parcial1solucions2022.pdf

alucero

Desenvolupament d'Aplicacions de Dades Massives

3º Grado en Ingeniería de Datos

Escuela de Ingeniería
Universidad Autónoma de Barcelona

1	2	3	4	5	6	7	8	9	10	Nota total

GEdD Ingeniería de Datos-UAB 30 de marzo de 2022

Aplicaciones masivas de datos **Teoria**

Apellidos:
Nombre: DNI o NIU:

1.- [1 punto] ¿Cuales son los principales requerimientos que tienen que cumplir los sistemas intensivos de datos? ¿Puedes definirlos brevemente?

Reliability: system should work correctly when errors happen
Scalability: how to deal with growth (data, traffic, complexity)
Maintainability: maintaining and improving behaviour should be a productive task

2.- [1 punto] Citar tres diferencias relevantes entre un sistema OLTP y uno OLAP

Property	Transaction processing systems OLTP	Analytic systems OLAP
Main read pattern	Small number of records per query, fetched by key	Aggregate over large number of records
Main write pattern	random-access, low latency writes from user input	Bulk import (ETL) or event stream
Primarily used by	End user/customer, via client application (web)	Internal analyst, for decision support
What data represents	Latest state of data	History of events that happened over time
Dataset size	Gigabytes to Terabytes	Terabytes to Petabytes

3.- [1 punto] Define el uso de una base de datos de forma "schema on read". Explica brevemente por qué es necesario este uso y pon un ejemplo de datos que necesiten ser procesados de esta forma

schema-on-read: the structure of the data is implicit, and only interpreted when the data is read, in contrast with schema-on-write: the traditional approach of relational databases, where the schema is explicit and the database ensures all written data conforms to it.

• Data structure is implicit
• Similar to Python's dynamic types
• Maximum flexibility

The schema-on-read approach is advantageous if the items in the collection don't all have the same structure for some reason (i.e., the data is heterogeneous)—for example, because:

- There are many different types of objects, and it is not practical to put each type of object in its own table
- The structure of the data is determined by external systems over which you have no control and which may change at any time.

4.- [1 punto] A la hora de explorar un perfil de un usuario en Linkedin. ¿Para qué es útil una BD relacional? ¿Cuándo nos da ventajas tener toda la información de un perfil en un único JSON?

Relational database: data has some entities and need to analyse relationships

- most queries are many-to-many
- Information structure will not change too much in the future
- Not a large amount of information changing in time
- Not thousands of entities

JSON document: data is mostly documents

- one-to-many relationship tree
- Load the tree just once
- Not many reference outside the tree

5.-[1 punto] ¿Qué operación realiza el algoritmo merge-sort en un indice SST? ¿Para qué sirve esta operación?

SST data segments require that the sequence of key-value pairs is sorted by key. We call this format Sorted String Table, or SSTable for short. We also require that each key only appears once within each merged segment file

The approach is like the one used in the mergesort algorithm: you start reading the input files side by side, look at the first key in each file, copy the lowest key (according to the sort order) to the output file, and repeat. This produces a new merged segment file, also sorted by key

Merging segments is simple and efficient, even if the files are bigger than the available memory. It allows the reduction of space and quick search for key-values.

6.- [1 punto] ¿Por qué no se actualizan los valores de las claves al usar índices hash?

why don't update the file in place, overwriting the old values with the new values?

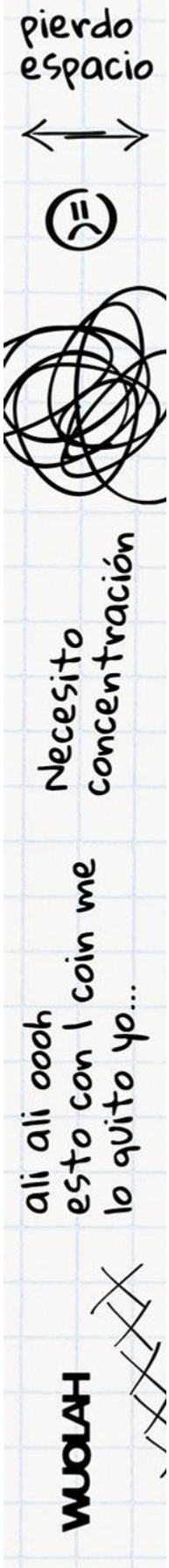
An append-only design turns out to be good for several reasons:

- Appending and segment merging are sequential write operations, which are generally much faster than random writes
- Concurrency and crash recovery are much simpler if segment files are append-only or immutable. For example, you don't have to worry about the case where a crash happened while a value was being overwritten, leaving you with a file containing part of the old and part of the new value spliced together.
- Merging old segments avoids the problem of data files getting fragmented over time.

7.- [1 punto] En un esquema de datos OLAP en estrella, ¿qué elementos encontramos en el centro? Cómo se usa el esquema en estrella para responder a consultas?
Center of the schema: fact table. Each element of the fact table represents an event of a particular time. Facts are captured as individual events

Example: fact_sales. Fact element: customer's purchase of a product.

Star-schema allows different types of analysis:
Fact table can become very large: petabytes of transaction history
Each column in fact table makes reference to other tables
Each dimension represents a way of analysing events history:
who
what
where
when



8.- [1 punto] Si hay que implementar en Redis un top 10 de canciones mas escuchadas, ¿qué estructura de datos usarías? Pon un ejemplo dibujando los campos que usarías y una operación de lectura y otra de escritura en tu ejemplo.

```
ZSET most-listened
<song-id, number of times it has been played>

ZADD most-listened 500 song-1234 99 song-0011

ZREVRANGE most-listened 0 9
```

9.- [1 punto] ¿Por qué conviene separar un data warehouse de los sistemas OLTP?

Less use of OLTP systems for analytics

Better: run analytics on a separate database

Data warehouse:

- NOT affect main business database
- Dump data to a new analytics data base for analysis
- Ready to receive data intensive questions

10.- [1 punto] ¿Cómo funciona la operación de escritura en un LSM-tree? Explica una posible implementación con la operacion write(doily, 42)

When a write comes in, add it to an in-memory balanced tree data structure. This in-memory tree is sometimes called a memtable.

When the memtable gets bigger than some threshold—typically a few megabytes —write it out to disk as an SSTable file. This can be done efficiently because the tree already maintains the key-value pairs sorted by key. The new SSTable file becomes the most recent segment of the database. While the SSTable is being written out to disk, writes can continue to a new memtable instance.