Federation Merger

Federation Merger	
Factory Name	com.searchtechnologies.aspire:aspire-federation
subType	merge
Inputs	Aspire Jobs
Outputs	Aspire Jobs

The *Federation Merger* stage takes a job (most likely originating from the Dispatcher) and *merges* the result sets from a number of different federated queries to form a single result set that can be sent back to the client.

The *Federation Merger* uses a tag in the document to identify the results to be merged and assumes that each child of this tag is a single result set from a federated query. The format of this child result set is detailed below.

The *Federation Merger* is able to use different merge methods, with the actual method used being specified in the incoming document. Once the result sets have been merged, the resulting set is added to the document and the source result sets are removed (in order to reduce the payload returned back to the client).

At the same time as merging the results set, the *Federation Merger* selects the appropriate page of results based on the incoming job parameters.

Document Format

The Federation Merger is designed to merge XML results sets from FAST search engines. If your search application is not FAST search, the *query application* specified in the Dispatcher should include a stage to convert the results to the FAST format.

The FAST XML format is shown below. This information is taken from the FAST documentation - ESP Query Integration Guide.

XML Template	Description

`<SEGMENTS>`
`<SEGMENT NAME="webcluster">`
	`Normally only one segment (cluster) returned`

`<QUERYTRANSFORMS>`
`<QUERYTRANSFORM NAME=`
`ACTION=`
`QUERY=`
`CUSTOM=`
`MESSAGE=`
`MESSAGEID=`
`INSTANCE= />`
`...`
`</QUERYTRANSFORMS>`
	`Query transformation block`
	`One query transformation feedback`
	`NAME element`
	`ACTION element`
	`QUERY element`
	`CUSTOM element`
	`MESSAGE element`
	`MESSAGE ID element`
	`INTANCE element`
	
	`Refer to Query Transformations in the Query Language and Parameters Guide for description of the elements.`

<NAVIGATION ENTRIES= > <NAVIGATIONENTRY NAME= USEDHITS= DISPLAYNAME= TYPE= UNIT= MODIFIER= SCORE= SAMPLECOUNT= HITCOUNT= RATIO= MIN= MAX= MEAN= ENTROPY= SUM= > <NAVIGATIONELEMENTS COUNT= > <NAVIGATIONELEMENT NAME= MODIFIER= COUNT= /> ... (more modifiers) </NAVIGATIONELEMENTS> </NAVIGATIONENTRY> ... (more navigators) </NAVIGATION>	Navigators Number of navigators Navigator name Number of used (considered) hits for each navigator Display name Navigator type Unit Modifier Score Sample count Hit count Ratio Min value Max value Mean value Entropy Aggregated sum of all values Navigator name Modifier Document count

<CLUSTERS> <CLUSTER TYPE= > <NODE ID= SUBMEMCNT= > <LABELS COUNT= > <LABEL>... </LABEL> ... (more labels) < /LABELS> <MEMBERS COUNT= > <MEMBER OFFSET= > ... (more members) < /MEMBERS> < /NODE> ... (more nodes) < /CLUSTER> ... (more clusters) < /CLUSTERS>	Clusters and cluster nodes. A cluster node ID (e.g. "S.0.1") Number of sub-members Cluster label Cluster member

<RESULTSET FIRSTHIT= LASTHIT= HITS= TOTALHITS= MAXRANK= TIME= > <HIT NO= RANK= SITEID= MOREHITS= > <FIELD NAME= > field_content </FIELD> ... (more fields) </HIT> ... (more hits) </RESULTSET>	Start of query result set. Index to first hit in result set Index to last hit in result set Number of hits presented Total number of hits for query MAXRANK is a theoretical maximum rank for a document for a specific query (if the document contained all the query terms close to each other, early in the document, in all the important fields, etc.). In practice the best document in the result set will usually have a rank score much lower then MAXRANK. Time used to process query Index to this result entry Rank value for result entry Field Collapse entries: SITEID = Field ID MOREHITS 1 if collapsed entries exist below the entry Field name and content End of this result entry
<PAGENAVIGATION> <NEXTPAGE FIRSTHIT= LASTHIT= URL= /> <PREVPAGE FIRSTHIT= LASTHIT= URL= /> </PAGENAVIGATION>	Information about next page in result set: First hit on next page (f) Last hit on next page (l) URL to retrieve next page (u)
</SEGMENT> </SEGMENTS>	Normally only one segment (cluster) returned

Important elements and Attributes

Certain information from the FAST XML results set are read or updated during the merge and the operation of the *merger* is undefined if these are not present. These attributes are detailed below:

Element	Description

NAVIGATION/@ENTRIES	Updated to hold the correct number of navigators.
NAVIGATION/NAVIGATIONENTRY/@NAME	Navigators from different result sets with the same name will be merged.
NAVIGATION/NAVIGATIONENTRY/NAVIGATIONELEMENTS/@COUNT	Updated to hold the correct number of elements for this navigator.
NAVIGATION/NAVIGATIONENTRY/NAVIGATIONELEMENTS /NAVIGATIONELEMENT/@COUNT	Updated to hold the correct number of hits for this navigator element.
RESULTSET/@FIRSTHIT	Updated to hold the hit number of the first hit in this result page.
RESULTSET/@LASTHIT	Updated to hold the hit number of the last hit in this result page.
RESULTSET/@HITS	Updated to hold the number of hits in this result page.
RESULTSET/@TOTALHITS	Updated to hold the total number of hits in this result set.
RESULTSET/@MAXRANK	Updated to hold the maximum rank this result set.
RESULTSET/HIT/@NO	Updated to hold the correct hit number for this hit.

Merging

The *Federation Merger* merges results set from the incoming Aspire document. The document includes a node containing a number of results sets (typically one from each server the query was federated too). The results sets should be in the FAST format described above. The merge process splits the results sets in to their constituent parts (QUERYTRANSFORMS, NAVIGATION, CLUSTERS and RESULTSET) and merges each in turn. A single result set is then re-created from the merged pieces.

Query Transforms

Merging of the query transforms simple concatenates the query transforms from each result set

Navigation

Navigation merge examines the navigators returned from each server in turn. For the first server, all navigators are simple added to the merged set. For subsequent servers the following approach is used:

- Get the navigator name from the *NAVIGATIONENTRY/@NAME* attribute
- Check if the merged list already contains this navigator (name)
- Add the navigator to the merged list if it doesn't exist
- If it does, merge the navigator elements in to the merged navigator list.

Merging is similar for the navigator elements

- Get the *NAVIGATIONELEMENT/@NAME* attribute
- Check if this element already exists in the navigator
- If it doesn't, add it
- If it does, update the @COUNT attribute to the appropriate value

The counts for the navigators as elements are also updated as part of the merge

Clusters

Merging of the clusters simple concatenates the clusters from each result set

Result Set Merging

Result set merging takes the results sets extracted from the incoming document and merges them using the schema suggested by the Dispatcher zone (or falling back to the default). The appropriate page of results (as requested by the query) is then selected.

The following types of merge are supported

Round robin

In the *round robin* merge method, a single hit is taken from each result set (from a specific server) in turn and added to a merged hit list. Once the hit list for a specific server is exhausted, then it is no longer considered and the lists for the remaining servers are used in turn until all results set from all servers are exhuasted. As hits are added to the list, the hit number is adjusted to the correct value. The total hists and max rank for the results set are also updated. The appropriate page of results is then selected.

Rank

In the *rank* merge method, the results are assume to be in descending rank order. The highest ranking hit from all result sets is removed and added to a merged hit list. This continues until all results set from all servers are exhuasted. As hits are added to the list, the hit number is adjusted to the correct value. The total hists and max rank for the results set are also updated. The appropriate page of results is then selected.

NOTE: the actual implementations of merge algorithms are optimised for performance and only collect the required page of results.

Configuration

The following configuration items are supported:

Element	Type	Default	Description
federationResultTag	String	aspireFederationResult	The tag in the document holding all of the results sets from the federated queries.
resultTag	String	SEGMENT	The tag of elements holding the individual result sets.
mergeType	String	robin	The default merge type to use if the merge type is not given in the document.

Example Configuration

```
<component subType="merge" name="Merger" factoryName="aspire-federation">
  <resultTag>SEGMENTS</resultTag>
  <federationResultTag>aspireFederationResult</federationResultTag>
  <mergeType>robin</mergeType>
  <debug>false</debug>
</component>
```