Canadian Interview Question – Scenario

Zirong Liu

How would you go about the following?

1. Coordinating efforts with the PMO

1.1 Schedule a kick-off meeting with the PMO (Program management office), acquiring the essential information such as:

- the problems need to be solved.
- the scope/objective of the project.
- the information PMO could provide.
- creating a project document that records the discussed details of the project.

1.2 After having an overview of the project's timeline, setting up a Gantt chart with mini-deadlines.

1.3 Setting up weekly/periodic follow-up meetings to better update the process or exchange information with the PMO. If the objective has been altered, these meetings will help adjust the project's direction in a timely manner.

2. Improving the data extraction and gathering process

2.1 Firstly, I would try to understand the current data extraction and gathering process, understand what the data is about, observe the correlation between data variables, check patterns contained within historical data.

2.2 Research relevant techniques which could improve the current data extraction and gathering process. I would use SQL to shorten the steps required to extract data or use specific APIs provided by different platforms/websites to replace manual extraction.

For example: when creating MS Power BI reports, the analyst could use SQL in Power Query to improve the data extraction process. In addition, governments/companies' websites would provide specific APIs for the users to acquire datasets, such as Statistics Canada's API, Google API.

2.3 There are a few things I would need to think before performing data extraction: what is the information clients/projects need, what variables in the data sets can provide useful insights, what are important metrics, identify the data sources or are there better data sources I could use, are the data extraction process being clients/projects focused, what is the frequency of extracting the data.

2.4 Another extra step that could be taken is implementing an accuracy standard for the datasets. Here, some statistics knowledge could be used. For example, use python Pandas/NumPy or R to run a few tests such as standard deviation ("sd" function), regression ("lm" function for the linear regression), hypothesis testing ("t.test" for T test) to see which variables are helpful, or employing means and sample size determination to see if current acquiring datasets have good quality.

3. Automating the data consolidation

3.1 I would automate the data consolidation by using the built-in functions provided by the required software.

3.2 For example, to consolidate data in excel, I would operate the following steps.

a) Select an empty sheet in the workbook as the master worksheet.
b) Use the consolidate built-in function in the data ribbon section. For this step, I could choose different consolidate built-in functions that Excel provides to consolidate the data. For example, Sum consolidate function. After putting every highlighted data in the reference consolidation dialogue, I would click ok to let the function run.

c) Copy the selected worksheet into a new workbook.
d) Before doing automation data consolidation, I would check out YouTube videos and forums to make sure I have no gap with the relevant knowledge. When running into issues I have not faced before, I would google and go through forums as most problems have already been solved/answered by others.

3.3 For example, to consolidate data in smartsheet, I would operate the following steps.

a) After researching, I found that Row Report in Smartsheet would aggregate row information from multiple sheets. The first step would be going into Solution Center and creating a row report directly.
b) Select criteria to include in the report. There is "Where?" Button to select the sheets and other buttons to help me to build target row report.
c) I need to choose which information could show in the row report based on the clients'/projects' needs.

d) I could preview the report, with two options as Gantt view and Calendar view.
e) Before doing automation data consolidation, I would check out YouTube videos and forums to make sure I have no gap with the relevant knowledge. When running into specific issues I have not faced before, I would google and go through forums as most problems have already been solved/answered by others. (For example: the troubleshoot issues sheet provided by the official website of the Smartsheet.)

4. Ensuring data quality and consistency

4.1 Understand the origin and timeline of the data. For example, the meaning of each variable, how the variables correlated with each other, what these variables mean for the final product, what are the data types for each variable: numbers or characters, is the data updated to the latest, etc.

To elaborate, communicate with the team (perhaps the PMO for this position) to see what are the sources of the data. Then using meta data to understand what the data is about. (Take the case study test's Excel sheet as example, I will try to figure out how many employee types are there) Put the data set in the MS Power BI, check if "Gender" column is "character" type.

4.2 Check the relevancy of the data. For example, I will eliminate variables for vegetables when the intended use of project is for animals.

4.3 Understand the margin of error of the data.

From random sampling surveys, there is the degree of errors existing. Checking websites to identify the average margin of error of the target survey samples, then compare with the survey samples/data we have to see if our data's margin of error falls into the average.

The calculation: check the sample size first, then use t-score or test out standard deviation. Then check population parameters to get standard error. Then use the critical value gained from running t-score to multiple standard deviation or stand error.

4.4 Set or check metadata (information on data) measurements.

When the data comes from multiple sources and is saved for future use, different departments and analysts may misinterpret the data. Hence, understanding/checking/setting a standard metadata measurement is important to prevent misunderstandings in interpreting data.

4.5 Fixing problems in the data.

The dataset might be required to suit the following standards:

Remove gaps within the data; delete duplicate data entries; change formats of the data; eliminate noise variables; fix typos; delete nicknames; check the format of the variables (for example: postcode variable has a particular format.) Therefore, it is crucial to filter out these biases using tools provided by various software such as Excel, Python Pandas, etc.

For example, in Excel, I could use the following functions:

- Use "Trim" function to reduce space between texts.
- Used "Conditional formation Duplicate Values" to remove duplicate rows.
- Use "Conditional formatting, new rules, cell value is Errors" to remove error values.

4.6 Normalizing the data to make sure the data is consistent.

For example, set one currency for the data collected from different countries. Check if the data from different departments have the same data calendar formats.

4.7 Operating a data quality report

Record the fixed problems, record the finding from data profiling, etc.

Another important step is communicating with the relevant teams regarding the problems they have faced/fixed/identified.

4.8 Automating repetitive tasks could save a lot of time for the future.

For this step, I could use software such as VBA or other software to automate maintaining the data quality and consistency.

5. Creating a visual representation in power BI to tell the data story of the data you have prepared and the data insights you want to communicate to leadership

I have a MS Power BI report regarding the office space optimization as an example contained in the following link. (The report is created from Power BI Desktop version)
GitHub Link: https://github.com/zirongl/zirongl

5.1 Use "Import Excel/CSV/Java/Azure SQL Database " function to import data tables into the software MS Power BI. I could also use "Direct Query" to direct using data which stored in the database. The advantage of the previous is the Power BI report could run faster. The advantage for the "Direct Query" is I would not need to update/refresh the dashboard to often. As long as the data stored in the database is refreshed, my Power BI report's data is refreshed.

5.2 Fixing issues contained within the data: the blanks, data table formats, the date format, etc. I would do this step in the Power Query as it is easier.

5.3 Understanding the data, checking the relationships between the variables of data tables. Using the built-in functions to connect separate data tables and variables.

For example, connecting the "user ID" in data table 1 and data table 2 (both are raw data tables extracted from the database) in the model view.

5.4 Brainstorming

What data insights/aspects of the data the clients/leadership would be most interested in.

What visuals could help the clients/leadership to understand the data more directly/easily/visual comfortably.

What visuals should I use: bar chart, maps, dual-axis chart, scatter plot, bubble chart, etc. Pie chart can show relative proportions of multiple data variables.

Map in the Power BI could directly present the level of aggregate values distributed around the world/provinces.

5.5 Experimenting with visuals and possible combinations of variables.

5.6 Using M language/Dax language adds extra information/variables to provide richer information for the clients/leadership. For example, using Dax language adds a new measure such as month over month increase (for month over month increase, I could just use “Quick Measure” built-in function)(or “calculate(parallelperiod())” Dax function).

5.7 Considering whether to add custom visuals provided by the official website, or add custom visuals by using python/R scripts.
For example, I could use Python Pandas Lambda function to add one extra variable for the data tables.

5.8 Troubleshooting the possible problems found while building the dashboards.
For instance, in a previous report I made, I found one line visual always show blanks when using certain variables. After use Google and check the data from “Data View”. The problem has be resolved by duplicate the variables with “character” format.

5.9 Rearranging the visuals to make the entire dashboard being more user-friendly.
For example, set up a cross-filter so the user could only manually adjust filters in the first page, and the rest pages are filtered at the same time.

6. Implementing and promoting the change on the users’ end

6.1 Based on the users’ identity, adjust the focus of the final product. For example, if the final product is a dashboard produced for the leadership, as the leadership need to consume rich information within a short period, the final product (the dashboard) would be better to contain a lot of visuals.
If the final product is a report for the analyzing team, it would be nicer to use bullet points summarizing key numbers.
If the final product is required to be uploaded on the official website for the public, I might need to communicate with the relevant team, and adjust the size/style of the final product.

6.2 Checking how the product is currently being implemented on the users’ end.

6.3 If there was no current product, consult the supervisor/client teams/documentation to see how previous similar products are implemented.

6.4 Ensuring the final product is easy to use for the users.

6.5 Being consistent with the styles, spacing, design elements of the final products.

6.6 Paying attention to the feedback from the users after setting up a feedback box.

6.7 Updating the products and eliminating confusion from users in a timely manner.

6.8 Checking the visits of the products through the website admin portal.

6.9 If it was needed, build a feedback survey (a few things need to aware: define a clear goal, don't make the survey too long, only use closed-ended questions, etc.)

6.10 Providing training to the users by recording YouTube videos/workshops/MS PowerPoint/meetings.

Thank you for your time to go over my answers. I am looking forward to discussing more details with you.