D

Data Aggregation

Tao Wen
Earth and Environmental Systems Institute, Pennsylvania State University, University Park, PA, USA

Definition

Data aggregation refers to the process by which raw data are gathered, reformatted, and presented in a summary form for subsequent data sharing and further analyses. In general, raw data can be aggregated in several ways, such as by time (e.g., monthly and quarterly), by location (e.g., city), or by data source. Aggregated data have long been used to delineate new and unusual data patterns (e.g., Wen et al. 2018). In the big data era, data are being generated at an unprecedentedly high speed and volume, which is a result of automated technologies for data acquisition. Aggregated data, rather than raw data, are often utilized to save storage space and reduce energy and bandwidth costs (Cai et al. 2019). Data aggregation is an essential component of data management, in particular during the "Analysis and Discovery" stage of the data life cycle (Ma et al. 2014).

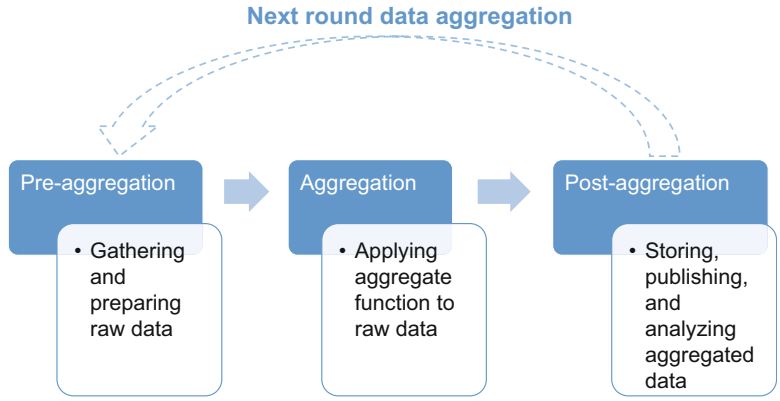
Data Aggregation Processes and Major Issues

The processes of transforming raw data into aggregated data can be summarized as a three-step protocol (Fig. 1): (1) pre-aggregation; (2) aggregation; and (3) post-aggregation. These steps are further described below.

Pre-aggregation

This step starts with gathering data from one or more data sources. The selection of data sources is dependent on both the availability of raw data and the goal of the "Analysis and Discovery" stage. Many search tools are available to assist researchers in locating datasets and data repositories (e.g., Google Dataset Search and re3data by DataCite). Some discipline-specific search tools are also available (e.g., DataONE for earth sciences). Data repositories generally refer to places hosting datasets. For example, Kaggle, an online repository, hosts processed datasets from a variety of disciplines. National Water Information System (NWIS) by the United States Geological Survey (USGS), and STOrage and RETrieval (STORET) database by the United States Environmental Protection Agency (USEPA) both provide access to water quality data for the entire United States. Incorporated Research Institutions for Seismology (IRIS) is a collection of seismology-related

L. A. Schintler, C.L. McNeely (eds.), *Encyclopedia of Big Data*,
https://doi.org/10.1007/978-3-319-32001-4_296-1



Data Aggregation, Fig. 1 General processes of data aggregation

data (e.g., waveform and seismic event data). Data downloaded from different sources are often not in a consistent format. In particular, data from different sources might be reported in different units (e.g., Niu et al. 2018), with different accuracy, and/or in different file formats (e.g., JavaScript Object Notation vs. Comma Separated Values). In addition, missing data are also very common. Before data aggregation at next step, data need to be cleaned and reformatted (noted as "preparing" in Fig. 1) into a unified format.

The most glaring issue in the pre-aggregation step might be related to data availability. Desired raw data might not be accessible to perform data aggregation. This situation is not uncommon, especially in business, since many of these raw data in business are considered proprietary. For example, the information of the unique identifier of persons clicking an Internet advertisement is often not accessible (Hamberg 2018). To resolve this problem, many communities, especially the academia, start to advocate open data and FAIR Principles (i.e., findable, accessible, interoperable, and reusable) when sharing data to the data users.

Aggregation

A variety of aggregate functions are available to summarize and transform the raw data into aggregated data. These aggregate functions include (but are not limited to) minimum, mean, median, maximum, variance, standard deviation, range, sum, and count. In general, raw data can be divided into two types: numeric and categorical. Numerical data are often measurements of quantitative features (e.g., air temperature, sulfate concentration, stream discharge), and they often have mathematical meaning. Unlike numerical data, categorical data are qualitative representations (e.g., city name, mineral color, soil texture). The functions listed above might not be applicable to all types of raw data. For example, categorical data can be counted but cannot be averaged. Additionally, raw data can be aggregated over time or over space (e.g., counting the number of Fortune 500 companies in different cities). The best way to aggregate data (e.g., which aggregate function to use) should be determined by the overarching goal of the study. For example, if a researcher is interested in how housing prices fluctuate on a monthly basis for a few given cities, they should consider aggregating their raw data in two steps sequentially: (1) spatially by city, and (2) temporally aggregating data of each city by month using mean or median functions.

Data can be aggregated into groups (i.e., level of segmentation) in many different ways, e.g., housing prices of the United States can be divided by state, by county, or by city. In the aggregation step, problems can arise if raw data were not aggregated to the proper level of segmentation (Hamberg 2018). Below, an example from a water quality study is provided to illustrate this problem.

In Table 1, a hypothetical dataset of sulfate concentration (on an hourly basis) from a USGS site is listed for 3 days: 01/01/1970–01/03/1970. To calculate the mean concentration over these 3

Data Aggregation, Table 1 Sulfate concentration (in milligram/liter; raw data) collected from 01/01/1970 to 01/03/1970 at a hypothetical USGS site

Sampling date and time	Sulfate concentration (milligram/liter)
01/01/1970, 10 AM	15
01/01/1970, 1 PM	10
01/01/1970, 4 PM	5
01/02/1970, 10 AM	2
01/03/1970, 10 AM	3

days, a researcher should first aggregate concentration by day (each of these 3 days will have a daily mean), and then aggregate these three daily means in order to get a more representative value. Using this approach, the calculated mean sulfate concentration over these 3 days is 5 milligram/liter. Due to the fact that more sulfate measurements are available on 01/01/1970, the researcher should avoid directly aggregating these five measurements of these 3 days since this approach gives more weight on a single day, i.e., 01/01/1970. In particular, direct aggregation of these five measurements yields a biased 3-day mean of 7 milligram/liter, which is higher than 5 milligram/liter by 40%.

Post-Aggregation

In this step, aggregated data might warrant further data aggregation, in which aggregated data from last round of data aggregation will be used as the input "raw data" in the next round. Alternatively, aggregated data might be ready for data analysis, publication, and storage. For example, in the above dataset of aggregated sulfate concentration on a monthly basis, time series analysis can be performed to determine the temporal trend of sulfate concentration, i.e., decline, increase, or unchanged.

Tools

Many tools are available for data aggregation. These tools generally fall into two categories: proprietary software and open-source software.

Proprietary software: Proprietary software is not free to use and might have less flexibility compared to open-source software; however, technical support is often more readily available for users of proprietary software. Examples of popular proprietary software include Microsoft Excel, Trifacta (Data) Wrangler, Minitab, SPSS, MATLAB, and Stata, all of which are mostly designed for preparing data (i.e., part of step 1: data cleaning and data reformatting) and aggregation (i.e., step 2). Some of these pieces of software (e.g., Excel and MATLAB) provide functions to retrieve data from varying sources (e.g., database and webpage).

Open-source software: Open-source software is free of cost to use although it might have steeper learning curve compared to proprietary software since programming or coding skills are often required to use open-source software. Open-source software can be either stand-alone program or package (or library) of functions written in free programming languages (e.g., Python and R). One example of stand-alone program is GNU Octave that is basically an open-source alternative to MATLAB, which can be used throughout all steps of data aggregation. Many programming packages are available for applications in the aggregation step (e.g., NumPy, SciPy, and Pandas in Python; dplyr and tidyr in R). These example packages can deal with data from a variety of disciplines. Some other packages including Beautiful Soup and html.parser help parse data from webpages. In certain disciplines, some packages are present to serve both steps 1 and 2, e.g., dataRetrieval in R allows users to gather and aggregate water-related data.

Conclusion

Data aggregation is the process where raw data are gathered, reformatted, and presented in a summary form. Data aggregation is an essential component of data management, especially nowadays when more and more data providers (e.g., Google, Facebook, National Aeronautics and Space Administration, and National Oceanic and Atmospheric Administration) are generating data at an extremely high speed. Data aggregation becomes particularly important in the era of big data

because aggregated data can save storage space, and reduce energy and bandwidth costs.

Cross-References

- ▶ Data Cleansing
- ▶ Data Sharing
- ▶ Data Synthesis

Further Readings

Cai, S., Gallina, B., Nyström, D., & Seceleanu, C. (2019). Data aggregation processes: a survey, a taxonomy, and design guidelines. *Computing, 101*(10), 1397–1429.

Hamberg, S. (2018). Are you responsible for these common data aggregation mistakes? Retrieved 21st Aug 2019, from https://blog.funnel.io/data-aggregation-101.

Ma, X., Fox, P., Rozell, E., West, P., & Zednik, S. (2014). Ontology dynamics in a data life cycle: Challenges and recommendations from a geoscience perspective. *Journal of Earth Science, 25*(2), 407–412.

Niu, X., Wen, T., Li, Z., & Brantley, S. L. (2018). One step toward developing knowledge from numbers in regional analysis of water quality. *Environmental Science & Technology, 52*(6), 3342–3343.

Wen, T., Niu, X., Gonzales, M., Zheng, G., Li, Z., & Brantley, S. L. (2018). Big groundwater data sets reveal possible rare contamination amid otherwise improved water quality for some analytes in a region of Marcellus shale development. *Environmental Science & Technology, 52*(12), 7149–7159.