IBM Content Analytics

Francisco J. Izquierdo

© 2011 IBM Corporation

Agenda

- ¿Qué es “Text Analytics”?
- ¿Qué es IBM Content Analytics?
- Minería de Texto con ICA
  - Demo
- Arquitectura de ICA
  - Componentes del Sistema
  - Modelo de Datos
  - Text Analytics y UIMA
  - Escalailidad
  - Query Syntax
- Preguntas

© 2011 IBM Corporation

¿Qué es “Text Analytics”?

Text Analytics engloba un conjunto de técnicas lingüisticas, estadísticas y de aprendizaje con las que podemos dar estructura a un contenido textual

Capacidades deText Analytics		
Extracción de Sintagmas	**Extracción de Hechos**	**Propuestas de Taxonomías**
Lematización	**Extracción de Metadatos**	**Análisis de Sentimientos**
Extracción de Raíces	**Resolución de Anáforas**	**Categorización y Clasificación**
Segmentación	**Resúmenes Automáticos**	**Etc, etc, etc**

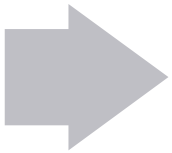
Herramientas y Técnicas de Text Analytics		
Redes Neuronales	**Autómatas Finitos**	**Extracción deExpresiones Regulares**
Clasificadores	**Modelos de Markoff**	**Análisis Semántico**
Árboles de Decisión	**Diccionarios**	**Etc, etc, etc**
Redes Bayesianas	**Máquinas de Inferencia**	

Aplicaciones de Text Analytics		
Análisis de Redes Sociales	**Análisis de Reputaciones**	**Análisis de Tendencias de Mercado**
Detección de Fraude	**Búsqueda de información**	**Registros de Pacientes**
		Etc, etc, etc

© 2011 IBM Corporation

Utilización de IBM Text Analytics para Análisis de informes Policiales

PC 143 **(Hunter)**
15 June 2006 23:47
Suspect **identified himself as** John Setsuko. **Matched description given by night club doorman (IC1, Male, Ag 22-24 yrs, blue Everton shirt). Stopped whilst driving** White Ford Mondeo, W563 WDL. Address **given as** 22 East Dene Ridge, Copdock, Ipswich. **Searched at scene and found in possession of** 1oz Cannabis Resin **and** lockable pocket knife.



Arresting_Officer	PC 143
Arrest_Date_Time	15/06/2006 : 23:47
Suspect_Forename	John
Suspect_Surname	Setsuko
Suspect_VRN	W563WDL
Suspect_Vehicle_Colour	White
Suspect_Vehicle_Make	Ford Mondeo
Suspect_Addr_Street	22 East Dene Ridge
Suspect_Addr_Town	Ipswich
Evidence_1_Description	1 oz Cannabis Resin
Classification	Drug possession

Con este análisis ahora ya es posible::

- **Comprobar errores e inconsistencias en las bases de datos existentes**
- **Proporcionar información útil a los diferentes departamentos**
- **Mejorar considerablemente las capacidades de búsqeuda documental**
- **Realizar tareas de resolución de identidades y relaciones**

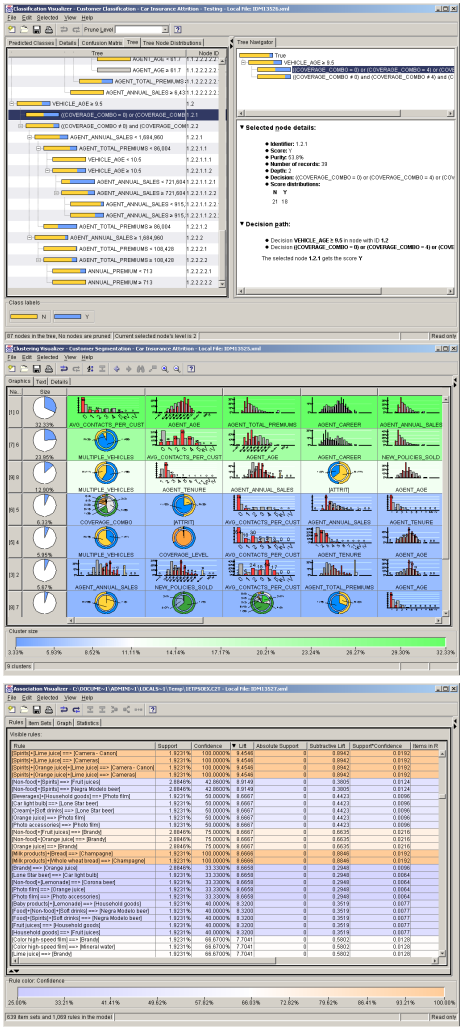
© 2011 IBM Corporation

¿Qué es la Minería de Textos?

- La Minería de Textos es el proceso de extración de información utilizable a partir de datos textuales no estructurados, mediante la identificación de conceptos, opiniones y tendencias.

Grupos de palabras

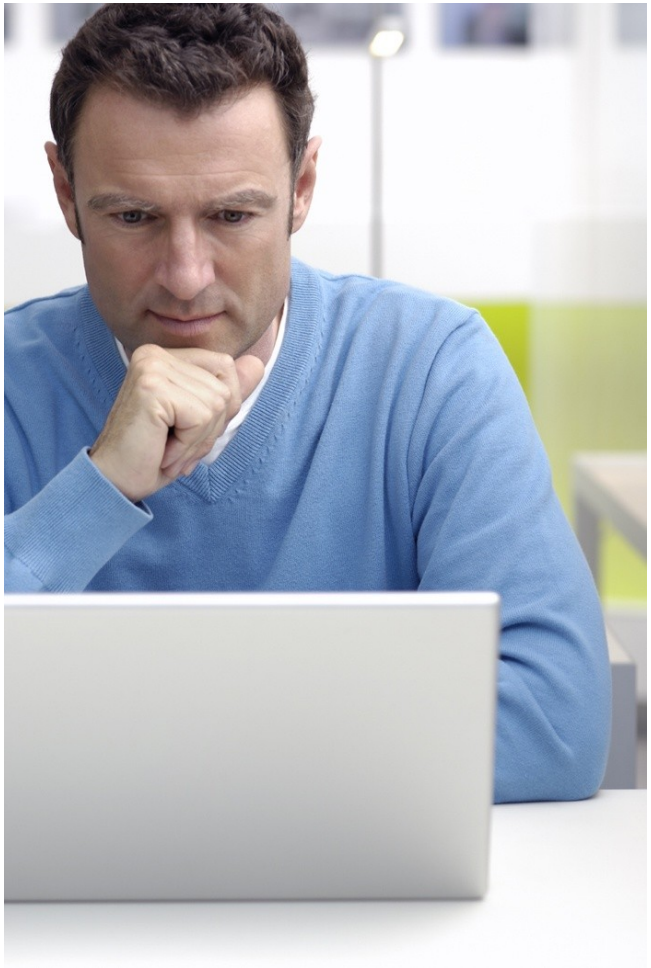
01100101
11101011
01101110
00010100



© 2011 IBM Corporation

¿Qué es IBM Content Analytics?

Un producto cuyo objetivo es ayudar en la percepción del negocio de nuestros clientes



Buscar información y ejecutar los mismos análisis sobre los datos no estructuras que siempre hemos realizado sobre los datos estructurados

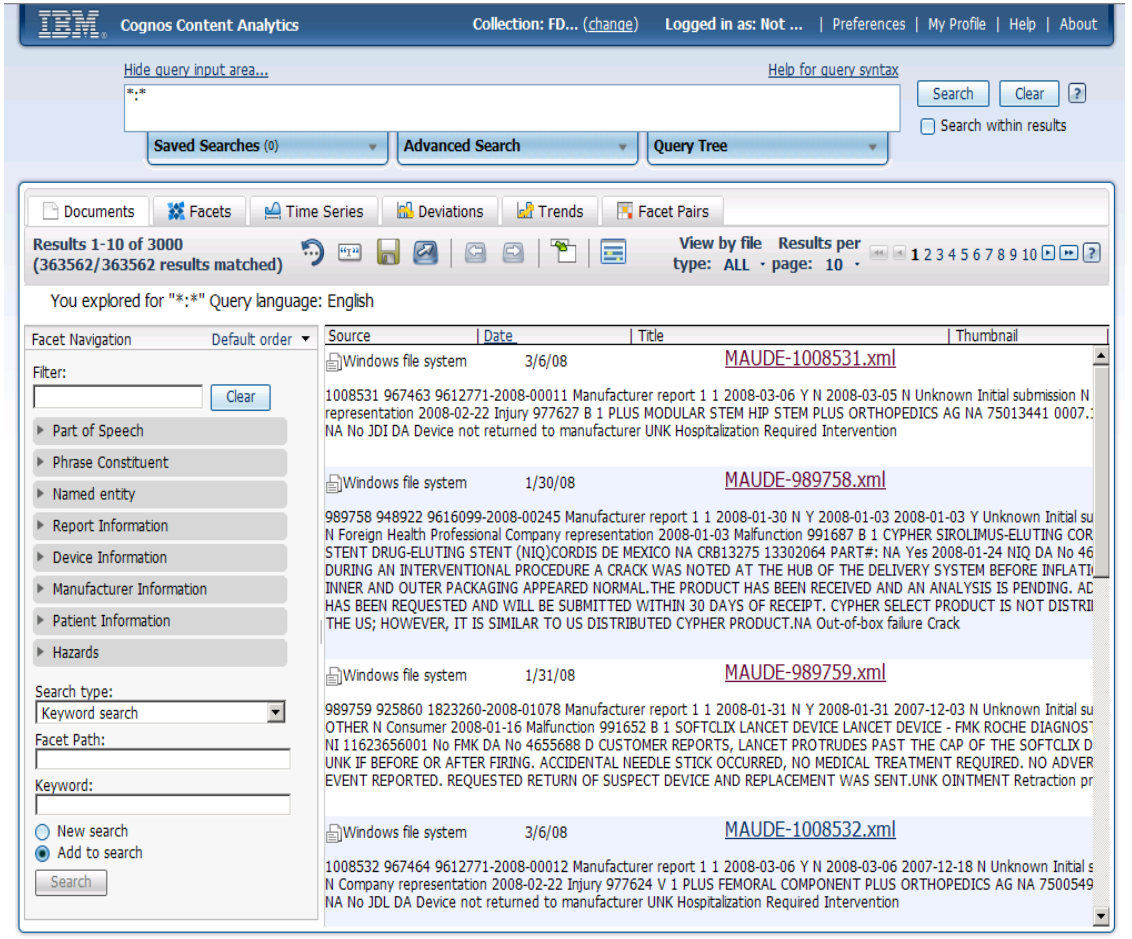
Características y Beneficios de ICA

- ✓ **Combina los datos estructurados y o estructurados para realizar el mismo tipo de análisis**
- ✓ **Permite analizar los contenidos desde múltiples puntos de vista**
- ✓ **Identifica de forma automática relaciones no habituales entre los datos, para realizar una investigación mas exhaustiva**

© 2011 IBM Corporation

IBM Content Analytics – Minería de Textos

- **Mínería de Textos**
  - Extraer información de calidad de los datos textuales
  - Combina métodos de búsqueda convencionales con herramientas de minería de datos
  - Detección de modelos y tendencias, para destacar desviaciones y cambios a investigar
- **Aporta una mejor percepción del negocio a partir contenidos no estructurados**
  - Combina la extracción de entidades y ayuda a categorizar los documentos y a obtener informes operativos
- **Punto de integración para información estructurada y no estructurada**
  - Se integra con Cognos 8 BI, InfoSphere Warehouse, IBM ECM, WebSphere Portal, y soluciones a medida
  - Proporciona un interface que permite varias formas de acceso a la documentación analizada



© 2011 IBM Corporation

Minería de Textos – Diferentes vistas

- **Minería de textos interactiva**
  - Facilita encontrar documentos según los valores de las facetas
  - Diferentes vistaas en función de lo que se quiera obtener
  - Gráficos orientados a ayudar entender cada situación
- **Nueve vistas**
  - **Documentos**
  - **Facetas**
  - **Series Temporales**
  - **Desviaciones**
  - **Tendencias**
  - **Pares de Facetas**
  - **Conexiones**
  - **Dashboard**
  - **Tiara Plugin**

Demostración de Minería de Texto

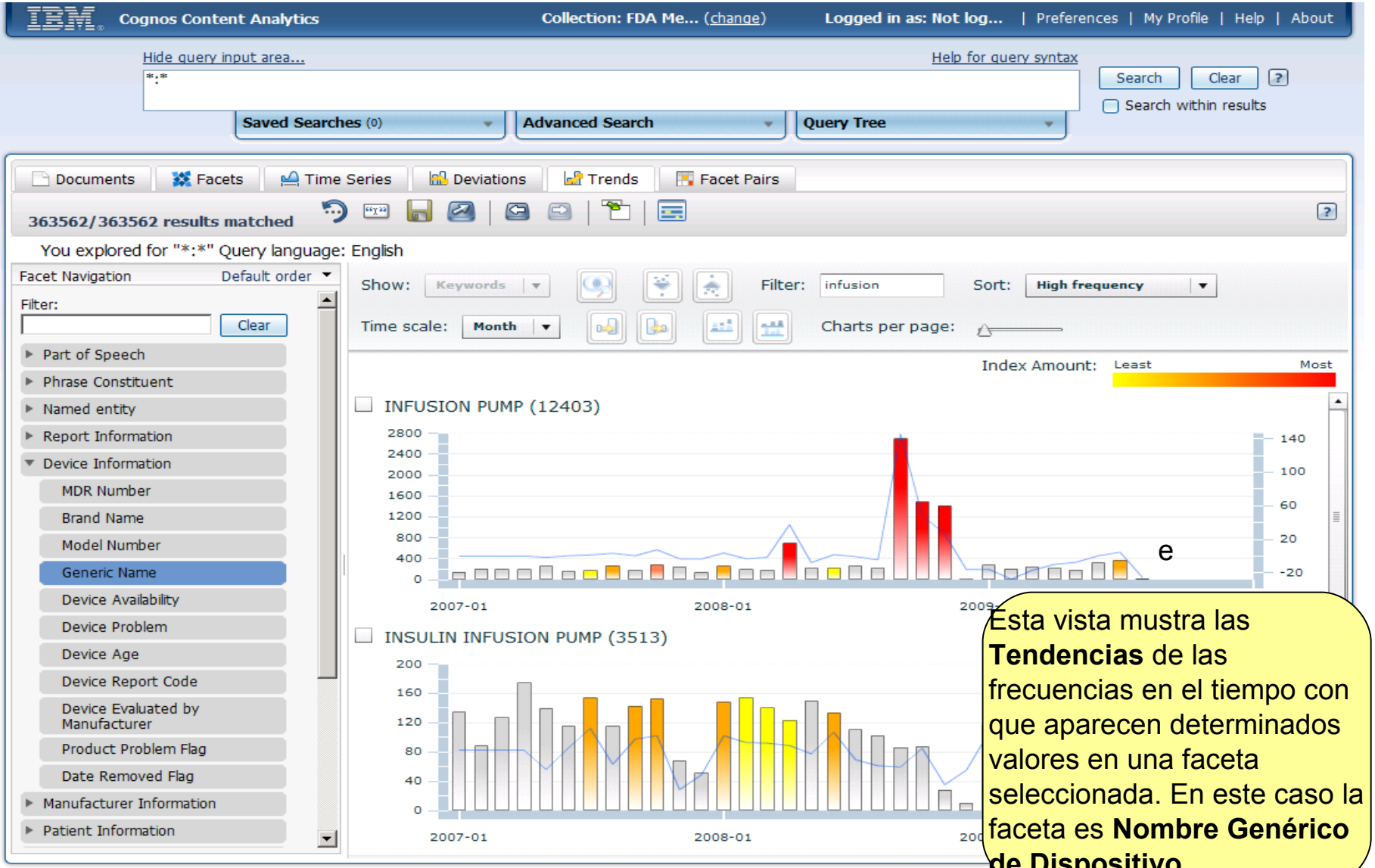
© 2011 IBM Corporation

El Programa MedWatch de la FDA

Los informes de Incidentes **MedWatch de FDA** son una fuente de datos de incidentes sobre instrumentos médicos que reportan los usuarios. Contienen tanto información **estructurada** como **no estructurada**. Adicionalmente, los fabricantes pueden **analizar sus documentos internos**, como reclamaciones o incidentes sobre instraumentos en garantía.

© 2011 IBM Corporation

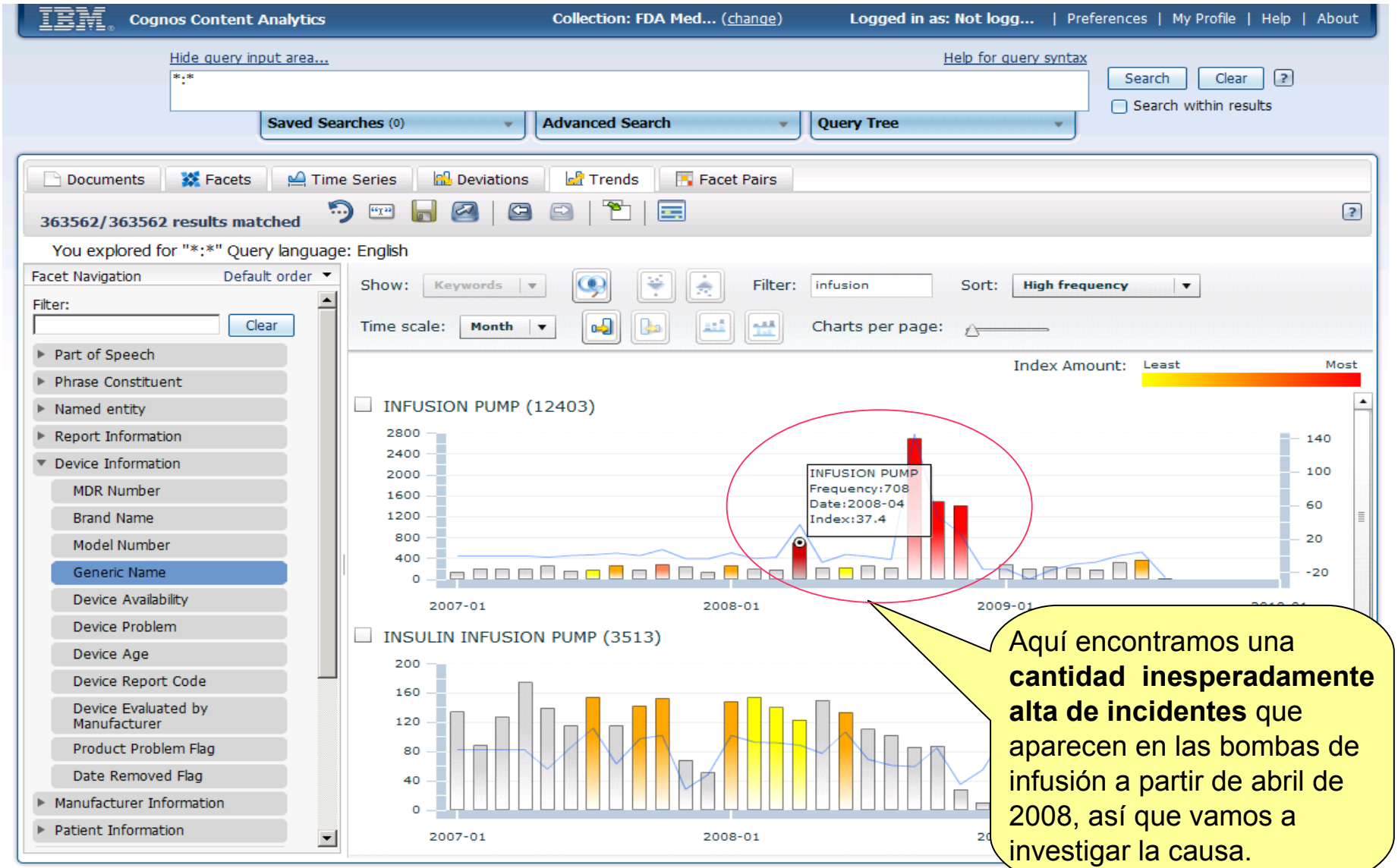
Análisis de Bombas de Infusión



Esta vista mustra las **Tendencias** de las frecuencias en el tiempo con que aparecen determinados valores en una faceta seleccionada. En este caso la faceta es **Nombre Genérico de Dispositivo**

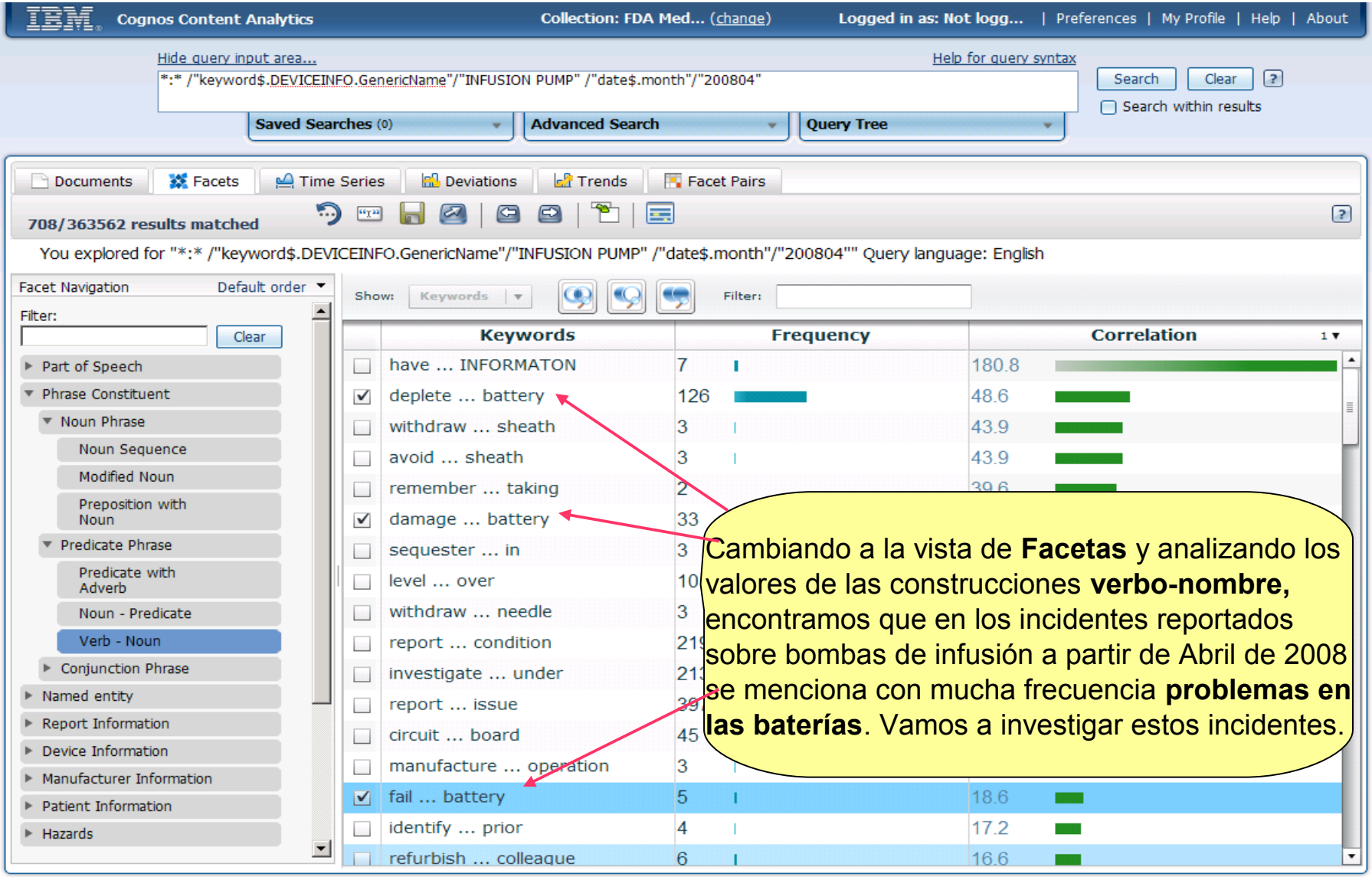
© 2011 IBM Corporation

Análisis de Bombas de Infusión



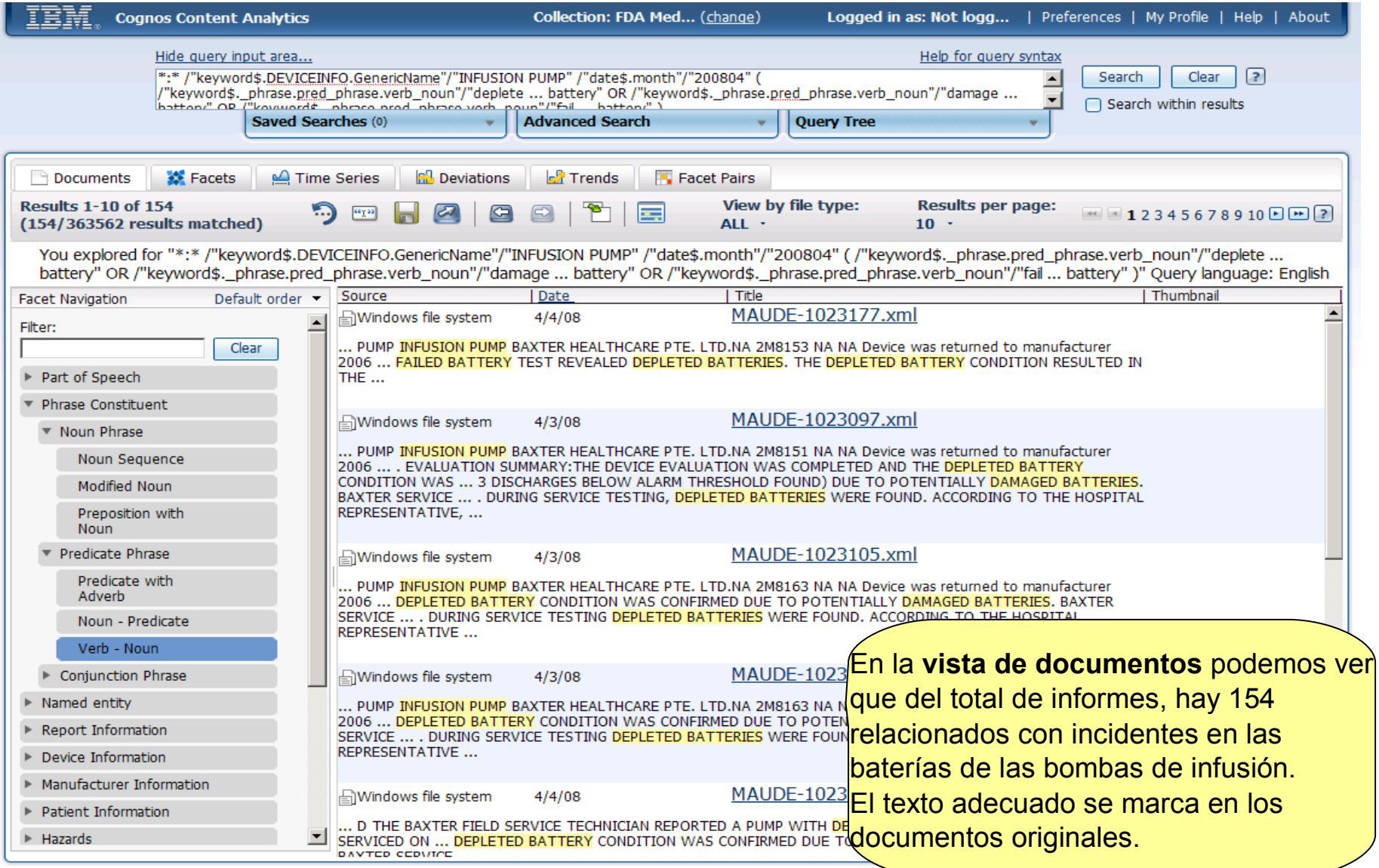
© 2011 IBM Corporation

Análisis de Bombas de Infusión



© 2011 IBM Corporation

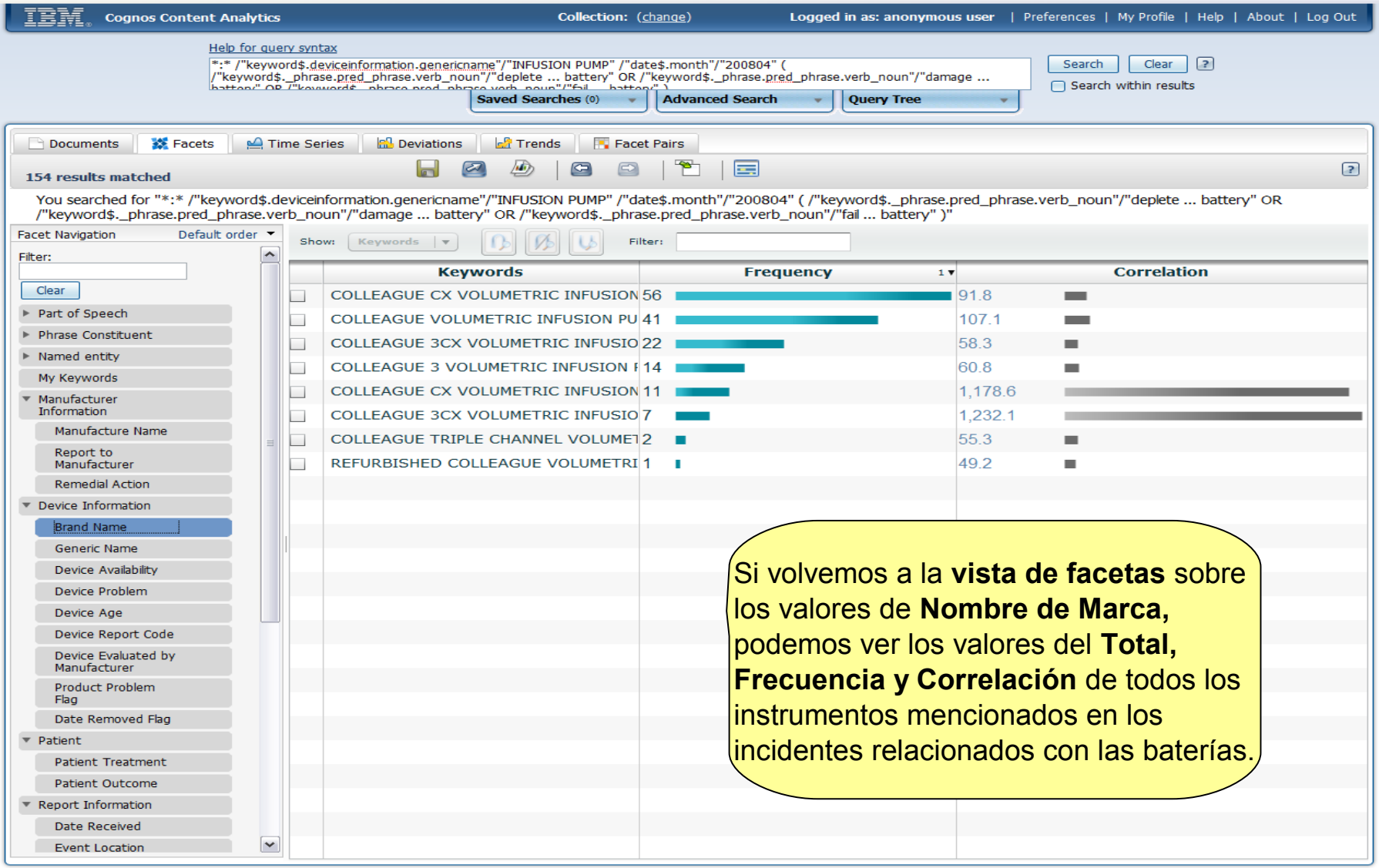
Análisis de Bombas de Infusión



En la **vista de documentos** podemos ver que del total de informes, hay 154 relacionados con incidentes en las baterías de las bombas de infusión. El texto adecuado se marca en los documentos originales.

© 2011 IBM Corporation

Análisis de Bombas de Infusión



Si volvemos a la **vista de facetas** sobre los valores de **Nombre de Marca,** podemos ver los valores del **Total, Frecuencia y Correlación** de todos los instrumentos mencionados en los incidentes relacionados con las baterías.

© 2011 IBM Corporation

Análisis de Bombas de Infusión

U.S. Food and Drug Administration

A-Z Index | Search

Home | Food | Drugs | Medical Devices | Vaccines, Blood & Biologics | Animal & Veterinary | Cosmetics | Radiation-Emitting Products | Tobacco Products

FDA Home > Medical Devices > Databases

Medical Device Recalls

CDRH SuperSearch

510(k) | Registration & Listing | Adverse Events | Recalls | PMA | Classification | Standards
CFR Title 21 | Radiation-Emitting Products | X-Ray Assembler | Medsun Reports | CLIA

New Search | Back to Search Results

Class 1 Recall Colleague, Colleague CX and CXE Triple Channel Infusion Pumps

See Related Information

Date Posted	March 02, 2009
Recall Number	Z-1038-2009
Product	Baxter Colleague Triple Channel Volumetric Infusion Pumps. Baxter Healthcare Corporation, Medication Delivery Division, Deerfield, IL 60015 U.S.A., Made in Singapore. Product codes: 2M8153, 2M8163, 2M9163, DNM8153, DNM9163, BRM8153, BRM8153T, BRM8153RT, PNM8153, PNM8153T, PNM8153RT, CNM8153, GNM8153, HNM8153, TRM8153 and WNM8153. Electronic infusion pumps are indicated for continuous or intermittent delivery of solution through clinically acceptable routes of administration intravenous (IV), intra-arterial (IA), subcutaneous, epidu spaces applications.
Code Information	All product numbers and all serial numbers.
Recalling Firm/ Manufacturer	Baxter Healthcare Corp. Rt. 120 & Wilson Rd Round Lake, Illinois 60073
For Additional Information Contact	Center for One Baxter 800-422-9837
Reason for Recall	Baxter has identified failure codes that lead to a delay or an interru of therapy, hazards associated with improper cleaning, and damaged battery messages for the Colleague Infusion Pumps

10 meses más tarde, la FDA emitió una orden de retirada de clase 1 para todas estas bombas de infusión, a causa de defectos en las baterías.

© 2011 IBM Corporation

THE WALL STREET JOURNAL.

Tuesday May 4th, 2010

Y 24 meses más tarde ...

- Baxter International Inc. said Monday it would recall the approximately 200,000 Colleague brand drug-infusion pumps that are on the market, after years of malfunctions with the device, along with patient injuries and deaths.
- The Colleague pumps have been widely used in hospitals, especially in the U.S., to deliver medication and other fluids to patients.

Baxter to Recall Infusion Devices

By Jennifer Corbett Dooren And Thomas M. Burton

Baxter International Inc. said Monday it would recall the approximately 200,000 Colleague brand drug-infusion pumps that are on the market, after years of malfunctions with the device, along with patient injuries and deaths.

The Colleague pumps have been widely used in hospitals, especially in the U.S., to deliver medication and other fluids to patients.

Baxter, of Deerfield, Ill., said the recall was requested by the Food and Drug Administration, which has been focusing its regulatory attention of late on infusion devices. The company, with $12.8 billion in annual sales, said it will record a pretax special charge of $400 million to $600 million in the first quarter for the estimated cost of the recall.

Baxter has had troubles with the Colleague infusion pump for several years, including a wide range of malfunctions. These included overinfusion and underinfusion, as well as electrical-shock hazard. At one time, the device had generated about $170 million in annual sales.

In 2005, more than 7,000 Colleague pumps were seized at the FDA's request after reports of 16 patient injuries and seven deaths from pump malfunctions. In 2006, the company entered into a consent decree with the FDA involving the Colleague, originally with the aim of fixing the device's problems.

Among other things, Baxter has in the past notified customers that Colleague pumps could sound an alarm and stop infusing critical medicines into patients.

In addition, the company told customers about "the possibility of the device overheating," resulting in smoke and fire, if improperly cleaned, according to one company statement last year.

Baxter said it expects to exchange existing Colleague pumps with Sigma Spectrum pumps that it distributes.

- **Se retiraron aproximadamente 200.000 unidades**
- **El coste estimado de esta operación estuvo entre 400 y 600 millones de dólares**

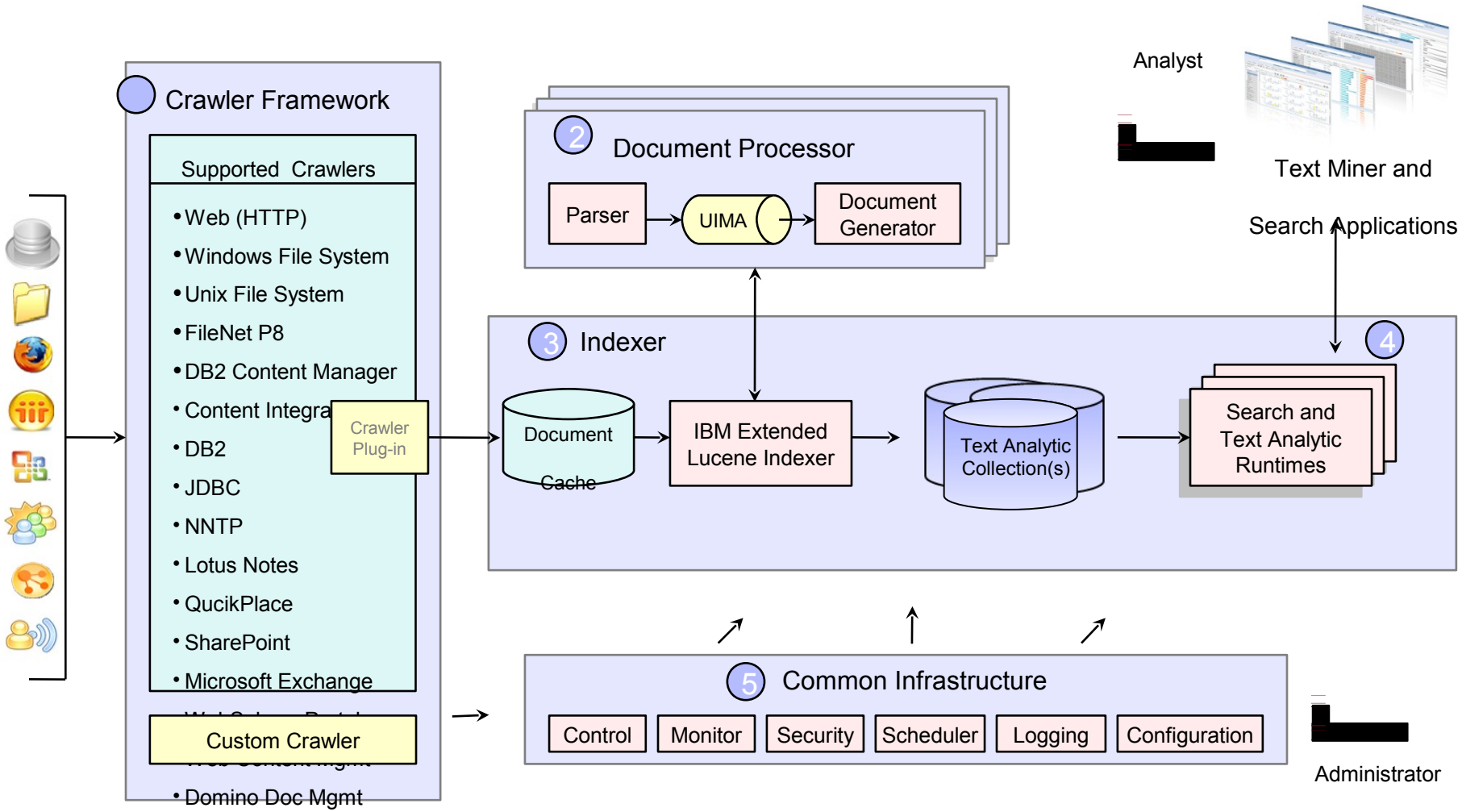
Septe

© 2011 IBM Corporation

Arquitectura del Sistema

© 2011 IBM Corporation

Arquitectura de ICA 2.2



© 2011 IBM Corporation

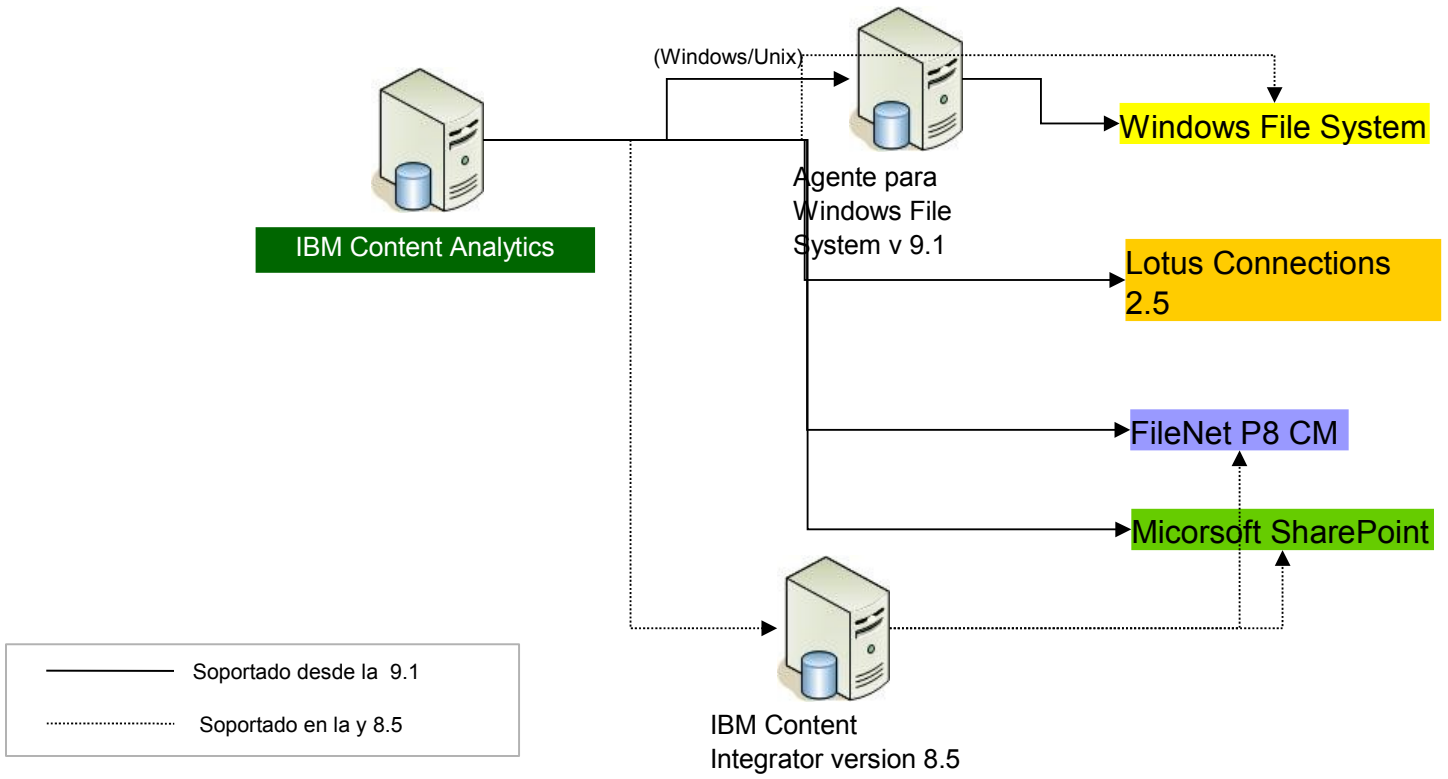
IBM Content Analytics: Rastreadores a Fuentes

- **Web Content**
  - HTTP/HTTPS including RSS and Atom support
  - News groups (NNTP)
  - WebSphere Portal, Web Content Management
- **Relational Data sources**
  - DB2 family (DB2 UDB, Informix, DB2 for iSeries, DB2 for z/OS)
  - Oracle, MS SQL Server, Sybase
  - VSAM, IMS, CA-Datacom, Software AG Adabas
- **Collaboration Systems**
  - Lotus Notes/Domino databases, Quick Place, Domino.Doc
  - Lotus Quickr, Lotus Connections
  - MS Exchange public folders
- **Content Management Systems**
  - DB2 Content Manager
  - Documentum, FileNet CS, FileNet P8, Hummingbird, LiveLink Open Text, Portal Document Manager (PDM), Microsoft Sharepoint
- **File System**
  - Unix File System
  - Windows File System

© 2011 IBM Corporation

Rastreadores Sobre Gestores de Contenidos

- Rastreador para FileNet P8
- Rastreador para SharePoint
- Rastreador para Windows File System



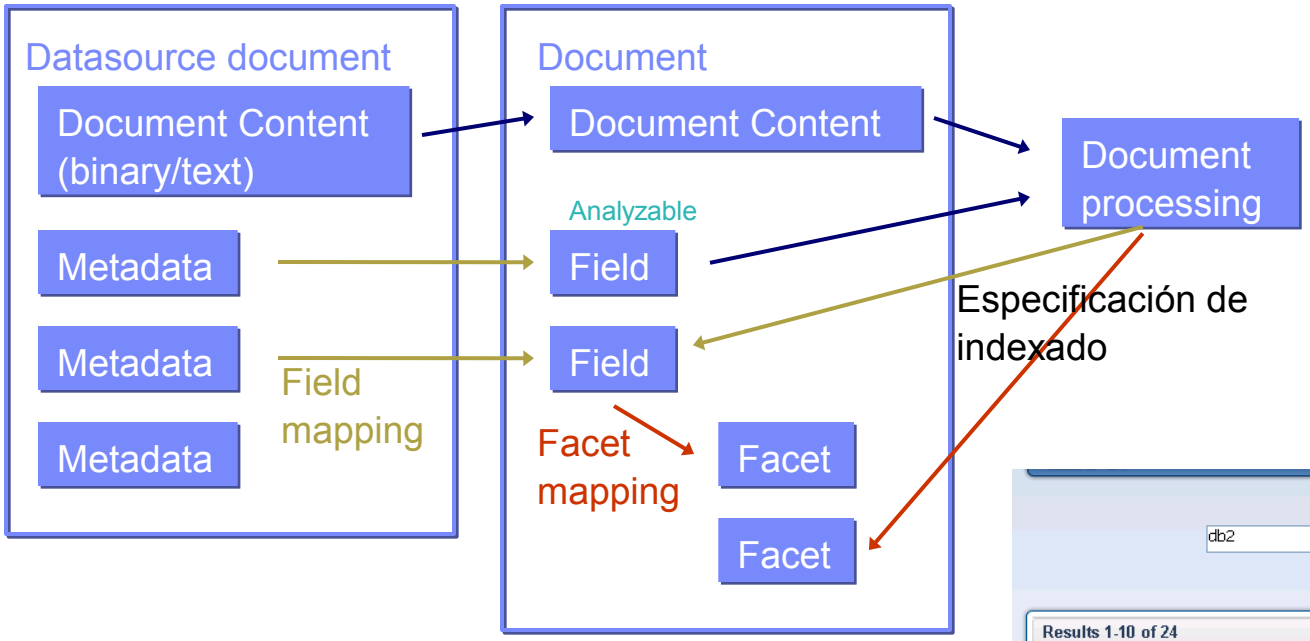
© 2011 IBM Corporation

Rastreadores para Seedlists

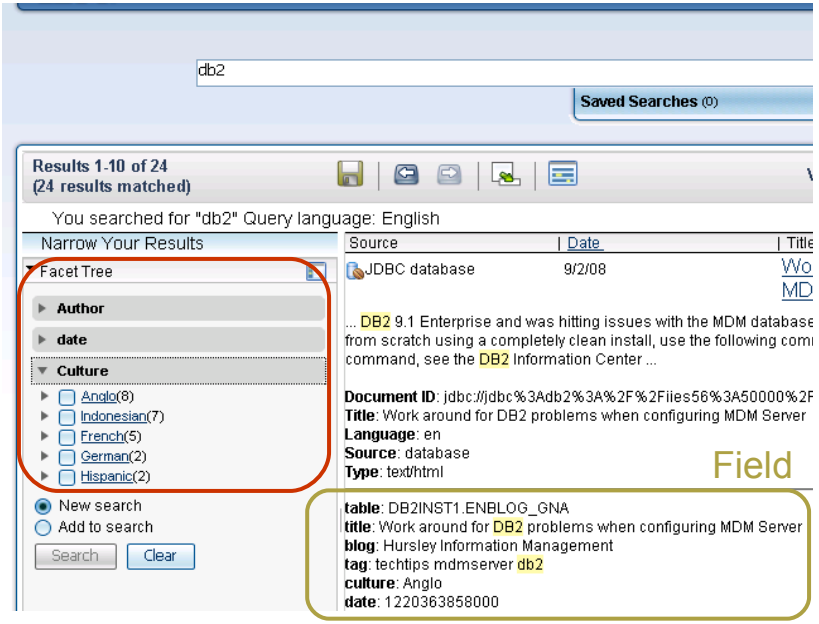
- El Rastreador para Seedlists soporta las siguientes fuentes de datos:
  - WebSphere Portal v6.1.x
  - Web Content Management v6.1.5
  - Lotus Quickr for Portal v8.x
  - Lotus Connections 2.0.x
  - Lotus Connections 2.5
- Relación entre campos
  - El Rastreador recoge todos los campos de la seedlist para que se puedan mapear libremente a los campos de búsqueda y a las facetas de ICA
  - A su vez, los campos y las facetas de ICA se pueden crear a partir de cualquier campo de la fuente de datos original

© 2011 IBM Corporation

IBM Content Analytics - Modelo de Datos

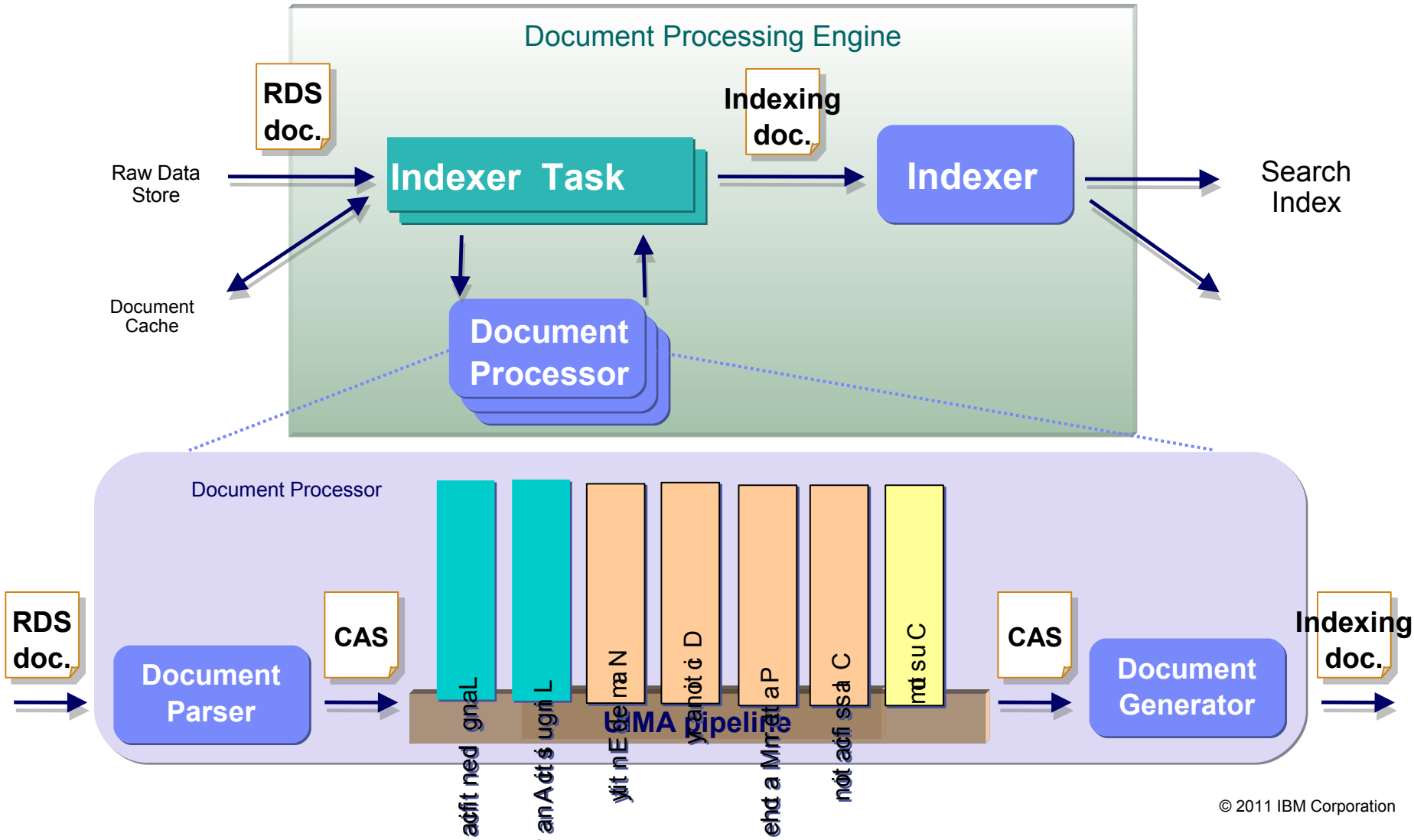


Contenidos, Campos, y Facetas (Vista de Usuario Final)



© 2011 IBM Corporation

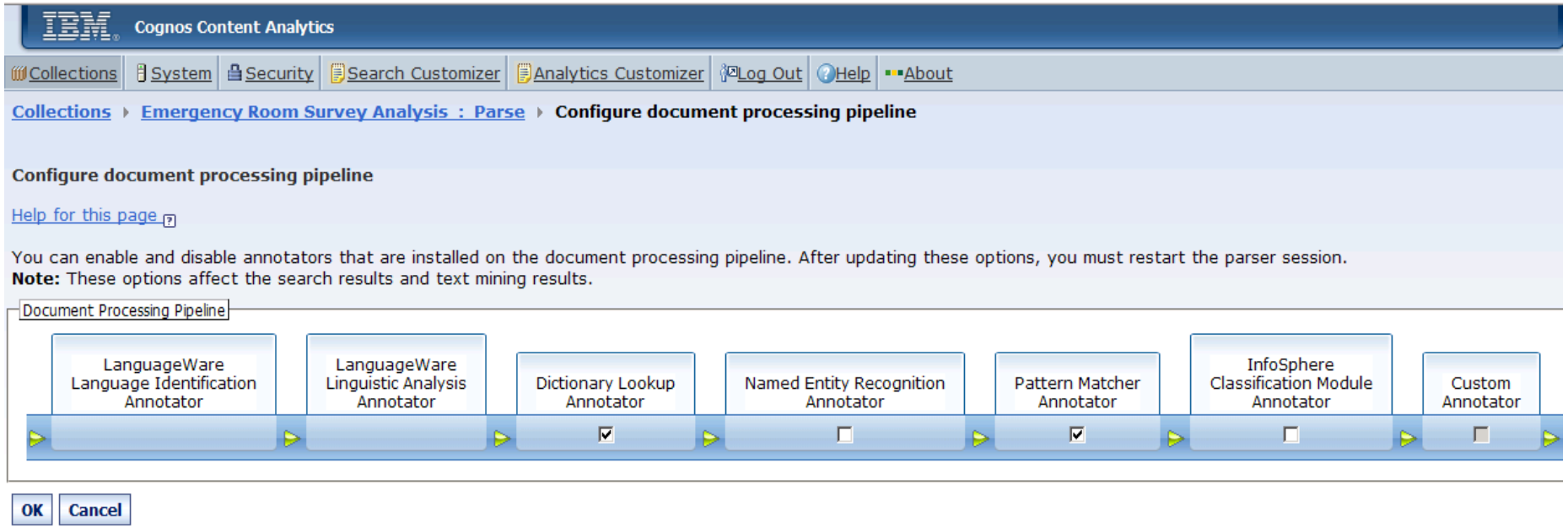
Estructura del Document Processing Engine



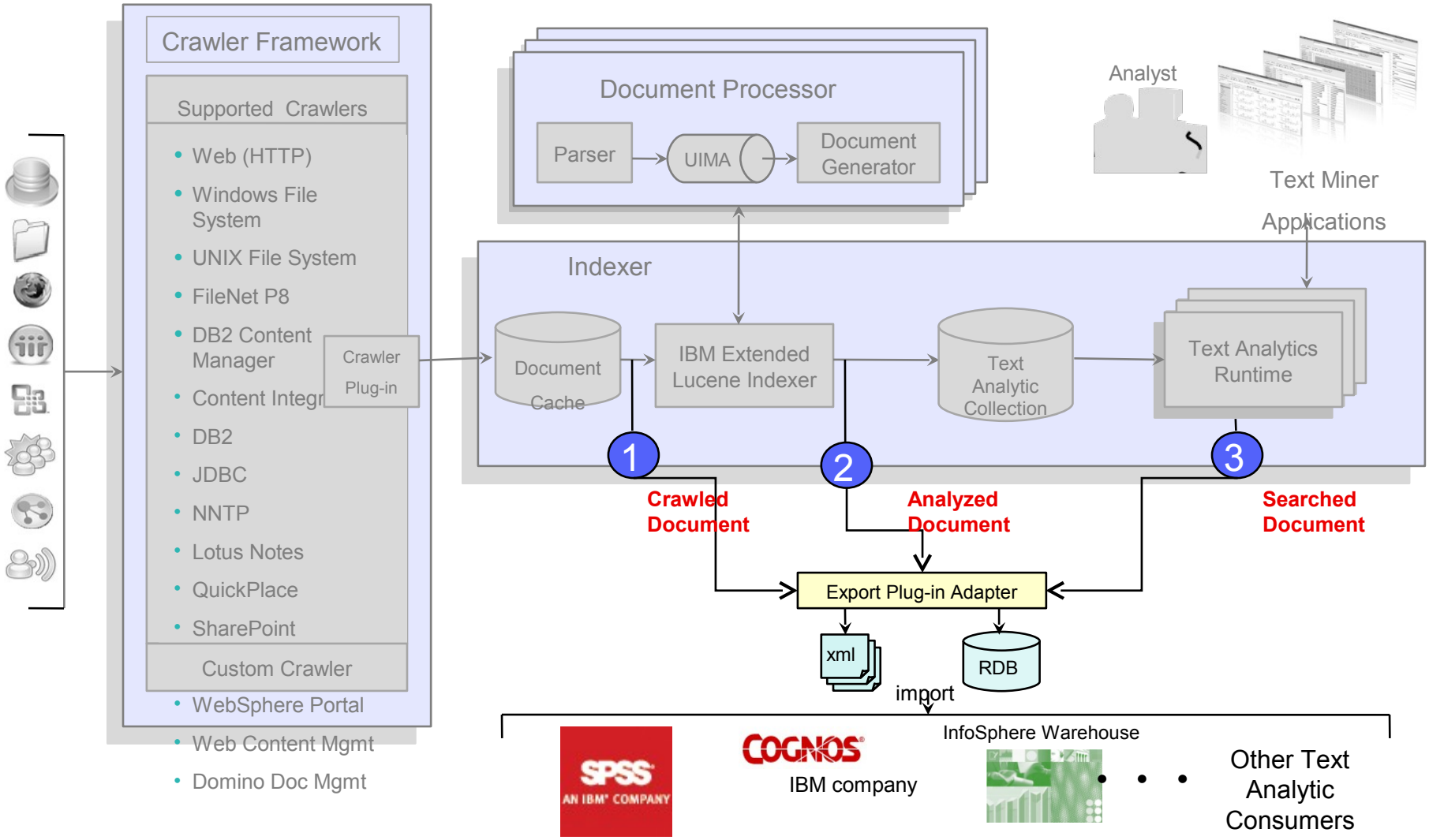
© 2011 IBM Corporation

Configuración de Anotadores UIMA

- **La herramienta de Administración permite a los usuarios habilitar o deshabilitar anotadores UIMA**
  - La búsqueda en Diccionarios y Reconocimiento de Patrones se habilitan por defecto
  - Por defecto el Anotador de Entidades con Nombre está deshabilitado
  - Los Anotadores de usuario son opcionales

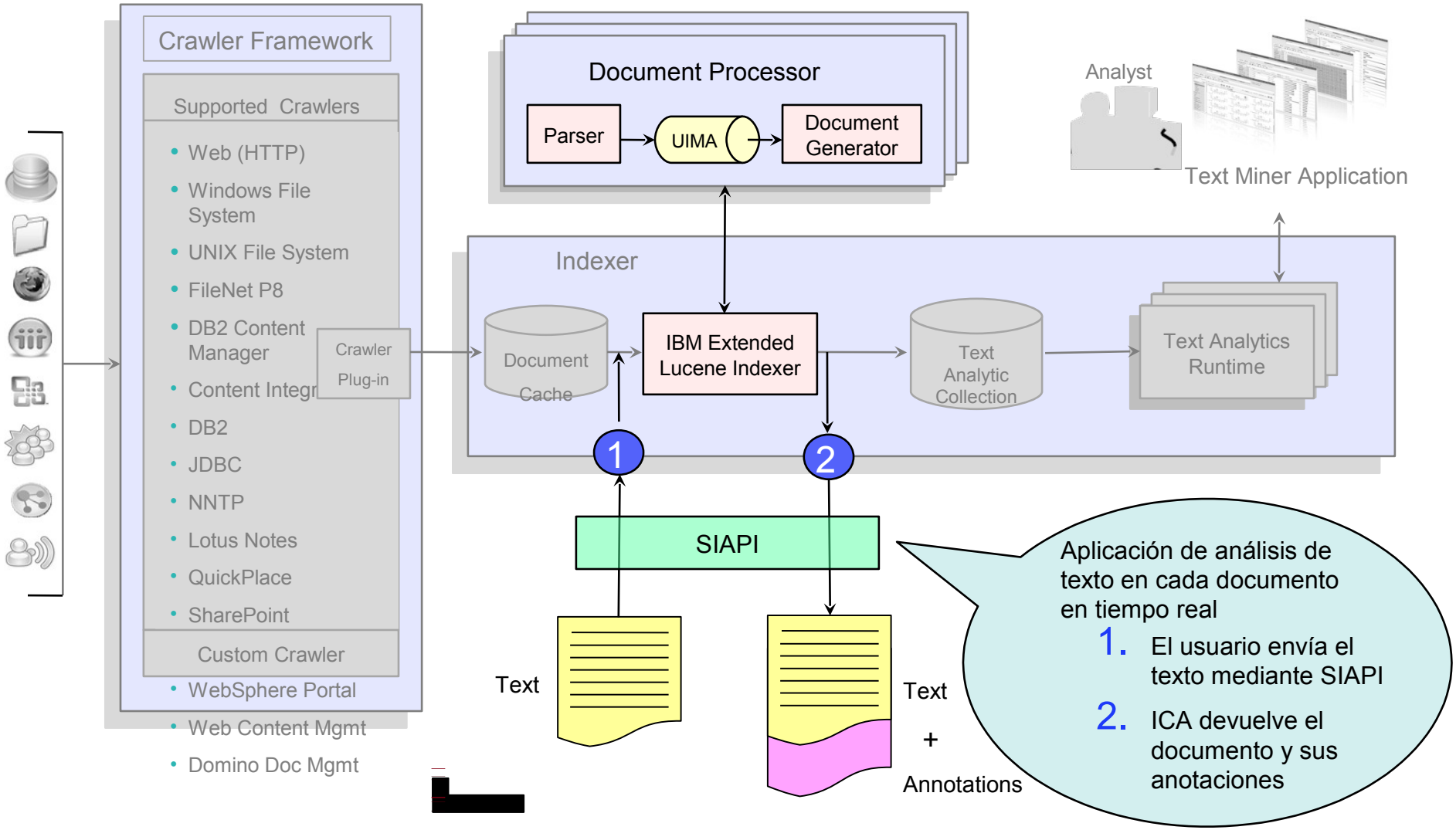


Capacidades de Exportación de Documentos



© 2011 IBM Corporation

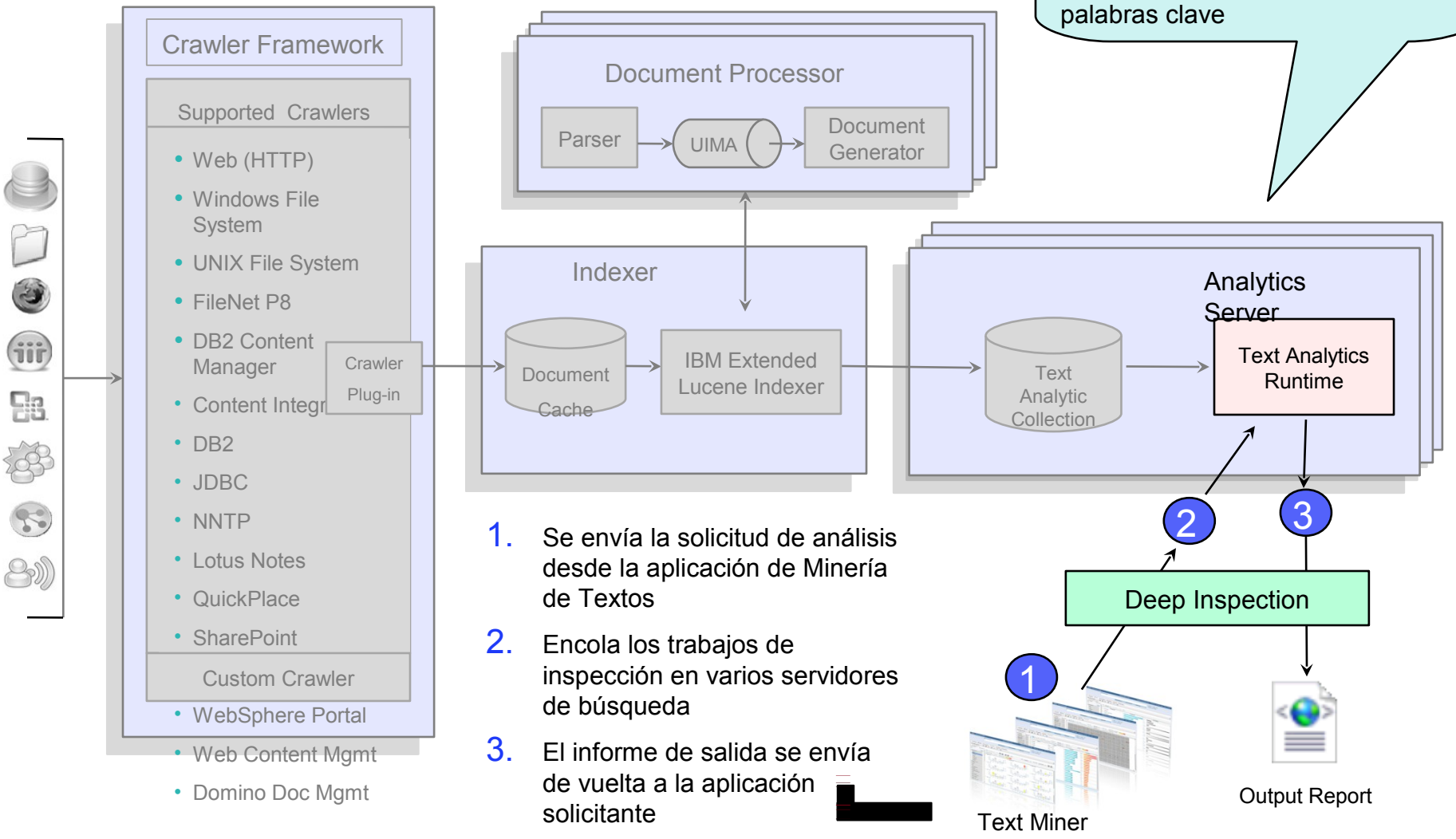
Análisis de Texto en Tiempo Real



© 2011 IBM Corporation

Deep Inspection

Esta facilidfdad permite ejecutar análisis de texto que requieren un elevado número de factas y palabras clave

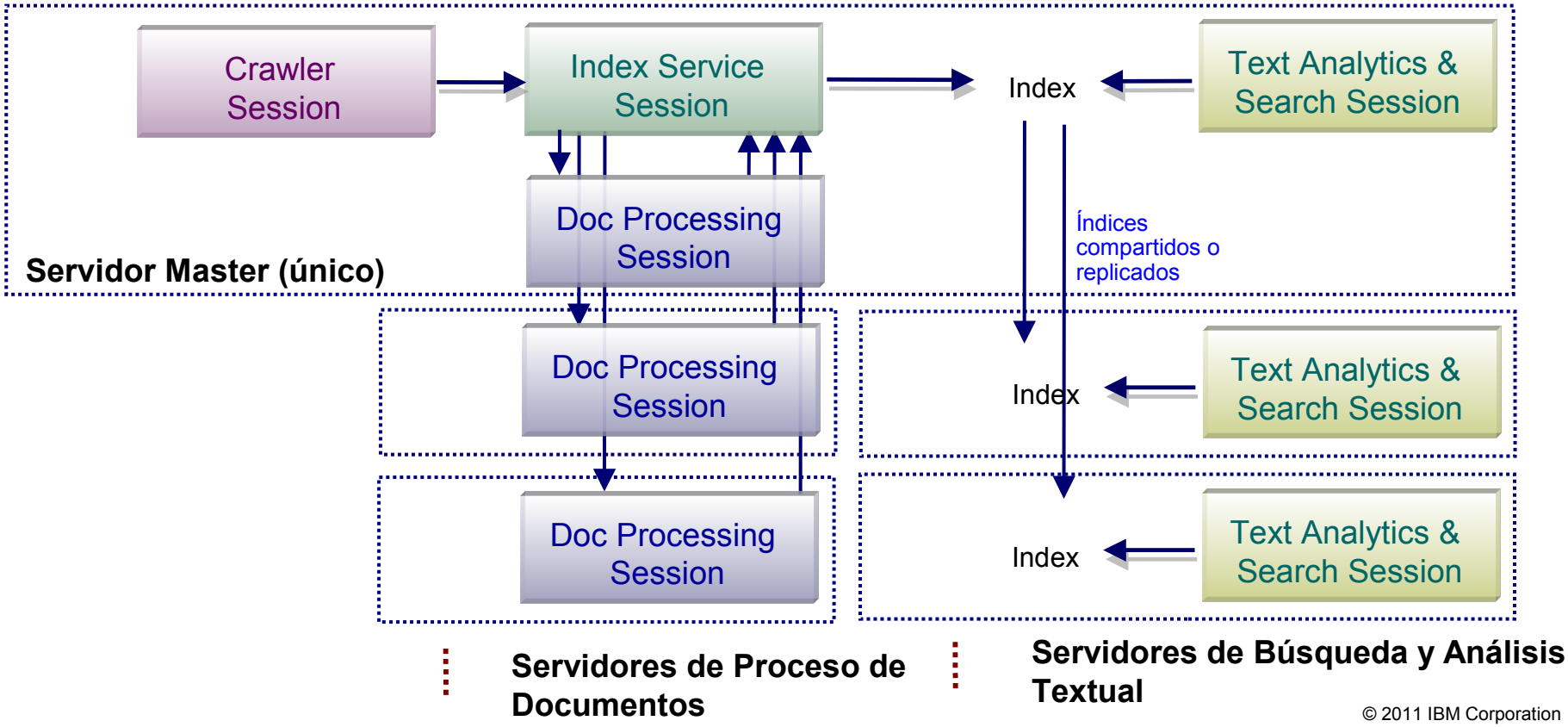


1. Se envía la solicitud de análisis desde la aplicación de Minería de Textos
2. Encola los trabajos de inspección en varios servidores de búsqueda
3. El informe de salida se envía de vuelta a la aplicación solicitante

© 2011 IBM Corporation

Configuración

- Se soportan configuraciones flexibles y escalables
- Con varios servidores de Procesado de Documentos
- Y también varios servidores de búsqueda
- El sistema permite añadior servidores sin necesidad de reiniciar



© 2011 IBM Corporation

Preguntas

© 2011 IBM Corporation