Arvados - Story #8912

[Data manager] Verbose reporting

04/08/2016 02:47 PM - Peter Amstutz

Status:	In Progress	**Start date:**	04/13/2016
Priority:	Normal	**Due date:**	
Assigned To:		**% Done:**	0%
Category:		**Estimated time:**	0.00 hour
Target version:	Arvados Future Sprints		

Description

In order to help diagnose Keep issues, data manager should have an option to provide detailed information about inconsistencies with blocks and collections:

1. List each keep server that provided an index, and dump the contents of the index that was received from each one.

1. List all blocks considered "not in any collection"
2. List all blocks considered "missing" and list the collection UUID that reference each missing block
3. List all blocks considered "over replicated"

Subtasks:

Task # 8976: Review 8912-missing-blocks-report **In Progress**

History

#1 - 04/08/2016 02:54 PM - Peter Amstutz

- Description updated

#2 - 04/08/2016 03:29 PM - Tom Clegg

Sounds like a good feature. How about just writing a CSV report, with one line per locator: locator,have,want

Perhaps (later?) a separate option for a CSV report with one line per collection: uuid,pdh,want,n=0,n=1,n=2,... where the n=2 column indicates #blocks in this collection that are currently at replication=2.

#3 - 04/08/2016 04:00 PM - Peter Amstutz

- Description updated

Proposed format (JSON):

```
{
  "blockhash": [size, have, want, ["collections"], ["keepservers"]]
}
```

#4 - 04/08/2016 04:02 PM - Peter Amstutz

Also want a dump of the indexes received by data manager:

```
{
  "keepstore": ["hash1", "hash2", ...]
}
```

#5 - 04/08/2016 06:31 PM - Peter Amstutz

Proposed approach:

- If given a flag, data manager produces a "missing blocks" report at a specified location. This consists of pairs of [block, collection uuid].
- Write a Python script which consumes the missing blocks report and reports precisely which files within each collection are affected by missing blocks.

#6 - 04/10/2016 08:34 PM - Peter Amstutz

Copied from #8878:

To help you recover while we continue to try and diagnose the underlying bug, I've added some additional reporting to datamanager, along with an auxiliary python script. These are in the 8912-missing-blocks-report branch and for your convenience I've attached a binary of datamanager and a copy of the python script keep_block_to_file.py to this ticket.

Running datamanager -dry-run -extra-reports will produce some timestamped files, the formats are timestamp_uuid_index.txt and timestamp_uuid_missing.txt. The former is the indexes returned by each keepstore to datamanager, the latter is the collections with missing blocks.

You then use keep_block_to_file.py *_missing.txt to get the list of specific files within each collection which have missing blocks.

#7 - 04/12/2016 06:57 PM - Brett Smith

- Target version set to Arvados Future Sprints

#8 - 04/13/2016 07:09 PM - Brett Smith

- Status changed from New to In Progress

- Assigned To set to Peter Amstutz

- Target version changed from Arvados Future Sprints to 2016-04-27 sprint

#9 - 04/14/2016 04:12 PM - Tom Clegg

Looking at 8912-missing-blocks-report at 48bafad...

datamanager flag description should mention that it will create (multiple) log files in CWD in each iteration. Calling it "-debug-extra-logs" might also be a helpful signal that it will create a bit of a mess.

The proposed "missing blocks" file format (e.g., "collection_uuid,missing_block\n") still seems better to me than the "separate file for each collection" approach implemented here. For example, it would avoid the issue of creating (potentially) thousands of log files at a time if storage volumes are down, and it would make it possible to use keep_block_to_file in a unix pipeline (e.g., "zcat report.gz | py") instead of relying on regexing the report filename to get the collection uuids.

Same goes for LogKeepIndex: just one file for the entire index at time X would be less sprawl, and you can always cut/grep later if you want separate files for some reason.

Please use "continue" in error handling, instead of "else" pyramids.

Missing idxfile.Close() in LogMissingBlocks().

Python script should have some sort of usage comment.

Python script output should be machine-readable. Perhaps

```
import csv

writer = csv.writer(sys.stdout)

if st in missingblocks:
    writer.writerow([collection, name, st])
```

#10 - 04/27/2016 08:33 PM - Brett Smith

- Target version changed from 2016-04-27 sprint to 2016-05-11 sprint

#11 - 04/27/2016 08:33 PM - Brett Smith

- Target version changed from 2016-05-11 sprint to Arvados Future Sprints

#12 - 06/08/2016 07:07 PM - Peter Amstutz

- Assigned To deleted (Peter Amstutz)