

# parcial1soluciones2021.pdf



alucero



Desenvolupament d'Aplicacions de Dades Massives



3º Grado en Ingeniería de Datos



Escuela de Ingeniería  
Universidad Autónoma de Barcelona

antes



**Descarga sin publi  
con 1 coin**



Después

**WUOLAH**





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

GEdD

Ingeniería de Datos-UAB

31 de marzo de 2021

**Aplicaciones masivas de datos****Teoría**

Apellidos:

Nombre:

DNI:

**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]** ¿Cuales son las métricas principales para definir el rendimiento de una aplicación? ¿Puedes definir las brevemente?

Throughput:

- number of records to process per second • time to process a job on a dataset

Response time:

- Time between a client sending a request and receiving a response

Latency:

- Duration that a request is waiting to be handled (request is latent waiting for service)

**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: Data structure is interpreted when reading

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

**4.- [1 punto]** ¿En qué consiste y por qué es necesaria la operación de compactar bloques de datos en un índice?

Block compact operation

- Eliminate duplicated keys and the end of log file
- Keep only the last key-value pair

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

SST: each block keeps (key,value) pairs sorted by key

- Each key has only one instance at each block
- Now we need a new kind of compact operation: merge and sort

Copy the smallest key to the output file and repeat

6.- [1 punto] ¿Cómo funciona la operación de escritura en un LSM-tree? Explica una posible implementación con la operación write(daily, 42)

Write new value: (daily, 42)

- add new key-value pair only to memtable

When memtable becomes larger than a limit size (few MBytes)

- Write memtable to disk becoming a new SSTable: flush operation
- new SSTable is most recent DB block

7.-[1 punto] ¿Puedes citar tres principales diferencias 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

8.- [1 punto] ¿Qué es una ETL y para qué se utiliza en un data warehouse?

Extract-Transform-Load (ETL)

- How to build large datasets for analytics?
- Read-only copy of data in different OLTP systems in a company
- Main operations to apply to datasets:
  - Extract input from OLTP databases
  - Transform data into an analysis-friendly schema
  - Clean-up and filter our errors
  - Load data into the warehouse

**9.- [1 punto]** ¿Cuales son los principales problemas a resolver en los sistemas de bases de datos en memoria?

Durability: what to do when system shutdowns?

Capacity: what to do when data needs are larger than memory space?

**10.- [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