Université de Rennes

Big Data - Caching

Author

- Laurent d'Orazio, Univ Rennes, CNRS, IRISA

1 Introduction

The application to be considered relies on a collection of movies. A sample of the data sets is illustrated by figure 1. The attributes of a movie are two identifiers (id and idimdb), the release data, the runtime, the budget and revenues, a link to a poster, an overview and a rating.

idimdb	title	release	runtime	budget	revenue	poster	overview	rating	last_update
tt0029583	Snow White and the Seven Dwarfs	1937-12-21	83	1488423	184925486	https://image.tmdb.org/t/p/original/1P9eGGlT7eV7kA...	A beautiful girl, Snow White, takes refuge in the ...	7.6	2022-02-27
tt0032910	Pinocchio	1940-02-23	88	2600000	84300000	https://image.tmdb.org/t/p/original/ynkEEi296ofs9k...	Lonely toymaker Geppetto has his wishes answered w...	7.4	2022-02-27
tt0033563	Dumbo	1941-10-31	64	812000	1600000	https://image.tmdb.org/t/p/original/hKDdIlsIMtsU9J...	Dumbo is a baby elephant born with over-sized ears...	7.2	2022-02-27
tt0034492	Bambi	1942-08-14	70	858000	267447150	https://image.tmdb.org/t/p/original/wV9e2y4myJ4KMF...	Bambi's tale unfolds from season to season as the ...	7.3	2022-02-27
tt0042332	Cinderella	1950-02-22	74	2900000	263591415	https://image.tmdb.org/t/p/original/4nssBcQUBadCTB...	Cinderella has faith her dreams of a better life w...	7.3	2022-02-27
tt0046183	Peter Pan	1953-02-05	77	4000000	87404651	https://image.tmdb.org/t/p/original/fJJOs1iyrhKfZc...	Leaving the safety of their nursery behind, Wendy,...	7.3	2022-07-13
tt0048280	Lady and the Tramp	1955-06-22	76	4000000	36359037	https://image.tmdb.org/t/p/original/wXKeGea4htM7Yz...	Lady, a golden cocker spaniel, meets up with a mon...	7.3	2022-10-18
tt0053285	Sleeping Beauty	1959-02-17	75	6000000	51600000	https://image.tmdb.org/t/p/original/j2mTyUukcLwDIe...	A beautiful princess born in a faraway kingdom is ...	7.2	2022-02-27
tt0055254	One Hundred and One Dalmatians	1961-01-25	79	4000000	85000000	https://image.tmdb.org/t/p/original/9klDisS1sVb5LI...	When a litter of dalmatian puppies are abducted by...	7.3	2022-02-27
tt0057546	The Sword in the Stone	1963-12-25	79	3000000	22182353	https://image.tmdb.org/t/p/original/7IyeeuhGAJSNXY...	Wart is a young boy who aspires to be a knight's s...	7.2	2022-02-27
tt0058331	Mary Poppins	1964-08-27	139	6000000	103082380	https://image.tmdb.org/t/p/original/yO8UyO86uNBq12...	A magical nanny employs music and adventure to hel...	7.8	2022-02-27

Figura 1: Sample

2 Basic concepts: data caching, hit, miss, replacement

Data caching consists in storing data temporarily in a cache, removing it automatically after a given period of time (notion of Time To Live TTL) or relying on a replacement policy when the cache is full. Caching is used in distributed systems, particularly web applications in order to improve performances. Indeed, on one hand caching enables data to be closer to the application or users, on the other hand it leads to reduce the number of queries to the database/storage system.

3 Exercises

3.1 Cache hits and misses

A Web (PHP for instance) application is used to store and retrieve movies using their IMDB Identifiers. To improve performance, a caching mechanism is deployed using Redis. To start with the cache is cold (there is no data within the cache).

The size of the cache is currently considered as unlimited.

Users are submitting the following queries: tt0325980, tt1477834, tt0325980, tt1477834, tt0325980, tt1477834, tt0325980.

1. What is the number of cache hits?
2. What is the number of cache misses?
3. What is the content of the cache at the end?

3.2 Replacement

The cache system uses a LRU (Least Recently Used) replacement (or eviction) policy and is used to consider the sequence of movie requests: tt0468569 tt0108052 tt0068646 tt0111161 tt0468569 tt0108052 tt0110912 tt0468569 tt0108052 tt0068646 tt0111161 tt0110912

1. What is the number of cache hits/misses with a cache of size 3?
2. What is the number of cache hits/misses with a cache of size 4?

The cache system now uses a FIFO (First In First Out) replacement policy.

1. What is the number of cache hits/misses with a cache of size 3?
2. What is the number of cache hits/misses with a cache of size 4?

4 Query caching

Several alternatives exist to store (SQL) queries results in a cache (for example in Redis), depending on the application's needs and the data structure to be used:

1. Full result caching: Entire result of the query in the cache using a unique key (for example the query as a string) to identify the result. Such an approach allows to retrieve the data directly from the cache (for example a key-value store) instead of executing the query every time.
2. Individual result caching: Each individual result (or record) using unique keys. For example, a cache's hashes can be used to store each row of the query, the key being a unique identifier of the record and the fields corresponding to the table columns.
3. Pagination caching: If the result is large, the cache can be used to store paginated data. That is to say dividing the results into pages and then storing the pages with a unique key per page. As a consequence, subsets of data can be quickly accessed without considering the whole result (and limiting the overhead of multiple small reads).

5 Exercise

5.1 Full result caching / Query caching

Let the query `SELECT * FROM MOVIE WHERE RUNTIME=143` associated to the following records:

```
tt0325980,Pirates of the Caribbean: The Curse of the Black Pearl,2003-07-09,143,140000000,655011224,htt
tt1477834,Aquaman,2018-07-06,143,160000000,1148461807,https://image.tmdb.org/t/p/original/xLPffWMhMj1l5
```

1. What is the cache entry if the result is stored as a string in the CSV format?
2. What is the cache entry if the result is stored as a string in the JSON format?

5.2 Individual result caching / Tuple caching

Let the 83 minute movie 'Snow White and the Seven Dwarfs', released on 1937-12-21, associated with the id tt0029583. Its budget was 1488423$, its revenue is 184925486 $ and its rating is 7.6. Its cover can be seen at https://image.tmdb.org/t/p/original/1P9eGGlT7eV7kAAvvSh9jfygr1C.jpg.

1. What is the cache entry if the result of 'Snow White and the Seven Dwarfs' as stored as a CSV string?

6 References

- **Lecture**: https://perso.univ-rennes1.fr/laurent.dorazio/data/teachings/big_data/big_data_redis.pdf