WGU

Josh Gaweda

**Capstone**

TABLE OF Contents

[1.  Summary of the problem 5](#_Toc67038535)

[2.  A description of how the data product benefits the customer and supports the decision-making process 5](#_Toc67038536)

[3. An outline of the data product 5](#_Toc67038537)

[4. a description of the data that will be used to construct the data product 5](#_Toc67038538)

[5. tHE OBJECTIVES AND HYPOTHESES OF THE PROJECT 5](#_Toc67038539)

[6. AN OUTLINE OF THE PROJECT METHODOLOGY 5](#_Toc67038540)

[7. FUNDING REQUIREMENTS 5](#_Toc67038541)

[8. THE IMPACT OF THE SOLUTION TO STAKEHOLDERS 5](#_Toc67038542)

[9. ethical and legal considerations and precautions that will be used when working with and communicating about sensitive data 5](#_Toc67038543)

[10. YOUR EXPERTISE RELEVANT TO THE SOLUTION YOU PROPOSE 6](#_Toc67038544)

[B.  Write an executive summary directed to IT professionals that addresses *each*of the following requirements: 6](#_Toc67038545)

[1.  THE DECISION-SUPPORT PROBLEM OR OPPORTUNITY YOU ARE SOLVING FOR 6](#_Toc67038546)

[2.  A DESCRIPTION OF THE CUSTOMERS AND WHY THIS PRODUCT WILL FULFILL THEIR NEEDS 6](#_Toc67038547)

[3.  EXISTING GAPS IN THE DATA PRODUCTS YOU ARE REPLACING OR MODIFYING (IF APPLICABLE) 6](#_Toc67038548)

[4.  THE DATA AVAILABLE OR THE DATA THAT NEEDS TO BE COLLECTED TO SUPPORT THE DATA PRODUCT LIFECYCLE 6](#_Toc67038549)

[5.  THE METHODOLOGY YOU USE TO GUIDE AND SUPPORT THE DATA PRODUCT DESIGN AND DEVELOPMENT 6](#_Toc67038550)

[6.  DELIVERABLES ASSOCIATED WITH THE DESIGN AND DEVELOPMENT OF THE DATA PRODUCT 6](#_Toc67038551)

[7.  THE PLAN FOR IMPLEMENTATION OF YOUR DATA PRODUCT, INCLUDING THE ANTICIPATED OUTCOMES FROM THIS DEVELOPMENT 6](#_Toc67038552)

[8.  the methods for validating and verifying that the developed data product meets the requirements and subsequently the needs of the customers 6](#_Toc67038553)

[9.  the programming environments and any related costs, as well as the human resources that are necessary to execute each phase in the development of the data product 6](#_Toc67038554)

[C.  Design and develop a fully functional data product that addresses your identified business problem or organizational need 6](#_Toc67038555)

[D.  Create *each* of the following forms of documentation for the product you have developed: 7](#_Toc67038556)

[1.  A BUSINESS VISION OR BUSINESS REQUIREMENTS DOCUMENT 7](#_Toc67038557)

[2.  RAW AND CLEANED DATA SETS WITH THE CODE AND EXECUTABLE FILES USED TO SCRAPE AND CLEAN DATA (IF APPLICABLE) 7](#_Toc67038558)

[3.  CODE USED TO PERFORM THE ANALYSIS OF THE DATA AND CONSTRUCT A DESCRIPTIVE, PREDICTIVE, OR PRESCRIPTIVE DATA PRODUCT 7](#_Toc67038559)

[4.  ASSESSMENT OF THE HYPOTHeSES FOR ACCEPTANCE OR REJECTION 7](#_Toc67038560)

[5.  visualizations and elements of effective storytelling supporting the data exploration and preparation, data analysis, and data summary, including the phenomenon and its detection 7](#_Toc67038561)

[6.  ASSESSMENT OF THE PRODUCTS ACCURACY 7](#_Toc67038562)

[7.  THE RESULTS FROM THE DATA PRODUCT TESTING, REVISIONS, AND OPTIMIZATION BASED ON THE PROVIDED PLANS, INCLUDING SCREENSHOTS 7](#_Toc67038563)

[8.  SOURCE CODE AND EXECUTABLE FILES 7](#_Toc67038564)

[9.  a quick start guide summarizing the steps necessary to install and use the product 7](#_Toc67038565)

[E.  Acknowledge sources, using in-text citations and references, for content that is quoted, paraphrased, or summarized. 7](#_Toc67038566)

A.  Create a letter of transmittal and a project proposal to convince senior, non-technical managers and executives to implement the data product you have designed. The proposal should include *each* of the following

March 16, 2021

Johnny Quest, Owner & Operator

X-Factor Enterprise Security

10101 Technology Way

Silicon Valley, CA 10101

Dear Mr. Quest,

This past year, X-Factor Enterprise Security has become the de facto platform for small and medium size businesses that require a focus on secure transmission of their data. While secure data transmission is critical to most companies, it doesn’t account for the type of security breaches found in small businesses. Over the last decade, the growth of Big Data has caused a large spike in the data discovery field. While this field has many different purposes that prove useful, it can also open the door for malicious behavior. Because of this, many companies exchange data without realizing the security risk and potential exposure of their customers personal information. Unfortunately, the large amounts of microtransactions involving customer information allow duplicate datasets to be captured by unauthorized users with a disturbing frequency. These unauthorized users are then able to compare other related datasets to find out new things about the exposed customer, thus violating their privacy. This scenario presents itself as a major threat to the big data community and can result in third-party entities possessing the data that the original party did not consent to. Because of this issue, our team has answered the call with a solution that can be quickly and easily integrated to your existing infrastructure.

Our product, the Epsilon-Delta Privacy Solution (ΕDPS), is based on one of the newer fields inside the machine learning community, Epsilon-Differential Privacy (εDP). Perhaps best described by Professor Christopher Brown of MIT University, “The only way to look at εDP is that it allows an analyst to know if a dataset has a noticeable impact on the outcome of a query. If not, then a user might as well contribute to that database.” To put it another way, applying εDP to a dataset shows that everything learned from that same dataset will remain unchanged whether you include yourself or not. This is only achievable with εDP. So, you might ask “what does it mean to apply εDP to a set of data?”. To put it in the fewest words possible, it is the process of taking small amounts of noise and applying it to random fields in the data. The data that the noise is applied to can ultimately be derived from the base data set or another known source. This process guarantees that your data, at least from a statistical viewpoint, doesn’t’ change much and allows authenticity verification of a single record in the dataset.

The objective of ΕDPS is to provide an easily implementable εDP solution that is comparable to flipping a coin. The coin flip approach is the most widely used variant of εDP and will be easy to scale to the needs of your large client base. ΕDPS works by reading a dataset and then asking which column views (aka fields) you want to apply εDP to. A coin flip then occurs for each value, and if heads, then the current value remains the same, and if the coin flip is tails, then the coin is flipped again. If the second flip results in heads, then the current value stays the same, and if the second flip results in tails, the value is replaced with that of another row in the same column. This method guarantees that around 25% of the dataset values are changed in the first dataset selected, with an upper bound of around 3% and lower bound roughly the same. The coin flip method can be applied to a second dataset as well, however, the second time it will be applied one nest level deeper. This will result in the second set of values being changed by about 12-13%, with an upper bound of about 1.5% and lower bound around the same percentage. When looking at a dataset overall, these changes will result in only a small change to the whole dataset. Initial tests show about 5-7% change in the overall dataset.

We understand any skepticism of our solution, but rest assured-- we reached out specifically to YOU because we are certain of your ethical commitment and dedication to customer privacy. EDPS firmly believes that by combing our services with your own you will be the among the very first providers to promote security and premium privacy to both consumer and customer markets. Over the years we have nurtured a strong reputation for creating software that is clean yet efficient, and always puts security and ethics as the top priority. We believe there is unmatched ethical value created by providing inherent data privacy. When a user subscribes to an email list or uses their real address in a form, they shouldn’t need to be worried about who will end up with this data in their hands.

Regarding the financial obligation of our privacy solution, we will need to restructure the demo we have provided to accept more robust data sets and scale it to your needs. To do this, we will break the project into two distinct phases:

1. Initial Deployment
2. Major product update to V1.0

Phase 1 will require $125,000 for developer hardware requirements and should take about 100 development hours to produce a working solution based on existing infrastructure. Phase 2 will require $125,000 and 150 development hours. We believe the upfront financial requirements will be immediately offset upon rolling out the solution, by generating over $250,000 in net income. Total time estimates but project completion at 8 weeks from the decided start date.

We are looking forward to discussing this with you further and hope to hear from you soon.

Best regards,

Joshua Gaweda, CEO EDPS

## 1.  Summary of the problem

X-Factor Enterprise Security currently offers a suite of tools that provide email and network encryption capabilities. The current problem with their suite is that they have no control of how their clients’ employees interact with data. The suite also fails to provide a tool that protects discovery of client data. By adding EDPS to it existing suite of tools, X-Factor Enterprise Security will become a full-service provider of data security.

## 2.  A description of how the data product benefits the customer and supports the decision-making process

EDPS takes existing data and applies noise in small amounts to reduce the validity below 100%. With less than 100% validity, any discovered data cannot be guaranteed accurate. The more data available in the dataset the better the algorithm works, since there is more data to manipulate. With just a single application of εDP, each person in the dataset can have complete privacy. This is possible because there is not a statistical advantage for any person in the dataset if they are removed vs if they are not removed.

The benefit of using EDPS is it will assist in removing company liability for compromised data because of unsecure data transmission. Incredible amounts of data move all over the word every day, and the likelihood that it becomes compromised somewhere along the way is increasingly high. By applying EDPS before data transmission, it ensures that the best precautions have been taken to protect client information. This provides obvious benefits to the customer by mitigating damage in a data breach event and increasing will drive up consumer confidence. Finally, EPDS can assist with decisions regarding the sale and sharing of data and help understand how valid any purchased data is.

## 3. An outline of the data product

## 4. a description of the data that will be used to construct the data product

## 5. tHE OBJECTIVES AND HYPOTHESES OF THE PROJECT

## 6. AN OUTLINE OF THE PROJECT METHODOLOGY

## 7. FUNDING REQUIREMENTS

## 8. THE IMPACT OF THE SOLUTION TO STAKEHOLDERS

## 9. ethical and legal considerations and precautions that will be used when working with and communicating about sensitive data

## 10. YOUR EXPERTISE RELEVANT TO THE SOLUTION YOU PROPOSE

# B.  Write an executive summary directed to IT professionals that addresses *each*of the following requirements:

## 1.  THE DECISION-SUPPORT PROBLEM OR OPPORTUNITY YOU ARE SOLVING FOR

## 2.  A DESCRIPTION OF THE CUSTOMERS AND WHY THIS PRODUCT WILL FULFILL THEIR NEEDS

## 3.  EXISTING GAPS IN THE DATA PRODUCTS YOU ARE REPLACING OR MODIFYING (IF APPLICABLE)

## 4.  THE DATA AVAILABLE OR THE DATA THAT NEEDS TO BE COLLECTED TO SUPPORT THE DATA PRODUCT LIFECYCLE

## 5.  THE METHODOLOGY YOU USE TO GUIDE AND SUPPORT THE DATA PRODUCT DESIGN AND DEVELOPMENT

## 6.  DELIVERABLES ASSOCIATED WITH THE DESIGN AND DEVELOPMENT OF THE DATA PRODUCT

## 7.  THE PLAN FOR IMPLEMENTATION OF YOUR DATA PRODUCT, INCLUDING THE ANTICIPATED OUTCOMES FROM THIS DEVELOPMENT

## 8.  the methods for validating and verifying that the developed data product meets the requirements and subsequently the needs of the customers

## 9.  the programming environments and any related costs, as well as the human resources that are necessary to execute each phase in the development of the data product

# C.  Design and develop a fully functional data product that addresses your identified business problem or organizational need

**[** [**ATTACHED**](Project%20Files) **]**

# D.  Create *each* of the following forms of documentation for the product you have developed:

## 1.  A BUSINESS VISION OR BUSINESS REQUIREMENTS DOCUMENT

## 2.  RAW AND CLEANED DATA SETS WITH THE CODE AND EXECUTABLE FILES USED TO SCRAPE AND CLEAN DATA (IF APPLICABLE)

## 3.  CODE USED TO PERFORM THE ANALYSIS OF THE DATA AND CONSTRUCT A DESCRIPTIVE, PREDICTIVE, OR PRESCRIPTIVE DATA PRODUCT

## 4.  ASSESSMENT OF THE HYPOTHeSiS FOR ACCEPTANCE OR REJECTION

## 5.  visualizations and elements of effective storytelling supporting the data exploration and preparation, data analysis, and data summary, including the phenomenon and its detection

## 6.  ASSESSMENT OF THE PRODUCTS ACCURACY

## 7.  THE RESULTS FROM THE DATA PRODUCT TESTING, REVISIONS, AND OPTIMIZATION BASED ON THE PROVIDED PLANS, INCLUDING SCREENSHOTS

## 8.  SOURCE CODE AND EXECUTABLE FILES

## 9.  a quick start guide summarizing the steps necessary to install and use the product

# E.  Acknowledge sources, using in-text citations and references, for content that is quoted, paraphrased, or summarized.