

Introduction

Traditionally, one of the major challenges organizations face when they wish to analyze the massive amounts of data they have collected is the necessity to first clean and standardize this data. Cleaning and standardization (also known as harmonization) is the process by which various data sets, often gathered by different stakeholders using varying methodologies, are transformed to become internally and externally consistent, and free of data errors. This process typically takes months to years to complete, depending on the size and complexity of the original data sets.

This ETL workflow (Extract-Transform-Load) is a linear process. Data managers first harmonize the original data, and then load the transformed data into a system, where it can be explored. Scientist end-users, who are domain experts and final data consumers, usually have to wait months, if not years, before they can finally explore the cleaned and standardized data and ultimately make scientific insights.

With Qigram™, we enable a new workflow to engage and incentivize scientists at a much earlier stage of data harmonization. Rather than adopting the linear ETL process, we encourage a more cyclical, iterative ELT (Extract- Load-Transform) process in which data can be loaded into Qigram largely “as is” before being transformed. Scientists can then explore the data by “drawing” diagrammatic queries using Qigram’s unique graphical interface. These queries allow users to immediately gain perspective on their data landscape and understand just how “messy” their data might be. Once scientists understand their data problems, they can then collaborate with data managers to harmonize the data.

This new approach has the inherent advantage of allowing the scientific domain experts to effectively

guide the data harmonization process. For example, scientists can place more focus on high-value subsets of the data, as well as eliminate data sets that are of low interest or low quality. As a result, data managers’ efforts are better optimized because of this added focus.

The scientist end-user also derives a second -- potentially equally important -- benefit from the ELT workflow: they can immediately begin to realize scientific value from their unclean data sets even as harmonization is ongoing. Unlike the ETL process which requires that scientists wait for the harmonization process to be completed before being able to explore their data, the ELT workflow enables scientists to have access to that data from day one. Exploration now becomes concurrent with harmonization and can even partially drive that harmonization. It is important to note that data cleaning and standardization is a multi-step iterative process that is ingrained into an overall data workflow. Figure 1 illustrates this data workflow in its entirety.

For the purposes of this discussion, we will focus on steps 3-10. We will only briefly touch on steps 1 and 2 (data acquisition), which are necessary precursors to data cleaning and standardization, as well as step 11 (making scientific insights into cleaned data), which is described in a separate white paper in more detail.

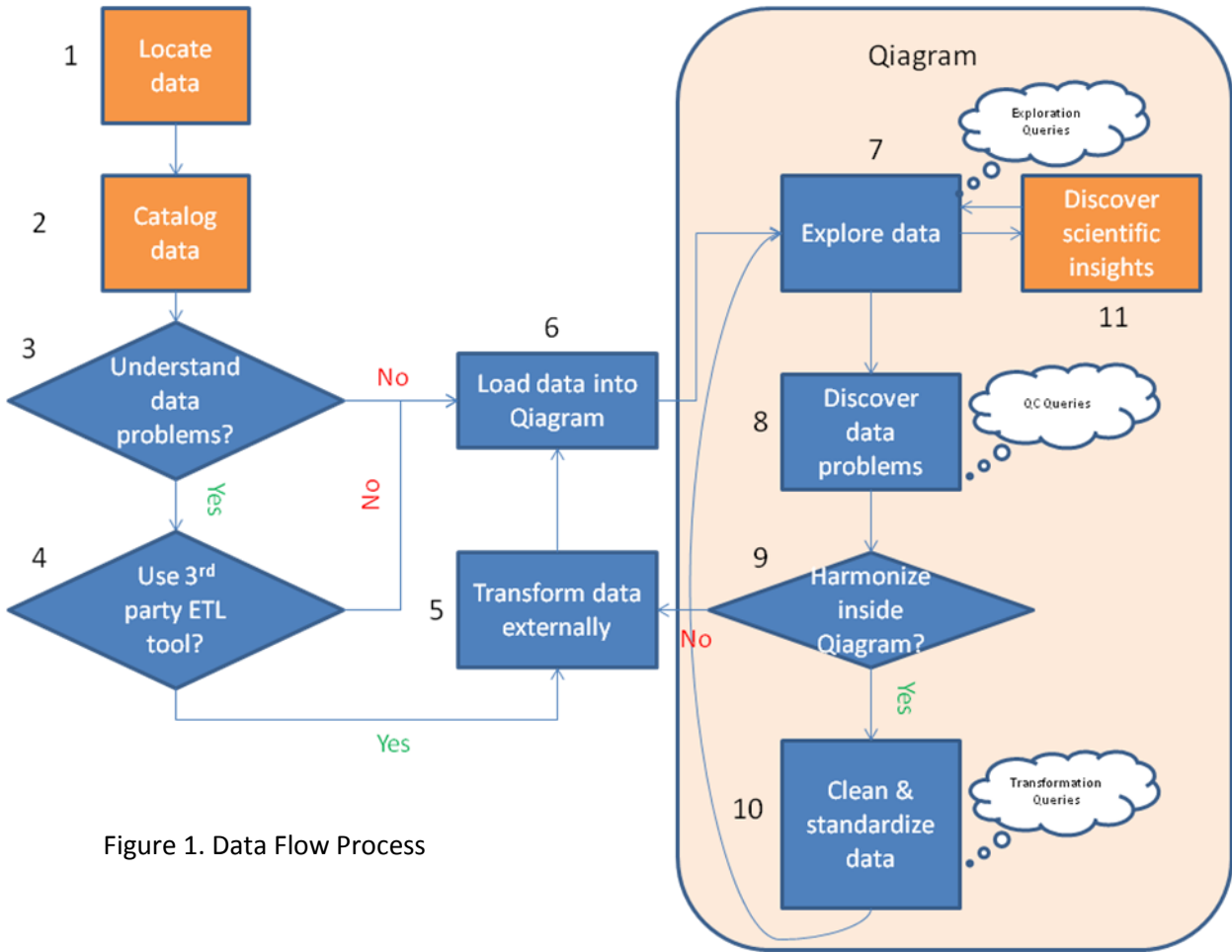


Figure 1. Data Flow Process

Data Acquisition

The first step in the data workflow is to define and locate the data sets of interest. This step usually begins with a series of meetings involving the data owners and various stakeholders in the data workflow process. The data sets often reside in myriad databases, Excel spreadsheets, or even physical notebooks.

Once all the relevant data sets have been located, they should be cataloged (step 2). Cataloging data sets allows downstream data processing tools to have easy access to all the data. This catalog can take many forms, from a file directory structure on a shared drive to a well structured database.

Data Harmonization outside Qigram

Once your data has been centralized, you must decide whether or not you understand your data problems well enough to clean the data before exploring it (step 3).

If you don't understand your data problems very well, we believe Qigram is the best way to help you determine what these problems are by leveraging its data exploration and harmonization tools. The following sections go into this process in much more depth. At this point, we recommended that you load your data into Qigram as is (step 6).

If you do understand your data problems, you may still decide that harmonization is best done within Qiagram. But, you also have other options (step 4). For example, you may choose to build your own custom application to do the harmonization, or you may use another third party ETL tool. It is worth noting that harmonizing your data outside Qiagram does not preclude you from doing further transformations within Qiagram (as described below) once you have had a chance to explore the data.

Data Problem Discovery within Qiagram

In this section, we assume that you have decided to follow the Qiagram-enhanced ELT workflow and have loaded your original data into Qiagram as is. Once you have done this, you will want to first explore your data (step 7) by building various queries, and then discover what problems exist (step 8). This cyclical process of exploration and problem discovery is ideally performed in a collaborative fashion involving the domain experts (scientists) and the data managers (IT, research informatics, etc.) to ensure the efficacy and efficiency of the process.

Qiagram has a number of features that allow you to explore your data and discover data problems.

- The data upload wizard preview screen is a quick way to determine whether or not data being loaded into the system is formatted properly (common culprits are delimiters, escape characters, and single/double quote characters.)
- Once data has been loaded, you can use the Qiagram data browser to analyze the distribution of values for all the fields in your data source. One common result of this analysis is finding data values at the tail end of the frequency distribution that might very well be typos, erroneous values, or mismatched data types.
- You can also build quality control (QC) queries in Qiagram to identify data problems specific to your data sets. Examples of problems that are commonly discovered using QC queries are:

- Field values are outside of the sensible range.
- Values that should be NULL actually contain spaces or some other undesired value.
- Logical relationships between related values are nonsensical.

Data Harmonization within Qiagram

Once you have discovered the problems within your data set, you are ready to fix them. You can choose whether or not you wish to harmonize this data within Qiagram (step 9). If you decide to harmonize within Qiagram (step 10), you do so by creating transformation queries using the following features. Typically, these transformation queries are created by data managers.

- You can define formulas to perform string or numeric transformations, for instance, to fix data key identifier mismatches between data sets. Some examples of such mismatches that can be fixed using formulas include:
 - The ID in data set 1 is S1234, but the same ID in data set 2 is S1234_Brain.
 - A “key” that allows multiple data sources to be related to each other must be a composite of multiple columns such as patient ID, visit date, and sample type.
- You can use mapped unions to harmonize data sources with schema mismatches. Two data sets may capture the same data but have different column names. For example, in two independent studies keeping track of patient information, study A has columns FName and LName while study B has columns First_Name and Last_Name.
- You can use the vocabulary (ontology) browser to relate synonymous but not identical terms between data sets. For example, in two independent studies keeping track of patient information, Study A may store gender values as Male/Female, while Study B stores values as M/F. Similarly, in drug development, the ID of a given compound changes as it moves through the pipeline, migrating from a compound registration name (e.g. M12345) to a

Types of queries	Users	Examples
Quality Control	IT, Scientists	Flag data values outside of normal range
Transformation	IT	Standardization of vocabulary terms; data pivots
Exploration	Scientists	Scientific insights and discovery; hypothesis validation
Operational	Scientists	Reports, dashboards, export to analytics

Table 1. Types of queries and user chart.

drug name X for indication 1, to drug name Y for indication 2.

- You can also use the vocabulary browser to hierarchically organize more complex relationships between concepts so that questions can be asked at varying levels of abstraction. For example, when exploring data about “Nervous system disorders”, you can choose to ask questions about specific subtypes of nervous system disorders or, instead, about the entire class of disorders.
- You can use the pivot operation to re-organize your data from a “short-and-fat” wide table format (many sparsely populated fields) to a “long-and-skinny” format (typically represented as entity-attribute-value triplets). For example, a clinical study where subjects participate in any number of a large battery of tests may have blank values corresponding to unadministered tests for a subject. Rather than storing all subject information in one row (which produces the blanks), it might be more useful to store this data as subject/test name/test result triplets for each test conducted on each subject. This schema can be used consistently for similar measurements across many studies. When it is time to query for data generated from these tests, the results can be pivoted back into the more user-friendly wide table format.
- You can use the set combine/filter features to combine columns from various data sources to create new data sources.

Once created, these transformation queries can be saved and then applied as needed to clean data problems discovered using the QC queries described previously. The combination of the QC queries and the transformation queries represents a repeatable work flow that iteratively produces clean data sets.

Data Exploration

Once you have cleaned your data, you can build exploration queries (or operational queries) in Qiagram to answer your domain-specific questions (or to monitor significant benchmarks) and make scientific insights (step 10). These insights can be made from the very beginning, once data has been loaded into Qiagram, even as data cleaning is ongoing. But, clearly the goal is for data cleaning to progressively facilitate better and easier insight discovery.

Table 1 illustrates the general types of queries that can be built in Qiagram as part of not only the exploration process but the cleaning and standardization process as well.

The following is a summary of the methodology described above:

Action	Description	Deliverable
Meet	Meeting of stake holders, including data holders.	<ol style="list-style-type: none"> 1. prioritized list of data sets of interest 2. clarity on data ownership and data sharing constrains 3. clarity on format and meaning of the data sets
Catalog	Catalog the data sets so that the project team has access to all the data sets from a central repository. This repository can take many forms, such as a shared drive with all the data files, or an Intranet site with links to data sets, or data marts and warehouses.	A central repository of all data sets that are in scope for the project. Data sets are present in this central repository in more or less their “native format”
Inspect	Import data into the data exploration platform largely “as is” to allow preliminary vetting of data to discover and understand data harmonization needs	Clear understanding and documentation of data harmonization needs
Harmonize	Incorporation or creation of standardized vocabularies for both data and meta data in the system. Map to the standard vocabularies. Map to a conceptual framework. Create data QC queries and reports.	<ol style="list-style-type: none"> 1. Controlled vocabularies 2. Data transformations 3. Conceptual framework for any given user base 4. QC queries and reports
Explore	Scientific data exploration	Allow researchers to ask questions
Customize	Any specialized needs that can be built on top of the platform.	TBD
Repeat	Apply the above steps iteratively for the current data sets, as well as any additional data sets	

Conclusion

Before scientists can produce valuable insight by exploring their data, their data must first undergo an intense, rigorous process of cleaning and standardization that prepares it for exploration. The historical approach to this harmonization process – known as ETL – sidelines scientists for months to years as they wait for the data to be transformed by data managers, who are themselves not domain experts. During this time scientists are unable to gain any value from their data.

With Qiagram, we introduce a much better approach, whereby data is first loaded into the system, and then more efficiently transformed through a collaborative

process involving scientists and data managers. Even as this transformation occurs, scientists already have access to the data from day one, and can therefore begin to explore it and make scientific insights using Qiagram’s easy-to-learn graphical interface. The result is that data managers are more efficient in their data cleaning efforts, while scientists are able to make earlier insights. Furthermore, with early participation by a large number of stakeholders in the data workflow, our approach can potentially leverage “crowd sourcing” to address the difficult task of data cleaning and standardization.

For more information please contact us at 443-276-2464 or info@biofortis.com