Oregon State University
College of Engineering

CS Capstone Problem Statement

October 9, 2017

CDK Data Stream AI

Prepared for

CDK Global

Chris Smith ______________________ ________
Signature Date

Prepared by

Group 65

Jacob Geddings ______________________ ________
Signature Date

Inhyuk Lee ______________________ ________
Signature Date

Juan Mugica ______________________ ________
Signature Date

Abstract

Our team has been assigned to assist in the development of AI for application to CDKs existing Data Streams. The goal in doing this is to gain insight from said data streams, use this data to predict future events, and detect when an anomaly is present. These three goals need to be independent of one another and be capable of functioning as a black box. This entails the ability to write applications and/or functions that are then fed through the system in various ways. Tools that will be necessary for project completion include the AWS platform, Docker, Linux platform, and several open source AI resources.

Contents

1 Defining the Problem

ROUGH DRAFT PRE-MEETING WITH CLIENT

CDK currently has a wide range of data streams that can be improved upon, and they believe that the introduction of AI is the solution. In its current form, the various data streams theyre using do not possess the ability to easily interpret data for predictive analysis or for detecting potential problems. This results in an area that is difficult to manage, a situation that is only exacerbated when considering the constant need to expand their existing structures. Team interpretations of project include the possibility of handling customer information in order to allow direct customer interaction with predictive analyses which interpret what they may want or need. Another potential depiction involves the interest in seeing data streams regulated between servers and mobile devices.

2 Proposed Solution

The proposed solution is to incorporate AI as a form of black box that all data must pass through. This AI will need to be versatile enough that it can handle multiple forms of data passing through it. As the data is fed through, it should be able to document that data and reference it against other data that has been passed through. In doing this, the AI must be able to determine likely future scenarios given past events. It must also detect when something is drastically outside of what is expected. This will be accomplished through the use of open source AI with a focus on analytics and business interests such as H20, Caffe, Deeplearning4j, and Apache SystemML. This will also require a platform off of which to operate, such as Docker, which enables a universal platform via a program that can be installed on multiple machines with minimal system conflicts. An environment to work on this program can also be provided through AWS services, which allow access to various integrated development environments and software development kits all under one roof. Another convenient functionality of AWS is the ability to utilize serverless developer tools such as AWS Serverless Application Model that natively supports CloudFormation. To also meet the need for the program to be portable, we will be utilizing Docker as a general platform. This will give us an easy-to-use container in which to store our program that will ideally result in minimal conflict with whatever system on which it is installed. For security and stability concerns, this will be written for *nix based platforms with a focus on CentOS, which provides a consistent, compatible, and reliable platform. Lastly, given there is an interest in cloud-based support, we will also need to establish proper network links for the program in the event that it needs to communicate with copies of itself or be accessed from various input locations. With network links, the program will need to be sufficiently tested to handle either TCP or UDP communications. It will also need to be tested for its ability to reliably communicate with wireless devices. The language itself will likely be a variant of Java with AWS providing the cloud network link for this project. Given the scope of this project, we will need to maintain close contact with our client to ensure it remains on track with the desired end goal. This will mean three key building blocks needing to be completed that will then be linked to create a uniform design: the AI will need to be developed to properly handle the data it is fed; the container in which the program will reside must be constructed; and the network must be constructed to bind all copies together.

3 Performance Metrics

There are a few hard requirements and several assumed requirements in regards to the methods of gauging project completeness. Hard requirements include the following. Each functionality must be independent such that analytics and error detection are not bundled together but can instead only use one or the other. This must follow a black box design which does not require the user of the end product to know the inner workings of the program; the user needs to only know what the expected input and output. Lastly, the system needs to work across several different machines. As for assumed requirements, the program will need to be capable of returning reports on the material CDK is interested in. It must also be capable of running for extended periods of time without failure to allow it to monitor data flow as well as accumulate sufficient data points to infer upon. It must correctly communicate with any cloud service or other networking method to work in tandem with all copies of the program able to share information. It needs to be optimized to a sufficient level that the performance impact on the existing structure is kept to a minimum. This will also require an easy method of maintenance or training for the program when requiring operator corrections should it give false positives or negatives. Anomaly detection will need to be extremely robust and capable of dealing with potential small errors in a time efficient manner. Documentation of what went into the program and how it functions will need to be suitably detailed such that a new team can take over as the maintenance team. In the event that the endeavor proves incapable of being completed, a detailed report will be required as to the what, where, why, and when said failure occurred. In either event, documentation must explicitly indicate every source used in the creation of the project with how it was used and where. During the development process, the success of the project can be gauged by several means. One such method will be through contact with CDK: should the project not go in a direction they want, it can be easily corrected through the established email contact method with our client. Communicating with Oregon State University will also help maintain a measure of success. This will include speaking with our teaching assistant and our instructors for guidance and instruction on these various new areas of study. Lastly, as the project begins to near completion, a traditional burn down chart should assist in determining what is left to do and in what areas.