Kepler - Bug #3578

optimize timing of data download by EML and other data source actors

10/29/2008 11:58 AM - Matt Jones

Status:	New	**Start date:**	10/29/2008
Priority:	Normal	**Due date:**	
Assignee:	Jing Tao	**% Done:**	0%
Category:	data access	**Estimated time:**	0.00 hour
Target version:	3.X.Y	**Spent time:**	0.00 hour
Bugzilla-Id:	3578		

Description

The EML actor and some other data source actors download data when the drag and drop on the workflow canvas occurs, or when a workflow is opened in the case of a previously saved workflow. For the EML actor, the Object Manager is first checked to see if the data file is already locally cached, and if not it is retrieved from the server. This is not always optimal, as sometimes it could make sense to delay the download to later in the workflow execution cycle.

Case 1: Optimal download at workflow loading time
When data objects are large but must be fully retrieved, it is best to retrieve the object as early as possible to avoid delaying the workflow execution.

Case 2: Optimal loading upon execution
When data objects are large but might not be fully downloaded (e.g., via a filtering SQL query on the remote host), it is better to postpone download until after the user has fully configured the actor, which should be complete by the time of workflow execution. Unfortunately, the EML actor does not yet support remote data subsetting, so there is no mechanism yet to support this case. When the Data Manager library is reincorporated in Kepler, this should then be possible and desirable.

Case 3: Optimal loading upon actor firing
When data objects are large but an actor is part of a distributed workflow, it is better to postpone loading data until the actor fires as the actor may actually execute on a different slave node rather than the master. Thus, prematurely downloading the data may cause the master to download data when in fact one or more slave nodes are actually the ones that need it locally.

There are probably other cases as well. The hard part is how Kepler can differentiate these cases with minimal user input in order to decide which case applies and therefore optimize the timing of the download via appropriate default behaviors for each case.

History

#1 - 03/27/2013 02:23 PM - Redmine Admin

Original Bugzilla ID was 3578