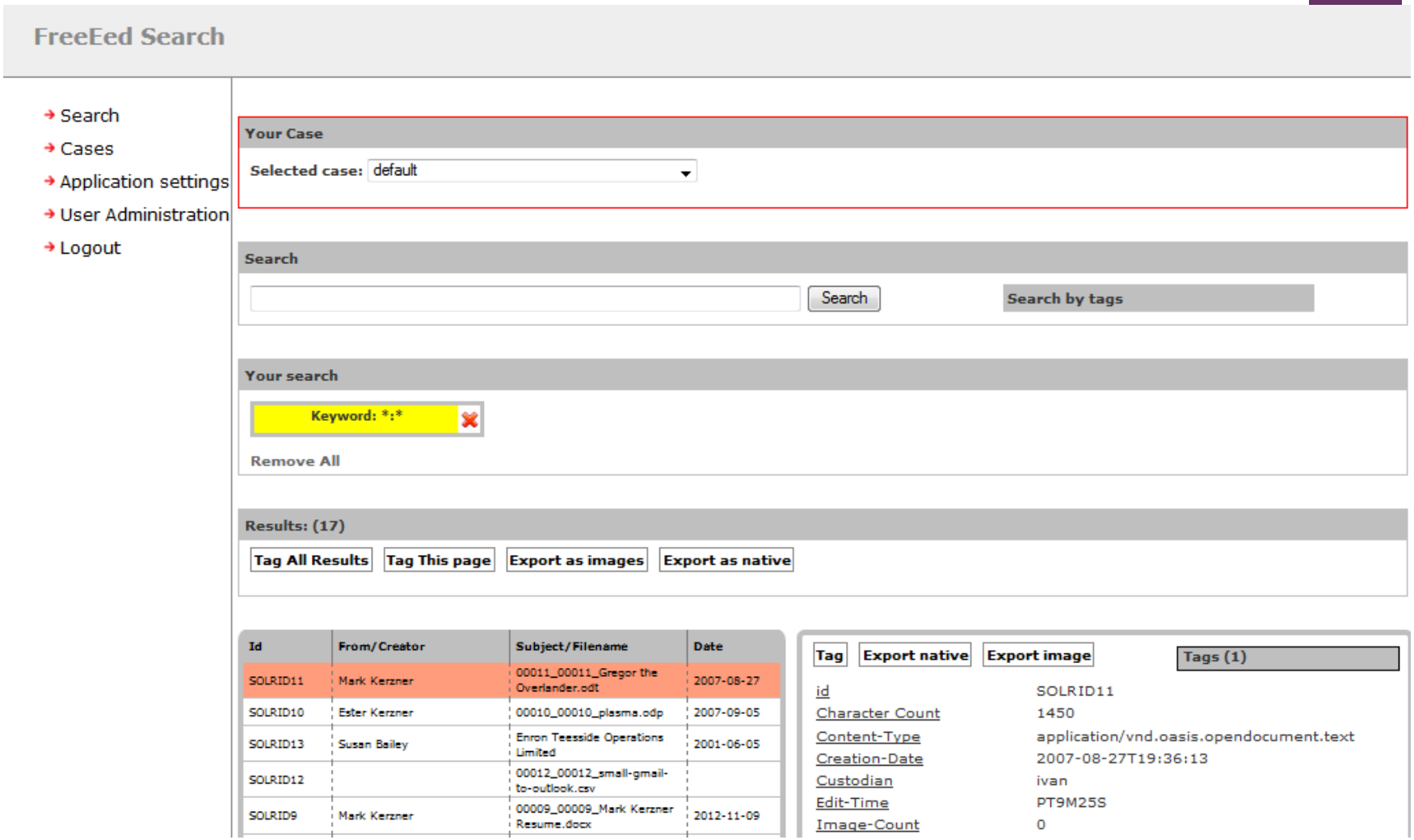
Next : Review Capabilities

- Features
- Processing
- Review Capabilities

Review Capabilities

- Review is an integral part of discovery
- Done by humans.. Need to be friendly ☺
- Search by keywords
- Cull down
- View text and metadata
- Tag documents (for further review)
- Export as images or as native files

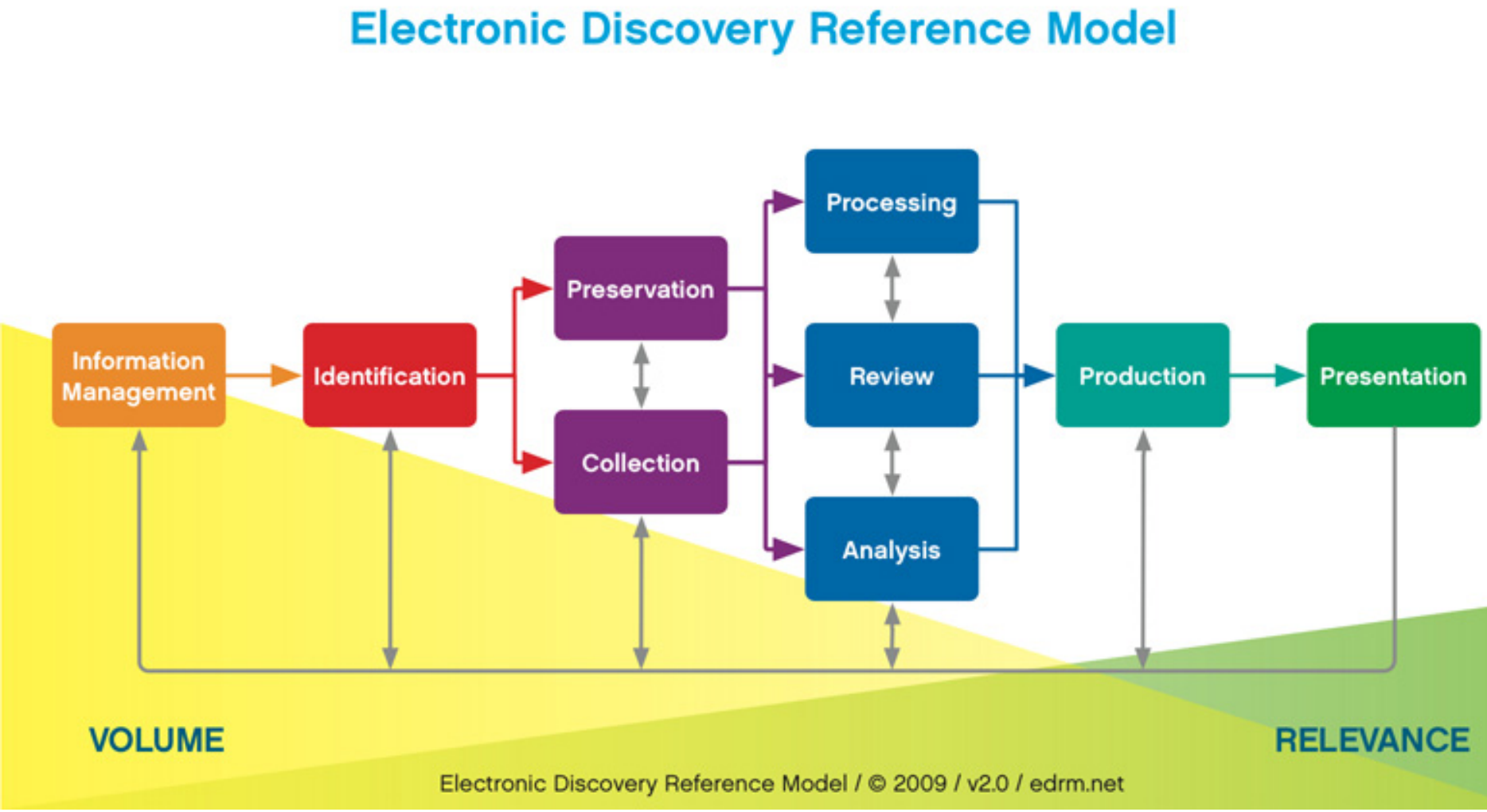
Review screen



Review Architecture

- Lucene / Solr
- Lucene indexes created in reducers and combined in Solr
- Custom GUI providing review / tagging functions

Eagle eye's view - EDRM



Distribution / Packaging

- FreeEed is built on Linux
- But what I don't have Linux?
- Download a virtual machine (VM)
  - Virtual Box container
  - Can run anywhere (Mac, Windows, Linux)
- Adjust VM capacity to fit data needs
  - More data → give more resources (CPU / Memory) to VM

Ease of Use…

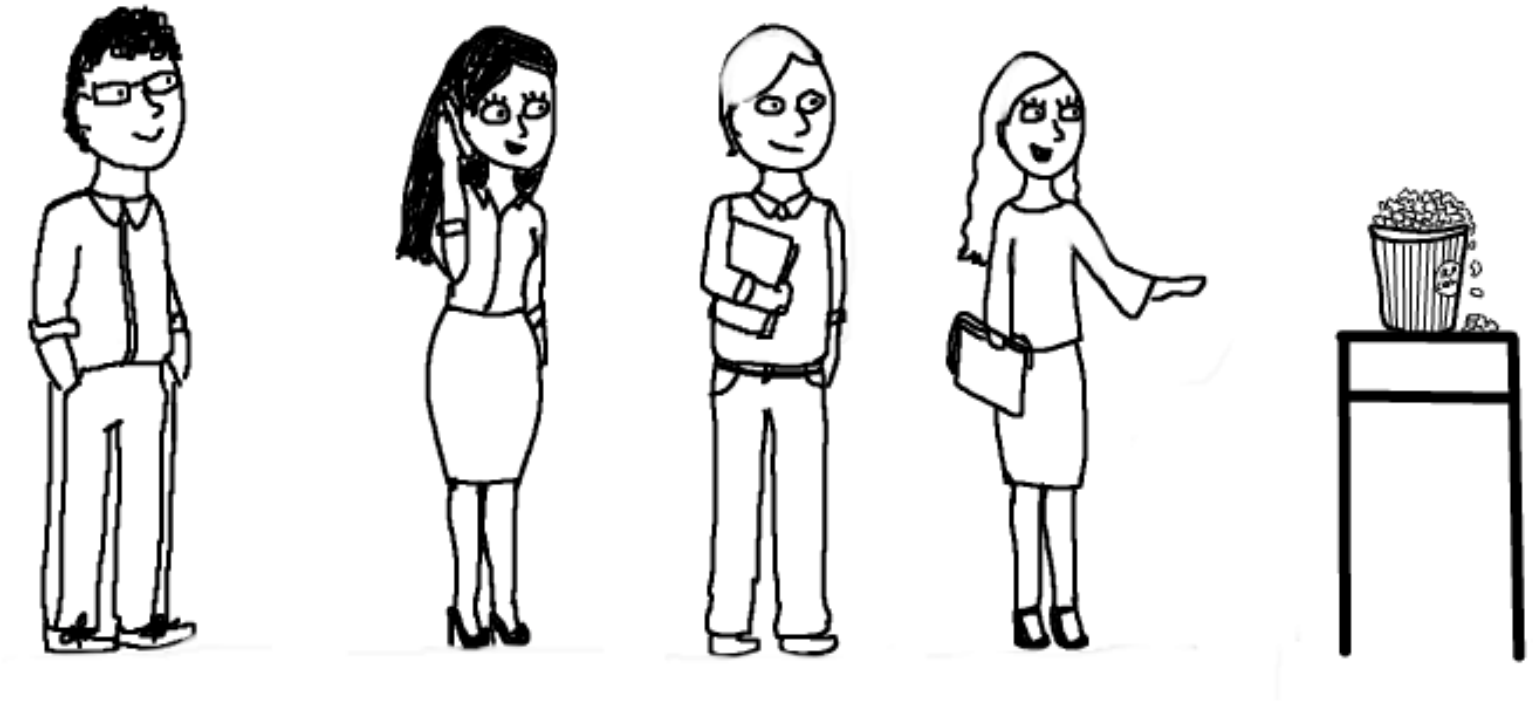
- But I don't know how to use Linux….
- No worries…
  - Project setup is wizard driven by UI
  - Review is done by UI (web based)
- VM is a nice bundle

We are calling it : Popcorn Maker!

FreeEed Popcorn Maker

- Deploy on laptops, servers or cloud
- One-node or any number of nodes
- Scalable storage
- Different cooking recipes
- No mixing of matters
- Easy archiving
- Easy deletion

FreeEed popcorn is very popular with lawyers, legal techs, IT, etc.



Capacity Planning

- Small cases, 5-10 GB: 1 VM, local or cloud
  - Stored and preserved
- Larger cases, 10-100 GB, 1 VM or 3-5 VM cluster
  - Store
  - Process
  - Preserve as cluster
- Large cases, 100GB +
  - Large clusters do it fast
  - Store locally or in the cloud (S3 is good)
- HBase and Lucene index storage is much smaller than data

FreeEed and data governance

- Virtualization for data preservation
- Scalable processing
- Archiving
- Documents groups not mixing
- Data format stored together with software that understands it

Next : Present & Future

FreeEed adoption

MAYER·BROWN

United States Coast Guard
U.S. Department of Homeland Security

haynesboone
Setting precedent.

Supreme Court of Oklahoma

Weatherford

Robinson Bradshaw

Locke Lord LLP

ECLIPSE AEROSPACE

BROWN

Disclaimer: not endorsed by these companies

FreeEed as a learning tool

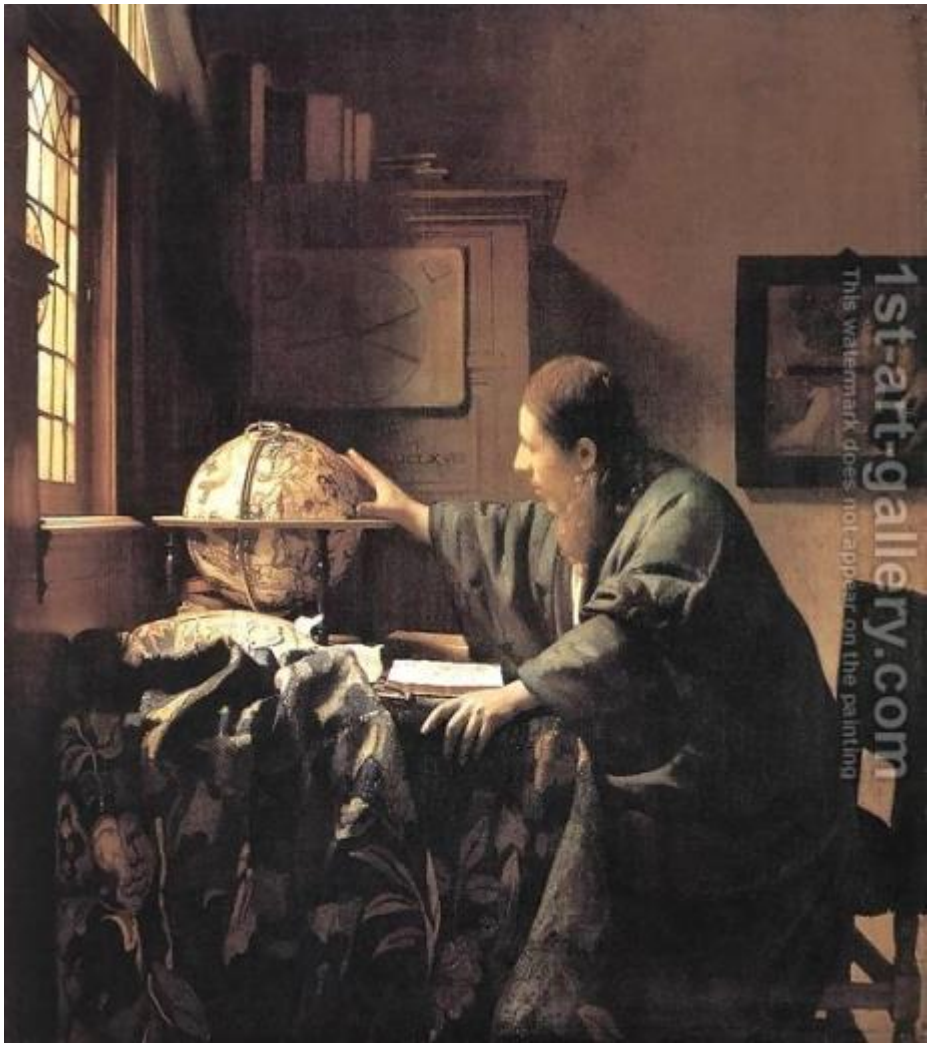
- 100's of downloads
- Dozens of active users
- Real-world Hadoop application
- Many developers download to learn
- Complex, real, but manageable

How you can use FreeEed

- For its intended purpose
  - Large law firms
  - Small firms and solos,
  - Pro-se
- Integrate in the IT legal
- Start a similar document management project

Looking forward

- Add
  - Collection
  - Analytics
- Community
- Integrations
- Implementations



Future Plans

- Use Hbase for storing documents
- Right now our processing is BATCH
- Do real time indexing of emails / documents as they come in
- Move to Solr-Cloud for scale
- Integrate with other cloud platforms

Hadoop & Big Data applications

- Other related applications
  - Financial – text analytics
  - Energy – documents and procedures analytics
  - Actual on-going projects

Q&A

- Thank you!

- Contact :
  - info@elephantscale.com
  - www.elephantscale.com
  - www.freeeed.org