Palantir

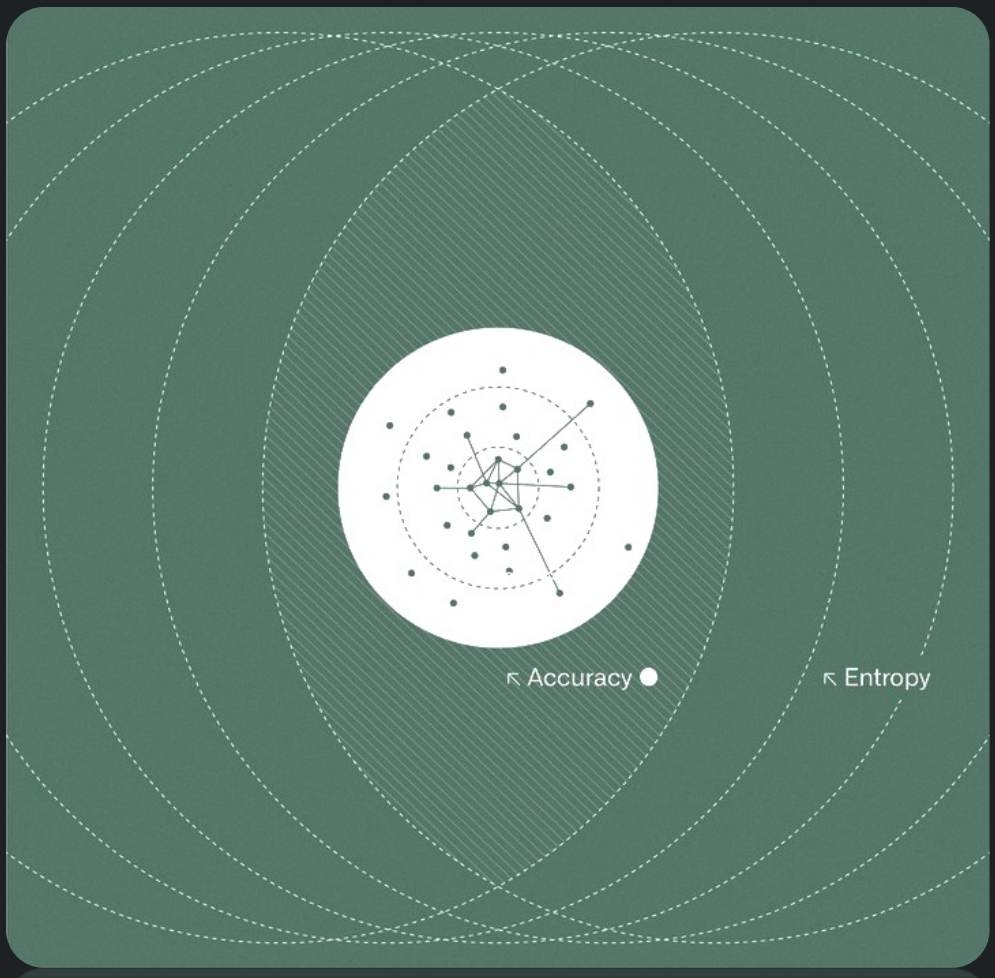
Trust in Data

Strategies to manage data quality, integrity, and transparency

palantir.com

Copyright © 2021
Palantir Technologies Inc.

All Rights Reserved



Introduction

Organisations around the world leverage Palantir's software for their most critical initiatives: from pandemic response to cancer research, supply chain operations and clean energy development. While these projects vary in scope and impact, they all begin with the same fundamental question: is the relevant data available? If so, what is the *quality* of this data, in terms of completeness, accuracy, and timeliness? How do we ensure our most important institutions don't succumb to data entropy, with poor data informing poor decisions? What can institutions do to protect their most valuable asset, ensuring it continues to inform their trajectory into the future?

In this whitepaper, we will explore why a high bar of data quality is imperative for every institution. We will highlight strategies to track and monitor data quality and data integrity, and then address how organisations can promote user trust in data. Finally, as a longstanding technology partner of governments, we will share our learnings from the COVID-19 response efforts, among others, on why the tenets of accountability and transparency are particularly crucial in crisis situations.

Why does it matter?

We start with data quality because it is the foundation upon which all successful data and analytics projects are built. A single bad dataset has the potential to compromise an entire data-driven initiative, triggering consumer distrust and rendering the effort moot. Worse, it could get into the hands of users and be misinterpreted or actioned upon.

Sometimes these lapses in data quality are simply a result of having the wrong source data (e.g., an incorrect mapping table), sometimes the data just represents a different perspective of correctness (e.g., different definitions of metrics between business groups), or sometimes it is due to mistakes in the data preparation (e.g., incorrect data cleaning logic).

When users are blind to assumptions or issues in their data, the decisions made upon that flawed data can propagate quickly throughout an organisation, with successive levels of users left none the wiser. This is why data quality must be paired with robust data transparency: all users, with relevant roles and permissions, should have the full context necessary to appropriately use and trust their data.

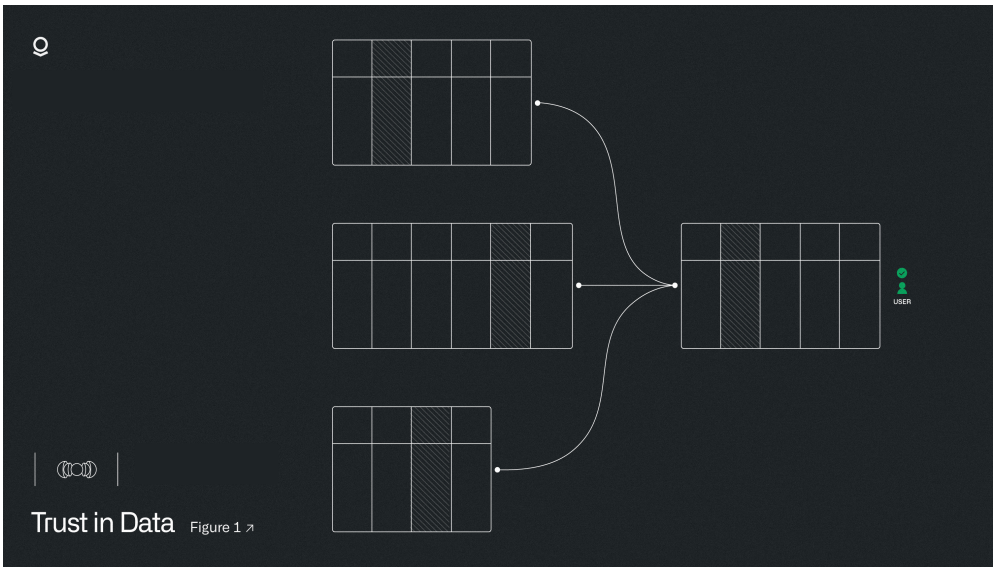
Copyright © 2021 Palantir Technologies Inc. and affiliates ("Palantir"). All rights reserved. The information in this document is proprietary and provided for informational purposes only and shall not create a warranty of any kind. Any data contained herein is notional and for demonstration purposes only.

Why is it difficult?

All data is not created equal, and rarely does it come packaged in a clean, reliable form that is easy to analyse. The reality is that there is no perfect data source. Sometimes, the source lacks the breadth or depth of necessary information. Other times, source systems are antiquated, rendering the data architecture, schemas, or structure inadequate to support a new use case. To compound the problem, many sources contain inaccurate data, get updated inconsistently, or are just altogether out-of-date. The examples are plentiful and these only scratch the surface.

Beyond the data itself, important context (metadata) surrounding the data can often be piecemeal or non-existent. All too often, updates are received in the form of one-off extracts or CSVs with no visibility into where the data came from, when it was last updated, or any indications of potential issues or shortcomings. In the event this information does exist, it might only be found in an email or verbal exchange with no lasting record.

These myriad issues can affect a single dataset. Now, imagine taking tens or hundreds of data feeds and harmonising them into a single source of truth. Differing reporting standards, definitions, schemas, and formats make it difficult to ensure the data is properly prepared, merged, and vetted before it goes in front of decision-makers.



Copyright © 2021 Palantir Technologies Inc. and affiliates ("Palantir"). All rights reserved. The information in this document is proprietary and provided for informational purposes only and shall not create a warranty of any kind. Any data contained herein is notional and for demonstration purposes only.

How we think about it

Whenever we commence a project, we encourage users and decision makers to ask the following fundamental questions of their data asset:

- What sources make up this data asset? How reputable are these data sources? What do we know about where this data comes from?
- How trustworthy are these data sources? Is this data skewed such that it might result in misleading interpretations?
- How consistent is this data? Will we know if the shape or quality of the data changes or deteriorates?
- When was this data last updated? Is there a lag or reporting? If so, how might it affect the data and decision-making?
- If there are multiple data sources that need to be joined together, which data source takes precedence if there is a conflict? How can we track the decision to overwrite or prioritise a data source?

In particular, data quality and data integrity questions come in all types of flavors and issues can arise all along the data lifecycle:

→ Data Collection: From the data collection and if the data collected is complete or if there are systematic gaps in the data collection

→ Data Content Consistency: to when the data is parsed and cleaned and how consistent the data elements are over time

→ Data Latency: to the timeliness of the data being received and published

→ Data Integration: to the harmonization of the data when merged with other sources

→ Data Legibility: to the output of the dataset being legible to the user for appropriate use

→ Data Validation: to the routine error checking on the data asset for consistency of data outputs over time.

Copyright © 2021 Palantir Technologies Inc. and affiliates ("Palantir"). All rights reserved. The information in this document is proprietary and provided for informational purposes only and shall not create a warranty of any kind. Any data contained herein is notional and for demonstration purposes only.

How we think about it (continued)

Answers to these questions are relevant not only for initial data integration, but at every stage of the subsequent data lifecycle, and are necessary for creating a trustworthy data asset. From our extensive work with enterprise customers over the years, we learned that answering these questions at the necessary scale and pace was extraordinarily difficult in the absence of structures and applications to help track and secure data quality.

As a result, we designed Foundry's data integration tools such that any data engineer is empowered to track the health of the data pipeline and so the data user can know where the data came from and how it should be interpreted.

Data quality, integrity, and transparency for producers in Foundry

As data quality and data integrity are the foundations for any data project, it is instrumental that we first empower data engineers with the tools and transparency they need to build a quality data pipeline.

Below, we highlight some of the unique tooling and capabilities of Palantir Foundry that help ensure data quality.

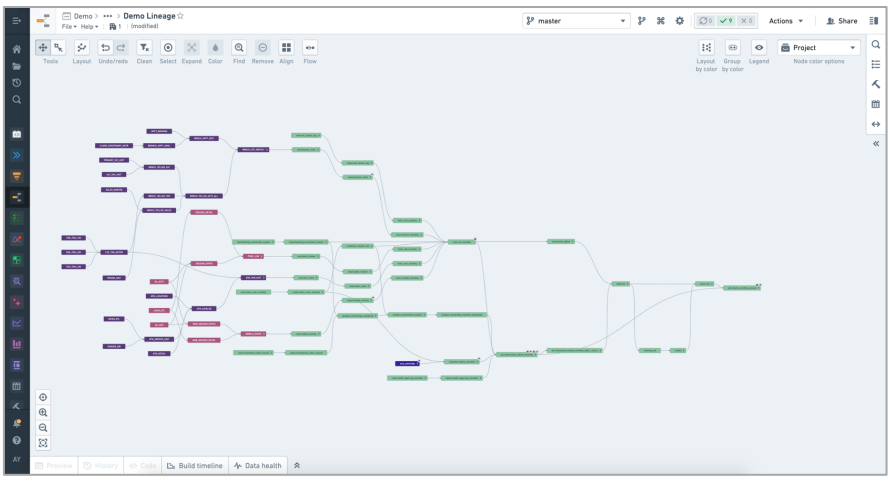
Data lineage tools

- **Data Lineage Tools:** Understanding where the data has been and where it is going allows data engineers to trace potential data quality issues across the platform, ensuring verified data is available to every applicable business division and team. For instance, what is the root source of the data? Which organisation owns this? When was it updated? What transformations, filters, or enrichments have been made on this data?
    - **Exploring Upstream Inputs:** Foundry allows data engineers and users to trace back, investigate, and resolve data quality issues at their source.
        - At a glance, data engineers can look upstream at every parent dataset that feeds into the current data asset, tracing every change and merge all the way back to the source. Users are able to inspect how the data has been transformed over time, providing critical debugging capabilities. For instance, if there are issues in the current dataset, the user can trace back to the source data and validate if it is an underlying data issue all within the platform.

Copyright © 2021 Palantir Technologies Inc. and affiliates ("Palantir"). All rights reserved. The information in this document is proprietary and provided for informational purposes only and shall not create a warranty of any kind. Any data contained herein is notional and for demonstration purposes only.

Tracking downstream effects

- **Tracking Downstream Effects:** Foundry's Data Lineage allows users to proactively trace where data quality issues may impact downstream artifacts.
  - If outputs look amiss downstream, Foundry enables users to search along the pipeline for where there might be issues with how the data was prepared with code. Some common issues include exploding joins, dropping or filtering data prematurely, or missing data. With Foundry, users can zero in on where in the pipeline this regression is introduced and remediate.
  - If an issue is identified, the team must investigate whether it has propagated downstream to other datasets, reports, or artifacts. Data Lineage allows data engineers to proactively alert downstream consumers and fix the data flow before issues compound.



Copyright © 2021 Palantir Technologies Inc. and affiliates ("Palantir"). All rights reserved. The information in this document is proprietary and provided for informational purposes only and shall not create a warranty of any kind. Any data contained herein is notional and for demonstration purposes only.

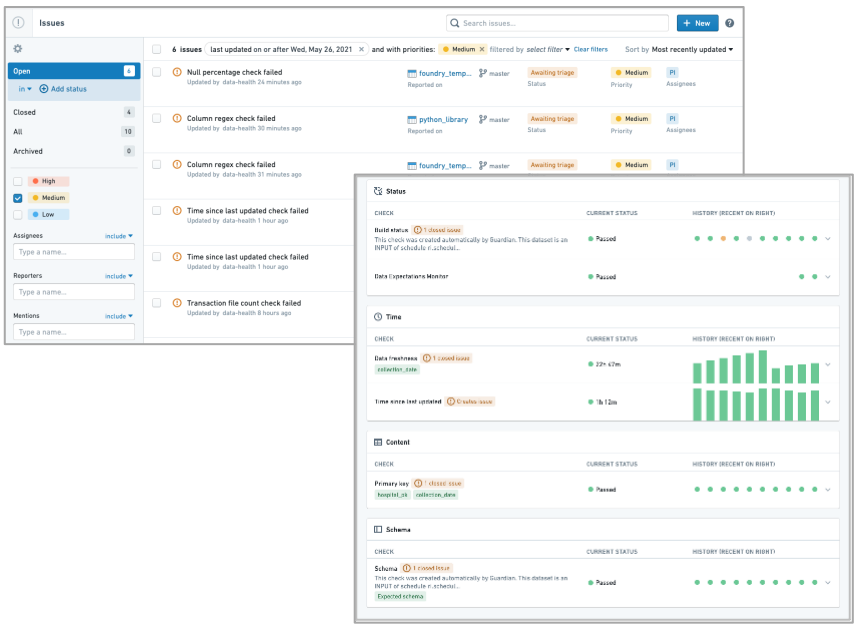
Automated data health checks

- **Automated Data Health Checks:** Foundry automates checks designed to detect aberrations in the data – whether timeliness of data updates, completeness, consistency, or even identify missing contents – to ensure robust data quality at scale. The pipelines can also listen for these checks and prevent propagation of the data downstream, if any of the checks fail.
    - For instance, users can define a target data schema, such as the column names and expected content types, and establish validations that ensure that changes in the data model are flagged early. If a validation fails, Foundry can trigger alerts for review and remediation. In addition to ingestion, Foundry also performs data pipeline health checks that trigger alerts all along the data pipeline. Some examples of data health checks include those listed below. This is not meant to be a comprehensive list, but rather a high-level view of some examples:
        - # of rows - Tracking whether data is within the range of expected rows. Data might be corrupted, backfilled, replicated or drop off — just to name a few. When this happens, proactive alerting allows the data engineer to be prompted immediately and root cause the issue.
        - Out-of-date - Metadata on when the last time the data was updated provides visibility into the latency of the data, that may make it out of date and no longer accurate. For instance, if there is a dataset that is generally updated daily, but does not update for over 24 hours, an engineer will have to do some investigation upstream to determine the cause, before this out-of-date data inaccuracies across the pipeline.

Copyright © 2021 Palantir Technologies Inc. and affiliates ("Palantir"). All rights reserved. The information in this document is proprietary and provided for informational purposes only and shall not create a warranty of any kind. Any data contained herein is notional and for demonstration purposes only.

Automated data health checks (continued)

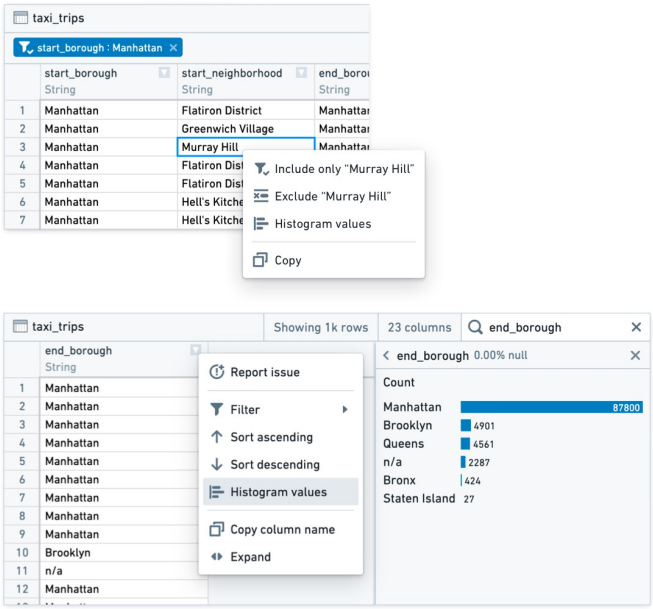
- Unique keys / joins - Particularly when data is joined or integrated across systems or tables, a common data integrity issue occurs when there are not unique keys, which may result in erroneous joins or data being dropped. For instance, when joining metrics to a list of geographical areas (e.g., counties in the US), users may confront the issue that, if any of the county codes change, the data may not map with the corresponding data and that data will be dropped. Having the ability to check for completeness in the keys on the matching table ensures that the data transformation processes do not inadvertently drop critical data.
- Null % - Another common issue is content within a table changing and resulting in disappearing data. Tracking when the shape of the data significantly changes can suggest a potential issue in the upstream data set or an issue with how the data is being transformed. This would trigger a notification to the data engineer to conduct a root cause analysis to figure out if the data is actually changing or if a change to the data source of data pipeline might have caused the issue.



Copyright © 2021 Palantir Technologies Inc. and affiliates ("Palantir"). All rights reserved. The information in this document is proprietary and provided for informational purposes only and shall not create a warranty of any kind. Any data contained herein is notional and for demonstration purposes only.

Ad hoc analyses for comparison

- **Ad Hoc Analyses for Comparison:** Foundry provides tooling for data engineers to conduct ad hoc analyses to better understand the shape of data and how it changes over time. This allows data engineers to proactively identify anomalous patterns and conduct root-cause analyses on data issues.
    - Some common ways Foundry enables ad hoc analyses include:
        - Creating histograms and tracking patterns – Users can inspect elements in the table to make sure that the distributions are as expected and the integrity of the dataset is not compromised. This gives data analysts an immediate sense of whether there are anomalies in the data (e.g., Are there more than expected nulls? Is the data skewed in one direction - if so, why?).



- Comparing transactions - Foundry is built to "version data like code" meaning that users can scroll back through to previous versions of a dataset and confirm if the dataset has changed over time. If large variations occur over time, Foundry may suggest some potential changes to the data source or pipeline that might have had deleterious effects on the outcome.

Copyright © 2021 Palantir Technologies Inc. and affiliates ("Palantir"). All rights reserved. The information in this document is proprietary and provided for informational purposes only and shall not create a warranty of any kind. Any data contained herein is notional and for demonstration purposes only.

Ad hoc analyses for comparison (continued)

- Differences between datasets - Another workflow built into Foundry is the ability to take two datasets and automatically find the differences between two. For instance, if there are any fields that exist in an upstream dataset and are missing in the corresponding downstream dataset and have minimal transformations applied to it, this might suggest some type of issue in the transformations or logic preparation between the two versions which would require investigation. This is particularly helpful when checking code transformations along the pipeline to make sure no data is dropped or lost.

- Inspecting data quality should be straightforward and easy for the data engineer who is preparing the data. Ensuring that these types of checks are part of the data pipeline preparation helps track the integrity of the data end-to-end. Access to tools that are purpose-built for this type of analysis speeds up development significantly.

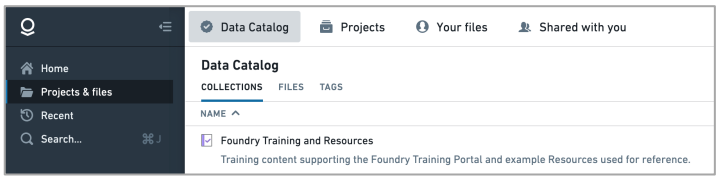
Copyright © 2021 Palantir Technologies Inc. and affiliates ("Palantir"). All rights reserved. The information in this document is proprietary and provided for informational purposes only and shall not create a warranty of any kind. Any data contained herein is notional and for demonstration purposes only.

Data quality, integrity, and transparency for consumers in Foundry

Even after methodically validating the data, the work is still not done -- the use and interpretation of data depends on the users, and users vary widely in their approach and needs. Some users are not required or expected to inspect data, others "trust but verify," and some want to know every step of how the data was prepared. Foundry is designed to meet the needs of many of these user profiles, providing tooling to surface the information critical to their work.

Data source trackers and data catalogs

- **Data Source Trackers and Data Catalogs:** Foundry allows users to explore permissioned data across the enterprise. Giving users visibility into what is in their data asset is key to establishing a baseline of trust and context on the platform. Foundry not only provides visibility into what data is on the platform, but also for critical metadata such as descriptions, classifications, and data handling and sharing guidance.



  - When users encounter any integrated data system, their first questions often seek to understand what data is available. For instance:
    - "What data is in here?" - In many Foundry instances, there is a configured data source tracker or curated data search tool to help users find and surface the data they need. Whether tagging data by topic or project labels, these tools are made for users to facilitate faster search and discovery within the platform. These provide critical context (e.g., data owners, data descriptions, data restrictions) to orient users quickly.
    - "Where is the data I'm looking for?" - Foundry also allows teams to configure data catalogs or surface "verified" datasets to users to quickly jump to analysis-ready datasets. This helps ensure users do not get lost in the sea of interim datasets before arriving at the latest version the user needs. Curating these sets of data by user group enables users to sign-in and immediately start analyzing and working with the data they need.

Copyright © 2021 Palantir Technologies Inc. and affiliates ("Palantir"). All rights reserved. The information in this document is proprietary and provided for informational purposes only and shall not create a warranty of any kind. Any data contained herein is notional and for demonstration purposes only.

Data profiles, documentation, and metadata

- **Data Profiles, Documentation, and Metadata:** Foundry can also surface to users critical contextual information alongside the data itself, such as data gaps, limitations, time range, or known data owner concerns, allowing all users to have the necessary context about the data inputs before using it to drive decisions.
    - Data Descriptions or Profiles
        - "What am I looking at?" - Data does not exist in a vacuum. Users need context to understand it. Annotations, like data descriptions, give users context regarding what they are looking at. This type of documentation is embedded alongside datasets in Foundry to provide users with a human intelligible summary of what each dataset is and flag any concerns on how the data should be used, as part of the user's analysis workflow. This can include information such as data descriptions/summaries, limits of the data, known anomalies, time period/inclusions/exclusions, update frequency, and details on the general shape of the data.
    - Metadata
        - "When was this dataset last updated? What is the shape of this data?" - Oftentimes, the metadata about data is more important than the data itself - as it provides critical information about the status and characteristics of the data. This is why Foundry surfaces this information alongside every dataset. From the number of rows in a table to the statistics about the data like distributions to the schema to when it was last updated to what sources feed into the dataset and on and on. For the user, all of this information is rich context about the dataset. Not only does this provide the user a better sense of what they are looking at, it can also help build a better intuition for how the user should interact with the data.

Copyright © 2021 Palantir Technologies Inc. and affiliates ("Palantir"). All rights reserved. The information in this document is proprietary and provided for informational purposes only and shall not create a warranty of any kind. Any data contained herein is notional and for demonstration purposes only.

Issues tagged to columns

- **Issues Tagged to Columns:** Regularly-updated datasets change and evolve over time: data that is healthy today may not be in the future. Platform-wide transparency about potential data issues gives platform users, data engineers, and data owners agency to ask questions or flag potential issues about the data, democratising data quality inspections.
    - "Has anyone noticed anything incorrect about this dataset?"- Foundry Issues gives users visibility into where there might be cause for concern about the data. If an issue is flagged on a dataset, this will be surfaced immediately to other users on the platform and send an alert for remediation or review. Alternatively, Foundry Issues can also be generated as part of the Data Health checks or logic mentioned above. This provides platform-wide transparency regarding any concern, whether surfaced by human operators or by automated checks. This type of collaboration is intended to build trust through transparency to users.

Visual data lineage

- **Visual Data Lineage:** End-to-end visibility from initial integration to data cleaning, aggregation, and analysis allows users to track how data is prepared and how it fits into the bigger picture. This includes reports, analyses, models, and everything in-between. This explicit transparency into data preparation and pipelines allows users to be confident that they can explain and interpret the data and its results.
    - "Where did this data come from and how is this data used?" - Similar to the section for producers, the Data Lineage tool also allows consumers to look both upstream and downstream at code transformations on the data. It also provides a visual means to review metadata instantaneously. From understanding whether the upstream datasets are up-to-date to seeing row counts and issues tagged on the dataset, Data Lineage gives users the ability to view this context simultaneously with the broader system.

Empowering users with transparency into data origins, preparation, and pipelines gives them confidence in the quality and lineage of their information, as well as the validity of their derived results.

Copyright © 2021 Palantir Technologies Inc. and affiliates ("Palantir"). All rights reserved. The information in this document is proprietary and provided for informational purposes only and shall not create a warranty of any kind. Any data contained herein is notional and for demonstration purposes only.

Data transparency for public trust

For organisations whose mandates and efficacy rely on public trust, such as government health departments, transparency is of utmost importance. To foster this trust, many organisations publish aggregated or deidentified data along with detailed notes on the origin and preparation of the information, providing the public with key insight into the inputs driving critical decisions.

Such documentation should include clear communication about the sources of all relevant information, how the data was prepared, and where data quality gaps exist. Being upfront about potential issues demonstrates that the organisation is acting in good faith, thoughtfully using available data, and being mindful of limitations.

We have seen this take shape both in the UK and US where providing public access to comprehensive COVID-19 data powered by Palantir Foundry enables not only immense utility for universities, local policy makers, and citizens, but also unprecedented visibility into the quality and completeness of the data being reported as it evolves over time. This level of public data reporting has also created a larger feedback loop as local communities use the same national data to inform local policies, which in turn, improve data quality and collection at the source.

We believe that trust in data is a public good that prevents institutions from seeming like they are operating in a black box. If the public is to trust their institutions, then trust must be built on a foundation of transparency.

Tying it all together

More than fifteen years of supporting customers tackling the world's hardest problems and the most complex data pipelines has made at least one thing clear: data quality requires a virtuous cycle between transparent data producers and informed data consumers. Making better, trustworthy decisions with data requires leveraging the right tools to automate and conduct reviews, providing documentation and feedback, and engaging with the full context of the data lineage and history. Foundry provides all of these tools, along with transparency from source to output. Our customers make some of the most consequential decisions in the world, and we are proud to say we are building the tools that allows society to trust in their data.

Copyright © 2021 Palantir Technologies Inc. and affiliates ("Palantir"). All rights reserved. The information in this document is proprietary and provided for informational purposes only and shall not create a warranty of any kind. Any data contained herein is notional and for demonstration purposes only.