NoSQL Assignment2
Comp7037

Joshua Nuttall, David Coughlan

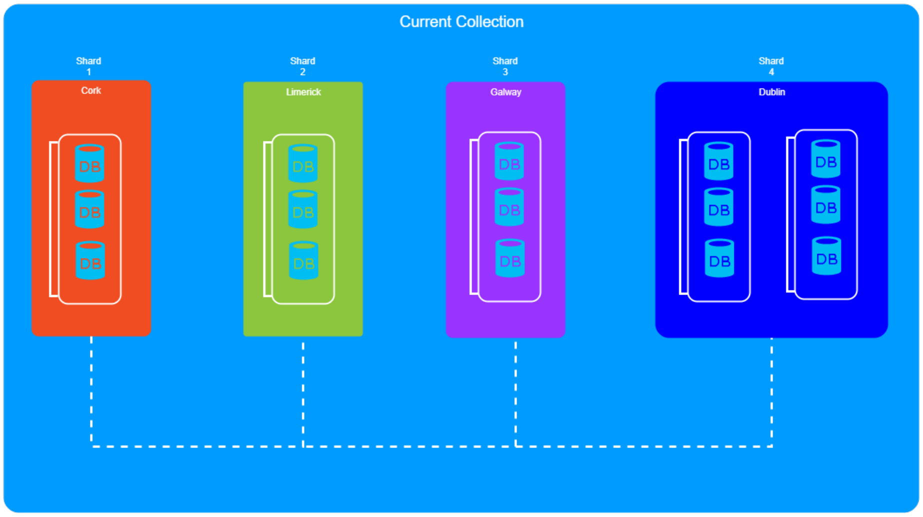
November 2017

1 Introduction

This report is to accompany queries.py, queries.js as well as other relevant Mongodb scripts and neo4j queries. The purpose of this report is to highlight alternate database configurations that may suit Rachel Allen's needs for present and future goals.

2 Current Configuration

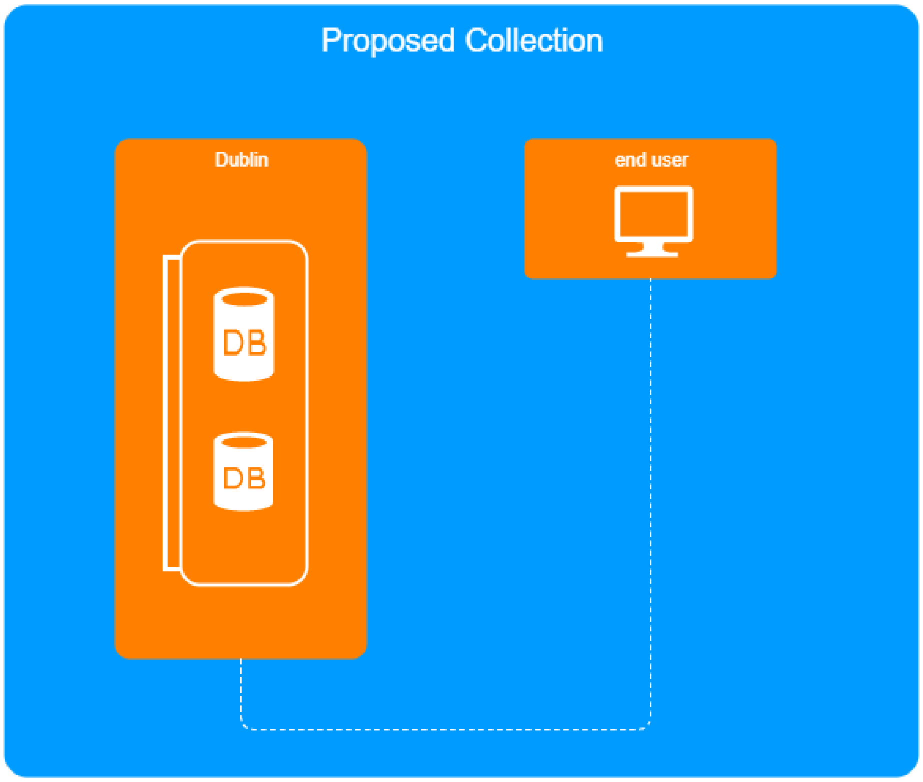
Rachel Allen's current configuration is comprised of a 4 shard system with 15 nodes in total, this setup includes 3 nodes per shard with 1 node acting as the primary and the other two acting as secondary nodes for purposes of redundancy with the exception of 6 nodes being based in Dublin that host the configuration servers containing the meta data for each primary node in the shards. This system is extremely complex for the data it is holding and has a large amount of administrative overhead due to this as well as being costly.



In total the data has a size of 11mb making this system extremely complex for the data it is holding, sharding is only usually used for the purpose of expanding your database in size or spreading workload across multiple machines for processing intensive tasks. The current system is sharding the data 5 times making each shard a size of 2.2mb on top of this we are replicating each node 2 times for redundancy meaning each shard now has a size of 6.6mb as well as the added overhead of extra computers maintenance and synchronization. While sharding can be used on small data sets where there is a expectation of rapid growth in the data however unfortunately this is not the case. The data could double in size and the current configuration could still not be justified. However the city of New York is highly unlikely to gain double the amount of restaurants over night. Taking price into consideration we can set the price of a cheap server as being 500 euro, we would have to spend a total of 7500 euro for the setup we currently have.

3 Proposed Configuration

While the above configuration does have its merits, there is no need for the sharding, the database simply does not have enough data and is not expected to grow at a rate that would justify the cost and complexity of such a system. The proposed solution would be to take just one replicate node of the database as to have a primary and secondary node. This would limit the processing power required for queries and synchronization of nodes as well as reducing cost to 1000 euro. Further more if expansion is required oppose to investing in new systems one could simply upgrade the current system reducing administrative costs of maintaining multiple systems although unlikely that this would be required.



4 Comparison

Comparing the proposed solution to the current solution the size of the database immediately stands out, whilst the given configuration is a good configuration for Big-Data solutions we are presented with a situation that simply does not have Big-Data hence the setup is not required, if the setup was in reverse where our system resources where consistently being pushed to the limits or storage was almost depleted then a sharded solution could most definitely be considered. One might argue that it is good practice to shard from initial setup for in such an event in the future where system resources might be more limited, however with the data in use even considering restaurants are opened and closed it is unlikely we will see growth in data rapid enough to even consider this for one city. With current technology allowing for terabytes of data available to one server it is unlikely such an expansion even including multiple cities would require more storage, at this point sharding may be considered for the benefits of extra ram and computing power, maybe allowing for sharding per state with replication.

As mentioned comparing our setups we see a price difference immediately, 1k oppose to 7.5k. our configuration does have less redundancy however it is unlikely to see a situation that a backup server would also become corrupt as well as the primary.

Administrative costs must also be considered the price of maintaining, updating and setting up the 15 systems used in the initial configuration oppose to the proposed 2 systems used in our configuration.