1. (10) Describe two interesting user cases/killer applications of Big Data and how the Big data characteristics (the four Big V’s) are reflected in these applications. If you, as a data scientist, are required to design a database system for these two applications, (a) what are the best data models (relational, XML, RDF, among others) you would use to represent the data and why? (2) Which one or more of the Big data characteristics are most important for the application and why?

2. (35) Select a public dataset (you can consider the dataset you will be using for your course project) and design a relational data model to present the data. Answer the questions below:
   A. [Schema Design] (5) Design and show the schema for your dataset with reasonable relation name, table names along with entity and referential integrity guarantees. For integrity, provide keys and possible foreign key constraints and give an algorithm on testing whether a set of attributes are keys/candidate keys.
   B. [Functional Dependencies] (5) Design and show a set of functional dependencies for your relational database in A.
   D. [Normalization] (5) Given a relation R(A₁, A₂, A₃, A₄), with FDs A₂, A₃ → A₄, A₃, A₄ → A₁, A₁, A₂ → A₃. Provide the 3NF and BCNF form of the schema and explain why.
   E. [Relational Algebra] (15) 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 π, Cartesian product X, join...)
      -Q1: who is the director of “The Hunger Games: Mockingjay”?
      -Q2: which theater feature “Minions”?
      -Q3: List the names and address of theaters featuring a film directed by Steven Spielberg.
      -Q4: What are the address and phone number of the Le Champo theater?
      -Q5: What movies are featured at the Le Champo at 21:00?
3. (25) In this part of the assignment, you will create your own XML and DTD representation of the dataset. Given the dataset you selected in problem 2, answer the following questions:
   A. [DTD] (10) Design and show a declaration of DTD to represent the dataset.
   B. [XML] (10) Select 5-10 entities from your data. Create a sample XML document constrained by the DTD. Submit the .xml file along with the assignment.
   C. [XML keys] (5). Describe a possible absolute and relative key for the XML document in B. 
      Hint: think about its relational counterpart in Question 1.

4. (30) In this part of the assignment, you will create an RDF schema and write an instance of RDF document.
   A. (10) Consider a set of natural language sentences collected from Web pages.
      i. A human can like another human.
      ii. A human can have a sex property of a man or a woman.
      iii. A man can be the father of another human.
      iv. A woman can be the mother of another human.
      v. A human can be married to another human.
      vi. A human can have a BirthYear property of type “xs:Year”.
      vii. If a human is married to another, then they like each other.
      viii. 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.

   B. (10) Write an instance of the RDF schema in A that express the following with a complete information that can be inferred from the schema.
      i. Mary is a woman and she is John’s wife.
      ii. Sophie, Sandra and Susan are women.
      iii. Mary and John has a son Frank.
      iv. John was born in 1950.
      v. Frank was born in 1980.
      vi. Susan is John’s daughter.
      vii. Susan was born in 1978.
      viii. Frank likes Sophie.
      ix. Sandra likes Frank.

      All other information for everyone else is unknown.

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

   C. (10) Read the following RDF instances.

   <rdf : RDF
   xmlns : rdf=http://www.w3.org/1999/02/22-rdf-syntax-ns#
xmns : rdfs=http://www.w3.org/2000/01/rdf-schema#">
Give a graph representation of the RDF schema. Note that do not miss the range and domain constraints.
About Course project Milestone 1.

Give a brief description on the dataset, 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 2.