Projeto de banco de dados de simulações numéricas

Ramon G. Costa, Fábio Porto, Bruno Schulze, Hermano Lustosa

LNCC - Laboratório Nacional de Computação Científica, Brasil
{ramongc, fporto, schulze, hlustosa}@lncc.br

Resumo. Com a rápida evolução dos sistemas computacionais, simulações numéricas baseadas em modelagem computacional têm alcançado soluções cada vez mais realistas. Ainda assim, o processo de simulação é complexo, exigindo grande capacidade computacional e produzindo muitos arquivos auxiliares com os resultados das simulações. Uma grande quantidade de arquivos, como os produzidos durante o processo de varredura de parâmetros, torna a gerência de experimentos uma tarefa bastante complexa, que associado ao cálculo da simulação, analisam seus resultados com programas específicos que precisam ser construídos e que fazem acesso ineficiente aos arquivos, muitas vezes em texto bruto. Por outro lado, a adoção de sistemas de gerência de bancos de dados (SGBD) em apoio à simulações numéricas traz seus próprios desafios. A representação de domínios espaço-temporal e de variáveis que representam quantidades físicas se mostra adequada para o armazenamento e consulta a esses tipos de dados. Neste contexto, este trabalho visa preencher uma lacuna existente na direção da adoção de SGBDs para o tratamento de dados de simulações numéricas. O focos deste trabalho está no processo de projeto do bancos de dados. Identifica-se, a partir do processo de projeto tradicional de banco de dados, extensões necessárias que permitem mapear o modelo conceitual de uma aplicações de simulação numérica em um projeto que combina modelo de dados relacional com modelo de matrizes multidimensionais. Apresenta-se um exemplo de aplicação de simulação do sistema cardiovascular humano e o processo de projeto de banco de dados implementado no sistema SciDB. Este trabalho tem como objetivo definir as fases de um projeto para o armazenamento de dados gerados por simulações numéricas, demonstrando a necessidade do armazenamento destes dados em array multidimensionais e o uso de técnicas tradicionais durante o processo de projeto do banco de dados. O conjunto de dados é representado através do uso do SciDB, que utiliza um modelo de dados em matrizes multidimensionais como a base de sua representação.

Categories and Subject Descriptors: G.1.10 [**Numerical Analysis**]: Applications; H.2.8 [**Database Management**]: Scientific databases; I.6.7 [**Simulation and Modelling**]: Simulation Support Systems

Keywords: Database Management, Information Storage and Retrieval, Numerical Analysis, Simulation and Modelling

1. INTRODUÇÃO

A simulação numérica de fenômenos naturais tem sido favorecida com os grandes avanços nas plataformas de computação de alto desempenho, obtendo representações mais realistas dos fenômenos modelados. No processo de modelagem computacional de fenômenos naturais, tais como o do sistema cardiovascular humano (SCH), utiliza-se um conjunto de equações diferenciais para representar as variações das quantidades no espaço-tempo. O modelo matemático é discretizado em um modelo computacional usando algum método numérico disponível. Do ponto de vista computacional, o estado da arte para o processamento de simulações numéricas utiliza arquivos texto para o armazenamento de dados. É fácil imaginar que a gerência de milhares de arquivos, como os produzidos durante o processo de varredura de parâmetros, torna a gerência de experimentos uma tarefa bastante complexa para o pesquisador. Ainda pior, uma vez calculada a simulação, cientistas analisam seus resultados, com programas de análise específicos que precisam ser construídos para o acesso ineficiente e não padronizado aos arquivos em formato de texto bruto.

O caráter científico de aplicações de simulações numéricas exige igualmente que os resultados obtidos possam ser reproduzidos e o processo de cálculo seja auditável. Neste contexto, dados de proveniência [Buneman et al. 2001] descrevem o processo experimental apoiando o cientista no processo de análise de seus resultados. Em particular, simulações numéricas requerem, além da gerência de dados de simulações propriamente dita, o tratamento de dados de proveniência.

Neste contexto, estender os benefícios do gerenciamento de dados em bancos de dados à aplicações de simulação numérica se torna essencial, uma vez que leva a esse domínio as vantagens já consagradas

na adoção de SGBDs em outros domínios, como na astronomia [Szalay et al. 2000]. Esta adoção, todavia, traz seus próprios desafios. A representação de domínios espaço-temporal determinando as variações admissíveis dos valores de quantidades físicas requer um modelo de dados mais apropriado do que simples relações n-árias [Stonebraker 2012]. Dentro deste contexto, o modelo de dados de matrizes multidimensionais implementado no sistema SciDB torna-se uma boa solução [Costa et al. 2012]. Naquela proposta apresentada, malhas espaciais tridimensionais e sua variação no tempo são mapeadas para dimensões em matrizes multidimensionais, enquanto as varíaveis são representadas em células. Neste trabalho, avançamos nesta proposta propondo um método para a derivação do projeto de bancos de dados para simulações numéricas a partir da representação conceitual do domínio, conforme adotado no projeto tradicional de bancos de dados.

Para ilustrar a utilização deste modelo no problema de simulações numéricas, considere a simulação do SCH desenvolvida pelo projeto Hemolab no LNCC [Blanco et al. 2009]. Nesta aplicação constrói-se uma representação geométrica das artérias do corpo humano na forma de uma malha multidimensional. Ao longo da artéria simulada, modelos com malhas em dimensões 1D e 3D são acoplados, segundo o interesse de pesquisa e a capacidade de processamento. Um modelo de simulação calcula, para cada instante de tempo e ponto da malha, valores de interesse como pressão e velocidade.

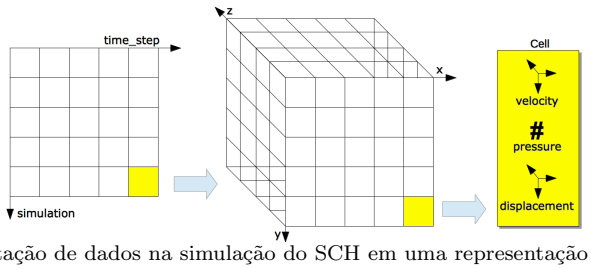

Fig. 1: Representação de dados na simulação do SCH em uma representação multidimensional

A Fig. 1 ilustra um modelo multidimensional para representar os dados calculados pela simulação do SCH. A ilustração apresenta as dimensões *simulation* e *time_step* e os pontos de uma malha tridimensional, referentes à simulação s e a um instante de tempo t. Na célula indicada, aparecem os valores das quantidades: velocidade, pressão e deslocamento. Assim sendo, o mapeamento da representação ilustrada na Fig. 1 para um modelo de dados de matrizes multidimensionais fica facilitado. As dimensões e os valores das células encontram correspondência direta nas estruturas de dimensão e células do modelo. A partir desta representação, podem-se elaborar consultas com base no espaço multidimensional construído, tal como obter os valores de velocidade e pressão em uma simulação s_i e tempo t_j para uma área da malha nas coordenadas $< x, y, z >$. Com isto, ao prover uma linguagem de alto-nível para expressões de consulta tem-se uma contribuição importante para cientistas que utilizam arquivos de texto bruto como resultado de suas simulações, obtendo ganhos maiores quando se utilizam estruturas de armazenamento e indexação eficientes [Zhang, Yi et al. 2011] para otimizar o acesso em disco.

O exemplo acima ilustra, de forma preliminar, a adequação do modelo de dados de matrizes multidimensionais no gerenciamento de dados de simulações numéricas. Uma questão fundamental, não explorada na literatura, é: *Como projetar um banco de dados para o modelo de matrizes multidimensionais, considerando-se as técnicas de projetos de bancos de dados*?

Neste contexto, este trabalho apresenta um método para o projeto de bancos de dados de simulações numéricas, tendo como alvo sua representação em modelos de dados de matrizes multidimensionais, bem como o tratamento de dados de proveniência produzidos durante o cálculo da simulação.

2. TRABALHOS RELACIONADOS

Vários sistemas têm surgido oferecendo suporte ao modelo de dados de matrizes multidimensionais [Stonebraker 2012] [Baumann et al. 1998], expondo as vantagens e os desafios no armazenamento e manipulação de *arrays* de forma nativa pelo sistema. Alguns trabalhos evocam a necessidade do uso de SGBDs para o armazenamento e gerenciamento de dados de simulação numérica, mas ainda utilizam

o modelo de dados relacional como forma de representação, como é o caso do *Turbulence Database Cluster* [Perlman et al. 2007]. Nosso trabalho considera que a utilização de matrizes multidimensionais são adequadas para o armazenamento e gerência de bancos de dados de simulações numéricas.

Trabalhos recentes discutem a necessidade de um tratamento mais eficiente para o armazenamento, indexação e processamento de dados científicos [Ogasawara et al. 2011] e [Zhang, Yi et al. 2011]. Stonebraker [Stonebraker et al. 2009] afirma que as atuais tecnologias de gerenciamento de dados são claramente incapazes de lidar com as exigências dos cientistas. Conhecer os requisitos destacados em projetos bem sucedidos se torna importante. Dentre estes projetos, podemos destacar o SciDB: um SGBD Orientado à Ciência [Stonebraker 2012] e os esforços de seus pesquisadores para especificar o conjunto de requisitos para a criação de um sistema de bancos de dados para a área científica. Este trabalho utiliza o SciDB para ilustrar o projeto físico do banco de dados de matrizes multidimensionais, demonstrando como um modelo conceitual pode ser mapeado para um esquema em *array*.

3. O PROJETO DE BANCOS DE DADOS PARA SIMULAÇÕES NUMÉRICAS

Para desenvolver o raciocínio que segue para a criação de bancos de dados de simulações numéricas, o processo será descrito a partir da análise dos requisitos que definem o ambiente do sistema, passando pelas fases de projeto conceitual até o projeto físico. Esta abordagem, chamada de processo de projeto *Top-down*, será utilizada neste trabalho para descrever o processo de projeto de bancos de dados de simulação numérica (Fig. 2). As fases do projeto são detalhadas nas subseções que se seguem.

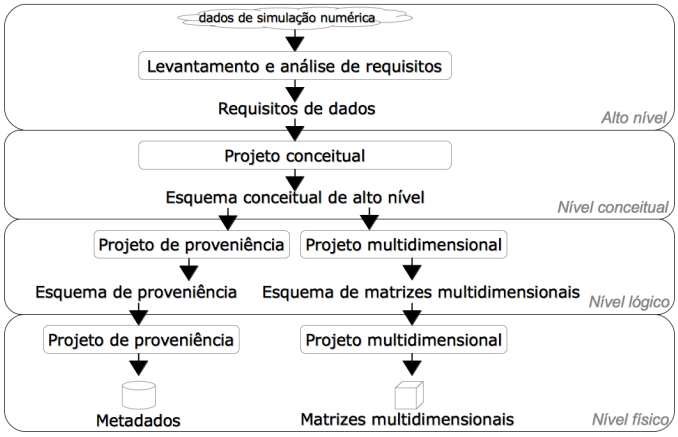

Fig. 2: Processo de projeto Top-Down para os dados gerados por simulações numéricas

3.1 O minimundo: dados de simulações numéricas

As simulações numéricas, nas quais o movimento de um fluido pode ser descrito pela especificação completa das propriedades a serem computadas em função de coordenadas espaço-temporais, são escopo de nossa pesquisa. Nas simulações numéricas de interesse, obtém-se informações do escoamento em função do que acontece em pontos fixos do espaço, conhecido como método de descrição euleriana do movimento [Batchelor 2000] e suas propriedades computadas referidas como quantidades físicas eulerianas.

Neste trabalho, os dados de simulação numérica são interpretados como sendo compostos por dois conjuntos de variáveis: variáveis dimensionais e quantidades físicas eulerianas. O primeiro conjunto define um sistema de coordenadas multidimensional, enquanto o segundo informa sobre as quantidades físicas calculadas em cada ponto de referência. Tipicamente, as variáveis dimensionais incluem espaço e tempo, além da dimensão identificadora de simulação. A dimensão espacial refere-se a uma malha que representa a topologia do domínio físico por meio de uma composição de simples objetos geométricos (por exemplo, um tetraedro). Uma malha é representada por um conjunto de pontos, referindo-se aos vértices dos objetos geométricos, e um conjunto de arestas que ligam os pontos e as faces do modelo, assim como em um grafo não direcionado.

No âmbito do projeto HeMoLab [HeMoLab 2013], a simulação do SCH pode se valer de um modelo tridimensional (3D) para representação espacial de seu domínio. Os modelos 3D são utilizados para obter um maior nível de detalhe sobre o comportamento do fluxo de sangue ao longo da artéria para uma estrutura especificada [Blanco et al. 2009].

A simulação computacional de um modelo do escoamento de sangue em uma artéria é feita através da resolução de um sistema de equações de Navier Stokes que descreve o movimento de fluidos [Formaggia et al. 2001]. Este processo é realizado por softwares informalmente referidos como *resolvedores numéricos*. Este último, recebe um conjunto de parâmetros que descrevem: a representação geométrica do domínio e parâmetros próprios da simulação. Cada período de simulação (geralmente um período refere-se a uma batida do coração) é subdividido em passos de tempo (um valor típico é 640) e para cada passo de tempo, um resolvedor numérico atualiza o arquivo que armazena os resultados da simulação com os valores de deslocamento, velocidade e pressão em cada ponto do volume. Por fim, durante o cálculo da simulação, armazenam-se os dados de proveniência informando sobre os programas utilizados, os parâmetros de entrada e os resultados obtidos.

3.2 O projeto do banco de dados de matrizes multidimensionais e de proveniência

Na abordagem adotada, inicia-se o projeto do banco de dados de simulações numéricas com o projeto conceitual dividido em duas partes. A primeira representa as dimensões envolvidas e as quantidades físicas associadas. A segunda diz respeito aos dados de proveniência associados ao cálculo da simulação.

Vários sistemas têm surgido oferecendo suporte ao modelo de dados de matrizes multidimensionais [Stonebraker 2012] [Baumann et al. 1998]. Como exemplo específico de implementação, podemos representar o esquema multidimensional usando o SciDB. Como as matrizes multidimensionais são a base de representação de dados no SciDB, um usuário pode especificar estruturas multidimensionais, fornecendo intervalos de valores para cada dimensão e uma lista de atributos para compor uma célula. Neste contexto, a estratégia de mapeamento a seguir, foi definida: *(i)* define-se o conjunto Δ de dimensões envolvidas; *(ii)* especifica-se a lista de quantidades eulerianas Π a serem computadas (cada conjunto de quantidades eulerianas correspondem a um fenômeno simulado em uma dada escala); *(iii)* cria-se uma matriz multidimensional contendo as dimensões Δ e os atributos Π.

Neste contexto, pode-se adotar a estratégia de anotação na representação conceitual Entidade-Relacionamento (ER) [Parent et al. 2006] para distinguir tais atributos neste modelo. Assim, na primeira parte, estende-se a representação conceitual para caracterizar cada atributo quanto ao seu propósito no ambiente de simulação numérica. Para cada atributo ou papéis de um relacionamento recursivo, deve-se identificar quais devem ser modelados como Metadados [M], Dimensões [D] ou Quantidades físicas eulerianas [Q]. Outros atributos ou papéis não devem conter identificação. A Fig. 3 apresenta um Diagrama ER de nível conceitual com as anotações [M], [D] e [Q].

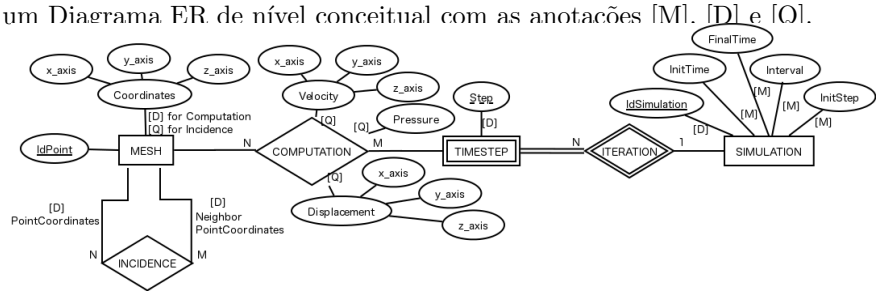

Fig. 3: Diagrama ER de Peter Chen com as anotações que direcionam o mapeamento para matrizes multidimensionais

No mapeamento do diagrama ER (Fig. 3) para uma representação lógica usando hipercubos (Fig. 4), as seguintes observações devem ser feitas: *(i)* identifique os dois grupos de informacões: atributos dimensionais [D] e atributos referentes às quantidades físicas eulerianas [Q]; *(ii)* as quantidades físicas eulerianas são mapeadas em células, onde os atributos correspondem às anotações [Q]; *(iii)* a dimensões espacial, temporal e identificadora das simulações devem ser representadas como dimensões do hipercubo; *(iv)* a matriz de adjacência entre os pontos da malha pode ser representada como um novo hipercubo.

A distinção entre atributos dimensionais e quantidade físicas requer a definição de conjunto de atributos determinantes funcionais. Para isso adota-se a representação de dependência funcional [Sadri and Ullman 1980] entre o conjunto de atributos. Sendo assim, Seja $A = \{a_1, a_2, \ldots, a_n\}$, tal que $a_i \subset d_i$, para todo $1 \leq i \leq n$, e d_i um domínio qualquer de valores, onde A é o conjunto de atributos envolvidos no cálculo de uma simulação numérica. Determinam-se os subconjuntos $D \subset A$ e $Q = A - D$, tais que D determina funcionalmente Q [Sadri and Ullman 1980]. O conjunto de atributos em D corresponde às dimensões do problema, enquanto o conjunto de atributos em A identifica as quantidade eulerianas. Neste caso:

`{Simulation, TimeStep, Point3D}` $\rightarrow$ `{velocity3D, pressure, displacement3D}`
`{Simulation, Point3D, NeighborPoint3D}` $\rightarrow$ `NeighborCoordinate3D`

A Fig. 4 apresenta a estrutura multidimensional resultante deste mapeamento.

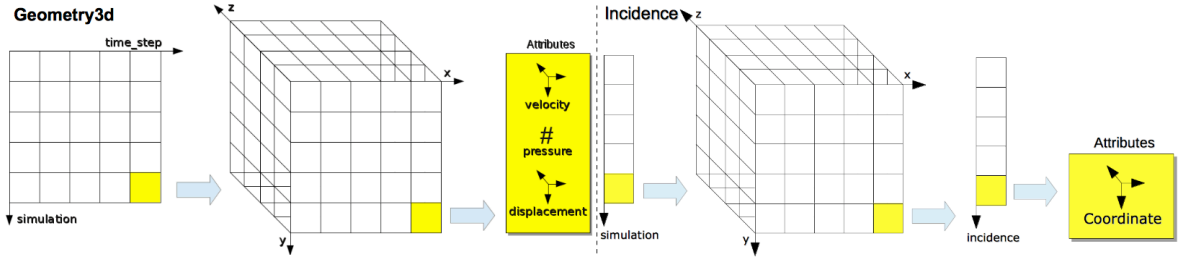

Fig. 4: Representação do modelo 3D do SCH em hipercubos

Através das linguagens AQL e AFL é possível seguir para o projeto físico e implementar o esquema interno que armazena dados em matrizes multidimensionais. A Tab. I apresenta o código para a criação do esquema interno do hipercubo *Geometry3d* apresentada na Fig. 4.

Tab. I: AQL utilizada para representar a matriz multidimensional *Geometry3d*

```
create empty array Geometry3d
<velocity_x:float, velocity_y:float, velocity_z:float, pressure:float,
displacement_x:float, displacement_y:float, displacement_z:float>
[simulation_number=0:9,1,0, time_step=0:30720,1921,0, x_axis(float)=155000,155000,0,
y_axis(float)=155000,155000,0, z_axis(float)=155000,155000,0];
```

Para o projeto do banco de dados de proveniência, deve-se levar em consideração as anotações referentes aos metadados [M] representados no diagrama da Fig. 3, e combinados aos dados de proveniência encontrados no problema, modelar um diagrama respeitando o domínio adotado. A Fig. 5(a) apresenta um diagrama, modelado para os dados de proveniência, usando o PROV-DM [Moreau et al. 2012].

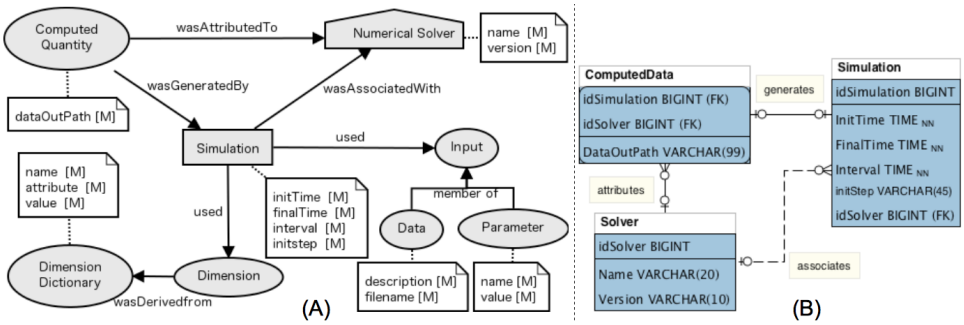

Fig. 5: (a) Diagrama PROV-DM; e (b) Diagrama *Crow's foot* para parte do conjunto de dados de proveniência.

No mapeamento do diagrama conceitual de proveniência (Fig. 5(a)) para o diagrama de nível lógico (Fig. 5(b)), as Entidades, os Agentes e as Atividades [Moreau et al. 2012] podem ser mapeadas como tabelas e as anotações [M] são mapeadas como atributos em cada tabela.

A Tab. II apresenta a implementação do esquema lógico representado na Fig. 5(b).

Tab. II: SQL utilizada para representar o catálogo de metadados

```
create database Cardio_catalog;
create table Cardio_catalog.Solver (
  idSolver serial primary key, name varchar, version varchar );
create table Cardio_catalog.Simulation (
  idSimulation serial primary key, InitTime time, FinalTime time,
  Interval time not null, initStep bigint,
  foreign key (idSolver) references Cardio_catalog.Solver(idSolver) );
create table Cardio_catalog.Computed_data (
  idSimulation bigint, idSolver bigint, DataOutPath varchar,
  primary key (idSimulation, idSolver),
  foreign key (idSimulation) references Cardio_catalog.Simulation(idSimulation),
  foreign key (idSolver) references Cardio_catalog.Solver(idSolver) );
```

4. CONCLUSÃO

Este trabalho propôs um processo de projeto *top-down* para o desenvolvimento de bancos de dados de simulação numérica. Dados desta natureza têm a característica de mudarem com o tempo e espaço e de serem envolvidos em operações próprias, tais como: multiplicação de matrizes, transposição de matrizes e estudo de convergências de valores. Devido a estas características, se torna adequado o uso de sistemas de bancos de dados que permitem o armazenamento de matrizes multidimensionais de forma nativa. Este trabalho demonstrou que os métodos tradicionais de projeto de bancos de dados podem ser utilizados em um processo que culmina na criação de matrizes multidimensionais para o armazenamento de dados de simulação numérica envolvidos durante o cálculo da simulação.

Muito se tem discutido a respeito do uso de estratégias diferentes das tradicionais para o armazenamento de dados científicos, mas nenhuma delas aborda o fato de definir um método para a especificação de bancos de dados de simulações numéricas. O resultado apresentado neste trabalho é uma proposta até então inexistente na área.

REFERENCES

Batchelor, G. K. *An Introduction to Fluid Dynamics*, 2000.

Baumann, P., Dehmel, A., et al. The multidimensional database system rasdaman. In *Proceedings ACM SIGMOD International Conference on Management of Data*. Seattle, USA, pp. 575–577, 1998.

Blanco, P. J., Pivello, M. R., Urquiza, S. A., and Feijóo, R. A. On the potentialities of 3d-1d coupled models in hemodynamics simulations. *Journal of Biomechanics* 42 (7): 919–930, 2009.

Buneman, P. et al. Why and where: A characterization of data provenance. In *ICDT*. pp. 316–330, 2001.

Costa, R. G., Porto, F., and Schulze, B. Towards analytical data management for numerical simulations. In *AMW*. pp. 210–214, 2012.

Formaggia, L., Gerbeau, J.-F., et al. On the Coupling of 3D and 1D Navier-Stokes Equations for Flow Problems in Compliant Vessels. *Computer Methods in Applied Mechanics and Engineering* 191 (6-7): 561–582, 2001.

HeMoLab. Hemodynamics modeling laboratory, 2013. `http://hemolab.lncc.br`.

Moreau, L., Missier, P., et al. Prov-dm: The prov data model. Tech. rep., 2012.

Ogasawara, E., Dias, J., de Oliveira, D., Porto, F., Valduriez, P., and Mattoso, M. An algebraic approach for data-centric scientific workflows. In *Proceedings of the VLDB Endowmt*. Vol. 4. Seattle, USA, pp. 1328–1339, 2011.

Parent, C., Spaccapietra, S., and Zimányi, E. *Conceptual modeling for traditional and spatio-temporal applications - the MADS approach*. Springer, 2006.

Perlman, E., Burns, R., Li, Y., and Meneveau, C. Data exploration of turbulence simulations using a database cluster. In *Proceedings of the IEEE conference on Supercomputing*. Vol. 23. ACM, New York, USA, pp. 1–11, 2007.

Sadri, F. and Ullman, J. D. The interaction between functional dependencies and template dependencies. In *SIGMOD Conference*. pp. 45–51, 1980.

Stonebraker, M. Scidb: An open-source dbms for scientific data. *ERCIM News* 2012 (89), 2012.

Stonebraker, M., Becla, J., Dewitt, D., Lim, K.-T., Maier, D., Ratzesberger, O., and Zdonik, S. Requirements for science data bases and scidb. In *Conference on Innovative Data Systems Research (CIDR)*. Asilomar, USA, 2009.

Szalay, A. S., Kunszt, P. Z., et al. Designing and mining multi-terabyte astronomy archives: the sloan digital sky survey. In *Proceedings of the ACM SIGMOD Conf. on Management of data*. New York, USA, pp. 451–462, 2000.

Zhang, Yi et al. Storing matrices on disk: Theory and practice revisited. In *VLDB*. Seatle, USA, 2011.