# Ritika Kumari-A20414073

# 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.

**Answer:** Dynamo is Amazon’s highly available key-value storage system. Dynamo provides a simple primary-key only interface to meet the requirements of applications that only need primary-key access to a data store such as those that provide best seller lists, shopping carts, customer preferences, session management, and product catalog etc. Dynamo’s use case for these services is providing a highly-available system that always accepts writes. This requirement forces the complexity of conflict resolution to data readers. Writes are never rejected.The paper focuses on how Dynamo combines several core distributed systems techniques to solve Amazon’s use case. This review focuses on the solutions to partitioning, replication, and data versioning. Dynamo is used to manage the state of services that have very high reliability requirements and need tight control over the tradeoffs between availability, consistency, cost-effectiveness and performance. Dynamo uses two well-known techniques: **Scalability and Availability** where Data is partitioned and replicated using consistent hashing, and **Consistency** which is facilitated by object versioning. The consistency among replicas during updates is maintained by a quorum-like technique and a decentralized replica synchronization protocol. Dynamo is a completely decentralized system with minimal need for manual administration. Storage nodes can be added and removed from Dynamo without requiring any manual partitioning or redistribution.

The main contribution of the paper is an evaluation of how various ideas like variation of consistent hashing with the concept of virtual nodes, Replication with eventual consistency, "Hinted Handoff" and "Sloppy Quorum", Vector clock for data versioning with application specific conflict resolution etc. which have been proposed in the literature can be put together to build a real-world distributed system. Each of these ideas have been optimized and varied specifically for Amazon's datastore. In addition, the paper also uses some innovative ideas, such as choosing coordinators for writes based on the latencies of a previous read operation.

Dynamo is a highly available and scalable data store, used for storing state of several core services of Amazon’s e-commerce platform. It provides the desired levels of availability and performance and has been successful in handling server failures, data center failures and network partitions. Dynamo is incrementally scalable and allows service owners to scale up and down based on their current request load. Dynamo presents an extremely simple interface to users, providing only get and put commands. Despite this simple interface, Dynamo allows a high degree of customization, allowing application level conflict handling, and tuning of R, W and N parameters per application.

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.

**Answer:** Cloud computing has emerged as a computational paradigm that can be used to meet the continuously growing storage and processing requirements of today’s applications such as Web technologies, mobile devices, sensory data, IoT etc. The article focuses on the storage aspect of cloud computing, specifically on data management in cloud environments.

The article reviews NoSQL and NewSQL data store solutions to provide a perspective on the domain by summarizing, organizing, and categorizing NoSQL and NewSQL solutions and to compare the characteristics of the leading solutions to provide guidance to practitioners and researchers to choose the appropriate data store for specific applications.

Specifically, the most prominent solutions are compared focusing on several dimensions, including Data Models (Key-Value stores, Column-family stores, Document stores, Graph Databases, New SQL), querying capabilities, scaling, and security attributes related capabilities. Features driving the ability to scale read requests and write requests, or scaling data storage are investigated, partitioning, replication, consistency, and concurrency control. Furthermore, use cases and scenarios in which NoSQL and NewSQL data stores have been used were discussed and the suitability of various solutions for different sets of applications was examined. The discussion of the use cases, together with the comparison of data stores, will assist practitioners in choosing the best storage solution for their needs. This study has also led to identification of challenges in the field, including the immense diversity and inconsistency of terminologies, limited documentation, sparse comparison and benchmarking criteria, occasional immaturity of solutions and lack of support and nonexistence of standardized query languages.

Traditional relational databases were designed in a different hardware and software era and are facing challenges in meeting the performance and scale requirements of Big Data.

NoSQL and NewSQL data stores present themselves as alternatives to traditional relational databases, capable of handling huge volumes of data by exploiting the cloud environment.

Because of the large number and diversity of existing NoSQL and NewSQL solutions, it is difficult to comprehend the domain and even more challenging to choose an appropriate solution for a specific task.

The article focuses on the storage aspect of cloud computing, specifically on data management in cloud environments.

The discussion of the use cases of implementation of NoSQL and NewSQL data stores, together with the comparison of data stores on several dimensions such as Data Models, Querying, Scaling, and Security, helps assist practitioners in choosing the best storage solution for their needs. In addition, it has identified challenges in the domain, including terminology diversity and inconsistency, limited documentation, sparse comparison and benchmarking criteria, occasional immaturity of solutions and lack of support, and non-existence of a standard query language.