Universidade Federal de Ouro Preto - UFOP

Instituto de Ciências Exatas e Biológicas - ICEB

Departamento de Computação - DECOM

CollA: Uma Plataforma Colaborativa para Análise de Grandes Volumes de Dados

Aluno: Bruno Cerqueira Hott
Matricula: 08.1.4133

Orientador: Joubert de Castro Lima

Ouro Preto
29 de março de 2012

Universidade Federal de Ouro Preto - UFOP

Instituto de Ciências Exatas e Biológicas - ICEB

Departamento de Computação - DECOM

CollA: Uma Plataforma Colaborativa para Análise de Grandes Volumes de Dados

Proposta de monografia apresentada ao curso de Bacharelado em Ciência da Computação, Universidade Federal de Ouro Preto, como requisito parcial para a conclusão da disciplina Monografia II (BCC391).

Aluno: Bruno Cerqueira Hott
Matricula: 08.1.4133

Orientador: Joubert de Castro Lima

Ouro Preto
29 de março de 2012

Resumo

A carência de teorias que descrevam o comportamento de determinadas abstrações nos faz armazenar dados para serem analisados posteriormente. A este tipo de abordagem damos o nome de *Data-Driven Approaches* ou Abordagens Orientadas por Dados. Nas últimas décadas o volume de dados estruturados, semi-estruturados e não estruturados cresceu significativamente, assim como o número de abordagens para analisá-los. De uma forma geral, a estratégia adotada para se analisar tal volume de dados é individual, feita a partir de uma cópia de um fragmento dos dados originais. Tal estratégia gera redundâncias e é pouco colaborativa.

Nosso principal objetivo é prover um conjunto de métodos, técnicas e ferramentas computacionais direcionadas a análise colaborativa de grandes volumes de dados. Assim, este trabalho se propõe a desenvolver uma plataforma de software denominada Collaborative Analysis (CollA), que irá permitir que artefatos da análise de dados sejam compartilhados por diversos especialistas em diversos ambientes. Os principais artefatos da plataforma CollA são: fontes de dados, nós de processamento (incluindo processador, memória principal e secundária), componentes de software, experimentos ou apenas os resultados dos experimentos.

Ao compartilhar artefatos permitimos que especialistas possam colaborar cada vez mais no processo analítico, cedendo CPUs, disponibilizando fontes de dados ou mesmo componentes de software, implementando experimentos ou obtendo novos resultados a partir de experimentos já implementados. Esta nova perspectiva de análise chama-se Análise Colaborativa de Dados.

Palavras-chave: Sistemas distribuidos. Programação de alto desempenho. Computação colaborativa.

Sumário

1 Introdução **1**

2 Justificativa **4**

3 Objetivos **5**
3.1 Objetivo geral 5
3.2 Objetivos específicos 5

4 Metodologia **6**

5 Cronograma de atividades **7**

Lista de Figuras

1 Compartilhamento e colaboração na plataforma CollA. 3
2 Processo de desenvolvimento Iterativo e Incremental. 6

Lista de Tabelas

1 Cronograma de Atividades. 7

1 Introdução

A espécie humana pode ser considerada a única que vive no presente, analisa o passado e tenta prever o futuro. Como estamos cercados de abstrações, tentamos a todo o momento predizer seus comportamentos. Existem duas alternativas científicas para predizer o comportamento de uma abstração, seja esta uma empresa, a bolsa de valores de um país, um indivíduo de uma espécie, a Amazônia, entre outros. A primeira usa teorias que descrevem o comportamento da abstração sendo modelada. Normalmente, tais teorias descrevem fenômenos físicos usando matemática como linguagem. Esta alternativa é representada por abordagens orientadas por teorias (Theory-Driven Approaches). A segunda usa o comportamento passado, descrito normalmente por um grande volume de dados, para predizer o comportamento. Esta alternativa é representada por abordagens orientadas por dados (Data-Driven Approaches [31, 34]), usadas quando teorias não existem.

As abordagens orientadas por dados são divididas em abordagens orientadas por hipóteses (Hypothesis-Driven Approaches [12, 15]) e abordagens orientadas por descobertas (Discover-Driven Approaches [6, 24, 27, 30]). De uma forma geral, abordagens orientadas por dados permitem identificar padrões desconhecidos, não triviais, potencialmente úteis e implícitos a partir de grandes volumes de dados.

O volume de dados é uma realidade que afeta sensivelmente os recursos computacionais utilizados pelas abordagens orientadas por dados, sejam estes processamento e memória. Uma pesquisa da universidade de Berkeley, datada em 2001, mostra que 93% da informação atual já nasceu digital e que estamos "'duplicando"' o volume de informação a cada ano. Segundo esta pesquisa de Berkeley, em 2003 tínhamos cerca de 53 bilhões de Gigabytes produzidos

Muitas abordagens foram desenvolvidas para analisar grandes volumes de dados. Podemos diferenciar tais abordagens de acordo com o tipo de repositório de dados usado, sejam estes bases relacionais, data warehouses, bases transacionais, arquivos planos, data streams e a própria Web.

Além dos repositórios, podemos diferenciar as abordagens de acordo com o tipo de padrão a ser minerado/identificado nos dados. Existem abordagens que associam os dados a classes ou conceitos mais abstratos. As classes ou conceitos são sumarizados em vários níveis, usando medidas estatísticas, por exemplo. Tais abordagens são denominadas Caracterizadores e Discriminadores [8, 10, 11, 12, 13, 15, 23, 25, 31, 32, 35, 36, 37, 38]. Existem abordagens para encontrar padrões freqüentes nos dados. Tais abordagens são denominadas Mineradores de Padrões Freqüentes, Associações ou Correlações [4, 5, 14, 18, 29, 19]. Abordagens mais sofisticadas, como Classificadores [17, 26, 33], também são desenvolvidas. Classificadores predizem modelos ou funções a partir de dados categóricos ou contínuos, respectivamente. Modelos descrevem ou distinguem classes ou conceitos para que os mesmos sejam usados para predizer objetos cuja classe ou conceito não se conhece. Existem também as abordagens usadas para gerar as classes e conceitos. A estas abordagens damos o nome de Geradores de Clusters [3, 7, 16, 20, 21, 22, 28]. Por fim, existem abordagens que analisam outliers e tendências nos dados [2, 9].

Infelizmente muito pouca pesquisa foi feita no sentido de tornar o processo de análise de dados mais colaborativo, já que o volume de dados e o número de abordagens para analisá-los tende a crescer. Partimos da hipótese de que serviços como chats, edi-

ção colaborativa, entre outros, tornam qualquer processo mais colaborativo, incluindo o processo de análise de grandes volumes de dados. De uma forma geral, existem inúmeras abordagens sendo disponibilizadas, algumas reunidas em ambientes de código aberto como o Weka [1], e um número crescente de fontes de dados disponíveis para serem copiadas e analisadas individualmente por tais abordagens ou ambientes, mas até o momento quase nenhum esforço foi feito no sentido de compartilhar mais do que simplesmente fontes de dados entre especialistas. Ainda não usufruímos efetivamente do fato de estarmos cada vez mais imersos na rede, o que torna a colaboração, antes considerada quase impossível, uma realidade que deve ser cada vez mais explorada.

Neste projeto de pesquisa proveremos um conjunto de métodos, técnicas e ferramentas computacionais para a análise colaborativa de grandes volumes de dados. Como resultado, propomos o desenvolvimento de uma plataforma de software denominada Collaborative Analysis (CollA). A plataforma CollA irá permitir que não somente fontes de dados sejam compartilhadas, mas também que componentes de software, experimentos, resultados de experimentos e nós de processamento (incluindo memória principal, secundária e processadores) possam ser usados de forma colaborativa por diferentes especialistas em diferentes ambientes criados na plataforma.

A análise colaborativa de dados deve permitir que ambientes possam ser criados e que artefatos comuns à análise possam ser disponibilizados em tais ambientes. Além disto, precisamos garantir que os ambientes sejam seguros, escaláveis e tolerantes a falhas, ocasionadas principalmente por nós de processamento que se tornam inativos em tais ambientes.

Na Figura 1, ilustramos um exemplo de compartilhamento e, conseqüentemente, colaboração que poderá ocorrer com a utilização da plataforma CollA. Suponhamos que um usuário A1, previamente cadastrado, tenha registrado no ambiente E1 dois nós de processamento n1 e n2, respectivamente. No nó de processamento n1 existe uma partição de uma fonte de dados Db e no nó de processamento n2 existe a segunda e última partição de Db. Além do usuário A1, temos no ambiente E1 o usuário A2. A2 almeja executar um experimento Exp a partir da fonte de dados particionada Db. O usuário A2 também disponibiliza um nó de processamento ao ambiente E1, denominado n3. O componente de software desenhado para a plataforma CollA poderá utilizar n1, n2 e, provavelmente, n3 para executar Exp. Uma vez executado, a saída de Exp estará disponível para A1 e A2. Suponhamos que A2 resolva explorar os dados gerados no experimento, conseguindo obter resultados interessantes do mesmo. Os resultados obtidos por A2 também podem ser compartilhados. De forma análoga, A1 pode explorar a saída de Exp para obter novos resultados ou visualizar os resultados já obtidos por A2. O uso de chats e edição colaborativa de documentos torna todo este processo colaborativo. A configuração de determinados componentes de software pode ocorrer somente à partir de um grupo de usuários e não um único usuário é outro exemplo de colaboração na plataforma CollA.

De uma maneira geral, é possível perceber que a plataforma CollA poderá prover compartilhamento e, conseqüentemente, colaboração ao processo analítico. Acreditamos que a melhor arquitetura para implementação da plataforma CollA seja a peer-to-peer, com nós de processamento se comportando como clientes e servidores simultaneamente. Um certo nível de centralização também será necessário, pois caberá a um conjunto de nós manter o estado dos variados ambientes criados, incluindo usuários, fontes de dados, experimentos, nós de processamento ativos, mapas de nós

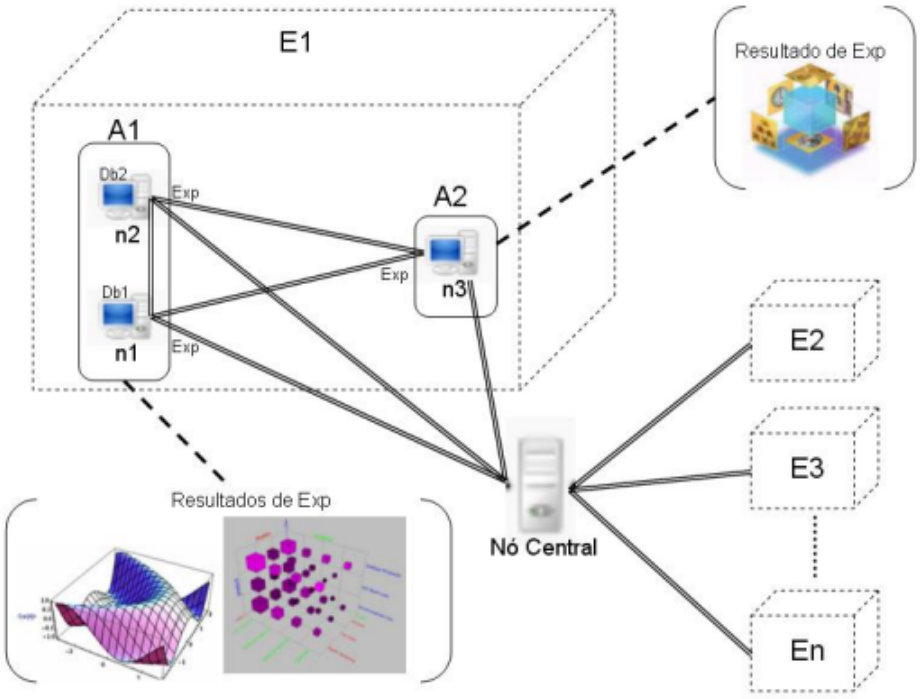

Figura 1: Compartilhamento e colaboração na plataforma CollA.

de processamento utilizados por experimento, mapa das partições das fontes de dados e suas réplicas, dentre outros.

Como em qualquer aplicação peer-to-peer, requisitos como tolerância a falhas, segurança, escalabilidade, entre outros, devem ser implementados como serviços básicos da plataforma. No exemplo ilustrado na Figura 1, um nó de processamento (n1 ou n2, por exemplo) pode se tornar inativo a qualquer momento. Caberá ao conjunto de nós de processamento que mantêm o estado dos diversos ambientes detectar tal falha e encontrar uma possível solução para a mesma ou reportar o problema aos usuários afetados.

Uma possível solução para a queda de nós de processamento seria manter um nível de redundância de fontes de dados dentro dos ambientes, ou seja, distribuir e replicar partições das fontes de dados em diversos nós de processamento. Replicação de fontes de dados gera problema de atualização, que por sua vez gera outros problemas. De uma forma geral, percebemos que o desenvolvimento da plataforma CollA deverá ocorrer de forma iterativa e incremental, pois os desafios são enormes.

Neste projeto reduzimos o escopo para o desenvolvimento do serviço de particionamento de fonte de dados. O referido serviço poderá ser configurado a partir de um grupo ou um único usuário. Os serviços básicos da plataforma CollA, sejam estes de distribuição, redundância, chats e alguma forma de editar documentos de forma colaborativa também serão desenvolvidos.

Na próxima seção descrevemos o objetivo geral e os específicos deste projeto de pesquisa. Na Seção 3, as justificativas e a relevância deste projeto de pesquisa são descritas. Na Seção 4, descrevemos a metodologia a ser adotada. Na Seção 5, descrevemos um cronograma de execução das tarefas macro do projeto. Finalmente, na Seção 6 apresentamos os resultados esperados com este projeto de pesquisa.

2 Justificativa

Este projeto de pesquisa poderá responder duas questões científicas extremamente relevantes: Será possível compartilhar artefatos de análise eficientemente? Será possível redesenhar as abordagens de análise clássicas, descritas na Seção 1, para que as mesmas sejam executadas num ambiente distribuído, heterogêneo e colaborativo ou teremos que propor novas abordagens?

O projeto produzirá dois artefatos tecnológicos: primeiro, uma plataforma de software denominada CollA, onde artefatos fundamentais a análise de dados poderão ser compartilhados. Segundo, um catálogo de abordagens analíticas que poderá ser utilizado por diferentes especialistas em diferentes ambientes de análise criados na plataforma CollA. Incialmente, este catálogo se resume a particionamento e consulta simples das fontes de dados distribuídas.

A plataforma CollA poderá incentivar o reuso de abstrações inovadoras. A criação de ambientes onde usuários podem ser provedores de fontes de dados, codificadores de experimentos, disponibilizadores de nós de processamento, fornecedores de componente de softwares ou simplesmente exploradores de resultados de experimentos, poderá mudar substancialmente a atual visão que temos sobre a análise de dados suportada por computador.

Pessoas físicas, laboratórios, universidades, empresas, entre outros, poderão se cadastrar em ambientes na plataforma CollA. CollA poderá também ser executada numa intranet de uma determinada empresa, órgão do governo ou universidade, onde funcionários criam ambientes, compartilham e cooperam seus processos analíticos. A academia, o governo e a iniciativa privada poderão encontrar ou criar ambientes direcionados a determinadas áreas, permitindo com isto que especialistas multidisciplinares compartilhem e cooperem cada vez mais.

3 Objetivos

3.1 Objetivo geral

- produzir métodos, técnicas e ferramentas computacionais para dotar especialistas, de qualquer área do conhecimento humano, de capacidade de compartilhar artefatos de análise e cooperar no processo analítico.

3.2 Objetivos específicos

- Desenvolver uma plataforma de software, denominada CollA, onde artefatos de análise possam ser compartilhados e onde o processo analítico possa ocorrer de forma colaborativa.

- Desenvolver o serviço de particionamento de fonte de dados na plataforma CollA para que usuários já possam usar para consultas do tipo SQL, porém em ambiente distribuído.

4 Metodologia

Neste projeto será adotada a metodologia de desenvolvimento Iterativo e Incremental. Trata-se de um processo de desenvolvimento de software criado em resposta às fraquezas do modelo em cascata, o mais tradicional. O padrão a ser utilizado é o Processo Unificado da Rational (RUP).

O RUP usa a abordagem da orientação a objetos em sua concepção e é projetado e documentado utilizando a notação UML (Unified Modeling Language) para ilustrar os processos em ação. É um processo considerado pesado e preferencialmente aplicável a grandes equipes de desenvolvimento e a grandes projetos, porém o fato de ser amplamente customizável torna possível que seja adaptado para projetos de qualquer escala.

A Figura 2 ilustra as fases existentes e as disciplinas a serem implementadas em cada fase. As fases indicam a ênfase que é dada no projeto em um dado instante. Concepção: ênfase no escopo do sistema; Elaboração: ênfase na arquitetura; Construção: ênfase no desenvolvimento; Transição: ênfase na implantação. Nota-se que múltiplas disciplinas são executas em cada fase.

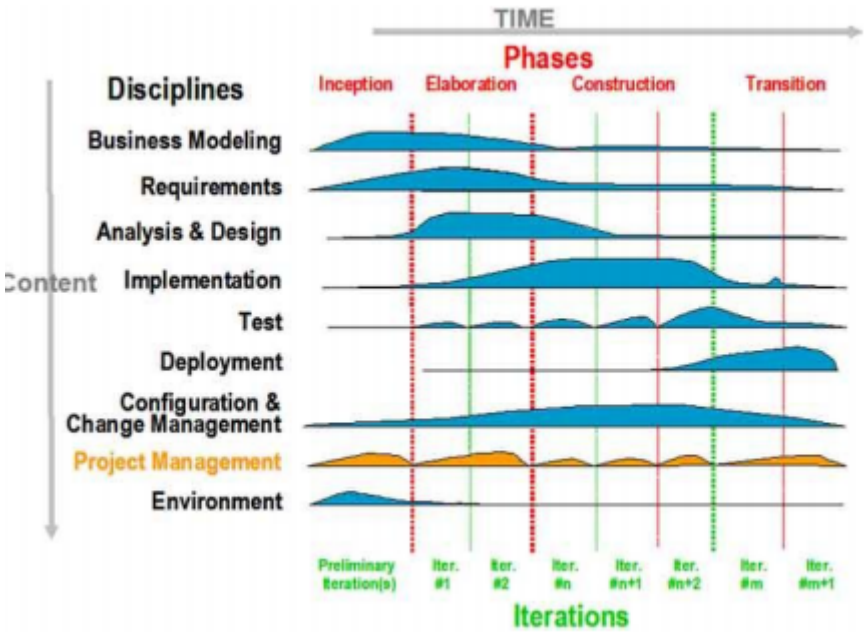

Figura 2: Processo de desenvolvimento Iterativo e Incremental.

5 Cronograma de atividades

Na Tabela 1, xxx.

Atividades	**Mar**	**Abr**	**Mai**	**Jun**	**Jul**
Passo 1	X	X			
Passo 2		X	X		
Passo 3			X	X	
Passo 4		X	X	X	
Redigir a Monografia			X	X	X
Apresentação do Trabalho					X

Tabela 1: Cronograma de Atividades.

Os seguintes passos serão utilizados para realização deste trabalho:

Passo 1: Desenvolvimento de componentes de software básicos para a plataforma CollA, sejam estes de distribuição (JXTA), redundância, criação de ambientes e compartilhamento de artefatos, chat e alguma forma de edição colaborativa de documentos.

Passo 2: Desenvolvimento do serviço de particionamento de fonte de dados na plataforma CollA.

Passo 3: Desenvolvimento de algum serviço de consulta à partir da fonte de dados distribuída. O projeto Apache HIVE (http://hive.apache.org/) que se propõe a oferecer o serviço de datawarehouse sobre o hadoop (http://hadoop.apache.org/) deverá ser considerado.

Passo 4: Escrita de relatórios e tutoriais sobre a plataforma CollA.

Referências

[1] Weka project. `http://www.cs.waikato.ac.nz/ml/weka/`, 2012.

[2] C.C. AGGARWAL and P.S. YU. Outlier detection for high dimensional data. *ACM SIGMOD*, pages 37–46, 2001.

[3] R. AGRAWAL, D. GEHRKE, J. GUNOPULOS, and P. RAGHAVAN. Automatic subspace clustering of high dimensional data for data mining applications. *ACM SIGMOD*, pages 94–105, 1998.

[4] R. AGRAWAL, T. IMIELINSKI, and SWAMI. A. mining association rules between sets of items in large databases. *SIGMOD*, pages 94–105, May 1993.

[5] R. AGRAWAL and R. SRIKANT. Fast algorithms for mining association rules. In *Int. Conf. Very Large Data Bases*, Santiago, Chile, set.

[6] E. ALPAYDIN. Introduction to machine learning (adaptive computation and machine learning). *MIT Press*, 2004.

[7] P. ARABIE, L.J. HUBERT, and G. DE SOETE. Clustering and classification. *World Scientific*, 1996.

[8] K. BEYER and R. RAMAKRISHNAN. Bottom-up computation of sparse and iceberg cubes. *SIGMOD Record (ACM Special Interest Group on Management of Data)*, 28.

[9] M.M. BREUNIG, H.P. KRIEGEL, R. NG, and J. SANDER. Lof: Identifying density-based local outliers. *ACM SIGMOD*, pages 93–104, 2000.

[10] Y. CHEN. Parallel rolap data cube construction on shared-nothing multiprocessors. in: Distributed and parallel databases. *IEEE Computer Society*, pages 219–236, 2004.

[11] F. DEHNE, T. EAVIS, S. HAMBRUSCH, and A. RAU-CHAPLIN. Parallelizing the data cube. distributed and parallel databases. 11, 2002.

[12] J. GRAY, S. CHAUDHURI, A. BOSWORTH, A. LAYMAN, D. REICHART, M. VENKATRAO, F. PELLOW, and H. PIRAHESH. Data cube: A relational aggregation operator generalizing group-by, cross-tab, and sub-totals.

[13] J. HAN, J PEI, DONG, and WANG. Efficient computation of iceberg cubes with complex measures. *IGMODIC: ACM SIGMOD International Conference on Management of Data*, 2001.

[14] J. HAN, J. PEI, and Y. YIN. Mining frequent patterns without candidate generation. *ACM-SIGMOD*, 2000.

[15] V. HARINARAYAN, A. RAJARAMAN, and J. D. ULLMAN. Implementing data cubes efficiently. *SIGMOD Record (ACM Special Interest Group on Management of Data)*, 25(2):205–216, June 1996.

[16] J.A. HARTIGAN. Clustering algorithms. *John Wiley & Sons*, 1975.

[17] T. HASTIE, R. TIBSHIARANI, and J. FRIEDMAN. The elements of statistical learning: Data mining, inference and prediction. *Springer-Verlag*, 2001.

[18] J. HIPP, U. GüNTZER, and G. NAKAEIZADEH. Algorithms for association rule mining - a general survey and comparison. *ACM SIGKDD Explorations*, pages 58–64, 2000.

[19] WANG J., HAN J., and PEI J. Closet+ : Searching for the best strategies for mining frequent closed itemsets. *SIGKDD*, August 2003.

[20] A.K. JAIN and R.C. DUBES. Algorithms for clustering data. *Prentice Hall*, 1988.

[21] A.K. JAIN, M.N. MURTY, and P.J. FLYNN. Data clustering: A survey. *ACM Computer Survey*, pages 264–323, 1999.

[22] L. KAUFMAN and P.J. ROUSSEEUW. Finding groups in data: An introduction to cluster analysis. *John Wiley & Sons*, 1990.

[23] L. V. S. LAKSHMANAN, J. PEI, and J. HAN. Quotient cube: How to summarize the semantics of a data cube. *VLDB. Morgan Kaufmann*, pages 778–789, 2002.

[24] P. LANGLEY. Elements of machine learning. *Morgan Kaufmann*, 1996.

[25] X. LI, J. HAN, and H. GONZALEZ. High-dimensional olap: A minimal cubing approach. proceedings of the thirtieth international conference on very large data bases. *Morgan Kaufmann*, pages 528–539, August 2004.

[26] D. MICHIE, D.J. SPIEGELHALTER, and C.C. TAYLOR. Machine learning, neural and statistical classification. *Ellis Horwood*, 1994.

[27] T.M. MITCHELL. Machine learning. *McGraw-Hill*, 1997.

[28] L. PARSONS, E. HAQUE, and H. LIU. Subspace clustering for high dimensional data: A review. *SIGKDD Explorations*, 6, 2004.

[29] N. PASQUIER, Y. BASTIDE, R. TAOUIL, and LAKHAL L. Discovering frequent closed itemsets for association rules. *ICDT*, January 1999.

[30] S. RUSSEL and P. NORVIG. *Artificial Intelligence: A Modern Approach.* 1995.

[31] Y. SISMANIS, A. DELIGIANNAKIS, N. ROUSSOPOULOS, and Y. KOTIDIS. Dwarf: Shrinking the petacube. *S.l.*

[32] W. WANG, H. LU, J. FENG, and J. X. YU. Condensed cube: An efficient approach to reducing data cube size. *IEEE Computer Society*, January 2002.

[33] S.M. WEISS and C.A. KULIKOWSKI. Computer systems that learn: Classification and prediction methods from statistics, neural nets, machine learning, and expert systems. *Morgan Kaufman*, 1991.

[34] I.H. WITTEN and E. FRANK. Data mining practical machine learning tools and techniques. *Morgan Kaufmann.*

[35] D. XIN, J. HAN, X. LI, Z. SHAO, and B. W. WAH. Computing iceberg cubes by top-down and bottom-up integration: The starcubing approach. *IEEE Trans. Knowl. Data Eng*, 19(1):111–126, 2007.

[36] D. XIN, Z. SHAO, J. HAN, and H. LIU. C-cubing: Efficient computation of closed cubes by aggregation-based checking. *Proceedings of the 22nd International Conference on Data Engineering*, 2006.

[37] D. ZHANG, C. ZHAI, and J. HAN. Topic cube: Topic modeling for olap on multidimensional text databases. *SIAM Int. Conf. on Data Mining*, 2009.

[38] Y. ZHAO, P. M. DESHPANDE, and J. F. NAUGHTON. An array-based algorithm for simultaneous multidimensional aggregates. *SIGMOD*, pages 159–170, June 1997.