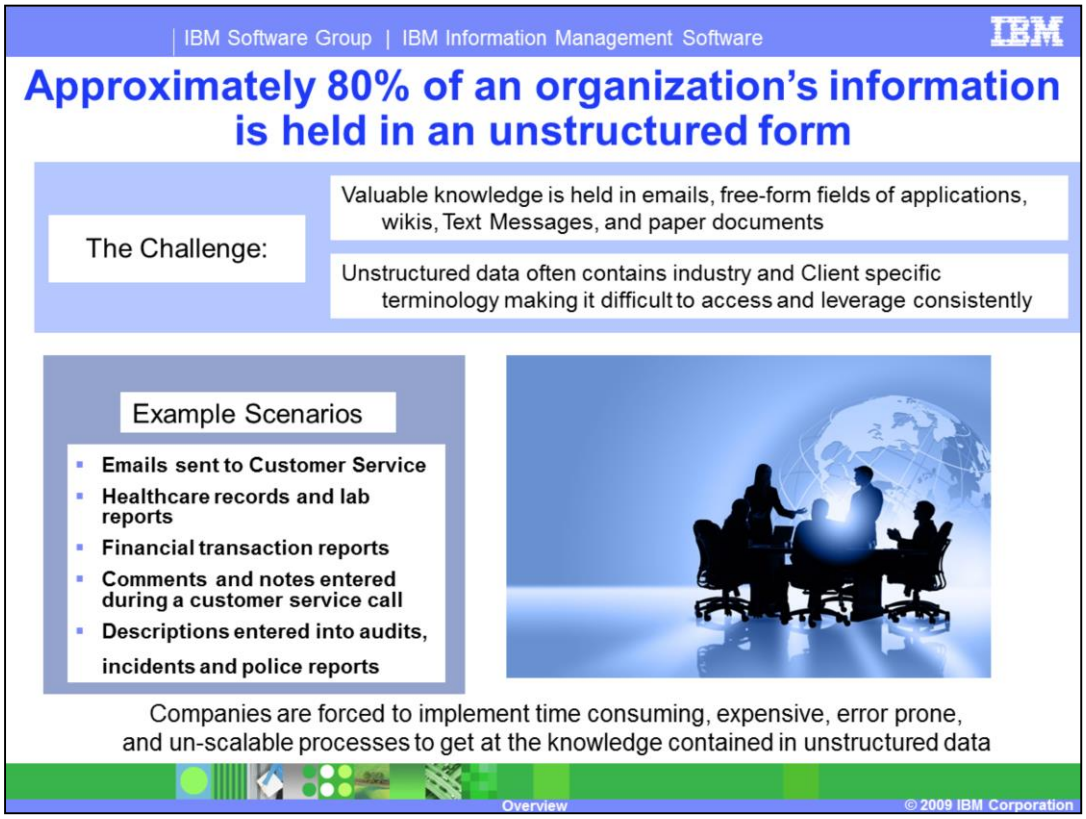
IBM Cognos Content Analytics V2.1

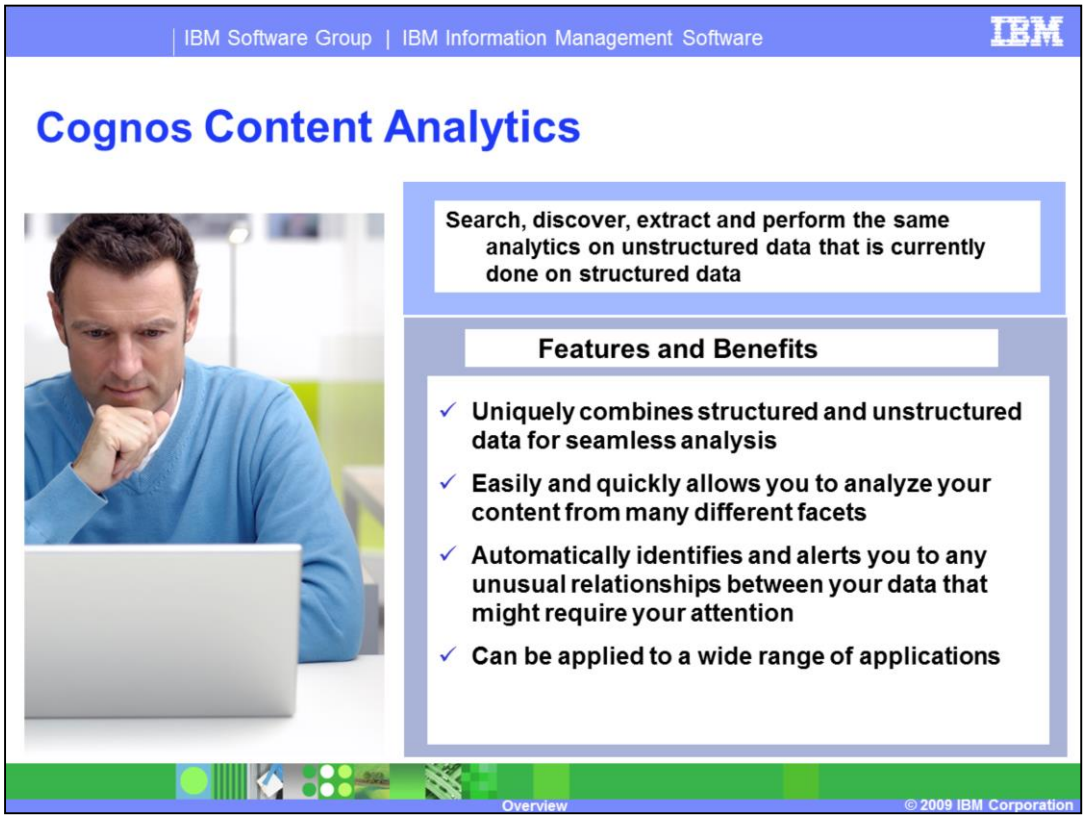
Overview

This presentation is an overview of IBM's Cognos Content Analytics version 2.1. It is intended for the users of Cognos Content Analytics and business and research analysts who want to learn how Cognos Content Analytics can bring value to their business through its unique content analytic features.



A commonly quoted statistic is that roughly 80% of a company's information is maintained as unstructured content, valuable assets such as e-mails, free-form fields on applications, wikis, text messages, and the like. Because this content lacks structure, it is very difficult to realize the benefits of automation when dealing with text.

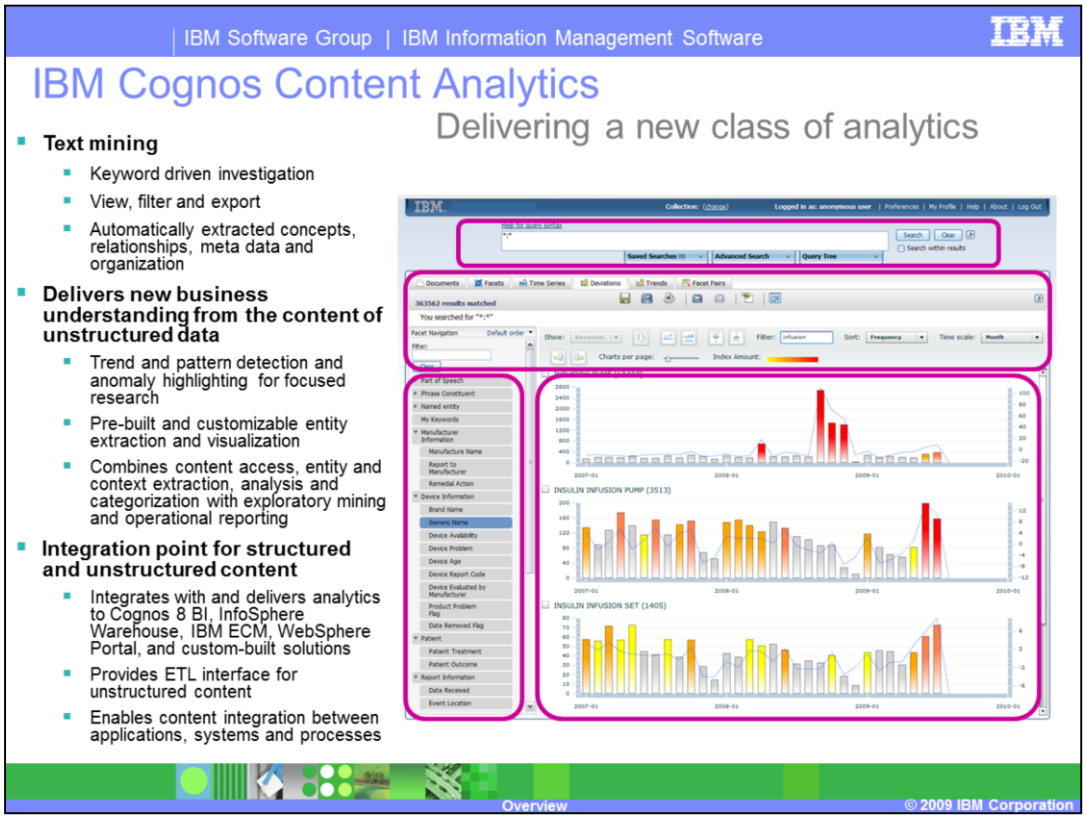
Typically text analysis requires a human being to read and understand what is being communicated. Human involvement can be an expensive and a time consuming component of your overall business process.



Cognos Content Analytics allows you to search, discover and perform the same analytics on unstructured data that is currently done on structured data. This allows you to gain the most from your unstructured content in ways never before obtainable.

The data and services extracted from your unstructured content are now accessible to a broad range of business applications that are independent of the analytics and data loading performed by Cognos Content Analytics.

For example, you can draw actionable conclusions from customer feedback, or improve your quality control process by identifying product defects in a more timely manner. The end result is making the most use out of all your data regardless of its structure.

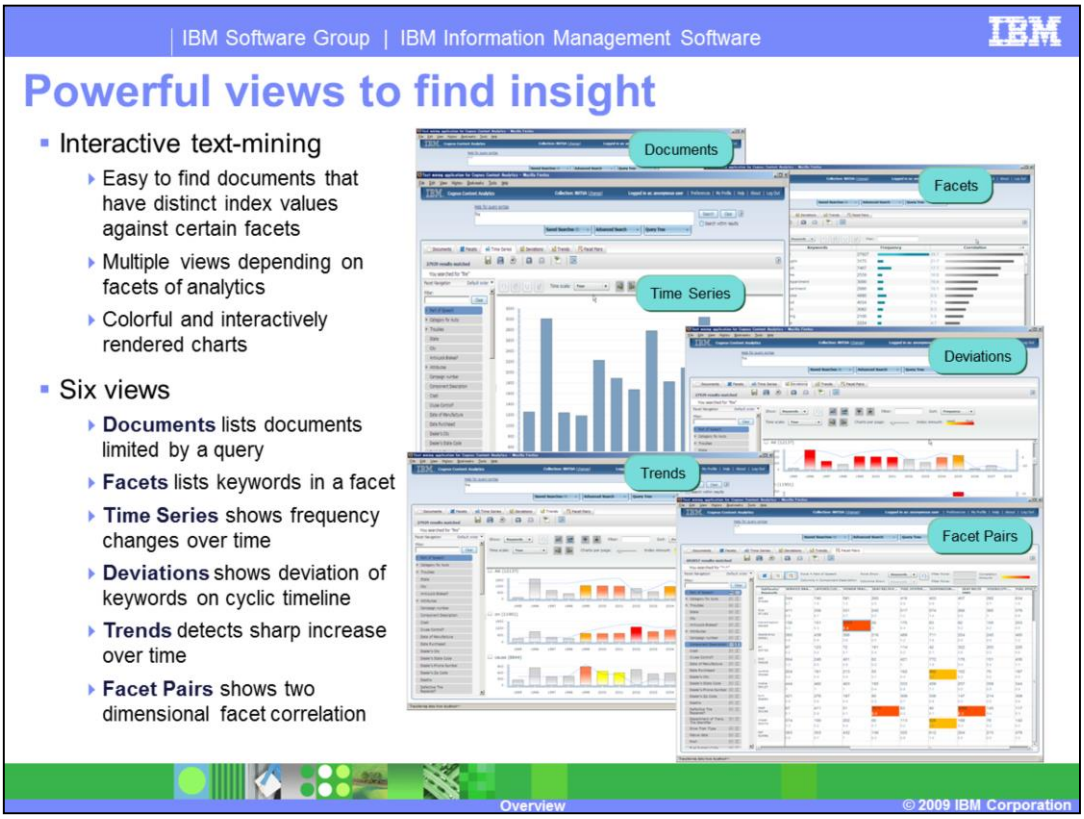


Cognos Content Analytics comes with a highly advanced and easy to use graphical interface that allows you to mine the textual content the same way you normally mine structured data. The interface shown on the right, offers several ways to explore your data, ranging from keyword driven investigation as seen at the top, to multi-faceted navigation that is shown on the left. As you explore your data with these techniques you can, at any time, add your current selections to your exploratory state incrementally building and focusing on only what is important to you. A backward and forward button enables you to quickly back track and try alternative explorations, much like the back button on a browser.

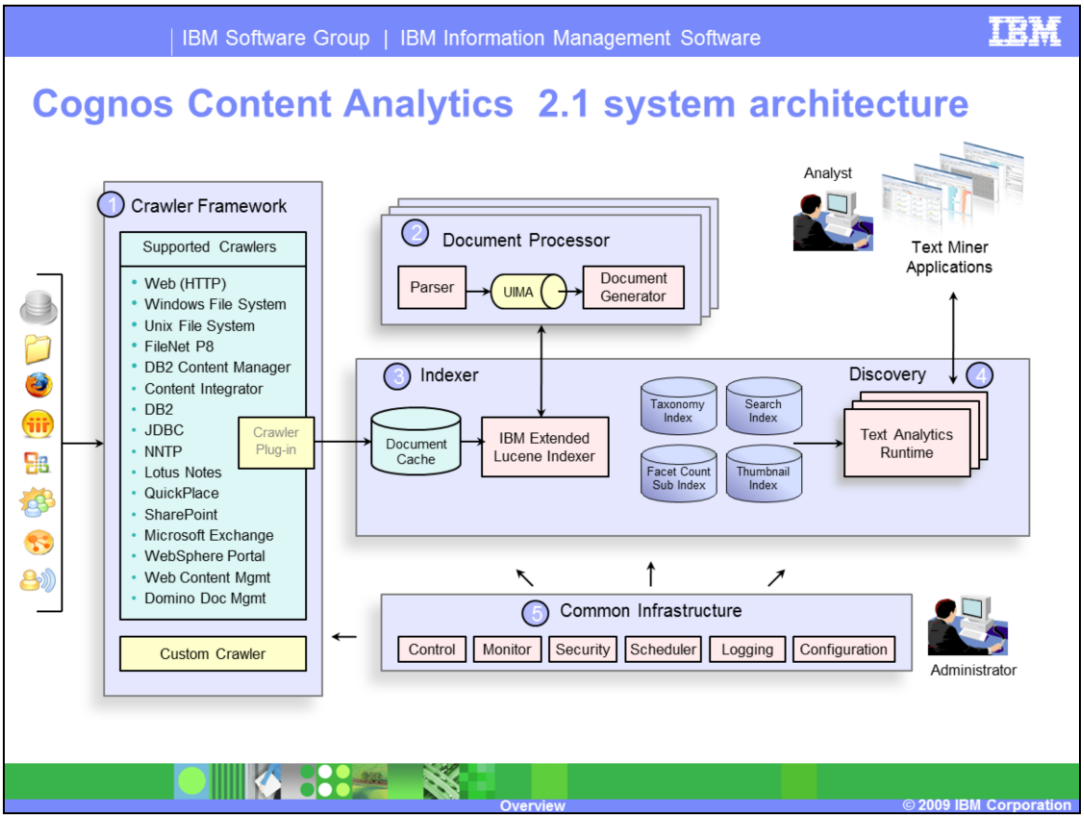
Your data, shown in the middle panel can be viewed in many different ways that provide powerful insight to your data. Switching between these views is achieved by clicking on the tab you want across the top. There is more detail about these views on the next slide.

As you explore your data, Cognos Content Analytics automatically identifies and alerts you with visual queues to any unusual relationships between your data that might require your attention. It is important to note that Cognos Content Analytics not only analyzes your unstructured content but also allows you to include your structured content for combined analysis.

At anytime you can export the results of your exploration for further analysis by other analytic applications such as Cognos 8 BI for operational reporting, InfoSphere Warehouse for cubing, or SPSS for deeper statistical analysis.

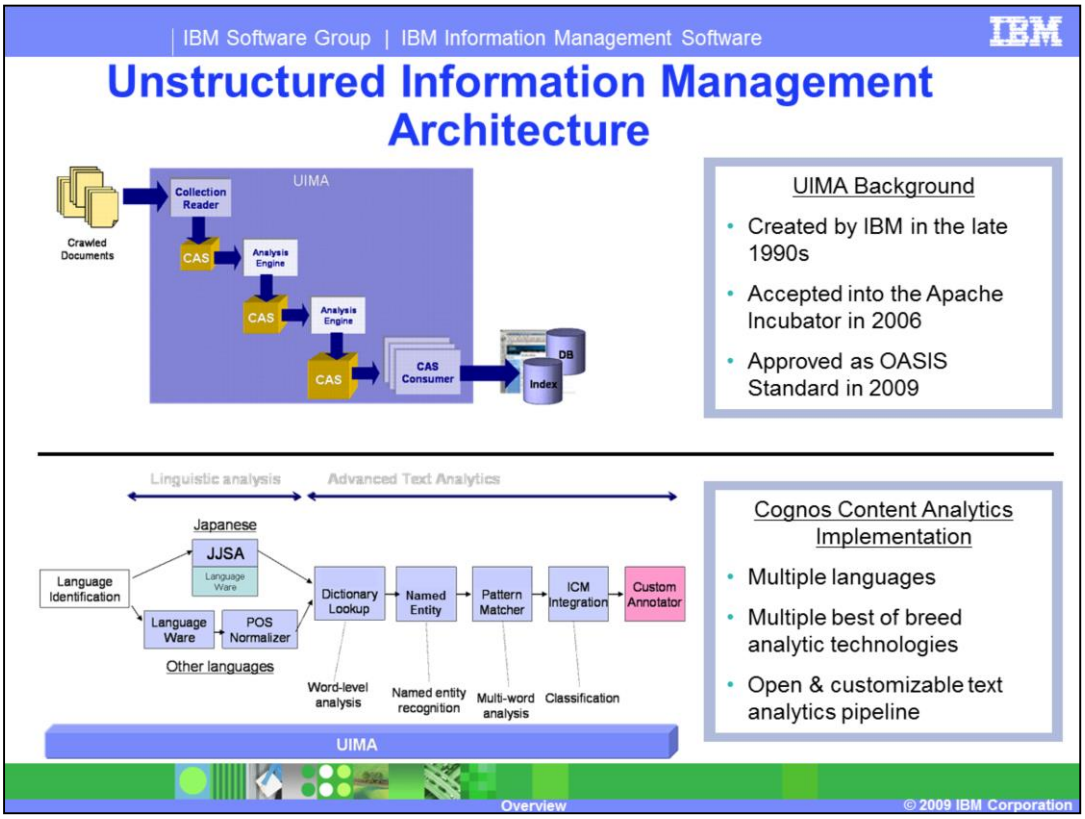


As previously mentioned, your data, shown in the middle panel, can be viewed in many different ways. There are currently six different views each corresponding to a tab across the top. The first view is the Documents view which lists and provides links to the original documents that match the current state of your exploration. This view allows you to quickly retrieve one or more original documents for visual inspection. The second view is the Facets view and requires that a facet be selected from the facet navigation panel on the left. For the selected facet, its values are listed in the middle panel along with counts for how often each value occurs and its correlation value. Correlation values are an important measure of uniqueness for a given set of exploratory factors. They can alert you when a particular facet value occurs more frequently than others when compared to the current exploratory state and its parameters. The next three views are temporal in nature and are useful if your data contains dates for specific events in time. The time series view will show how the currently matched documents are distributed across the values of a given date field in your documents. The deviations view is similar to the time series view but instead is limited to the values of a selected facet. Each facet value is given a separate time series graph for easy comparison and shows the rate of change, or deviation, between each time period. The trends view is similar to the deviations view except that it shows the direction of change, or trend, between each time period. The last view is the facet pairs view which allows you to perform a two dimensional comparison of frequency and correlation values between any two facets. As with the other views, Cognos Content Analytics will automatically alert you with visual queues as to when the combination of two facet values are highly correlated.



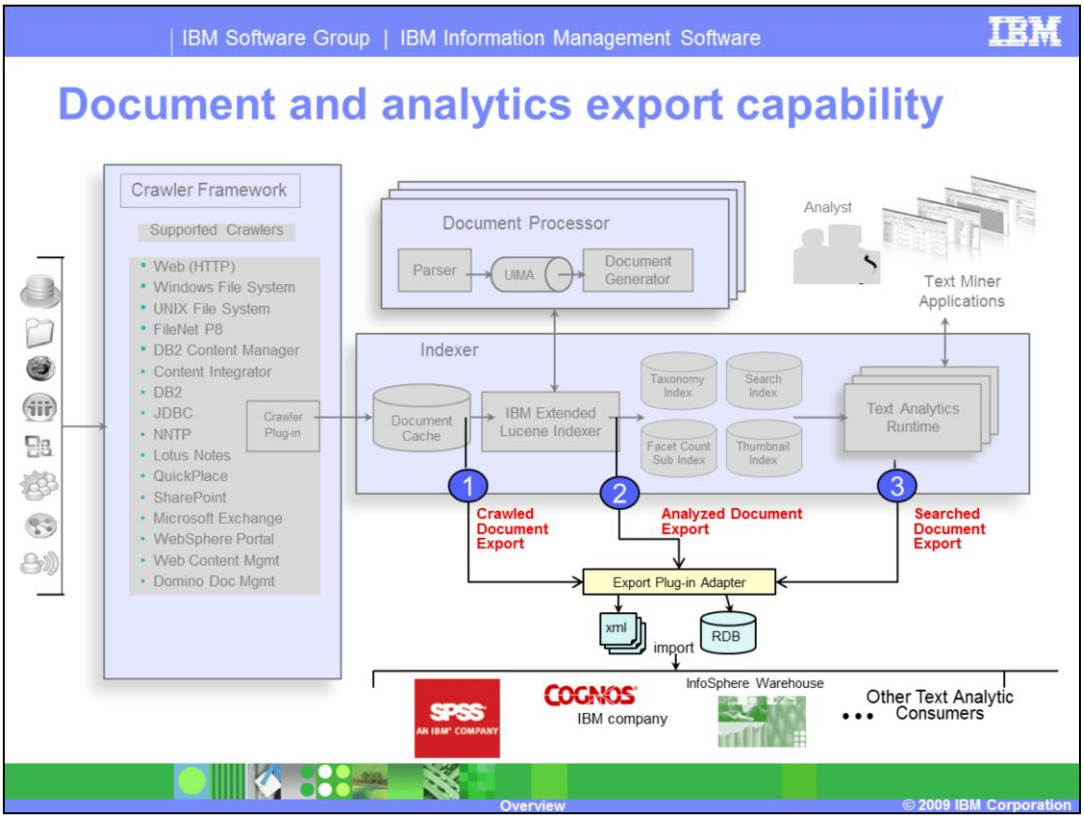
This slide displays the overall system architecture of Cognos Content Analytics. It consists of four key technologies. One technology is the crawling framework for content acquisition from a broad range of enterprise sources. The second technology is one or more document processors for the parsing and analysis of text with advanced analytics for information extraction. The third technology is the IBM extended Lucene indexer for the efficient storage of the document content, analytical results and associated metadata. The fourth technology is the discovery runtime that supports the fast and concurrent multi user exploration of the stored information.

All of these components are managed through a common infrastructure and administrative console.



Adding text analytics to Cognos Content Analytics is achieved using the Unstructured Information Management Architecture or UIMA. UIMA is an open interface that enables the plug and play of your own or vendors components for advanced text analysis. Each text analytic component is referred to as an “Annotator” because it ultimately annotates the document with additional information that it discovers. These annotators can then be plugged into Cognos Content Analytics using the administrative GUI and is subsequently invoked just before the documents are ready to be indexed.

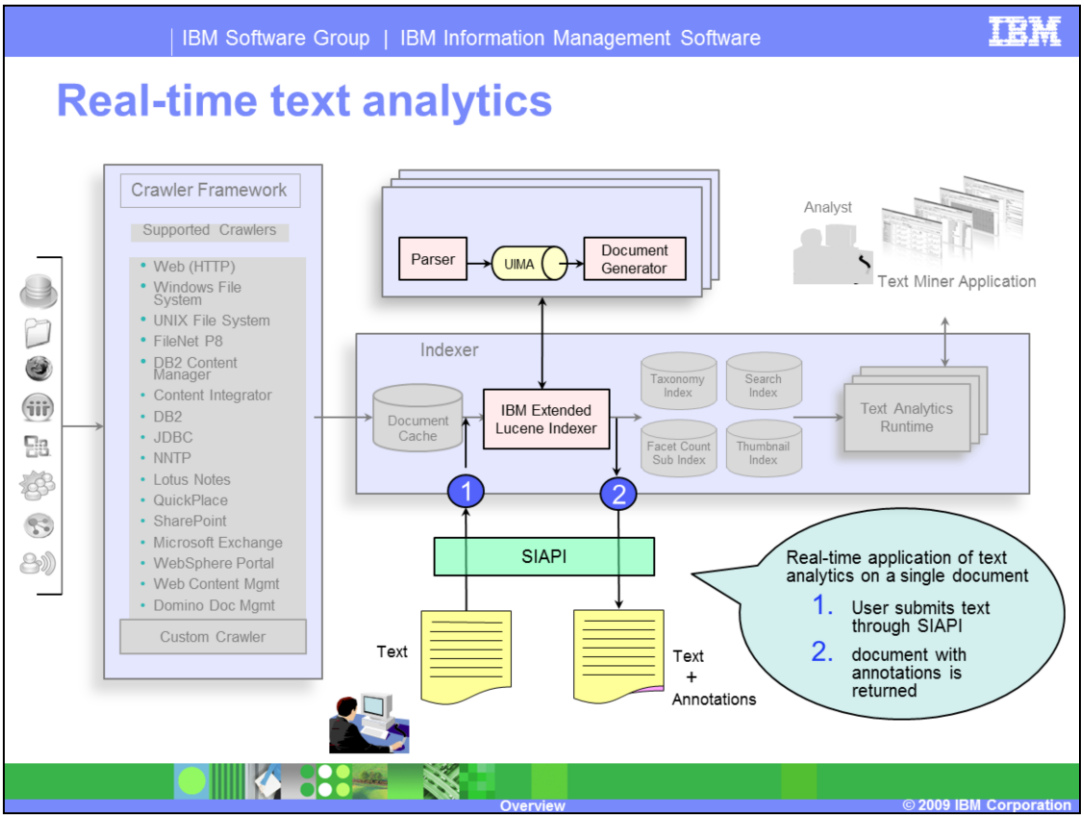
Cognos Content Analytics comes with two annotators that are always invoked. One to identify the language of the document and then a parser that knows how to extract words or tokens depending on the language. Other annotators can be enabled in the chain to perform additional tasks such as dictionary lookup, named entity extraction, and classification using IBM’s classification module. You also have the ability to plug in your own custom annotator.



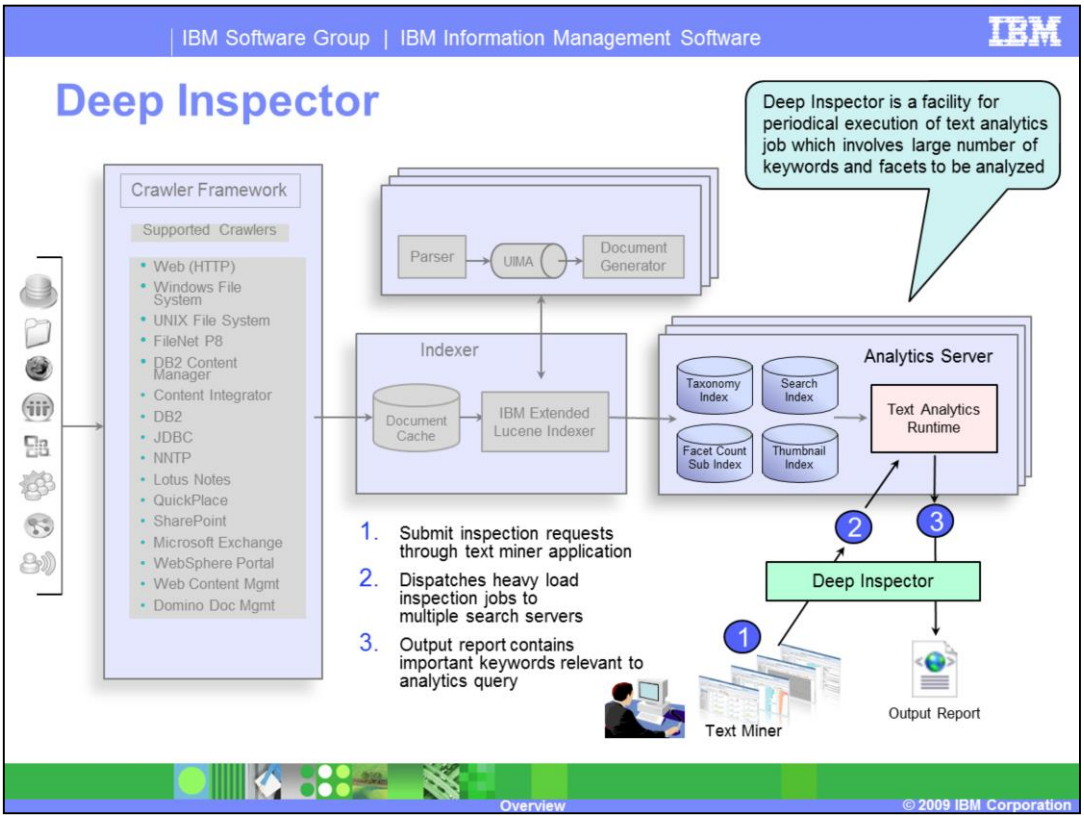
Cognos Content Analytics also has a very flexible export capability enabling it to integrate with other analytic applications such as Cognos 8 BI for operational reporting, InfoSphere Warehouse for cubing, or SPSS for deeper statistical analysis. You have the choice of exporting the content to the file system as XML files or directly into a relational database. There is also an export plug-in interface that allows you to write your own adapter for redirecting exported content to a consumer of your choice.

You also have three points in system architecture from which to export content. The first is after a crawl has been performed. Here the content of the document and its metadata is exported as retrieved by the crawler but before any text analytics have been performed.

The second option is to export the documents after it has been indexed. This includes any of the annotations that have been added to the documents from the text analytics performed in the UIMA pipeline. The third option is to export your search results which contains only the documents, metadata and annotations that match your current exploration state.

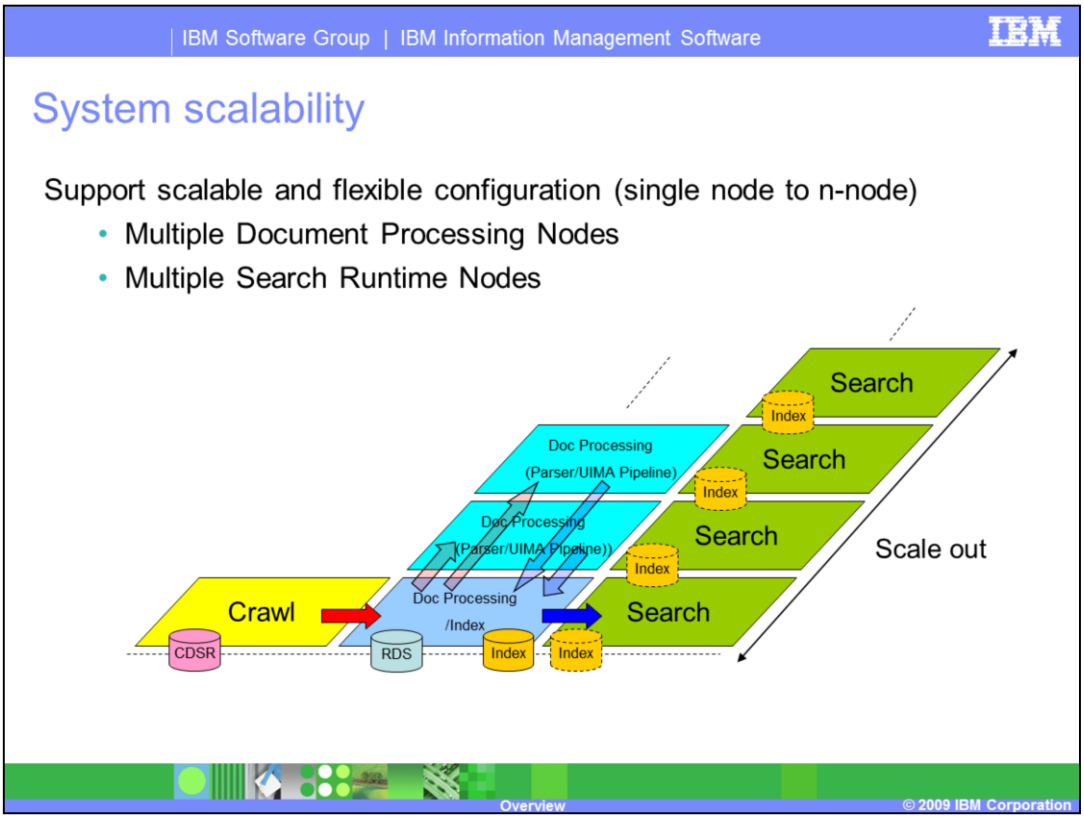


There are certain cases where an application will want to take advantage of the advanced text analytics used by Cognos Content Analytics but without employing the full processing flow as dictated by the system architecture. In this case, an application already has the content to be analyzed and has its own disposition for what is to be done with the results instead of having it indexed by Cognos Content Analytics. Under these circumstances, the application can call one of Cognos Content Analytics' Search and Index APIs, referred to as SIAPI, to analyze a unit of unstructured text. Once analyzed, the results of the analysis is returned. Note that SIAPI is a remote Java API so in this way Cognos Content Analytics can be treated as a kind of Text Analytics service.



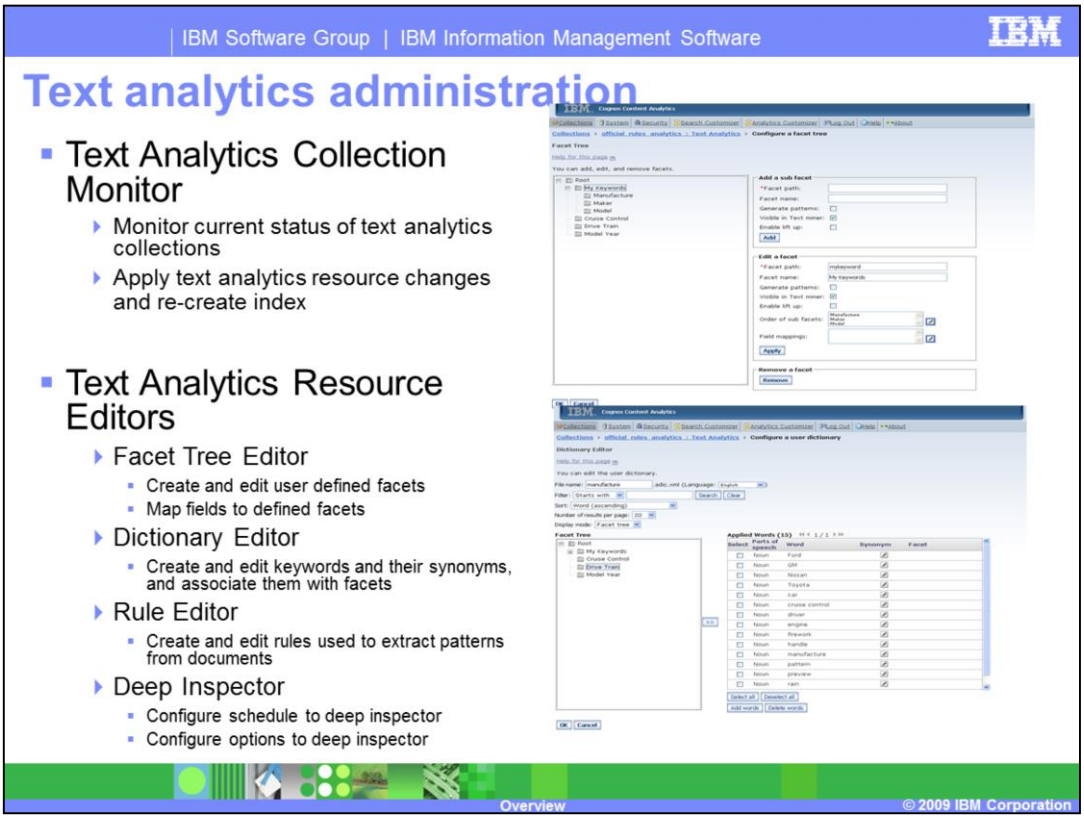
The Text Miner application is an interactive tool that allows you to explore your data in real time. This is possible when the number of facets to your data is 100 or less. As the number of facets increases so does the compute time needed to calculate the correlations across the entire corpus of data. Eventually, you will reach a point where the responsiveness of the text miner is not suitable for real time interaction.

Under these circumstances, where there is a large number of keywords and facets to be analyzed, you have the option to invoke the Deep Inspector which will submit to you an analysis request in batch mode. The Deep Inspector takes care of dispatching your request across the multiple analytic search servers that you have configured and monitoring them for the completion of their respective tasks. Upon completion of the entire request, an output report of the analysis is delivered.

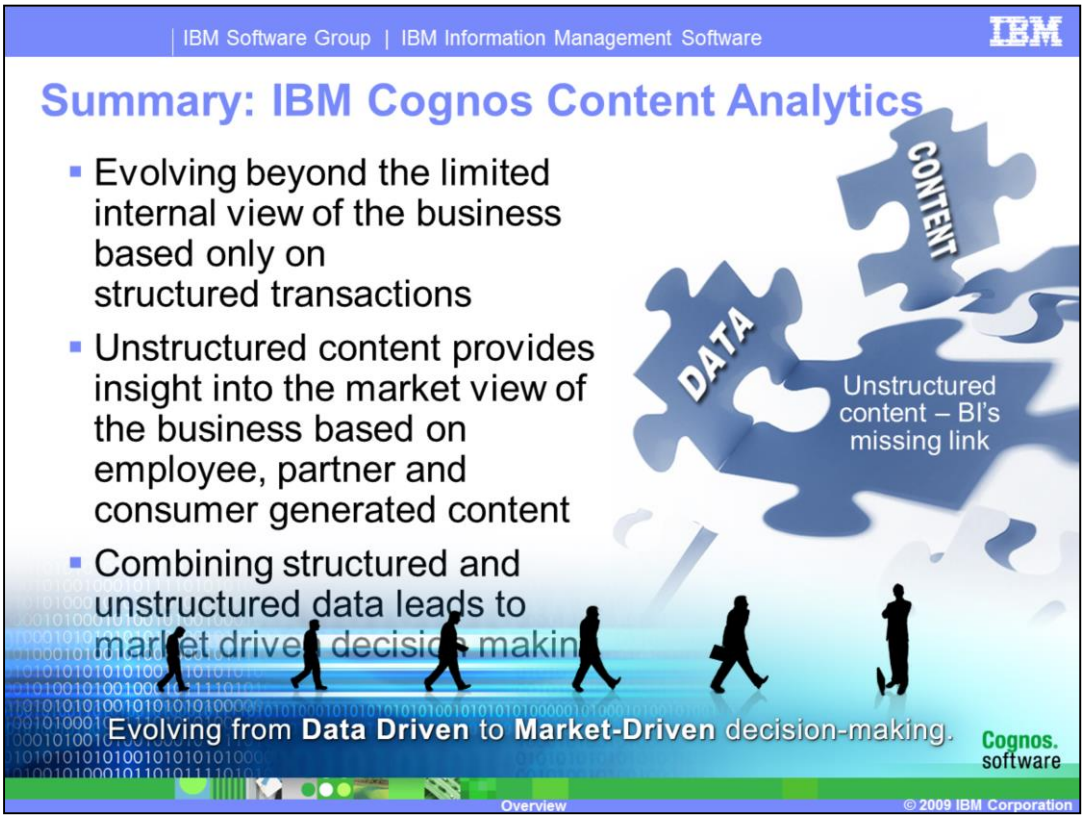


Text Analytics can be a very compute intensive task both during the document processing phase and the exploratory phase as described during the Deep Inspection discussion.

Cognos Content Analytics has accordingly been engineered to accommodate an increasing volume of documents to be processed and an increasing number of users accessing Cognos Content Analytics. As shown in this diagram, you can add any number of document processing servers to account for the increase in the number of documents to be analyzed. Likewise, on the exploratory side you can add any number of search servers to support an increasing number of users or facets as in the Deep Inspector case.



Cognos Content Analytics also comes with an easy to use and robust Administration console. All phases of the system life cycle can be configured and managed through this console. The administration console also supports multiple administrators with varying roles and responsibilities allowing a single installation of Cognos Content Analytics to service the various analytic needs of your organization. Roles can be applied that determine the degree of functions that can be performed. For example, allowing certain administrators to only change and monitor collections that they have been assigned. There is also a useful set of editors for the facet tree, dictionary, and rule pattern matcher.



In summary, Cognos Content Analytics allows you to evolve beyond the limited internal view of the business based only on structured transactions. Unstructured content provides insight into the market view of the business based on employee, partner and consumer generated content. Also, combining structured and unstructured data leads to market driven decision making.

Trademarks, copyrights, and disclaimers

IBM, the IBM logo, ibm.com, and the following terms are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both:

Cognos DB2 Domino FileNet IBM InfoSphere Lotus
Lotus Notes QuickPlace WebSphere

If these and other IBM trademarked terms are marked on their first occurrence in this information with a trademark symbol (® or ™), these symbols indicate U.S. registered or common law trademarks owned by IBM at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries. A current list of other IBM trademarks is available on the Web at "Copyright and trademark information" at http://www.ibm.com/legal/copytrade.shtml

Microsoft, SharePoint, Windows, and the Windows logo are registered trademarks of Microsoft Corporation in the United States, other countries, or both.

UNIX is a registered trademark of The Open Group in the United States and other countries.

Java, JDBC, and all Java-based trademarks and logos are trademarks of Sun Microsystems, Inc. in the United States, other countries, or both.

Other company, product, or service names may be trademarks or service marks of others.

Product data has been reviewed for accuracy as of the date of initial publication. Product data is subject to change without notice. This document could include technical inaccuracies or typographical errors. IBM may make improvements or changes in the products or programs described herein at any time without notice. Any statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only. References in this document to IBM products, programs, or services does not imply that IBM intends to make such products, programs or services available in all countries in which IBM operates or does business. Any reference to an IBM Program Product in this document is not intended to state or imply that only that program product may be used. Any functionally equivalent program, that does not infringe IBM's intellectual property rights, may be used instead.

THE INFORMATION PROVIDED IN THIS DOCUMENT IS DISTRIBUTED "AS IS" WITHOUT ANY WARRANTY, EITHER EXPRESS OR IMPLIED. IBM EXPRESSLY DISCLAIMS ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE OR NONINFRINGEMENT. IBM shall have no responsibility to update this information. IBM products are warranted, if at all, according to the terms and conditions of the agreements (for example, IBM Customer Agreement, Statement of Limited Warranty, International Program License Agreement, etc.) under which they are provided. Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products in connection with this publication and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products.

IBM makes no representations or warranties, express or implied, regarding non-IBM products and services.

The provision of the information contained herein is not intended to, and does not, grant any right or license under any IBM patents or copyrights. Inquiries regarding patent or copyright licenses should be made, in writing, to:

IBM Director of Licensing
IBM Corporation
North Castle Drive
Armonk, NY 10504-1785
U.S.A.

Performance is based on measurements and projections using standard IBM benchmarks in a controlled environment. All customer examples described are presented as illustrations of how those customers have used IBM products and the results they may have achieved. The actual throughput or performance that any user will experience will vary depending upon considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve throughput or performance improvements equivalent to the ratios stated here.