DATA 514 - Final Project - Mock Doc

Project Overview

This project involves designing a database schema and writing queries for the BDD100K dataset, a large-scale diverse driving video database. Our focus will be on a subset of the data to create a manageable project scope.

Dataset Description

BDD100K

- **Total Size**: 1.8TB of video clips, 100K videos.
- **Annotations**: JSON files for object detection, drivable area, lane markings, and segmentation tasks.
- **Key Features**:
  - Videos from diverse locations and conditions.
  - Rich annotations including bounding boxes, segmentation masks, and GPS trajectories.

Data Governance Issues

- **Sensitive Information**: GPS data and video footage.
- **Protection Measures**: Anonymization of personal data, secure storage, and restricted access.
- **Future Concerns**: Compliance with data privacy laws and ethical use of autonomous driving data.

Schema Design

Entities:

- **Videos**: `video_id`, `location`, `weather`, `time_of_day`, `duration`.
- **Annotations**: `annotation_id`, `video_id`, `type`, `data`.
- **Objects**: `object_id`, `annotation_id`, `category`, `bounding_box`.
- **Trajectories**: `trajectory_id`, `video_id`, `gps_data`.

ER Diagram: [Insert ER Diagram]

Example Use Cases & Queries

1. **Object Detection Frequency**:
   - **Description**: Count the number of each object category in the dataset.
   - **Query**:

```
SELECT category, COUNT(*) AS frequency
FROM Objects
GROUP BY category;
```

2. **Video Metadata Retrieval**:

- **Description**: Retrieve metadata for videos recorded in rainy weather.
- **Query**:

```
SELECT video_id, location, time_of_day
FROM Videos
WHERE weather = 'rainy';
```

3. **Annotations by Type**:

- **Description**: List all annotations of a specific type (e.g., lane markings).
- **Query**:

```
SELECT annotation_id, video_id, data
FROM Annotations
WHERE type = 'lane_marking';
```

Query Logical Plan Analysis

- **Naive RA Tree**: [Insert RA Tree for one of the queries]
- **Optimized RA Tree**: [Insert optimized RA Tree]
- **Performance Differences**: The optimized RA tree reduces the number of joins and filters data earlier in the process, leading to improved performance.

Indexes

- **Video Metadata Index**: Index on `weather` column in `Videos` table for faster retrieval.
- **Object Category Index**: Index on `category` column in `Objects` table to speed up frequency queries.

DBMS Selection

Recommended DBMS: PostgreSQL

- **Considerations**:
  - **Performance**: Supports large datasets and complex queries efficiently.
  - **Features**: Advanced indexing and support for JSON data.
  - **Scalability**: Capable of handling the scale of BDD100K.