

## **Inference 4.0: Accelerated Review and Analysis**

April 2009



Review and analysis have always been the most challenging and expensive components of the eDiscovery process. The reason is simple: massive amounts of digital data require armies of reviewers. These reviewers operate in-house and in external law firms, poring over reams of electronic views. The results are missed opportunities for early planning, a huge bill for corporations, and precious little control over the entire eDiscovery process.

The solution is to gain highly effective visibility and control into the entire review and analysis process. By applying powerful analytics and review tools from early case assessment to production, the corporation will see very significant ROI. They will be able to hire fewer reviewers, speed up the process significantly, render more accurate and earlier review results, and efficiently oversee even the largest matters.

Inference Data provides this level of control to corporations and law firms with features for early assessment, case preparation, and accelerated review and production. By quickly targeting key documents and speeding review rates, customers realize significant cost savings from beginning to end of the review process.

### **eDiscovery Review Challenges**

eDiscovery is hard enough to do these days in light of large data sets and onerous civil rules and procedures. And the larger the individual matter, the harder it becomes. Let's take a quick look at some of the major challenges surrounding large cases.

eDiscovery search software can help execute the process of searching but bad queries are still bad queries. One of the sticking points in eDiscovery search is that many legal professionals are fixated on the classic keyword search. However, poorly chosen keywords and keyword-only searches yield less than ideal information. And many attorneys still wield keywords

like a weapon, attempting to force opposing counsel into spending even more money and time to search hundreds of keywords that may or may not return relevant documents.

This impacts the eDiscovery process at all stages, including early case assessment or ECA. In fact, "early" case assessment often happens anything but early. This is because many attorneys still lack analytic tools that will allow them to quickly and accurately review data collections *before* the meet-and-confer and search term negotiations. This means that attorneys are going into pre-trial meetings without a handle on eDiscovery results, even though this is the time to set search parameters and terms.

**P R O D U C T P R O F I L E**

Simplistic search and analysis also impacts the corporation's ability to leverage eDiscovery processes. The average enterprise handles 200 litigation cases or more throughout the year. Many of these cases involve the same universe of relevant data, so it makes perfect sense to replicate processes across large and multiple matters. However, traditional review and analysis platforms handle matters on a one-off basis. The corporation that wants reusable and sustainable processes for ancillary matters is out of luck.

*"This Opinion should serve as a wake-up call to the Bar in this District about the need for careful thought, quality control, testing, and cooperation with opposing counsel in designing search terms or 'keywords' to be used to produce emails or other electronically stored information... it appears that the message has not reached many members of our Bar."*

*– Magistrate Judge Andrew J. Peck*

Scalability – or the lack thereof – also looms large as a challenge these days. Scalability used to be IT's provenance, but a poorly scaled review platform will slow to a crawl around multiple review teams and large data collections. This is a growing

problem as ESI grows larger and larger, and traditional review and analysis platforms lack the architecture to catch up. These platforms were originally built to accommodate a single law firm or corporate department reviewing an average-sized data set. They were never meant to scale to modern data set sizes and large numbers of reviewers.

Another serious and growing challenge is the inability to centrally manage large eDiscovery matters. A few years ago this was not much of a problem, since the corporation would simply collect relevant data and throw it over the wall to their law firms for review and analysis. The law firms would undertake the review and analysis and return the produced data along with huge invoices. Now the corporate client is demanding oversight over their cases and law firms, but this level of control has eluded many of them.

Finally there is the challenge of making sense of competing vendor claims around analytics. There is a pitched battle going on between some companies who are founded on the left-hand side of the EDRM and those who are founded on the right. The left is the provenance of Information Management, Identification, and Preservation/Collection. Processing can occur on this side as well but traditionally leads off the right-hand side of the EDRM. It is followed by Review and Analysis, Production, and Presentation.

However, left-hand side analytics also exist

## PRODUCT PROFILE

and can operate as early as the Information Access stage through Identification and Collection. Left-hand vendor analytics claims are therefore perfectly legitimate – until these vendors start asserting that their analytics replace full-blown Review and Analysis toolsets. They do not, and cannot.

### Inference Meets the Challenge

We find that Inference 4.0 meets these challenges and more. Inference is a powerful and highly scalable eDiscovery review platform that serves large data collections and multi-party matters. Inference offers a suite of conceptual search and analytic tools that let reviewers quickly and cost-effectively locate key documents even from large data sets. Inference also provides configurable and collaborative eDiscovery workflow management throughout the process.

We identify its core capabilities as 1) powerful conceptual search and analysis, 2) targeted early case assessment and risk reduction, 3) full eDiscovery review toolset, and 4) exceptionally high processing acceptance and scalability.

- **Core #1: Suite of conceptual search and analytics tools.** Advanced search and analytics technology enables users to effectively review even the largest volumes of electronic documents. Inference technology is highly accurate and repeatable, which significantly reduces overall review time and effort.

- **Core #2: Targeted early case assessment and risk reduction.** Analytics reach back in the eDiscovery process to enable true early case assessment. Reviewers can quickly and accurately run first review passes on even large data collections. Attorneys can confidently enter meet-and-confers knowing what the discovered data reveals – and what it doesn't.
- **Core #3: Complete eDiscovery toolset.** The “review and analysis” phrase sells a lot of eDiscovery products, but some lack even basic review tools. Inference's review toolset is comprehensive and feature-rich, including customizable issues coding, redaction, production, and native review features.
- **Core #4: Exceptionally high scale and search speed.** Inference's hosted architecture uses high volume, distributed indices to handle multi-party matters and large data collections. The distributed architecture enables Inference to maintain consistent speed and responsiveness in even the largest and most complex matters.

### A Deeper Look

We find that Inference is a powerful platform for analytics-centric litigation assessment and accelerated review. We will take a more in-depth look at the platform's capabilities.

**P R O D U C T P R O F I L E**

## **Powerful Conceptual Search and Analytics**

Basic keyword search is still a standard component of the search process, but by itself is inadequate to today's discovery demands. Inference retains Boolean search but expands it to include features like word stemming, proximity searching, wildcard, and parametric field values.

Inference then goes beyond keyword search to proactive conceptual search. Inference conceptual search analyzes data by combining definition, frequency, and contextual placement of metadata and content. This means that conceptual search does not depend on an individual reviewer's skill at constructing queries – which differs wildly among reviewers. Instead, conceptual search analyses the data under review and returns an index of concepts prioritized by percentage of similarity to team members for further review. This offers invaluable guidance for reviewers faced with large sets of potentially relevant data, especially at the early case assessment stage with its severe time restrictions.

For example, let's contrast both types of search using the concept of "backdated stock options." Boolean search will be able to find instances of the word "options." This keyword will likely produce large numbers of un-prioritized results along with completely useless ones such as "hotel options" or "lunch options." Reviewers can improve these poor results by using

combinations like "stock options," "backdated options," or "backdated stock options." Some platforms (including Inference) also offer keyword searches in combination with metadata, so the reviewer could choose to search for all of the above but only in data produced by certain individuals. In this case, the more parameters the reviewer adds to the search, the more accurate the search results might be. This level of search is fine as far as it goes and is still standard fare in many cases.

*FRE 502 states "a party that uses advanced analytical software applications and linguistics in screening for privilege and work product may be found to have taken 'reasonable steps' to prevent inadvertent disclosure of protected communications or information."*

Now let's take a look at this same type of keyword search using conceptual search. Using "options" in the sense of backdated stock options, the system can track each time the term is mentioned along with the relevant context. References to the term "options" are returned to the user with a percentage of similarity to the desired concept. Therefore, documents containing the terms hotels or restaurants, in relation to "options" would appear at the bottom of the list if at all, but mentions of options in relation to stocks, backdating, individuals, or compliance would appear prominently

**P R O D U C T P R O F I L E**

in the list. Furthermore, Inference will suggest query refinements, suggestions, related words, terms and concepts, in addition to prioritizing query results for even more accurate search analysis.

This level of search has distinct advantages including highly accurate and targeted results, an automated process for much faster searches, and query help for every level of reviewer. This enables even the most junior attorney to construct powerful queries and also accelerates the entire review process.

### **Complete Review Toolset**

Inference's review tools prioritize and accelerate the entire document review and analysis process. By quickly identifying and clustering key documents, Inference saves millions of dollars in review costs versus traditional linear review.

The entire review process also speeds up enormously by reducing processing time and by quickly identifying and prioritizing key documents for reviewers. Inference ingests at the rate of half a million documents a day and can configure review sites immediately. This enables review teams to start online review within 48 hours of receiving standard load files for upload and indexing. This is in sharp contrast to the usual scenario of hiring law firms and contract attorneys to hit the ground running... as they wait and wait and wait for the review platform to process the data set. Additional tools such as native review, query refinement, hot keys, and

redaction all improve and accelerate the review process.

### **Collaboration and Centralized Management**

When it comes to distributed review teams, collaboration and management have always been difficult issues. Corporations are understandably reluctant to grant access rights to outside reviewers, and managing the workflow among companies and law firms is equally difficult.

A highly efficient and cost-effective solution is to provide centralized management and collaboration using the highest levels of security and control. The Inference model runs on a highly secure hosted platform and offers a browser-based portal that enables multiple case insight and management from a single view. It can concurrently monitor all users connected with a matter or a group of matters. For example, the Inference Metrics portal enables the corporation to manage reviewers in different workgroups and law firms with real-time monitoring and case analysis tools.

Using this model, the customer controls data access permissions without having to admit external law firms behind the corporate firewall. This makes for a compelling argument for a secure hosted platform that extends its benefits across the distributed corporation, its contract attorneys, and its law firms.

**P R O D U C T P R O F I L E**

Security is critical in this model and Inference secures application, network and physical layers. Specific functions include secure transaction via SSL encryption, secure access by IP address identification, real-time virus protection intruder monitoring, and fully redundant hardware and network architecture. The co-locations are secure against physical intrusion and require card key biometric service levels for access. Inference actively monitors and backs up systems and data, and deploys redundant indices for fully secure storage of client work product.

*“I bring to the parties’ attention recent scholarship that argues that concept searching as opposed to keyword searching is more efficient and likely to produce the most comprehensive results.”*

*– Magistrate Judge  
John M. Facciola*

Another dimension of security is user access control and oversight. Access level security enables administrators to customize access so the interface presents different screens and fields depending on user and user group settings, and function settings as well. In addition to individual feature and access level rights, Inference utilizes “closed user groups” to maintain security between and among different levels of users, working groups and firms –

even opposing counsel. This ability to fine-tune permissions provides an extremely high level of oversight and control throughout all tiers of the review process.

**Early Case Assessment and Reducing Risk**

Inference analytics lets reviewers quickly accomplish meaningful review quite early in the eDiscovery process. Traditional ECA happens well after the meet-and-confer because attorneys simply lack the scalable tools to do meaningful early analysis on their data collections. Inference’s scalable indices make huge results sets highly efficient, and can break them down by concept for the reviewers. This enables review teams to correctly prioritize the most relevant sets for early review, making it possible to do accurate ECA in the early stages of a case.

For example, Data Classification works immediately upon upload to de-duplicate, map, and associate data. This results in smaller and more efficient data collections, which accelerate the review process. From the user side, the Concept Suggest feature presents conceptual search recommendations. This helps even entry-level reviewers to quickly and accurately determine concepts found within the data collection, and to review entire data set early on in the eDiscovery process. Other features such as Dynamic Concepts, Show Similar and Clustering also present highly relevant search results to reviewers, allowing them to successfully analyze data from the ECA through Production stages.

## PRODUCT PROFILE

### Unmatched Scale and Search Speed

Inference leverages Autonomy's powerful IDOL engine to achieve high levels of scalability across multiple indices. These index engines efficiently and dynamically update their environments and all queries and analytics occurring inside the indexes via rolling deliveries.

Inference automatically parses queries and sends them to individual engines to spread processing loads. With matters easily approaching the 1,000,000 document mark – and even going beyond it -- it is vital that these large amounts do not impact system availability or performance. Up to 169 engines can be added without disturbing workflow, for a total of 169 million documents being actively stored and processed. Upload speeds of half a million documents per day save significant money and time as reviewers do not have to wait through long load times. Rolling uploads and productions do not disrupt the process.

The Inference platform does not require proprietary processing and is data agnostic. It can accept Concordance-type loads or can ingest delimited load files containing metadata and text. It also supports a broad range of data formats including structured, semi-structured and unstructured data.

### Benefits and Use Cases

#### 1. Scenario: Quickly identify key

**documents for ECA.** A review team was charged with locating all emails associated with four different users across a particular time period. Traditionally the reviewers would have had to work with IT to recover Exchange archives for the periods in question, and use recovery software to locate and view all separate messages concerning those four. Using Inference, the indices ingested Exchange archive data within hours and the reviewers commenced their search. The review team easily returned only those results containing the required users within the queried date ranges. The messages were linked, threaded, and prioritized for immediate results. Using this information, the review team was able to quickly pass sensitive review results to the lead attorneys for early case planning.

#### 2. Customer Scenario: Reduce processing and review costs.

A corporation was astounded at the invoice amounts from its outside law firms. They adopted Inference and directed their law firms to use the hosted platform for the corporation's matters. The corporation quickly saw high ROI in a single large case. During the course of the matter, Inference cut attorney review time from more than 23,000 hours to less than 8000. At the first level of review alone, where attorney salaries are around \$65, this resulted in a savings of more than \$800,000 per case just in the first

**P R O D U C T P R O F I L E**

review tier. The savings expanded when including second and third tier review and higher priced attorneys.

3. **Customer Scenario: Reduce processing time for faster review starts.** Defense attorneys went up against the U.S. federal government. The pre-trial stage required the defense to review and produce a large data collection of more than 300,000 documents, and to do it within 90 days. Making matters worse, these were government documents and hardly models of clarity. The defense used Inference to quickly analyze and prioritize the document set for relevance, which gave them strong leverage in pre-trial activities. The law firm met the deadline in short order and their results favored the client. As a result, the judge excluded an informant's evidence and barred a key plaintiff witness from testifying at trial.

### **Taneja Group Opinion**

Review and analysis activities represent the corporation's biggest eDiscovery pain point. Of all the eDiscovery stages, review and analysis is by far the most costly in terms of expense and resources. It is also the stage that represents the highest risk for adverse judgments and sanctions if not done well. Because of this, "review and analysis" is the buzzword term that loosens eDiscovery budget. Many vendors claim a share of the pie by presenting their search and review products.

However, corporations and law firms must invest in platforms that provide search and analysis well beyond basic keyword search. They must also be certain that the platform can secure and scale large data collections, can greatly accelerate the review process, and can significantly save on review process resources. We find that Inference 4.0 meets these standards and more.

---

*NOTICE: The information and product recommendations made by the TANEJA GROUP are based upon public information and sources and may also include personal opinions both of the TANEJA GROUP and others, all of which we believe to be accurate and reliable. However, as market conditions change and not within our control, the information and recommendations are made without warranty of any kind. All product names used and mentioned herein are the trademarks of their respective owners. The TANEJA GROUP, Inc. assumes no responsibility or liability for any damages whatsoever (including incidental, consequential or otherwise), caused by your use of, or reliance upon, the information and recommendations presented herein, nor for any inadvertent errors which may appear in this document.*