**CSP 554 – Assignment #8**

Name: Adarsh Mathad Vijayakumar

CWID: A20424847

Email ID: avijayakumar@hawk.iit.edu

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.

* At Amazon world-wide e-commerce platform reliability is one of the biggest challenges as even the slightest outage has significant financial consequences which impacts customer’s trust. To meet the reliability and scalability needs Amazon has developed a number of storage techniques such as S3 and Dynamo.
* Many services on Amazon’s platform that only need primary-key access to a data store such as those that provide best seller lists, shopping carts, customer preferences, session management, sales rank, and product catalog. Dynamo provides a simple primary-key only interface to meet the requirements of these applications. It is a decentralized distributed data store with high availability and scalability. It partitions and replicated data using consistent hashing and is facilitated by object versioning. The consistency among replicas during updates is maintained by a quorum-like technique and a decentralized replica synchronization protocol.
* The main assumptions/ requirements by Dynamo is storing objects that are relatively small and require a simple query model with no complexities of the relational database. It relaxes the ACID properties and does not provide any isolation guarantees and permits only single key updates. Configuration support in Dynamo to ensure services can always achieve their latency and throughput requirement measured at the 99.9th percentile of the distribution.
* Design Considerations: Dynamo is designed to be an eventually consistent data store; that is all updates reach all replicas eventually. It pushes the complexity of conflict resolution to the reads in order to ensure that writes are never rejected. It allows applications perform the process of resolving conflicts else uses “last write wins” as a fallback strategy. It supports easy scale of one storage system at a time. Every node in the system has same set of responsibilities as its peers. It is decentralized and heterogeneous.
* In its system interface, it exposes two operations: get() to read an object based on its key, put() to update an object associated with a key. It treats both key and object as an opaque array of bytes. It applies a MD5 hash on the key to generate a 128-bit identifier, which determines the storage nodes that are responsible for serving the key.
* It uses consistent hashing while partitioning so that it can scale incrementally. It provides eventual consistency, which allows for updates to be propagated to all replicas asynchronously.
* It provides configuration parameters such as N, W and R, which can be fine-tuned based on the service needs. Different configurations work well to achieve their desired levels of performance, availability and durability. For instance, the value of N determines the durability of each object. A typical value of N used by Dynamo’s users is 3. The values of W and R impact object availability, durability and consistency.
* Dynamo has provided the desired levels of availability and performance and has been successful in handling server failures, data center failures and network partitions. It is incrementally scalable. It has demonstrated over the past one year that decentralized techniques can be combined to provide a single highly-available system.

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.

- This paper focuses on the NoSQL and NewSQL data store as a storage solution for Big Data. It reviews NoSQL and NewSQL solutions with the objective to provide a perspective in the field, provide guidance to choose the appropriate data store, and identifying challenges and opportunities in the field.

- Big Data is a term used to refer to massive and complex datasets made up of a variety of data structures, including structured, semi-structured, and unstructured data. It can be defined by 3Vs: volume, velocity, and variety.

- Cloud computing platform has emerged as a paradigm to solve the immense processing and storage requirements of today. However, it imposes new requirements to data management such as Scalability and high performance, Elasticity, Ability to run on commodity heterogeneous servers, Fault tolerance, Availability, Security and privacy features.

- NoSQL and NewSQL are an answer to the above requirements that can handle this huge volume of data and provide the required scalability. NoSQL solution typically have the following characteristics: Simple and flexible non-relational data models, they offer flexible schemas or are sometimes completely schema-free and are designed to handle a wide variety of data structures. Ability to scale horizontally over many commodity servers. Provide high availability. They do not support ACID transactions as provided by RDBMS. However, some NoSQL data stores, such as CouchDB provide ACID compliance.

- The family of data stores belonging to the NoSQL category can be further sub-classified based on their data models, such as key-value stores (ex: Memcached and Redis), column-family stores (ex: Cassandra), document stores (ex: MongoDB), and graph databases.

- NewSQL solutions are by definition based on the relational model. (ex: VoltDB, Clustrix and NuoDB).

- The paper presents us with a table of comparison of various NoSQL and NewSQL solutions in the market based on data model, querying capability, Partitioning, replication, consistency, and concurrency control capabilities, and security features which will help practitioners to pick the right datastore based on their application needs. It also discusses some general guidelines to be followed while picking a data store solution. Example: key-value data stores are appropriate for scenarios in which applications access a set of data as a whole using a unique value as the key

- It presents us with the opportunities and challenges available in the area such as establishing a common terminology in this area, establishing a SQL like query interface, establishing standardized performance benchmarking, establishing sophisticated security and privacy mechanism, increasing the documentation and support, experimenting using more than one NoSQL data store for a single application need.

- This paper as covered the reason for moving to NoSQL and NewSQL data store solution, it has highlighted their main characteristics, provided the knowledge to compare solutions in this area and has opened up areas for further research opportunities.