**Duplicate URLs:** You have 10 billion URLs. How do you detect the duplicate documents? In this case, assume "duplicate" means that the URLs are identical.

**Size:** 10 billion URL = 10 ^ 9 records

1. If the size is a concern then we need to use multi pass algorithm were we use buckets and then try to accommodate the URLs in each of these buckets for the first pass and then if needed we can try to find the duplicate in the second pass itself or go for further screening during the subsequent passes so that finally when the list becomes relatively small we can using hash map or other technique to find the duplicate for a smaller set.
2. If size is not a concern then we can use a hash based indexing and then separate the database to a distributed one (sharding with hash index) because even though we currently have only 10 billion records we might reach a point where the test data might be more than this and so it requires a scaling. So we can implement a splitting of database now rather than later.

**Approach #1:** Hashing

Calculate the hash of a URL and seek in the corresponding server for the entry.

If the entry already exists then record the URL in the duplicate list and proceed to the next

If the entry is not present then add and proceed to the next

Repeat it for 10 ^ 9 records

**Approach #2:** Parallel Computing

Create a thread for example 10 threads and perform the above process parallel and handle the concurrency issues, disk seek time delay, lock delay (mutex latency)