**CPTS 415   
Big Data**

Assignment 1

1. [Big Data concept] (10) Give one example of Big Data application you know. Use the detailed example to explain each of the four Big V’s . If you are required to design a database system for this application, what are the best data models (relational, XML, RDF, among others) you would use to represent the data and why?
2. [Relational Data Model] (30) As of January 2017, the OpenFlights Airports Database (https://openflights.org/data.html) contains over 10,000 airports, train stations and ferry terminals spanning the globe. Each entry in the Airport table contains the following:  
   -------------------------------------------------------------------------------------------------------------------------------------------

**Airport ID** Unique OpenFlights identifier for this airport.

**Name** Name of airport. May or may not contain the City name.

**City** Main city served by airport. May be spelled differently from Name.

**Country** Country or territory where airport is located. See countries.dat to cross-reference to ISO 3166-1 codes.

**IATA** 3-letter IATA code. Null if not assigned/unknown.

**ICAO** 4-letter ICAO code.

**Latitude** Decimal degrees, usually to six significant digits. Negative is South, positive is North.

**Longitude** Decimal degrees, usually to six significant digits. Negative is West, positive is East.

**Altitude** In feet.

**Timezone** Hours offset from UTC. Fractional hours are expressed as decimals, eg. India is 5.5.

**DST** Daylight savings time. One of E (Europe), A (US/Canada), S (South America), O (Australia), Z (New Zealand), N (None) or U (Unknown). See also: Help: Time

**Tz** **database** **time** **zone** Timezone in "tz" (Olson) format, eg. "America/Los\_Angeles".

**Type** Type of the airport. Value "airport" for air terminals, "station" for train stations, "port" for ferry terminals and "unknown" if not known. In airports.csv, only type=airport is included.

**Source** Source of this data. "OurAirports" for data sourced from OurAirports, "Legacy" for old data not matched to OurAirports (mostly DAFIF), "User" for unverified user contributions. In airports.csv, only source=OurAirports is included.  
-------------------------------------------------------------------------------------------------------------------------------------------

(a) (5) Consider the following terms: *relation schema, relational database schema, domain, attribute, attribute domain, relation instance.* Give what these terms are with the above Airport database. Give one small (4-5 tuples) instance of the Airport table.

(b) (10) There are three databases in the OpenFlight dataset: Airport, Airline, and Route. Give the schema of these three databases and mark the primary keys, foreign keys and provide examples of functional dependencies you identified over the three tables. [You may draw a diagram to illustrate the schema, PKs, FKs and FDs]

(c) [FD inferencing] (10)

**Recall Armstrong’s axioms.**

**1. Reflexivity rule: if Y ⊆ X then X → Y**

**2. Augmentation rule: if X → Y then XZ → YZ**

**3. Transitivity rule: if X → Y and Y → Z then X → Z**

(1) Give two examples for using Armstrong’s inference rules to induce new FDs from the set of FDs you designed in question 2 (b).

(2) Prove the following inference rules also hold, using FD definition and Armstrong’s Axioms.

a. decomposition rule: **if X → YZ then: X → Y and X → Z**

b. Psuedo transitivity: **if X → Y and YW → Z then: XW → Z**

(d) [Normalization] (5)Given a relation R(A1, A2, A3, A4), with three FDs A2, A3 → A4 ; A3, A4 → A1; A1, A2→ A3. Provide the 3NF and BCNF form of the schema and explain why.

1. [Relational Algebra] (20)Consider the following database schema:

**Movies** (Title, Director, Actor);

**Location** (Theater, Address, Phone number);

**Schedule** (Theater, Title, Time).

Express the following queries in relational algebra (select σ, project ∏ , Cartesian product X, join (theta-join))

-Q1: which theaters feature “Zootopia”?

-Q2: List the names and address of theaters featuring a film directed by Steven Spielberg.

-Q3: What are the address and phone number of the Le Champo theater?

-Q4: List pairs of actors that acted in the same movie. (\* you want to use renaming on Movies and join the Movies with its copy Movie’).

1. **[XML and RDF] (40)**(a) (10) Consider the database instance you gave in Question 2 (a). Assume now that you don’t have any schema. Give an XML document to represent the tuples as the fact about the airports.

(b) (10) Consider the relational schemas you gave in Question 2 (b). Give an XML schema representation of each relational schema. How do you encode keys? Foreign keys?

(c) (20) Consider a set of natural language sentences collected from Web pages.

* + 1. A human can like another human.
    2. A human can have a sex property of a man or a woman.
    3. A man can be the father of another human.
    4. A woman can be the mother of another human.
    5. A human can be married to another human.
    6. A human can have a BirthYear property of type “xs:Year”.
    7. If a human is married to another, then they like each other.
    8. If a human is a mother or father, the human is a parent.

Write a RDF schema and give a graphical presentation to describe these relationships.

Write an instance of the RDF schema in A that express the following with a complete information that can be inferred from the schema.

* + 1. Mary is a woman and she is John’s wife.
    2. Sophie, Sandra and Susan are women.
    3. Mary and John has a son Frank.
    4. John was born in 1950.
    5. Frank was born in 1980.
    6. Susan is John’s daughter.
    7. Susan was born in 1978.
    8. Frank likes Sophie.
    9. Sandra likes Frank.

All other information for everyone else is unknown.

Write a RDF document and give a graphical presentation to describe these facts.

*Course project Milestone 1: Give a brief description on the dataset, and a brief description of the project you will be implementing and the tools you have played with. At this stage, you may also send me any questions regarding the project. Prepare for project Milestone*

*Submission of Milestone reports is optional -- completion of milestones in time guarantees a good progress of your course project.*