Best Practices in Data Analytics

Daniel Aguilar
AguilarTech.com.au

AguilarTech.io

Introduction

In today's data-driven world, effective data analytics is essential for making informed decisions. This comprehensive 12 step guide outlines best practices in data analytics to help you clean, structure, and analyse data efficiently. Whether you are a novice or an experienced data professional, these guidelines will assist you in maintaining data integrity, optimizing performance, and delivering actionable insights.

1 Clean Data

Ensure data is clean, well-structured, and organized. This minimizes errors and simplifies analysis.

Best Types of Data

The ideal data format is a table with columns (data fields) where the column names remain consistent over time. Each column should contain only one type of data, and all necessary cells should be filled in. Here are some key points to consider:

- System-generated data is preferable to manually entered data, as it tends to be more consistent and reliable.
- Common formats include SQL, CSV, Excel, and JSON.
- PDF data often requires significant cleanup but can still be extracted effectively using tools like Power Query.
- Avoid defining your datasets inconsistently; always use the same definitions on your data to maintain consistency.

Figure 1: Data Sources

2 Unique Key Values

Use unique key values wherever possible. This ensures tables can be joined correctly and without confusion.

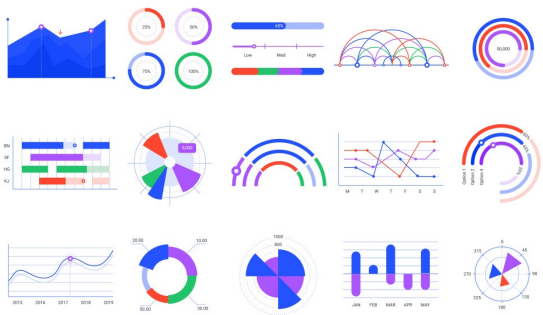
3 Structured Queries

Figure 2: Organise your data

Queries should be structured cleanly and organised using meaningful naming conventions so that others can reuse them. Use accurate and complete names for all queries, tables, columns, and variables.

- **Use Meaningful Names:** Assign descriptive and meaningful names to all queries, tables, columns, and variables. Avoid using generic names like "Query1" or "Table2". Instead, use names that reflect the content and purpose, such as "SalesByRegion" or "CustomerDetails".
- **Avoid Redundancy:** Structure your reporting system in a way that avoids querying the same data multiple times. This reduces inefficiencies and optimises the performance of your reporting system. Use joins and subqueries effectively to consolidate data retrieval.
- **Document Your Queries:** Include comments and documentation within your queries to explain their purpose and logic. This helps others understand the query's functionality and makes it easier to troubleshoot and maintain.
- **Consistent Formatting:** Follow consistent formatting guidelines for writing queries. This includes indentation, line breaks, and spacing. Consistent formatting improves readability and makes it easier to identify errors.

4 Minimalistic Data Design



Avoid having more columns or rows than necessary. By reducing the number of columns and rows, the data becomes easier to read and interpret. End users can quickly locate the information they need without wading through unnecessary data. Smaller datasets are quicker to process, which improves the performance of queries and reports.

- **Identify Essential Data:** Determine which data points are crucial for your analysis and reporting. Focus on including only these essential columns and rows.
- **Remove Redundancy:** Eliminate any duplicate or redundant data that does not add value to your analysis. For example, if a column contains information that can be derived from another column, consider removing it.
- **Use Aggregation:** Where possible, aggregate data to provide summaries rather than detailed transactional data. This can significantly reduce the number of rows in your dataset.

5 User-Centric Design

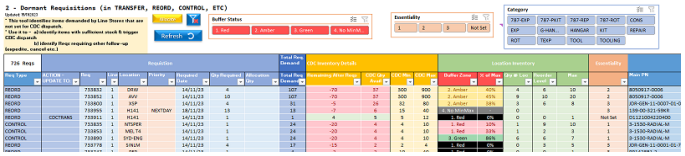
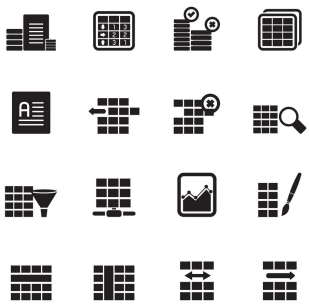


Figure 3: Make tools easy for users to get the actionable data that they need

Always build analytics and tools with the end-users in mind. Reports should effectively communicate insights to drive decision-making.

- Only include what is needed, in the order that it is needed.
- Use color, cleanly labeled columns, and correct information.
- Strive for end users only needing at most one-click for refreshing the data in your reports and tools.

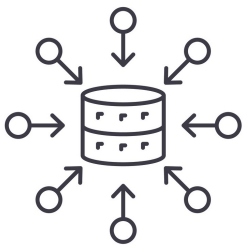
6 Understanding Data Types



Understanding the different types of data is crucial for effective data analysis. Here are the main categories:

- Master Data: Like location and product master data.
- Transactional Data: Such as orders and requisitions.
- Time Series Data: Like orders over time, which can be thought of as cube-type data due to the time dimension.
- Snapshot Data: Data that captures a specific moment in time, like orders for a particular day.

7 Aggregation Knowledge



Understand when to use aggregation functions like 'Group By' and when not to. Recognize which fields can be aggregated (quantities, values) and which cannot (categories, product descriptions).

8 Loyalty to Original Naming

Stick to the original naming conventions of fields and columns. Changing them can lead to confusion.

9 Avoid Hardcoding Filters

Instead of hardcoding specific values, like locations, use reference tables that can be updated as values change over time.

10 Versatility in Tools

Familiarize yourself with a variety of tools to give yourself the toolset knowledge to be able to choose the most suitable tool for a given task rather than choosing a tool just because it is the one you already know. This includes SQL, PowerQuery, Python, Excel, PowerBI, stored procedures, Power Automate, and more.

11 Naming Conventions

Adopt clear and standardised naming conventions in all your work. Being organized and systematic not only enhances your efficiency but is also pivotal for collaborative efforts. If you're careless or inconsistent with naming or organization, it will lead to added work and confusion in the long run. Embracing methodologies such as the Japanese 5S system can be beneficial in this context, especially for eliminating clutter and streamlining processes.

Figure 4: The ISO 8601 international standard for dates is YEAR-MONTH-DATE

12 Normalized vs Denormalized Data

Grasp the concepts of normalized and denormalised data structures. Normalised data reduces data redundancy and improves data integrity by organizing data into separate tables based on their relationships. On the other hand, denormalised data combines multiple data tables into one, often for performance reasons, especially in data warehousing scenarios. Being aware of when to use each structure is essential for efficient data storage and query performance.

For instance, denormalised data might include repeated details that are only needed once. By focusing on normalised data, you can streamline your data sets, making them more efficient to handle and analyse.

- Real world Example: Instead of pull data with both 'part number' and 'part description' in your dataset, request only the 'part number' and merge in the 'part description' only once from your master data table.

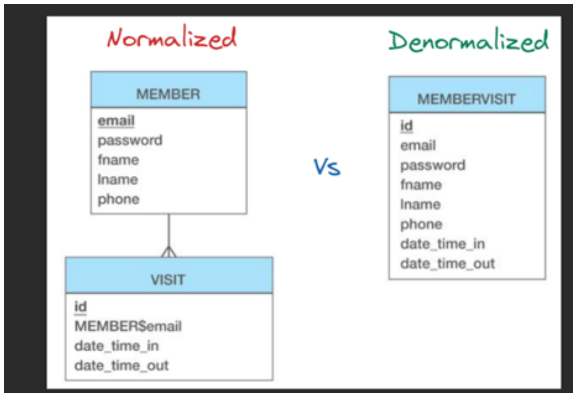


Figure 5: Normalised vs Denormalised data

Requesting a subscription for data from a system with denormalised data can add unnecessary bloat to the data. This means you might receive repetitive information that is not needed, which can significantly increase the size of the data and the resources required to work with it.

- Real world Example: A report that was originally 100 MB in size and causing performance issues was reduced to 2 MB after normalisation, leading to improved performance.

With these principles in mind, data professionals can ensure a comprehensive approach to data analytics, catering to various challenges and needs in the field.