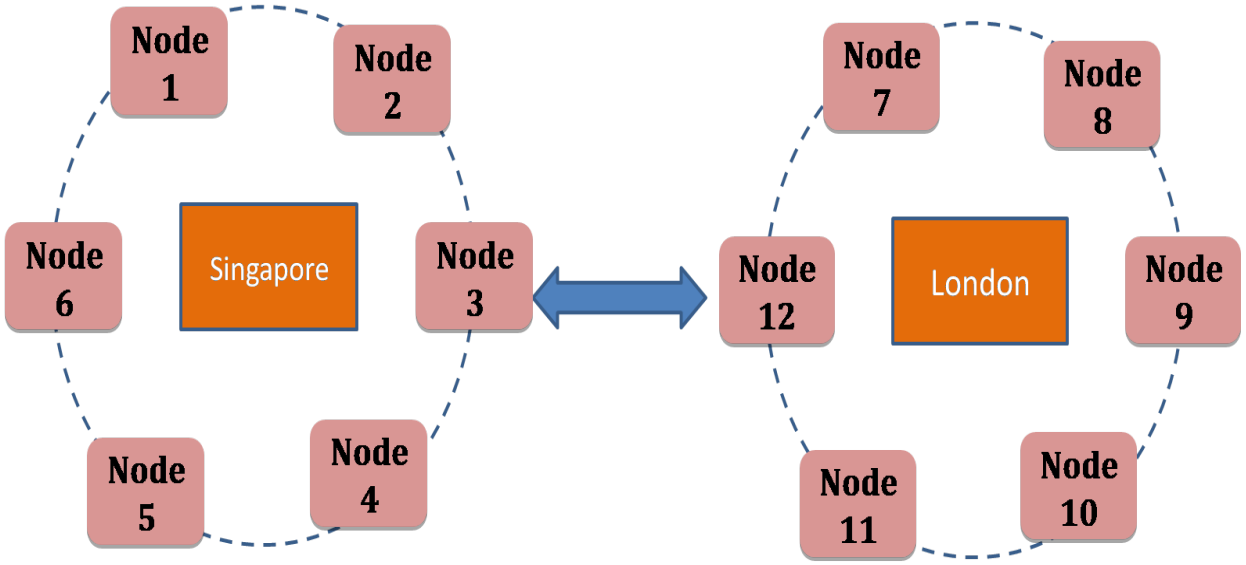
Reflection
and Sample Project Training

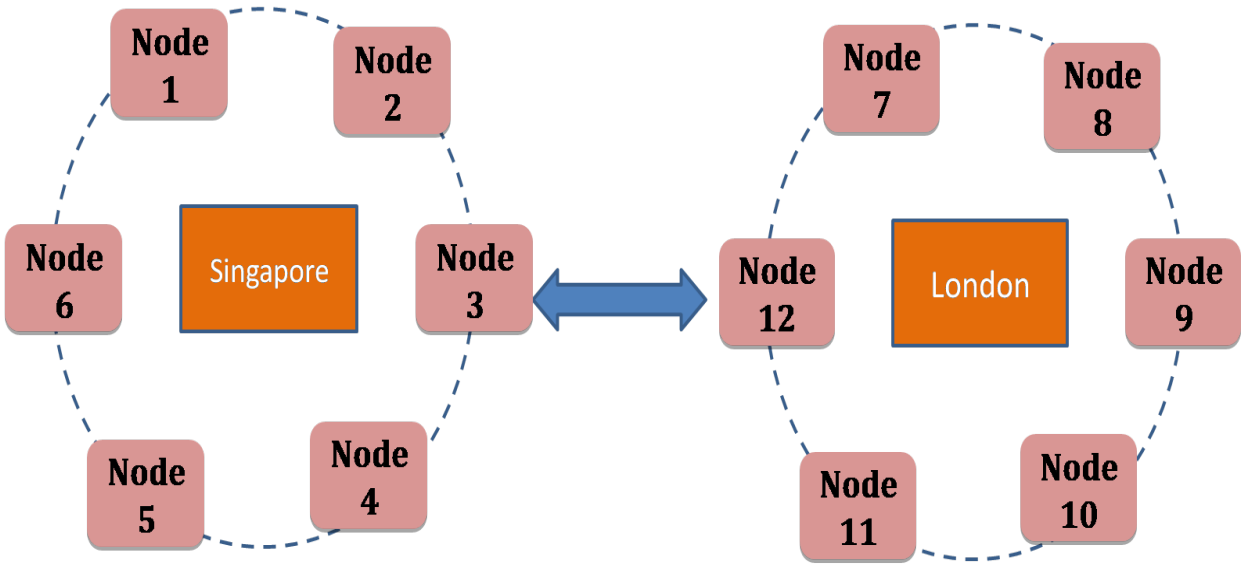
Discussion session of Unit 5

Q:



The graphic depicts one distributed NoSQL store that has 6 nodes in Singapore and 6 nodes in London. While the system is a single noSQL store, the data stored in the Singapore cluster of nodes serves the Pacific Rim while the data in the London cluster of nodes serves Europe. Expect there to be 20% overlap, so that each cluster holds 20% of the data held by the other cluster.

A:



Using the use case in the graphic and on the previous slide, think through how each of the below concepts might be implemented.

Data sharding: data are partitioned across nodes so that each node carries a similar user access (query) load as the other nodes

Replication: data are replicated across nodes so that a failure of one node does not result in a failure of the system

Sample 3000 Profiles data

- In the file, there are 3000 user profiles.
- In the file, each line represents a user profile.
- The format is as follows:

Format
[PROFILE 1]
[PROFILE 2]
….

Sample 3000 Profiles data

- A profile contains many useful information about a user, including User ID, User Name, Friend Count, Follow Count, Status Count, Favorite Count, Account Age and User Location. The format for a [PROFILE] is as follow

Format
[User ID] \tab\ [USER NAME] \tab\ [FRIEND COUNT] \tab\ [FOLLOW COUNT] \tab\ [STATUS COUNT] \tab\ [FAVORITE COUNT] \tab\ [ACCOUNT AGE] \tab\ [USER LOCATION]

Sample 3000 Profiles data

100008949 esttrellitta 264 44 6853 0 28 Dec 2009 18:01:42 GMT El Paso,Tx.
100009841 ChelseaBex 152 50 394 0 28 Dec 2009 18:05:43 GMT
100012792 ErinPattisonn 984 666 5003 0 28 Dec 2009 18:19:39 GMT under your bed.
100013967 TUBeautifulRosa323 251 1269 0 28 Dec 2009 18:24:51 GMT on Twitter ..... ahaahaa !
100014135 GeenaJohnson 144 130 9789 0 28 Dec 2009 18:25:37 GMT Arkansas
100015928 GooSau 93 286 8075 0 28 Dec 2009 18:33:59 GMT
10001882 rjwilson 1 340 6358 0 6 Nov 2007 15:54:47 GMT iPhone: 39.053871, 95.674576
100019750 HovMinajJackson 135 136 6022 0 28 Dec 2009 18:51:29 GMT neverland
100020433 MattieBX 131 97 2610 0 28 Dec 2009 18:54:40 GMT zundert
100024321 KatieStepek 64 93 503 0 28 Dec 2009 19:13:08 GMT Hamilton
……

Transfer package into Instance

- Mac

```
Scp SampleDataProjectTutorialData.tar.gz
xsede_username@instance_ip_address:/home/your_username/
```

- Window
  - WinSCP to transfer package into Instance directory

Package information

- Extract documents from zipped package at SampleProject directory

```
tar –zxf SampleDataProjectTutorialData.tar.gz
```

- reformat.sh
  - Reformat the user profile dataset from ISO-8859-1 to UTF-8 format
- import_mongodb.sh
  - Import tab-separated value file into MongoDB
- users_3000.txt
  - 3000 user profiles dataset

Create Database and Collection

- Open your MongoDB and
- Create Database ‘sampleProject’ and Collection ‘sample’ for this sample dataset

```
use sampleProject
db.createCollection(‘sample’)
```

Reformat the dataset

- The raw txt file of user profiles is encoded in ISO-8859-1 format. This is a format that the MongoDB NoSQL store does not accept, a common problem. So you will need to convert the txt file into the UTF-8 format that MongoDB accepts. You need to do this before you can store the Twitter user profiles into the MongoDB database.

Reformat the dataset

- Reformat the user profile twitter dataset from ISO-8859-1 to UTF-8 format by running the following reformatting script

```
./reformat.sh <input file> <output file>
```

- Sample command

```
./reformat.sh users_3000.txt users_3000.tsv
```

Headline of User Profile

- Use vi editor to open the file you created. Add the following line as the first line to the newly reformatted Twitter data file (it becomes the "headline", something MongoDB understands). Be sure that you use tabs to split the fields.

```
user_id user_name friend_count follow_count status_count
favorite_count
account_age user_location
```

Import data into MongoDB

- The tab-separated values (tsv) file can be imported directly into MongoDB, however, proper headerlines (fields) must be defined so that MongoDB can give structure to the data when converting it to its internal format

```
./import_mongodb.sh <db name> <collection name> <import file type>
<import file>
```

- <db name> is ‘sampleProject’, <collection name> is ‘sample’, <import file type> is ‘tsv’, <import file> is users_3000.tsv

Query on MongoDB

- Go to the MongoDB
- Do some queries on inserted dataset.
  - How many data you have insert into sample collection?

```
db.sample.find().count()
```

  - How many users has less than 100 friends?

```
db.sample.find({'friend_count' : {$lt : 100}}).count()
```

  - How many users has less than 100 friends and less than 30 followers?

```
db.sample.find({'friend_count' : {$lt : 100}, 'follow_count' : {$lt : 30}}).count()
```