SNS COLLEGE OF TECHNOLOGY

Coimbatore-35
An Autonomous Institution

Accredited by NBA – AICTE and Accredited by NAAC – UGC with 'A++' Grade Approved
by AICTE, New Delhi & Affiliated to Anna University, Chennai

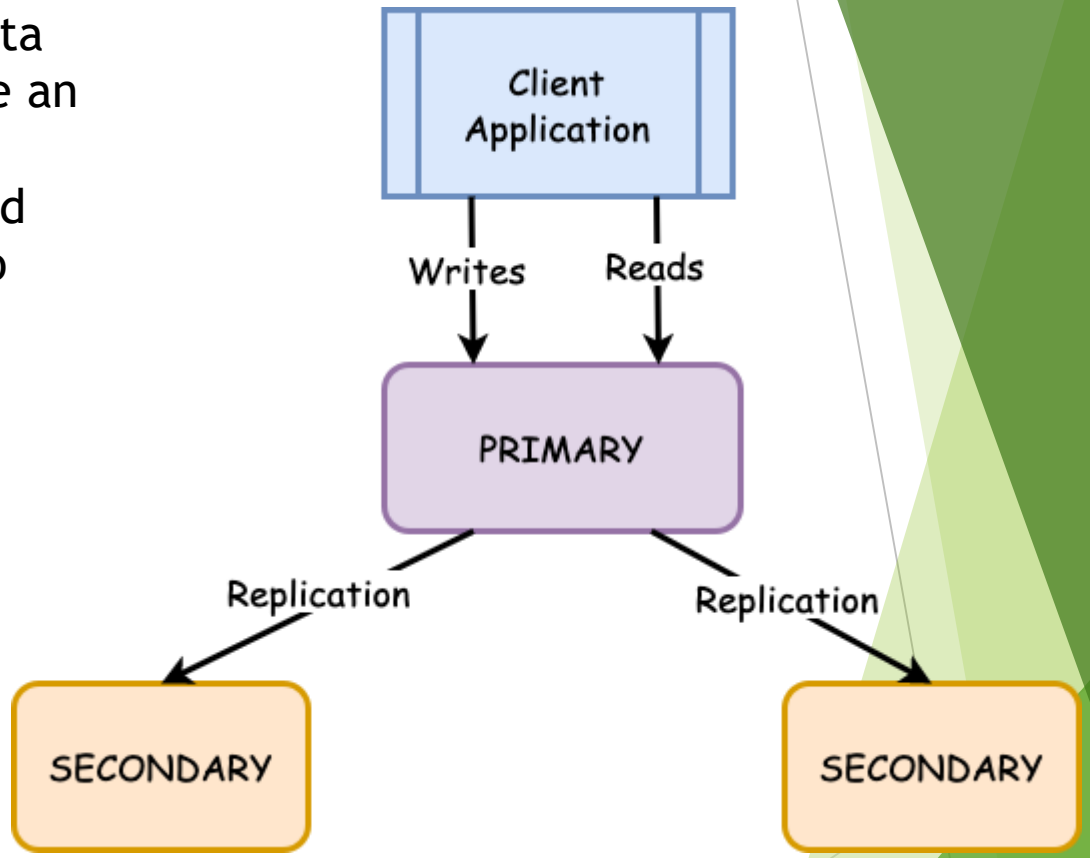
DEPARTMENT OF COMPUTER APPLICATIONS

19CAE716 – DATA SCIENCE
II YEAR III SEM

UNIT – V: MAPREDUCE

Replication

Replication is the method of duplication of data across multiple servers. For example, we have an application and it reads and writes data to a database and says this server A has a name and balance which will be copied/replicate to two other servers in two different locations.

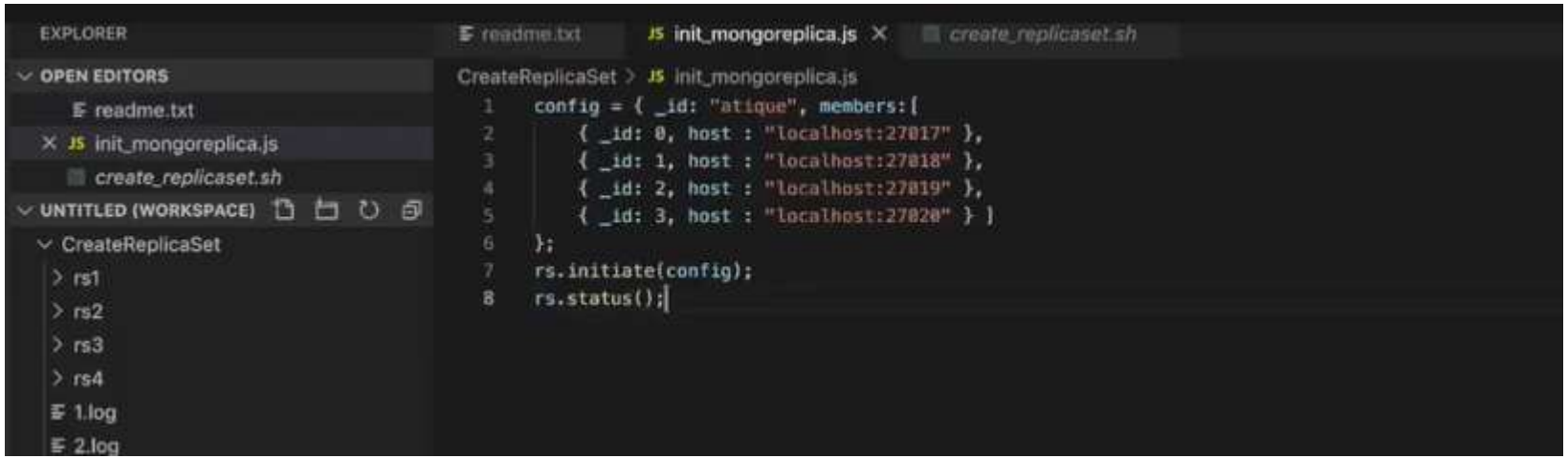


How replication is formed?

```
create a ".sh" file create_replicaset.sh
and init_mongoreplica.js
```

Then run the following script :

```
./create_replicaset.sh
```

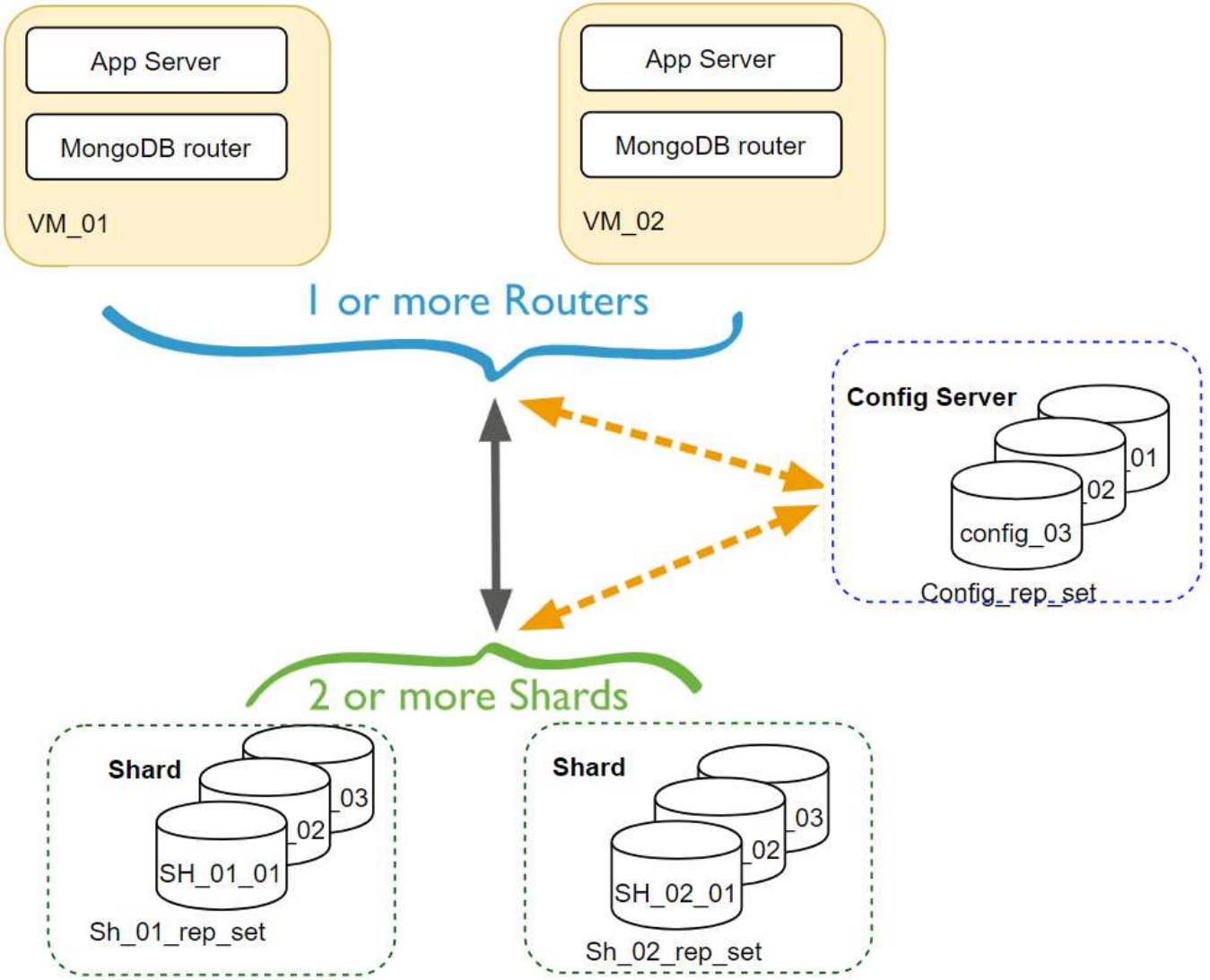


Database Sharding

The need for database sharding arises because as data grows, it becomes more challenging to store it on a single server. It is because the server has limited storage and processing power. As a result, the performance of the database can start to deteriorate.

There are a few things to consider when sharding a database:

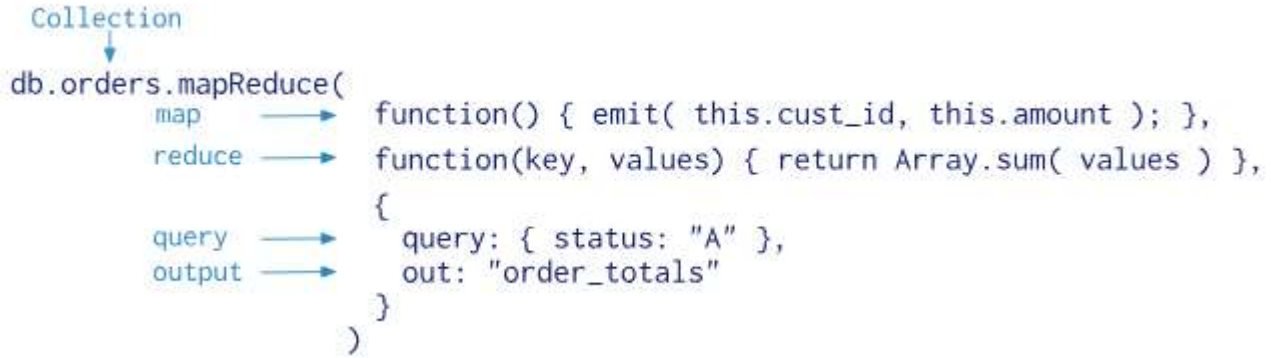
1. The data must be divided up in a way that makes sense.
2. The system must handle queries that span multiple servers.
3. The system must be able to recover from failures.

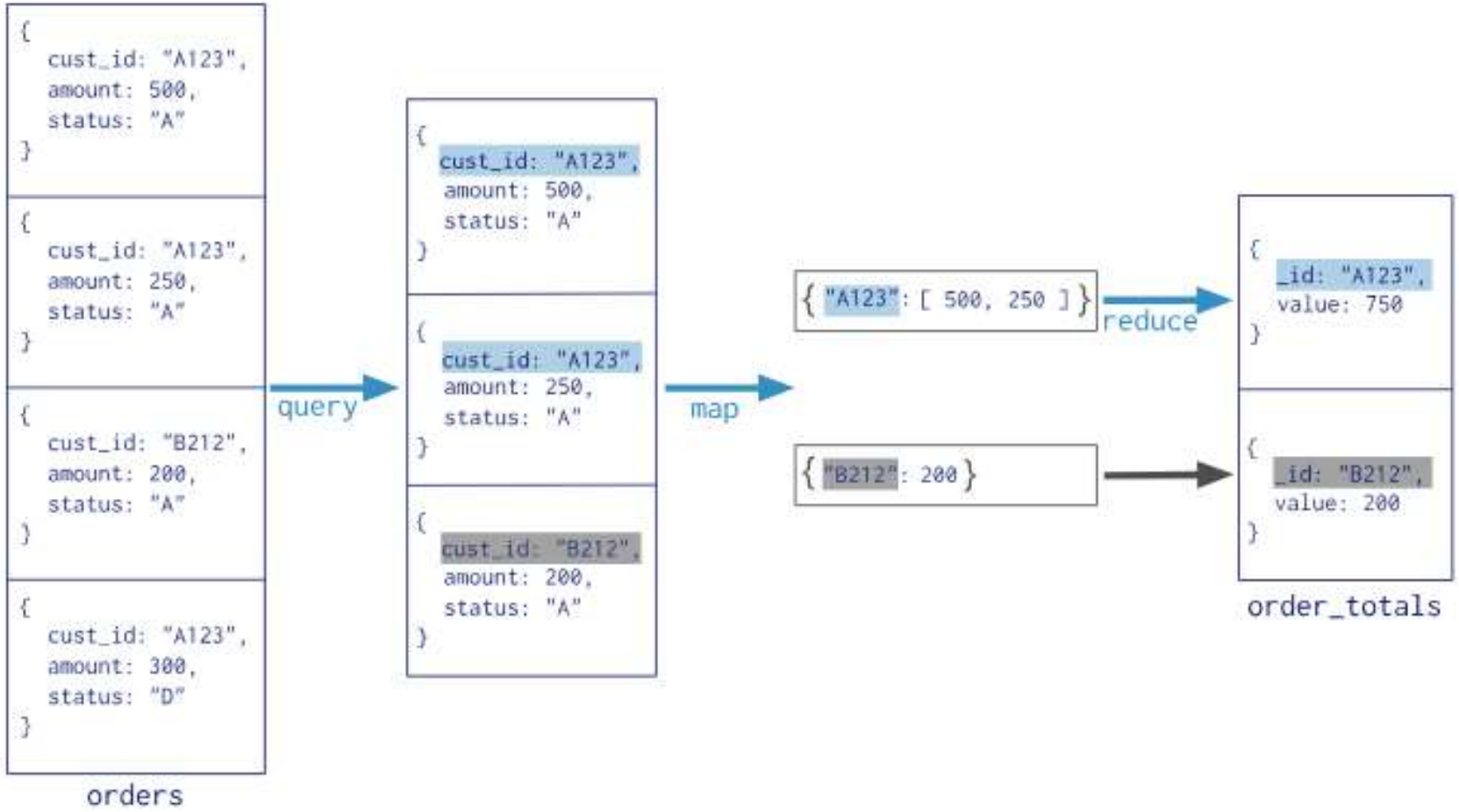

App Server
MongoDB router
VM_01
App Server
MongoDB router
VM_02
1 or more Routers
Config Server
config_01
config_02
config_03
Config_rep_set
2 or more Shards
Shard
SH_01_01
SH_01_02
SH_01_03
Sh_01_rep_set
Shard
SH_02_01
SH_02_02
SH_02_03
Sh_02_rep_set

Map-Reduce

Map-reduce is a data processing programming model that helps to perform operations on large data sets and produce aggregated results.

Consider the following map-reduce operation:




{ cust_id: "A123", amount: 500, status: "A" }
{ cust_id: "A123", amount: 250, status: "A" }
{ cust_id: "B212", amount: 200, status: "A" }
{ cust_id: "A123", amount: 300, status: "D" }
orders
query
{ cust_id: "A123", amount: 500, status: "A" }
{ cust_id: "A123", amount: 250, status: "A" }
{ cust_id: "B212", amount: 200, status: "A" }
map
{ "A123": [500, 250] }
{ "B212": 200 }
reduce
{ _id: "A123", value: 750 }
{ _id: "B212", value: 200 }
order_totals

map() function: It uses **emit()** function in which it takes two parameters key and value key. Here the key is on which we make groups like groups by in MySQL. Example like group by ages or names and the second parameter is on which aggregation is performed like avg(), sum() is calculated on.

reduce() function: It is the step in which we perform our aggregate function like avg(), sum().

query: Here we will pass the query to filter the resultset.

output: In this, we will specify the collection name where the result will be stored.

```
> db.employee.find()
{ "_id" : ObjectId("60192cb40cf217478ba935a2"), "name" : "eren", "age" : 25, "ra
nk" : 1 }
{ "_id" : ObjectId("60192cb40cf217478ba935a3"), "name" : "mikasa", "age" : 24, "
rank" : 2 }
{ "_id" : ObjectId("60192cb40cf217478ba935a4"), "name" : "jean", "age" : 26, "ra
nk" : 4 }
{ "_id" : ObjectId("60192cb40cf217478ba935a5"), "name" : "conny", "age" : 23, "r
ank" : 7 }
{ "_id" : ObjectId("60192cb40cf217478ba935a6"), "name" : "sasha", "age" : 25, "r
ank" : 6 }
{ "_id" : ObjectId("60192cb40cf217478ba935a7"), "name" : "armin", "age" : 24, "r
ank" : 3 }
>
```

Any questions?