Advanced Databases: Exam

Pierre Senellart

14 December 2022

This exam lasts three hours. The only documents allowed are ten A4 sheets (both sides), with the content of your choice. Communicating devices are strictly forbidden.

When writing code, imprecision in the syntax of the languages will be tolerated.

The exam is formed of a single problem, graded out of 20 points.

Cataloging and archiving TV programs

We consider in this problem an archival institution (such as INA in France) who wishes to catalog and archive a (potentially very large) collection of TV programs of different *types* (unique or serialized dramas, news, documentaries, talk shows, game shows, etc.). The information we want to record is:

- metadata about each individual program: title, type, summary, duration
- the persons involved in the creation of the program, with their names and role (e.g., actor, host, game participant, producer, camera jockey, etc.)
- when a program belongs to a series (e.g., the series of all 8:00 news of a given channel, a serial TV drama, a game show), metadata about the series as a whole (title, description, type) and the list of episodes of the series (with season number and episode number)
- all scheduled programmings of a particular individual program with information about the TV channel, and the start and end date and times
- the actual video of the TV show, with possibly multiple soundtracks and subtitle tracks

All items except the last are called the *meta-information*, while the last item is the *video information*.

1. (1 point) Let us first consider the sizes involved by doing an order-of-magnitude computation. Assume the archival institution wants to archive TV programs from a collection of 100 different TV channels, that a channel will typically schedule around 20 different programs per day over 20 hours, that all meta-information about an individual program takes on average 10 000 bytes (including information about series and schedulings), and that one hour of archived video takes 10^9 bytes, audio and subtitle tracks included.

   How much data is required to store 50 years worth of:

   - meta-information;
   - video information.

   Do not hesitate to make (reasonable) approximations in this computation, we are only interested in a rough order of magnitude.

   Comment on the hardware required to store such amounts of data, including backups.

2. (3 points) Propose a complete relational (SQL) schema (i.e., a list of tables, with their attributes and types) to represent all meta-information but not the video information. Is it feasible to store this amount of meta-information in a regular database management system such as Oracle and PostgreSQL? What about the video information?

3. (1 point) Write a SQL query that retrieves all participants to all episodes of the *Fort Boyard* game show, with the earliest date that particular episode appeared on TV.

4. (1 point) How many updates would typically be issued on such a database (say, per day)? Is a storage solution that implements strong ACID guarantees required for such an application? Explain.

5. (1 point) Assume that we want to use BigTable/HBase to store the video information. Propose a way to organize this information into the HBase data model: what should be the key? what should be the columns? Given the previous computation, what would be a reasonable number of computers/nodes in a BigTable cluster storing this data?

6. (1 point) Same question but using a key–value store, such as a DHT.

7. (1.5 point) We want to compute a representative image (*thumbnail*) of every video. Propose a MapReduce program (in pseudo-code) that computes these thumbnails on top of the video information stored in HBase, and stores the result within HBase as well.

8. (4 points) Choose 2 out of the 3 following pairs of data model and query languages:
   - XML and XQuery;
   - Graph databases and Cypher;
   - RDF and SPARQL.

   For the 2 technologies you have chosen, illustrate (using a diagram) how the meta-information about a specific episode of the TV drama of your choice would be represented in this data model, and how you would write a query that retrieves all different types of programs that were scheduled on January 1st, 2000, at 00:00.

9. (1.5 point) Comment on the advantages and disadvantages of each of the 3 technologies from the previous question for this particular use case, compared to the relational model.

10. (1 point) Assume that the types of TV programs are given in a complex ontology. Give a specific example where ontology-based querying would be needed.

11. (2 points) Conclude by giving a detailed proposal (a couple of paragraphs long) on how you would design the entire architecture of your solution for cataloging and archiving TV programs over 50 years: what technologies would you use to store and query the meta-information? for the video information? Justify your choices of technologies.

12. (2 points) Assume the data is entered into the database by TV production companies (for video data and data about episodes and series) and TV channels (for scheduling information). How could you make sure, when issuing a query on the meta-information, to determine what the original source of the data is? Explain in detail how this would work.