Federated sharing and processing of genomic datasets for tertiary data analysis

Arif Canakoglu, Pietro Pinoli, Andrea Gulino, Luca Nanni, Marco Masseroli, Stefano Ceri

Supplementary material

Federated GMQL queries

Here, we present the exact queries that we used in the manuscript and the execution log of the BEST query. All the queries are ready to use on the dedicated server (GeCo as LOCAL).
We present 4 distributed strategy, 3 centralized and the best strategy, which has been presented in the paper. We excluded externalized strategy, because the AWS instances are accessible only on demand.

DIST-1:

Distributed query 1: In this example, all the unary operations are on the machine where the dataset is selected. The binary operations are executed on the DEIB instance.

```
######################## DIST-1 (J,M:DEIB) ########################

#  1
AccRnaseq = SELECT(manually_curated__tumor_tag == "acc" AND
                   manually_curated__tissue_status == "tumoral"; at:CINECA)
                                                 CINECA.HG19_TCGA_rnaseqv2_gene;

#  2
AccExp = COVER(1, ANY; aggregate: mean_exp as AVG(normalized_count); at:CINECA)
AccRnaseq;

#  3
AccExpFilt = SELECT(region: mean_exp > 1000; at:CINECA) AccExp;

#  4
Myc = SELECT(gcm_curated__cell_line == "H1-hESC" AND
             target__name == "MYC-human" AND
             file__output_type == "conservative idr thresholded peaks"; at:DEIB)
                                                      DEIB.HG19_ENCODE_NARROW_2019_01;

#  5
GeneMyc = JOIN(dist < 0; output: left_distinct; at:DEIB) AccExpFilt Myc;

#  6
myMutation = SELECT(at:LOCAL) myMutation;

#  7
myMutationMerge = MERGE(at:LOCAL) myMutation;

#  8
GeneMycMut= MAP(count_name: mut_count; at:DEIB) GeneMyc myMutationMerge;

#  9
ResGenes = SELECT(region:mut_count > 0; at:DEIB) GeneMycMut;

# 10
MATERIALIZE ResGenes INTO ResGenes;

######################## DIST-1 (J,M:DEIB) ########################
```

DIST-2:
Distributed query 2: In this example, all the unary operations are on the machine where the dataset is selected. The binary operations are executed on the CINECA instance.

```
######################## DIST-2 (J,M:CINECA) ########################

#  1
AccRnaseq = SELECT(manually_curated__tumor_tag == "acc" AND
                   manually_curated__tissue_status == "tumoral"; at:CINECA)
                                                 CINECA.HG19_TCGA_rnaseqv2_gene;

#  2
AccExp = COVER(1, ANY; aggregate: mean_exp as AVG(normalized_count); at:CINECA)
AccRnaseq;

#  3
AccExpFilt = SELECT(region: mean_exp > 1000; at:CINECA) AccExp;

#  4
Myc = SELECT(gcm_curated__cell_line == "H1-hESC" AND
             target__name == "MYC-human" AND
             file__output_type == "conservative idr thresholded peaks"; at:DEIB)
                                                    DEIB.HG19_ENCODE_NARROW_2019_01;

#  5
GeneMyc = JOIN(dist < 0; output: left_distinct; at:CINECA) AccExpFilt Myc;

#  6
myMutation = SELECT(at:LOCAL) myMutation;

#  7
myMutationMerge = MERGE(at:LOCAL) myMutation;

#  8
GeneMycMut= MAP(count_name: mut_count; at:CINECA) GeneMyc myMutationMerge;

#  9
ResGenes = SELECT(region:mut_count > 0; at:CINECA) GeneMycMut;

# 10
MATERIALIZE ResGenes INTO ResGenes;

######################## DIST-2 (J,M:CINECA) ########################
```

DIST-3:

Distributed query 3: In this example, all the unary operations are on the machine where the dataset is selected. The binary operations, JOIN and MAP, are executed on the DEIB and GeCo (LOCAL) instance, respectively.

```
######################## DIST-3 (J:DEIB,M:GECO) ########################

#  1
AccRnaseq = SELECT(manually_curated__tumor_tag == "acc" AND
                   manually_curated__tissue_status == "tumoral"; at:CINECA)
                                                 CINECA.HG19_TCGA_rnaseqv2_gene;

#  2
AccExp = COVER(1, ANY; aggregate: mean_exp as AVG(normalized_count); at:CINECA)
AccRnaseq;

#  3
AccExpFilt = SELECT(region: mean_exp > 1000; at:CINECA) AccExp;

#  4
Myc = SELECT(gcm_curated__cell_line == "H1-hESC" AND
             target__name == "MYC-human" AND
             file__output_type == "conservative idr thresholded peaks"; at:DEIB)
                                                    DEIB.HG19_ENCODE_NARROW_2019_01;

#  5
GeneMyc = JOIN(dist < 0; output: left_distinct; at:DEIB) AccExpFilt Myc;

#  6
myMutation = SELECT(at:LOCAL) myMutation;

#  7
myMutationMerge = MERGE(at:LOCAL) myMutation;

#  8
GeneMycMut= MAP(count_name: mut_count; at:LOCAL) GeneMyc myMutationMerge;

#  9
ResGenes = SELECT(region:mut_count > 0; at:LOCAL) GeneMycMut;

# 10
MATERIALIZE ResGenes INTO ResGenes;

######################## DIST-3 (J:DEIB,M:GECO) ########################
```

DIST-4:

Distributed query 4: In this example, all the unary operations are on the machine where the dataset is selected. The binary operations, JOIN and MAP, are executed on the CINECA and GeCo (LOCAL) instances, respectively.

```
######################## DIST-4 (J:CINECA,M:GECO) ########################

#  1
AccRnaseq = SELECT(manually_curated__tumor_tag == "acc" AND
                   manually_curated__tissue_status == "tumoral"; at:CINECA)
                                                 CINECA.HG19_TCGA_rnaseqv2_gene;

#  2
AccExp = COVER(1, ANY; aggregate: mean_exp as AVG(normalized_count); at:CINECA)
AccRnaseq;

#  3
AccExpFilt = SELECT(region: mean_exp > 1000; at:CINECA) AccExp;

#  4
Myc = SELECT(gcm_curated__cell_line == "H1-hESC" AND
             target__name == "MYC-human" AND
             file__output_type == "conservative idr thresholded peaks"; at:DEIB)
                                                    DEIB.HG19_ENCODE_NARROW_2019_01;

#  5
GeneMyc = JOIN(dist < 0; output: left_distinct; at:CINECA) AccExpFilt Myc;

#  6
myMutation = SELECT(at:GeCo ) myMutation;

#  7
myMutationMerge = MERGE(at:LOCAL) myMutation;

#  8
GeneMycMut= MAP(count_name: mut_count; at:LOCAL) GeneMyc myMutationMerge;

#  9
ResGenes = SELECT(region:mut_count > 0; at:LOCAL) GeneMycMut;

# 10
MATERIALIZE ResGenes INTO ResGenes;

######################## DIST-4 (J:CINECA,M:GECO) ########################
```

CENT-1:

Centralized query 1: In this example, all the selection operations run on the machine where the dataset is selected. All the others run on DEIB instance.

```
######################## CENT-1 (DEIB) ########################

#  1
AccRnaseq = SELECT(manually_curated__tumor_tag == "acc" AND
                   manually_curated__tissue_status == "tumoral"; at:CINECA)
                                                 CINECA.HG19_TCGA_rnaseqv2_gene;

#  2
AccExp = COVER(1, ANY; aggregate: mean_exp as AVG(normalized_count); at:DEIB) AccRnaseq;

#  3
AccExpFilt = SELECT(region: mean_exp > 1000; at:DEIB) AccExp;

#  4
Myc = SELECT(gcm_curated__cell_line == "H1-hESC" AND
             target__name == "MYC-human" AND
             file__output_type == "conservative idr thresholded peaks"; at:DEIB)
                                                    DEIB.HG19_ENCODE_NARROW_2019_01;

#  5
GeneMyc = JOIN(dist < 0; output: left_distinct; at:DEIB) AccExpFilt Myc;

#  6
myMutation = SELECT(at:LOCAL) myMutation;

#  7
myMutationMerge = MERGE(at:DEIB) myMutation;

#  8
GeneMycMut= MAP(count_name: mut_count; at:DEIB) GeneMyc myMutationMerge;

#  9
ResGenes = SELECT(region:mut_count > 0; at:DEIB) GeneMycMut;

# 10
MATERIALIZE ResGenes INTO ResGenes;

######################## CENT-1 (DEIB) ########################
```

CENT-2:

Centralized query 2: In this example, all the selection operations run on the machine where the dataset is selected. All the others run on CINECA instance.

```
######################## CENT-2 (CINECA) ########################

#  1
AccRnaseq = SELECT(manually_curated__tumor_tag == "acc" AND
                   manually_curated__tissue_status == "tumoral"; at:CINECA)
                                                 CINECA.HG19_TCGA_rnaseqv2_gene;

#  2
AccExp = COVER(1, ANY; aggregate: mean_exp as AVG(normalized_count); at:CINECA)
AccRnaseq;

#  3
AccExpFilt = SELECT(region: mean_exp > 1000; at:CINECA) AccExp;

#  4
Myc = SELECT(gcm_curated__cell_line == "H1-hESC" AND
             target__name == "MYC-human" AND
             file__output_type == "conservative idr thresholded peaks"; at:DEIB)
                                                    DEIB.HG19_ENCODE_NARROW_2019_01;

#  5
GeneMyc = JOIN(dist < 0; output: left_distinct; at:CINECA) AccExpFilt Myc;

#  6
myMutation = SELECT(at:LOCAL) myMutation;

#  7
myMutationMerge = MERGE(at:CINECA) myMutation;

#  8
GeneMycMut= MAP(count_name: mut_count; at:CINECA) GeneMyc myMutationMerge;

#  9
ResGenes = SELECT(region:mut_count > 0; at:CINECA) GeneMycMut;

# 10
MATERIALIZE ResGenes INTO ResGenes;

######################## CENT-2 (CINECA) ########################
```

CENT-3:

Centralized query: In this example, all the selection operations run on the machine where the dataset is selected. All the others run on GeCo (LOCAL) instance.

```
######################## CENT-3 (GECO) ########################

#  1
AccRnaseq = SELECT(manually_curated__tumor_tag == "acc" AND
                   manually_curated__tissue_status == "tumoral"; at:CINECA)
                                                 CINECA.HG19_TCGA_rnaseqv2_gene;

#  2
AccExp = COVER(1, ANY; aggregate: mean_exp as AVG(normalized_count); at:LOCAL) AccRnaseq;

#  3
AccExpFilt = SELECT(region: mean_exp > 1000; at:LOCAL) AccExp;

#  4
Myc = SELECT(gcm_curated__cell_line == "H1-hESC" AND
             target__name == "MYC-human" AND
             file__output_type == "conservative idr thresholded peaks"; at:DEIB)
                                                    DEIB.HG19_ENCODE_NARROW_2019_01;

#  5
GeneMyc = JOIN(dist < 0; output: left_distinct; at:LOCAL) AccExpFilt Myc;

#  6
myMutation = SELECT(at:LOCAL) myMutation;

#  7
myMutationMerge = MERGE(at:LOCAL) myMutation;

#  8
GeneMycMut= MAP(count_name: mut_count; at:LOCAL) GeneMyc myMutationMerge;

#  9
ResGenes = SELECT(region:mut_count > 0; at:LOCAL) GeneMycMut;

# 10
MATERIALIZE ResGenes INTO ResGenes;

######################## CENT-3 (GECO) ########################
```

BEST:

Best query: In this example, all the selection operations run on the machine where the dataset is selected, and also the cover operation run on CINECA instance. All the others run on GeCo (LOCAL) instance.

```
######################## BEST ########################

#  1
AccRnaseq = SELECT(manually_curated__tumor_tag == "acc" AND
                   manually_curated__tissue_status == "tumoral"; at:CINECA)
                                                 CINECA.HG19_TCGA_rnaseqv2_gene;

#  2
AccExp = COVER(1, ANY; aggregate: mean_exp as AVG(normalized_count); at:CINECA)
AccRnaseq;

#  3
AccExpFilt = SELECT(region: mean_exp > 1000; at:LOCAL) AccExp;

#  4
Myc = SELECT(gcm_curated__cell_line == "H1-hESC" AND
             target__name == "MYC-human" AND
             file__output_type == "conservative idr thresholded peaks"; at:DEIB)
                                                    DEIB.HG19_ENCODE_NARROW_2019_01;

#  5
GeneMyc = JOIN(dist < 0; output: left_distinct; at:LOCAL) AccExpFilt Myc;

#  6
myMutation = SELECT(at:LOCAL) myMutation;

#  7
myMutationMerge = MERGE(at:LOCAL) myMutation;

#  8
GeneMycMut= MAP(count_name: mut_count; at:LOCAL) GeneMyc myMutationMerge;

#  9
ResGenes = SELECT(region:mut_count > 0; at:LOCAL) GeneMycMut;

# 10
MATERIALIZE ResGenes INTO ResGenes;

######################## BEST ########################
```

The log file of run of the Best execution

```
2019-07-29 15:49:38,350 Starting Federated query job_canakoglu_best_20190729_154937
2019-07-29 15:49:38,350 Splitting the computation DAG
2019-07-29 15:49:38,359 Getting DAGs to execute remotely
2019-07-29 15:49:38,363 Starting the federated query
2019-07-29 15:49:38,387 Sending sub-query to CINECA
2019-07-29 15:50:34,404 Execution time at remote (CINECA): 56 s.
2019-07-29 15:50:34,404 Moving temp_0 (metadata) from CINECA to LOCAL
2019-07-29 15:50:35,427 Execution time of moving from CINECA to LOCAL: 1 s.
2019-07-29 15:50:35,450 Sending sub-query to CINECA
2019-07-29 15:55:16,591 Execution time at remote (CINECA): 281s.
2019-07-29 15:55:16,591 Moving temp_1 (region) from CINECA to LOCAL
2019-07-29 15:55:19,616 Execution time of moving from CINECA to LOCAL: 3 s.
2019-07-29 15:55:19,638 Sending sub-query to DEIB
2019-07-29 15:55:58,969 Execution time at remote (DEIB): 39s.
2019-07-29 15:55:58,969 Moving temp_2 (metadata) from DEIB to LOCAL
2019-07-29 15:55:59,990 Execution time of moving from DEIB to LOCAL: 1 s.
2019-07-29 15:56:00,011 Sending sub-query to DEIB
2019-07-29 15:56:46,803 Execution time at remote (DEIB): 46s.
2019-07-29 15:56:46,803 Moving temp_3 (region) from DEIB to LOCAL
2019-07-29 15:56:47,827 Execution time of moving from DEIB to LOCAL: 1 s.
2019-07-29 15:56:47,828 Executing local query
2019-07-29 15:57:32,025 Execution time at local: 44 s.
2019-07-29 15:57:32,025 Total response time: 473 s.
```