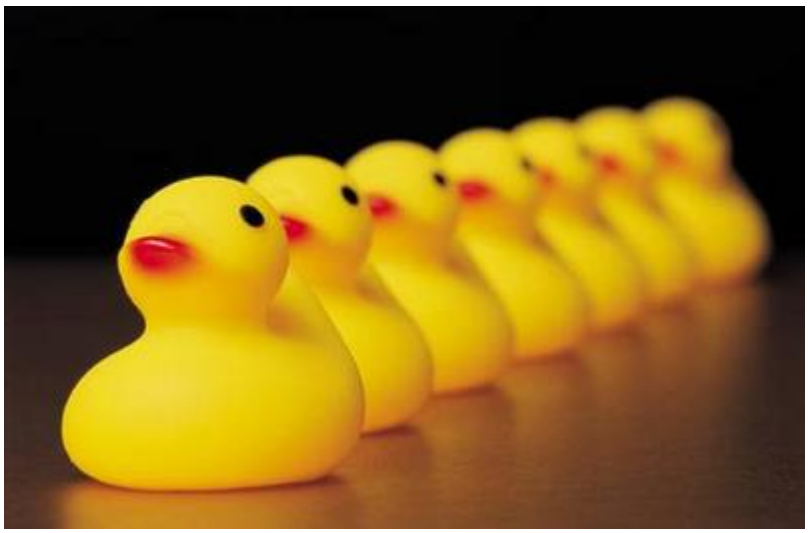
10 Best Practices in Data Preparation



1. A word on Data Governance

Data governance is not data preparation per se, but it's a necessary "wrapper" that defines the business objectives, business glossary definitions, data quality, data auditing, and data lineage standards that data preparation efforts must meet.

Ultimately, business executive stakeholders must own data governance efforts, which requires that they see data as a strategic asset for their business. Some organizations even have a **Data Governance** department on the same level as HR, Finance, Operations, and IT departments. Without this level of focus and organizational commitment to data governance, data preparation efforts will not be as effective as they otherwise could be.

2. Start with good "Raw Material"

It's easy to jump into prepping data without thinking about where the data comes from and the reliability of the source. However, for cases where you'll have to repeatedly load data, the quality, accessibility, and format of the data source can have a big impact on your analytics.

Data sourcing roughly breaks down into three steps:

1. **Defining the data needed for a given business task**
2. **Identifying potential sources of that data, along with its business and IT owner(s)**
3. **Confirming that the data will be sufficiently available with the frequency required by the business task**

There is usually some political wrangling and negotiation included in this step, but it's necessary to secure a reliable data source.

3. Extract Data to a good "Work Bench"

Once you've identified a reliable data source, you need to pull this data into an environment where it can be safely analyzed and manipulated. Smaller data files that have a relatively good native structure can be opened with text editors or spreadsheets.

Larger and/or more complicated data sets will require more **powerful profiling tools**, the likes of which are included with many **Extraction/Transformation/Load (ETL) tools, high-end statistical software, or enterprise-class Business Intelligence packages**.

The point here is to get the data into an environment where it can be closely examined, which is not usually the case with most original data formats.

4. Spend the right amount of time on Data Profiling

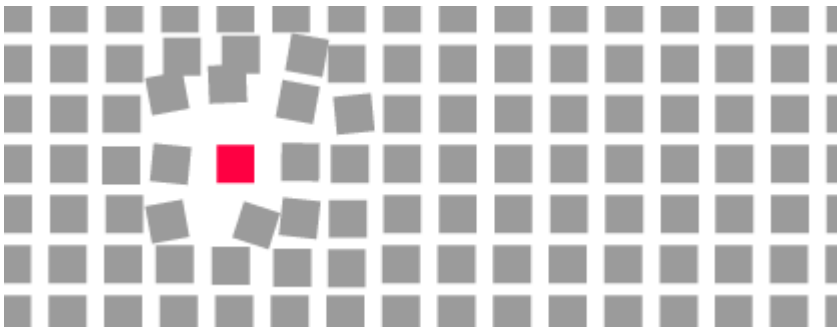
This is the crucial but often overlooked step in data preparation: you really need to get to **know your data** before you can properly prepare it for downstream consumption. Beyond simple visual examination, you need to **profile**, **visualize**, **detect outliers**, and find **null values** and other junk data in your data set.

The first purpose of this profiling analysis is to decide if the data source is even worth including in your project. As data warehouse guru Ralph Kimball writes in his book *The Data Warehouse Toolkit* "Early disqualification of a data source is a responsible step that can earn you respect from the rest of the team, even if it is bad news."

If the data source is deemed worthy of inclusion, results from **data profiling** this source will help you evaluate the data for overall quality and estimate the ETL work effort to adequately cleanse the data for downstream analysis.

5. Start Small

In the **Big Data** era, preparing large data sets can be **cumbersome** and **time consuming**. So start with a random sample of your data for **exploratory analysis** and **data preparation**. Developing data preparation rules on a valid sample of your data will greatly speed your time-to-insight, as it will reduce the **latency** associated with iterative exploration of a very large data set.



This step is a bit of both **art** and **science**. The data analyst should be very familiar with both source data and the business analytics task at hand to zero in on the right columns and rows to sample and eventually prep for further analysis.

6. Zero in on Data Types

Explore the columns you have in your data set and verify that the actual **data types** match the data that should be in each column. For example, a field titled “sales_date” should have a value in a common data format like MM/DD/YYYY.

Similarly, you should understand the generic data type each field represents. If it’s a **numeric** field, is it **discreet** or **continuous**? If it’s a **character** field, is it **categorical** or a **nominal** free text field? Knowing these distinctions will help you better understand how to prep the data contained therein.

7. Your Data ought to be in Pictures

Graphing key fields can be a great way to get to know your data. Use **histograms** to get a feel for the distributions of key fields, **pie charts** to see values as a percent of the whole, and **scatter plots** for the all-important outlier detection. Graphing data has the added benefit of making explanations of data profiling results to non-technical users much faster and more productive.

8. Don’t forget the Sanity Check

Are five-year-old's granted driver’s licenses? Are gas prices $1257 per gallon? Is the average summertime high temperature in San Antonio, Texas -12 degree Fahrenheit? **Sanity checking** means understanding what certain columns represent, knowing a *“ballpark range”* of values that would be appropriate for those columns, and using this understanding and range of values to apply some common sense to the data set.

START_DATE	CUSTOMER_AGE	STATUS
Jan 2002 - Oct 2011	1 - 12	2 Categories
2007-01-03	7	ACTIVE
2004-10-27	9	ACTIVE
2010-02-18	2	CANCELLED
2007-06-30	6	ACTIVE
2006-08-04	7	ACTIVE
2003-06-01	10	ACTIVE
2004-04-23	9	ACTIVE

Additionally, use **automated tools** and **graphing functionality** to find outliers. The challenge with **outliers** is that they can wildly **distort metrics** that use a mean of the data, which can lead to some rather awkward conversations with business stakeholders if you haven’t identified and accounted for those outliers. So, find the outliers, run analysis both with and without them, and present the findings to stakeholders as the beginning of a collaborative, constructive conversation on how to handle them.

9. Iteratively Cleanse and Filter

Based on your knowledge of the end business analytics goal, experiment with different data cleansing strategies that will get the relevant data into a usable format.

Again, start with a small, statistically-valid sample to iteratively experiment with different data prep strategies, refine your record filters, and discuss with business stakeholders.

Upon finding what seems to be a good approach, take some time to rethink the subset of data you really need to meet the business objective. Running your data prep rules on the entire data set will be much more time consuming, so think critically with business stakeholders about which columns you do and don't need, and which records you can safely filter out.

10. Lather, Rinse, Repeat: Bathe your Data

Now that you've developed a data preparation approach on a sample set, run your data preparation steps on the entire data set and examine the results again. Even if you properly sample the test data set, the full data set may still contain unusual cases that could throw off your results, so be ready to iteratively validate and tweak your data preparation steps. Again, be ready for this step to take some time, but the quality of analysis and use trust in the data it will cultivate will be well worth it.

Data preparation is a messy but ultimately rewarding and valuable exercise. Taking the time to evaluate data sources and data sets up front will save considerable time later in the analytics project. Guided by data governance principles and armed with sampling techniques, profiling tools, visualizations, and iterative stakeholder engagement, you can develop an effective data preparation approach that will build trust in the data and earn respect from business stakeholders.