MTA Számitástechnikai és Automatizálási Kutató Intézete, Közlemények 24/1980.

ON A CERTAIN DISTRIBUTED DATA BASE MODEL

Józef Kubit
The Academy of Economics in Cracow
Department of Computer Science
Poland

ABSTRACT

The paper presents a formal model of distributed data base. Present work is an attempt at defining fundamental notion connceted with decompositions of data base model. Distributed data base model consists of family of individual (local) data base models and administrator of distributed data base model. Model is based on a definition of relational schema of data. Data are the object of widely understood measurement.

1. INTRODUCTION

In recent yeast many different models have been presented describe the data base at the conceptual (mathematical, logical) level. The rapid development of large computer networks led to investigations about distributing data base throughout such a system. Looking at these trends we can identify two basic approaches [Neuhold 1977]:

1) Using already existing networks of large computer systems,
2) As an alternatice, we may build a new data base management system which takes into account the existence of a computer network but does not rely on already existing indivudual (local) data base systems.

We envision for thenext future, that computer networks commonly will encompass both large minicomputers even including inteligent terminals.

The reason for the interest in distributed data base systems is because the provide a solution to very real problems for the geographically distributed organization which needs to preserve a unified information--sharing and processing system.

*The paper has been presented on The International Colloquium COMPCONTROL'79 in Sopron on November, 1979.

In general, the advantage of centralized approach are the disadvantages of a distributed approach and vice versa.
They are, in broad form [Champine 1977]:
Distributed advantages /centralized disadvantages

- communication failsoft capability
- lower communications data rate and cost
- configuration flexibility
- high system performance (fast response and high transition rate)
- modular implementation
- modulat up-grade

Generalized advantages / distributed disadvantages

- operations economy
- hardware economy of scale
- unified control
- easy updata / retrieval
- compatibility.

Gradually along with a vigorous development of data base systems based on Codd's relational model of data, a new theory of relations has emerged. The birth of this theory must be assigned to Cood's papers [1970] and [1971] where the basic notions were introduced and collected together.

Depending on delimited purposes of realized work the notion of data base model is not used similarly. It concerns for example the works devoted theory of i.s.r systems [Marek and Pawlak 1976].

General comparison of Codd's and Pawlak's and Marek's approach was done by Koczkodaj [1978] who showed that they lead to essentially same consequences.

Many data models have already been proposed, wach having its own concepts and terminology. In studying the various models is becomes apparent that they have similarities and differences which are not trivial to analyze [Kerschberg, Klug and Tsichritzis 1976].

The notion of data base model is connected with the notion of information system model. Information system model in [Kubit 1979] consists of four specialized system models. Two of them constitute a certain data base model.

Information system model is component of controlled economic system model.

The purpose of the paper is a presentation of divided - distributed data base model.

In each of distributed data base cases we deal as follows:

The base B is decomposed into smaller entities $B_1,\dots,B_i,\dots,B_n$ and we add new object called administrator ($\bar{A}$). The administrator of queries. Same of i'S could be empty.

The response for i is the settheoretical union of the responses of B_i to i_i.

Of particular importance may be a formulation of relations occuring between notions of data base model and administrator of data base model.

2. NOTION OF DATA

Let's take an attribute names **N**. This set will be named alphabet of data names. From elements of this alphabet we'll make a word set $\mathbf{N}^*$.

Let N be a subset of $\mathbf{N}^*$ set. N set will be named a set of data names. Data names will be marked by N while for all N there will be

$$N = < n_1,\dots,n_j,\dots,n_{r_N} > \quad n_j \in \mathbf{N}.$$

Denote $<n_1,\dots,n_j,\dots,n_{r_N}>$ will be also defined relational schema of data.

Let's take a nonempty family of domains $\{M_i\}_{i \in I}$ and mapping m.

The mapping m need not be one-to-one. This is onto mapping

$$m: \mathbf{N} \twoheadrightarrow \{M_i\}_{i \in I}$$

According to what was said above for relational schema of data will correspond a finite relation

$$R_N = m(n_1)\mathbf{X} \dots \mathbf{X} m(n_j)\mathbf{X} \dots \mathbf{X} m(n_{r_N})$$

The set of data values corresponding relational schema N will be defined as $\gamma(N)$, $\gamma(N)=\mathbf{P}(R_N)$ where $\mathbf{P}(R_N)$ R_N.

By V will be denoted an element of subset $\mathbf{P}(R_N)$, $V \in \mathbf{P}(R_N)$. Element V will be named data value and

$$V = \langle v_1,\ldots,v_j,\ldots,v_{r_N} \rangle$$

DEFINITION 1

Data d will be named a pair composed of data name and data value $d = \langle N, V \rangle$.

Names n_j are the names of specified attributes, whereas parts v_j of data value V are the values of these attributes. Attributes which is defined by n_j includes the definition of scale and that of value unit v_j. Further on will be identified as equivalence class of data $\dot{d} = \| d \|$.

Data will be obtained in the process of measurement in the broadest sense of the notion. Data measurement inotion will be defined by means of data measurement function. Data measurement function will be referred to as isomorphism from real objects (portions of productive factors) algebra, onto data algebra

$$h: \mathbf{A} \rightarrowtail\!\!\!\twoheadrightarrow \mathbf{D}$$

where: A - universe of algebra of equivalence classes of real objects $\mathbf{A} = A \lesssim O_1,\ldots,O_n \rangle$

D - universe of data algebra $\mathbf{D} = \langle D, O_1',\ldots,O_n' \rangle$ (cf. [Kobayashi 1975])

Dated $\dot{d} \in D$ are data equivalence classes because of defined equivalence relation. Algebras $\mathbf{A}$ and $\mathbf{D}$ are similar algebras. For these two algebras the following theorem holds true.

Theorem 1

Let f be one-to-one transformation of A_0 set of algebra generators onto D_0 set of similar algebra **D** generators. If there exists homomorphism h of algebra **A** into algebra **D** being an extension f and also homomorphosm g of algebra **D** into algebra **A** being an extension f^{-1}, then h is isomorphism **A** onto **D** and $h^{-1}=g$.

3. SYSTEM OF DATA RETRIEVAL

We assume that I_N denotes a set of attribute names according to the relational schema of data (data name) N. By m_N will be understood restriction of m function to I_N set of attribute names

$$m_N = M \uparrow I_N$$

Let $\bar{U}$ be a set of all functions m_N for set **N** of data names

$$\bar{U} = \{m_N\}_{N \in \mathbf{N}}$$

Let I_N^* denote a set of all words overthethe alphabet I_N of attribute names of data N. The elements of I_N set will be marked as $\bar{n}$.

By $\mathbf{P}(R_N)[\bar{n}]$ we shall define a set of all data value V restrictions to a set of attribute names $\bar{n}$ according to a relational schema of data name N

$$\mathbf{P}R_N[\bar{n}] = \{V\,\bar{n} : V \in \mathbf{P}(R_N)\}$$

Let $\mathbf{P}(R_N)^*$ be understood as a set of all projections $\mathbf{P}(R_N)$ for all $\bar{n} \in I_N^*$. By **P** we shall mark a set of all projections $\mathbf{P}(R_N)^*$ for each data name N

$$= \{(R_N)^*\}_{N \in \mathbf{N}}$$

We assume the Θ_N^j will denote a set of two-argument order relations j-type values domain assigned to the data name N. Two-argument order relations Θ_N^j give a set Θ of order relations of data name attribute values which is expressed in the following way

$$\Theta = \{\{\Theta_N^j\}_{j\in\{1,\dots,r_N\}}\}_{N\in \mathbf{N}}$$

By Q we shall mark a set of set theory operations [Neuhold 1974] defined for **P** set.

D E F I N I T I O N 2

System of data retrieval for D set of data will be marked seventuple

$$S_{DR} = \langle D, \mathbf{N}, N, \bar{U}, \quad , \Theta, Q \rangle$$

where:

- D - universe of data algebra
- **N** - alphabet of data names
- N - set of data names
- $\bar{U}$ - set of functions m_N (assigning attribute values domains) for data defined by relational schema from set
- **P** - set of all projections of data values from data set D
- Θ - set of two-argument order relations of attribute values
- Q - set of set theory operations for set of data values projections.

4. SYSTEM OF CONTROL SYSTEM SETS OF DATA RETRIEVAL

Let set **N** be alphabet of subset of data names for set D. From elements N of set **N** we compose of set of all words $\mathbf{N}^*$. Let $\hat{\mathbf{N}}$ be a subset of $\mathbf{N}^*$ set. $\hat{\mathbf{N}}$ will be names a set of names of data subsets. The names of data subsets will be expressed by $\hat{N}$, where for each $\hat{N}$ there will be

$$\hat{N} = \langle N_1, \dots, N_k, \dots, N_{w_N} \rangle \; ; \qquad N_k \in \mathbf{N}$$

The denote $\hat{N}$ will be defined as a name of subset or relational schema of data subsets (sets).

We say that $\hat{N}$ is a lattice with respect of the operation $\cup$ (join) and $\cap$ (meet) if the following equations hold (axioms of lattice) (cf. [Kuratowski and Mostowski 1976]).

(1) $\hat{N} \cup \hat{N} = \hat{N}$ $\qquad$ $\hat{N} \quad \hat{N} = \hat{N}$

(2) $\hat{N} \cup \hat{N}' = \hat{N}' \cup \hat{N}$ $\qquad$ $\hat{N} \cap \hat{N}' = \hat{N}' \cap \hat{N}$

(3) $\hat{N} \cup (\hat{N}' \cup \hat{N}'') = (\hat{N} \cup \hat{N}') \cup \hat{N}''$ $\qquad$ $\hat{N} \quad \hat{N} \quad \hat{N} = \hat{N} \quad \hat{N} \quad \hat{N}$

(4) $\hat{N} \cap (\hat{N} \cup \hat{N}') = \hat{N}$ $\qquad$ $\hat{N} \cup (\hat{N} \cap \hat{N}') = \hat{N}$

We call a lattice distributive if

(5) $\hat{N} \cap (\hat{N}' \cup \hat{N}'') = (\hat{N} \cap \hat{N}') \cup (\hat{N} \cap \hat{N}'')$ $\qquad$ $\hat{N} \cup (\hat{N}' \cap \hat{N}'') = (\hat{N} \cup \hat{N}') \quad (\hat{N} \cup \hat{N}'')$

We introduce an order relation between elements of lattice:

(6) $\hat{N} \leq \hat{N}' \equiv \hat{N} \cup \hat{N}' = \hat{N}'$

or, equivalenty,

(7) $\hat{N} \leq \hat{N}' \equiv \hat{N} \cap \hat{N}' = \hat{N}$

Similarly we define the elements o and i as the elements satisfying conditions

(8) $\hat{N} \cup o = \hat{N},$ $\qquad$ $\hat{N} \cap i = \hat{N}$

It is easy to show o is the smallest element in $\hat{N}$ and that i is the largest, namely for every $\hat{N} \in \hat{N}$

(9) $o \leq \hat{N} \leq i$

Let's take a set I of control system queries. We assume that for every query correspond a finite nonempty set $\hat{N}_i \in \hat{N}$, such a way that

$$\bigcup_{i \in I} \hat{N}_i = \hat{N}$$

and for all $i, j \in I$, $i \neq j$ $\hat{N}_i \cap \hat{N}_j = \emptyset$

Instead of speaking about the partition $\{\hat{N}_i\}_{i \in I}$ we may consider the equivalence relation R_I $\hat{N} \times \hat{N}$ on $\hat{N}$ that the family of its equivalence classes is indexed by the set I and

$$\hat{N}/R_I = \{\hat{N}_i\}_{i \in I}$$

DEFINITION 3

A system of control system sets of data retrieval is a sixtuple

$$S_{CSSDR} = <D, \hat{N}, \cup, \cap, R_I, U \qquad \text{(cf. [Raś 1978])}$$

where:

D - data set

$< \hat{N}, \cup, \cap >$ - lattice of names of data subsets

R_I - an equivalence relation in and I is set of control systems queries

$U: \hat{N} \to 2^D$ - monotonic function satisfying the following conditions:

$$(\forall \hat{N}, \hat{N}' \in \hat{N})(\hat{N} \leq \hat{N}') = U(\hat{N}) \subseteq U(\hat{N})$$

$$(\forall d \in D)(\exists \hat{N} \in \hat{N})(d \in U(\hat{N}))$$

D E F I N I T I O N 4

Let S_{CSSDR} be a system of control system sets of data retrieval. Let $D' \subseteq D$.

a) D' is said to be describe within S_{CSSDR} iff there is $\hat{N} \in \hat{\mathbf{N}}$ such that $U(\hat{N})=D'$.

b) $\mathbf{B}(S_{CSSDR})$ is the family of allsubsets of D descriable within S_{CSSDR}

D E F I N I T I O N 5

A system is selective iff for all $d \in D$, $U(\hat{N}_d)=\{d\}$

Theorem 2 (cf. [Lipski and Marek 1975])

A system S_{CSSDR} is selective iff $\mathbf{B}(S_{CSSDR}) = 2^D$ (recall that we consider only the case when I (set of control system queries), and consequently D, are finite).

P r o o f:

If $\mathbf{B}(S_{CSSDR}) = 2^D$ then obviously S_{CSSDR} is selective. If S_{CSSDR} is selective then all $d \in D$ have different descriptions, hence D is finite. For any $D' \subseteq D$ we have then

$$D' = U(\sum_{d \in D} \hat{N}_d),$$

i.e. D' is descriable.

5. A DATA BASE MODEL.

DEFINITION 6

Data base model of productive object is a pair

$$DB = \langle S_{DR} , S_{CSSDR} \rangle$$

where:

S_{DR} - system of dara retrieval

S_{CSSDR} - system of control system sets of data retrieval

The following properties of data base model presented above can be distinguished:

1. Model is based on a definition of relational schema of data as proposed by Codd [1971].
2. Data are the object of widely understood measurement.
3. For the presented model i.s.r, system theory can be applied.
4. Identification of data base follows the identification of information system for productive object.
5. Presented data base model can be used in the construction of distributed data base model.

Remark

Presented data base model is a proposal resulting from necessity for complex analysis of problems connected with its modelling. It concerns primarly the contents of data base.

6. DISTRIBUTED DATA BASE MODEL AND ITS PROPERTIES

D E F I N I T I O N   7

System of storage and retrieval of documents (cf. [Marek and Pawlak 1976]) is quadruple

$$I = \langle X, B, R_J, V \rangle$$

where:

- X - set of documents
- B - set of descriptors
- R_J - equivalence relation on B of finite index
- V - maps B into 2^X $(V: B \to 2^X)$ and satisfied the following two conditions:
  1) if $aR_j b$ $a \neq b$, then $V(a) \cap V(b) = \emptyset$
  2) $\cup\{V(b) : bR_J a\} = X$ for each $b \in B$.

D E F I N I T I O N   8

Let $S_{CSSDR} = \langle D, \hat{\mathbf{M}}, \cup, \cap, R_I, U\rangle$ be system of control system sets of data retrieval and I be system of storage and retrieval of documents. System I coverage S_{CSSDR} IFF

1) there exists an one-to-one function Ψ such that Ψ: $J \to I$
2) there exists a function

$$\phi : \bigcup_{j \in J} \hat{N}_{\Psi(j)} \to B \quad \text{such that}$$

$$V(b) = \cup\{U(\hat{N}) : \hat{N} \in \phi^{-1*}\{b\}\} \qquad \text{and}$$

$$\hat{N} \in \hat{\mathbf{N}}_{\Psi(j)} \Rightarrow \phi(\hat{N}) \in B_j$$

The intuition which is connected with this following: I is a "presystem" classifying the names of data subsets in according to some documents from set of documents X.

DEFINITION 9

Let S_{CSSDR} be system of control system sets of data retrieval and I be system of storage and retrieval of documents. We say thet strongly covers S_{CSSDR} iff covers S_{CSSDR} and

$$1) \quad (\forall b)_B \quad card(\phi^{-1}*\{b\}) \geq 2$$

and

$$2) \quad \forall i_I \quad card(\hat{N}_i) \leq 2)$$

Let K be set of indexes of existing individual (local) systems of control system sets of data retrieval.

DEFINITION 10

Let S_{CSSDR} be system of control system sets of data retrieval. Let $\{I_k\}_{k\in K}$ be a partition of the set I. An induced family $\{S_{CSSDR_k}\}_{k\in K}$ of individual (local) systems of control system sets of data retrieval is formed as follows:

$$S_{CSSDR_k} = \langle D_k, \hat{N}^{(k)}, U_k, \cap_k, R_{I_k}, U_k \rangle$$

$$1) \quad N^{(k)} = \bigcup_{i\in I_k} \hat{N}_i$$

$$2) \quad R_{I_k} = R_I \cap (\hat{N}^{(k)} \times \hat{N}^{(k)})$$

$$3) \quad U_k = U \uparrow \hat{N}^{(k)}; \quad \cap_k = \uparrow \hat{N}^{(k)}$$

$$4) \quad U_k = U \uparrow \hat{N}^{(k)}$$

$$5) \quad D_k = \hat{N}^{(k)} \in \hat{N}^k \quad U_k(\hat{N}^{(k)})$$

By S_{DR_k} will be understood restriction of system of data retrieval S_{DR} to set of data D_k

$$S_{DR_k} = \langle D_k, \mathbf{N}, \mathbf{N}^{(k)}, U_k, \mathbf{P}_k, \Theta_k, Q_k \rangle$$

where:

D_k - set of data

$\mathbf{N}$ - alphabet of data names

$\mathbf{N}^{(k)}$ - set of names of data set D_k

U_k - set of functions m_N (assigning attribute values domains) for data defined by relational schema from $\mathbf{N}^{(k)}$ set

$\mathbf{P}_k$ - set of all projections of data values for data set D_k

Θ_k - set of all two-argument order relations of data name attribute values for $\mathbf{N}^{(k)}$ set of data names

Q_k - set of set theory operations for $\mathbf{P}_k$ set of data values projections.

D E F I N I T I O N 11

Individual (local) data base model DB_k is a pair

$$DB_k = \langle S_{DR_k} , S_{CSSDR_k} \rangle$$

where: S_{DR_k} - individual (local) system of data retrieval

S_{CSSDR_k} - individual (local) system of control system sets of data retrieval

D E F I N I T I O N   12

Divided data base model is a quadruple

$$DDB = <\{DB_k\}_{k\in K} , I , \Psi, \phi >$$

where: $\{DB_k\}_{k\in K}$ - family of individual (local) data base models

I - system of storage and retrieval of documents (covering system)

Ψ , ϕ - covering functions

Administrator of divided data base model will be named triple

$$\bar{A} = < I, \Psi, \phi > .$$

Final remark

Presented divided data base modes is a proposal resulting from necessity for analysis of very real problems for the geographically distributed organization of large data bases for an enterprise.

7. LITERATURE

1. Champine, G.A.: "Six approaches to distributed data bases". Datamation 5(1977) pp. 69-72.
2. Codd, E.F.: "A relational model of data for large shared data banks". Comm. ACM, Vol.13, No.6, June 1970, pp.377-387.
3. Codd, E.F.: " Further normalization of the data base relational model", Couran Comp. Sc. Symp. 6: Data base systems, Prentice-Hall, N.J., May 1971, pp. 65-98.
4. Kerschberg, L. - Klug, A. - Tsichritzis, D.: "A taxonomy of data models"; System for Large Data Bases; Lockemann, P.C.-Neuhold,E.J. (eds.), Norht-Holland Publishing Company, 1976.

5. Kobayashi, I.: "Information and information processing structure"; Information Systems 1(1975), pp.39-49.

6. Koczkodaj, W.W.: "A relational model of data and its connections with the i.s.r. systems"; ICS PAS Reports 306. Warsaw 1978.

7. Kubit, J.: "On a certain information system model"; 5th Symposium of algorithms-ALGORITHMS'79, April 1979, Strebske Pleso, pp.309-315.

8. Kuratowski, K. - Mostowski, A.: "Set theory" ; North-Holland Publishing Company, Polish Scientific Publishers, Warsaw 1976.

9. Lipski, W. - Marek, W.: "On information storage and retrieval systems"; CC PAS Report 200, Warsaw 1975.

10. Marek, W. - Pawlak, Z.: "Information storage and retrieval systems - Mathematical foundations"; Theoretival Computer Science 2(1976), pp. 331-354.

11. Marek, W. - Rode - Babczenko, I.: "A decomposition of informational systems"; CC PAS Reports 212, Warsaw 1975.

12. Neuhold, E.J.: "Formal properties of data bases"; University of Stuttgart, Mathematical Centre Tracts 63, 1974, pp.121-177.

13. Neuhold, E.J. - Biller, H.: "POREL: A distributed data base on an inhomogeneous computer network"; Third International Conference on Very Large Data Bases; Tokyo, Japan, October 1977, pp. 380-389.

14. Raś, Z.: "Algebraic foundations of information storage and retrieval system II"; (in Polish), ICP PAS Reports 339, Warsaw 1978.

ÖSSZEFOGLALÁS

Egy osztott adatbázis modell

Józef Kubit

Osztott adatbázisok egy formális modelljét mutatjuk be, a felbontásaival kapcsolatos alapvető jelöléseket definiáljuk. Lokális adatbázis modellek egy családját és az osztott adatbázis modell adminisztrátorát értjük osztott adatbázis modellen. A modell a relációs adatséma egy definicióján alapul.

Об одной модели распределенных баз данных

Йожеф Кубит

В данной работе представляется некоторая формальная модель распределенных баз данных, и предлагаются определения основных понятий связанных с ее декомпозицией. Модель распределенной базы данных состоит из администратора и семейства локальных моделей. Предложенная модель основана на некотором определении реляционной схемы данных.