# CSP554—Big Data Technologies

## Assignment #7

**Exercise 1) Read and then write a half page summary of the article “Lambda architecture for cost-effective batch and speed big data processing” found on the Blackboard in the “Articles” section.**

From the paper:

The paper uses the Lambda Architecture design pattern on Amazon Web Services- EC2 to handle online data processing of Big Data datasets so that there is high throughput and network cost maintenance is reduced. It uses the sensor data ESnet network data as an example to show their results. The use of lambda architecture is that the user knows which data needs to be done online processing and which needs to be done batch processing. The online processing is used to detect anomalies and batch processing helps to detect data patterns by using the previous data. Also, the author states that this data processing carried out was easier in AWS as compared to Azure. The author uses DynamoDB instead of S3 because the query processing speed is more which makes it cost effective as a whole when considering the long-term goal. In the initial implementation Spark SQL was used doing the map and reduce operation for the batches in a timely manner for faster query analysis. In this system the data arrives at 30 seconds interval and maximum and average values is calculated then passed to a new input stream for visualization. Batch processing is carried out on these data sets in a consolidated manner then these outputs are stored in separate S3 buckets. The author has shown results for 3 months which states that instead of using large machines we can replace them with smaller machines to give same results but in a cost-effective manner. The paper concludes that when local testing is done, we can solve the errors easily before deploying it on EC2 clusters. The paper helps in producing cost optimized solution by using lambda architecture where processing is done when data arrives followed by data aggregations that reduce time for visualizations.

**Exercise 2) Read and then write a half page summary of the article “Azure Data Lake Store: A Hyperscale Distributed File Service for Big Data Analytics” found on the Blackboard in the “Articles” section.**

From the paper:

The paper highlights about the Azure Data Lake Store that consists of multiple storage tier, scalability security and data sharing and is mainly used when there are numerous parallel read and writes. ADlS is a type of PaaS service. The ADLS and Azure Data Lake Analytics (ADLA) are used to handle workloads of Hadoop, Hive, Spark and Microsoft’s Scope and U-SQL.The weakness of the Big data analytics is solved by the paper using data location services on storage tiers by using a single ADLS layer. It keeps the location of data transparent thus reducing the problems for the user. The design of ADLS is made in such a way that it supports local tiers stored in ADLS clusters that can be used for job computation and remote tiers that reside outside ADLS cluster. ADLS supports partial file that are sealed. The ADLS consists of three components: backend nodes, microservice nodes and front-end nodes. The back-end nodes are used for storing data on local tiers where ADLA carries out its computation. The front-end act as gateway for scheduling requests and the microservices are used for hosting the microservices. The RSL-HK ring structure is the most important one for supporting large files and folders that provide scalability, high availability and security. The Secure Store Services (SSS) acts as an entry point into ADLS that creates boundaries between microservices and storage providers. Every request has a secure token that identifies caller and to do the authorization checks. Secret Management Service is used to handle high throughput and low latency and also including the functioning of microservcies. ADLS provides encryption of data that means each block that is appended is separately encrypted by using a unique key such that cypher generated is small and as every block contains metadata it allows block level integrity check and algorithm identification. The paper concludes that ADLS has improvements over Cosmos mainly in the security feature. The big data overloads have been replaced and can be used externally with Azure. The shortcoming of this paper is data migration for large datasets that would be addressed in the future