examen-parcial-1-solucio-TE06042...

annahidalgo

Desenvolupament d'Aplicacions de Dades Massives

3º Grado en Ingeniería de Datos

Escuela de Ingeniería
Universidad Autónoma de Barcelona

EdD Enginyeria de Dades-UAB 7 abril 2021

Desenvolupament d'aplicacions massives de dades **1er Control**

Cognoms:
Nom: DNI/NIE:

1. [1,5 punts] Quines son les tres principals característiques que s'han de gestionar quan es treballa amb sistemes de processament de dades? Pots donar una definició breu de cadascun?

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

Reliability/Fiabilitat: continuar funcionant correctament inclús quan tenim algun problema
Escalabilitat/scalability: com gestionar el creixement de les dades, del tràfic, de la complexitat
Manteniment/Maintanability: mantenir i millorar el comportament d'un sistema hauria de ser una tasca realitzable/productiva

2. [1,5 punts] Què significa el concepte fan-out quan estem analitzant l'activitat d'un usuari a Twitter? Quan tenim N usuaris i M seguidors per a cada usuari, quin és el principal problema que hem de gestionar i quines solucions es poden aplicar?

Fan-out: un usuari segueix a N altres usuaris i cada usuari és seguit per M altres usuaris. Cada missatge publicat te un potencial de N*M número de missatges a gestionar

Pincipal problema: creixement quadràtic de missatges amb un número limitat de recursos (capacitat d'emmagatzematge, temps de processament). Hem d'evitar haver de processar N*M missatges per construir la visió dels missatges dels usuaris (timeline quèries).

La solució principal passa per aprofitar el fet de que la ràtio de missatges publicats és 100 vegades menor que la ràtio de lectura del home timeline. Per tan podem construir una cache amb el timeline de cada usuari quan van arribar els missatges nous. Quan l'usuari vol llegir el seu timeline ja està preparat.

3. [0,5 punts] Quin tipus d'esquema (schema on read/schema on write) s'adapta millor a aquests tipus de dades?

- JSON google maps: schema on read
- Compte corrent bancària: schema on write
- Missatge de twitter: schema on read
- Llistat d'alumnes matriculats a assignatura: schema on write

4. [0,5 punts] Tria les característiques de cada tipus de llenguatge de consultes (imperatiu/declaratiu)

- Especifica un ordre particular de realitzar les operacions: imperatiu
- Oculta els detalls d'implementació: declaratiu
- No es defineix cap ordre en les operacions a realitzar: declaratiu

5- [1 punt] En una base de dades simple on inserim registres al final d'un fitxer, quin és el principal problema de rendiment i per què?

Escrivim elements clau,valor: (1234, Barcelona) al final i llegim valors a partir d'una clau: db_get(1234)

La funció db_get(key) te un baix rendiment si la BD te molts registres, ja que el cost de la lectura depèn del número de registres: O(n)

6- [1 punt] Quin és l'objectiu que es persegueix en la elaboració d'un Data Warehouse?

Business needs:
-Provide required information to know so that end users can take informed business decisions
-Indication of what sources of information the ETL process will have
-Reveal information to the end users that may be of their interest

7- [1 punt] Si en un projecte s'ha d'emmagatzemar 4TB de dades que s'han de guardar fora de les oficines durant 5 anys. En el cas que l'organització necessiti accedir a les dades aquestes han de estar disponibles en menys de 8 hores. Quins son els requeriments de durabilitat i de disponibilitat? Quin servei d'emmagatzemament AWS seria el més recomanable i per què?

Durabilitat: percentatge de pèrdua de dades (fitxers o objectes) durant un any
Disponibilitat: percentatge de temps en el que un objecte està disponible per a ser llegit

En aquest cas, volem durabilitat màxima (no volem perdre cap valor), per tant serà interessant recomanar el servei S3, que te la durabilitat més alta de tots els serveis de dades

Respecte a la disponibilitat, com que no cal que sigui immediata, és acceptable que necessitem fins a 8 hores per tenir les dades llestes per ser llegides podem triar el servei Glacier dintre de les opcions de S3.

8- [1 punt] Per què la arquitectura serverless encaixa molt bé amb l'arquitectura de pipelines dels fluxes de dades?

Per que el concepte principal és el mateix: implementar un fluxe de dades des de l'origen de les dades a la seva destinació final amb una sèrie de etapes al mig on es van implementant certes tasques de millora com filtres i processament de la informació.

9- [1 punt] Quin és el principal mecanisme per a facilitar l'escalabilitat horitzontal de les bases de dades al cloud? En que consisteix? Quines son les seves principals avantatges?

El principal mecanisme és la replicació: mantenir una còpia de les mateixes dades en múltiples màquines connectades en xarxa

Principals avantatges:
Mantenir les dades a prop dels usuaris: menys latència
El sistema pot funcionar si algun component falla: disponibilitat
Incrementem el número de màquines que poden donar servei a les peticions de lectura: millor throughput de lectura

10- [1 punt] Què és la consistència eventual? En quin moment i amb quines condicions es produeix?

Si tenim un sistema líder i diverses rèpliques ens pot passar que alguns followers asíncrons vagin actualitzant els canvis que venen del líder de forma lenta i continguin dades no actualitzades: ***replication lag***. Si llegim les mateixes dades, no tindrem el mateix resultat al líder o a la rèplica. Tenim un problema de consistència global del nostre sistema.

La consistència eventual gestiona aquest estat de manca de consistència de dades del sistema actualitzant les rèpliques de forma retardada en el temps. Les rèpliques contindran les dades correctes en el curs d'un petit interval de temps i les aplicacions distribuïdes han de saber treballar amb aquest interval de temps.