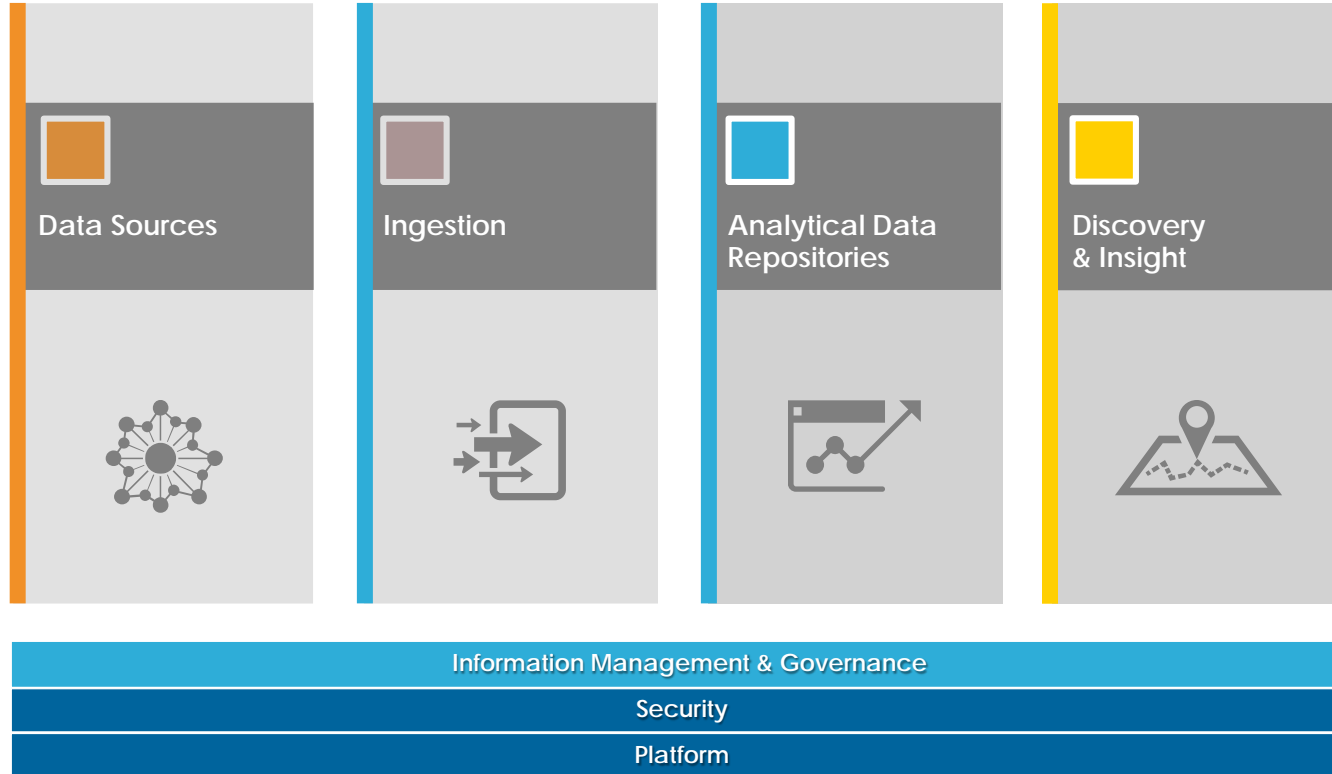


Buscadores

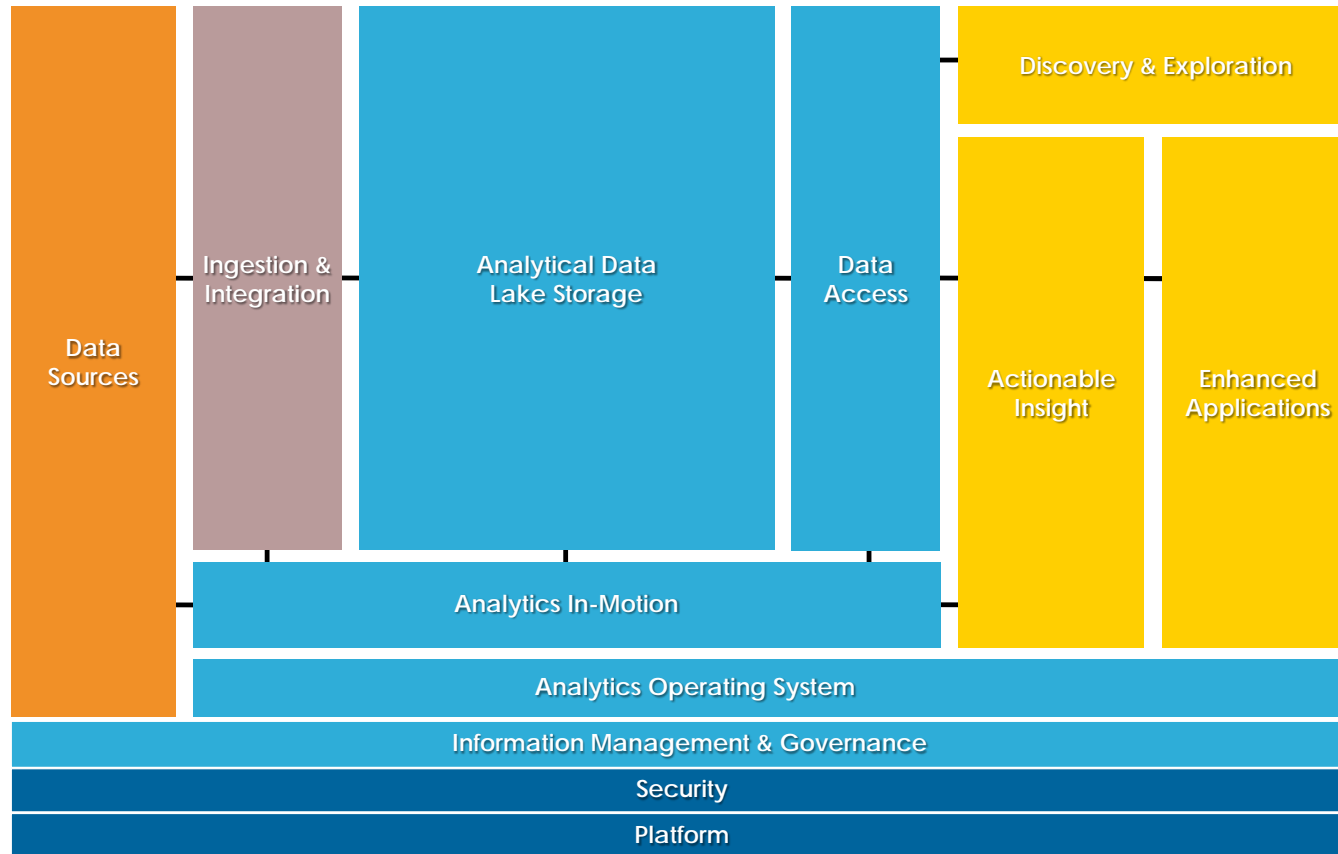
Sesion I

IBM Analytics Platform

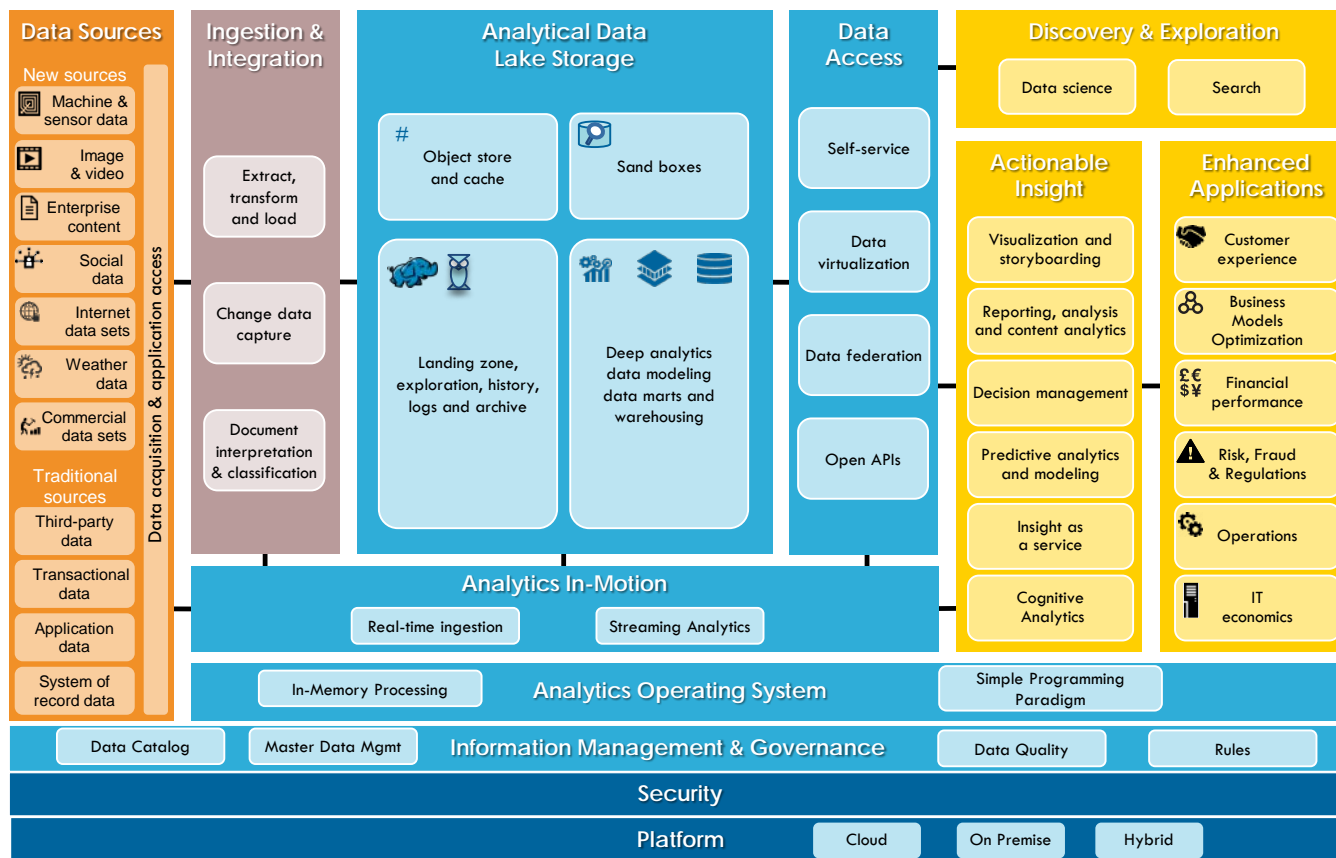
IBM Analytics platform overview



Componentes



Capacidades

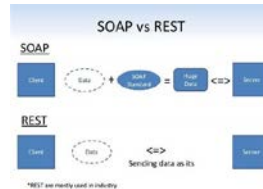


¿Y en proyectos de Big Data?

- Si prestamos especial atención al término **Big**:
 - Haremos uso de las tecnologías que están disponibles para sacar partido de los grandes volúmenes de datos
- Si prestamos especial atención al término **Data**:
 - Dialogaremos con los datos con el fin de entenderlos e interpretar qué nos dicen. Buscaremos patrones de conocimiento
 - Recurriremos a esos datos para extraer modelos o utilizarlos como respaldo para el resultado que queramos mostrar

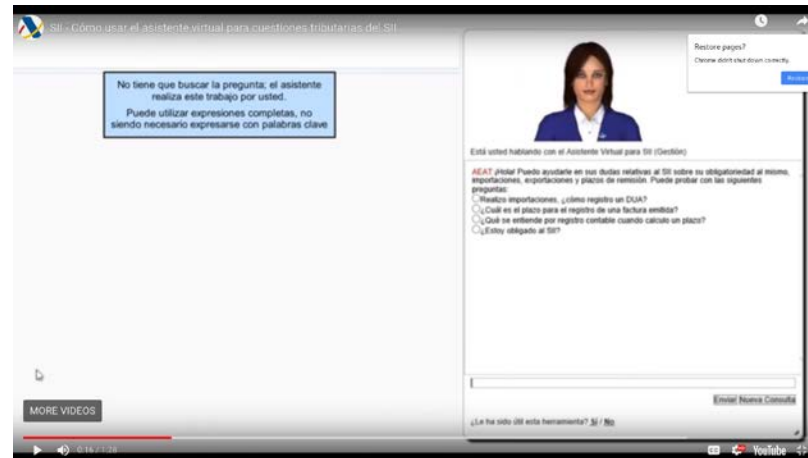
Una solución analítica

- Deberá prestar especial atención a la parte de explotación:
 - Habilitando el consumo del conocimiento específico adquirido
 - Mostrando el conocimiento de forma global, en toda su dimensión
 - Permitiendo al usuario final que profundice o sea capaz de ver la información original y ser capaz de interpretarla por él mismo.



Por ejemplo...

- Un asistente virtual para cuestiones tributarias
 - El agente va haciendo preguntas con el fin de desambiguar la intención inicial del solicitante cuando es necesario
 - En el momento final **se recurre a la documentación disponible para respaldar su recomendación**



Patrón común de esas soluciones

- Es habitual que en algún paso del flujo de información se incluya un buscador (o recuperador de información)
 - Tecnología que habilita el almacenamiento de información no estructurada y que
 - Permita su consulta rápida con un lenguaje con características adaptadas a la fuente de la información
- Es importante diferenciarlo de bases de datos documentales cuya principal funcionalidad es soportar el ciclo de vida del documento en toda su dimensión, no solo actuar como repositorio y consulta... aunque en muchos casos la frontera es difusa y también su uso en los proyectos

Introduction to Information Retrieval: <https://nlp.stanford.edu/IR-book/>

¿Y qué hay en el mercado?

DB-Engines Ranking of Search Engines

The DB-Engines Ranking ranks database management systems according to their popularity. The ranking is updated monthly.

This is a partial list of the [complete ranking](#) showing only search engines.

Read more about the [method](#) of calculating the scores.



☐ include secondary database models

21 systems in ranking, December 2021

Rank			DBMS	Database Model	Score		
Dec 2021	Nov 2021	Dec 2020			Dec 2021	Nov 2021	Dec 2020
1.	1.	1.	Elasticsearch	Search engine, Multi-model	157.72	-1.36	+5.23
2.	2.	2.	Splunk	Search engine	94.32	+2.02	+7.32
3.	3.	3.	Solr	Search engine, Multi-model	57.72	+3.87	+6.48
4.	4.	4.	MarkLogic	Multi-model	8.94	-0.40	-2.00
5.	5.	5.	Algolia	Search engine	8.24	+0.03	+0.41
6.	6.	7.	Sphinx	Search engine	8.01	+0.10	+1.69
7.	7.	6.	Microsoft Azure Search	Search engine	7.15	-0.22	+0.30
8.	9.	10.	Virtuoso	Multi-model	5.07	+0.25	+2.48
9.	8.	8.	ArangoDB	Multi-model	4.75	-0.35	-0.76
10.	10.	9.	Amazon CloudSearch	Search engine	2.21	-0.03	-0.85
11.	11.	12.	CrateDB	Multi-model	0.91	+0.01	+0.02
12.	12.	11.	Xapian	Search engine	0.76	-0.02	-0.25
13.	13.	13.	Alibaba Cloud Log Service	Search engine	0.56	-0.02	+0.15
14.	14.	14.	SearchBlox	Search engine	0.35	+0.01	-0.04
15.	15.	16.	Weaviate	Search engine	0.14	+0.00	+0.09
16.	16.	15.	Manticore Search	Search engine	0.06	+0.00	-0.01
17.	17.	17.	Exorbyte	Search engine	0.04	-0.01	+0.01
18.	18.	18.	FinchDB	Multi-model	0.03	+0.00	0.00
19.	19.	20.	Indica	Search engine	0.00	±0.00	±0.00
19.	19.	20.	Rizhiyi	Search engine, Multi-model	0.00	±0.00	±0.00
19.	19.	19.	searchxml	Multi-model	0.00	±0.00	-0.01

<https://db-engines.com/en/ranking/search+engine>

¿Qué es Solr (pronounced as “Solar”)?



- Solr es un motor de búsqueda completo
 - Está combinado con Lucene (el motor de indexación que lo respalda)
 - Sus principales características incluyen búsqueda de texto completo [por palabras clave], búsqueda facetada, búsqueda por campos, resaltado de aciertos, agrupación dinámica, integración de bases de datos y manejo de documentos enriquecidos (p. ej., Word, PDF)..
 - SolrCloud permite la búsqueda distribuida (índice o fragmentos particionados) y la replicación de índices. Como resultado, Solr es altamente escalable.
 - Se puede usar con o sin HDFS
 - Fácil de instalar y utilizar

<https://lucene.apache.org/solr/>

¿Qué aporta cada componente?



➤ Solr

- Capacidades de búsqueda de texto
- Optimizado para tráfico web alto
- Basado en estándares: XML, JSON, HTTP...
- Interfaz de administración vía navegador
- Estadísticas disponibles vía JMX
- Escalabilidad casi lineal



➤ Lucene

- Motor de indexación inversa de alto rendimiento
- Basado en Java
- Algoritmos de búsqueda potentes, precisos y eficientes
- Búsqueda por campos y por rangos. Facetado flexible, agrupación de resultados
- Arquitectura de pluggins o transformadores



Habitualmente, en entornos RDBMS

Documento	Contenido
1	Esta es la primera línea
2	Y esta es ya la primera, digo, la segunda
3	¿Llegaremos a la tercera?
4	Disfruta, anda, que ya llegó el final

- Se extrae de los documentos la información que se quiere indexar
- Buscar un texto supone conocer la/s columnas en las que buscar y, la consulta, supone leer cada registro (todos) y buscar en los registros el texto deseado
- La indexación, si se hace, es directa y se hace por cada columna en la que se plantea realizar la búsqueda

La indexación inversa, por el contrario...

Contenido	Índice
es	[1,2], [2,3]
y	[2,1]
la	[1,3], [2,5], [3,3]
primera	[1,4], [2,6]
...	[<i>documento, posición</i>],

- Registra de cada palabra en qué documento se encuentra y las veces que ocurren
- El índice puede ser realmente grande y se hace necesario, por optimización, el *repartirlo* entre nodos o sistemas que colaboren juntos: sharding
- Por disponibilidad, además queremos que se repliquen para que un índice siempre esté disponible aunque algún nodo se caiga

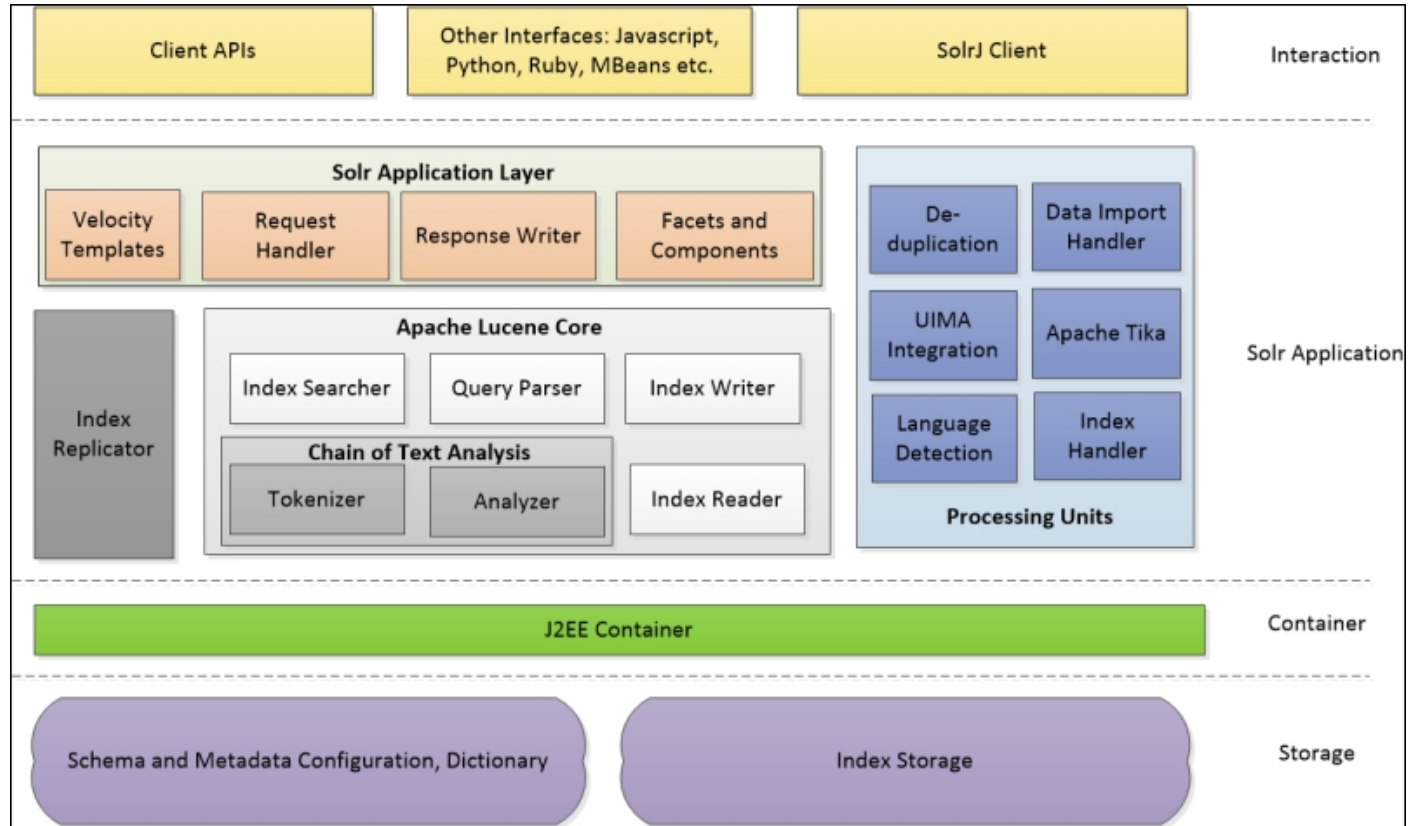
¿Pero es suficiente?

- Consideremos un texto como “El envío del correo se hace así...”
- Queremos que el buscador sea capaz de responder si la búsqueda es: “¿Como enviar correo?”
- ¿Lo habría hecho siguiendo el procedimiento anterior?...
- Se hace necesario añadir pasos para optimizar la búsqueda. Por ejemplo:
 - Eliminar puntuación, acentos
 - Conjuguar verbos, diferenciar singular y plural, etc.
 - Añadir sinónimos
 - Tokenizar
 - Etc.
- Y esos pasos se deben seguir tanto a la hora de indexar como a la hora de buscar

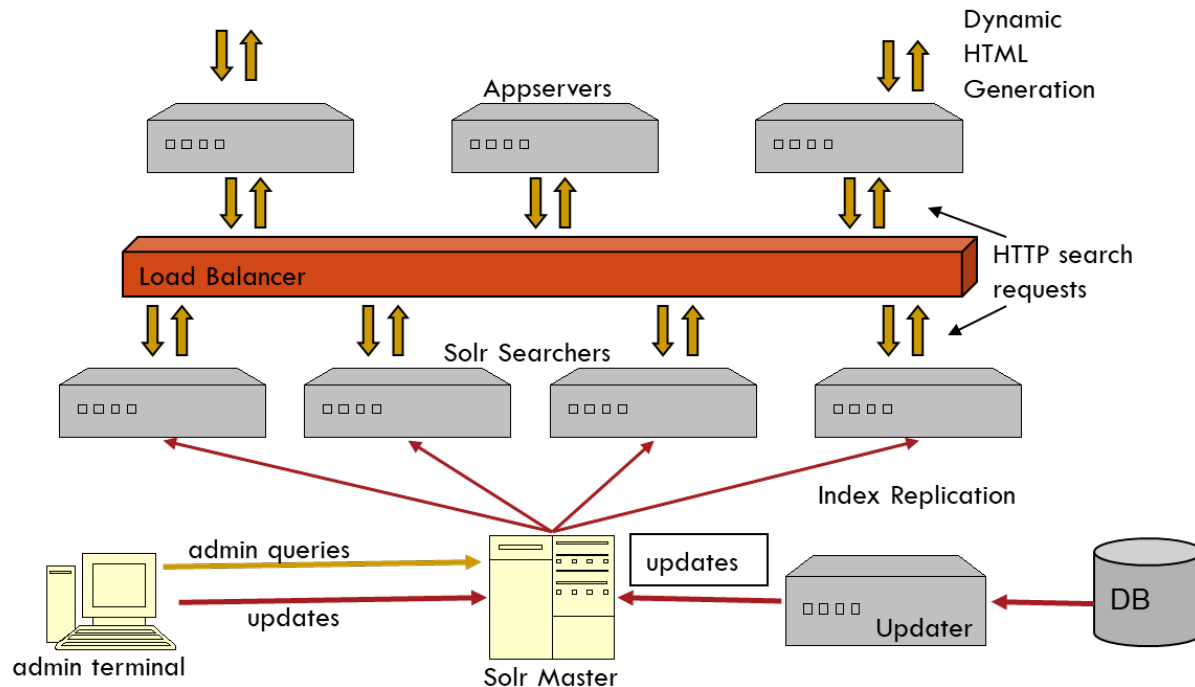
¿Y se resuelven todos los casos?

- Hay ciertas búsquedas que no se resuelven del todo bien, como por ejemplo el NOT IN
- Tener una gran distribución de los índices puede ser un inconveniente importante
- A veces las transformaciones realizadas en la indexación pueden ser diferentes a las de la búsqueda. Se hace necesario poder diferenciarlas y, además, optimizarlas

Arquitectura por componentes



En cloud y Alta disponibilidad




Visualización por facetas

DESKTOPS

You found 1045 items for System type: [Budget desktop system](#)
Too few results? Click a link above to remove that filter, or [remove all filters](#).

Find by price <ul style="list-style-type: none">▸ Less than \$400 (76)▸ \$400 to \$699 (337)▸ \$700 to \$999 (468)▸ \$1000 to \$1299 (5)	Find by manufacturer <ul style="list-style-type: none">▸ Dell, Inc. (43)▸ Lenovo (490)▸ HP (342)▸ Acer America Corp. (28)▸ Cyberpower Inc (22)▸ See all manufacturers	Find by processor manufacturer <ul style="list-style-type: none">▸ Intel (804)▸ AMD (122)▸ Motorola (1)	Or find by <ul style="list-style-type: none">▸ Clock speed▸ Graphics processor▸ RAM installed▸ Hard drive size▸ OS provided▸ See all
--	---	--	--

Sort by: [Product name](#) | [Lowest price](#) | [Editors' rating](#) | [Review date](#) Check products to [Compare](#) ↓



Reviewed on 05/05/2006

Dell Dimension B110 Desktop Computer for Home (Cel-D 2.53GHz/160GB/512MB)

Dell's entry-level Dimension B110 series features aging technology and a dated design, but its members will suffice as second PCs for basic tasks.


Specs: Celeron D (2.53 GHz), 512 MB, 160 GB, 15 in, Microsoft Windows XP Home Edition

⊕ [Add to my products](#) **New!** [What is this?](#)

\$479
at 1 store

[Check prices](#)

COMPARE >>> ☐



Dell Dimension B110 Desktop Computer for Home (Cel-D 2.53GHz/80GB/256MB)

Dell's entry-level Dimension B110 series features aging

\$349
at 1 store

[Check prices](#)

COMPARE >

Ejemplos de sintaxis de búsquedas

- Términos con campos y frases
 - Title:*right* and text: go
 - Title:*right* and go (go deberá aparecer en el campo “text” por defecto)
 - Title: “*the right way*” and go
- Proximidad
 - “*quick fox*”~4
- Wildcards
 - *pla?e* (válida para las palabras *plate* o *place* o *plane*)
 - *practic** (válida para las palabras *practice* o *practical* o *practically*)
- Fuzzy (considera la distancia máxima de letras para validar el resultado)
 - *plantin*~0.75 (válida para *granting* or *planning*)
 - *roam*~ (por defecto es 0.5)

Ejemplos de sintaxis de búsquedas

- Rango
 - `date:[05072007 TO 05232007]` (inclusive)
 - `author: {king TO mason}` (exclusive)
- Clasificación de acuerdo al peso de palabras ^
 - `title:"Bell" author:"Hemmingway" ^3.0`
 - Default boost value 1. May be <1 (e.g 0.2)
- Operadores booleanos: AND, "+", OR, NOT and "-"
 - "Linux OS" AND system
 - Linux OR system, Linux system
 - +Linux system
 - +Linux -system
- Agrupación
 - Title: (+linux +"operating system")

Admin

The screenshot displays the Apache Solr Admin web interface in a browser. The address bar shows the URL `bdvs087.svl.ibm.com:8983/solr/#/~cloud`. The left sidebar contains a menu with the following items: Dashboard, Logging, Cloud (selected), Tree, Graph, Graph (Radial), Dump, Core Admin, Java Properties, and Thread Dump. Below the menu is a 'Core Selector' dropdown. The main content area shows a diagram of a Solr cloud configuration. A central node labeled 'collection1' is connected to two shard nodes: 'shard1' and 'shard2'. Both shard nodes are marked with a solid black dot, indicating they are Leaders. The shard1 node is also labeled with the IP address 'bdvs087.svl.ibm.com', and the shard2 node is labeled with 'bdvs088.svl.ibm.com'. A legend in the bottom right corner explains the status icons: a solid black dot for 'Leader', a green circle for 'Active', a yellow circle for 'Recovering', an orange circle for 'Down', a red circle for 'Recovery Failed', and a grey circle for 'Gone'. At the bottom of the interface, there are links to Documentation, Issue Tracker, IRC Channel, Community forum, and Solr Query Syntax.

Demo

- Descargar e instalar el producto
- Utilizar el fichero de películas de Kaggle
- Crear una colección con el fichero
- Indexar su contenido
- Realizar búsquedas dentro de él

Usar el producto

➤ Básicamente dos opciones:

➤ Docker: https://hub.docker.com/_/solr

➤ docker pull solr

➤ docker run -p 8983:8983 -t solr

Existe la opción de utilizar docker-compose para arrancar varios servidores solr en un cluster y configurarlos conjuntamente.

Es una instalación vacía, sin colecciones, y es una instalación de un solo servidor.

➤ O realizar una instalación a partir del código binario

➤ <https://lucene.apache.org/solr/mirrors-solr-latest-redirect.html>

Más flexible en términos de poder contar con documentos para las colecciones, arrancar un cluster, etc.

Instalación del producto

- Descarga del binario:
 - <https://lucene.apache.org/solr/mirrors-solr-latest-redir.html>
- Diferentes versiones según plataforma: Windows, Linux/Mac
- Empaquetado como tar.gz o zip
- Desempaquetar y ejecutar:
 - `solr start -e cloud`
 - Configurad un cluster con dos servidores, dos shardings y dos replicas de los índices
 - Para la colección utilizaremos el fichero movies.csv que hemos comentado en otros casos
 - Para pararlo: `solr stop -all`
 - Para reutilizarlo podemos volver a lanzar `solr start -e cloud` e indicamos que queremos reutilizar la colección o revisa al final de esta presentación los comandos `solr start` (pag. 36)

El resultado es parecido a...

```
[umaster@ibmuamdocker bin]$ ./solr start -e cloud
*** [WARN] *** Your open file limit is currently 1024.
It should be set to 65000 to avoid operational disruption.
If you no longer wish to see this warning, set SOLR_ULIMIT_CHECKS to false in your profile or solr.in.sh
*** [WARN] *** Your Max Processes Limit is currently 4096.
It should be set to 65000 to avoid operational disruption.
If you no longer wish to see this warning, set SOLR_ULIMIT_CHECKS to false in your profile or solr.in.sh
```

Welcome to the SolrCloud example!

This interactive session will help you launch a SolrCloud cluster on your local workstation.
To begin, how many Solr nodes would you like to run in your local cluster? (specify 1-4 nodes) [2]:

Ok, let's start up 2 Solr nodes for your example SolrCloud cluster.

Please enter the port for node1 [8983]:

Please enter the port for node2 [7574]:

```
Creating Solr home directory /home/umaster/Master_Platform/solr-8.11.1/example/cloud/node1/solr
Cloning /home/umaster/Master_Platform/solr-8.11.1/example/cloud/node1 into
/home/umaster/Master_Platform/solr-8.11.1/example/cloud/node2
```

Starting up Solr on port 8983 using command:

```
"/home/umaster/Master_Platform/solr-8.11.1/bin/solr" start -cloud -p 8983 -s "/home/umaster/Master_Platform/solr-8.11.1/example/cloud/node1/solr"
```

started solr server on port 8983 (pid=11541). Happy searching!

Starting up Solr on port 7574 using command:

```
"/home/umaster/Master_Platform/solr-8.11.1/bin/solr" start -cloud -p 7574 -s "/home/umaster/Master_Platform/solr-8.11.1/example/cloud/node2/solr" -z localhost:9983
```

```
*** [WARN] *** Your open file limit is currently 4096.
It should be set to 65000 to avoid operational disruption.
If you no longer wish to see this warning, set SOLR_ULIMIT_CHECKS to false in your profile or solr.in.sh
*** [WARN] *** Your Max Processes Limit is currently 4096.
It should be set to 65000 to avoid operational disruption.
If you no longer wish to see this warning, set SOLR_ULIMIT_CHECKS to false in your profile or solr.in.sh
```

Waiting up to 180 seconds to see Solr running on port 7574 [-]

Started Solr server on port 7574 (pid=11549). Happy searching!

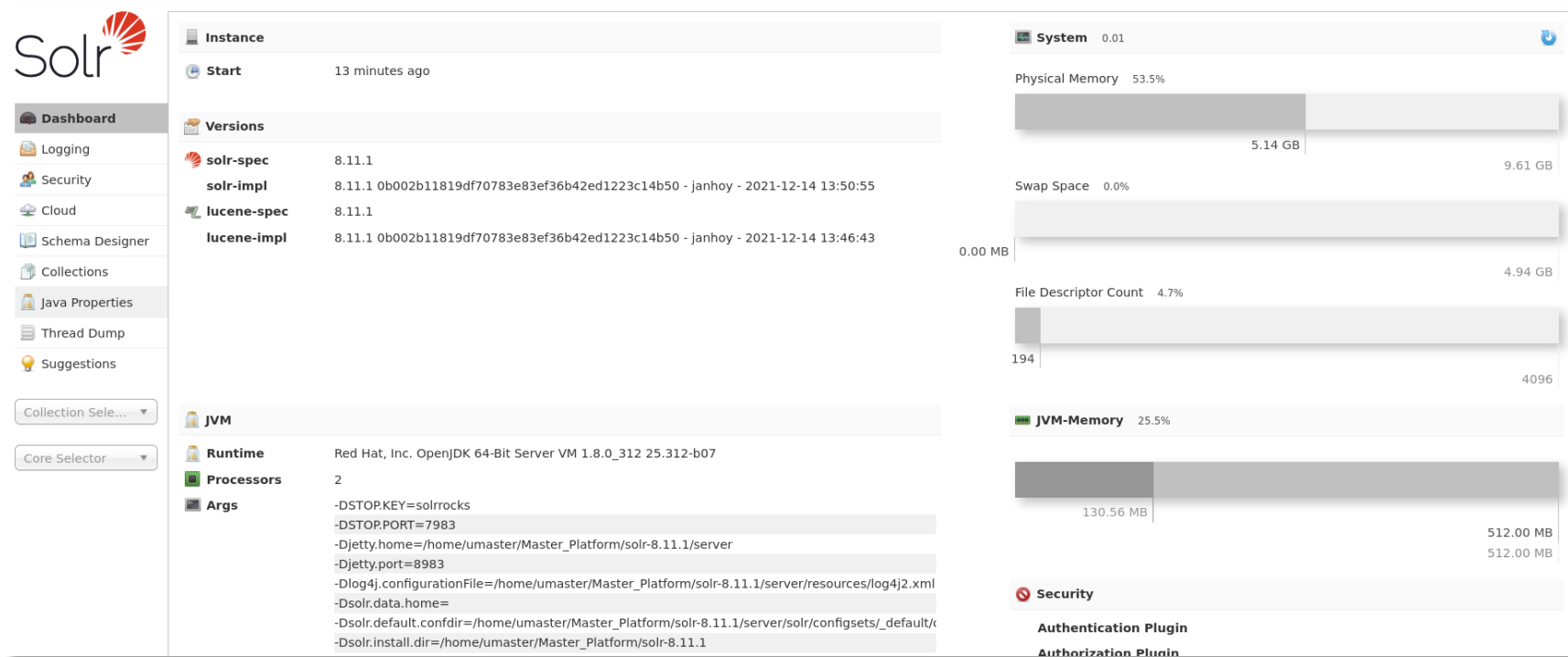
Enabling auto soft-commits with maxTime 3 secs using the Config API

```
POSTing request to Config API: http://localhost:8983/solr/gettingstarted/config
{"set-property":{"updateHandler.autoSoftCommit.maxTime":"3000"}}
Successfully set-property updateHandler.autoSoftCommit.maxTime to 3000
```


SolrCloud example running, please visit: <http://localhost:8983/solr>



Solr (I)



Solr (II)



[Dashboard](#)
[Logging](#)
[Security](#)
[Cloud](#)
[Nodes](#)
[Tree](#)
[ZK Status](#)
[Graph](#)
[Schema Designer](#)
[Collections](#)
[Java Properties](#)
[Thread Dump](#)
[Suggestions](#)


Collection Sele... ▾

Core Selector ▾

[Refresh](#) [Show all details](#)

Hosts 1 - 1 of 1. Filter by: Show hosts per page.

Host	Node	CPU	Heap	Disk usage	Requests	Collections	Replicas
10.0.2.15 Linux 9.6Gb Java 1.8 Load: 0 show details...	7574_solr Uptime: 13m show details...	0%	57%	138.0b	RPM: 0.17 p95: 28ms	gettingstarted	gettingstarted_s1r1 (0 docs) gettingstarted_s2r5 (0 docs)
	8983_solr Uptime: 14m show details...	0%	26%	138.0b	RPM: 0.24 p95: 35ms	gettingstarted	gettingstarted_s1r2 (0 docs) gettingstarted_s2r7 (0 docs)



[Dashboard](#)
[Logging](#)
[Security](#)
[Cloud](#)
[Nodes](#)
[Tree](#)
[ZK Status](#)
[Graph](#)
[Schema Designer](#)
[Collections](#)
[Java Properties](#)
[Thread Dump](#)
[Suggestions](#)


[Refresh](#) [Toggle details](#)

Status: green
ZK connection string: localhost:9983
Ensemble size: 1
Ensemble mode: standalone
Dynamic reconfig enabled: false

	localhost:9983
ok	true
clientPort	9983
secureClientPort	-1
zk_server_state	standalone
zk_version	3.6.2
zk_approximate_data_size	449207
zk_znode_count	149
zk_num_alive_connections	3

[Documentation](#) [Issue Tracker](#) [IRC Channel](#) [Community forum](#) [Solr Query Syntax](#)


SolR (III)







- Dashboard
- Logging
- Cloud
- Collections**
- Java Properties
- Thread Dump
- Suggestions


Collection Sele....

Core Selector



UAM_Movies

 **Collection: UAM_Movies**

Shard count:2


configName:UAM_Movies

replicationFactor:2

maxShardsPerNode:-1



router:compositeld



autoAddReplicas:false


 **Shard: shard1**


state:active

range:80000000-ffffff

 Replica: core_node3



 Replica: core_node5






 **Shard: shard2**

state:active

range:0-7ffffff

 Replica: core_node7

 Replica: core_node8



[Documentation](#) [Issue Tracker](#) [IRC Channel](#) [Community forum](#) [Solr Query Syntax](#)

UAM
UNIVERSIDAD AUTÓNOMA
DE MADRID

Máster en Big Data y Data Science

Ciclo de Vida Analítico del Dato

27

solrconfig.xml

- Es el fichero de configuración del propio Apache Solr. Alojado en el directorio conf
- Define, entre otras cosas:
 - Opciones de indexación
 - RequestHandlers
 - Highlighting
 - Correctores de texto
 - Servicios de infraestructura, como JMX, etc.

```
[umaster@ibmuamdocker solr-8.11.1]$ find . -name "*solrconfig*" -print
./example/example-DIH/solr/atom/conf/solrconfig.xml
./example/example-DIH/solr/db/conf/solrconfig.xml
./example/example-DIH/solr/mail/conf/solrconfig.xml
./example/example-DIH/solr/solr/conf/solrconfig.xml
./example/example-DIH/solr/tika/conf/solrconfig.xml
./example/files/conf/solrconfig.xml
./server/solr/configsets/_default/conf/solrconfig.xml
./server/solr/configsets/sample_techproducts_configs/conf/solrconfig.xml
[umaster@ibmuamdocker solr-8.11.1]$
```

Managed-schema

- Define los campos a ser indexados y el tipo del campo (texto, entero, etc.)
- Por defecto puede ser manejado en tiempo real a través de las llamadas al API correspondientes pero la colección también puede ser configurada con un esquema estático que se carga en el momento de arrancar la colección

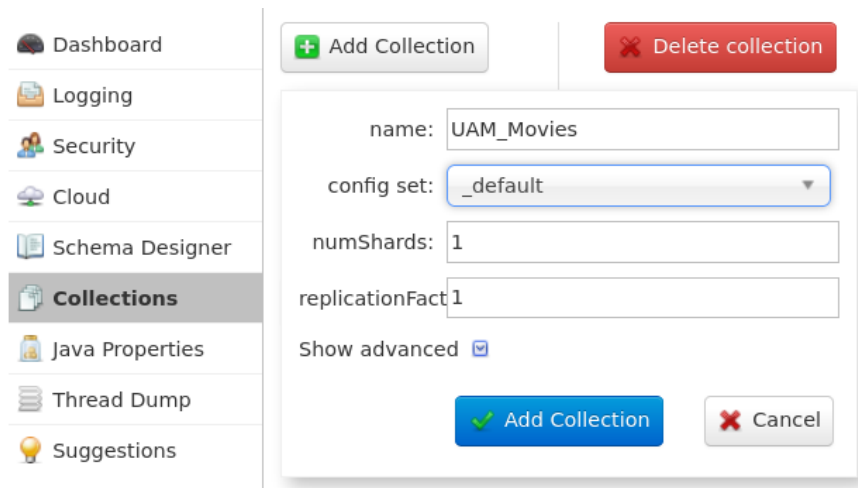
```
[umaster@ibmuamdocker solr-8.11.1]$ find . -name "*managed-schema*" -print
./example/example-DIH/solr/atom/conf/managed-schema
./example/example-DIH/solr/db/conf/managed-schema
./example/example-DIH/solr/mail/conf/managed-schema
./example/example-DIH/solr/solr/conf/managed-schema
./example/example-DIH/solr/tika/conf/managed-schema
./example/files/conf/managed-schema
./server/solr/configsets/_default/conf/managed-schema
./server/solr/configsets/sample_techproducts_configs/conf/managed-schema
[umaster@ibmuamdocker solr-8.11.1]$
```

Siguiendo con la demo...

- Indicamos que utilizábamos la configuración de `_default`
- SolR sigue una estrategia `schemaless`. Podremos llegar a cargar los datos directamente sin tener que especificar ningún formato.
- No es aconsejable hacerlo en producción o en un entorno no sencillo (mismo campo en distintos documentos), indexación compleja, etc.

Cargamos los datos de movies.csv directamente

- Utilizando la herramienta de administración crea una colección denominada UAM_Movies



The screenshot displays the MongoDB Admin interface. On the left is a sidebar menu with the following items: Dashboard, Logging, Security, Cloud, Schema Designer, Collections (highlighted), Java Properties, Thread Dump, and Suggestions. The main area shows a modal dialog for adding a new collection. At the top of the dialog are two buttons: '+ Add Collection' and 'Delete collection'. The dialog contains the following fields: 'name:' with the value 'UAM_Movies', 'config set:' with a dropdown menu showing '_default', 'numShards:' with the value '1', and 'replicationFact' with the value '1'. There is a 'Show advanced' checkbox which is checked. At the bottom of the dialog are two buttons: 'Add Collection' (with a green checkmark icon) and 'Cancel' (with a red X icon).

Cargamos los datos de movies.csv directamente

- Una vez creada cargamos los datos

```
./post -c UAM_Movies ../datos/movies.csv
```

```
[umaster@ibmuamdocker bin]$ ./post -c UAM_Movies ../datos/movies.csv
java -classpath /home/umaster/Master_Platform/solr-8.11.1/dist/solr-core-8.11.1.jar -Dauto=yes -Dc=UAM_Movies -Ddata=files org.apache.solr.util.SimplePostTool ../datos/movies.csv
SimplePostTool version 5.0.0
Posting files to [base] url http://localhost:8983/solr/UAM_Movies/update...
Entering auto mode. File endings considered are xml,json,jsonl,csv,pdf,doc,docx,ppt,pptx,xls,xlsx,odt,odp,ods,ott,otp,ots,rtf,htm,html,txt,log
POSTing file movies.csv (text/csv) to [base]
1 files indexed.
COMMITting Solr index changes to http://localhost:8983/solr/UAM_Movies/update...
Time spent: 0:00:00.386
[umaster@ibmuamdocker bin]$
```

Se pueden enviar parámetros en la carga como el tipo de separador, si hay campos con múltiples valores, etc.

Esta carga es, por tanto, totalmente dinámica, no estamos aplicando ningún formato por defecto a la carga de los datos

Consultémosla con la consola de administración

The screenshot displays the Solr Admin Console interface. On the left, a sidebar contains navigation links: Dashboard, Logging, Cloud, Collections, Java Properties, Thread Dump, Suggestions, and a dropdown menu for 'UAM_Movies' with sub-links for Overview, Analysis, Dataimport, Documents, Files, and Query (which is selected). Below these is a 'Core Selector' dropdown. The main panel is titled 'Request-Handler (qt)' and shows the following configuration:
 - Request-Handler: /select
 - common:
 - q:
 - fq:
 - sort:
 - start, rows: 0 to 10
 - fl:
 - df:
 - Raw Query Parameters: key1=val1&key2=val2
 - wt: json
 - indent off
 - debugQuery
 - dismax
 - edismax
 - hl
 - facet
 - spatial
 - spellcheck
 - Execute Query button
 The right panel shows the JSON response from the query:
 http://localhost:8983/solr/UAM_Movies/select?q=%3A*&wt=json
 {
 "responseHeader": {
 "zkConnected": true,
 "status": 0,
 "QTime": 24,
 "params": {
 "q": "*",
 "wt": "json",
 "_: 1572464641564"
 }
 },
 "response": { "numFound": 45843, "start": 0, "maxScore": 1.0, "docs": [
 {
 "movieId": [4],
 "title": ["Waiting to Exhale (1995)"],
 "genres": ["Comedy|Drama|Romance"],
 "id": "d4cd4878-7028-4a9a-bfc9-b731af060a16",
 "_version_": 1648848572428320768
 },
 {
 "movieId": [6],
 "title": ["Heat (1995)"],
 "genres": ["Action|Crime|Thriller"],
 "id": "d6b94ddf-d656-4785-b69b-ccf144c36416",
 "_version_": 1648848572455583744
 },
 {
 "movieId": [7],
 "title": ["Sabrina (1995)"],
 "genres": ["Comedy|Romance"],
 "id": "21e3802e-9221-46f1-9a22-dabbfaef4f19",
 "_version_": 1648848572456632320
 },
 {
 "movieId": [10],
 "title": ["GoldenEye (1995)"],
 "genres": ["Action|Adventure|Thriller"],
 "id": "8182afe0-7ea2-4112-867d-98472c2374b0",
 "_version_": 164884857245680896
 },
 {
 "movieId": [12],
 "title": ["Dracula: Dead and Loving It (1995)"],
 "genres": ["Comedy|Horror"],
 "id": "e309f660-3d0e-433a-af8e-93d1e7b773e0",
 "_version_": 1648848572458729472
 }
]
 }

Consultémosla con la consola de administración

The image displays the Solr Admin Console interface with two different query configurations for the 'UAM_Movies' collection.

Top Configuration:

- Request-Handler (qt):** /select
- q:** title:"Matrix"
- fq:** (empty)
- sort:** (empty)
- start, rows:** 0 to 10
- fl:** (empty)
- df:** (empty)
- Raw Query Parameters:** key1=val1&key2=val2
- wt:** csv
- Execute Query** button

URL: `http://localhost:8983/solr/UAM_Movies/select?q=title%3A%22Matrix%22&wt=csv`

Results:

```
movieId,title,genres,id_version
2571,"Matrix\, The (1999)",Action|Sci-Fi|Thriller,c1b51648-c398-4767-a508-31cf2b49912f,1648848573823975424
157721,Armitage: Dual Matrix (2002),Action|Adventure|Animation|Sci-Fi|Thriller,e02fbaf9-2c55-48ba-b277-be9e6f71d3b1,1648848580264329217
6365,"Matrix Reloaded\, The (2003)",Action|Adventure|Sci-Fi|Thriller|IMAX,43933250-c0af-4541-9171-dc8ee2666c4b,1648848574750916609
6934,"Matrix Revolutions\, The (2003)",Action|Adventure|Sci-Fi|Thriller|IMAX,af6e6fe2-a50f-425a-9200-8429cc09fcb6,1648848574857871360
172255,The Matrix Revisited (2001),Documentary,932a6900-7f84-4e62-8c11-2df69d425e53,1648848580901307404
132490,Return to Source: The Philosophy of The Matrix (2004),Documentary,c0295dee-efca-4064-9e4d-4b40ddf84fdf,1648848578708242433
```

Bottom Configuration:

- Request-Handler (qt):** /select
- q:** title:"Matrix" and NOT genres:"Documentary"
- fq:** (empty)
- sort:** (empty)
- start, rows:** 0 to 10
- fl:** (empty)
- df:** (empty)
- Raw Query Parameters:** key1=val1&key2=val2
- wt:** csv
- Execute Query** button

URL: `http://localhost:8983/solr/UAM_Movies/select?q=title%3A%22Matrix%22%20and%20NOT%20genres%3A%22Documentary%22&wt=csv`

Results:

```
movieId,title,genres,id_version
2571,"Matrix\, The (1999)",Action|Sci-Fi|Thriller,c1b51648-c398-4767-a508-31cf2b49912f,1648848573823975424
157721,Armitage: Dual Matrix (2002),Action|Adventure|Animation|Sci-Fi|Thriller,e02fbaf9-2c55-48ba-b277-be9e6f71d3b1,1648848580264329217
6365,"Matrix Reloaded\, The (2003)",Action|Adventure|Sci-Fi|Thriller|IMAX,43933250-c0af-4541-9171-dc8ee2666c4b,1648848574750916609
6934,"Matrix Revolutions\, The (2003)",Action|Adventure|Sci-Fi|Thriller|IMAX,af6e6fe2-a50f-425a-9200-8429cc09fcb6,1648848574857871360
```

Alguna búsquedas adicionales...

- Crea una nueva colección: *Documentos*

```
./solr create -c Test_Documentos -d sample_techproducts_configs
```

- Carga los documentos que están disponibles en el directorio *exampledocs*. Por ejemplo con un comando del tipo:

```
./post -c Test_Documentos /home/umaster/Master_Platform/solr-8.11.1/example/exampledocs
```

- Realiza ahora una búsqueda en esa nueva colección
 - Busca la palabra panic
 - Observa el resultado... indica qué documento contiene esa palabra
 - Carga otros ficheros PDF y realiza búsquedas en él

Cómo parar y arrancar Solr

- Para parar:

```
./bin/solr stop -all
```

- Para arrancar (si utilizaste la configuración por defecto propuesta por la configuración cloud: dos nodos, etc.)

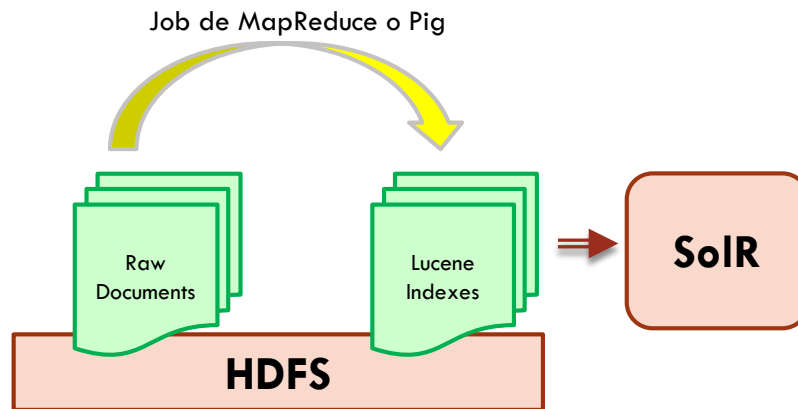
```
./bin/solr start -c -p 8983 -s example/cloud/node1/solr
```

```
./bin/solr start -c -p 7574 -s example/cloud/node2/solr -z localhost:9983
```

Opción de involucrar Hadoop. Indexar en HDFS

➤ Ingesta a través de Jobs MapReduce

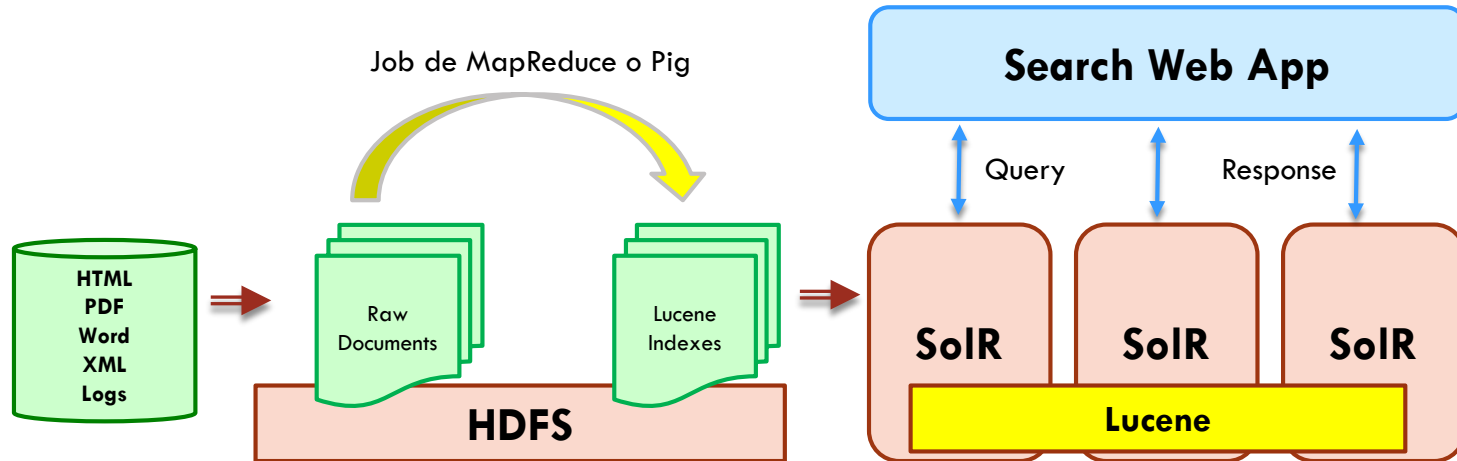
- CSV
- Microsoft Office
- Grok
- Zip
- Solr XML
- Ficheros secuenciales
- WARC



➤ Procesamiento: con Pig

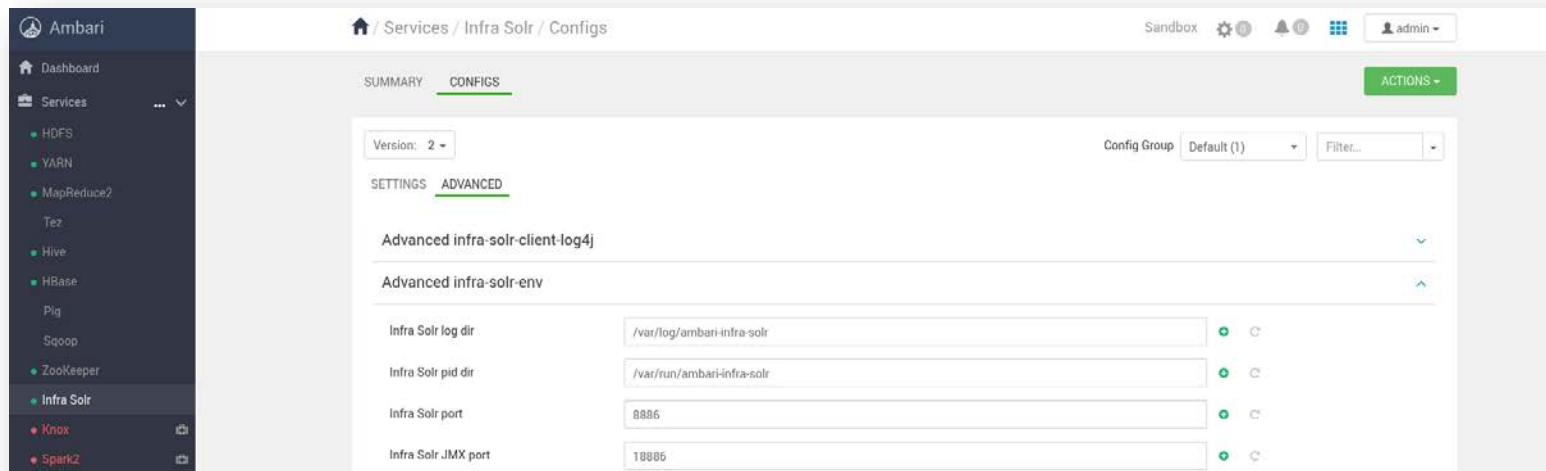
- Se escriben scripts en pig para indexar el contenido:
 - Pig lo utilizamos para precisar y unir ficheros
 - La salida se lanza ya a Solr

Se puede ampliar con un cluster adicional de SolR

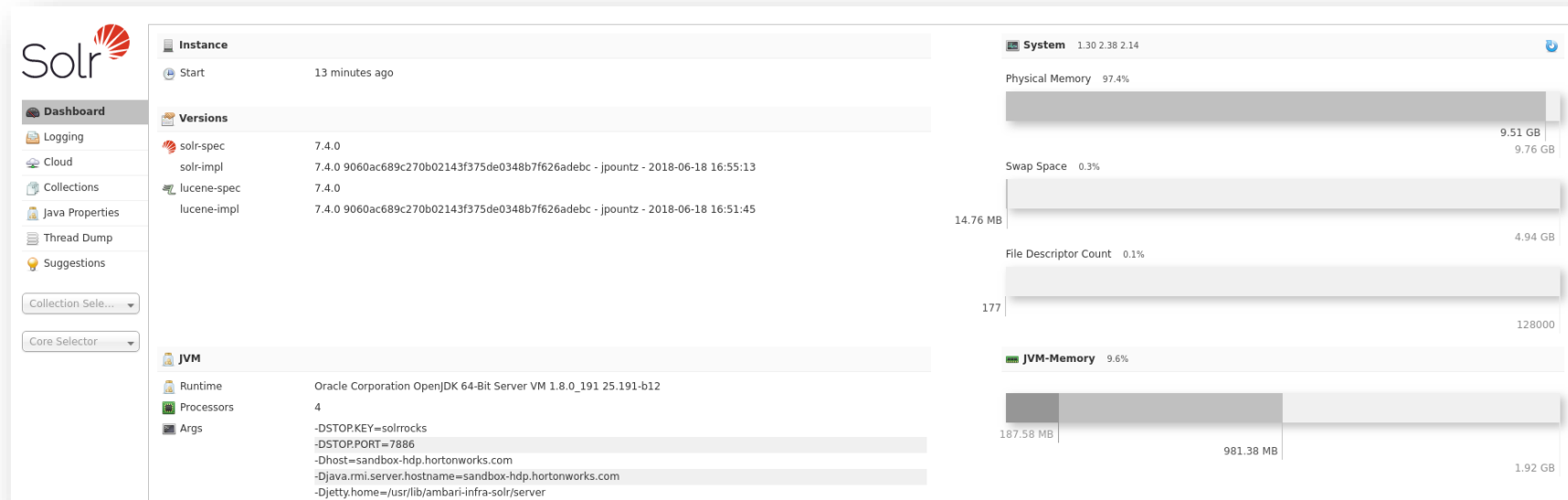


Se incluía en Hortonworks Data Platform

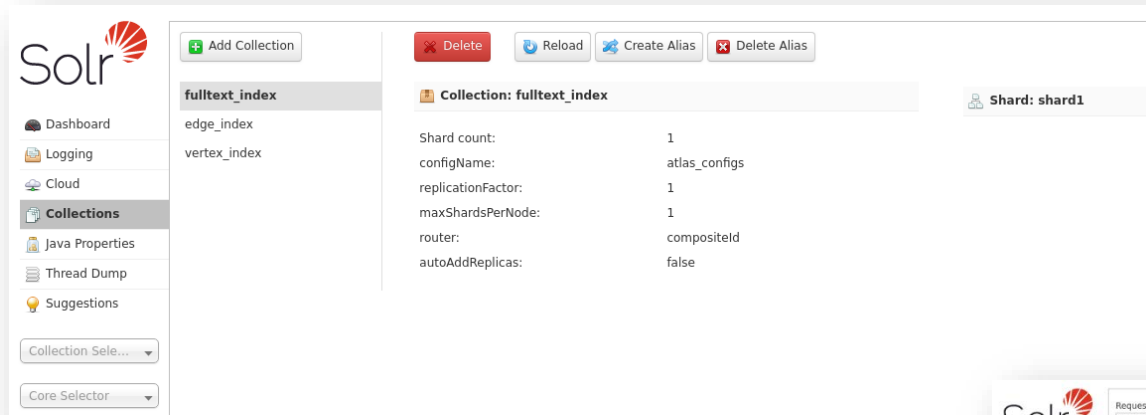
- Anteriormente incluido en HDP 2.X
- Incluido en HDP 3.0 formando parte de la infraestructura como habilitador para Ambari (consola de gestión de Hadoop en HDP)



SolR en HDP 3.0 (I)

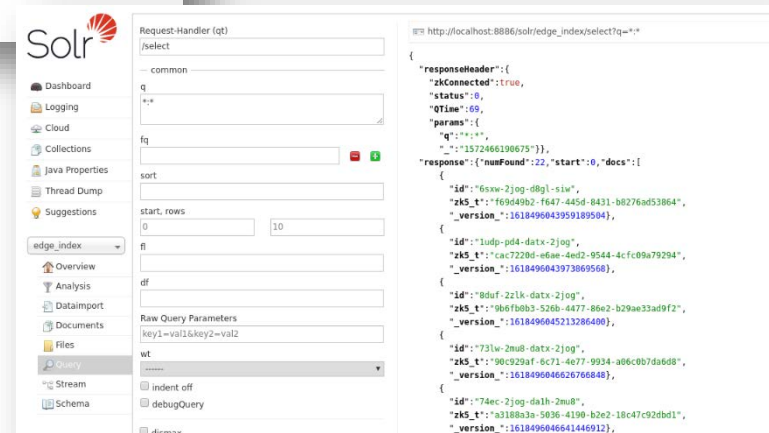


Solr en HDP 3.0 (II)



The Solr Admin UI shows the 'Collections' page for the 'fulltext_index'. The left sidebar contains navigation links: Dashboard, Logging, Cloud, Collections (selected), Java Properties, Thread Dump, and Suggestions. Below the sidebar is a 'Collection Sele...' dropdown and a 'Core Selector' dropdown. The main content area displays the 'fulltext_index' collection details. At the top, there are buttons: 'Add Collection', 'Delete', 'Reload', 'Create Alias', and 'Delete Alias'. Below these, the collection name 'fulltext_index' is shown. To the right, the shard information is displayed: 'Shard: shard1'. The collection configuration is listed below:

Property	Value
Shard count:	1
configName:	atlas_configs
replicationFactor:	1
maxShardsPerNode:	1
router:	compositeld
autoAddReplicas:	false



The Solr Admin UI shows the 'Query' page for the 'edge_index' collection. The left sidebar contains navigation links: Dashboard, Logging, Cloud, Collections, Java Properties, Thread Dump, Suggestions, Overview, Analysis, Dataimport, Documents, Files, Query (selected), Stream, and Schema. Below the sidebar is a 'Collection Sele...' dropdown and a 'Core Selector' dropdown. The main content area displays the query interface. At the top, there are buttons: 'Request-Handler (qt)', 'common', 'q', 'fq', 'sort', 'start, rows', 'ff', 'df', 'Raw Query Parameters', 'key1=...&key2=...', 'wt', 'indent off', 'debugQuery', and 'diagnostics'. The query parameters are set to: 'q=*:*', 'fq=*:*', 'sort=*:*', 'start, rows=0, 10', 'ff=*:*', 'df=*:*', 'Raw Query Parameters: key1=...&key2=...', 'wt=json', 'indent off', 'debugQuery', and 'diagnostics'. The response is displayed on the right, showing a JSON object with 'numFound' and 'docs'.

```
http://localhost:8886/solr/edge_index/select?q=*:*

{
  "responseHeader": {
    "zkConnected": true,
    "status": 0,
    "QTime": 69,
    "params": {
      "q": "*:*",
      "fq": "*:*",
      "sort": "*:*",
      "start": 0,
      "rows": 10
    }
  },
  "response": {
    "numFound": 22,
    "start": 0,
    "docs": [
      {
        "id": "6xxx-2jog-d8gl-siw",
        "zk5_t": "f69449b2-f647-445d-8431-b8276ad53864",
        "_version_": 1618496043959189594,
        {
          "id": "ludp-pd4-datx-2jog",
          "zk5_t": "cac7220d-e6ae-4ed2-9544-4cfc09a79294",
          "_version_": 1618496043973869568,
          {
            "id": "8duf-2zlk-datx-2jog",
            "zk5_t": "06efb0b3-526b-4477-8ae2-b29ae33ad9f2",
            "_version_": 161849604523286400,
            {
              "id": "73lw-2mu8-datx-2jog",
              "zk5_t": "90c929af-6c71-4e77-9934-a06c0b7da6d8",
              "_version_": 161849604662676840,
              {
                "id": "74ec-2jog-da1h-2mu8",
                "zk5_t": "a3188a3a-5036-4190-b2e2-18c47c92dbd1",
                "_version_": 1618496046641446912,

```

Mas información

- [Reference Guide for SolR](#)
- [Apache SolR website](#)
- [Resources](#)