Frank, A. U. "Distributed Databases for Surveying." *Journal of Surveying Engineering* 111, no. 1 (1985): 79-88.

Distributed Databases for Surveying

6

By Andrew U. Frank[1]

Abstract: Space related data is gathered in different places and by different services. Considerable duplication of effort to maintain the same data is obvious. The use of a distributed database management system to manage such data collections is proposed. Information and data collection have special characteristics which must be taken into account. They are expensive to update, but can easily be duplicated or transported from one place to another. Different users may produce or use the data and data may be an important asset for a user group. A distributed database may be adapted to such situations and most adequately suit managerial considerations. Users may maintain control over data of importance to them but nevertheless let other users gain access to it.

Introduction

Information is gathered at different locations and within different organizations. Not always is it the same unit of an organization which produces the data and also uses them. An exchange of data between the data producing unit and the unit that needs the data is thus necessary. Computer networks are the most advanced method for such exchanges. In fact, the ease and speed for data transfer in a computer network makes many applications relying on exchange of data possible. Information is becoming one of the most important commodities of modern society. The proper management of data is of prime interest for every organization. It is, therefore, important to analyze flows of data between the producers and the users. Where producer and user of data are separate units, data must be stored and transported between the two. That is the point where databases enter the picture. A common solution is to concentrate all data in a central database where all users can find the data needed. This is only feasible if producer and user are part of an overall organization, and even there, the rational of large centralized facilities is questionable. In many cases it is desirable to form several, independently managed databases that can be used by all users together. Of course, this is only possible if the databases are linked via a computer network.

This solution seems extremely helpful for the management of surveying and mapping data, including land ownership and other legal data (i.e., for Land Information Systems).

Typically, different departments within a town administration, within the public utilities and even private companies gather data describing properties. These data are then used for assessment and taxing, issuing building permits, servicing utility lines, etc. A computerized network and a distributed database are the necessary tools to realize a system to manage such data with minimal duplication and thus least cost.

[1]Dept. of Civ. Engrg., Surveying Engrg. Program, Univ. of Maine at Orono, Orono, Me. 04469.

Note.—Discussion open until August 1, 1985. To extend the closing date one month, a written request must be filed with the ASCE Manager of Journals. The manuscript for this paper was submitted for review and possible publication on October 23, 1983. This paper is part of the *Journal of Surveying Engineering*, Vol. 111, No. 1, March, 1985. ©ASCE, ISSN 0733-9453/85/0001-0079/$01.00. Paper No. 19591.

Outline

The goal of this paper is to present the concept of the distributed database and give some indications how it can be applied. Technical solutions are described, but more stress is placed on the managerial aspects of the problem.

First, a terminology is introduced and the database concept explained. Then the idea of distributed systems is introduced and applied to databases. This leads to an examination of advantages of different solutions and shows the conditions for a specific approach.

Finally, it is briefly explained how the design of a land information system may profit from a distributed database system.

Data and Information

In the following, we will attempt to use the terms data and information with a slight difference in meaning: **information** describes a human being's understanding of the world, whereas **data** means an encoded form of information, suitable for storage, transmission, and treatment in a formal way (i.e., using mechanical means like computers). In many instances, both words can be used, stressing either the idea embodied in information or limiting the point of view to the mere symbols used for encoding the ideas.

Data are very unlike most other goods we deal with in daily life, and these special characteristics of data must always be born in mind. Data collecting is extremely costly, as it is a tedious, labor intensive operation. Data, once collected, can be easily copied, with cost of producing the copies virtually nil. It becomes, therefore, inexpensive to accumulate very large quantities of data. However, such accumulations are utterly worthless unless they are organized to permit human users' access to the information they seek. We may say that a data collection is only worth its access method. If we cannot find, within a reasonable amount of time, the facts we look for, the collection is of no economic value to us.

Organizing large collections of data has proven to be a very difficult task, mainly because it borders organizing information, a thing not well understood today. Management of larger collections of data is difficult and expensive—the larger the collection, the more so. Furthermore, it is not sufficient to organize the collection once, but data must be permanently kept up to date. This requires careful monitoring of the facts of reality and new data gathering and encoding if any changes are observed, which can be even more expensive than the initial data collection effort. A copied version of the initial data becomes worthless if changes in the original are not transmitted and changed in the copy. Maintaining a large data collection up to date, even if no direct data gathering is done, is a very involved process. In a way, we can see that as an application of the third law of thermodynamics to data. A data collection tends to degrade (approaching worthlessness) if no positive actions are carried out permanently in order to keep it up to date.

As a conclusion, data collections are only valuable if they are kept up to date; accumulations of data that are not organized nor updated are worthless (and may be dangerous). Economic value of data collections is determined by their use. A data collection has no intrinsic value nor is the initial cost to gather the data an indication for its value. The only measurement of value is by its use, either how much users are willing to spend to use the data or how much can be saved using the data in the data collection.

Situations of Data Exchange

In the following, we will deal with units which produce, store, or use data. In this context, we define a **unit** as a person or a small group of persons who carry out a certain function within a larger organization. An **organization** is an aggregate of several units with a centralized decision power so that logically connected orders to different units are possible. An example may be the different levels of governmental organizations where the single local agencies form the units and a central decision making unit can direct them. Similar organizational structures are found in large corporations with many partially independent units, which are, however, subject to central decisions.

A unit may **produce** data, either by observing reality and encoding and recording its findings, or by taking over already recorded data from other units and transforming and encoding them in a form suitable for its task. A unit is said to **use** data if knowledge of certain information is necessary for its functioning.

If the producer of a class of data is its sole user, only storage of data may be necessary. However, if production and use of information happen to be done by different units, means of **transporting** the data from producer to user are needed.

Several policies are possible:

1. Data may be kept at its origin and maintained by the producer. Other users can access these data over the network only through the intermediate of the producer.
2. Data may be broadcast over the network after collection; a producer transmits all data collected to all possible users. Users store the data and have unlimited access to them.
3. Data may be collected centrally for several producers. Users can then access the data directly at the centralized locations.

Databases

Large collections of data are difficult to maintain without the help of computers and appropriate software. Computers can store nearly unlimited amounts of data on external disks, and programs can retrieve any data element within less than a second. Generalized software for maintaining any type of data has been designed and is readily available on the market to fulfill this task. Generalized database systems feature a **data description language** to describe the data that must be stored, and **data manipulation language** to store and retrieve data. The user has to set up a description of the database in the data description language, called the **database schema**, and the basic functions for storing and retrieving data are immediately available (1,5). This schema description should be primarily a logical description of the data, augmented by a separate specification for storage methods. Complete database systems contain not only the function of smaller programs for data storage on personal computers (file managers), but provide also techniques to protect the data against mishap. As mentioned before, the data collection, if properly maintained, is an extremely valuable asset and must be protected against misuse by nonauthorized persons (i.e., **data protection**), and against loss through hardware or software failure (i.e., **data security**), and all data entered must conform to certain set standards (i.e., **data integrity**).

A special threat to a database is posed if several users can access the database at the same time to change data (**concurrent access**). A program must control these accesses and check that they do not interfere with each other. It may happen that the changes of one user wipe out changes another user introduced just before, or, even worse, an interference between two users may disturb the data storage system and cause loss of large parts of the database. If more than one user can possibly access a database at the same time, a secure method of control of concurrency must be in place; a method, that prevents accidents most of the time is insufficient. Especially, it is not acceptable to rely on user actions, not controlled by the computer, as only the database system can detect such interferences, since they may happen even if the information the two users treat are seemingly not connected.

Distributed Systems

Reasons to combine several computers into a cooperating system are: (1) Constructing systems of higher performance from simpler and less expensive machines, especially achieving systems with extremely high performance through combination of hundreds of smaller microprocessors; (2) building systems that are fault tolerant, i.e., continue to work even if some hardware parts fail; (3) distributing systems over several locations, in order to provide most of the services locally and to reduce data communication cost.

In an overall view, several computers connected in a network working at a common task can be regarded as one single system. We must then examine how the different functions are shared among the hardware components. In order for the machines to cooperate and fulfill the common task, they need a means of exchanging data. Two solutions are often used:

1. Sharing of memory.—The machines use a common part of data memory to deposit data and to read data deposited by other participating machines.
2. Messages.—The machines pass messages between them.

Coupling machines using shared memory is typically found in large multi-processor mainframe systems where several processors, situated near each other (e.g., in the same room), work together to achieve higher

Machines in a spatially distributed system communicate with messages, transmitted over the **communication network** (2). The actual distance between the machines does not influence the basic method much. Decisive is the fact that cooperating machines must communicate using unreliable network links. It is not technically possible to design communication links that are, at the same time, free of communication errors and without delays. However, we may trade between the two goals.

Using lower level network protocols (cf. Ref. 4) it is possible to check for transmission errors and provide logical network connections that are (to any desired degree) error free. This, however, increases variable time delays in the transmission.

By the term distributed system we understand a collection of cooperating computers connected by unreliable communication links. The amount of functions distributed may vary and include either the complete operating system (distributed operating systems) or only certain functions, e.g., the database (distributed database). The distributed function may be a single program, installed on several machines of the same or different manufacturers running the same or different operating system. Obviously, the case of a uniform hardware is simpler to solve.

Motivation for Distributed Databases

In situations where several units produce data that are then used by other units to base their decisions on, we have to study how to make available the necessary data to each using unit.

If there is only a small amount of data involved and changes to the data are relatively infrequent, traditional methods may be used to best advantage. Examples are lists of zip codes, which are easiest made available to all users in printed form, as only a modest amount of information is involved and changes occur seldom. The telephone directory is already bordering the limits of such methods since the regular distribution in printed form is wasteful. Most of the data received by the subscriber are never used, but very often the update interval is too long, and data for new numbers not present in the directory are sought. The French telephone company has undertaken a study to evaluate whether other forms of making this information available to clients would not be more advantageous, and found that a central database that the subscriber can investigate would be a cheaper solution.

The distribution of all data to all potential users can also be implemented using computerized systems. Included among the benefits of a system where each user has his own copy of all data in the system is the very fast and independent access of each user unit to all data it ever needs. However, each unit will most likely have to cope with very large collections of data, much larger than it actually requires; most of the data it receives are not relevant to its use.

There will be more equipment needed to store the data and, more importantly, more complex software to organize the larger amounts. We call such a policy a system with **full replication** of all data. The problem with full replication is updating. Every change that is made to a data element must be broadcast to all other copies, i.e., to all other units. This creates not only a heavy load on the computer network linking the any unit, and has, therefore, the same load to carry as would one central computer serving all units at once.

This leads to another policy, the **centralized system**. All data are stored in one central location where all units deposit their new or updated data, and from where they can get the data they need for their decisions. This changes the load pattern; retrieval of data must now pass over the network to the central unit, and the response must travel back. That may create a heavy load on the central computer. However, updates of data are transmitted and processed only once in the central computer without involving the other units.

A third policy is to **decentralize data** so that each unit holds a subset of the data that does not overlap with the data held in other units (no replication). If a unit needs data or wants to update data, it has first to find out on which computer these data are stored, and then transmit its order to the appropriate unit. If the subset of data used by different units is stable, and data can be distributed in a manner to have each unit primarily use data it manages locally and only rarely access data managed by other units, then this solution leads to optimal performance, as each unit deals mainly with its own load, and there is only a small additional load to service other units' request. This also leads to minimal load on the communication network, since most operations can be processed locally without communication.

Unfortunately, practical problems seldom fall into one of these three ideal categories. Most of the time, combinations of three policies are required. The most general case, which is usually understood when the term **distributed database system** is used, is a generalization of the decentralized system, including some replication of data that are heavily used by more than one unit.

The data in a distributed database are described globally in an overall database schema. This schema will not only contain the logical data description, but must also indicate how data storage is to be distributed among the different machines. It is also possible that a separate schema is established for each machine and then communicated to the participating machines. This may involve some complex data translation methods to convert data from one form of representation to another.

Operational Data and Statistical Data

Data may be used, in daily operations, as a base for concrete decisions in specific cases. This use shall be called **operational**. On the other hand, management uses individual data to abstract statistical information from it which is then used for strategic planning. Use of data to this end shall be termed **statistical**.

Data used for operational decisions must be accurate. If parts are shipped to a client, the number and type of parts he ordered must be known exactly; if a construction permit is considered, the number of stories recorded in the database must be the same as in the original document, lest an erroneous decision should be taken, and complicated and expensive actions of correcting will ensue. When data are used for ag-

Data used for such applications are therefore not required to respond to the same level of accuracy.

The difference in accuracy required may affect the amount of precautions taken to prevent erroneous data in the collection. As such actions are generally expensive, it may be worthwhile to apply them only where needed.

Managerial Considerations

Data are important assets in our society. Thus, control over data means increased importance and power. A unit within an organization will therefore not easily give up control over data, as this would result in a loss of power and influence to the unit.

Additionally, we have recognized that data are only valuable if carefully managed and permanently updated. If new data are introduced without checks, a data collection can easily become worthless—not only the erroneous data are worthless, but also the rest of the data, which are, in principle, correct, becomes worthless, as it is too difficult for a user to find out which data contain reliable information and which do not.

Units that depend on data that they collect will, therefore, not be willing to let other units change their data, fearing they might introduce errors, which in turn will affect the performance of the unit using the data. In conclusion, we see that data collections are valuable assets for which clear responsibilities must be identified. Technical precautions must be taken in order to assure the units that the domains of their responsibility are secure, and that no accidental or malicious changes can be introduced into the data for which they are responsible.

If several users are cooperating, even within one unit (the preceding definition of a unit is intentionally open to allow for consideration of units within units), it will be necessary to record changes to the data, together with the name of the person executing them, the data of execution, and (perhaps) an authorizing approval from a superviser. Present day database management systems do not cater to such requirements, and only recently the so-called "version" problem has been introduced into database research. Additional investigations in this area are done under the heading of "office automation."

Only in organizations with strong central decision making and control over units, a centralization of data from different units may be achieved. Very often, such centralization is limited to statistical, aggregated data, to reduce the amounts of data managed centrally and to minimize dependency of decentral units on central resources.

Technical Advantages of Solutions

Full Replication.—If there is only a modest amount of data that are seldomly changed and if most operations are retrieval operations, data replication is generally a good solution. Only the rare updates will pass over the network, and all the retrieval operations are executed locally. On the other hand, if data are very often updated, replication is not an appropriate solution, as all updates have to be broadcast to all other users.

Systems with replicated data can continue to furnish data even if parts of the system are not functioning properly. Updates in a situation when only part of the machines are working (or some network links are "down") are not simple, and complex schemes have been devised to reduce influence of a single failure.

Distribution of Data over Nodes with No or Only Partial Replication.—To retrieve data in a system with incomplete replication, knowledge about where a certain data element is stored is necessary. If the data describing the distribution of data is centrally maintained, failure of this central node puts the whole system out of work. For many practical applications, it can be decided in advance where each datum is to be stored, and this description is seldom changed. It is then advantageous to replicate this schema information and store it in each node.

Centralized Storage.—Centralized storage may be useful if no patterns of local use are discernable. If the likelihood that a unit needs a data element is uniform and is independent of the unit and the data element, then central storage may be recommended. Obviously, such a system is vulnerable to failure of this central node, but other methods to reduce down time may be appropriate (hot standby). Typical examples in this category are the reservation systems of passenger airlines.

Application to Surveying Data

There is much interest in storing data related to land, including ownership data, land utilization and zoning, etc., and making it available for different services. Ideally all such data would be acquired only once and then provided for all interested users. Such a system, usually called multi-purpose cadastre or land information system (3), will result in considerable savings as duplication of efforts to collect and update data by several services can be avoided.

Data describing properties are gathered today by different departments and services within a town and county government, and each unit presently collects and maintains all the data it needs to carry out its function. Evidently, the same data are collected and maintained in many files, and even then cannot be used for a new task (e.g., statistics for planning purposes).

A common database would reduce the duplication of efforts necessary today for maintaining these duplicated files and would also prevent discrepancies between the different service files.

A single centralized database is a technically possible solution, but does not cater to the independence of the units concerned. A distributed database can solve these managerial requirements in an optimal way. Independent units can retain control over "their" vital data but at the same time make them available to other users. They are guaranteed that no other unit can change data they are responsible for, and they do not

depend on some other units' operations (e.g., [illegible]) located out-of-house).

Land Information Systems are relatively easy to implement as distributed databases for the following reasons:

1. Data are static and updates seldom. This makes replication of data attractive for all data that are routinely used by several units.
2. Responsibility for data can be clearly determined in advance. A unit is usually responsible for updating certain categories of data (ownership data, building characteristics, utility lines) within an exactly determined area. Overlapping of areas of interest is very seldom, as jurisdictions are not overlapping.
3. Many units are not interested in individual data but in statistical averages. Aggregation can easily be carried out when transferring data from the collecting unit to the user. Privacy regulations may restrict a collecting unit's rights to pass data to other units, but such restrictions do not apply to anonymous aggregated data.

Distributed databases for this application can be built when the connecting network is available. The requirements can be satisfied by today's technology.

In order to proceed, an analysis of the intended use is necessary. Then the data to be included in the system and their logical connections (data structure) have to be identified. In addition to the regular description of data, it must be determined (for each data element) which unit is responsible for collecting and updating data, where data are to be stored and who may use them. It will be necessary to decide on common methods of encoding information to minimize possible misinterpretation of the data by other units.

Conclusions

Using data that are stored in different computers belonging to different users as if they were stored in as a single large database is feasible, if the computers are linked using a network. User interfaces are available that do not require the user to understand the distribution of the data over the machines, and enable him to use the data as if it were one central database. Nevertheless, the data can be distributed and each organizational unit can maintain full control over "his" data.

In order to organize such a shared use of data collection, computer networks must be installed and database software must be used that support data distribution. On the nontechnical side, the structure of the data, their precision, and timeliness must be defined, and then the organizational unit responsible for collection and updating of the data identified. Such a solution has great organizational advantages and is technically feasible today.

Appendix.—References

1. Date, C. J., *An Introduction to Database Systems*, Addison-Wesley Publishing Co., Vols. 1 and 2, 3rd ed., 1982.

2. Davies, D. W., et al., *Distributed Systems—Architecture and Implementation*, Springer Verlag, Berlin, Germany, 1981.
3. FIG (International Federation of Surveyors), Resolution 301, XVI International Congress, Montreux (Switzerland), 1981.
4. ISO, *Reference Model of Open Systems Interconnection*, ISO/TC97/SC16 N227.
5. Ullman, J., *Principles of Database Systems*, Computer Science Press, 1982.

DISCUSSION

Note.—This paper is part of the *Journal of Surveying Engineering*, Vol. 111, No. 1, March, 1985. ©ASCE, ISSN 0733-9453/85/0001-0091/$01.00. Paper No. 19523.