2nd Provenance Challenge – CESNET Team Results

Aleš Křenek, Jiří Sitera, František Dvořák, Jiří Filipovič, Zdeněk Šustr et al.

www.eu-egee.org

Job Provenance

- see our former presentations/papers for details
- queriable **archive of jobs** executed on the Grid
- job (process) is the primary entity of interest
  - all data items assigned to a particular job
- logical data view – attributes
  - namespace:name = value
- physical data store
  - individual namespace:name = value tags
  - bulk files – processed by plugins to retrive attributes
- part of EGEE gLite middleware
  - targeted at 1M jobs/day throughput

Data dependence detection

- First Challenge – explicit dependences between workflow processes
  - workflow implemented with gLite DAG
  - dependence data directly available
  - callenge query implementation depends on them
- Second Challenge – **can't rely on explicit dependences**
  - not present in all formats
  - definitely not available accross boundaries of split workflow parts
- detection of data dependence
  - B depends on A iff there is F being output of A and input of B simultaneously
  - stored in JP explicitely $\Rightarrow$ **functional query implementation**
  - currently implmented as external script wrt. JP (recursive query)

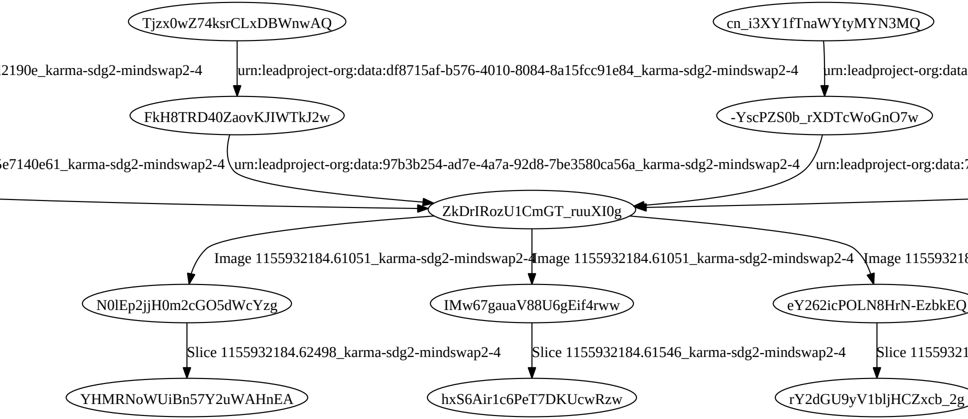
Data dependence detection

- First Challenge – explicit dependences between workflow processes
  - workflow implemented with gLite DAG
  - dependence data directly available
  - callenge query implementation depends on them
- Second Challenge – **can't rely on explicit dependences**
  - not present in all formats
  - definitely not available accross boundaries of split workflow parts
- detection of data dependence
  - B depends on A iff there is F being output of A and input of B simultaneously
  - stored in JP explicitely ⇒ **functional query implementation**
  - currently implmented as external script wrt. JP (recursive query)

Running the challenge

- convert used worklflow parts to our format
- adjust filenames (for challenge purpose only)
  - rename files to match on workflow part boundaries
  - eventually append unique suffixes to filenemes to allow multiple debugging imports of the same data
- import adjusted files
  - assign new unique job identifiers in our format
  - original ones can be presereved as process attributes
- run data-dependence detection
- run challenge queries

Combined workflow example


Tjzx0wZ74ksrCLxDBWnwAQ
cn_i3XY1fTnaWYtyMYN3MQ
12190e_karma-sdg2-mindswap2-4
urn:leadproject-org:data:df8715af-b576-4010-8084-8a15fcc91e84_karma-sdg2-mindswap2-4
urn:leadproject-org:data
FkH8TRD40ZaovKJIWTkJ2w
-YscPZS0b_rXDTcWoGnO7w
5e7140e61_karma-sdg2-mindswap2-4
urn:leadproject-org:data:97b3b254-ad7e-4a7a-92d8-7be3580ca56a_karma-sdg2-mindswap2-4
urn:leadproject-org:data:7
ZkDrIRozU1CmGT_ruuXI0g
Image 1155932184.61051_karma-sdg2-mindswap2-4
Image 1155932184.61051_karma-sdg2-mindswap2-4
Image 11559321
N0lEp2jjH0m2cGO5dWcYzg
IMw67gauaV88U6gEif4rww
eY262icPOLN8HrN-EzbkEQ
Slice 1155932184.62498_karma-sdg2-mindswap2-4
Slice 1155932184.61546_karma-sdg2-mindswap2-4
Slice 11559321
YHMRNoWUiBn57Y2uWAHnEA
hxS6Air1c6PeT7DKUcwRzw
rY2dGU9yV1bljHCZxcb_2g

Queried systems

- five systems chosen
  - ES3, SDG, Mindswap, Karma, MyGrid
  - subjective first choice, most suitable for conversion to our format
- chosen combinations
  - CESNET-Karma-SDG, ES3-CESNET-Karma, ES3-MyGrid-SDG, MyGrid-ES3-SDG, Karma-SDG-Mindswap
- virtually **no issues specific to cross queries**
  - file naming not considered important
  - in real-world either consistent naming would be used or unique name mapping available
- all other issues related to particular system
  - running also **homogeneous** queries to demonstrate them

Smooth queries

#1 *Find the process that led to Atlas X Graphic.*

#2 *Find the process that led to Atlas X Graphic, excluding everything prior to softmean outputting the Atlas Image.*

#3 *Find the Stage 3–5 details of the process that led to Atlas X Graphic.*

- OK in all homogeneous and heterogeneous queries
- missing warping parameters (ES)
  - not critical for these queries, used only in #1
  - affects all combinations taking ES3 part 1
- executables are named differently (#2)
  - appropriate name (matching workflow part 2) must be provided

Not so smooth queries

#4 *Find all invocations of procedure align_warp that have ever occurred in the system using a twelfth order nonlinear 1365 parameter model that ran on a Monday.*

#6 *Find all images ever output from softmean where the images were align_warped using a twelfth order nonlinear 1365 parameter model.*

- depend on presence of warping parameters
  - ES3: parameters missing, query impossible
  - Mindswap: wrong format "-m -12", query possible after fixing
  - MyGrid: not implemented but present, query possible
  - SDG, Karma: query possible
- affects all combinations according to part 1

Not so smooth queries

#5 *Find all Atlas Graphic images outputted from workflows where at least one of the input Anatomy Headers had an entry global maximum = 4095*

- depends on `global maximum` header entry
  - ES3, Karma, Mindswap: missing, query impossible
  - MyGrid: encoded ambiguously, not implemented
  - SDG: query possible
- affects all combinations according to part 1

Unimplemented queries

#7 *A user has run the workflow twice on the same input files, in the second instance replacing each convert procedure in the final stage . . .*

- done partially in First Challenge, not addressed in Second Challenge

#8 *A user has annotated some anatomy images with a key-value pair center=UChicago. . .*

- pure data annotation, out of scope of JP
  - but cf. MyGrid using "pseudoprocesses" producing even the inputs

#9 *A user has annotated some atlas graphics with key-value pair where the key is studyModality. . .*

- not implemented
- ES3, SDG, Mindswap: annotation not present, query impossible
- Karma, MyGrid: query possible

Unimplemented queries

#7 *A user has run the workflow twice on the same input files, in the second instance replacing each convert procedure in the final stage . . .*

- done partially in First Challenge, not addressed in Second Challenge

#8 *A user has annotated some anatomy images with a key-value pair center=UChicago. . .*

- pure data annotation, out of scope of JP
  - but cf. MyGrid using “pseudoprocesses” producing even the inputs

#9 *A user has annotated some atlas graphics with key-value pair where the key is studyModality. . .*

- not implemented
- ES3, SDG, Mindswap: annotation not present, query impossible
- Karma, MyGrid: query possible

Unimplemented queries

#7 *A user has run the workflow twice on the same input files, in the second instance replacing each convert procedure in the final stage . . .*

- done partially in First Challenge, not addressed in Second Challenge

#8 *A user has annotated some anatomy images with a key-value pair center=UChicago. . .*

- pure data annotation, out of scope of JP
  - but cf. MyGrid using “pseudoprocesses” producing even the inputs

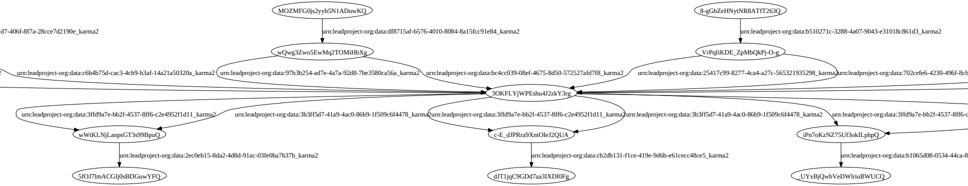
#9 *A user has annotated some atlas graphics with key-value pair where the key is studyModality. . .*

- not implemented
- ES3, SDG, Mindswap: annotation not present, query impossible
- Karma, MyGrid: query possible

Extraneous data

- processing details, file properties
- redundant data → process links
- additional processes (eg. MyGrid)

- double arcs (eg. Karma)


MOZMFG0js2yyb5N1ADuwKQ
8-gGbZeHNytNR8ATfT263Q
d7-406f-8f7a-28cce7d2190e_karma2
urn:leadproject-org:data:df8715af-b576-4010-8084-8a15fcc91e84_karma2
urn:leadproject-org:data:b510271c-3288-4a07-9043-e31018c861d3_karma2
wQwg3Zwo5EwMq2TOMiHbXg
VrPqIiKDE_ZpMbQkPj-O-g
urn:leadproject-org:data:c6b4b75d-cac3-4cb9-b3af-14a21a50320a_karma2
urn:leadproject-org:data:97b3b254-ad7e-4a7a-92d8-7be3580ca56a_karma2
urn:leadproject-org:data:bc4cc039-08ef-4675-8d50-572527afd7f8_karma2
urn:leadproject-org:data:25417c99-8277-4ca4-a27c-565321935298_karma2
urn:leadproject-org:data:702cefe6-4230-496f-8cb
3OKFLYjWPEshu4J2zkY3rg
urn:leadproject-org:data:3ffd9a7e-bb2f-4537-8ff6-c2e4952f1d11_karma2
urn:leadproject-org:data:3b3ff5d7-41a9-4ac0-86b9-1f509c6f4478_karma2
urn:leadproject-org:data:3ffd9a7e-bb2f-4537-8ff6-c2e4952f1d11_karma2
urn:leadproject-org:data:3b3ff5d7-41a9-4ac0-86b9-1f509c6f4478_karma2
urn:leadproject-org:data:3ffd9a7e-bb2f-4537-8ff6-c
wWtKLNjLanpsGT3s99BpuQ
c-E_dJPRra9XntOIeJ2QUA
iPn7oKzNZ75Uf3okILphpQ
urn:leadproject-org:data:2ec0eb15-8da2-4d8d-91ac-03fe0ba7b37b_karma2
urn:leadproject-org:data:cb2db131-f1ce-419e-9d6b-e61cecc48ce5_karma2
urn:leadproject-org:data:b1065d08-0534-44ca-8
5fOJ7lmACGIj0sBDGuwYFQ
dJT1jqC9GDd7aa3IXDI0Fg
UYxBjQwbVeDWlriuBWUCQ

Imperfect implementation

- “sewing” script (data dependence)
  - Challenge: standalone, invoked manually
  - production: agent, listening on JPPS feed interface for any input/output assingments
- import external formats
  - Challenge: external converter & import program, generating name=value tags
  - production: JPPS plugin, reading external format directly

Imperfect implementation

- “sewing” script (data dependence)
  - Challenge: standalone, invoked manually
  - production: agent, listening on JPPS feed interface for any input/output assingments
- import external formats
  - Challenge: external converter & import program, generating name=value tags
  - production: JPPS plugin, reading external format directly

Conclusions

- five foreign formats successfully imported
- queries on homogeneous workflows too
- no significant differences in query results
  - clearly show format-specific issues
- virtually no issues specific to cross-queries
- local effect on possibility to perform query and query results
  - e.g. warping parameters – part 1, queries #1, 4, 6
  - cross-query results follow those on homogeneous workflows
- process vs. data provenance
  - still apparent but less serious than expected
  - some convergence seen – pseudoprocesses "generate" wf. inputs