## 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

For data preprocessing task I found a dataset “Aviation Data and Documentation from the NTSB Accident Database System”.

URL for the dataset: <https://catalog.data.gov/dataset/aviation-data-and-documentation-from-the-ntsb-accident-database-system>

Public: This dataset is intended for public access and use.

License: No license information was provided. If this work was prepared by an officer or employee of the United States government as part of that person's official duties it is considered a U.S. Government Work.

This dataset was initially in XML format, I had to export it into CSV to work with it. The dataset has 84088 records and 31 variables:

\_EventId - 0 - Event identifier, each event should have an id. Field is not useful for further analysis as it is unique.

\_InvestigationType - 4 - Type of investigation (Accident, Incident). Some events have no type. NULL value could be replaced with Unknown.

\_AccidentNumber - 0 - Number (identification) of an event in some recording systems. Field is not useful for further analisys as it is unique.

\_EventDate - 0 - Date of an event. Cannot be NULL.

\_Location - 76 - Location of an event (city, state in case of the US or other identifiable location if outside of the US). Some locations are NULL. There are other variables for locations data.

\_Country - 507 - some locations do not have county filled. It could have NULL values when an event happened outside of any country (i.e. under an ocean or a sea).

*\_*Latitude – 54039

\_Longitude – 54048

\_Latitude and \_Longitude are geographical coordinates of an event. It should not be empty, but there are other location variables that could substitute them. If we really need them we need to work with other variables to get exact coordinates of an event.

\_AirportCode - 36439 - each airport has a code (3 or 4 letters). An event could happen outside of airport, so it could have NULL values. This variable could be used to define geographic location of an event.

\_AirportName - 33735 - Each airport has a human readable name. In case an event of outside of any airport this field would have NULL value. I would ignore Location variables. An event could happen at any location. But for general statistics these values could be useful to find which locations have more incidents.

\_InjurySeverity - 0 - Each event has assigned severity. Cannot be NULL.

\_AircraftDamage - 2676 - Describes damage to an aircraft. NULL value indicates no damage, NULL should be replaced with "NODAMAGE".

\_AircraftCategory - 56751 - Describes category of an aircraft like airplane, balloon, etc. NULL value indicates that category was not defined or not recorded. Missing value could be determined by Make and \_Model variables.

\_RegistrationNumber - 3778 - Categoriacal value (non-numeric). NULL value indicates that aircraft was not registered. For analysis I would remove this column. Replacing NULL with something like "NOT REGISTERED" would affect results of analysis: lot of events would be falsely attributed to it.

\_Make - 70 - Each aircraft should have a producer like Boeing, Airbus, Cessna, etc. But some are made by people and do not have registered Maker.

\_Model - 99 - Some aircrafts do not have models because they were built by private persons.

\_AmateurBuilt - 592 - indicates if aircraft is built by amateur builder. NULL value indicates that data was not recorded for this variable, NULL could be replaced with "NO"

\_NumberOfEngines - 4970 - Some aircrafts has no engines (like baloon or glider). But sometimes aircraft definetely have 1 or more engines but information was not collected.

\_EngineType - 4280 - Describes engine types. NULL indicates either there is no an engine, or data is missing.

\_FARDescription - 57056 - Describes if an aircraft performed a specific function (i.e. Armed Forces or General Aviation, etc.). NULL indicates no data is recoded, it should be changed to Unknown.

\_Schedule - 72269 - Indicates if a flight was scheduled (like AA20 from DFW to London Heathrow), unscheduled (i.e. for Armed Forces interception) or unknown. NULL values should be replaced with "NO".

\_PurposeOfFlight - 4791 - Describes the purpose of flight. NULL values should be replaced by "Unknown".

\_AirCarrier - 79927 - Names air carrier. Most of events have no air carrier, because event happened with personal aircrafts that do not perform regular passengers’ transfers. I would ignore this column.

\_TotalFatalInjuries - 27068 - Describes number of total fatal injuries. NULL values should be replaced with "0".

\_TotalSeriousInjuries - 29726 - Describes number of total serious injuries. NULL values should be replaced with "0"

\_TotalMinorInjuries - 28523 - Describes number of total minor injuries. NULL values should be replaced with "0"

\_TotalUninjured - 14507 - Describes number of total uninjured. NULL values show that there are no data. Normally it should show total people on board minus all injures combined. But there is no variable for total passengers, so we cannot calculate this value. I would ignore it.

\_WeatherCondition - 3060 - This variable describes weather conditions. NULL values indicate that there is no data on file, or weather had nothing to do with an event.  
BroadPhaseOfFlight - 6691 - Describes phase of flight when an event happened. NULL values should have "UNKNOWN"  
ReportStatus - 0 - Describes status of a report. Cannot have NULL values.  
\_PublicationDate - 14236 - Date of a report publication. NULL value shows that report is not published (could be "Secret" or not ready for publication)

## PART IV: Machine Learning: Supervised

## PART V: Machine Learning: Supervised

Dataset adult\_salary.csv

Most of string values has a space symbol at the beginning. With Excel I run a replacement function to remove all extra space symbols

## PART VI: Machine Learning: Unsupervised

## PART VII: Evaluate and Compare Machine Learning Models

## PART VIII: Final Presentation Videos: YouTube Links