Jeonghyeon Woo

CPTS 415 Big Data

HW 4

1.

a. The four types of noSQL are: key-value, document, big tables, and graph.

   According to Wikipedia, Oracle nosql database is a type of key-value model, Azure tables is a type of wide column store, Arango DB is a type of document model, and Microsoft SQL server is a type of graph db.

b. Key-value will be good at login system since it can quickly match the key(id, password) with values. However, it will not really good at systems that demand difficult relations between data such as search engine.

   Big table will be good at storing user profile. Big table can give every attributes of the key, and some keys may have more, less, and different attributes. However, it will not good at what key-value is good at.

   Document will be good at web programs like the airport example given in hw1. It can connect schemas with primary keys and foreign keys. However, it will be not good when ACID is important.

   Graph will be good at searching engines as it can handle difficult relations. However, it will not as fast as the other models.

c. Ex. User: fruit, juice, song

   Key value returns: <woo (primary key), lemon> <woo (primary key), sprite>

   Big table returns: <woo, fruit: lemon, song: jazz>, <j, fruit: lemon, juice: sprite>

   Document returns: <woo, lemon>; <lemon; sour, sprite>. <woo, sprite>

d. Graph returns: woo→lemon→sour, woo→sprite→sour, woo→jazz

   So, the key has child node of its attributes.

2.

a. Cap is about consistency, availability, and partition tolerance.

   For example, databases that store the data of toy will have cap when:

   toy1 is out in north mart(s1) and north mart(s1) notice south mart(s2) that toy1 is out, so the customers of both marts can know that toy1 is out(consistence).

   both customers at south mart(s2) and north mart(s1) are able to see what toys are available (available).

   toy2 is moved from north mart(s1) to south mart(s2), but both marts know they

have toy2(partition tolerance).

b. ACID is about atomicity, consistency, isolation, and durability.

For example, 2 accounts have ACID when:

there is no transaction, the account balance is not changed (durability)

transaction is in process, atm can not access both accounts at the same time (isolation).

transaction from 123 is failed, balance of both accounts is not changed (atomicity).

123 sends 200 dollars and 456 gets 200 dollars (consistency).

Durability is violated when account balance is changed without transaction. Isolation is violated when atm process transfer of 100 dollar from 789 to 456 while transaction between 123 and 456 is happening. Atomicity is violated when balance of 123 is reduced by 200, but balance of 456 is still the same. Consistency is violated when the 123 sends 200 dollars and 456 gets 250 dollars (total balance in the bank is changed).

c. BASE is basically available, soft state, and eventually consistent. BASE has weak consistency. It will eventually have consistency anyways, but it does not have consistency all the times. It is focused on availability, so it has fast access to the data. However, data may be expired if it is not modified. ACID is like meet a teller to check your account, and BASE is like asking your mom how much money you have in your account (Mom will say, “I think it is 5 to 7 bucks”, and you will check your account and realize that is true).

3.

a. Column store stores the data of same attribute in the same page. Operators of column store is similar to row store’s, but it needs different iterations to make operation. While row store iterates all attributes, column store iterates only selected attributes. However, its query processing needs access to the other attribute pages. For example, when key is sorted, the value must be also sorted, but column store can not sort the value without accessing the pages storing values.

b. Quorum is a concept that is for keeping CAP of distributed system. Suppose there are n copies of data, put() gets w copies of the data, and get() gets responses from r copies of the data. If w+r is bigger than n, which means it is sure that there is at least one copy of data that goes both put() and get(), then the consensus is made, so it is safe to say the data is consistent.