csc343 assignment1
Jiani Li 1003847867
Mingyu Zheng 1003797661

February 7, 2020

constraints explanations

- $\pi_{species}(Artifact) - \pi_{species}(Species) = \emptyset$.
  A species of an artifact must be a species in all species. Because what can be collected as an artifact must belong to a recorded species.

- $\pi_{rank}(Staff) \subseteq \{$'technician', 'student', 'pre-tenure', 'tenure'$\}$.
  The rank of a staff must be one of technician, student, pre-tenure or teunured. Because this is the zoological institute staff ranking schema.

- $\pi_{family}(Genus) - \pi_{family}(COL) = \emptyset$.
  Families in Genus belongs to families in COL. COL families must include families which a random genus belongs to. Since all of the families are recorded in COL.

- $\pi_{genus}(Species) \subseteq \pi_{genus}(Genus)$.
  Genuses in Species belong to genuses in Genus. The genus of Species must be included in the genus of Genus.Since Species are subsets of Genus.

- $\pi_{CID}(Collected) = \pi_{CID}(Collection)$.
  The CID attributes in the Collected and Collection relations are the same thing. Because they are all referring to the collection ID. We need the same CID appearing in both relations because 'Collection' records how the collection was collected, while 'Collected' shows more about what was collected.

- $\pi_{AN}(Artifact) = \pi_{AN}(Collected)$.
  The AN attributes in both Artifact and Collected relations represent the same thing. Because Artifacts must be Collected, and after Collected it would be recorded as an artifact.

- $\pi_{SID}(Collection) \subseteq \pi_{SID}(Staff)$.
  The SIDs in Collection all belong to SIDs in Staff. Because the person who collected a collection must be a staff.

- $\pi_{SID}(Artifact) \subseteq \pi_{SID}(Staff)$.
  The SIDs in Artifact belong to SIDs in Staff. Because the technician who maintains an artifact must be a staff.

- $\pi_{type}(Artifact) \subseteq \{'tissue','image','model','live'\}$
  As described in relations, the type of an artifact must be one of tissue, image, model or live.This is how the institute categorize Artifact.

- $\pi_{AN}(Published) \subseteq \pi_{AN}(Artifact)$
  The AN in Published must be the AN in Artifact. Beacuse the published artifacts must be in the collected artifacts in the field.

queries

Write relational algebra expressions for each of the queries below. You must use notations from this course and operators:

$$\pi, \sigma, \rho, \bowtie, \bowtie_{condition}, \times, \cap, \cup, -, =$$

You may also use constants:

today (for current date)   ∅ (for the empty set)

In your queries pay attention to the following:

- All relations are sets, and you may only use relational algebra operators covered in Chapter 2 of the course text.
- Do not make assumptions that are not enforced by our constraints above, so your queries should work correctly for any database that obeys our schema and constraints.
- Other than constants such as 23 or "lupus", a select operation only examines values contained in a tuple, not aggregated over an entire column.
- Your selection conditions can use arithmetic operators, such as $+, \leq, \neq, \geq, >, <$ and friends. You can use logical operators such as $\vee, \wedge$, and $\neg$, and treat dates and numeric attributes as numbers that you can perform arithmetic on.
- Use good variable names and provide lots of comments to explain your intentions.
- Return multiple tuples if that is appropriate for your query.

There may be a query or queries that cannot be expressed in the relational algebra you have been taught so far, in which case just write "cannot be expressed." The queries below are not in any particular order.

1. Rationale: Performance reviews include seeing how current the work is of staff who have held their current rank for a long time.

   Query: Find the most recent collection date of any artifact collected by a staff member who has held their current rank the longest. Keep ties.

   Find SID of all the staff that aren't held their current rank the longest

   $$shorterThan(SID) := \pi_{s1.SID}(\sigma_{s1.date>s2.date}((\rho_{s1}Staff) \times (\rho_{s2}Staff)))$$

   Find SID of the staff that held his current rank the longest

   $$longestStaff(SID) := \pi_{SID}Staff - shorterThan$$

   Find all the collection dates of the longest current rank staff

   $$dates(date) := \pi_{date}(\sigma_{Collection.SID=longestStaff.SID}(Collection \times longestStaff))$$

   Find the most recent collection dates of the longest current rank staff

   $$mostRecent(date) := \pi_{d1.date}(\sigma_{d1.date<d2.date}((\rho_{d1}dates) \times (\rho_{d2}dates)))$$

2. Rationale: Staff who maintain every artifact in some collection should be considered favourably in performance reviews.

   Query: Find all staff who maintain all artifacts in at least one collection.
   Link Collected.CID with Artifact.SID, since each artifact is related with one SID, only need SIDs to find all artifacts

   $$temp(CID, SID) := \pi_{CID,SID}(Collected \bowtie Artifact)$$

   If a staff maintain all artifacts in one collection, then for same CID, there should be only one SID. if not, resulting in collections that are not maintained by one staff

   $$notByOneStaff(CID) := \pi_{t1.CID}(\sigma_{t1.CID=t2.CID \wedge t1.SID \neq t2.SID}(\rho_{t1}temp \times \rho_{t2}temp))$$

   Get collections that are maintained by one staff

   $$byOneStaff(CID) := \pi_{CID}(Collected) - notByOneStaff$$

   Get the SID of the staff who maintain all artifacts in one collection

   $$maintainAllArtifact(SID) := \pi_{SID}(byOneStaff \bowtie temp)$$

3. Rationale: An artifact collected and maintained by the same staff may have some special requirements that should be investigated.

   Query: Find all artifacts that were collected by the same staff who maintains them.
   Link Collection.SID with Collected.AN

   $$Combine1(CID, date, SID, AN) := Collection \bowtie Collected$$

   Link Collection.SID with Artifact.SID by Collection.AN=Artifact.AN(collected by same one maintained)

   $$Combine2(CID, date, Collection.SID, AN, species, type, location, Artifact.SID) :=$$

   $$Combine1 \bowtie_{Combine.AN=Artifact.AN} Artifact$$

   Get artifacts that collected and maintained by same person(SID)

   $$Answer(AN) := \pi_{AN}(\sigma_{Artifact.SID=Collection.SID}Combine2)$$

4. Rationale: Identify multi-talented field workers.

   Query: Find all staff who have collected at least 3 artifacts from every species in some family.
   Extract AN and species from Artifact

   $$ArtifactNew(AN, species) := \pi_{AN,species}Artifact$$

   Get a Temp with joining attributes

   $$Temp(CID, AN, date, SID, species, genus, family) :=$$

   $$(((Collected \bowtie Collection) \bowtie ArtifactNew) \bowtie Species) \bowtie Genus$$

Rule out the species with less than 3 artifacts

$$actualAtLeast3Artifacts(SID, species, family) :=$$

$$\pi_{T1.SID,T1.species,T1.family}(\sigma_{T1.SID=T2.SID=T3.SID,T1.AN<T2.AN<T3.AN}$$
$$((\rho_{T1}Temp) \times (\rho_{T2}Temp) \times (\rho_{T3}Temp)))$$

Get all possible pairs of species and family and store them in Temp2

$$Temp2(species, family) := \pi_{species,family}(Genus \bowtie Species)$$

Find what the relation should be to meet the problem statement

$$should(SID, species, family) :=$$

$$(\pi_{SID,family} actualAtLeast3Artifacts) \bowtie_{actualAtLeast3Artifacts.family=Temp2.family} (Temp2)$$

Find those didn't always occur(i.e. failures) by using 'should' - 'those did occur'

$$Failures(SID, species, family) := should - actualAtLeast3Artifacts$$

Substract failures from all to get the answer, here we keep family to avoid rule out SID that satisfying some families while not satisfying some other families

$$Answer'(SID) := \pi_{SID,family} actualAtLeast3Artifacts - \pi_{SID,family} Failures$$

$$Answer(SID) := \pi_{SID} Answer$$

5. Rationale: Which publications might have some specialized niche focus?

   Query: Find all publications that have used exactly 2 of our artifacts.
   exactly 2 = more than 2 - more than 3; Published.AN is same, Published.journal is different

$$twoOrMore(journal) := \pi_{journal}(\sigma_{p1.AN<p2.AN \wedge p1.journal=p2.journal}$$

$$((\rho_{p1}Published) \times (\rho_{p2}Published)))$$

$$threeOrMore(journal) :=$$

$$\pi_{journal}(\sigma_{p1.AN<p2.AN \wedge p1.journal=p2.journal \wedge p2.AN<p3.AN \wedge p2.journal=p3.journal}$$
$$(((\rho_{p1}Published) \times (\rho_{p2}Published)) \times (\rho_{p3}Published)))$$

$$exactlyTwo(journal) := twoOrMore - threeOrMore$$

6. Rationale: Identify motherlode locations.

   Query: Find all locations where at least one artifact from every family has been collected.
   fail=actual-should; answer=actual-fail;

$$actual(location, family) := \pi_{location,family}((Artifact \bowtie Species) \bowtie Genus)$$

$$should(location, family) := actual \bowtie COL$$

$$fail(location, family) := should - actual$$

$$Answer(location) := \pi_{location} actual - \pi_{location} fail$$

7. Rationale: Exclusively tissue sample collectors may need extra support for special reagents and shipping costs.

   Query: Find all staff who have collected only tissue samples.
   Find all staff that have collected types other than tissue, and get staff who only collected tissues by subtracting notOnly from all

$$temp(SID, type) := \pi_{T1.SID,type}(\sigma_{T1.AN=Artifact.AN}(\rho_{T1}(Collection \bowtie Collected)) \times Artifact)$$

$$notOnly(SID) := \pi_{SID}(\sigma_{type='image' \vee type='model' \vee type='line'} temp)$$

$$Answer(SID) := \pi_{SID} temp - notOnly$$

8. Rationale: Collection staff who should be encouraged to diversify their network.

   Query: Find all staff pairs who have worked only with each other on collections.
   Step 1.Find all work pairs; Step 2.Find all solos; Step 3.Find all staff who have worked with at least 2 people; Step 4.Find all duplicates in work pairs; Step 5.Use 1-2-3-4
   Step 1.Find all work pairs;

$$Temp(CID, collector, maintainer) :=$$

$$\pi_{T.CID,T.SID,Artifact.SID}(\rho_T(Collected \bowtie Collection) \bowtie_{T.AN=Artifact.AN} Artifact)$$

$$CMPair := \pi_{collector,maintainer} Temp$$

$$MMPair := \pi_{T1.maintainer,T2.maintainer}(\sigma_{T1.CID=T2.CID \wedge T1.maintainer<T2.maintainer}(\rho_{T1} Temp \times \rho_{T2} Temp))$$

$$workPairs1(SID1, SID2) := CMPairr \cup MMPair$$

   To avoid duplicates like(1,2)(2,1):

$$workPairs2(SID1, SID2) := \pi_{SID2,SID1} workPairs1$$

$$workPairs3(SID1, SID2) := \pi_{SID2,SID1} workPairs1 \cup workPairs2$$

$$workPairs(SID1, SID2) := \sigma_{SID1<SID2} workPairs3$$

Step2.Find all solos

$$notSoloCollection(CID) :=$$

$$\sigma_{T1.CID}(\sigma_{T1.CID=T2.CID \wedge (T1.collector \neq T2.maintainer \vee T1.maintainer \neq T2.maintainer)}(\rho_{T1} Temp \times \rho_{T2} Temp))$$

$$soloCollection(CID) := \pi_{CID} Temp - notSoloCollection$$

$$soloStaff(SID) := \pi_{collector}(soloCollection \bowtie Temp)$$

Step 3. Find all work with multiple people

$$workMultiple(SID1, SID2) :=$$

$$\pi_{T1.collector, T1.maintainer}(\sigma_{T1.CID=T2.CID \wedge T1.collector \neq T1.maintainer \neq T2.collctor \neq T2.maintainer}(\rho_{T1} Temp \times \rho_{T2} Temp))$$

$$workMultiple(CID) := \pi_{SID1} workMultiple \cup pi_{SID2} workMultiple$$

Step 4. take care of "only with each other"
To avoid (1, 2) (1, 3) cases

$$fail1_1(SID) := \pi_{p1.SID1}(\sigma_{p1.SID1=p2.SID1 \wedge p1.SID2 \neq p2.SID2}(\rho_{p1} workPairs, \rho_{p2} workPairs))$$

$$fail1_2(SID) := \pi_{p1.SID2}(\sigma_{p1.SID1=p2.SID1 \wedge p1.SID2 \neq p2.SID2}(\rho_{p1} workPairs, \rho_{p2} workPairs))$$

$$fail1_3(SID) := \pi_{p2.SID2}(\sigma_{p1.SID1=p2.SID1 \wedge p1.SID2 \neq p2.SID2}(\rho_{p1} workPairs, \rho_{p2} workPairs))$$

$$fail1(SID) := fail1_1 \cup fail1_2 \cup fail1_3$$

To avoid (1, 3) (2, 3) cases

$$fail2_1(SID) := \pi_{p1.SID1}(\sigma_{p1.SID1 \neq p2.SID2 \wedge p1.SID2=p2.SID2}(\rho_{p1} workPairs, \rho_{p2} workPairs))$$

$$fail2_2(SID) := \pi_{p1.SID2}(\sigma_{p1.SID1 \neq p2.SID2 \wedge p1.SID2=p2.SID2}(\rho_{p1} workPairs, \rho_{p2} workPairs))$$

$$fail2_3(SID) := \pi_{p2.SID1}(\sigma_{p1.SID1 \neq p2.SID2 \wedge p1.SID2=p2.SID2}(\rho_{p1} workPairs, \rho_{p2} workPairs))$$

$$fail2(SID) := fail2_1 \cup fail2_2 \cup fail2_3$$

To avoid (1, 2) (2, 3) cases

$$fail3_1(SID) := \pi_{p1.SID1}(\sigma_{p1.SID1 \neq p2.SID2 \wedge p1.SID2=p2.SID1}(\rho_{p1} workPairs, \rho_{p2} workPairs))$$

$$fail3_2(SID) := \pi_{p1.SID2}(\sigma_{p1.SID1 \neq p2.SID2 \wedge p1.SID2=p2.SID1}(\rho_{p1} workPairs, \rho_{p2} workPairs))$$

$$fail3_3(SID) := \pi_{p2.SID2}(\sigma_{p1.SID1 \neq p2.SID2 \wedge p1.SID2=p2.SID1}(\rho_{p1} workPairs, \rho_{p2} workPairs))$$

$$fail3(SID) := fail3_1 \cup fail3_2 \cup fail3_3$$

Step 5. Substraction

$$failWorker(SID) := soloStaff \cup workMultiple \cup fail1 \cup fail2 \cup fail3$$

$$possibleFailPairs(SID1, SID2) := \pi_{failWorker.SID, Staff.SID}(failWorker \times Staff)$$
$$\cup \pi_{Staff.SID, failWorker.SID}(failWorker \times Staff)$$

$$Answer(SID1, SID2) := workPairs - possibleFailPairs$$

9. Rationale: Track the influence of a given staff member.

   Query: Staff member SID_1 is influenced by staff member SID_2 if (a) they have ever worked together on a collection or (b) if SID_1 has ever worked with a staff member who is influenced by SID_2. Find SIDs of staff members influenced by SID 42.

   Cannot Be Expressed!

your constraints

For each of these constraints you should derive a relational algebra expression of the form $R = \emptyset$, where R may be derived in several steps, by assigning intermediate results to a variable. If the constraint cannot be expressed in the relational algebra you have been taught, write "cannot be expressed."

1. No species is also a genus.
   $\sigma_{Genus.genus=Species.genus}(Genus \times Species) = \emptyset$

2. No genus belongs to more than one family.
   $\sigma_{g1.genus=g2.genus \wedge g1.family \neq g2.family}((\rho_{g1}Genus) \times (\rho_{g2}Genus)) = \emptyset$

3. All publications must be published after all artifacts they use have been collected.
   $\sigma_{Collection.date>=Published.date}((Collection \bowtie Collected) \bowtie_{Collected.AN=Published.AN} Published) = \emptyset$

4. Students may not catalogue live artifacts.
   $\sigma_{Staff.rank='student' \wedge Artifact.type='live'}(Artifact \bowtie Staff) = \emptyset$

submissions

Submit **a1.pdf** on MarkUs. One submission per group, whether a group is one or two people. You declare a group by submitting an empty, or partial, file, and this should be done well before the due date. You may always replace such a file with a better version, until the due date.

Double check that you have submitted the correct version of your file by downloading it from MarkUs.

marking

We mark your submission for correctness, but also for good form:

- For full marks you should add comments to describe the *data*, rather than *technique*, of your queries. These may help you get part marks if there is a flaw in your query.
- Please use the assignment operator, “:=” for intermediate results.
- Name relations and attributes in a manner that helps the reader remember their intended meaning.
- Format the algebraic expressions with line breaks and formatting that help make the meaning clear.