DIAS: Data-Intensive Applications and Systems Laboratory
School of Computer and Communication Sciences
Ecole Polytechnique Fédérale de Lausanne
Building BC, Station 14
CH-1015 Lausanne
URL: http://dias.epfl.ch/

Lab 1 Description

CS-300 Data-Intensive Systems
Spring 2025

Table of Contents

Summary

This document describes lab 1.

We provide the ER model, the tables available in the database, as well as the queries that you should write. This document explains the elements we provide as the common starting point for every student. This document should also be considered educational and instructional, and we hope you find it useful.

DIAS: Data-Intensive Applications and Systems Laboratory
School of Computer and Communication Sciences
Ecole Polytechnique Fédérale de Lausanne
Building BC, Station 14
CH-1015 Lausanne
URL: http://dias.epfl.ch/

READ THE DOCUMENT FULLY AND CAREFULLY.

Introduction

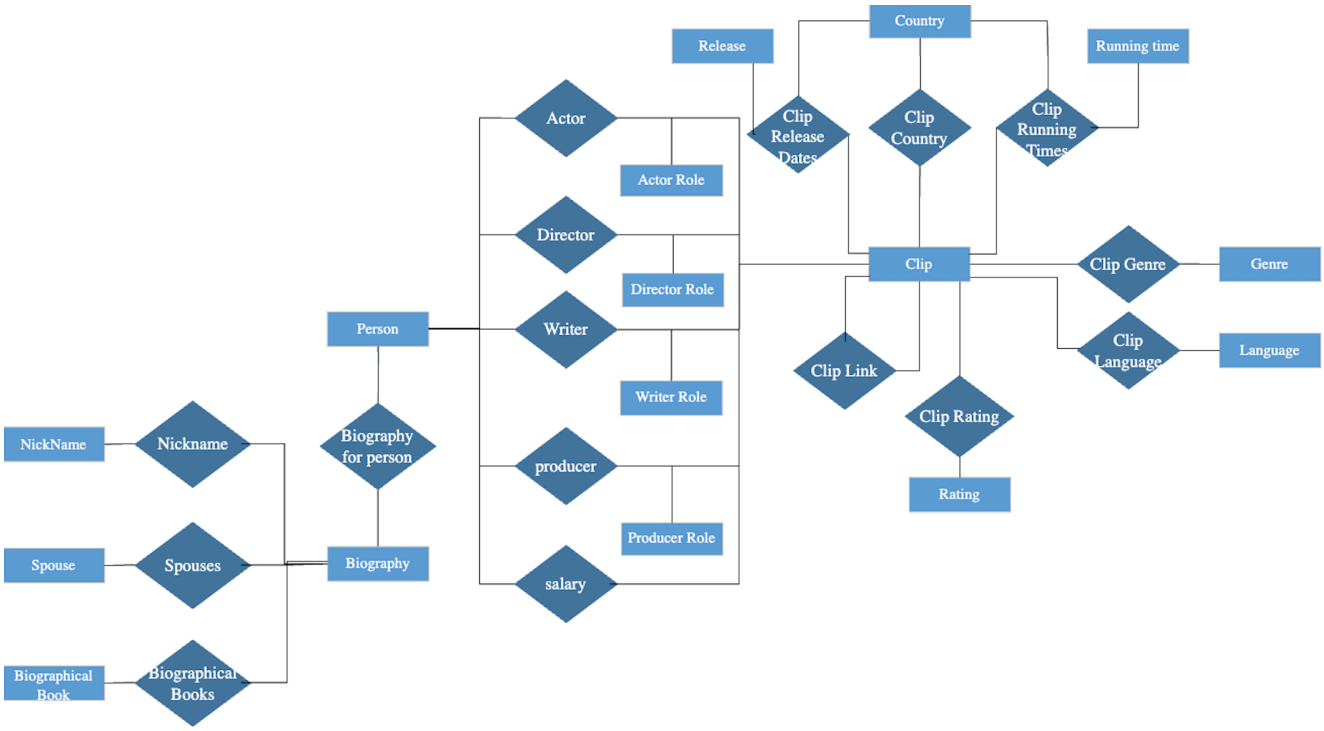
This document presents the ER model, and reasoning for the model simplifications. The purpose of this document is to serve as a starting point for **lab 1** and to be instructive and educative regarding some common practices that we will explain.

In this document, we provide the following:

1. ER model,
2. Tables available in our DBMS,
3. Queries for you to write and run.

NOTE: This is not the only, best, or ideal ER model. Modeling often requires simplifications and adapting to a particular use case of existing entities and relationships that can be captured in a best effort by the data related to some real-world process or informed by domain experts.

ER Model



DIAS: Data-Intensive Applications and Systems Laboratory
School of Computer and Communication Sciences
Ecole Polytechnique Fédérale de Lausanne
Building BC, Station 14
CH-1015 Lausanne
URL: http://dias.epfl.ch/

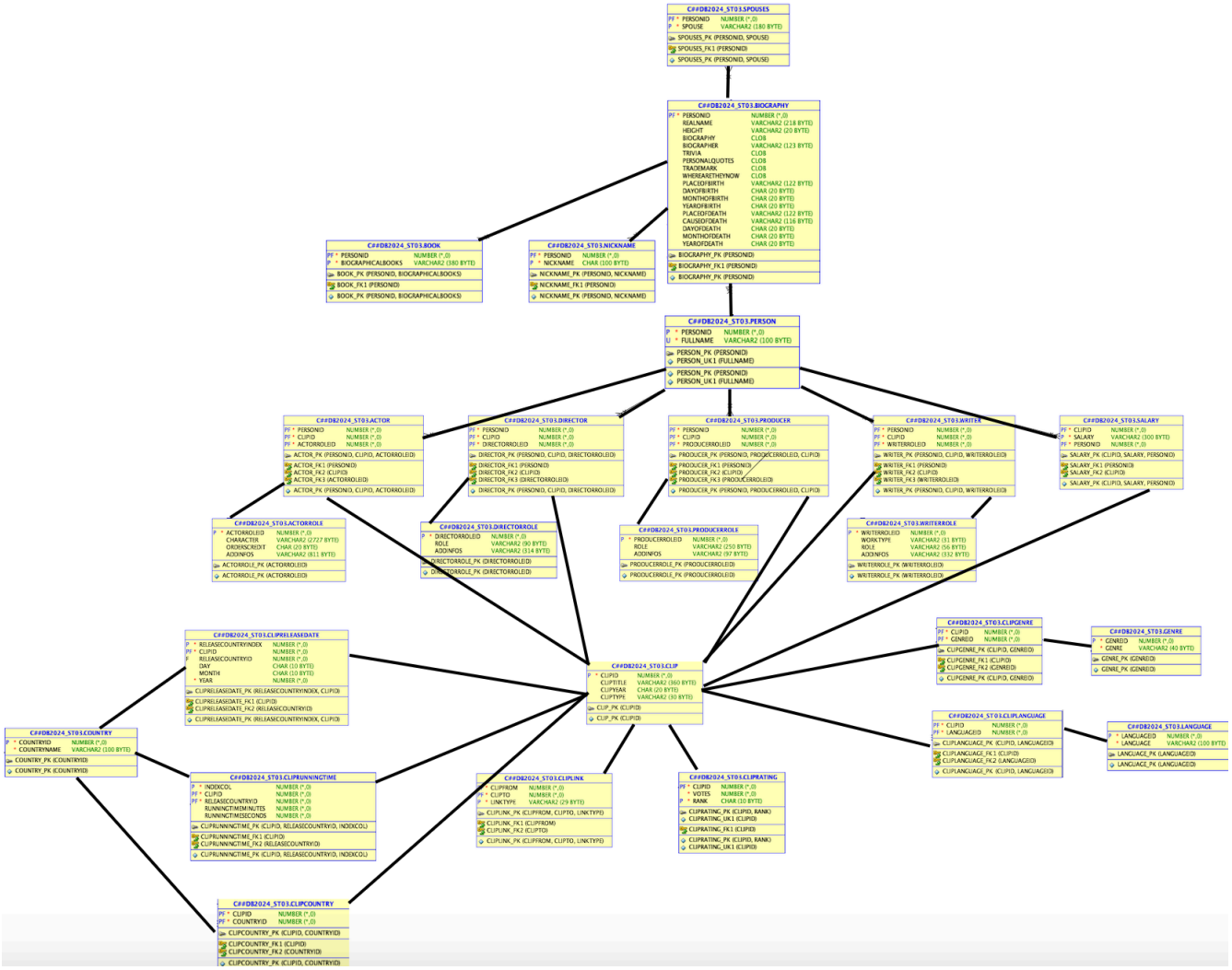
Different relationship types(1-to-1, many-to-many, etc) between entities and constraints are not represented, and weak entities are not marked.

One common design for ER model schema is a snowflake schema – where fact tables (that contain and connect the information) and dimension tables (attributes or sets of attributes grouped as entities further describing the data) are normalized and organized so they resemble a snowflake. The effect of such design is normalization – where data is split up to avoid redundancy but introducing the need for joins. Formally, you can check out normalization and normal forms. 1st normal form is the basis of the relational model: each table column must have a single value, and sets of values or nested records are not allowed.

Database Tables

The tables in the database, containing information about the attributes per table, their primary keys, their constraints and with which tables they are connected, are provided below. **This is not the ER model.**

DIAS: Data-Intensive Applications and Systems Laboratory
School of Computer and Communication Sciences
Ecole Polytechnique Fédérale de Lausanne
Building BC, Station 14
CH-1015 Lausanne
URL: http://dias.epfl.ch/



Database Connection

For connecting, we will provide the username and password - do not forget to use EPFL VPN if not on the EPFL network. The database is stored in the **C##DB2025** schema (i.e., all table names should be prefixed with (C##DB2025.). The **C##DB2025_STUDENT** user that we provide you does not own the tables and only has SELECT grants on the **C##DB2025** schema tables, so you have to explicitly set the **C##DB2025** schema in your SQL IDE to introspect the tables.

Every student has access with the same username and password.

hostname: cs322-db.epfl.ch, **port:** 1521, **SID:** ORCLCDB

username: C##DB2025_STUDENT

password: password: DB2025Lab1

NOTE: You cannot modify the provided tables.

DIAS: Data-Intensive Applications and Systems Laboratory
School of Computer and Communication Sciences
Ecole Polytechnique Fédérale de Lausanne
Building BC, Station 14
CH-1015 Lausanne
URL: http://dias.epfl.ch/

Queries (100 points)

Since the data is already loaded, you can proceed with the queries (10 points per query).

Make sure to follow the query specification carefully with the attributes, order, and number of the results. Write each query in a separate .sql file, for example, 1.sql is the correct file name for the 1st query. You must follow this file naming convention, otherwise your work will not be evaluated. Put the query files in the root of your GitHub Classroom assignment repository for Lab 1. Finally, push your submission using git. Your queries will be automatically graded every time you push updates to your repository.

1. For each FULLNAME, compute the total number of distinct WRITERROLEID values across all clips. Print FULLNAME and the total count of unique WRITERROLEIDs. Order by the total descending and print only the first 10 rows. Output should be in the format FULLNAME, TOTAL_WRITER_ROLES as in the example below (the provided rows are part of the correct answer):

   ```
   FULLNAME              TOTAL_WRITER_ROLES
   --------------------  ------------------------------
   Scott Jeffrey (III)   116
   Simenon Georges       114
   ```

2. Print the top 10 languages that appear in the greatest number of CLIPs based on CLIPLANGUAGE. Print LANGUAGE and the count of CLIPs as CLIP_COUNT. Order by CLIP_COUNT descending. Output should be in the format LANGUAGE, CLIP_COUNT as in the example below (the provided rows are part of the correct answer):

   ```
   LANGUAGE      CLIP_COUNT
   ---------------  -----------------
   ENGLISH       458025
   SPANISH       54049
   ```

3. For each CLIPID, compute the total number of directors who co-directed the clip (i.e., had the 'co-director' role). Print CLIPID and the total count of co-directors. Order by the total descending and print only the first 10 rows. Output should be in the format CLIPID, NUM_CO_DIRECTORS as in the example below (the provided rows are part of the correct answer):

   ```
   CLIPID          NUM_CO_DIRECTORS
   -------------   ------------------------------
   1203403         17
   398165          15
   ```

4. Find the 10 clips with the maximum total running time. Compute total running time in seconds using RUNNINGTIMEMINUTES and RUNNINGTIMESECONDS. Print CLIPID, RELEASECOUNTRYID, and total

running time in seconds. Order by TOTAL_RUNNINGTIME descending. Output should be in the format CLIPID, RELEASECOUNTRYID, TOTAL_RUNNINGTIME as in the example below (the provided rows are part of the correct answer):

CLIPID	RELEASECOUNTRYID	TOTAL_RUNNINGTIME
1796377	202	2678400
440379	69	535800

5. Compute the average running time (in seconds) of clips in each country. Convert minutes to seconds and add RUNNINGTIMESECONDS. Print COUNTRYNAME and the computed average. Round the average to seconds using the ROUND function. Order by COUNTRYNAME ascending and print only the first 10 rows. Output should be in the format COUNTRYNAME, AVERAGE_RUNNINGTIME_SECONDS as in the example below (the provided rows are part of the correct answer):

COUNTRYNAME	AVERAGE_RUNNINGTIME_SECONDS
Albania	2825
Argentina	3518

6. For each PERSONID, compute the total number of distinct ACTORROLEID values across all clips. Print PERSONID and the total count of unique ACTORROLEIDs. Order by the total descending and print only the first 10 rows. Output should be in the format PERSONID, DISTINCT_ACTOR_ROLES as in the example below (the provided rows are part of the correct answer):

PERSONID	DISTINCT_ACTOR_ROLES
210135	998
173938	869

7. For each CLIPID, compute the total number of languages associated with the clip. Print CLIPID and the total count of languages. Order by the total descending and print only the first 10 rows. Output should be in the format CLIPID, NUM_LANGUAGES as in the example below (the provided rows are part of the correct answer):

CLIPID	NUM_LANGUAGES
1457468	20
1457467	19

DIAS: Data-Intensive Applications and Systems Laboratory
School of Computer and Communication Sciences
Ecole Polytechnique Fédérale de Lausanne
Building BC, Station 14
CH-1015 Lausanne
URL: http://dias.epfl.ch/

8. Identify the languages with the largest number of distinct clips featuring at least one actor in a leading role (ACTORROLEID = 1). Print LANGUAGE and the total count of such clips. Order by the total descending and print only the first 10 rows. Output should be in the format LANGUAGE, CLIP_COUNT as in the example below (the provided rows are part of the correct answer):

LANGUAGE	CLIP_COUNT
English	1184
Spanish	65

9. Find all PERSONIDs that appear as both a director and a producer for a clip for at least 2 distinct clips. Print PERSONID only. Order by PERSONID ascending and print only the first 10 rows. Output should be in the format PERSONID as in the example below (the provided rows are part of the correct answer):

PERSONID
2
335

10. For each PERSONID, compute the total number of distinct clips in which they appear in any of the following roles: ACTOR, WRITER, or DIRECTOR. Print PERSONID and the total count of unique clips. Order by the total descending and print only the first 10 rows.Output should be in the format PERSONID, TOTAL_CLIPS as in the example below (the provided rows are part of the correct answer):

PERSONID	TOTAL_CLIPS
164704	3569
1528998	3432

Submission

The deadline for the submission is on Monday 24 March at 9.00 am. Each student needs to create a repository by joining the Github Classroom assignment for Lab 1 and submit 10 SQL files to their repository, one for each query. **If you don't use the above mentioned submission format, your work will not be evaluated. Additionally, make sure to follow the query specifications carefully with the attributes, order, and number of the output rows requested.**

DIAS: Data-Intensive Applications and Systems Laboratory
School of Computer and Communication Sciences
Ecole Polytechnique Fédérale de Lausanne
Building BC, Station 14
CH-1015 Lausanne
URL: http://dias.epfl.ch/

Frequently Asked Questions

Can I collaborate?

The lab assignment is individual. The SQL code you write should be your own and should not be copied from someone else, or the internet. We will run plagiarism checks once the deadline expires to catch any offenders. Any violation of the honesty policy will be met with strict action.

What if I submit the deliverable late?

As soon as the deadline passes, your submission is marked **LATE** and we will not evaluate it.

When can I ask questions about lab 1?

The weekly lab session is the intended place for questions. Otherwise, please use the forum for questions that are of interest to your colleagues too.

GOOD LUCK!