



Engineering Challenges in Legal Technology

Anthony Cassandra, Ph.D.

Director of Engineering

February 1, 2019

Outline








- Who am I?
- Where am I?
- What do I do?
- What are my problems?
- What are your questions?

Who am I?

Education



University of Buffalo	<i>Nothing</i>	
Suffolk County Community College	A.S.	
Stony Brook University	B.S.	
Brown University	M.Sc.	
Brown University	Ph.D.	

Work Experience



machinist, inspector	factory	7+ yrs.	
researcher	industrial lab	3 yrs.	
founder, engineer, director	failed startup	2 yrs.	
adjunct professor	large university	1 yr.	
engineer, researcher	industrial lab	1 yr.	
assistant professor	small university	3 yrs.	
researcher	academic lab	1 yr.	
founder, engineer, CTO	successful startup	8 yrs.	
engineer, architect, manager	medium-sized company	3 yrs.	
engineer, architect, director	successful startup	2 yrs	

Where am I?

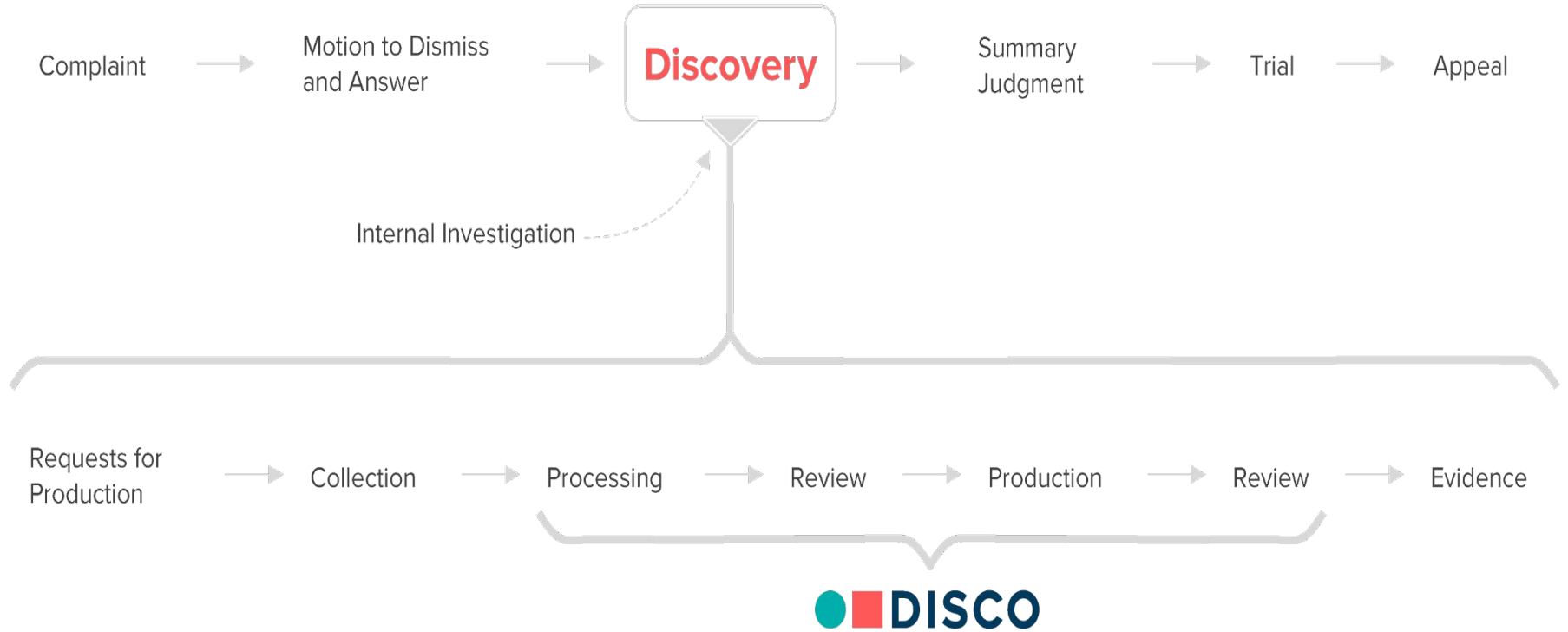
Current Company



- CS DISCO, Inc.
- Legal Technology
- e-Discovery Product
- Cloud-based
- Recent funding round.
 - Expanding to support more of the legal processes.



The Legal Discovery Process



Searching During e-Discovery



DISCO MENU

Search

1 - 10 of 19,144 View: Default Sort by: Similar Count

SIMILAR COUNT	INDICATORS	TAG COUNT	DOC ID	BATES NUMBER	INFO	EXCERPT
<input type="checkbox"/>	EMAIL 0 0 0 0	3	1	EnronDemo040033, EnronDemo087189, EnronDemo080203	DATE 10/20/2017 3:15 PM UTC SUBJECT Do not delete Organizer note - recor...	Sent: Tuesday, January 1, 1980 8:00 AM UTC To: Subject: Do not delete Orga ***** EDMR Enron Email Data Set has been produced in EML, PST a
<input type="checkbox"/>	EMAIL 0 0 0 0		3	EnronDemo080231, EnronDemo040058, EnronDemo087213	DATE 10/20/2017 3:15 PM UTC SUBJECT Do not delete Organizer note - Tabl...	Sent: Tuesday, January 1, 1980 8:00 AM UTC To: Subject: Do not delete Orga EDRM Enron Email Data Set has been produced in EML, PST and NSF forma
<input type="checkbox"/>	EMAIL 0 0 0 0		4	EnronDemo087187, EnronDemo080201, EnronDemo040031	DATE 10/20/2017 3:15 PM UTC SUBJECT Do not delete Organizer note - recor...	Sent: Tuesday, January 1, 1980 8:00 AM UTC To: Subject: Do not delete Orga ***** EDMR Enron Email Data Set has been produced in EML, PST a
<input type="checkbox"/>	EMAIL 0 0 0 0		5	EnronDemo080219, EnronDemo040049, EnronDemo087204	DATE 10/20/2017 3:15 PM UTC SUBJECT Do not delete Organizer note - Tabl...	Sent: Tuesday, January 1, 1980 8:00 AM UTC To: Subject: Do not delete Orga EDRM Enron Email Data Set has been produced in EML, PST and NSF forma
<input type="checkbox"/>	EMAIL 0 0 0 0		6	EnronDemo080208, EnronDemo040038, EnronDemo087193	DATE 10/20/2017 3:15 PM UTC SUBJECT Do not delete Organizer note - Tabl...	Sent: Tuesday, January 1, 1980 8:00 AM UTC To: Subject: Do not delete Orga Enron Email Data Set has been produced in EML, PST and NSF format by ZL
<input type="checkbox"/>	EMAIL 0 0 0 0		7	EnronDemo087205, EnronDemo080220, EnronDemo040050	DATE 10/20/2017 3:15 PM UTC SUBJECT Do not delete Organizer note - Tabl...	Sent: Tuesday, January 1, 1980 8:00 AM UTC To: Subject: Do not delete Orga Enron Email Data Set has been produced in EML, PST and NSF format by ZL
	EMAIL			EnronDemo080214, EnronDe		

Document Reviewing and Tagging



The screenshot displays a document review interface. On the left is a dark sidebar with the following sections:

- 8 of 27** (Page indicator)
- TAGS**: Apply changes to: [icon]. Includes tags: **FERC** x **-99** **Broadband** x, **96** **Enron Outside USA** x.
- PREDICTIONS**: Hide. Includes prediction: **97** **Enron - Oil, Gas or Energy** +.
- RECENT DECISIONS**: Hide. Includes: **FERC**, **Broadband**, **Enron Outside USA**.
- FIELDS (No Data)**: [dropdown arrow]
- ACTIONS**: [dropdown arrow]. Includes: **Add Redaction**, **Add a Note or Privilege Note**, **Folder**, **See History**, **Add to Exhibit Set**, **Download Original** (18.50 KB).
- RELATED DOCUMENTS**: [dropdown arrow]. Includes: **Conversation (5 of 6)**, **Show Attachments**, **Email not collected**.

The main content area shows an email titled **RE: Amending Red Rock Contracts**. The email header is:

From: Porter Gregory J.
Sent: Friday, October 5, 2001 11:24 AM CDT
To: Lokey Teb
Subject: RE: Amending Red Rock Contracts
Importance: Low

The email body contains the following text:

Tab,

I don't disagree with your comment however, my recollection is that one of the main reasons we made the decision we did was because the commercial folks didn't want to have to reopen the agreement with a rate amendment. The concern I though was that such an overture would possibly send the wrong message to the customer or possibly create a perception that the customer had some leverage. My question is, not that we need to amend the agreement for other reasons should we add the discounted rate... do it all at once? Please provide your thoughts. Thanks. Greg

-----Original Message-----
From: Lokey, Teb
Sent: Friday, October 05, 2001 9:09 AM
To: Miller, Mary Kay; Kirk, Steve; Kilmer III, Robert; Porter, Gregory J.; Darveaux, Mary; Hass, Glen
Subject: RE: Amending Red Rock Contracts

At a meeting held 2-3 weeks ago, it was decided that the P&C Agreements modified the terms of the base contracts and that separate rate amendments were not required. I believe this was the consensus of the attendees in both reg and others attended via videoconference in Omaha.

A small tooltip at the bottom left shows: **SUBJECT** FW: Amending Red Rock Contracts

What do I do?

Platform Team Name: Atlas



- **AWS Aurora:** main system of record (SoR) for documents:
 - > 1 billion documents (and growing); and
 - > 125 attributes (+ text) per document (and growing).
- **Elasticsearch:** search is the “heart” of our product:
 - > 140 data nodes (across 4 clusters); and
 - > 150 TB storage across all indices (and growing).
- **Services / APIs:** for insert, update, lookup and search, etc.

Sample of Our Technologies



- Python / C# / Bash
- AWS: Aurora/RDS, S3, ECS, ECR, SQS, Lambda, EC2
- Elasticsearch, Redis, Consul, Celery, Flask, Git
- Datadog, Kibana, Packetbeat, Filebeat, Logstash, Logspout
- Docker, Jenkins, Terraform, Code Pipeline, OpsWorks

Everyone on the team has to be familiar with all of these.

Unique Domain Problems



- Previous retail product search experience:
 - tens of millions of users;
 - simple queries; with
 - low data durability and recall requirements.
- Current legal search experience:
 - thousands of users;
 - highly sophisticated and very complex queries; with
 - very high data durability and recall requirements.

Team's Technical Challenges



- Underlying data model has reached its limits.
- Redesigning infrastructure to scale further.
- Continue to support new features of the review product.
- Supporting new products.
- Scaling team, organization and operations.

These are typical for a startup that has survived.

What are my problems?

Similar Documents Feature



The screenshot displays an email client interface. On the left, a sidebar shows a list of 'Similar Documents (1,405)'. The top of the sidebar indicates '1 of 16' items. Below this, there are tags: 'testing', 'Cheetah', and 'Scarab'. A 'Show predictions' button is visible. The main list shows 16 items, all titled 'Schedule Crawler: HourAhead...', with a similarity score of '≥ 95%' and a count of '(309)'. Each item has a '97%' similarity score and a small icon. The right pane shows an email preview for 'EMAIL Schedule Crawler: HourAhead Failure'. The email header includes 'PDF', 'Text', and a download icon. The email body contains the following text:

SUBJECT
Schedule Crawler: HourAhead Failure

SENDER SCHEDULE CRAWLER
Schedule Crawler 01/17/2002 8:36 AM CST

RECIPIENTS
TO: Pete Davis
CC: Bert Meyers; William Williams III; Craig Dean; Geir Solberg

TAGS
Tagging in progress...

Near Duplicate (95-99.9% similar, but not exact duplicate)

★ Inclusive Email

Schedule Crawler
Friday, January 18, 2002 3:36 AM CST
Pete Davis
Bert Meyers; William Williams III; Craig Dean; Geir Solberg
Schedule Crawler: HourAhead Failure
Low

18/02; HourAhead hour: 2; HourAhead schedule download failed. M

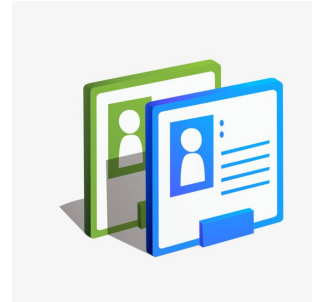
SAGES:
LE --> O:\Portland\WestDesk\California Scheduling\ISO Final Sch

Error: dbCaps97Data: Cannot perform this operation on a closed database
!!!Unknown database.
Alias: dbCaps97Data
!!!Unknown database.
Alias: dbCaps97Data
!!!Unknown database.
Alias: dbCaps97Data
Error: dbCaps97Data: Cannot perform this operation on a closed database
!!!Unknown database.
Alias: dbCaps97Data
!!!Unknown database.

The Similar Documents Project



- Original quick-n-dirty, start-up version:
 - poor quality; and
 - does not scale.
- Replace with a better version:
 - shingling;
 - min-hash; and
 - locally sensitive hashing (LSH).



Base Requirements



- Pagination and sorting requirements.
- Strong ($< 100ms$) service level agreements (SLAs).

Similar Documents (1,405)		^
<u>≥ 95%</u> (309)		↑↓
Schedule Crawler: HourAhead...	2	97%
Schedule Crawler: HourAhead...	1	97%
Schedule Crawler: HourAhead...	2	97%
Schedule Crawler: HourAhead...	1	97%
Schedule Crawler: HourAhead...	1	97%

Design Challenge

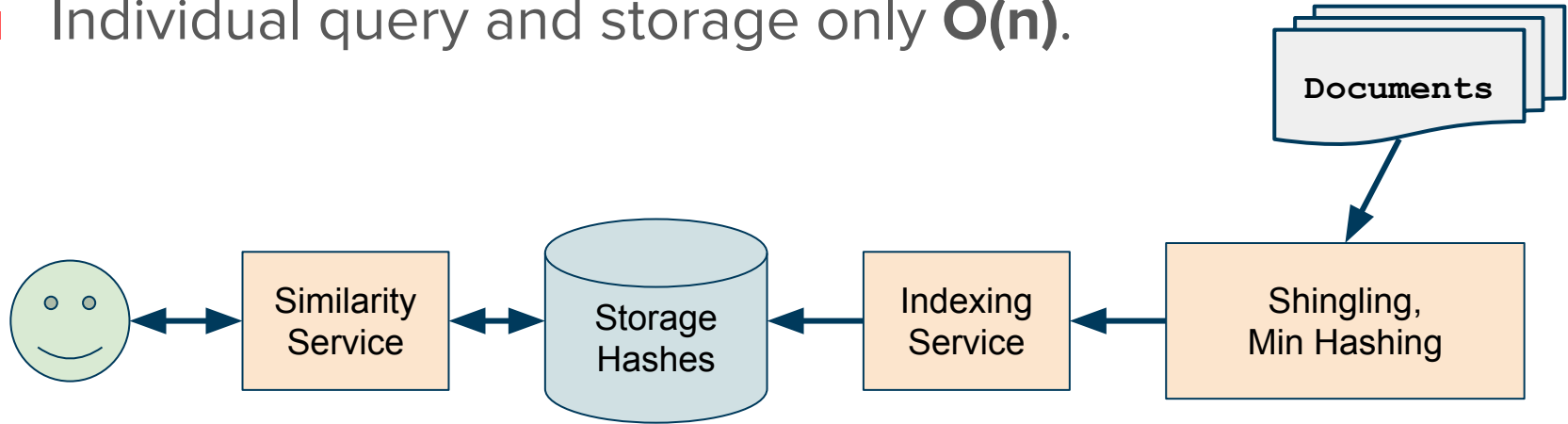


- Leaves us with a problem requiring $O(n^2)$ comparisons.
- We can have n = tens of millions of documents.
- Our CS training tells us to avoid anything $O(n^2)$.

The “Right” Solution



- Pre-compute and cleverly index min hashes.
- Do comparisons dynamically at query time.
- Individual query and storage only $O(n)$.



Problem I - Search Sorting Requirements



The screenshot shows a search interface with a dropdown menu open. The menu is titled "Search columns" and contains the following options:

- Subject
- RELATED DOCUMENTS
 - Attachment Count
 - Conversation Count
 - Similar Count (highlighted in blue and pointed to by a red arrow)
- INGEST
 - Sessions
 - Processing Status
 - Ingest Type
 - Processing Info
 - Processing Details

The background interface shows a search results table with columns for "INDICATORS", "INFO", and "FILENAME". The "Sort by" dropdown is set to "Relevance".

Problem I - Search Sorting Requirements



1 - 10 of 2,147		View: Default	Sort by: Similar Count			
SIMILAR COUNT	INDICATORS	TAG COUNT	DOC ID	BATES NUMBER	INFO	
<input type="checkbox"/> 22	WORD 0 0 1 22	2	1	EnronDemo040033, EnronDemo087189, EnronDemo080203	DATE 10/20/201 SUBJECT Do not	
<input type="checkbox"/> 11	PDF 0 0 2 11	2	3	EnronDemo080231, EnronDemo040058, EnronDemo087213	DATE 10/20/201 SUBJECT Do not	
<input type="checkbox"/> 11	PDF 0 0 2 11	2	4	EnronDemo087187, EnronDemo080201, EnronDemo040031	DATE 10/20/201 SUBJECT Do not	
<input type="checkbox"/> 9	PDF 0 0 2 9	2	5	EnronDemo080219, EnronDemo040049, EnronDemo087204	DATE 10/20/201 SUBJECT Do not	
<input type="checkbox"/> 9	PDF 0 0 2 9	2	6	EnronDemo080208, EnronDemo040038, EnronDemo087193	DATE 10/20/201 SUBJECT Do not	

Problem II - Search Syntax Requirements



DISCO | MENU

custodian(mary) & similarCount(5)

1 - 2 of 2 | View: **Default** | Sort by: **Relevance**

INDICATORS	TAG COUNT	DOC ID	BATES NUMBER	INF
<input type="checkbox"/> PDF 0 0 2 5	2	1	EnronDemo040033, EnronDemo087189, EnronDemo080203	DAT SUB
<input type="checkbox"/> PDF 0 0 1 5	2	3	EnronDemo080231, EnronDemo040058, EnronDemo087213	DAT SUB
			EnronDemo087187, EnronDe	

Bottom Line



- We *must* pre-compute similar documents (and counts).
- It is the only way we can:
 - efficiently sort by counts; and
 - know the counts while searching.
- Needs $O(n^2)$ computation.
- Needs $O(n^2)$ storage.

Worst Case Analysis



- Documents in a legal case:
 - 3×10^7
 - = 30,000,000 (30 million)
- Possible similar document relations:
 - $(3 \times 10^7) * (3 \times 10^7)$
 - = 900,000,000,000,000 (900 trillion)



Strategy



- Raise the Alarm!



- Talk to the Product Manager!

Business Reaction



- Worst case analysis does not impress anyone.
- Business is fuzzy: it is about risks and trade-offs.
- Needs to be a practical problem, not a theoretical one.

Practically Speaking ...



- Average case is much, much lower.
- Average case is within in realm of possible.
- But still need limits on a per-document basis still needed.
- Worst case for design is still hundreds of millions.
- Back to the Product Manager for more “negotiations”.

This is What Success Looks Like



- Requirement Changed: 10K maximum per document.
- New worst case:
 - $(3 \times 10^7) * (1 \times 10^4) = 300,000,000,000$ (300 billion)
- Ain't that better?
- Average case estimates:
 - 10's of millions for large legal cases.
- What if we are wrong?

What Next?



- How to generate all the similar docs relations?
 - Answer: Parallelize and distribute (by different team).
- But final “counts” require global knowledge.
 - Must collate after parallel computation is done.
- How to do all this “fast” enough?
 - The customer is waiting.
- How to store all this data?
 - There’s going to be a lot of it.

Data Storage Choices



- AWS Aurora?
- AWS Dynamo?
- Apache Cassandra?
- Other key-value stores? (e.g., Mongo, Redis)
- Other columnar stores? (e.g., Vertica)

Dangers in Choosing Technologies



- Engineers' Biases:
 - Newest
 - Coolest
 - Most interesting
 - Most familiar
 - Theoretical best
 - Resume building

Practical Considerations



- Do we want to introduce a new platform dependency?
- Do we want to introduce a learning curve for the platform?
- Do we have the operational expertise for the platform?
- Amount of maintenance is needed by the platform?
- Time to market considerations?
- Tends to be a buzz-kill for younger engineers.
- All these led to benchmark Aurora vs. Dynamo.

Performance and Cost Testing



- Both Aurora and Dynamo should be able to do the job.
- AWS has different cost models and different performance:
 - Annoying amount of time spent dealing with pricing.
 - CPU, read/write request, GB stored, data transfer, etc.
- Need to tweak and benchmark them to meet SLAs.
 - Hard to know the right price point without testing.
- Only then can we determine the costs and choose.
 - Spoiler Alert: Aurora Won

Can we Afford It?



- Cost of Goods Sold (COGS): important business metric.
- “Cost” in this context is our AWS bill.
- We charge customers by their data size (GB / month).
- Our financial models depend on estimates of costs per GB.
- Material change in COGS is of great interest the business.

Now, we do the math ...



- The storage cost can grow considerably, but:
 - the more similar docs storage we need,
 - the more documents there are, and
 - the more we would be charging.
- This adds much less than 1% to costs per GB.
- i.e., Storage truly is cheap.

Now We Build It



- Distributed System Design
- Interface Definitions
- Logging
- Monitoring
- Alerts
- Metrics / Dashboards
- High Availability (HA)
- Disaster Recovery (DR)
- Deployment plan
- Release plan
- Test plan
- Documentation
- Component diagrams
- Sequence diagrams
- Architecture Review
- Configuration
- Project task breakdown
- Time estimates
- Cost estimates
- Cross-team coordination
- Scheduling

Solution



- An expensive, inelegant and inefficient system, but it
- enables features that will be a competitive advantage and
- will not materially cut into our profit margins.



A Conclusion



- There is no “right” way to build a piece of software.
- Business context dictates what is “right”.
- The business context will change as the company grows.

“Today’s bad hack was yesterday’s good decision.”

What are your questions?

Thank You

www.csdisco.com