# Desarrollo de una Aplicación de Big Data para Explorar y Descubrir Datos

Consejos, técnicas y pautas prácticas para ayudarlo a comenzar

Explorar big data y datos empresariales tradicionales es un requisito común de muchas organizaciones. En este artículo, señalamos un método y pautas para indexar big data gestionados por una plataforma basada en Hadoop para utilizarla con una solución de descubrimiento de datos. Específicamente, describimos la forma en que los datos almacenados en IBM InfoSphere® BigInsights™ (una plataforma basada en Hadoop) pueden ser trasladados a InfoSphere Data Explorer, una herramienta sofisticada que les permite a los usuarios empresariales explorar y combinar datos de múltiples orígenes, tanto empresariales como externos.

**Introducción**

Si ha estado siguiendo muchos de los estudios de casos recientes relacionados con big data, tal vez haya llegado a creer en la frase "usted no sabe lo que no sabe". En efecto, las aplicaciones de big data frecuentemente se concentran en obtener información empresarial a partir de datos que de otra manera podrían ser descartados o ignorados por distintas razones. Cada vez más, las compañías están buscando desarrollar una estrategia integral de gestión de la información que implique más que solo explorar o analizar big data. Específicamente, desean integrar big data en sus estrategias generales de gestión de la información junto con los sistemas de datos existentes, incluidos DBMSs relacionales, sistemas de gestión de contenido empresariales, depósitos de datos, etc.

Este artículo examina un aspecto de ese desafío, señalando una arquitectura y un método para indexar orígenes de big data y datos tradicionales, y proporcionando una interfaz basada en Web para descubrir nueva información a través de estos orígenes de datos dispares. En particular, describe cómo Data Explorer, una plataforma de descubrimiento de datos, puede indexar datos gestionados por InfoSphere BigInsights, permitiendo que las formas persistentes de big data sean combinadas con los datos empresariales existentes. Data Explorer y BigInsights son componentes clave de la plataforma de big data de IBM, así que comenzaremos con una visión general de esta plataforma y de estos dos componentes fundamentales.

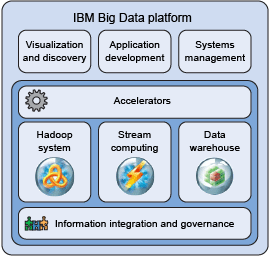
[Volver arriba](http://www.ibm.com/developerworks/ssa/data/library/app_bigdata/872864.html#ibm-pcon)

**Visión general de la plataforma de big data de IBM**

La plataforma de big data de IBM está diseñada para ayudar a las organizaciones a explorar, analizar y gestionar una gran variedad de datos, incluidos los datos de modalidad continua, datos empresariales tradicionales y datos "no convencionales" o auxiliares que anteriormente han sido difíciles de incorporar en plataformas de inteligencia y análisis de negocio. Observemos esta plataforma brevemente antes de concentrarnos en los dos componentes clave — InfoSphere Data Explorer e InfoSphere BigInsights — para el resto de este artículo.

Figura 1 representa la arquitectura de la plataforma de big data de IBM, que difiere de otras ofertas comerciales en la amplitud de sus capacidades. Desde la parte superior hacia abajo, verá que la plataforma de IBM presenta herramientas y tecnologías para visualizar y descubrir información a través de diversos orígenes de datos, desarrollar aplicaciones de análisis y gestionar su entorno. Data Explorer ofrece capacidades clave de visualización y descubrimiento para la plataforma de big data de IBM, así que discutiremos ese componente con más detalle en breve. Los aceleradores que se muestran en la Figura 1 son kits de herramientas provistos por IBM que incluyen decenas de artefactos de software preincorporados para ayudar a las compañías a desplegar rápidamente soluciones para analizar datos de redes sociales y de equipos (por ejemplo, registros). Tres motores de procesamiento de datos les permiten a las organizaciones trabajar de forma efectiva con la variedad, el volumen y la velocidad inherentes a big data. Estos motores incluyen un sistema basado en Hadoop (BigInsights, el cual discutiremos con detalle más adelante), una plataforma de informática continua (InfoSphere Streams) y una plataforma de depósito de datos (como PureData™ para Analytics o DB2®). Finalmente, la plataforma de big data de IBM incluye conectividad con otros software empresariales populares, incluidos DBMSs relacionales, plataformas para extraer/transformar/cargar, herramientas de inteligencia de negocio, sistemas de gestión de contenido y más.

**Figura 1. Arquitectura de la plataforma de big data de IBM**



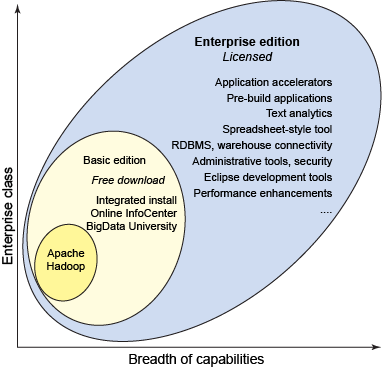
**Visión general de InfoSphere BigInsights**

InfoSphere BigInsights es la plataforma de IBM que admite y analiza muchas formas de big data. Con base en el proyecto de código abierto Apache Hadoop, BigInsights está diseñado para ayudar a las compañías a descubrir y analizar información empresarial oculta en grandes volúmenes de datos que de otra manera podrían ser ignorados o descartados, debido a que es poco práctico o muy difícil procesarlos a través de medios tradicionales. Ejemplos de dichos datos incluyen registros, trayectorias de navegación, datos de redes sociales, transferencias de noticias, emails, salidas de sensores electrónicos e incluso algunos datos transaccionales.

Para que las empresas aprovechen el valor de estos tipos de datos de forma eficiente, BigInsights Enterprise Edition incluye varios proyectos de código abierto del ecosistema Hadoop, así como un número de tecnologías desarrolladas por IBM que mejoran y amplían el valor de este software de código abierto. Como indica la Figura 2 , estas tecnologías van desde aceleradores de aplicaciones hasta recursos de análisis, herramientas de desarrollo, plataformas mejoradas e integración de software empresarial. Por ejemplo, los clientes de BigInsights pueden utilizar capacidades sofisticadas de análisis de texto para extraer contenido y contexto de documentos, emails y mensajes. Los desarrolladores de aplicaciones pueden emplear asistentes basados en Eclipse para acelerar el desarrollo de aplicaciones personalizadas Java™ MapReduce, Jaql, Hive, Pig y análisis de texto. Los administradores pueden gestionar y monitorear sus entornos de BigInsights a través de una consola Web integrada y los usuarios empresariales pueden lanzar aplicaciones provistas por IBM o desarrolladas de forma personalizada a través de un catálogo basado en Web.

En este artículo, nos concentraremos en un subconjunto de funciones de BigInsights, como las herramientas de análisis de texto y ciclo de vida de la aplicación. Para obtener más información sobre BigInsights, vea [Recursos](http://www.ibm.com/developerworks/ssa/data/library/app_bigdata/872864.html#resources).

**Figura 2. Arquitectura de InfoSphere BigInsights**



**Visión general de InfoSphere Data Explorer**

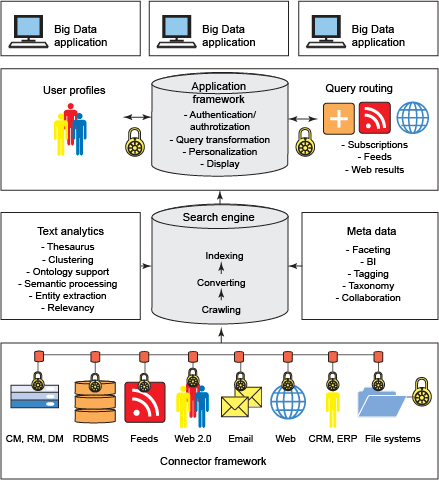
InfoSphere Data Explorer le permite indexar grandes conjuntos de datos estructurados, no estructurados y semiestructurados desde orígenes de datos dispares. También ofrece la posibilidad de desarrollar aplicaciones para la exploración de big data y aplicaciones de información de 360 grados. InfoSphere Data Explorer le permite al usuario crear una vista de información relevante sobre distintas entidades como clientes, productos, eventos, socios, etc., a partir de grandes conjuntos de datos almacenados en distintos repositorios internos y externos sin tener que mover datos.

Un desafío clave para las empresas de hoy es que los usuarios no pueden encontrar rápidamente la información que necesitan para solucionar un problema empresarial o completar una tarea. Con frecuencia, los datos están dispersos en diversos sistemas para soportar aplicaciones específicas gestionadas por diferentes organizaciones. Además, nuevos orígenes de datos están surgiendo como recursos fundamentales que las personas pueden tener que considerar en su trabajo diario y para la toma de decisiones importantes, como las redes sociales, las notificaciones de dispositivos móviles, Twitter, etc.

Como ejemplo, la información del cliente, tal como información de contacto, productos comprados, tiquetes de servicio abiertos e información sobre la garantía, se almacena en distintas aplicaciones empresariales como CRM, sistemas de tiquetes de soporte, portales de marketing, etc. Imagine a un vendedor que desea llamar a un cliente para incrementar las ventas. Primero, tiene que iniciar sesión en 10 aplicaciones para agregar los datos sobre el cliente o hablar con cinco personas para entender toda esta información.

Data Explorer se encarga de este desafío clave. La información se almacena en varios sistemas y silos distintos, pero los usuarios necesitan una forma consistente para ver todos los datos y navegar rápidamente hacia lo que es más relevante para ellos. El desafío es entregar información en el punto de impacto en el que los empleados más la necesitan para tomar decisiones.

**Figura 3. Arquitectura de InfoSphere Data Explorer**



**Integración de BigInsights y Data Explorer**

BigInsights y Data Explorer se complementan entre sí, permitiéndoles a las organizaciones ampliar el alcance de la información que pueden analizar de forma consistente y coherente. Por ejemplo, BigInsights se utiliza frecuentemente para almacenar contenido no estructurado y semiestructurado. Más aún, la necesidad de explorar y navegar por el contenido se está volviendo cada vez más crítica — frecuentemente en una interfaz similar a la de búsqueda. Esto permite que la información sea más consumible para el usuario de línea empresarial. Por ejemplo, si se almacenan datos de equipo, tal vez un usuario final quiera navegar por fechas de contenido, buscar tipos específicos de fallas de equipo, etc. Por otra parte, si se almacenan datos de redes sociales, un usuario final tal vez quiera buscar los sentimientos del usuario en relación con los productos. Todo esto requiere capacidades de indexación enriquecidas. Además de la indexación, Data Explorer puede ofrecer una experiencia enriquecida para el usuario, incorporando contenido de BigInsights y otro contenido empresarial para permitir una exploración completa de big data.

[Volver arriba](http://www.ibm.com/developerworks/ssa/data/library/app_bigdata/872864.html#ibm-pcon)

**Escenario de muestra**

Para implementar esta arquitectura, necesitamos realizar varios pasos. Los resumiremos aquí y los exploraremos con mayor detalle más adelante:

* Recolectar y preparar los datos de redes sociales para el análisis
  + BigInsights proporciona una variedad de mecanismos de recolección de datos a través de aplicaciones preincorporadas. Una vez que las publicaciones de redes sociales basadas en texto residan en BigInsights, es necesario extraer información de interés de forma que pueda ser fácilmente indexada y explorada más adelante. BigInsights proporciona capacidades de análisis de texto sofisticadas para ayudarle a extraer sus entidades de interés, incluyendo productos, personas y sentimientos sobre productos.
* Modele entidades empresariales y relaciones de interés
  + Una aplicación puede impulsar este proceso al especificar un modelo de entidad para Data Explorer para ayudar a establecer diversas opciones de configuración que mostraremos muy pronto. Este modelo de entidad es crítico para el éxito general de su escenario de aplicación.
  + El modelo de entidad capturará el conjunto de entidades y relaciones empresariales importantes que sus analistas empresariales estarán interesados en buscar, descubrir y explorar en Data Explorer. Así, un diseño efectivo de modelo de entidad supondrá una comprensión de cómo y qué querrán buscar y explorar los analistas empresariales.
  + El modelo de entidad capturará el conjunto de configuraciones importantes de su clúster de Data Explorer para reflejar su capacidad y planificación de despliegue. Un poco más adelante, verá cómo capturamos productos y tuiteos como entidades empresariales de interés clave, especificamos con mayor detalle las relaciones entre estas entidades y proporcionamos la información de despliegue de topología del clúster de Data Explorer.
* Desarrolle su primera aplicación de indexado para indexar datos extraídos de redes sociales en Data Explorer
  + Estará listo para desarrollar su aplicación de indexado al aprovechar el ciclo de vida del desarrollo de la aplicación de BigInsights, lo cual le permite crear, publicar y desplegar su aplicación con un esfuerzo mínimo. Una vez desplegada, la extracción de entidad de sus datos de redes sociales será llevada a una colección de búsqueda de Data Explorer y estará lista para una exploración más detallada utilizando la función de búsqueda por facetas de Data Explorer y para compilar una aplicación de vista de 360 grados.
* Utilizando Data Explorer para visualización
  + Data Explorer Application Builder proporciona una forma de compilar una aplicación que reúne la información relevante sobre datos esparcidos en distintos sistemas. En nuestro escenario de muestra, un ejecutivo de planificación de productos podría interesarse en un producto o familia de productos, así que una aplicación de vista de 360 grados podría incluir retroalimentaciones, problemas de producto e interacciones anteriores con el cliente.

[Volver arriba](http://www.ibm.com/developerworks/ssa/data/library/app_bigdata/872864.html#ibm-pcon)

**Recolecte y prepare sus datos de redes sociales en BigInsights para análisis**

BigInsights proporciona una variedad de mecanismos de recolección de datos a través de aplicaciones preincorporadas, como las aplicaciones de Boardreader.

**Figura 4. Aplicaciones de Boardreader**



Puede recolectar sus redes sociales y almacenarlas en BigInsights, aprovechando un número de opciones de almacenamiento, incluyendo los sistemas de archivos distribuidos y el motor de almacenamiento como HBase.

**Figura 5. Sistemas de archivos distribuidos y motores de almacenamiento de BigInsights**



Una vez que las publicaciones de redes sociales basadas en texto residan en BigInsights, es necesario extraer la información de interés de modo que pueda ser fácilmente indexada y explorada posteriormente. BigInsights proporciona capacidades de análisis de texto sofisticadas para extraer sentimientos relacionados con productos y perfiles de usuarios de redes sociales. La siguiente figura muestra fragmentos del resultado de la extracción de entidad y sentimiento de los datos de redes sociales, resaltando algunos campos de interés para esta extracción de entidad, que incluyen *Category, Brand, Product, Source, IsSentiment, IsCustomerOf, Polarity, Created Time, FullName, Screenname,UserID,Text*.

**Figura 6. Sentimiento del tuiteo con respecto al producto**

Image shows Tweet sentiment on product

[Volver arriba](http://www.ibm.com/developerworks/ssa/data/library/app_bigdata/872864.html#ibm-pcon)

**Diseñe y gestione su modelo de entidad de aplicación**

Una vez que tiene la extracción de entidad de análisis de texto de BigInsights de la sección anterior, estará listo para diseñar el modelo de entidad de Data Explorer.

Esta sección cubre un conjunto de elementos de su modelo de entidad que debe considerar en su solución y diseñarlos en su modelo. Este proceso garantizará que su solución de aplicación cumpla con el patrón de acceso y exploración de datos que sus analistas empresariales necesitan para entregar un entorno de búsqueda escalable para sus big data. Resumiremos el conjunto de etapas de este proceso de diseño y las exploraremos con mayor detalle más adelante:

* Determine el conjunto importante de entidades y relaciones empresariales que le interesa soportar para una búsqueda y exploración más detalladas en Data Explorer e identifique los distintos orígenes que estará proporcionando donde podrían estar esparcidas estas entidades empresariales.
* Capture estas entidades y relaciones en el modelo de entidad de su escenario.
* Determine la escalabilidad de su clúster de Data Explorer y diseñe estas especificaciones en su modelo de entidad. Estas especificaciones determinarán la escalabilidad de su despliegue de Data Explorer.
* Despliegue su modelo de entidad en el clúster Zookeeper para la gestión centralizada de valores de configuración.

**Determinación del conjunto de entidades y relaciones empresariales importantes para compilar información contextual**

En nuestro escenario de muestra, hemos acumulado datos internos sobre clientes existentes y nuestros productos. Estos datos se almacenan en una DBMS relacional. Además, hay tuiteos que recolectamos y de los que hemos extraído los sentimientos de los usuarios con respecto a nuestros productos utilizando el análisis de texto de BigInsights. Nuestros analistas empresariales tal vez quieran tener una visión más completa de la forma en que los clientes perciben nuestros productos y qué tan visibles son esos productos en el mercado en general. Combinar datos empresariales y de redes sociales puede así proporcionar más información a nuestros analistas empresariales. Identificamos que las siguientes entidades serán de gran interés para nuestro analista empresarial:

* Sentimiento de usuarios sobre productos extraído de datos de redes sociales en BigInsights
* Datos de productos almacenados en una base de datos relacional
* Clientes online almacenados en una base de datos relacional

Igualmente importante es la necesidad de proporcionar la información contextual correcta a los analistas empresariales. Para alcanzar esta meta, es necesario definir el conjunto de relaciones entre las entidades. La relación es el elemento crucial que utiliza Data Explorer Application Builder para enlazar la interacción entre entidades y ofrece el beneficio fundamental de compilar la información contextual. Por ejemplo, en nuestro escenario, necesitamos capturar el hecho de que los tuiteos están asociados con usuarios específicos (clientes) y algunos tuiteos pueden estar relacionados con productos.

**Cómo capturar estas entidades y cualquier relación importante en el modelo de entidad**

El modelo de entidad de Data Explorer está en formato XML. Utilice el editor de XML de su preferencia para crear un nuevo archivo para el modelo de entidad:

* Añadir la entidad *sentiment:*
  + El fragmento para añadir la entidad sentiment al modelo de entidad se verá como el siguiente listado. Incluye la información adicional para los campos que deseamos capturar a fin de que puedan ser utilizados en la aplicación de búsqueda que Data Explorer Application Builder puede compilar.

**Listado 1. Definición de la entidad *sentiment***

<entity-definition default-searchable="true" identifier="@hash"

name="tweet" store-name="tweet-search-store">

<field external-name="Category" name="Category"/>

<field external-name="Brand" name="Brand"/>

<field external-name="Product" name="Product"/>

<field external-name="IsSentiment" name="IsSentiment"/>

<field external-name="Polarity" name="Polarity"/>

<field external-name="CreatedTime" name="CreatedTime"/>

<field external-name="Screenname" name="Screenname"/>

</entity-definition>

* Añadir la entidad *product:*
  + El fragmento para añadir la entidad product y los campos relacionados con el modelo de entidad se verá como el siguiente listado:

**Listado 2. Definición de la entidad *product***

<entity-definition default-searchable="true" identifier="@hash"

name="product" store-name="product">

<field external-name="BRAND" name="BRAND"/>

<field external-name="PRODUCT\_NUMBER" name="PRODUCT\_NUMBER"/>

<field external-name="PRODUCT\_BRAND\_CODE" name="PRODUCT\_BRAND\_CODE"/>

<field external-name="PRODUCT\_DESCRIPTION" name="PRODUCT\_DESCRIPTION"/>

</entity-definition>

* Añadir las relaciones esenciales entre nuestras entidades:
  + La relación es el elemento crucial que Data Explorer Application Builder utiliza para enlazar la interacción entre entidades y ofrece el beneficio fundamental de compilar la información contextual. Como ejemplo para nuestro escenario, tal vez deseemos capturar el hecho de que algunos tuiteos están relacionados con productos. Las definiciones de relación podrían verse como el siguiente listado:

**Listado 3. Definición de relación**

<entity-definition default-searchable="true" identifier="@hash"

name="product" store-name="product">

...

<association-definition name="feedback" to="FEEDBACK\_TYPE">

<link from-field="BRAND" fuzzy="false" to-field="Brand"/>

</association-definition>

</entity-definition>

**Definición de las especificaciones de topología para su clúster de Data Explorer**

* Especificar el almacenamiento de colección de clúster para Data Explorer:
  + Una vez que ha identificado las entidades de interés y las relaciones, es necesario desarrollar un índice para soportar la búsqueda, el descubrimiento y el análisis. Para ello, es necesario especificar un mecanismo de almacenamiento — un*almacenamiento de colección*— para este índice. Para casos de uso que involucran a BigInsights, como nuestro escenario, deseamos utilizar un almacenamiento de colección de clúster. Es uno de los muchos tipos de almacenamientos de colección soportados por Data Explorer. Elegir el tipo de almacenamiento de colección de clúster le permitirá al motor de Data Explorer aprovechar un clúster de máquinas para escalar de forma horizontal para manejar una escala mayor de indexación para datos de BigInsights.
  + El siguiente fragmento muestra cómo especificar el almacenamiento de colección de clúster para indexar datos de redes sociales provenientes de BigInsights. La otra entidad que representa los datos de la DBMS relacional utilizará la colección simple más común.

**Listado 4. Almacenamiento de colección de clúster para datos de BigInsights**

<cluster-collection-store activity-collection="false"

collection-name="tweet-search-store"

monitor-activities="false"

name="tweet-search-store"

base-collection="default-push"

n-shards="2"/>

<collection-store activity-collection="false"

collection-name="gssdb-product"

monitor-activities="false"

name="product"/>

* Especificar la escalabilidad del clúster de Data Explorer:
  + Añadir particiones a la aplicación de búsqueda permite que los datos sean particionados de forma horizontal, particularmente cuando estas particiones están esparcidas en múltiples instancias físicas de Data Explorer. El rendimiento general puede incrementarse al manejar grandes cantidades de datos a medida que las operaciones de indexación y búsqueda son distribuidas en un entorno en clúster. Al reutilizar el ejemplo de modelo de entidad anterior, especificamos el número de particiones en nuestro almacenamiento de colección de clúster y las dividimos entre dos instancias físicas distintas de Data Explorer.

**Listado 5. Especificación de la escalabilidad para datos de BigInsights**

<velocity-instance url="http://velocity1.domain.com:9080/vivisimo/cgibin/

velocity?v.app=api-soap&amp;wsdl=1&amp;use-types=true&amp;"

username="api-user"

password="password">

<serves name="tweet-search-store"

shard="1"

n-shards="2"

port="9081"/>

<serves name="tweet-search-store"

shard="2"

n-shards="2"

port="9082"/>

</velocity-instance>

<velocity-instance url="http://velocity2.domain.com:9080/vivisimo/cgibin/

velocity?v.app=api-soap&amp;wsdl=1&amp;use-types=true&amp;"

username="api-user"

password="password">

<serves name="tweet-search-store"

shard="1"

n-shards="2"

port="9081"/>

<serves name="tweet-search-store"

shard="2"

n-shards="2"

port="9082"/>

</velocity-instance>

**Uso de ZooKeeper para gestionar el modelo de entidad de Data Explorer**

Data Explorer utiliza ZooKeeper para gestionar el modelo de entidad de su aplicación. Zookeeper es un servicio centralizado para mantener la información de configuración, proporcionar sincronización distribuida y proporcionar servicios de grupo. Ahora que hemos definido nuestro modelo de entidad de aplicación, necesitamos ponerlo a disponibilidad de la aplicación subiéndolo a un clúster de ZooKeeper. Este clúster de configuración será utilizado por la aplicación para descubrir la topología de despliegue en uso:

* Subida del modelo de entidad al clúster de ZooKeeper:
  + Una vez que el clúster de ZooKeeper esté configurado, puede subir y gestionar su modelo de entidad de aplicación de Data Explorer utilizando este clúster de ZooKeeper. El JAR bigindex incluido en la carpeta lib de Data Explorer BigIndex API ZIP es un ejecutable que se puede utilizar como herramienta de línea de comando básica para subir y gestionar el modelo de entidad en ZooKeeper. A continuación, se muestra el uso de la línea de comando.

**Listado 6. Subida del modelo de entidad al clúster de ZooKeeper**

java -jar bigindex-2.0.0.jar

--properties-file zookeeper.properties

--import-file scenario\_entity\_model.xml

--export-to-screen --legacy-model

Tenga en cuenta que si utiliza la IU administrativa de Data Explorer Application Builder para gestionar el modelo de entidad de aplicación, puede omitir la etapa anterior y, alternativamente, dirigir su aplicación hacia la misma instancia de servidor y espacio de nombres de ZooKeeper utilizados por Data Explorer Application Builder. Puede encontrar más detalles sobre la configuración de ZooKeeper que se está utilizando en el archivo zookeeper.yml ubicado en IBM/IDE/AppBuilder/wlp/usr/servers/AppBuilder/apps/AppBuilder/WEB-INF/config.

[Volver arriba](http://www.ibm.com/developerworks/ssa/data/library/app_bigdata/872864.html#ibm-pcon)

**Desarrollo de la primera aplicación de indexación de BigInsights con Data Explorer**

Una vez que haya terminado de diseñar el modelo de entidad de Data Explorer, estará listo para aprovechar el ciclo de vida del desarrollo de la aplicación de BigInsights para desarrollar la primera aplicación de indexación con el objetivo de llevar sus datos de redes sociales a una colección de búsqueda en Data Explorer. La infraestructura de la aplicación de BigInsights le permite crear, publicar y desplegar la primera aplicación de indexación con un mínimo esfuerzo.

**Cómo crear un proyecto de BigInsights y una nueva clase Java**

Es necesario crear un proyecto apropiado para su aplicación, como es de esperar de cualquier esfuerzo de desarrollo de aplicaciones basado en Eclipse. Vea el artículo "Developing, publishing, and deploying your first Big Data application with InfoSphere BigInsights" para conocer las etapas rápidas para crear un proyecto de BigInsights (vea [Recursos](http://www.ibm.com/developerworks/ssa/data/library/app_bigdata/872864.html#resources)). Después de crear el proyecto de BigInsights, deberá añadir una nueva clase Java al proyecto. Para hacerlo, desde su entorno de Eclipse, seleccione **File > New > Java > Class**. Complete la información para su clase (nombre del paquete, etc.) . Cuando termine, haga clic en **Finish**.

**Cómo utilizar APIs de BigIndex nuevas para indexar datos de BigInsight en Data Explorer**

Su aplicación invocará un conjunto de APIs de Java indexadas proporcionadas por Data Explorer (APIs de BigIndex) para mover los datos de BigInsights. Los siguientes pasos muestran las diversas piezas clave de las APIs para lograr esta meta:

* Recuperación de la topología de despliegue de Data Explorer para indexación:
  + Como mencionamos en la sección anterior sobre el modelo de entidad, la topología de clúster de Data Explorer es capturada en el modelo de entidad subido a ZooKeeper. La aplicación de indexación necesitará establecer una conexión con este clúster de ZooKeeper para recuperar la topología y ubicar el clúster de Data Explorer para la indexación. El siguiente listado es un fragmento de código que permite llevar a cabo esta tarea.

**Listado 7. Establecimiento de la conexión con el clúster de ZooKeeper**

ZookeeperConfiguration zookeeperConfiguration =

new ZookeeperConfiguration("namespace\_sample\_big\_data\_app",

new ZookeeperEndpoint("zkhost1.domain.com", 2181));

* Uso de un analizador de campo para procesar el formato de datos de entrada:
  + Una vez que la instancia de Data Explorer que utilizaremos para la indexación esté establecida, estaremos listos para procesar nuestros datos de entrada para la indexación. Para nuestro escenario de muestra, los datos de redes sociales son capturados en formato CSV. Tal vez, necesitemos utilizar el analizador OpenCSV de código abierto para analizar cada archivo CSV y procesar cada fila en una lista de valores clave. Esto preparará los datos en el formato correcto para el motor de indexación de Data Explorer. A continuación, hay un fragmento de código para mostrar la lógica de aplicación para analizar datos CSV. **Nota:** deberá utilizar el analizador de campo apropiado para procesar el formato de datos de sus datos de entrada para indexación.

**Listado 8. Código de muestra para analizar datos CSV**

// Read each CSV input file stored on BigInsights HDFS

for (FileStatus fStatus : listFilesFromHDFS(inputDirectory)) {

// For each CSV file, parse each row into a list of key values

CSVReader reader =

new CSVReader(new InputStreamReader(fs.open(fStatus.getPath())));

// For each key, we will show later how to index it into

// the Data Explorer index record

while((listOfFields = reader.readNext())!= null){

...

* Definición del esquema de registro para datos de redes sociales de muestra:
  + El indexador de Data Explorer también espera que la aplicación defina el esquema de un registro de índice. El siguiente listado muestra el fragmento de código de Java utilizando la API BigIndex de Data Explorer para definir el esquema de registro para los diversos campos clave de nuestros datos de Tweet. **Nota:** en la siguiente llamada a addRecordType(), el valor de entrada debe coincidir con el nombre de la entidad tal como se define en el modelo de entidad. En nuestro escenario, el nombre de entidad es "tweet".

**Listado 9. Definición del esquema de registro**

RecordSchema recordSchema = new RecordSchemaBuilder()

.addRecordType("tweet")

.addTextField("Category").retrievable(true).sortable(true)

.addTextField("Brand").retrievable(true).sortable(true)

.addTextField("Product").retrievable(true).sortable(true)

.addTextField("isSentiment").retrievable(true).sortable(true)

.addDateField("CreatedTime").retrievable(true).sortable(true)

.addTextField("Screenname").retrievable(true).sortable(true)

.build();

* Indexación de los registros en el motor de Data Explorer:
  + Ahora que finalizó la definición del esquema de registro, podemos añadir cada campo del registro en el índice y continuar con el siguiente registro hasta que haya terminado de indexar todos los registros de los datos de tuiteos. A continuación, se muestra el fragmento de código para la indexación de registros en el motor de Data Explorer. **Nota:** en la llamada a newRecordBuilder(), el valor debe coincidir con el nombre de la entidad tal como se define en el modelo de entidad. Por ejemplo, en nuestro escenario de muestra, el nombre de entidad es "tweet".

**Listado 10. Indexación de registros**

RecordBuilderFactory recordBuilderFactory =

new RecordBuilderFactory(recordSchema);

// In the following call, provide the name of

// the entity as defined in the entity model

RecordBuilder recordBuilder =

recordBuilderFactory.newRecordBuilder("tweet");

// For each CSV row that's been parsed into a list of fields

while((listOfFields = csvreader.readNext())!= null){

recordBuilder.id(String.valueOf(recordId++));

// For each field, set field name and field value

while (i != listOfFields.length){

String fieldName = listOfFieldNames[i];

String fieldValue = listOFields[i];

// Add the field to the indexing record

recordBuilder = recordBuilder.addField(fieldName, fieldValue);

...

}

}

// Finally, call to generate the record with the

// current data and add it to the indexer

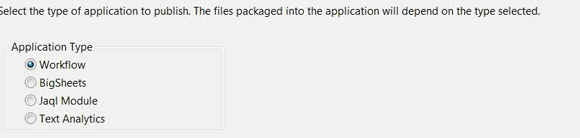
RequestStatus status = indexer.addOrUpdateRecord(recordBuilder.build());

**Cómo publicar y desplegar la aplicación de indexación**

Después de desplegar la aplicación de indexación, estará listo para publicarla en el catálogo de aplicaciones de BigInsights. Empaquetar y publicar la aplicación de indexación le permite definir el flujo de trabajo de la aplicación, especificar los parámetros como datos de entrada (sus datos de redes sociales) y el punto final de ZooKeeper de Data Explorer. Vea el artículo "Developing, publishing, and deploying your first Big Data application with InfoSphere BigInsights" para conocer los pasos rápidos para crear un proyecto de BigInsight (vea [Recursos](http://www.ibm.com/developerworks/ssa/data/library/app_bigdata/872864.html#resources)) y obtener una visión general de los pasos para publicar la aplicación de BigInsights. Durante este proceso de publicación, especificará la siguiente información para la aplicación de indexación:

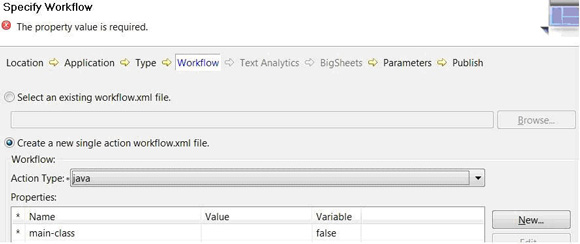
* Tipo de aplicación:
  + Seleccione el flujo de trabajo del tipo de aplicación, como se muestra en la siguiente Figura.

**Figura 7. Tipo de aplicación**



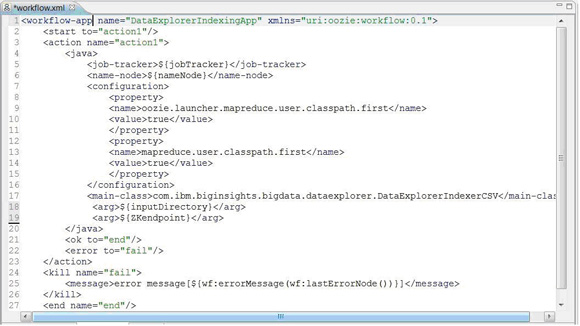
* Definición del flujo de trabajo de Oozie:
  + La consola Web de BigInsights genera un flujo de trabajo de Oozie para ayudar a gestionar los trabajos de MapReduce. En la pestaña Workflow, acepte los valores predeterminados para que el asistente cree una nueva acción: workflow.xml. En el menú desplegable, cambie el tipo de flujo de trabajo a Java, como se muestra a continuación.

**Figura 8. Tipo de acción de flujo de trabajo de Oozie**



* Indexación de parámetros de aplicación:
  + En la página Parameters, especifique el parámetro para su aplicación de indexación, incluido el directorio de entrada. Opcionalmente, puede proporcionar la información de punto final de ZooKeeper como un parámetro de entrada para la aplicación de indexación, en lugar de tenerla como código en la aplicación. El flujo de trabajo final puede verse como la siguiente figura.

**Figura 9. Muestra de flujo de trabajo de Oozie**



**Configuración de bibliotecas de cliente de Data Explorer en el clúster de BigInsights**

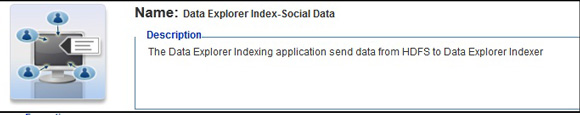
Antes de ejecutar la aplicación de indexación, es necesario configurar ciertas bibliotecas de cliente de Data Explorer en el clúster de BigInsights.

* Copie la carpeta install-dir/AppBuilder/bigindex.zip desde la instalación de su clúster de Data Explorer en el sistema local de archivos de clúster de BigInsights.
* Descomprima la carpeta bigindex.zip. Verá la lista de los archivos JAR de dependencia de Data Explorer.
* Cree un directorio de HDFS como el siguiente: /biginsights/oozie/sharedLibraries/DataExplorer.
* Copie los archivos JAR de dependencia de Data Explorer en el directorio HDFS /biginsights/oozie/sharedLibraries/DataExplorer utilizando el comando de copia de Hadoop (por ejemplo, hadoop fs -copyFromLocal \*jar /biginsights/oozie/sharedLibraries/DataExplorer/) o utilice la consola de BigInsights para subir los archivos al directorio de HDFS.

**Supervisión de la aplicación de indexación**

Una vez que la aplicación está desplegada, aparecerá en la consola Web de BigInsights como se muestra en la siguiente figura. Puede utilizar la consola Web para inspeccionar detalles de la aplicación y ejecutarla. Puede aprender más sobre la consola Web de BigInsights para supervisar el flujo de trabajo en el artículo "Exploring your InfoSphere BigInsights cluster and sample applications" (vea [Recursos](http://www.ibm.com/developerworks/ssa/data/library/app_bigdata/872864.html#resources)).

**Figura 10. Aplicación de indexación de BigInsights**



[Volver arriba](http://www.ibm.com/developerworks/ssa/data/library/app_bigdata/872864.html#ibm-pcon)

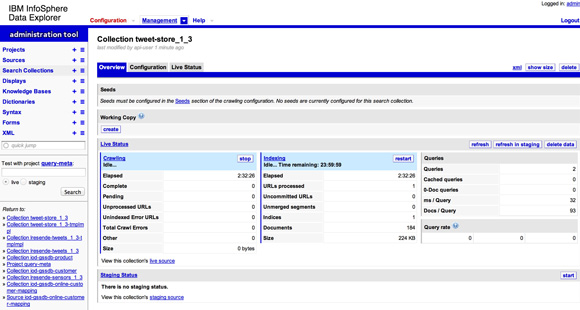
**Visualización con Data Explorer**

**Verificación de los datos de redes sociales en el índice de Data Explorer**

Una vez que sus datos de redes sociales pasan de BigInsights a una colección de búsqueda en Data Explorer, debe poder utilizar la IU administrativa de Data Explorer Engine para inspeccionar los datos indexados. Por ejemplo, puede verificar visualmente que los distintos campos de interés estén indexados correctamente. Para acceder a la IU de administrador, siga estos pasos:

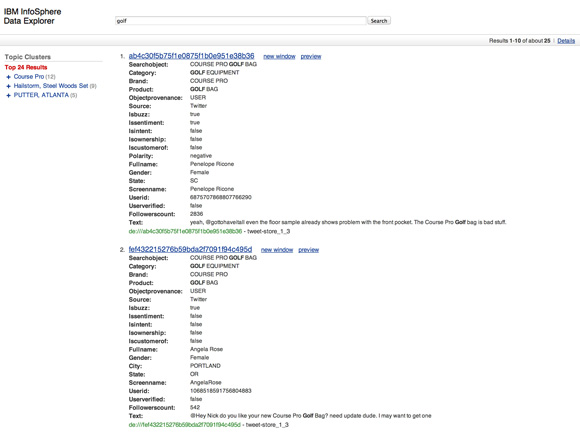
* Inicie sesión en la IU administrativa de Data Explorer Engine.
* Seleccione **Search Collection** en el menú izquierdo.
* Busque el almacenamiento de colección para los datos de redes sociales tal como se especifica en su modelo de entidad.
* Abra la colección de búsqueda y haga clic en el botón **Search** en el panel izquierdo.

**Figura 11. IU administrativa de Data Explorer Engine**



* Búsqueda de tuiteos de usuarios sobre un producto:
  + En el recuadro de búsqueda, un usuario puede escribir palabras clave como golf para realizar una búsqueda de texto utilizando la interfaz existente como se muestra en la siguiente figura.

**Figura 12. Búsqueda de mediante texto de tuiteos de usuarios relacionados con el golf**

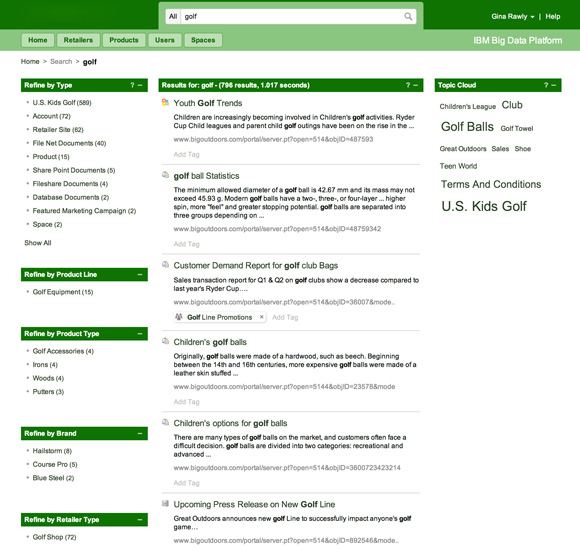


**Aprovechamiento de Data Explorer Application Builder**

Application Builder es nuevo en Data Explorer y proporciona la infraestructura para desarrollar una aplicación de exploración de datos convincente, como búsqueda por facetas, así como aplicaciones de información de 360 grados que pueden reunir la información relevante sobre los datos esparcidos entre los sistemas.

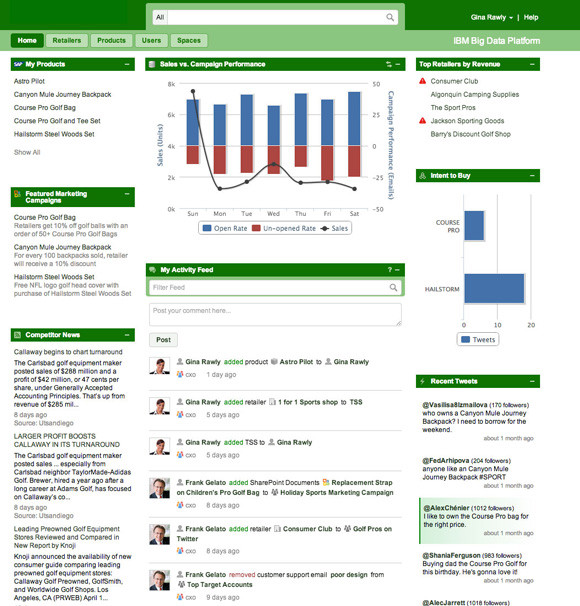
* Búsqueda por facetas:
  + La siguiente figura ilustra el widget de búsqueda que se puede desarrollar utilizando Application Builder para proporcionar una aplicación de búsqueda por facetas intuitiva para que los usuarios exploren sus datos de redes sociales. La búsqueda por facetas permite navegar fácilmente por los conjuntos de resultados de un tema específico utilizando un conjunto de refinamientos. En este ejemplo, exploramos los tuiteos de usuarios sobre productos de golf.

**Figura 13. Búsqueda por facetas de tuiteos de usuarios relacionados con el golf**



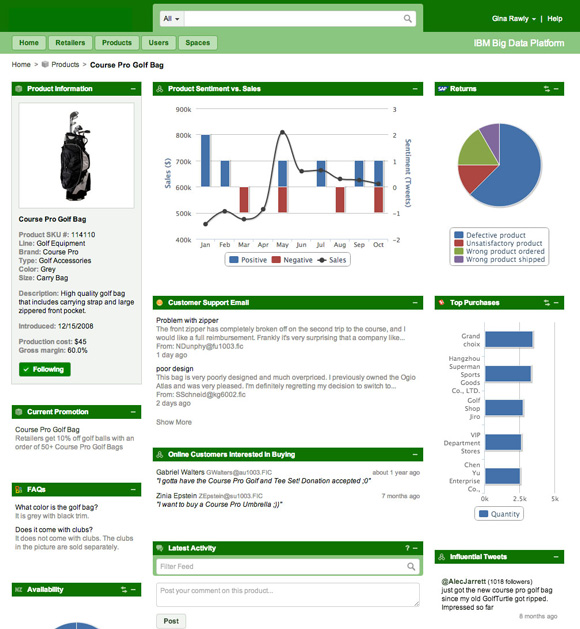
* Aplicación de información de 360 grados:
  + Después de explorar los aspectos de los datos de redes sociales, también pueden relacionarse con tipos adicionales de datos, tales como datos de clientes o de productos, que se pueden extraer de otros sistemas. Data Explorer proporciona conectividad y capacidades de rastreo para diversas bases de datos relacionales, sistemas de CRM empresariales, recursos compartidos de archivos y otros. Data Explorer Application Builder ofrece la forma de desarrollar una aplicación de visualización de 360 grados que reúne la información relevante sobre datos esparcidos entre sistemas dispares, todo dejando los datos donde residen originalmente.

**Figura 14. Aplicación de visualización de 360 grados**



En nuestro escenario de muestra, un analista de marketing podría estar interesado en un producto o familia de productos, de modo que una aplicación de visualización de 360 grados podría incluir retroalimentación de usuarios y detalles de productos. La siguiente figura ilustra una página de producto donde se reúnen múltiples widgets, mostrando la información de producto con comentarios relacionados de los usuarios. Para obtener más información sobre el desarrollo de páginas de entidad con múltiples widgets que se muestran aquí, vea [Recursos](http://www.ibm.com/developerworks/ssa/data/library/app_bigdata/872864.html#resources).

**Figura 15. Aplicación de visualización de 360 grados — Página de entidad**



Tenga en cuenta que las figuras 14 y 15 muestran una aplicación de 360 grados e incluyen datos de distintos sistemas que no fueron mencionados en este artículo.

[Volver arriba](http://www.ibm.com/developerworks/ssa/data/library/app_bigdata/872864.html#ibm-pcon)

**Resumen**

Este artículo exploró una arquitectura de software que les permite a los analistas empresariales explorar datos derivados de una variedad de orígenes dispares con facilidad y eficiencia. En particular, examinamos cómo InfoSphere Data Explorer puede indexar datos grandes de redes sociales gestionados por InfoSphere BigInsights, así como también datos estructurados gestionados por los orígenes de datos empresariales más tradicionales. La indexación de los datos permite un acceso eficiente, mientras que las capacidades de búsqueda por facetas de Data Explorer proporcionan una forma intuitiva de explorar estos datos, analizar relaciones y obtener conocimientos para quienes no son programadores.