

➤ Differences Between Retrieve Superstep, Assess Superstep, Process Superstep, Transform Superstep

Retrieve Superstep: This method is used to import external data sources into the processing ecosystem, particularly a data lake with various data^[1].

Assess Superstep: Involves actions like dropping columns with missing values, keeping rows that contain specific elements, and other data assessment tasks^[2].

Process Superstep: The Process Superstep in data science involves adapting the assess results of the retrieve versions of data sources into a highly structured data vault. This structured data vault forms the basic data foundation for further analysis and processing in the data science workflow^[1].

Transform Superstep: This step allows data scientists to take data from the data vault and formulate answers to questions raised by investigations^[4].

Data Science Process:

🔗 *Roots of Data Science:*

🔗 Data science is at its core about curiosity and inquisitiveness.

🔗 This core is rooted in the 5 Whys.

🔗 The 5 Whys is a technique used in the analysis phase of data science.

🔗 Benefits of the 5 Whys:

🔗 The 5 Whys assist the data scientist to identify the root cause of a problem and determine

the relationship between different root causes of the same problem.

🔗 It is one of the simplest investigative tools—easy to complete without intense statistical

analysis.

🔗 When Are the 5 Whys Most Useful?

🔗 The 5 Whys are most useful for finding solutions to problems that involve human factors

or interactions that generate multilayered data problems.

🔗 In day-to-day business life, they can be used in real-world businesses to find the root

causes of issues.

❓ **How to Complete the 5 Whys:**

- ❓ Write down the specific problem. This will help you to formalize the problem and describe it completely.
- ❓ It also helps the data science team to focus on the same problem.
- ❓ Ask why the problem occurred and write the answer below the problem.
- ❓ If the answer you provided doesn't identify the root cause of the problem that you wrote down first, ask why again, and write down that answer.
- ❓ Loop back to the preceding step until you and your customer are in agreement that the problem's root cause is identified.
- ❓ Again, this may require fewer or more than the 5 Whys.

4.1.12 5-Whys Technique

Data science is at its core about curiosity and inquisitiveness. This core is rooted in the 5Whys.

The 5 Whys is a technique used in the analysis phase of data science.

4.1.12.1 Benefits of the 5 Whys

The 5 Whys assist the data scientist to identify the root cause of a problem and determine the

relationship between different root causes of the same problem. It is one of the simplest

investigative tools—easy to complete without intense statistical analysis.

4.1.12.2 When Are the 5 Whys Most Useful?

The 5 Whys are most useful for finding solutions to problems that involve human factors or

interactions that generate multi-layered data problems. In day-to-day business life, they can be

used in real-world businesses to find the root causes of issues.

4.1.12.3 How to Complete the 5 Whys?

Write down the specific problem. This will help you to formalize the problem and describe it

completely. It also helps the data science team to focus on the same problem. Ask why the

problem occurred and write the answer below the problem. If the answer you provided doesn't

identify the root cause of the problem that you wrote down first, ask why again, and write down

that answer. Loop back to the preceding step until you and your customer are in agreement that

the problem's root cause is identified. Again, this may require fewer or more than the 5 Whys.

5 Whys Example:

❓ Problem Statement: Customers are unhappy because they are being shipped products

that don't meet their specifications.

❓ 1. Why are customers being shipped bad products?

❓ • Because manufacturing built the products to a specification that is different from what

the customer and the salesperson agreed to.

❓ 2. Why did manufacturing build the products to a different specification than that

of sales?

❓ • Because the salesperson accelerates work on the shop floor by calling the head of

manufacturing directly to begin work.

❓ An error occurred when the specifications were being communicated or written down.

❓ 3. Why does the salesperson call the head of manufacturing directly to start

work instead of following the procedure established by the company?

❓ • Because the "start work" form requires the sales director's approval before work can

begin and slows the manufacturing process (or stops it when the director is out of the

office).

❓ 4. Why does the form contain an approval for the sales director?

❓ • Because the sales director must be continually updated on sales for discussions with the

CEO, as my retailer customer was a topten key account.

❓ In this case, only four whys were required to determine that a non-value-added signature

authority helped to cause a process breakdown in the quality assurance for a key account.

❓ Fishbone Diagrams:

❓ The diagram is drawn up as you complete the 5 Whys process, as you will discover that

there are normally many causes for why specific facts have been

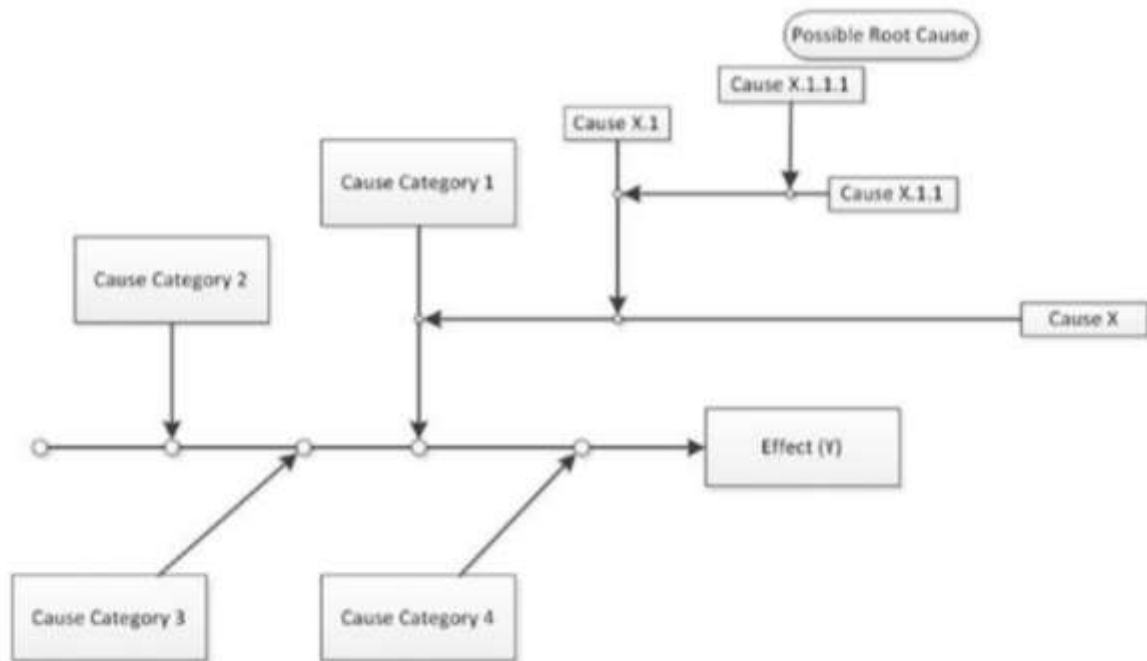


Figure 4.1-8. Fishbone diagram

recorded.

❓ The ten cans are the effect (Y), but the four root causes of the purchase are

❓ 1) I was hungry, so I bought ten tins. I did not like the brand of curry that I bought 10

cans of the previous week.

❓ 2) My neighbor needed five cans, as she was no longer able to walk, and she requested

the brand that I purchased.

❓ 3) I fed two cans to the dog, because I feel dog food is not nutritious, but I was not

prepared to buy a more expensive brand of canned beef curry for the dog.

☐ 4) I put three cans in the charity bin outside the local school.

What's good about fishbone diagrams?

- ❑ The fishbone diagram strives to pinpoint everything that's wrong with current market offerings so that you can develop an innovation that doesn't have these problems.
- ❑ Finally, the fishbone diagram is also a great way to look for and prevent quality problems before they ever arise

4.1.19 Data Science

Shorter Version.....

Basic Steps:

1. Start with a question, addressing the 5 Whys.
2. Follow a good pattern to formulate a model.
3. Gather observations and generate a hypothesis.
4. Use real-world evidence to judge the hypothesis.
5. Collaborate early and often with customers and subject matter experts.

Data Science work best when approved techniques and algorithms are followed.

After performing various experiments on data, the result must be verified and it must have support.

Data sciences that work follow these steps:

Step 1: It begins with a question.

Step 2: Design a model, select prototype for the data and start a virtual simulation. Some

statistics and mathematical solutions can be added to start a data science model.

All questions must be related to customer's business, such a way that answer must provide an insight of business.

Step3: Formulate a hypothesis based on collected observation. Based on model process the observation and prove whether hypothesis is true or false.

Step4: Compare the above result with the real-world observations and provide these results to real-life business.

Step 5: Communicate the progress and intermediate results with customers and subject expert and involve them in the whole process to ensure that they are part of journey of discovery