December 20, 2018
Arnab Chakrabarti, Rihan Hai

Implementation of Databases (WS 18/19)

Exercise 6

Due until January 15, 2019, 10am.
Please submit your solution *in a single PDF file* before the deadline to the L^{2}P system!
Please submit solutions in groups of three students.

Exercise 6.1 (Cost Estimation using Spark) (9 pts)

Consider the Chinook database from Exercise 1.2 and the 5th Query *Q5*.

1. Install Spark 2.4[1]. In Exercise 1.2 you have data stored in PostgreSQL database. Export the data from tables **Track** and **InvoiceLine** individually to CSV files. Load these two CSV files into Spark. Alternatively, you can also load the data frames directly from PostgreSQL. Provide your codes. **(3 pts)**

2. Run the previous query Q5 from Exercise 1.2.5 in Spark and provide the logical plan and the physical plan. Note if you have directly connect to PostgreSQL in last task, then for this task note that the query has to be processed within Spark. Provide your codes. **(4 pts)**

3. Compare the query plans of Q5 by Spark and PostgreSQL (Exercise 1.2.5), and itemize the differences. **(2 pts)**

   Important: Include the screenshots regarding every intermediate step.

Exercise 6.2 (Datalog) (8 pts)

1. Given the following extensional database:

   - teamNBA(X): X is a NBA basketball team
   - coach(X, Y): X is a coach of team Y
   - player(X,Y): X is a basketball player of team Y
   - taller(X,Y): X is taller than Y.

   Please formalize the following rules using Datalog and the given predicates (notably, there might be also non-NBA teams in this database):

   (a) shorterTeammate(X,Y): X and Y are players in the same NBA basketball team, and X is shorter than Y. **(1 pts)**

[1] `https://spark.apache.org/downloads.html`

(b) coachOrPlayer(X,Y): X is a coach or a player of a NBA basketball team Y. **(1 pts)**

(c) coachTallest(X,Y): X is a coach of a NBA team Y, and X is taller than all the players in the team Y. **(2 pts)**

2. You have given the below facts F, rules R, and a query Q.

   (a) Decide if the program given by the rules R is stratifiable. State why or why not it is stratifiable (corresponding graph with strati and statement). **(1 pts)**

   (b) Given the facts F and rules R do a fixpoint computation with all intermediate steps. Based on the computation, write down all answers to query Q. **(3 pts)**

F: s(a,b) s(b,c) s(c,b) r(a) r(c) r(d)

R:
p(X,Y) ← s(X,Y), NOT r(Y)
q(X,Y) ← p(Y,X), s(X,Y)
q(X,Y) ← q(Y,X), r(X)
t(X,Y) ← r(X), q(X,Y)

Q: t(X,Y)

Exercise 6.3 (Data Integration) **(6 pts)**

Given is the following global schema with three relations:

$$Hospital(HospitalName, Street, CityName, Postcode)$$
$$Doctor(DoctorName, HospitalName, Disease)$$
$$Patient(PatientName, Age, AttendingDoctorName)$$

There are four data sources:

- DS1: AachenHospital(HospitalName, Street,Postcode): hospitals in Aachen.
- DS2: DoctorsAndPatients(DoctorName, PatientName, Disease): doctors and their patients.
- DS3: TeenagePatients(PatientName, Age): patient information with the patient age less than 18.
- DS4: Hospital(HospitalName, Street, Postcode, CityID), City(CityID, CityName, Country): hospitals and cities.

1. Provide the LAV mappings between the source DS1 and the global schema.
2. Provide the LAV mappings between the source DS2 and the global schema.
3. For DS3 and the global schema, which mapping is more precise, a GAV mapping or a LAV mapping? Why? Also provide the mapping you think is more precise.
4. Consider DS4 and the global schema. Rewrite the below query on the global schema to a query on the schema of DS4.

   $q(CityName) :- Hospital('FrancisHospital', \_, CityName, \_)$.

Exercise 6.4 (Answer questions briefly) **(7 pts)**

1. What is the goal of systems like Pig Latin or Hive? **(2 pts)**
2. When do you need to do shuffling in Spark? What are narrow and wide dependencies? **(2 pts)**
3. Sketch a data integration architecture. What is a mediator? What is the task of a mediator in a data integration architecture? **(2 pts)**
4. What is the Herbrand Base and the Herbrand Model? **(1 pts)**