## ADTA5340 Final Project

## PART I: A Strategy to Employ Machine Learning in a Firm

### Define data to collect

First thing we need to determine what data we would collect. We need to take samples of datasets and estimate their current volume and growth. We need to check if data requires additional preprocessing. We need to plan for enough storage and processing power and we should be able to manage capacity at real-time without affecting processing capabilities.

The system should not be used as data store for any other business critical application, rather it should be able to collect / consume data from those systems.

### High-level architecture

The system would have several independent layers of processing. Independent levels would ensure resiliency and provide ability to scale each layer independently.

#### Data collection layer.

This layer would be used to collect and store initial data from other systems. This layer should have enough storage capacity to collect data for a period of time before it would be moved to other layers and processed there. Capacity of this system should be kept at maximum of 50%. Depends on how much data coming each second, we would need to create a storage with fast write.

#### Data preprocessing layer.

The purpose of this layer is to verify, clean, and get data ready for further processing. Also, this layer would be used as a quality assurance system. If there would be large amount of preprocessing required for some dataset, we would be able to alert data quality issues to respective systems’ owners.

#### Data storage layer.

This is long-term storage. It would require high capacity and fast read time. We should be able to increase capacity fast without affecting currently stored data.

#### AI/ML processing layer.

This layer requires highest number of CPU and memory to constantly build and run models. Storage requirements are not high, it should have some storage to keep some temporary files, but most of the time it should be able to keep everything in memory.

#### Reporting layer.

This is separate reporting layer. We need it to generate and store reports. In terms of CPU, memory and storage it would have mediocre requirements.

#### Web Access layer.

We need this layer to perform following functions:

* Overall system management, including access management;
* User access to run models and get results;
* Provide access to stored report and to report management functions;

#### API Access layer.

Modern organizations are data driven. Lot of systems produce, process, and store different data, that consumes by other systems in automated manner. We need a layer that could provide data for other machines in a machine-readable way.

#### System Alerts layer.

This layer is required to send alerts related to the system itself (health, reports readiness, security events, etc.), as well as alerts related to various external systems (i.e. data quality)

### System diagram

### Technologies to be used.

#### Hardware

Hardware should be based on commodity systems of x86/x64 architecture. This would provide ability to replace systems fast with low cost. Different components should be packed into standard racks with ability to add new racks into the system.

Also, we should replace components by racks. When certain percent of systems within a rack have issues would should replace the entire rack. System should be able to rebalance itself automatically.

#### Software

Software should be mostly based on open-source software. It would provide ability to both: change code of applications or develop new code, and at the same time use achievements from other software developers.

## PART II: Big Data, Artificial Intelligence, and Machine Learning

### The history of artificial intelligence until now

(3 pages min including images)

### Select three different sectors of the U.S. economy, do research, and discuss the impacts of **big data** and **machine learning** on **each** of them

(3 pages min including images)

### Discuss **in detail** the three major styles of learning in machine learning: (1) Supervised Learning, (2) Unsupervised Learning, and (3) Semi-Supervised Learning

## PART III: Data Preprocessing

## PART IV: Machine Learning: Supervised

## PART V: Machine Learning: Supervised

## PART VI: Machine Learning: Unsupervised

## PART VII: Evaluate and Compare Machine Learning Models

## PART VIII: Final Presentation Videos: YouTube Links