



Integrazione fra eXtraWay ed Elasticsearch



Versione Elasticsearch supportata: **5.2.2** (o superiore se mantenuta la major version) con **JAVA 8**

Caratteristiche integrazione

- L'integrazione fra eXtraWay ed Elasticsearch è gestita a livello applicativo (strato broker)
- Per ogni archivio eXtraWay (che deve essere indicizzato su Elasticsearch) deve essere creato uno specifico indice su Elasticsearch (relazione 1-1 fra archivi eXtraWay ed indici di Elasticsearch, ad ogni archivio su eXtraWay corrisponde un indice su Elasticsearch)
- Su eXtraWay vengono mantenuti tutti gli indici primari e i seriali, mentre ogni altro dato da indicizzare viene lasciato ad Elasticsearch
- Tutte le scritture su eXtraWay vengono replicate sull'indice Elasticsearch (tramite specifico componente sul broker)
 - In caso di errore in scrittura sull'indice di Elasticsearch viene registrato un file di errore sul file system
 - Ogni minuto un processo verifica la presenza di eventuali file di errore e ritenta l'operazione indicata (in modo da mantenere l'allineamento fra i dati memorizzati su archivio eXtraWay e quelli presenti sull'indice di Elasticsearch)
- Le ricerche vengono realizzate:
 - Su eXtraWay se i filtri di ricerca riguardano solo campi seriali indicizzati su eXtraWay
 - Su Elasticsearch in ogni altro caso
- Ogni ricerca in formato eXtraWay ricevuta dal broker viene opportunamente parsata e, se deve essere rediretta su Elasticsearch, convertita in formato JSON
- **N.B.:** Non esiste più il concetto di selezione conosciuto in eXtraWay. La paginazione dei risultati di una ricerca su Elasticsearch comporta una nuova esecuzione della stessa (possono quindi variare anche il numero totale di risultati) → [Questo porta a differenze di comportamento a livello di interfaccia dell'applicativo](#)
- L'integrazione è stata inclusa sulle seguenti applicazioni:
 - DocWay4 (DocWay4-service)
 - 3diWS
 - MailArchiver (MSA)
- La gestione applicativa di tutte le selezioni derivanti da query ha comportato un maggiore utilizzo di memoria da parte delle webapp (o altre applicazioni) Java. Per questo motivo è stata integrata una libreria specifica per gestire la cache applicativa (mantenimento in memoria fino ad una determinata soglia e successiva serializzazione su file system): **Apache JCS**

Creazione dell'indice su Elasticsearch

Per creare un nuovo indice su Elasticsearch e popolarlo con i dati contenuti in un archivio eXtraWay occorre procedere nel modo seguente:

1. Creare l'indice su Elasticsearch (indicando il json di mapping per l'archivio). Si consiglia di chiamare l'indice con un nome del tipo '[nome_archivio]_ddMMyyyy', in modo da poter gestire il reindex senza dover stoppare Elasticsearch.
2. Assegnare all'indice creato '[nome_archivio]_ddMMyyyy' l'alias '[nome_archivio]'. In questo modo sarà possibile accedere all'indice tramite '[nome_archivio]'
3. Lanciare la procedura di importazione dati da eXtraWay

Mapping su Elasticsearch

L'operazione di mapping consiste nel definire tutte le risorse e i campi (con relativa modalità di indicizzazione) da gestire su Elasticsearch per uno specifico archivio.

Alla pagina corrente sono stati allegati due esempi di file di mapping (relativi agli archivi DocWay e ACL):

- File Mapping DocWay
- File Mapping ACL

Esempio di comando di creazione indice su Elasticsearch:

```
curl -XPUT "http://localhost:9200/xdocwaydoc_27062017?pretty=true" -d @xdocwaydoc-mapping.json
```

Assegnazione Alias

```
curl -XPOST 'localhost:9200/_reindex?pretty' -H 'Content-Type: application/json' -d '{
  "source": {
    "index": "xdocwaydoc_27062017"
  },
}
```



```
"dest": {  
  "index": "xdocwaydoc"  
}  
},  
,
```

Per maggiori info consultare la pagina: <https://www.elastic.co/guide/en/elasticsearch/reference/current/docs-reindex.html>

Importazione dati da eXtraWay

L'importazione dei dati su un indice Elasticsearch dato un archivio eXtraWay già popolato, può essere realizzata tramite una nostra specifica app.

Progetto GIT: <http://gitlab.bo.priv/docway5/it.tredi.xway2elastic-import>

Repository Nexus: **it.tredi.xway2elastic-import**



- **Il processo di importazione dati da eXtraWay ad Elasticsearch si aspetta che il nome dell'archivio eXtraWay e dell'indice su Elasticsearch siano identici**
- E' richiesto *Java8* per l'esecuzione del comando

Configurazione

I file di configurazione dell'applicazione sono presenti nella directory [XWAY2ELASTIC-HOMEDIR]/classes. Il principale è 'application.properties':

```
# Connessione ad elasticsearch  
elasticsearch.host=127.0.0.1  
elasticsearch.port=9200  
elasticsearch.threads.count=  
elasticsearch.client.sniff=false  
#elasticsearch.username=  
#elasticsearch.password=  
elasticsearch.connectionRequestTimeout=30000  
elasticsearch.maxRetryTimeout=30000  
elasticsearch.socketTimeout=30000  
  
# Connessione ad eXtraWay  
xway.host=127.0.0.1  
xway.port=4859  
# Il nome del db eXtraWay e l'eventuale query per filtrare i record devono essere specificati  
# come argomento del comando shell  
xway.dbname=  
  
# Classe di implementazione della conversione di documenti da XML a JSON per salvataggio su  
Elasticsearch. Questo parametro e' obbligatorio, puo' essere definito anche come  
# parametro di avvio dell'applicazione da shell (parametro '-doc2json.impl=')  
doc2json.impl=  
#doc2json.impl=it.tredi.xway2elastic.parser.DocWayDoc2Json  
  
# Elenco di query (e relativi ordinamenti) in base alle quali recuperare i documenti da eXtraWay  
per la sincronizzazione  
# su Elasticsearch  
#xway.query.1=([UD,/xw/@UdType/]="doc") AND ([/doc/@tipo/]="interno" OR "varie" OR "partenza")  
#xway.sort.1=  
#xway.query.2=([UD,/xw/@UdType/]="doc") AND ([/doc/@tipo/]="arrivo")  
#xway.sort.2=  
#xway.query.3=([UD,/xw/@UdType/]="fascicolo" OR "raccoglitore" OR "seduta")  
#xway.sort.3=  
#...  
#xway.query.n=  
#xway.sort.n=
```

Per poter avviare l'importazione dei dati è necessari specificare la query tramite la quale recuperare i documenti da eXtraWay. Da quanto si può notare dal file di properties precedente, è possibile specificare più istanze di query (eventualmente con relativo sort) in modo da avviare più thread di inserimento e quindi diminuire la durata dell'importazione.



Maggiore è la distribuzione dei risultati delle query, minore sarà il tempo totale di importazione.

Utilizzo

WINDOWS:

```
[XWAY2ELASTIC-HOMEDIR]/bin/xway2elastic.bat -dbname=xdocwaydoc ...
```

LINUX:

```
sh [XWAY2ELASTIC-HOMEDIR]/bin/xway2elastic.sh -dbname=xdocwaydoc ...
```

Parametri supportati:

- **-dbname=**, nome del db eXtraWay da sincronizzare su Elasticsearch (l'indice su Elasticsearch deve già essere stato creato)
- **-doc2json.impl=**, classe di implementazione della conversione di documenti da XML a JSON per salvataggio su Elasticsearch (*parametro obbligatorio*, es.: `it.tredi.xway2elastic.parser.DocWayDoc2Json`)
- **-titlePageSize=**, numero di salvataggi da includere in una singola richiesta su Elasticsearch (*BulkRequest*) (default = 1000)
- **-enableXwFiles=**, abilita o meno l'indicizzazione del contenuto degli allegati `xw:file` (*true / false*, default = *true*)
- **-xwFileIndexByBulkRequest=**, true per indicizzare gli `xw:file` tramite *BulkRequest*, false tramite richiesta singola (*true / false*, default = *true*)
- **-disableRefresh=**, disabilita il refresh automatico dell'indice durante l'inserimento bulk (*true / false*, default = *true*)



Per maggiori informazioni sui comandi di sincronizzazione dati di `xway2elastic-import` si rimanda al file **README.md** presente nella directory `[XWAY2ELASTIC-HOMEDIR]/bin`.

Configurazione dell'integrazione sul broker

Cache applicativa con Apache JCS

Indicizzazione di Files



TODO