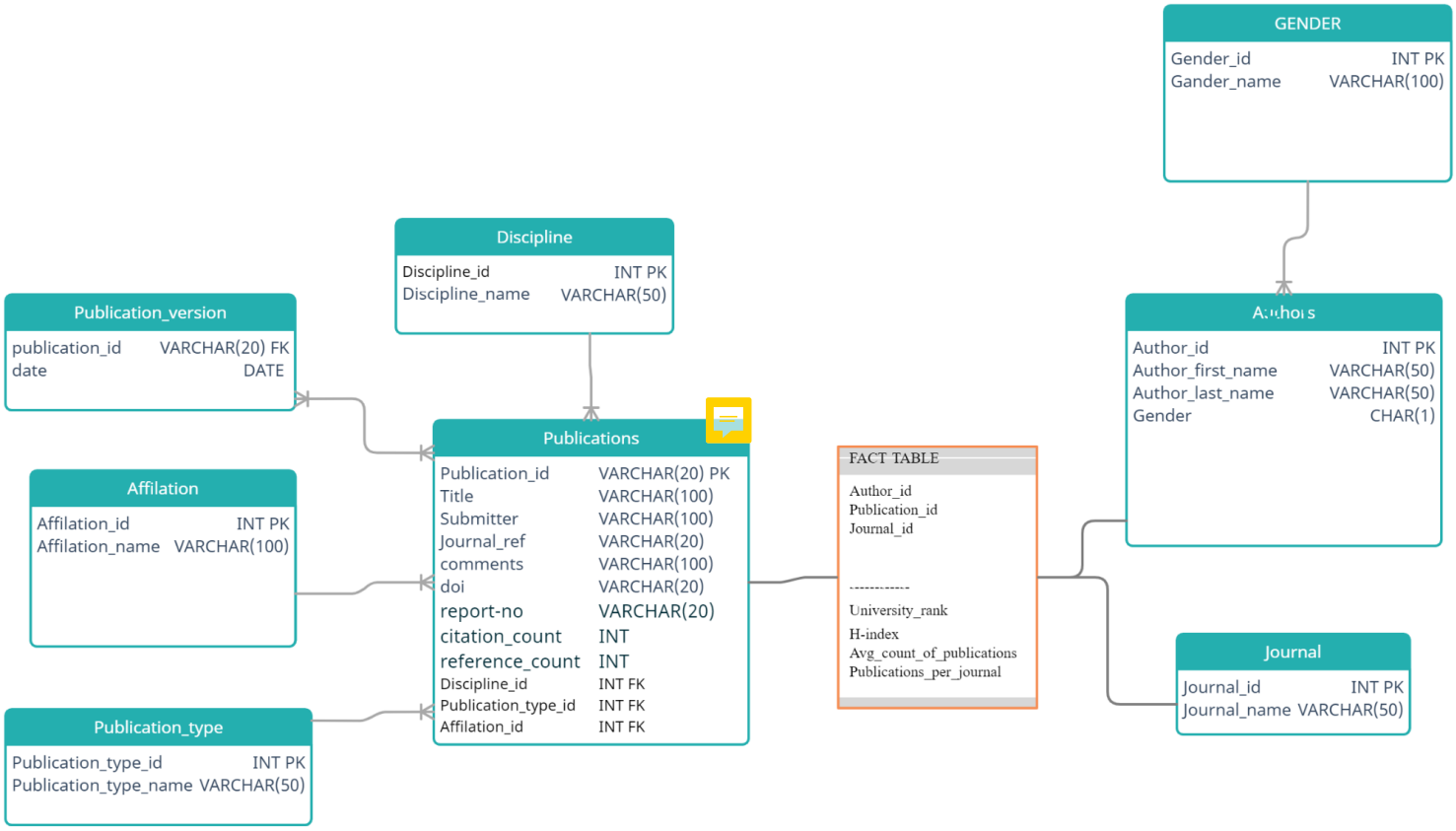
LTAT.02.007
Data Engineering
Group 1
Project design document

1. The Data Warehouse - snowflake schema

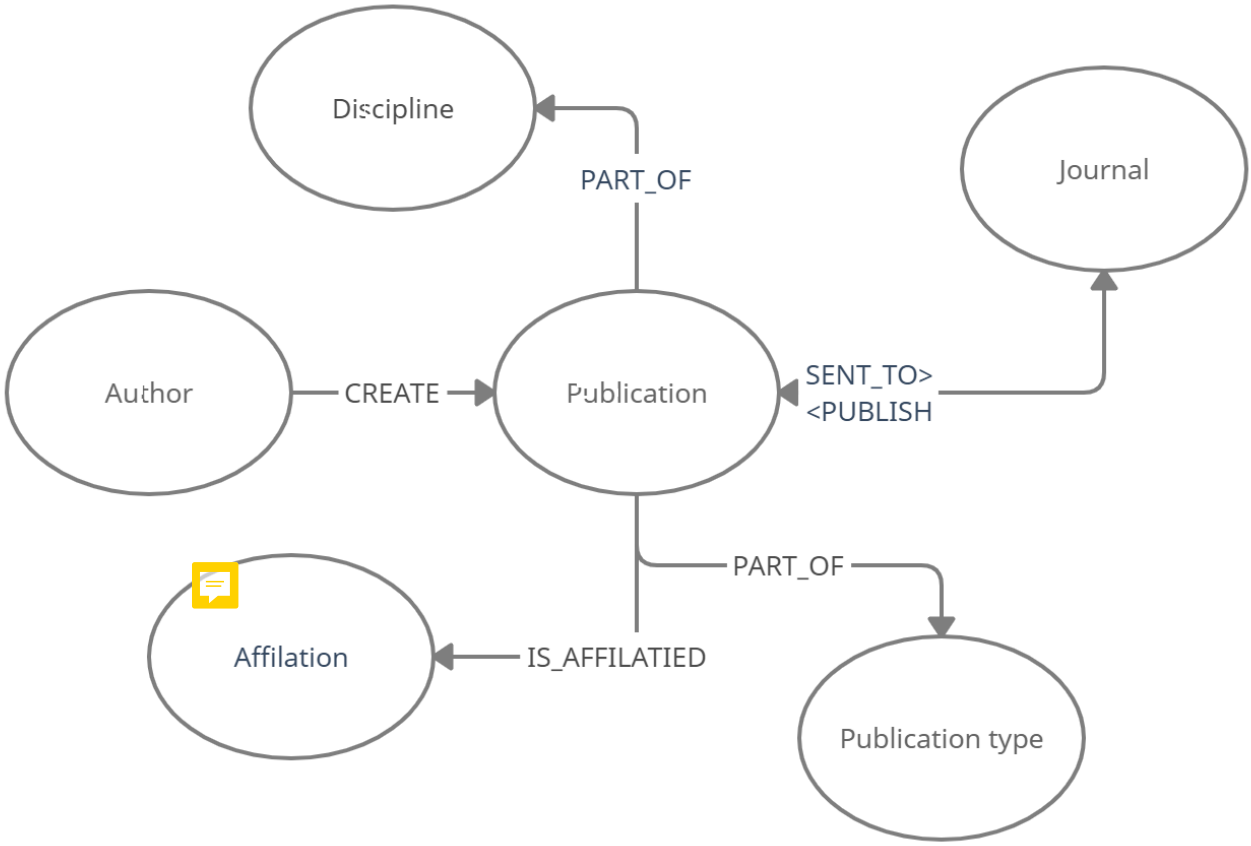


The DWH schema will answer the following BI queries. Any additional query might be executed to find out more insight about the data later on.

BI queries:

1. Rank of authors in the given discipline ~ h-index
2. Popularity of affiliations ~ average_of_publications for each affiliation
3. Popularity of disciplines ~ total_number_of_publications within the discipline
4. Popularity of journal (publication venue) ~ publications_per_journal
5. Gender proportion in specific disciplines ~ rate_of_gender within the discipline

2. Graph view

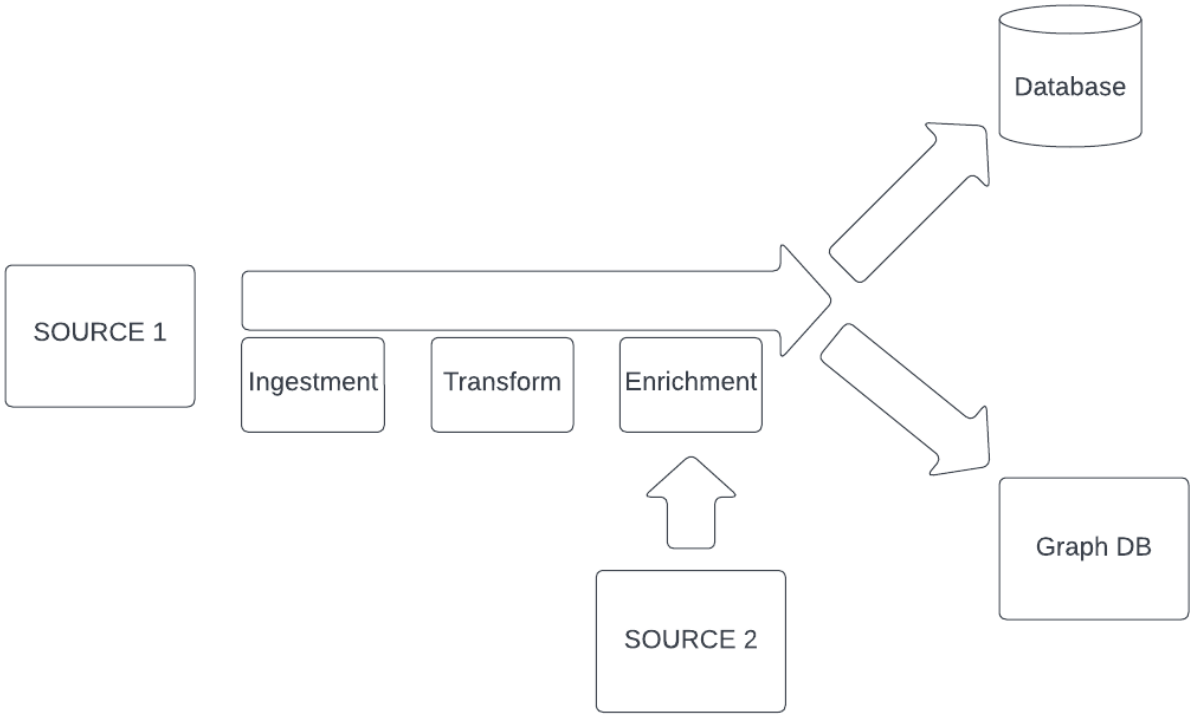


Author: can create Publication, and sent it to Journal
Journal: can receive and publish Publication
Publication: is a part of Discipline and Publication type
Affiliation: sponsors Publication
Publication type: is part of publication and appoints publication type
Discipline: is part of publication and appoints publication discipline

Graph queries:

1. Discovery of co-authors ~ find out the authors who co-authored with same person, so the person can introduce those authors to work on the relative discipline to contribute
2. Discovery of journals which are commonly used by top 5 authors
3. Discovery of relation between the affiliation and journals

3. Pipeline and sources



Ingestment: Importing JSON data from Kaggle (data stored in local/cloud machine) ~ SOURCE 1

Transformation:

- Data cleaning
    - drop publications with very short titles, e.g. one word, with empty authors
    - drop the abstract as it is not required in the scope of this project
    - In the process of transformation we will explore the data thoroughly to find out what we need to do more.
- Data manipulation (taking several key points from one column, to fill into a few columns separately)

Enrichment:

SOURCE 2 represents: any additional source that is mentioned on the following points

- Adding citation and reference count by REST API
- Affiliations will be imported from scholar.py (possibly from other sources as well)
- Journal names scholar.py(possibly from other sources as well)
- Author gender: https://gender-api.com/