CMPU4003 Advanced Databases
Continuous Assessment Project
Initial Dataset Description

Initial Database - Entities and Relationships

A TV production company is interested in analyzing some statistics related to an annual music contest in Ireland. The contest is held on the last Saturday of May each year.
During the last 10 years, the company has collected information about televoting from its viewers. The company now wishes to analyze trends related to the viewers' voting activity over several years. It also wishes to investigate using a relational database, document store or distributed column store to achieve this.
The schema of the database is as follows:

AGEGROUP: Stores information about viewer age groups, such as age group IDs and descriptions (e.g., 18-25, 25-30, 30-50, >50).
COUNTY: Represents counties, with each entry containing a unique county ID and county name (a subset of counties is being used).
EDITION: Represents each annual edition of the competition, storing the year and the presenter's name for that edition.
VIEWERCATEGORY: Categorizes viewers into different voting categories ("Jury" or "Audience").
PARTICIPANTS: Contains details about the participants (performers or bands), including their names and the county they represent.
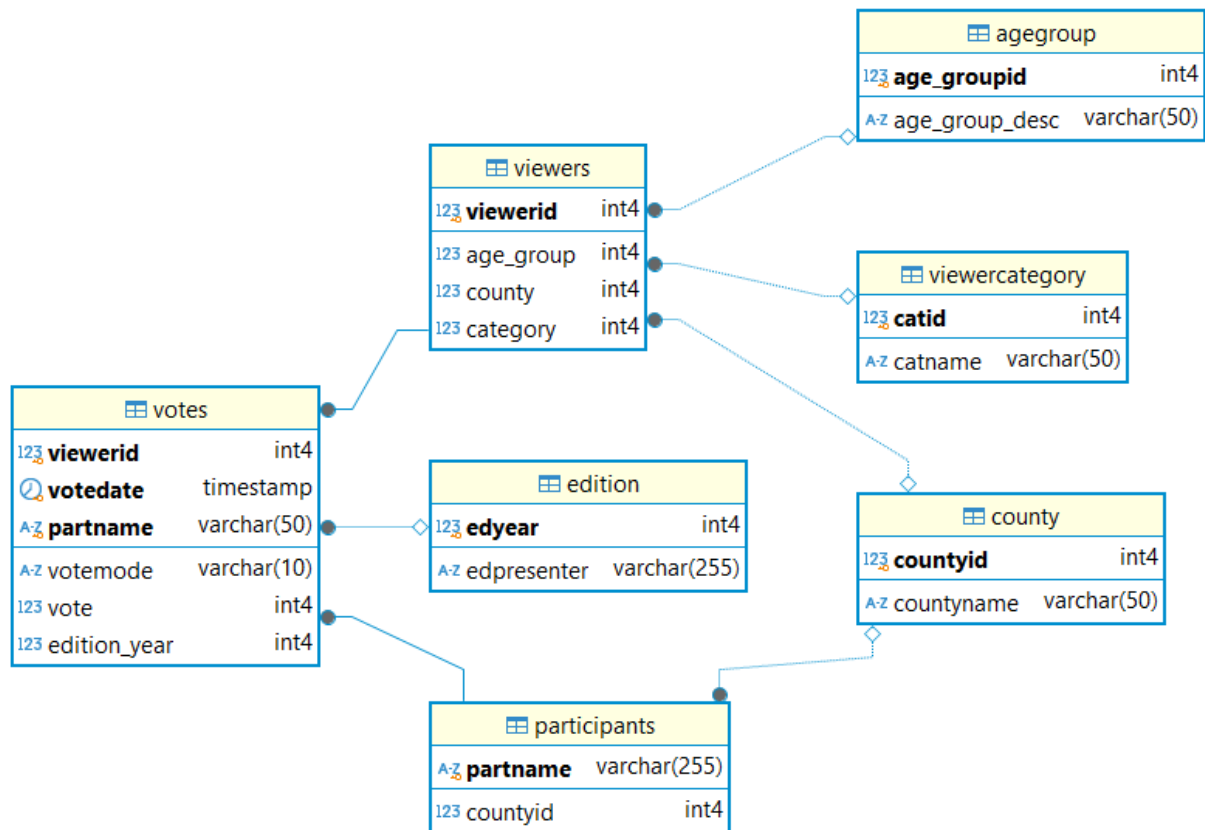**VIEWERS**: Holds information about viewers who participate in the voting process, including their age group and county of residence.
VOTES captures the details of each vote cast by viewers, including:
- The viewer's ID voting.
- The voting mode or the way in which the vote was cast
  - "Phone", "Facebook", "Instagram", "Web" where Web indicates the TV Company's website was used.
- The date when the vote was expressed.
- The vote value (ranging from 1 to 5).
- The edition year – the year of the competition the vote was cast for.
- The name of the participant voted for.
- The table also ensures data integrity by linking to the **VIEWERS**, **VIEWERCATEGORY**, **Edition**, and **PARTICIPANTS** tables through foreign keys.

Each viewer can vote for up to five participants in each edition allocating a vote for each from 1 to 5 but they are not required to use all their votes, they can for example simply vote for one participant giving a vote of 3 if they wish.

Each vote is associated to a specific edition of the TV program (e.g. '2020', '2021', …). The TV program is conducted at the end of May.

agegroup

123	**age_groupid**	int4
A-Z	age_group_desc	varchar(50)

viewers

123	**viewerid**	int4
123	age_group	int4
123	county	int4
123	category	int4

viewercategory

123	**catid**	int4
A-Z	catname	varchar(50)

votes

123	**viewerid**	int4
⊘	**votedate**	timestamp
A-Z	**partname**	varchar(50)
A-Z	votemode	varchar(10)
123	vote	int4
123	edition_year	int4

edition

123	**edyear**	int4
A-Z	edpresenter	varchar(255)

county

123	**countyid**	int4
A-Z	countyname	varchar(50)

participants

A-Z	**partname**	varchar(255)
123	countyid	int4

Partial Data Warehouse Schema

This schema includes a fact table plus a set of dimensions which will help answer the three questions posed below. The purpose of this warehouse is to allow investigation of votes cast, and income earned, for each participant in each edition by age group, jury/audience viewers, vote modes and viewer county.

At this point the design includes a fact working at the level of granularity of the total votes achieved for each participant from each voting category from each voting mode from each age group in each county.

agedimension

123	**age_groupid**	int4
A-Z	age_group_desc	varchar(50)

categorydimension

123	**catid**	int4
A-Z	catname	varchar(50)

votesfact

123	**editionyear**	int4
A-Z	**participantname**	varchar(100)
A-Z	**votemode**	varchar(10)
123	**viewercat**	int4
123	**vieweragegrp**	int4
123	totalvotes	int4

editiondimension

123	**editionyear**	int4
A-Z	presentername	varchar(100)

participantdimension

A-Z	**participantname**	varchar(100)
A-Z	countyname	varchar(100)