BIGS BIG IMAGE DATA ANALYSIS TOOLKIT
RELEASE 0.2b mar 2012

PROCESS MODEL

A Stage of a Pipeline defines:

- A task (and concrete configuration parameters) *KMeans, SumSVM*
- A list (hierarchy) of Task-Containers (and concrete configuration parameters) *Iterative, CrossValidation*
- An input datasource
- An output datasource

For instance:

```
stage.01.task: bigs.modules.ml.KMeans
stage.01.container.01: bigs.modules.containers.IterativeTaskContainer
stage.01.container.02: bigs.modules.containers.DataPartitionTaskContainer
stage.01.input.source: bigs.modules.storage.HBaseDataSource
stage.01.input.table: dataset.CLEF2012
stage.01.output.source: bigs.modules.storage.HBaseDataSource
stage.01.output.table: models.CLEF2012

stage.01.KMeans.numberOfCentroids: 20
stage.01.Iteration.numberOfIterations: 2
stage.01.DataPartition.numberOfPartitions: 2
```

From this, BIGS

(1) allows each TaskContainer to tag input data as desired
(2) establishes a schedule to process all input data grouped by tags
(3) establishes execution priorities according to whether TaskContainers are parallel or sequential
(4) provides workers to to execute the schedule

PROCESS MODEL

A **Pipeline** is made of a set of consecutive **Stages**

A schedule for a Stage is hierarchy of **TaskContainers**, each container composed of a set of identical **Tasks**

A TaskContainer defines parameters and generic behavior placeholders for itself and its Task

Examples of TaskContainers: *Iteration, DataPartition, CrossValidation*

A TaskContainer defines whether it executes its blocks sequentially or in parallel

A TaskContainer also defines how input data is tagged

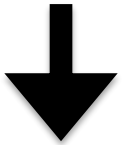
A **Task** defines concrete behaviour for the placeholder defined by certain TaskContainers (and not necesarily for all TaskContainers)

Examples of Tasks: *KMeans, RGBFeaturesExtractor, SummationFormSVM*

PROCESS MODEL

```
stage.01.task: pilot.modules.ml.KMeans
stage.01.container.01: pilot.modules.containers.IterativeTaskContainer
stage.01.container.02: pilot.modules.containers.DataPartitionTaskContainer
stage.01.input.source: bigs.modules.storage.HBaseDataSource
stage.01.input.table: dataset.CLEF2012
stage.01.outpu.source: bigs.modules.storage.HBaseDataSource
stage.01.output.table: models.CLEF2012

stage.01.KMeans.numberOfCentroids: 20
stage.01.IterativeTaskContainer.numberOfIterations: 3
stage.01.DataPartitionTaskContainer.numberOfPartitions: 3
```



```
12/04/12 12:11:11 INFO bigs: RULIX Stage 1
12/04/12 12:11:11 INFO bigs: RULIX configured task: KMeans [numberOfCentroids=20]
   TopLevelTaskContainer []
        IterativeTaskContainer [numberOfIterations=3, iterationNumber=1]
             DataPartitionTaskContainer [numberOfPartitions=3, partitionNumber=1]
             DataPartitionTaskContainer [numberOfPartitions=3, partitionNumber=2]
             DataPartitionTaskContainer [numberOfPartitions=3, partitionNumber=3]
        IterativeTaskContainer [numberOfIterations=3, iterationNumber=2]
             DataPartitionTaskContainer [numberOfPartitions=3, partitionNumber=1]
             DataPartitionTaskContainer [numberOfPartitions=3, partitionNumber=2]
             DataPartitionTaskContainer [numberOfPartitions=3, partitionNumber=3]
        IterativeTaskContainer [numberOfIterations=3, iterationNumber=3]
             DataPartitionTaskContainer [numberOfPartitions=3, partitionNumber=1]
             DataPartitionTaskContainer [numberOfPartitions=3, partitionNumber=2]
             DataPartitionTaskContainer [numberOfPartitions=3, partitionNumber=3]
```

Provided by **IterativeTaskContainer** implements **TaskContainer**

Provided by **DataPartitionTaskConatiner** implements **TaskContainer**

```
12/04/12 12:11:11 INFO bigs: RULIX Stage 1
12/04/12 12:11:11 INFO bigs: RULIX configured task: KMeans [numberOfCentroids=20]
000   TopLevelTaskContainer [].preSubContainers
001      IterativeTaskContainer [numberOfIterations=3, iterationNumber=null].preMyContainers
002         IterativeTaskContainer [numberOfIterations=3, iterationNumber=1].preSubContainers
003            DataPartitionTaskContainer [numberOfPartitions=3, partitionNumber=null].preMyContainers
004               DataPartitionTaskContainer [numberOfPartitions=3, partitionNumber=1] LOOP processDataBlock
004               DataPartitionTaskContainer [numberOfPartitions=3, partitionNumber=2] LOOP processDataBlock
004               DataPartitionTaskContainer [numberOfPartitions=3, partitionNumber=3] LOOP processDataBlock
005            DataPartitionTaskContainer [numberOfPartitions=3, partitionNumber=null].postMyContainers
006         IterativeTaskContainer [numberOfIterations=3, iterationNumber=1].postSubContainers
007         IterativeTaskContainer [numberOfIterations=3, iterationNumber=2].preSubContainers
008            DataPartitionTaskContainer [numberOfPartitions=3, partitionNumber=null].preMyContainers
009               DataPartitionTaskContainer [numberOfPartitions=3, partitionNumber=1] LOOP processDataBlock
009               DataPartitionTaskContainer [numberOfPartitions=3, partitionNumber=2] LOOP processDataBlock
009               DataPartitionTaskContainer [numberOfPartitions=3, partitionNumber=3] LOOP processDataBlock
010            DataPartitionTaskContainer [numberOfPartitions=3, partitionNumber=null].postMyContainers
011         IterativeTaskContainer [numberOfIterations=3, iterationNumber=2].postSubContainers
012         IterativeTaskContainer [numberOfIterations=3, iterationNumber=3].preSubContainers
013            DataPartitionTaskContainer [numberOfPartitions=3, partitionNumber=null].preMyContainers
014               DataPartitionTaskContainer [numberOfPartitions=3, partitionNumber=1] LOOP processDataBlock
014               DataPartitionTaskContainer [numberOfPartitions=3, partitionNumber=2] LOOP processDataBlock
014               DataPartitionTaskContainer [numberOfPartitions=3, partitionNumber=3] LOOP processDataBlock
015            DataPartitionTaskContainer [numberOfPartitions=3, partitionNumber=null].postMyContainers
016         IterativeTaskContainer [numberOfIterations=3, iterationNumber=3].postSubContainers
017      IterativeTaskContainer [numberOfIterations=3, iterationNumber=null].postMyContainers
018   TopLevelTaskContainer [].postSubContainers
```

IterativeTaskContainer declared as Sequential, **DataPartitionTaskContainer** declared as Parallel


```
12/04/12 14:56:41 INFO bigs: RULIX Stage 1
12/04/12 14:56:41 INFO bigs: RULIX configured task: KMeans [numberOfCentroids=20]
000   TopLevelTaskContainer [].preSubContainers
001      IterativeTaskContainer [numberOfIterations=3, iterationNumber=null].preMyContainers
002         IterativeTaskContainer [numberOfIterations=3, iterationNumber=1].preSubContainers
003            DataPartitionTaskContainer [numberOfPartitions=3, partitionNumber=null].preMyContainers
004               DataPartitionTaskContainer [numberOfPartitions=3, partitionNumber=1] LOOP processDataItem
004               DataPartitionTaskContainer [numberOfPartitions=3, partitionNumber=2] LOOP processDataItem
004               DataPartitionTaskContainer [numberOfPartitions=3, partitionNumber=3] LOOP processDataItem
005            DataPartitionTaskContainer [numberOfPartitions=3, partitionNumber=null].postMyContainers
006         IterativeTaskContainer [numberOfIterations=3, iterationNumber=1].postSubContainers
002         IterativeTaskContainer [numberOfIterations=3, iterationNumber=2].preSubContainers
003            DataPartitionTaskContainer [numberOfPartitions=3, partitionNumber=null].preMyContainers
004               DataPartitionTaskContainer [numberOfPartitions=3, partitionNumber=1] LOOP processDataItem
004               DataPartitionTaskContainer [numberOfPartitions=3, partitionNumber=2] LOOP processDataItem
004               DataPartitionTaskContainer [numberOfPartitions=3, partitionNumber=3] LOOP processDataItem
005            DataPartitionTaskContainer [numberOfPartitions=3, partitionNumber=null].postMyContainers
006         IterativeTaskContainer [numberOfIterations=3, iterationNumber=2].postSubContainers
002         IterativeTaskContainer [numberOfIterations=3, iterationNumber=3].preSubContainers
003            DataPartitionTaskContainer [numberOfPartitions=3, partitionNumber=null].preMyContainers
004               DataPartitionTaskContainer [numberOfPartitions=3, partitionNumber=1] LOOP processDataItem
004               DataPartitionTaskContainer [numberOfPartitions=3, partitionNumber=2] LOOP processDataItem
004               DataPartitionTaskContainer [numberOfPartitions=3, partitionNumber=3] LOOP processDataItem
005            DataPartitionTaskContainer [numberOfPartitions=3, partitionNumber=null].postMyContainers
006         IterativeTaskContainer [numberOfIterations=3, iterationNumber=3].postSubContainers
007      IterativeTaskContainer [numberOfIterations=3, iterationNumber=null].postMyContainers
008   TopLevelTaskContainer [].postSubContainers
```

IterativeTaskContainer declared as Parallel, **DataPartitionTaskContainer** declared as Parallel


```
12/04/12 15:02:28 INFO bigs: RULIX Stage 1
12/04/12 15:02:28 INFO bigs: RULIX configured task: KMeans [numberOfCentroids=20]
000   TopLevelTaskContainer [].preSubContainers
001      IterativeTaskContainer [numberOfIterations=3, iterationNumber=null].preMyContainers
002         IterativeTaskContainer [numberOfIterations=3, iterationNumber=1].preSubContainers
003            DataPartitionTaskContainer [numberOfPartitions=3, partitionNumber=null].preMyContainers
004               DataPartitionTaskContainer [numberOfPartitions=3, partitionNumber=1] LOOP processDataItem
005               DataPartitionTaskContainer [numberOfPartitions=3, partitionNumber=2] LOOP processDataItem
006               DataPartitionTaskContainer [numberOfPartitions=3, partitionNumber=3] LOOP processDataItem
007            DataPartitionTaskContainer [numberOfPartitions=3, partitionNumber=null].postMyContainers
008         IterativeTaskContainer [numberOfIterations=3, iterationNumber=1].postSubContainers
009         IterativeTaskContainer [numberOfIterations=3, iterationNumber=2].preSubContainers
010            DataPartitionTaskContainer [numberOfPartitions=3, partitionNumber=null].preMyContainers
011               DataPartitionTaskContainer [numberOfPartitions=3, partitionNumber=1] LOOP processDataItem
012               DataPartitionTaskContainer [numberOfPartitions=3, partitionNumber=2] LOOP processDataItem
013               DataPartitionTaskContainer [numberOfPartitions=3, partitionNumber=3] LOOP processDataItem
014            DataPartitionTaskContainer [numberOfPartitions=3, partitionNumber=null].postMyContainers
015         IterativeTaskContainer [numberOfIterations=3, iterationNumber=2].postSubContainers
016         IterativeTaskContainer [numberOfIterations=3, iterationNumber=3].preSubContainers
017            DataPartitionTaskContainer [numberOfPartitions=3, partitionNumber=null].preMyContainers
018               DataPartitionTaskContainer [numberOfPartitions=3, partitionNumber=1] LOOP processDataItem
019               DataPartitionTaskContainer [numberOfPartitions=3, partitionNumber=2] LOOP processDataItem
020               DataPartitionTaskContainer [numberOfPartitions=3, partitionNumber=3] LOOP processDataItem
021            DataPartitionTaskContainer [numberOfPartitions=3, partitionNumber=null].postMyContainers
022         IterativeTaskContainer [numberOfIterations=3, iterationNumber=3].postSubContainers
023      IterativeTaskContainer [numberOfIterations=3, iterationNumber=null].postMyContainers
024   TopLevelTaskContainer [].postSubContainers
```

IterativeTaskContainer declared as Sequential, **DataPartitionTaskContainer** declared as Sequential

PROCESS MODEL

```
12/04/12 12:11:11 INFO bigs: RULIX Stage 1
12/04/12 12:11:11 INFO bigs: RULIX configured task: KMeans [numberOfCentroids=20]
000   TopLevelTaskContainer [].preSubContainers
001      IterativeTaskContainer [numberOfIterations=3, iterationNumber=null].preMyContainers
002         IterativeTaskContainer [numberOfIterations=3, iterationNumber=1].preSubContainers
003            DataPartitionTaskContainer [numberOfPartitions=3, partitionNumber=null].preMyContainers
004               DataPartitionTaskContainer [numberOfPartitions=3, partitionNumber=1] LOOP processDataItem
004               DataPartitionTaskContainer [numberOfPartitions=3, partitionNumber=2] LOOP processDataItem
004               DataPartitionTaskContainer [numberOfPartitions=3, partitionNumber=3] LOOP processDataItem
005            DataPartitionTaskContainer [numberOfPartitions=3, partitionNumber=null].postMyContainers
006         IterativeTaskContainer [numberOfIterations=3, iterationNumber=1].postSubContainers
. . . . .
```

WORKER LOGIC:

Can take over the execution of any Schedule Item that:

(1) Its parent has finished

(2) Its siblings with lower priority have finished

PipelineStage

methods
- fromProperties

fields
- configuredTask
- topTaskLevel

Task

@BIGSParam

TaskContainer

@BIGSParam

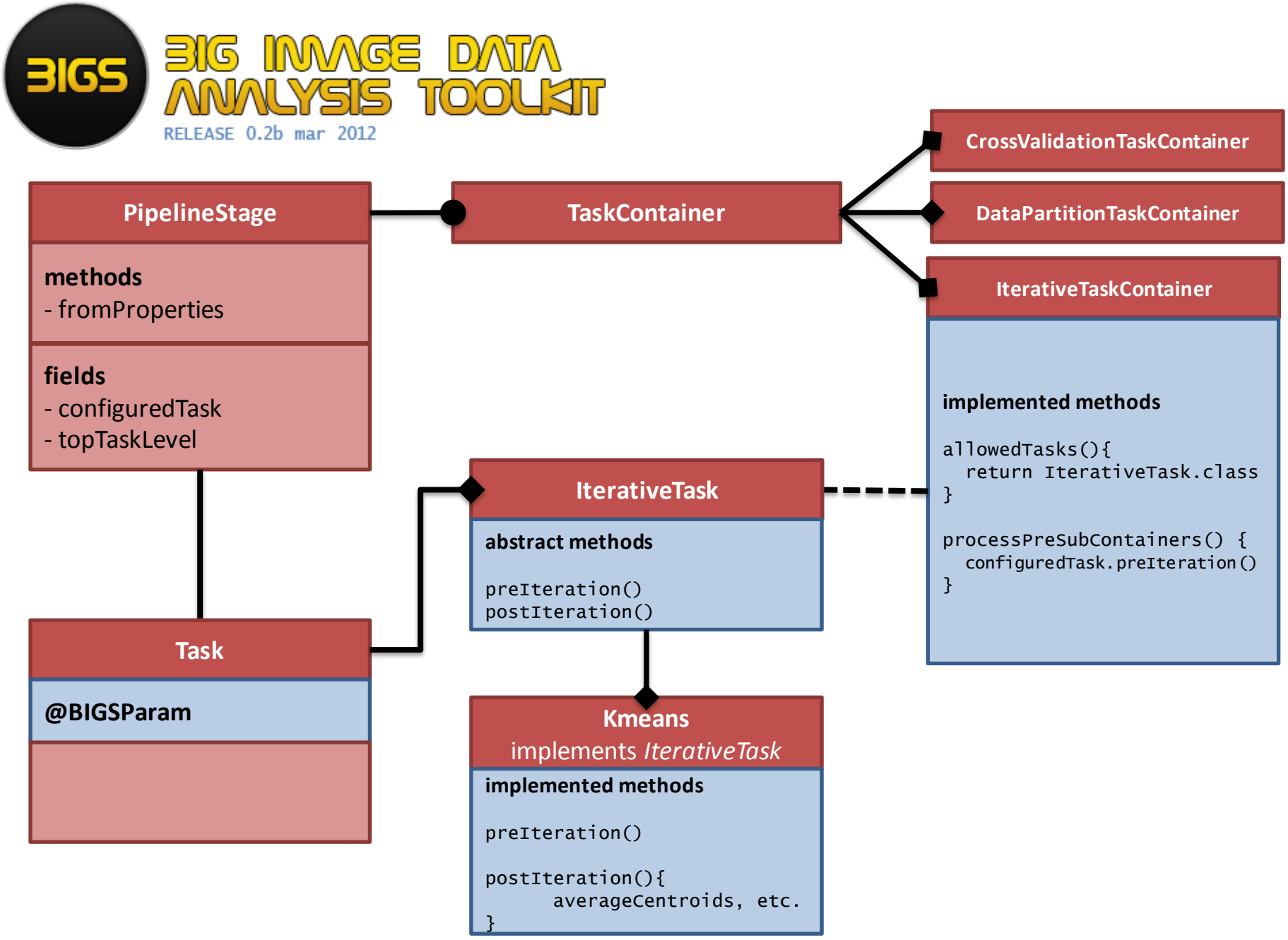
abstract methods
- supportsParallelization
- allowedTasks
- allowedTaskContainers
- allowedTags
- processPreSubContainers
- processPostSubContainers
- processPreMyContainers
- processPostMyContainers
- processPreDataBlock
- processDataItem
- processPostDataBlock
- List<String> tagDataItem(item)

fields
- taskContainers (list)
- parentTaskContainer

CrossValidationTaskContainer

DataPartitionTaskContainer

IterativeTaskContainer


BIGS
BIG IMAGE DATA ANALYSIS TOOLKIT
RELEASE 0.2b mar 2012
PipelineStage
methods
- fromProperties
fields
- configuredTask
- topTaskLevel
TaskContainer
CrossValidationTaskContainer
DataPartitionTaskContainer
IterativeTaskContainer
implemented methods
allowedTasks(){
return IterativeTask.class
}
processPreSubContainers() {
configuredTask.preIteration()
}
IterativeTask
abstract methods
preIteration()
postIteration()
Task
@BIGSParam
Kmeans
implements IterativeTask
implemented methods
preIteration()
postIteration(){
averageCentroids, etc.
}

BIGS BIG IMAGE DATA ANALYSIS TOOLKIT
RELEASE 0.2b mar 2012

DATA FLOW

Each **TaskContainer** is given the chance to produce tags for the data

Key	Content	Tags by IterativeTaskContainer Iteration 1	Iteration 2	Tags by CrossValidationTaskContainer Fold	Function	Tags by DataPartitionTaskContainer Split
1	...	1	2	1	TRAIN	1
2	...	1	2	1	TRAIN	1
3	...	1	2	1	TRAIN	2
4	...	1	2	1	TRAIN	2
5	...	1	2	1	TEST	1
6	...	1	2	1	TEST	1
7	...	1	2	2	TRAIN	2
8	...	1	2	2	TRAIN	2
9	...	1	2	2	TRAIN	1
10	...	1	2	2	TRAIN	1
11	...	1	2	2	TRAIN	2
12	...	1	2	2	TRAIN	2
13	...	1	2	2	TEST	1
14	...	1	2	2	TEST	1

Then, all data rows with SAME TAG SET are grouped in the same processing BLOCK

```
12/04/12 12:11:11 INFO bigs: RULIX Stage 1
12/04/12 12:11:11 INFO bigs: RULIX configured task: KMeans [numberOfCentroids=20]
000   TopLevelTaskContainer [].preSubContainers
001      IterativeTaskContainer [numberOfIterations=3, iterationNumber=null].preMyContainers
002         IterativeTaskContainer [numberOfIterations=3, iterationNumber=1].preSubContainers
003            DataPartitionTaskContainer [numberOfPartitions=3, partitionNumber=null].preMyContainers
004               DataPartitionTaskContainer [numberOfPartitions=3, partitionNumber=1] LOOP processDataItem
004               DataPartitionTaskContainer [numberOfPartitions=3, partitionNumber=2] LOOP processDataItem
. . . . .
```

BIGS BIG IMAGE DATA ANALYSIS TOOLKIT
RELEASE 0.2b mar 2012

DATA FLOW

STATE IS PASSED ON FORWARD BY THE FRAMEWORK TO CHILD TASKS

```
12/04/12 12:11:11 INFO bigs: RULIX Stage 1
12/04/12 12:11:11 INFO bigs: RULIX configured task: KMeans [numberOfCentroids=20]
000   TopLevelTaskContainer [].preSubContainers
001      IterativeTaskContainer [numberOfIterations=3, iterationNumber=null].preMyContainers
002         IterativeTaskContainer [numberOfIterations=3, iterationNumber=1].preSubContainers
003            DataPartitionTaskContainer [numberOfPartitions=3, partitionNumber=null].preMyContainers
004               DataPartitionTaskContainer [numberOfPartitions=3, partitionNumber=1] LOOP processDataBlock
004               DataPartitionTaskContainer [numberOfPartitions=3, partitionNumber=2] LOOP processDataBlock
004               DataPartitionTaskContainer [numberOfPartitions=3, partitionNumber=3] LOOP processDataBlock
005            DataPartitionTaskContainer [numberOfPartitions=3, partitionNumber=null].postMyContainers
006         IterativeTaskContainer [numberOfIterations=3, iterationNumber=1].postSubContainers
007         IterativeTaskContainer [numberOfIterations=3, iterationNumber=2].preSubContainers
008            DataPartitionTaskContainer [numberOfPartitions=3, partitionNumber=null].preMyContainers
009               DataPartitionTaskContainer [numberOfPartitions=3, partitionNumber=1] LOOP processDataBlock
009               DataPartitionTaskContainer [numberOfPartitions=3, partitionNumber=2] LOOP processDataBlock
009               DataPartitionTaskContainer [numberOfPartitions=3, partitionNumber=3] LOOP processDataBlock
010            DataPartitionTaskContainer [numberOfPartitions=3, partitionNumber=null].postMyContainers
011         IterativeTaskContainer [numberOfIterations=3, iterationNumber=2].postSubContainers
012         IterativeTaskContainer [numberOfIterations=3, iterationNumber=3].preSubContainers
...........
```

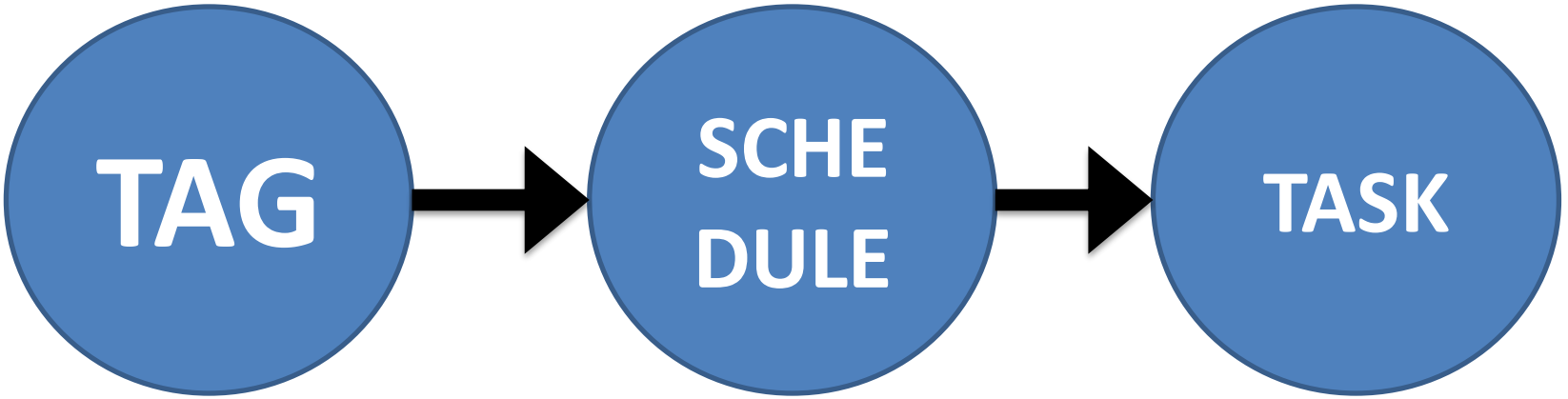
TASK RESULTS and DATA ARE GATHERED AND PASSED ON BY THE FRAMEWORK ACCORDING TO SCHEDULE PRIORITIES AND PARENTSHIP. Must have two channels

- **For processed data** (such as for feature extraction)
- **For process results** (such as for Kmeans centroids)

BIGS
BIG IMAGE DATA ANALYSIS TOOLKIT
RELEASE 0.2b mar 2012

TAG-SCHEDULE-TASK

processing model



By **TaskContainers** managed by BIGS

By **BIGS**, according to **TaskContainers** parallelization support

By **Tasks** through **TaskContainers** managed by BIGS **WORKERS**