Assignment 3 Solution
CIS 670 Data Science, Winter 2019

1. In a mediator (data integration) system, the mediated schema is

$Customer(ID, name, statecode)$
$State(statecode, statename)$

There are two data resources:

S1:

$Customer(ID, first\_name, last\_name, region\_code)$
$Region(region\_code, region\_name)$

S2:

$Customer(ID, full\_name, region)$

Assume “name” in the mediated schema means “full_name” in S2. “Region” in data resources means “state” in the mediated schema. For example, “statecode” or “region_code” of “Oregon” is “OR.” Can you do schema mediation (mapping) in Global-as-View or Local-as-View, or both? Write down your solution(s) in SQL views.

Write a query “List all customer names in ‘Oregon’” in SQL based on the mediated schema. Can the query be answered with your views?

Solution:

1) Global-as-View:

```
Create View Customer as
Select S2.Customer.ID AS ID, S2.Customer.full_name AS name, S1.Region.region_code AS statecode
From S2.Customer, S1.Region
Where S2.Customer.region = S1.Region.region_name
Union
Select ID, CONCAT(first_name, “ ”, last_name) AS name, region_code AS statecode
from S1.Customer
```

```
Create View State as
Select region_code AS statecode, region_name AS state_name
from S1.Region
```

However, if join across tables in different databases is hard to implement. Another way is:

Create State View first:

```
Create View State as
Select region_code AS statecode, region_name AS state_name
from S1.Region
```

Then use it in the Customer view definition:

```
Create View Customer as
Select S2.Customer.ID AS ID, S1.Customer.full_name AS name, State.statecode AS statecode
From S2.Customer, State
Where S2.Customer.region = State.state_name
Union
Select ID, CONCAT(first_name, “ ”, last_name) AS name, region_code AS statecode
from S1.Customer
```

2) Local-as-View

If we have a way extract first name and last name correctly from a full name, then we can define that

```
Create View S2.Customer AS
Select Customer.ID, Customer.name, State.statename
From Customer, State
Where Customer.statecode = State.statecode
```

```
Create View S1.Customer AS
Select ID, ExtractFirstName (name) AS first_name, ExtractLastName (name) AS last_name, statecode AS region_code
from Customer
```

```
Create View S1.Region AS
Select statecode AS region_name, statename AS region_name
From State
```

However, it is hard to use existing functions in SQL to create ExtractFirstName and ExtractLastName, considering there are different types full names, such as “Last, First”, “First Middle Last”, or the names with “de,” “von” etc.

Query:

Select name
From Customer, State
Where Customer.statecode = State.statecode AND State.statename= Oregon

2. Prove the following statement: Given two LAV data integration systems $\mathcal{I}_1 = \langle \mathcal{G}, \mathcal{S}_1, \mathcal{M}_1 \rangle$ and $\mathcal{I}_2 = \langle \mathcal{G}, \mathcal{S}_2, \mathcal{M}_2 \rangle$, $\mathcal{I}_1$ is **p-contained** in $\mathcal{I}_2$ if, for each query Q, $cert_{[Q,\mathcal{I}_1]}$ equivalent to Q implies $cert_{[Q,\mathcal{I}_2]}$ equivalent to Q.

Proof: Given two LAV data integration systems $\mathcal{I}_1 = \langle \mathcal{G}, \mathcal{S}_1, \mathcal{M}_1 \rangle$ and $\mathcal{I}_2 = \langle \mathcal{G}, \mathcal{S}_2, \mathcal{M}_2 \rangle$, $\mathcal{I}_1$ is **p-contained** in $\mathcal{I}_2$ means that all the queries that can be answered by the views defined in $\mathcal{I}_1$ should also be answered by the views defined in $\mathcal{I}_2$.

For each query Q, $cert_{[Q,\mathcal{I}_1]}$ equivalent to Q means that $cert_{[Q,\mathcal{I}_1]}$ is an equivalent rewriting of Q wrt $\mathcal{I}_1$. It means that Q can be equivalently rewritten in terms of $\mathcal{S}_1$ and $\mathcal{M}_1$, which are a set of views based on G (definition of LAV). The answers for $cert_{[Q,\mathcal{I}_1]}$ based on the set of views are the answers for Q based on G. For the same query Q, if $cert_{[Q,\mathcal{I}_2]}$ is also an equivalent rewriting of Q wrt $\mathcal{I}_2$ (i.e., $cert_{[Q,\mathcal{I}_2]}$ equivalent to Q), it means that Q can be equivalently rewritten in terms of $\mathcal{S}_2$ and $\mathcal{M}_2$, which are another set of views based on the same G. The answers for $cert_{[Q,\mathcal{I}_1]}$ based on the set of views are the answers for Q based on the same G. Therefore, if Q can be answered by the views in $\mathcal{I}_1$, Q can be answered by the views in $\mathcal{I}_2$.

3. Give a real world multimodal data fusion system which is not mentioned either in the lecture or in the survey paper by (Atrey *et al.* 2010).

Any system from research papers or industry reports is good.

4. What is the relationship between Deep Learning and Big Data? Give a real world Big Data application which Deep Learning technique(s) show advantages than other traditional machine learning techniques (e.g., Bayes Networks, SVM), why? Give an example Deep Learning technique or application that Big Data help optimize or improve the performance, why?

Examples from research papers or industry reports are good.