Current State Analysis of Your Data – Part 2 – Data Freshness

Published: February 2, 2022 | 12:11 am · Vanessa Lam

This article is the second in a series taking a deep dive on how to do a Current State Analysis on your data. (see first article here) This article focusses on Data Freshness: what it is, why it's important, and what questions to ask to determine its current state.

The questions are organized by stakeholder group to facilitate usability; hopefully you can use this as a template to start your Current State Analysis journey. A few definitions before we begin – note that these groups are not mutually exclusive:

People who Input Data: These are people who collect and/or input data into the system. For example, sales people inputting their sales numbers, or survey creators.

People who Manipulate and Analyze Data: These are people who organize the data and create analyses. This includes Data Engineers, Business Intelligence Professionals, and Data Analysts.

People who Make Decisions based on Data: These are the people who use the data to make decisions. This may be a sales manager deciding where to invest resources, a product manager understanding product use demographics, or an executive trying to cut costs.

What is Data Freshness?

Data Freshness refers to whether data is available when you need it to be. This includes the cadence of data input and data refreshes as well as the reliability of processes and tools. Data Freshness issues may be

introduced at any point of the process: during data collection, at any point in the ETL or data pipeline process, or even at the display stage. If a report relies on salespeople to input their time data weekly and some people have not yet filled out their timesheets, this would be an example of data collection delays. Data pipeline or ETL delays can be introduced in many ways, including buggy code, tool or server failures, or other (often technical) issues. The display stage, such as BI tools or Excel worksheets, can cause issues if the underlying data has changed and the dashboard doesn't update, or occasionally because there are manual steps that need to be performed before data can be displayed.

Fixing Data Freshness is usually about ensuring the process is being correctly followed, both by people and in the technical implementation. It is often helpful to write out the entire data workflow, from all the data input sources to each tool and update process the data undergoes before it is displayed. This can help find bugs and make quality checking your data easier because you can check each stage and isolate where the data differs from expectation.

Why is Data Freshness Important?

Data Freshness is one of the core elements for data trust. If people are unsure of the last time the data was updated, they will feel hesitant to use it and may not be interested in making decisions based on it or sharing it with their managers.

However, it is nearly impossible to always keep the data fresh. Data refreshes are expensive and require a lot of computing power, so "real-time" data is often unachievable. It is important to set realistic cadences for which the data will be updated. This means taking the technical limitations in to account while also understanding the business needs. Many organizations have a daily refresh where data is set to update overnight when the servers are not as busy. However, talking to the business about how often they need data is a good practice – you may find that daily updates are unnecessary and that they only use this report once per quarter when doing company updates. If that is the case, it is possible to reduce the number of updates and save money and computing power.

Even if you set your schedule optimally, it is not uncommon for something to go wrong and for data to be stale. In this case, it is important that there is some way to alert people that the data is not up to date. This could be another dashboard, an email, or an automatic alert. Alerting data users about data freshness is important because they will be frustrated if they make reports or decisions from stale data. Although it might seem like bad PR to message out data failures, it will actually build trust in the data team because stakeholders will feel like they are being kept in the loop and they will not unknowingly make decisions based on stale data.

Questions to Determine Current State of Data Freshness

To Those Who Input Data

These questions are designed to understand the process of data collection at the organization. It is important to understand how manual data is being input because if the data isn't making it into the system, no amount of technical savvy can recover it on the back end. Manual data collection is the backbone of many data organizations, so it is essential that data collection is designed to be as easy as possible, removing any roadblocks or redundancies. Additionally, it is important to understand if there is any data being kept or exchanged outside the system and why that may be happening. In an ideal world, all data is kept centrally, where everyone can access it, so if there are external data exchanges there may be a process or tool that requires reassessment.

- How often are you inputting data?
- Are you keeping any data in manual trackers, outside the typical data input process?
- Are you ever providing manual data to data users, bypassing the typical data input process?
- What is the hardest part about inputting data? What prevents you from inputting data regularly?

To Those Who Manipulate and Analyze Data

Technical Data Freshness issues will usually be centered around people in this group. The people who manipulate data are usually the ones who are acutely aware of stale data, so they may be able to more quickly pinpoint where the issues are arising. We are able to get a lot of knowledge about processes, tools, and where there may be large data freshness vulnerabilities.

- Is your data being input or refreshed in a timely manner?
- What is your ideal refresh schedule for the data? What is the current refresh schedule?
- What are the blockers for faster refreshes or more timely data?
- What processes and tools are currently in place to ETL data and create data pipelines?
  - Are they reliable?
  - Are they brittle? What input changes could occur that could break these processes?
- What tools (such as BI tools) are you using to deliver analyses?
  - What changes can occur so that the BI tools do not display the correct values (even if the data is correct)?

To Those Who Make Decisions based on Data

Data decision makers (or their analysts) are ultimately the people who will have specific data freshness requirements, so it is important to understand when they need data so that it can be refreshed on time. They may also feel the impact of stale data as a stakeholder, so they may be in a good position to explain its impact.

- How often do you need to make decisions with this data?
- Are there specific times (eg., End of Quarter, before a Board meeting) where it is essential to have the data up to date?

- Do you find yourself seeing the data and overriding a data-backed decision because you think the data is out of date?

Conclusion

Data Freshness is essential for building trust and reliance on data. If people feel that it is unreliable and frequently out-of-date, they will stop using it to make decisions. While it often isn't possible for data to be real time, there are other ways to meet stakeholder expectations. Understand when they need updated data and understand which datasets need to be updated often and which ones can be on a more spaced-out schedule. Additionally, understand the data input schedule and assess to see if there are ways to streamline the process and increase frequency. If things go wrong, be sure to communicate so that stakeholders do not end up using stale data to do analyses or make decisions. Although data freshness is usually seen as a problem for technical people, it requires input and communication from everyone in the workflow to optimize.

This article is the second in a series discussing the important considerations when assessing your Current State of Data. Follow along for the next article about Data Culture – the way that people in the organization interact with data!

ALSO ON **TDAN**

The Data-Centric Revolution: Avoiding …
2 years ago · 1 comment
Gartner has put "Knowledge Graphs" at the peak of inflated expectations. If …

Celebrating 25 Years of TDAN.com!
10 months ago · 1 comment
Twenty-five years ago today, I published the first issue of The Data Administration …

All in the Data: The Same Difference of …
9 months ago · 1 comment
What does your head do when someone uses the expression "same …

All in Gove
2 years
I last p Gover in a TI

Current State Analysis of Your Data – Part 1

Published: January 19, 2022 | 12:09 am · Vanessa Lam

Whether you're stepping into a new organization as a data lead or trying to overhaul your data infrastructure, the first step in the process is to understand how your organization currently uses data. While that may sound simple, it can be an intimidating process to start. This is the beginning of a series of articles meant to give you a jumping-off point to understanding your data organization. In this series, I will be outlining the important categories to consider as well as some example questions to ask to better understand the current state of your data organization.

There are four main categories to consider when doing a current state analysis of an organization's data: Data Quality, Data Freshness, Data Culture, and Data-Driven Outcomes. This first article will give you definitions of these four terms and why they are important, and also address the first term, Data Quality. In future articles, I will provide some questions (and some of their potential answers) to better understand these categories.

Data Quality

Data Quality refers to the accuracy and precision of your data. Is the data being input properly? Is there high enough coverage on data fields to trust the data? Good data quality is essential in order to know that insights drawn from this data can be trusted.

Data Freshness

Data Freshness refers to whether or not data is available where and when you need it to be. It also used to assess the fit of the tools and processes to the business needs. Practically, we will recognize these as the

cadence of data input and data refreshes as well as the reliability of processes and tools.

Data Culture

Data Culture refers to the way that people in the organization interact with data. Who is able to interact with it? How do they feel about it? How do they speak about it? This is critical to understand if employees are willing and interested to work with data or if increased education and discussions are required to improve adoption of healthy data practices.

Data Outcomes

Data Outcomes refers to an organization's main goal for their data. Is it a revenue driver? Does it dictate decision making? Do they need it for fundraising? These questions will indicate how data is being used now, but also potentially areas for it to grow.

These are the four major categories that we will be focusing on when doing a current state analysis on the data. As you can see, this is a wholistic look at data; we are not merely looking at the data itself, but the treatment and sentiment toward data across the organization. It is essential to consider all these factors when trying to build a sustainable data organization. Follow along for future articles that will dive deeper into the questions we can ask to better understand these four categories!

This section of this article will focus on Data Quality: what it is, why it is important, and what questions to ask to determine its current state.

The questions are organized by stakeholder group to facilitate usability; hopefully you can use this as a template to start your Current State Analysis journey. A few definitions before we begin – note that these groups are not mutually exclusive:

People who Input Data: These are people who collect and/or input data into the system. For example, salespeople inputting their sales numbers, or survey creators.

People who Manipulate and Analyze Data: These are people who organize the data and create analyses. This includes Data Engineers, Business Intelligence Professionals, and Data Analysts.

People who Make Decisions based on Data: These are the people who use the data to make decisions. This may be a sales manager deciding where to invest resources, a product manager understanding product use demographics, or an executive trying to cut costs.

What is Data Quality?

Data quality is a measure of the condition of your data, including accuracy, consistency, and completeness. Data quality can be affected in any part of the process, from collection to ETL to analysis. Accuracy refers to how well your data reflects the truth while consistency refers to how well individual data elements match each other. An example of accuracy and consistency in data collection is when a Sales Representative fills out their sales region (e.g., Pacific Northwest). If the input is a manual text field, it is possible for them to spell the region incorrectly (Pacific Nortwest) or even capitalize the region incorrectly (pacific northwest). This is a failure of data consistency. A possible correction here would be to supply a drop-down list of the possible regions (e.g., Pacific Northwest, Midwest, South, East). However, this scenario still provides opportunities for inaccuracy in choosing the incorrect region. An example of incomplete data here would be the Sales Representative leaving this field blank or not filling out the sales record entirely.

Potential solutions to these issues could be making the Region field mandatory or even pre-filling the field if Sales Representatives are assigned to specific regions. However, be aware that putting these rules on data collection may slow down data collection or even disincentivize people from completing forms if they think that it is too difficult or restrictive. When assessing data quality rules, it is important to evaluate whether guardrails and processes can be added to make positive change, or if the restrictions will cause unintentional side-effects.

Why is Data Quality Important?

Good data quality is essential to know that insights drawn from the data can be trusted. Most analysts have the experience of their manager questioning the numbers from the data, especially if they have some anecdotal knowledge that implies that the data is completely incorrect. Better data quality will reduce the number of times that this happens as the data will be more accurate and complete, meaning that it is a better reflection of the truth. In the cases where there are still disagreements, they may be more substantive and provide opportunities to talk about where anecdotal evidence and collected data may differ.

Questions to Determine Current State of Data Quality

To Those Who Input Data

These questions are designed to better understand the guardrails around data collection. It is important for data to have strong validations, such as choosing from a dropdown list instead of allowing free text; however, be aware that this could backfire if the validations prevent accurate data entry (for example, missing categories in a dropdown) or make it too difficult to fill out (making long forms with many mandatory fields can make a form unappealing to fill out). Often, data quality issues are rooted in poor data collection.

- What is your process for inputting data?
- Are there specific fields for inputting commonly collected information or is it largely free text?
- Are there selection options for appropriate fields? (e.g., A dropdown with “True” or “False” selections or a free text field where you can type “True” or “False”)

- Are there automatic or pre-loaded values in the fields? Are you typically using the pre-loaded value, or do you frequently need to change it?
- What is your biggest frustration when inputting data?

To Those Who Manipulate and Analyze Data

These questions are designed to understand how the data looks to someone who is in charge of cleaning, understanding, and verifying the numbers to be presented to business stakeholders. They are often the ones who feel most impacted by poor data quality, and therefore likely have a good idea of where the data is weak.

- When doing analyses, how confident do you feel about the outcome based on the data?

- Do you feel like you need to put caveats on your data based on low confidence intervals?
- Are there data fields that are being used for multiple purposes? Are there data fields that still exist in the data, but everyone says to avoid?

To Those Who Make Decisions based on Data

These questions are designed for people who are using data to make decisions. Often, they are not the ones who have done the analysis themselves, so they may not know the intricacies of the data as well. However, they are usually keenly aware of the business implications of their decisions and have a good sense of what aggregate numbers make sense. For example, a sales executive will likely have a good idea if the number presented is approximately correct based on their experience with the industry, company, and knowledge of their sales representatives. It is important to hear their opinion because ultimately these are the customers of the data. If they aren't happy with the quality, then some changes may need to take place.

- Do you find yourself seeing the data and overriding data-backed decisions because you don't think that what you're reading is a reflection of reality?
- What parts of the data are you confident in? What parts are you concerned by?

Conclusion

Data quality is a bedrock of your data systems. Without good quality data, analyses cannot be trusted, decisions cannot be made with data, and eventually people will stop relying on data altogether. Before attempting to implement a strong data ecosystem, it is important to have a plan to improve data quality throughout the organization, from data collection and data processes to analyses. This article is the first in a series discussing the important considerations when assessing your Current State of Data. Follow along for the next article about Data Freshness – measuring whether the data is up-to-date and reliable!

Current State Analysis of Your Data – Part 3 – Data Culture

Published: March 2, 2022 | 12:07 am · Vanessa Lam

This article is the third in a series taking a deep dive on how to do a current state analysis on your data. This article focuses on data culture, what it is, why it is important, and what questions to ask to determine its current state. The first two articles focused on data quality and data freshness.

The questions are organized by stakeholder group to facilitate usability; hopefully you can use this as a template to start your Current State Analysis journey. A few definitions before we begin – note that these groups are not mutually exclusive:

People Who Input Data: These are people who collect and/or input data into the system. For example, sales people inputting their sales numbers, or survey creators.

People Who Manipulate and Analyze Data: These are people who organize the data and create analyses. This includes Data Engineers, Business Intelligence Professionals, and Data Analysts.

People Who Make Decisions Based on Data: These are the people who use the data to make decisions. This may be a sales manager deciding where to invest resources, a product manager understanding product use demographics, or an executive trying to cut costs.

What is Data Culture?

Data Culture refers to the way that people in an organization interact with data. Whether or not it has been intentionally curated, every organization has a Data Culture. It can be found in the way that people speak

about data, if they are afraid of it, or how they include it in their decision making. A poor data culture can lead to confusing communication, inconsistent decision-making, and non-actionable insights, while a good one promotes robust, actionable, and data-driven insights.

Although good technologies and processes are important for responsible data use, it is equally important to understand how humans interact with data. Often, organizations will implement new data technologies and procedures, but will not think through how it affects the people who produce, use, and make decisions based on that data. Cultivating a strong Data Culture refers to the human side of data management – making sure that data is trusted and that everyone is using data responsibly.

Why is Data Culture Important?

Data Culture is important because even though data is often viewed as technical, it is ultimately consumed, used, and interpreted by people. If people are afraid of data or do not trust it, they will not use it to make decisions. Additionally, if they are not using the same vocabulary and metrics when discussing data, they will end up with confusion and miscommunication. Time will be wasted discussing the correctness of data rather than using it to make decisions.

Conversely, a strong data culture means that people will turn to data first when making decisions. There will be less time questioning the data and visualizations and more time spent interrogating the insights and decisions. There will be a strong foundation on which to base discussions because people trust and feel confident in the data.

Data Culture can often spiral, both for the better and worse. Left unchecked, different departments – and even individuals – can make different decisions with data, leading to inconsistencies. These consistencies lead to mistrust, and therefore discourage people from using data to make decisions. Each department becomes more entrenched in their data ways, which can lead to a downward spiral for data culture. Conversely, when people trust their data, they want to use it to make decisions. The more people discuss data in a positive light, the more data practices will converge across the organization, further engendering trust and encouraging data-driven decisions. Read more about Data Culture here.

Questions to Determine Current State of Data Culture

To Those Who Input Data

These questions are centered around understanding how connected the people who input data are to the data insights that are produced from their inputs. Oftentimes, the people who input data are not the same as those who consume it, and therefore it becomes a lower priority to the people who input the data. A strong data culture would try to close the feedback loop, showing those who input the data dashboards and metrics based on the data that they input. With more feedback and visibility, the potentially tedious task of data collection may gain more meaning.

- Do you know how the data that you input is used? Do you know why it is important?
- Who is the primary consumer of the data that you input?
- Are you a consumer of any data or dashboards? How does that affect the way you do your role?
- Is data a positive or negative influence in your role?

To Those Who Manipulate and Analyze Data

For the people who analyze the data every day, the Data Culture is most of their working culture. Not only do they reap the benefits of trustworthy and consistent data in their own work, but they also will be the ones to hear about any data issues from stakeholders across the organization. These are the people who have the strongest understanding about the Data Culture and have the most power to change it from the ground up.

These questions are meant to understand how much they trust data as users, but also how they think others use and perceive data across the organization. We are trying to understand if people are communicating about data in a positive way, but also if they are using consistent metrics so that everyone is speaking the same language.

- Do you trust the data that you are using? Do you understand it?
- Are people across the organization using the same vocabulary to refer to the same things?
- If there are issues in the data, are they well communicated across the organization?
- Are there discussions about data across the organization? Are they generally positive or negative?
- How is your data organization currently structured? Do you have a democratized or centralized data team? (Read more about data organizations here)

To Those Who Make Decisions Based on Data

It is very easy for data decision makers to be detached from the source of their data. While it is not essential for them to understand the intricacies, it is good to know where their level of understanding is with the data. This helps to know if they are appropriately assessing the level of trust they should have in the data. Some decision makers may blindly trust the data because they have no reason to believe it is wrong, even though analysts have wide confidence intervals and issue many warnings with their analyses. Others may forgo data in their decision making altogether because they do not trust it, even if their analysts have a high level of trust. It is good for decision makers to have an ear to the ground for upstream data issues and have a high-level understanding of where the data comes from.

- Do you understand where the data comes from (at a high level?)
- Do you understand what goes into the metrics that are being used to make decisions and measure performance?
- What factors would cause you to override a decision that is supported by data?

Conclusion

Data Culture is what happens when people interact with data. Data Cultures influence data outcomes just as much as Quality and Freshness, but is less discussed because it is not a technical issue. Without focus on Data Culture, you may have perfect data and no one to use it. Even worse, you could have people using it in inconsistent ways across the organization, causing confusion and slowing down processes. When doing a Current State Analysis on data, Data Culture is an integral piece of the puzzle.

This is the third in a series discussing the important considerations when assessing your Current State of Data. Follow along for the next article about Data Outcomes – how people use data to drive their organization's goals and mission.

ALSO ON **TDAN**

All in the Data: Empower Stewards with a Data …

3 months ago · 1 comment

Most organizations that are implementing formal data governance programs …

Celebrating 25 Years of TDAN.com!

10 months ago · 1 comment

Twenty-five years ago today, I published the first issue of The Data Administration …

The Tower of Babble – Where Truth Goes to Die

2 months ago · 1 comment

Once upon a time (the way every good fairy tale begins), a book or a paper by a Ted …

How Secu

7 mon

The p busin efficie