Federated GMQL queries

Introduction:

Here, we present the multiple version of a query described below to show possible execution strategies on different conditions. All the queries are ready to be used on the dedicated server (GeCo as LOCAL).
In this query, we showed a case study on Adenoid Cystic Carcinoma (**ACC**), an uncommon form of malignant neoplasm that arises within secretory glands of the head and neck. The researcher has several **MUTATION** samples in her private instance, and she likes to know the highly expressed and highly mutated genes associated with **MYC** transcription factor.
She integrates her experimental dataset with public data from TCGA and ENCODE, available in CINECA and DEIB instances respectively, by performing a Federated GMQL query. This case study shows the relevance of the Federated GMQL system and also the expressive power of GMQL in building queries of biological interest.

We present 4 distributed strategies (DIST-1 to DIST-4), 3 centralized ones (CENT-1 to CENT-3), and the BEST strategy. We also showed 1 example for protective directive and 2 examples for

Distributed 1:

In this example, all the unary operations are on the machine where the dataset is selected. The binary operations, JOIN and MAP, are both executed on the DEIB instance.

```
######################## DIST-1 (JOIN, MAP: 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) Example_Mutation;

#  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 (JOIN, MAP: DEIB) ########################
```

Distributed 2:

In this example, all the unary operations are on the machine where the dataset is selected. The binary operations, JOIN and MAP, are both executed on the CINECA instance.

```
######################## DIST-2 (JOIN, MAP: 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) Example_Mutation;

#  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 (JOIN, MAP: CINECA) ########################
```

Distributed 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 (JOIN: DEIB, MAP: 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) Example_Mutation;

#  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 (JOIN: DEIB, MAP: GeCo) ########################
```

Distributed 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 (JOIN: CINECA, MAP: 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:LOCAL) Example_Mutation;

#  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 (JOIN: CINECA, MAP: GeCo) ########################
```

Centralized 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) Example_Mutation;

#  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) ########################
```

Centralized 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) Example_Mutation;

#  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) ########################
```

Centralized 3:

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) Example_Mutation;

#  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 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, i.e., where the covered dataset is selected. 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) Example_Mutation;

#  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 ########################
```

Protected dataset:

In this example, we set the protected dataset and the without setting any other location rather than first selection of the data, the system uses the default policy, distributed. It guaranties that the *protected* dataset will not move to other instances.

```
######################## Protected ########################
@protected Example_Mutation

#  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)) AccRnaseq;

#  3
AccExpFilt = SELECT(region: mean_exp > 1000) 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) AccExpFilt Myc;

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

#  7
myMutationMerge = MERGE() myMutation;

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

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

# 10
MATERIALIZE ResGenes INTO ResGenes;

######################## Protected ########################
```

Distributed policy:

In this example, we set only the locations of selection of the datasets with distributed policy, which is default.

```
######################## Distributed policy ########################
@policy distributed
#  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)) AccRnaseq;

#  3
AccExpFilt = SELECT(region: mean_exp > 1000) 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) AccExpFilt Myc;

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

#  7
myMutationMerge = MERGE() myMutation;

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

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

# 10
MATERIALIZE ResGenes INTO ResGenes;

######################## Distributed policy ########################
```

Centralized policy:

In this example, we set only the locations of selection of the datasets with centralized policy at DEIB instance. The query is equivalent to the Centralized 1:. By changing the centralized location into CINECA and LOCAL, we will have equivalent query of Centralized 2: and Centralized 3:, respectively.

```
######################### Centralized policy at DEIB ########################
@policy centralized 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)) AccRnaseq;

#  3
AccExpFilt = SELECT(region: mean_exp > 1000) 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) AccExpFilt Myc;

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

#  7
myMutationMerge = MERGE() myMutation;

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

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

# 10
MATERIALIZE ResGenes INTO ResGenes;

######################### Centralized policy at DEIB ########################
```