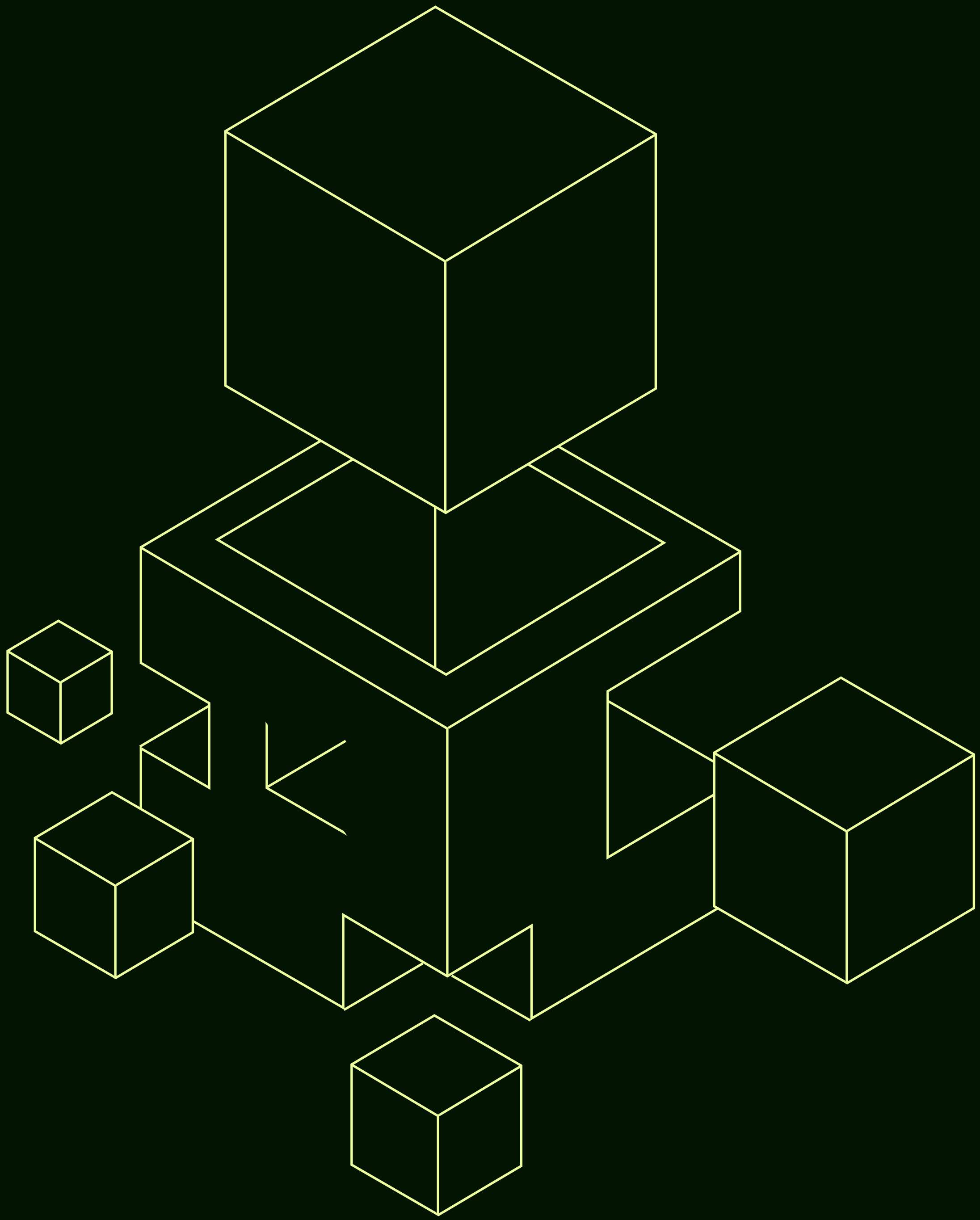
DATA TRANSFORMATIONS: ADDING VALUE TO YOUR TELEMETRY DATA



mezmo

WHITE PAPER

INTRODUCTION

The first step in observability is collecting the telemetry data necessary to derive relevant insights about complex systems. However, in many cases, data in its original form is not enough for achieving the insights in an efficient or impactful way.

That's why businesses seeking to get the very most value out of their data bake data transformation processes into their telemetry pipelines. In many cases, data transformation can dramatically increase the efficiency and effectiveness of observability by improving the quality and usefulness of data before it is analyzed, stored, or otherwise processed.

This white paper explains how to leverage data transformation inside a telemetry pipeline to increase the value of your data. It begins by defining data transformation and its role within data pipelines, then compares specific types of data transformations and data transformation use cases. It also offers tips and best practices for data transformation within telemetry pipelines, and summarizes the key impacts of data transformation on observability success.

UNDERSTANDING DATA TRANSFORMATION

Data transformation is the modification of the format or structure of data. In other words, when you transform data, you take "raw" data generated by a data source, such as an application or an infrastructure resource, and alter the data in a way that changes its form or format.

The main purpose of data transformation is to make data easier to work with. It often happens that the format or structure in which data appears when it is output by a source is not ideal for data processing needs. Data transformation provides a way of modifying the data so that it can be interpreted more easily. Sometimes, data transformation also helps achieve other goals, such as reducing the cost of data storage or making it easier to organize data.

DATA TRANSFORMATION EXAMPLE

As a simple example of data transformation, consider an application that contains multiple microservices, each of which generates its own log file. Rather than analyzing each log file individually, an organization merges the logs into a single data set, then feeds them into analytics tools.

In this way, the organization achieves what's known as aggregation, which is one form of data transformation. Aggregating data from multiple sources into a single file can speed and simplify analytics because it reduces the number of files that analytics tools need to ingest.

WHICH TYPES OF DATA CAN YOU TRANSFORM?

Metrics, traces, logs, and virtually any other type of data that an organization can collect from its IT estate can be transformed. As long as engineers understand how data is formatted originally and how to achieve their desired transformations, they can deploy tools that will modify data within the telemetry pipeline such that it undergoes the desired changes as it moves from its source to the tools that process it.

The only type of data that typically cannot be transformed efficiently is data that is formatted in a non-standardized way, such as a log file produced by an application not conforming to any mainstream logging standards. Fortunately, in today's standards-centric world, it's rare to run into data sources that don't conform to norms. And even if you are dealing with non-standard data, it's still usually possible to write data transformation rules for it, although doing so may require a bit more effort because you'll need to understand the data source's bespoke data formatting processes first.

BENEFITS OF DATA TRANSFORMATION

When you use data transformation to change the format or structure of data, you can reap a variety of benefits, such as:

1. **Enhanced data quality.** Data transformation can remove duplicate entries or correct missing information within a data set. This improves data quality, which in turn increases the reliability of data analytics processes.

2. **Cost savings.** By optimizing the format of data prior to analyzing it, you can reduce the computational resources required to complete analyses. This translates to lower processing costs. Additionally, using a pipeline to transform the same piece of data to meet the specific needs of the destination can reduce the number of agents or logs generated at the source down to one, which not only reduces the strain on the source of data but also allows companies to easily test and deploy new destination solutions. Transformation may also reduce data volume, which saves on storage.

3. **Faster time-to-value.** By transforming data to meet the specific needs of the destination, you can weed out unnecessary noise and provide only the appropriate, curated data to the destination in question. This optimized data processing results in faster insights and reduces the load on the destination solutions, allowing for more efficient processing and analysis of the data. This, in turn, enables faster time-to-value for businesses using the data.

4. **Improved compliance.** In some cases, data transformation can help organizations conform with data compliance rules by, for example, anonymizing sensitive information.

The bottom line: Data transformation makes data more usable, which in turn increases the value that you can derive from data while simultaneously reducing the time and effort required to derive the value.

LIMITATIONS OF DATA TRANSFORMATION

It's worth noting that data transformation is not always necessary or appropriate. If the original format of your data lends itself well to your goals, there is no need to transform it. You should also consider the cost and effort required to perform data transformation, which in certain cases may outweigh the benefits.

In general, however, data transformation is well worth the effort in any situation where the data that is produced by your sources is not ideally suited for processing by an array of destinations. Even if performing transformations requires some time and adds complexity to your observability pipeline, those costs will typically be repaid many times over thanks to the added processing efficiency and accuracy that you'll achieve by transforming your data. By moving the heavy lifting of data transformation to the pipeline, you can lock down the data sources, reducing the need for change management on the source side. This not only simplifies your pipeline but also ensures that the data is consistently processed and prepared for downstream analysis, making it more accurate and reliable.

In that sense, data transformation is an investment that yields rich dividends later in the observability pipeline. While transformation is not a cost or effort-free process, investing in data transformation during the earlier stages of your pipeline will significantly increase the value of the processes that take place later in the pipeline.

DATA TRANSFORMATION IN ACTION

Metrics, traces, logs, and other data can be subject to a wide variety of transformations. The following are the most common:

REDUCE

Reduction is the combination of multiple events over time into one based on a set of criteria to achieve a specific goal. For example, if you have an authentication log that contains entries for authentication attempts by multiple users, and the entries are ordered chronologically, you might filter the data by grouping each user's attempts into a single section. Doing so could make it easier to analyze authentication patterns on a per-user basis.

AGGREGATION

As noted above, aggregation means merging multiple data sources, such as a series of log files, into one. Aggregation can make it more efficient to ingest and process data by reducing the number of individual files that your analytics tools have to parse.

ENRICHMENT

When you enrich data, you add contextual information that makes it easier to interpret. As an example, consider a network log file that records the IP addresses of endpoints. If you have a separate file that maps IP addresses to host names, you could enrich the original file with the hostnames of each endpoint, which might make it easier to identify different types of systems than IP addresses alone.

COMPACTING FIELDS

Compacting fields means merging two or more fields within a data set together. For instance, if hostnames and IP addresses for endpoints appear as separate fields within a log file but it's not important to you to make a distinction between each type of data, you could compact that information into a single field.

FLATTENING

Flattening is the opposite of compacting. To flatten data, you take data that currently exists as a single field, such as an IP address-hostname mapping, and break it into separate fields. Flattening is useful if, for example, you need to analyze hostnames and IP addresses separately because your network hostname assignments have changed over time and you therefore can't treat them as a single field.

DEDUPLICATION

Deduplication means removing redundant information from a data set. For example, if a bug in an application causes it to log the same event twice, deduplication would allow you to remove the redundant entries. In turn, your analytics results would be more accurate because your tools would not misinterpret the two entries as distinct events.

DROPPING

Dropping is the removal of certain types of information from a data set. For example, you might drop certain columns from a database if they contain sensitive information (like customer names) that you don't need to retain for analytics purposes. Data dropping can help to achieve data compliance. It can also speed analytics by reducing the total volume of data that needs to be analyzed.

FIELD ENCRYPTION

Field encryption means the encryption of specific fields within a data set. This is a way of protecting the data inside the field from unauthorized access while still retaining it so that any tools configured to decrypt the data can read it.

SAMPLING

Data sampling is the process of extracting only a certain portion of data from a larger data set. When you are dealing with large volumes of information, sampling can help to speed analytics and lower analytics costs by reducing the total amount of information you need to analyze. As long as the sampled data is representative of the broader data set, your analytics results should still be accurate.

PARSING

Data parsing is the conversion of data from one format (such as JSON) to another (such as CSV). Parsing can help to prepare data for analytics tools that only support one format. It can also standardize data formats in order to provide greater consistency and make it easier to write queries that will work across multiple data sources.

DATA TRANSFORMATION USE CASES: THREE EXAMPLES

To add context to the ways in which the various types of data transformations described above can assist with observability goals, consider the following observability goals and the role data transformation plays within them.

REDUCING DATA VOLUME TO CONTROL COSTS

In general, the more data you ingest into analytics tools or storage platforms, the higher your costs. Most analytics tools charge based partly on the total volume of ingested data.

By reducing the load on a destination that is hosted within a customer's locale, they can reduce the costs of deployment.

To reduce costs, you can leverage data transformation processes like deduplication (which removes redundant data, leading to a reduction in total data volume), dropping (which makes data sets smaller by eliminating unnecessary data), and sampling (which allows you to analyze only a subset of your total data). Each of these transformations could significantly lower your total analytics bill.

SUMMARIZING DATA

If you have a large volume of data from disparate sources, making sense of it all can be difficult. Even if you are able to analyze the data automatically, you may not be sure which data sources to analyze or how to relate them to each other.

In this situation, parsing the data so that it all appears in the same format may be a first step toward understanding how to make it more actionable. From there, you might aggregate data from similar sources in order to reduce the total data sets you have to contend with. Sampling and compacting could also help to reduce complexity. Ultimately, these processes would add up to data that is easier to make sense of, and therefore easier to analyze in a systematic way.

PROTECTING SENSITIVE DATA

If your data contains sensitive information, such as personal names, credit card numbers, and addresses, data transformation can help you to reduce the risk of misusing the data or accidentally exposing it in a way that violates compliance requirements. With a single pipeline, you can encrypt sensitive fields within data that will be stored, while dropping the fields out to destinations that would cause a compliance violation. This enables you to protect sensitive information while ensuring that the necessary confidential data is available to different sources within your organization. By using data transformation in this way, you can minimize the risk of data breaches and maintain compliance with regulations.

BEST PRACTICES FOR DATA TRANSFORMATION

Given the many types of data transformations you can perform and the many reasons you might perform them, it's important to think strategically about how to get the greatest value out of data transformation. The following practices can help.

AUTOMATE DATA TRANSFORMATION

The single most important best practice for data transformation is to automate the process. Although most data transformations could be performed manually, doing so is not practical if you operate at scale. Instead, you should employ data transformation tools that can automatically modify data based on rules you configure.

TRANSFORM WITHIN YOUR PIPELINE

Wherever possible, automated data transformations should take place within the telemetry pipeline. Rather than using one set of tools to collect data, a different, disconnected set to transform the data, and a third set to analyze it, construct a pipeline that does it all – data collection, transformation, and analytics – using an integrated set of tools and processes.

By integrating data transformation into your pipeline, you can move data as efficiently as possible both before and after transformation, which leads to better manageability and faster processing. An integrated pipeline also helps to reduce processing costs because you don't need to export and import data between discrete sets of tools.

ESTABLISH DATA GOVERNANCE RULES

Rather than leaving it to individual teams to decide how to leverage data transformations, provide organization-wide guidance in the form of data governance rules. Organization-wide governance rules that define which types of data transformations to use and when to use them help to ensure a consistent approach to data transformation.

DON'T COMPROMISE QUALITY

In some cases, data transformation has the potential to reduce, rather than enhance, data quality. For example, if you drop an important column from a data set, you may lose context that you need to make sense of other data within the set.

For that reason, it's important to assess whether there will be any negative data quality repercussions before transforming data. Be sure you can justify your transformations.

ASSESS DATA TRANSFORMATION COSTS

If your data contains sensitive information, such as personal names, credit card numbers, and addresses, data transformation can help you to reduce the risk of misusing the data or accidentally exposing it in a way that violates compliance requirements. With a single pipeline, you can encrypt sensitive fields within data that will be stored, while dropping the fields out to destinations that would cause a compliance violation. This enables you to protect sensitive information while ensuring that the necessary confidential data is available to different sources within your organization. By using data transformation in this way, you can minimize the risk of data breaches and maintain compliance with regulations.

CONCLUSION

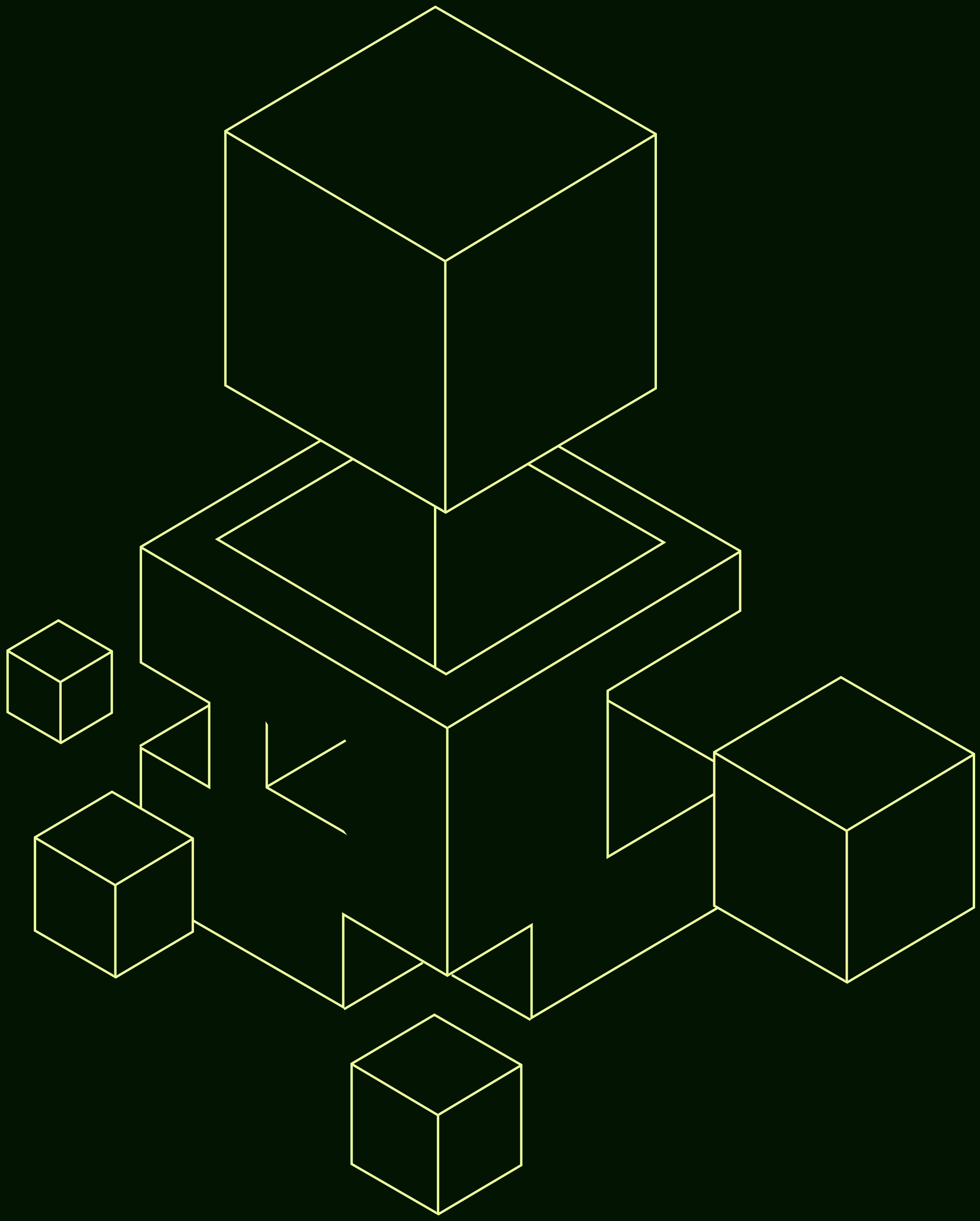
Although not every data source needs to be transformed to achieve every goal, data transformation can in many cases make observability data much easier and more efficient to analyze. It also increases the accuracy and value of analytics results and reduces costs.

In other words, data transformation can significantly increase the value of telemetry data while simultaneously decreasing the time, effort, and money required to use the data. If you're not transforming data within your telemetry pipeline, you're likely missing out on important opportunities and value.

Mezmo Telemetry Pipeline makes it easy to add data transformations to your observability operations. By allowing you to collect data from virtually any source, transform it according to rules you define, and then route it to destinations of your choice, Mezmo gives you maximum control over how your data is optimized for any and all use cases you need to support.

Request a demo to learn more about Mezmo Observability Pipeline.

SALES CONTACT:	OUTREACH@MEZMO.COM
SUPPORT CONTACT:	SUPPORT@MEZMO.COM
MEDIA INQUIRIES:	PRESS@MEZMO.COM



mezmo