# CSP554—Big Data Technologies

## Assignment #8

**Exercise 1) Read and provide a half page summary and analysis of this article available on the blackboard in the ‘Articles’ section: Dynamo: Amazon’s Highly Available Key-value Store.**

From the paper:

The paper tells about “Dynamo: Amazon’s Highly Available Key-value Store” a reliable service that is always available to users but does so by sacrificing its consistency. Dynamo achieves this by using consistent hashing and this hashing is achieved by object versioning. It is decentralized and has less human intervention. The system has services like Query Model: that does read and write operation by using a unique key the ACID properties: to guarantee reliability and efficiency to meet strict requirements. Service Level Agreement is like a contract between the client and service where the service guarantees that the requested distribution for API and its latency. This is an essential component in the decentralized architecture of Amazon. The paper mainly references Amazon’s 99.9th percentile of distributions. Dynamo focuses on “always writeable” data store i.e. it is always available for write operations which helps to solve the customer update issue. Dynamo supports incremental scalability(scale out storage host), symmetry(peers sharing same responsibility), decentralization and heterogeneity. Systems like Ficus and Coda replicate files which ensures availability but at the cost of consistency. They are not affected by network partitions and outages and they perform system level conflict resolution. Dynamo mainly has key/value pairs that maintain high availability. Dynamo focuses on “always writeable” scheme where updates won’t be discarded because of network failures. Dynamo has two operations get() for locating the object replicas with key and put() to decide where to put the replicas with its associated key. Dynamo scales incrementally that dynamically partitions the data. It also replicates data along all its nodes. Data versioning is Dynamo helps to maintain consistency among its replicas asynchronously. It uses “sloppy quorum” meaning that the all read and write operations will be performed on first N healthy nodes of the preference list. Dynamo has 3 main software components: request coordination, membership and failure detection, a local persistence engine which are all implemented in Java. This, Dynamo states that decentralized techniques can be combined to form a single engine that is highly available. It has proved it to be a success in challenging application environments that require high availability.

**Exercise 2) Read and provide a half page summary and analysis of this article available on the blackboard in the ‘Articles’ section: Data management in cloud environments: NoSQL and NewSQL data stores.**

From the paper:

The paper mainly focuses on 3 objectives for NoSQL and NewSQL solutions with the objective of providing perspective, providing guidance to researchers to choose a data store and identifying the challenges and opportunities in the field. Initially the data was stored on database models which could not be processed for storing large data set values as there were issues with performance and scalability. So, to address these issues there was introduction of NoSQL and NewSQL. It states that NoSQL offers flexible schemas and has data models that can be divided into: key/value pairs, document stores, column family stores and graph databases. The characteristics of this make it suitable for cloud management systems. Due to tremendous amount of data being generated it becomes difficult to understand domain and is challenging to select a correct solution for the processing of such data. Hence to handle huge volumes of data NoSQL and NewSQL were introduced. The analysis of the size of data and its environment were done to choose a strategy. There are huge amounts of data being generated and 90% come from social networking sites. NoSQL and NewSQL support MapReduce. They are also used to add servers into the data pool i.e. they can scale horizontally. They also have a feature of replication hence they can be stored in distributed systems also. The last feature they use is multi-version concurrency control that uses advanced techniques like optimistic locking. There are various strategies for page replacement and partitioning schemes that are developed to handle this data. Hence the paper states that NoSQL and NewSQL are the solutions to the researchers to provide solution to their problem and find a data store that could be used and solve their challenges. The solutions proposed in this paper address security and scalability concerns and data models. It also states that NoSQL and NewSQL although address these solutions diversity is also an issue. It requires that common terminology be established if there is same type of data model. Various use cases and scenarios were provided for NoSQL and NewSQL data sets that were suitable for various application environment. It gives solution for scalability issues and security issues with respect to authorization, authentication and encryption. There are still problems of limited documentation non-existence of query language etc. that needs to be looked upon.