ICM Report

Contents

1 Intercontinental Cargo Moving Corporation Evaluation

1.1 Background

Intercontinental Cargo Moving Corporation (ICM) operates a large seaport and has hired our team for data and analytics (DA) system evaluation. ICM wants us to measure the maturity of their current DA system and provide a solid plan to optimize their DA capabilities. Using our company's model, ICM hopes to install customer trust and confidence in their data practices. We can strengthen customer trust and confidence by supplying ICM with tools to more effectively manage their people, technologies, and processes.

To provide ICM with the aforementioned tools, our team is tasked with implementing a model and providing a report that will provide:

- Metrics to measure the current DA system maturity level. This should be a calculation of KPIs used to measure the success of their people, technologies, and processes.
- Demonstrations of the model's application and how results from the model could be used to recommend future changes to the system allowing for the company to maximize the potential of their data assets.
- Suggest protocols that ICM can enact to measure their system's effectiveness.
- Demonstrate how the model can be applied to subsequent seaports of varying sizes. We can also analyze how this model can be adapted to other industries, how it could benefit clients of ICM, and how ICM can benefit by interacting with clients using this model.

1.2 Model

1.2.1 People

This section is meant to address the hiring needs and concerns of Intercontinental Cargo Moving (ICM) corporation. Including the implementation of a logistic regression model specifically designed to evaluate current talent and identify key areas of need for future employees for the Data & Analysis (D&A) team. The model being considered will consist of predicting how likely an employee is to be working on a project best suited to their skills and to ICM's needs. The increasing challenges of hiring and evaluating employees has never been more important and similarly, the ability to use models to create an intellectually stimulating and innovative work environment is at the forefront of all companies needs. This section will address the concerns as identified by ICM, including how to evaluate current and future employees, where and what kinds of people to hire, and how to best employ an efficient workforce.

1.3 Introduction

>We begin by examining a framework in which ICM can get a complete picture of how their employees are performing. Namely, the evaluation of current and future talent at ICM will consist of identifying the ratio between cost and time of given projects. That is, given a project, how much is it costing the company versus how much is ICM gaining from the completion of the project. This can be revenue or experience gained. Cost in this sense is referring to both the financial cost, as well as cost in terms of number of employees assigned to a project. In other words, the goal is to identify projects that are high cost and high time, and make adjustments that reduce cost/time based on which employees are working on the project. High cost, high time project will often yield significantly worse results then projects that are low cost and low time, which in turn, effects the overall productivity of the company. Ideally, all projects would be designated low cost and low time. Again, low cost in this sense, does not strictly refer to financial cost, but to an optimization of having the right employees working on the right projects. To achieve the goal of optimizing ICM's workforce, there must be a plethora of data in which employees can be accurately regarded. For the sake of simplicity and clarity, the initial model will consists of five key skill identifiers. Including, general/specific skills: Data Analysis, Quantitative reasoning, Programming ability, Project Management, and Communication skills. Additionally, years of experience will be a parameter to consider, and finally, is the employee actively involved in a low cost low time project. This is subject to change, and clearly, somewhat subjective, which is intended. By identifying what projects are currently viewed as low cost and time, as well as who is working on those projects, the skills that are involved in ensuring the success of those projects can be leveraged in such a way that all employees can be assigned to projects that would increase the likely hood of the project either continuing to be or moving to being a low cost and time project.

1.4 Model

>The logistic regression model is useful when attempting to predict the probability of an event occurring in either a True or False scenario, which will be denoted as the response variable. In this case, the model will be used to determine the probability of an employee being involved in a low cost and low time project, so the response variable is involved in 'low cost low time project' or not. The same model can be used for predicting employees involved in high cost high time projects. The goal in doing this, as mentioned previously, is to identify which persons are generating revenue and boosting productivity as opposed to employees who are not. This is both to highlight the abilities of a number of employees whilst exploiting weaknesses in projects or employees. This will then allow for adjustments to be made, training to be implemented, and general needs to be met. The general equation of logistic equation is represented as:

$$\log[\frac{p(X)}{(1-p(X))}] = \beta_0 X_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

Where X_j is the j^{th} predictor variable and β_j is the coefficient estiamte for the j^{th} predictor variable, for $1 \leq j \leq n$.
$ $
The formula on the right side of the equation predicts the log odds of the response variable taking on a value of 1.

Thus, when we fit a logistic regression model we can use the following equation to calculate the probability that a given observation takes on a value of 1:

$$p(x) = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_p X_p}}{(1+e^{\beta_0 x_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_p X_p)}}$$

We then use some probability threshold to classify the observation as either 1 or 0. For example, given a data set with 40 people, the model used would output something that predicts with approximately 70% accuracy an employee being involved in a low cost low time project. From here ICM can examine what skills that employee has, as well as how man years of experience they have. Now, it should be emphasized that this model will improve with more data, that is, the more parameters that are used to associate employees with successful projects the better the model will be at identifying high or low need areas.

1.5 Outcomes

>Once areas of concern are addressed and adequate skills are identified. The next step is to hire employees that fill the gaps. It is recommended that ICM immediately begins an internship program. From the model, as seen above, years of experience correlate to more successful projects. This means that by starting an internship program ICM can gain a competitive advantage by training undergraduates or recently graduates and exposing them to employees that have a high probability of being involved in low cost low time projects, so by the time they are ready to being working full-time, they already possess some of the key skills ICM is looking for. How many individuals that are hired should be proportional to the number of low cost low time projects that ICM is working on. For example, by hiring too many people it is likely that the risk of inflating low cost low time projects increases, and conversely for not hiring enough people. SO the general guideline that would be suggested is to continually check the model to ensure that if a project is shifting toward being high cost high time, it may be time to hire more people, and similarly, by removing people from projects, if the project remains as a low cost low time project, then it is likely that too many people are involved on the project. \ > Aside from universities, getting involved in conferences and looking for people that specialize in certain tasks will be beneficial, as well providing context for contracting out positions versus full-time hiring. For example, suppose there exists a project that is highly quantitative and necessitates programming skills. It would clearly be advantageous to hire somebody who already possesses those skills and can efficiently turn a high cost high time project into a low cost low time project. However, it could be that ICM only has a small number of these projects at a given time, where hiring a full-time employee increases the risk of inflating projects and creating bloat. This would be a good time to contract-out these temporary positions without compromising the integrity of the model. Instead, focusing attention on projects that need full-time employees.

1.6 Summary

>Addressing the complex and comprehensive needs of a corporation like ICM must rely on the use of data-driven approaches. By modeling the various projects ICM is involved in using cost and time, ICM can identify key skills and characteristics of their employees and use that information to make data informed decisions. It should also be noted that not all employees share the same skills, so by moving an employee who has high programming or quantitative skills to an area that is in need of those skills the chances of improving the productivity of the workplace increases. This also exposes employees to a variety of different skills and serves as a "On-the fly" training experience. Hiring, aside from the usual outlets, should consist of locating talented individuals through an internship program or university/conference setting to find specialized talent they can outsource for optimal efficiency. Training can be done on the job when employing a model that specializes in putting the right people on the right projects. Therefore, by utilizing a logistic regression model, as well as some variation of skill, as determined by what projects are having success, ICM can create an environment that will not only benefit themselves, but their employees by putting them in positions to succeed.

1.6.1 Technology

1.6.1.1 Criteria Since the framework for evaluating technical solutions is more important than choosing them at this time, specific criteria need to be established as metrics to measure the effectiveness of a specific solution, now or in the future. The six criteria measured in this model are as follows, with Feature Match having the highest weighting in the final calculation:

- Price
- Support
- Ease of Use
- Feature Match
- Trial Options
- Accessibility (operating system availability, options for contrast, etc.)

For each solution, the criteria are evaluated and given a score from 0 to 100 with 100 being the highest score. Consideration of what is considered "good" for a metric is also taken into account. For example, a low priced

solution with good support would have high scores in both categories. The scores are input into the program which displays the scores on a radar chart separated by solution. Additionally, a weighted score is calculated for each solution.

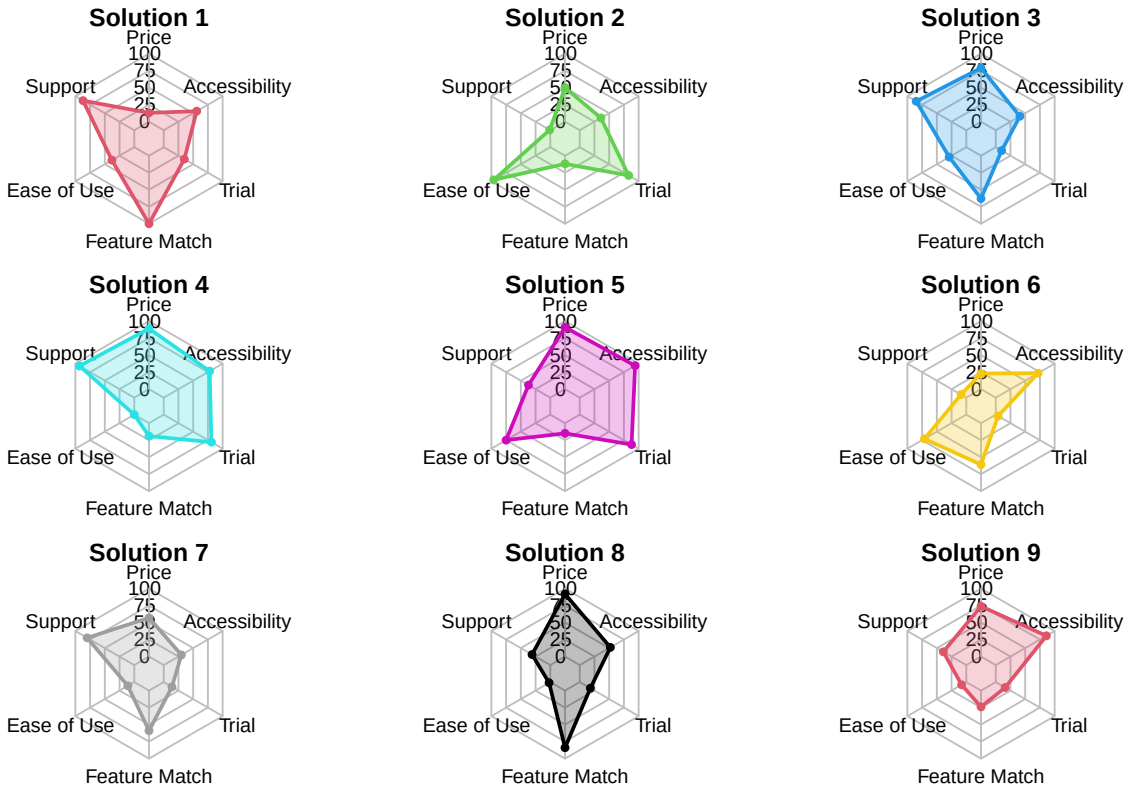
The model can work for physical and digital forms of technology and can indicate if multiple products are needed. For example, if two products fulfill different aspects of a needed solution, but both have good overall scores, then they might both be a good fit to use together.

The following code and diagrams are examples made with simulated data and do not reflect real technical solutions.

1.6.1.2 Initial Pass

The first step is to display all the proposed solutions next to each other to create an initial comparison. This provides an overview of all possible options before the list gets trimmed down.

```
set.seed(12345)     # Sets seed for repeatable results
num_prog <- 9       # Sets number of solutions

categories <- c("Price", "Support", "Ease of Use",
                "Feature Match", "Trial", "Accessibility")
solution_names <- c("Max", "Min", paste("Solution", 1:num_prog))
size <- length(categories)

# Creates data frame with score data. Uses placeholder random data
tech_analysis <- data.frame(rbind(rep(100, size), rep(0, size),
                                  matrix(sample(0:100, size * num_prog), nrow = num_prog)))

# Sets column and row names
colnames(tech_analysis) <- categories
row.names(tech_analysis) <- solution_names

# Sets margins to allow for multi-graph display
op <- par(mar = c(1, 1, 1, 1))
height <- ((nrow(tech_analysis) - 2) / 3)
par(mfrow = c(height, 3))

colors <- 2:10

# Loops through all rows except first 2
for (i in 3:nrow(tech_analysis)) {
  fmsb::radarchart(tech_analysis[c(1, 2, i),],               # Creates radar chart
                   cglty = 1,                                # Grid line type
                   cglcol = "gray",                          # Grid line color
                   pcol = colors[i-2],                       # Color for each line
                   plwd = 2,                                 # Width for each line
                   plty = 1,                                 # Line type for each line
                   pfcol = scales::alpha(colors[i-2], 0.25), # Fill color
                   title = rownames(tech_analysis)[i],       # Set titles
                   axistype = 1,                             # Set axis type
                   axislabcol = "black",                     # Set axis color
                   caxislabels = seq(0, 100, 25))            # Set axis labels
}
```




```
par(op)
score <- c()

cat_weights <- c(0.05, 0.2, 0.15, 0.5, 0.05, 0.05)    # Sets weights for each category

# Calculates weighted total score for each row
for (i in 3:nrow(tech_analysis)) {
  row <- i
  score_list <- as.numeric(tech_analysis[row, ])
  weighted_score <- sum(score_list * cat_weights)
  score <- c(score, weighted_score)
}

# Creates new data frame with weighted scores
solution_data <- cbind(tech_analysis, "Eval" = c(-1, -1, score))

# Displays solutions with all data including the weighted score
table1 <- solution_data[3:nrow(solution_data),] %>%
  tibble::rownames_to_column("Solutions") %>%
  flextable::flextable() %>%
  flextable::theme_zebra() %>%
  flextable::autofit(add_w = 0, add_h = 0)
table1
```

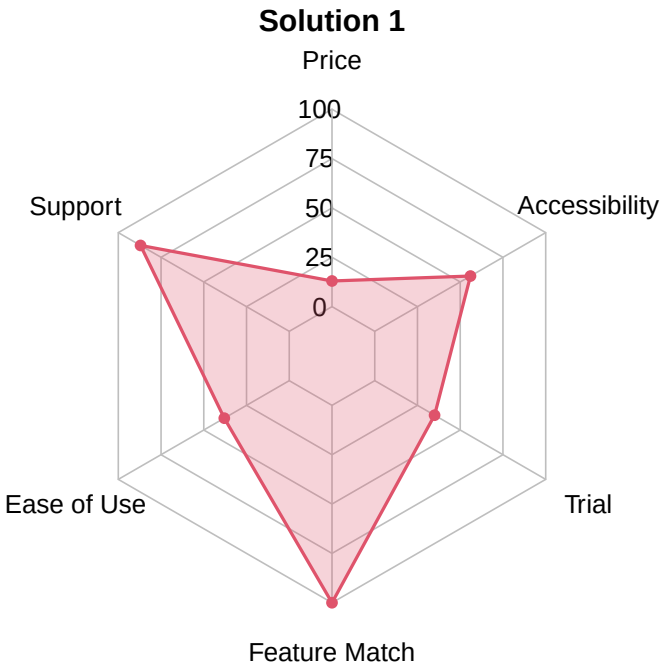
Solutions	Price	Support	Ease of Use	Feature Match	Trial	Accessibility	Eval
Solution 1	13	87	38	100	35	56	78.30

Solutions	Price	Support	Ease of Use	Feature Match	Trial	Accessibility	Eval
Solution 2	50	1	96	12	83	36	29.05
Solution 3	79	85	29	63	10	41	59.35
Solution 4	89	93	0	19	81	78	40.50
Solution 5	91	37	75	15	88	94	39.80
Solution 6	23	9	71	61	4	72	47.90
Solution 7	57	80	11	59	14	30	52.20
Solution 8	92	31	2	84	18	52	56.60
Solution 9	74	39	8	24	16	86	29.80

A viewer can easily use the charts to compare the strengths and weaknesses of each solution or they can use the table for a more detailed breakdown of each solution. By displaying the data in this fashion, incompatible or unnecessary tools can be dismissed quickly, resulting in less cluttered decision making.

1.6.1.3 Filtering After all solutions are evaluated and displayed, the data set is filtered for solutions with high evaluation scores. The program is currently configured to pull solutions with a weighted score of 75 or higher.

```
# Filters out all solutions with a weighted score >= 75 and removes score column
high_scores <- solution_data %>%
  dplyr::filter(Eval >= 75 | Eval == -1) %>%
  dplyr::select(-Eval)

# Sets margins to allow for multi-graph display
op <- par(mar = c(1, 1, 1, 1))
height <- ceiling((nrow(high_scores)-2) / 3)
par(mfrow = c(height, 1))

colors <- 2:10

# Loops through all rows except first 2
for (i in 3:nrow(high_scores)) {
  fmsb::radarchart(high_scores[c(1, 2, i),],                  # Creates radar chart
                   cglty = 1,                                 # Grid line type
                   cglcol = "gray",                           # Grid line color
                   pcol = colors[i-2],                        # Color for each line
                   plwd = 2,                                  # Width for each line
                   plty = 1,                                  # Line type for each line
                   pfcol = scales::alpha(colors[i-2], 0.25),  # Fill color
                   title = rownames(high_scores)[i],          # Set titles
                   axistype = 1,                              # Set axis type
                   axislabcol = "black",                      # Set axis color
                   caxislabels = seq(0, 100, 25))             # Set axis labels
}
```




```
par(op)

# Displays solutions with all data including the weighted score
table_frame <- solution_data %>% filter(Eval >= 75 | Eval == -1)
table2 <- table_frame[3:nrow(table_frame),] %>%
  tibble::rownames_to_column("Solutions") %>%
  flextable::flextable() %>%
  flextable::theme_zebra() %>%
  flextable::autofit(add_w = 0, add_h = 0)
table2
```

Solutions	Price	Support	Ease of Use	Feature Match	Trial	Accessibility	Eval
Solution 1	13	87	38	100	35	56	78.3

1.6.1.4 Summary Shrinking the data set to only a few options, or in this case one, allows for a better look at this solution. In this example, we can see that this tool has all the features needed and has great support even though it might be expensive. Additionally, it has decent accessibility and seems that it would not need a copious amount of training to use. A trial period might not be as necessary in this case because the solution fits enough criteria.

With these systems and visualizations in place, it is easy to see how different technical solutions compare to each other, as well as where they excel and what they lack. By using this system, the Information Technology (IT) department at ICM can make informed decisions about technical solutions and prioritize what criteria are most important to the company.

1.6.2 Process

1.6.2.1 Data Governance Now that we've discussed methods to evaluate the previous components of ICM's DA system, we can dive into the processes that ICM should follow to ensure a strong data governance program. As mentioned within the problem guidelines, a good data governance program provides oversight of data resources, access levels, modifications to data sources and tables, and follows consistent protocols.

Processes that encourage strong data governance are supported by the people within the department and the technologies employed by the company. In these situations, its required that ICM utilizes an ERP that helps manage data by providing a single source of truth.

To prescribe a plan of action to ICM's Information Security Officer (ISO), we first start by outlining proper structure and content of data sources to eliminate any inconsistencies between independent data sources derived from port operations, transportation schedules, customs inspections, and container storage. We want to make sure that operations data is maintained and easily accessible to our DA systems team when we need to gather insights from our activity. Additionally, its important to plan for company longevity since our data will grow considerably with company growth. As our data continues to increase with time, it should remain just as manageable and easily accessible by the DA systems team. In the event of an error or oversight the data collected and utilized by ICM should also follow protocols for data archival. Archiving operations data ensures that if something bad happens there is minimal downtime and data recovery is quick.

We can evaluate ICM's current program by measuring their current ability to access and gain consistent insights from their operations data:

- It's given that independent data processes are creating inconsistencies between operations data sources. Because of the inconsistent data, reports that are derived from each data set will result in different figures leading to visibility of operations activity being impossible. This is an area that needs to be promptly addressed, otherwise management will lack the guidance needed to run the company successfully.
- We're currently unaware of any protocols that are in practice by the DA systems team that would enforce strong internal controls relating to modifying and maintaining data sets among each area of operations. To implement a strong data governance system its important that the team follows proper protocols so that data is correctly modified, maintained, and archived. Tables should be used to communicate the activity within each area of operations so that managers can view a single status report with accountable figures.
- There current process does not include protocols for data management throughout the data lifecycle. We can ensure the longevity of the data as the company grows by enacting processes that ensure we have the capacity to store and access big data generated by ICM.
- Finally, we want to make sure that there is a clear structure for the DA systems team so that staff roles are clearly defined and individuals have the appropriate access levels. With data governance in mind, ICM should enforce a least-privileged system where data engineers are given the ability to maintain and modify data sources. Other staff roles such as data scientists, data analysts, and machine learning engineers should only have the ability to access the data for analysis. These permissions will help reduce unwanted or unforeseeable side-effects from inexperienced users, therefore reducing the possibility of downtime.

Prescriptions for a strong data governance program would be a direct response to pitfalls outlined in the previous assessment. We would strongly encourage ICM to generate data using the capabilities offered by their ERP and collect that subsequently generated data for analysis of operations. By accessing the data from a single source of truth and following strict data engineering processes to ensure that the data is maintained, data analysts and data scientists within the DA systems team can produce reports of accountable figures to management for visibility of operations system status. Additionally, strict protocols for access levels, and archival will provide minimal downtime for data recovery. Finally, we can suggest that data is maintained and flexible for longevity as the company continues to grow.

Within the following section we show how a system following a single source of truth can be used to track incoming and outgoing freight that the port will interact with while in operation. This data would live within properly managed data containers in data warehouses set up by the data engineering staff.

We create two tables that track imports and exports so that we can provide management with insights as to which containers are entering and exiting the port. This would allow us to track the number of containers we are currently holding, when they should be scheduled to leave the port, whether or not they should be going through customs, and the owner of the containers.

Table 3: Import Dimensio

transaction_id	transaction_type	client_id	client_name	num_containers	container_ids	i
A-001-45A-000371	Import	A-001	Metal Manufacturers, Inc.	3	45A-000371	2
A-001-45A-000372	Import	A-001	Metal Manufacturers, Inc.	3	45A-000372	2
A-001-45A-000373	Import	A-001	Metal Manufacturers, Inc.	3	45A-000373	2

An example of import data that could be contained within the appropriate database container.

Table 4: Export Dimensional Table

transaction_id	transaction_type	client_id	client_name	num_containers	container_ids	c
A-001-45A-000371	Export	A-001	Metal Manufacturers, Inc.	2	45A-000371	2
A-001-45A-000372	Export	A-001	Metal Manufacturers, Inc.	2	45A-000372	2

An example of export data that could be contained within the appropriate database container.

Similarly, we can audit where specific containers are transported after we receive them. Using a dimensional table that tracks where containers are relocated after we receive them, or before and after going through the customs process would be a useful tool to audit which containers we still have in our possession and where. To illustrate this idea, we include a matrix of hypothetical grid locations within the port used for storing these containers.

```
## `dim_port_transportation` looks at the transportation schedules of the items within the port.
## storage_location should be a function of schd_export_date so that we're storing items in the appropri
v <- c(paste0("A", 1:6), paste0("B", 1:6), paste0("C", 1:6)) %>%
  matrix(., ncol = 3, byrow = TRUE)

dim_port_transportation <-
  dplyr::tibble(
    transaction_type = "Relocate",
    client_id = "A-001",
    client_name = "Metal Manufacturers, Inc.",
    num_containers = 3,
    container_ids = c("45A-000371", "45A-000372", "45A-000373"),
    transport_date = "2022-02-20",
    transport_from = "A1",
    transport_to = "B3"
  ) %>%
  dplyr::add_row(
    transaction_type = "Relocate",
    client_id = "A-001",
    client_name = "Metal Manufacturers, Inc.",
    num_containers = 2,
    container_ids = c("45A-000371", "45A-000372"),
    transport_date = "2022-02-26",
```

```
    transport_from = "B3",
    transport_to = "A1"
  ) %>%
  dplyr::mutate(
    transaction_id = paste0(client_id, "-", container_ids)
  ) %>%
  dplyr::relocate(transaction_id)
```

```
## `fact_operations` allows managers to gain visibility into the daily operations occuring within the p
fact_operations <-
  dim_imports %>%
  dplyr::full_join(
    dim_exports,
    by = c(
      "transaction_id",
      "transaction_type",
      "client_id",
      "client_name",
      "num_containers",
      "container_ids"
    )
  ) %>%
  dplyr::full_join(
    dim_port_transportation,
    by = c(
      "transaction_id",
      "transaction_type",
      "client_id",
      "client_name",
      "num_containers",
      "container_ids"
    )
  )
```

1.6.2.2 Metadata Metadata is equally as important as some of the items outlined in the previous data governance section. To reduce the time spent by members of the DA system team investigating the use and contents of data sets, we can provide metadata that explains the use and contents of the data. ICM can create effective metadata by drafting documentation that lists each table contained within a database, their uses, and which area of the business they pertain to. Within each table of the database, ICM should create a list of all columns of that table, the data types of each column, the primary keys associated with that table, and the contents that each column represents. Strong documentation can assist with the longevity and archival of data as tables outlive their usefulness, are superseded or deprecated, or become irrelevant if the data contained within the table is no longer needed for analysis.

1.7 Letter to Interested Parties