Report from Dagstuhl Seminar 11361

Data Warehousing: from Occasional OLAP to Real-time Business Intelligence

Edited by
Markus Schneider[1], Gottfried Vossen[2], and Esteban Zimányi[3]

1 **University of Florida, US,** `mschneid@cise.ufl.edu`
2 **Universität Münster, DE,** `vossen@uni-muenster.de`
3 **Université Libre de Bruxelles, BE,** `ezimanyi@ulb.ac.be`

Abstract

This report documents the outcomes of the Dagstuhl Seminar 11361 "Data Warehousing: from Occasional OLAP to Real-time Business Intelligence". In the past, data warehousing and analytical processing (OLAP) have produced fundamental and important technologies for the design, management, and use of information systems for decision support. Indeed, many industrial and administrative organizations have successfully used data warehouses to collect essential indicators that help them improve their business processes and their decision making efforts. Recent developments like column stores instead of row stores at the physical level, real-time data warehousing and Business Intelligence applications at the conceptual level, and decision support for new emerging applications have raised new research questions. This seminar has focused on the following five main topics: (i) Real-Time Data Warehouses and Business Intelligence, (ii) Spatio-Temporal Data Warehousing, (iii) Situational Business Intelligence, (iv) Query Processing in Data Warehouses, and (v) Knowledge Extraction and Management in Data Warehouses. These topics were discussed in parallel groups and each group identified open research issues and new challenges.

Seminar 4.–9. September, 2011 – www.dagstuhl.de/11361
1998 ACM Subject Classification H.2.7 Database Administration – Data warehouse and repository.
Keywords and phrases Business Intelligence, Data Warehouses, OLAP, Spatio-temporal information, ETL, Service Orientation, Query optimization
Digital Object Identifier 10.4230/DagRep.1.9.1

1 Executive Summary

Markus Schneider
Gottfried Vossen
Esteban Zimányi

This Dagstuhl Seminar brought together 37 researchers from 14 countries across disciplines that study data warehousing. The seminar can be seen as a successor of the Dagstuhl Perspectives Workshop 04321 "Data Warehousing at the Crossroads" (http://www.dagstuhl.de/04321) co-arranged by one of the organizers in 2004. After seven years (in 2011), we felt that it was time to take stock again, determine the status quo, and reflect on the future of data warehousing. Further, the seminar in 2004 was a perspectives workshop with the

<table>
<tr><th></th><th>Monday</th><th>Tuesday</th><th>Wednesday</th><th>Thursday</th><th>Friday</th></tr>
<tr><td>09:00–10:30</td><td>Opening
Short
Introductions</td><td>Parallel
Groups</td><td>Group
Reports</td><td>Parallel
Groups</td><td>Group
Conclusions</td></tr>
<tr><td>10:30–11:00</td><td colspan="5">Coffee break</td></tr>
<tr><td>11:00–12:00</td><td>Short
Introductions</td><td>Parallel
Groups</td><td>IBM Industrial
Presentation</td><td>Parallel
Groups</td><td>Group
Conclusions</td></tr>
<tr><td>12:00–14:00</td><td colspan="5">Lunch</td></tr>
<tr><td>14:00–15:30</td><td>Short
Introductions</td><td>Parallel
Groups</td><td rowspan="3">Excursion
to
Trier</td><td>Parallel
Groups</td><td rowspan="3"></td></tr>
<tr><td>15:30–16:00</td><td colspan="2">Coffee break</td><td>Coffee break</td></tr>
<tr><td>16:00–17:30</td><td>Short
Introductions</td><td>Parallel
Groups</td><td>Parallel
Groups</td></tr>
<tr><td>18:00–···</td><td colspan="4">Dinner</td><td></td></tr>
</table>

Figure 1 Timetable of the seminar

exclusive goal of predicting the possible future direction of data warehousing and OLAP. This seminar has a different scope and mainly dealt with current and recent research results. This does not exclude to look into the future and to determine what the field has achieved in the meantime.

The participants of the seminar were clustered around five thematic areas as follows

- Real-Time Data Warehouses and Business Intelligence,
- Spatio-Temporal Data Warehousing,
- Situational Business Intelligence,
- Query Processing in Data Warehousing, and
- Knowledge Extraction and Management in Data Warehouses.

This clustering was arranged prior to the seminar so that participants came to the seminar with short introductions describing both themselves and the research topics they work on, as well as identify challenging questions in these thematic areas. The whole Monday was devoted to these introductions.

Tuesday and Thursday were devoted to parallel working groups, each group discussing the state of the art in its thematic area and identifying the open questions. The organizers asked to each group to answer the following four questions

1. Identify most pressing research issues.
2. Is the topic purely industrial or also academic?
3. Where do you expect to be 5 years from now?
4. What would you want a new PhD student to work on?

Further, longer research presentations were given within each group for focusing the groups' work around a common approach. Each group presented a short ongoing report of the work realized in Wednesday morning, and a final presentation of their results on Friday morning.

It is worth noting that the Data Warehouse domain is both an active research domain but also drives a very intense activity in the industrial world. One objective of the seminar was also to establish bridges between these two communities. The seminar attracted participants from major companies active in the Data Warehouse domain (Sybase, HP, IBM, and EMC). Another concrete step in this respect was an industrial presentation given on Wednesday morning by Knut Stolze, from IBM Germany.

All five groups have committed to produce a paper to be published in a special issue of a journal. After discussion among the participants, the organizers started negotiating with different journals about this possibility, and finally the journal chosen was the International Journal of Data Warehouse and Mining. It is expected that the papers will be submitted to the journal in January 2012 so that the publication of the special issue will be at the end of 2012.

2 Table of Contents

Working Groups

Open Problems

3 Overview of Talks

3.1 Service Oriented BI

Alberto Abello (Universitat Politècnica de Catalunya – Barcelona, ES)

The traditional way to manage Information Technologies (IT) in the companies is having a data center, and licensing monolithic applications based on the number of CPUs, allowed connections, etc. This also holds for Business Intelligence environments. Nevertheless, technologies have evolved and today other approaches are possible. Specifically, the service paradigm allows to outsource hardware as well as software in a pay-as-you-go model. In this work, we will introduce the concepts related to this paradigm and analyze how they affect Business Intelligence (BI). We analyze to which extent it is possible to consider Business Intelligence just a service and use techniques like Cloud Computing, Service Oriented Architectures (SOA) and Business Process Modeling (BPM). Finally, we store the other way round. Since special attention must be paid to service companies' characteristics and how to adapt BI techniques to enhance services.

3.2 Data Mining for Business Intelligence: from Relational to Graph representation

Marie-Aude Aufaure (Ecole Centrale – Paris, FR)

Traditional data mining methods, e.g. clustering, association rules, etc. are based on a tabular representation of input data and mainly ignore the relations between objects. Data, always growing, are mainly unstructured (80distributed and come from heterogeneous data sources in an unpredictable way. Business Intelligence methods and tools need to take this big data challenge into account. Graphs are everywhere (social networks, web, biology, etc.) and can be considered as a way for managing structured, semi-structured and unstructured data. This structure can be used for traversing linked information, finding shortest paths, doing semantic partition, recommending and discovering potential interesting linked information. Methods and tools can be defined to exploit the graph structure of large repositories such as digital libraries, web, databases or data warehouses with their associated metadata. This presentation focuses on the extraction of graphs from relational databases and their aggregation. Open problems still remain: many graph models exist so the most appropriate one has to be chosen, how can we combine data mining methods with graphs algorithms to find communities that not only takes into account links between individuals, but also their similarities based on their own attributes, how can we summarize and aggregate these graphs and maintain their consistency.

3.3 Open Research Areas in Data Warehouses

Cristina Ciferri (University of Sao Paulo, BR)

In this talk, I highlight my main research areas of interest, especially those related to data warehousing. Regarding geographic data warehouses, I have proposed: (i) the SB-index and the HSB-index structures; (ii) the Spadawan and the Spatial Star Schema benchmarks; and (iii) a systematic way to investigate spatial data redundancy. Another research area of interest is the storage and recovery of images in data warehousing environments. I have been investigating how to store images and how to perform the ETL process, as well as how to handle aggregation levels and how to provide efficient query processing over image data warehouses. I also have interest in horizontal and vertical fragmentation of data warehouses, view materialization using OLAP signatures and GRASP, and analytical processing over XML documents.

Finally, other research topics include mining of medial data, data integration and provenance, and the proposal of index structures for biological databases, complex data such as image, audio and video, and time series data.

3.4 Spatial Data Warehouses (SDW): Research Topics and Open Issues

Ricardo Ciferri (University of Sao Carlos, BR)

In this talk, I presented the main research topics in data warehouse area that the researchers of the Database Group (DBG) of the Federal University of São Carlos (UFSCar), Brazil, have been investigating. The main research topic is spatial data warehouse. For this topic, it was proposed techniques to improve SOLAP query processing performance over star schemas through specialized indices, such as the SB-index and the HSB-index. Also, an improved strategy to process drill-across SOLAP queries was briefly described in this talk. The performance evaluation of SOLAP query processing also was investigated through the proposal of two benchmarks:

The Spadawan Benchmark and The Spatial Star Schema Benchmark.

Nowadays, the Database Group of the Federal University of São Carlos (Brazil) is investigating data warehouses that store vague spatial objects and spatio-temporal data warehouse performance.

3.5 My Current Research Interests

Maria Luisa Damiani (Università di Milano, IT)

This brief contribution is to illustrate my area of research and some research questions related to the use of spatial data warehouses. My current research is carried out in two main areas, both related to spatial and spatio-temporal data management. The first addresses issues related to location-based security and location privacy, i.e., how to use location information to regulate the access to sensitive resources and how to preserve the privacy of personal position data, respectively. The second line is concerned with spatio-temporal data modeling, in particular the conceptual modeling of spatial datawarehouse and the recent paradigm of semantic trajectories, i.e., how to enrich the movement information with semantics. Spatial data warehouses have a potential in all these applications. Spatial data warehousing has been a “promising” area of research since early 2000. The key question is whether this promise has turned into a concrete result. While map visualization is extensively used in conventional data warehouses, it is still not clear whether any real progress has been made in the direction of building real and large scale spatial data warehouses. Moreover, it is not clear whether the paradigm founded on the notion of dimension measure and fact is sufficient to deal with the specific features of novel spatio-temporal representations like semantic trajectories. These are questions that are worth being discussed.

3.6 Towards spatio-temporal datawarehouses

Maria Luisa Damiani (Università di Milano, IT)

The widespread use of positioning technologies makes it relatively easy to track objects across indoor and outdoor settings. Large amounts of data on the traces of entities can be easily collected, possibly reporting not only their time varying positions but also information about the context in which the movement takes place, for example acquired through sensors. In this way the behavior of moving entities can be more easily analyzed. An emerging paradigm for the representation of the behavior of moving entities is that of semantic trajectories. A semantic trajectory is a trajectory which is temporally bounded (i.e., has a start and an end) and which represents the movement carried out by an object for some specific purpose. A semantic trajectory is a sequence of stops and moves where both stops and move are annotated semantically, for example a stop can represent a railway station while the move can be a particular train line. The question we pose is how to obtain aggregated information out from large bulks of semantic trajectories so as to build models of collective behavior. This calls for the definition of novel OLAP algebras defined over sequences of multidimensional data and integrating visual analytics functionalities.

3.7 Privacy and performance of cloud data warehouses

Jerome Darmont (Université Lumière Lyon 2, FR)

We address in this presentation two critical issues in cloud BI. With respect to data privacy and availability, we distribute data over several cloud service providers (CSP) using Shamir's Secret Sharing cryptographic scheme so that:

(1) each CSP cannot exploit the data it stores; (2) a majority of CSP need to band together to break the secret, or fail at the same time to render data unavailable. With respect to query performance, we propose to balance the increase of CPU power (and cost) with classical performance optimization data structures, i.e., materialized views (at a storage cost). Our aim being to find the best cost tradeoff, we designed cost models that help output an optimal configuration of CPU instances and materialized views, thus providing flexible control over both power and budget.

3.8 BI/OLAP research themes

Todd Eavis (Concordia University – Montreal, CA)

In this talk I provide an overview of a number of current research issues in the OLAP/BI domain. Topics include: Parallel Computing, Indexing, query ptimization compression, exploitation of GPUs, OLAP languages, native language (Java) querying, Query rewriting, Web-based interfaces, drag and drop BI queries, and conceptual modelling.

3.9 Knowledge Discovery & Management: My Point Of View

Cécile Favre (Université Lumière Lyon 2, FR)

URL http://eric.univ-lyon2.fr/ cfavre/DAGSTUHL/-Favre_DAGSTUHL_KnowledgeDiscovery&Management_MyPointOfView.pdf

This talk was prepared in order to begin the discussion in the group. This includes (1) a first list for using Knowledge Discovery and Management in Data Warehouses, (2) the detail about my own works concerning the topics of the group, (3) the links with the other groups, and at the end (4) other issues for the group.

3.10 Research Interests

Cecile Favre (Université Lumière Lyon 2, FR)

This is an overview of my past, current and future research work related to datawarehousing. This includes Personalization in Data Warehouses, Data Warehouse Modeling and the combination of Social Networks and Data Warehouses.

3.11 Column Stores and Query Processing for real-time BI

David Fekete (Universität Münster, DE)

Column Stores, as column-based Database Management Systems (DBMS), have come to be often regarded as alternative to typical row-based DBMS (Row Stores) in Data Warehousing (DWH) and Business Intelligence (BI) scenarios. Column Stores promise to improve read-performance in those read- intensive, analytical (On-Line Analytical Processing, OLAP) scenarios significantly. But especially query processing is an integral part in realizing this speed-up potential and subject to many optimization techniques, like Block Iteration, Late Materialization or Compression. A query executor has to utilize the columnar database structure to enable the promised performance advantages. On the way from "occasional OLAP" to "real-time" BI there are several possible issues with regard to query processing. Those include the contribution of these Column Store technologies to real-time BI or the impact of novel hardware technologies, such as Solid State Drives (SSD) and multi-core CPUs.

3.12 Collaborative Business Intelligence

Matteo Golfarelli (Università di Bologna, IT)

Cooperation is seen by companies as one of the major means for increasing flexibility and innovating. Business intelligence (BI) platforms are aimed at serving individual companies, and they cannot operate over networks of companies characterized by an organizational, lexical, and semantic heterogeneity. In this presentation we summarize different approaches to collaborative BI and we propose a framework, called Business Intelligence Network (BIN), for sharing BI functionalities over complex networks of companies that are chasing mutual advantages through the sharing of strategic information. A BIN is based on a network of peers, one for each company participating in the consortium. Peers are equipped with independent BI platforms that expose some querying functionalities aimed at sharing business information for the decision-making process. After proposing an architecture for a BIN, we outline the main research issues involved in its building and operating.

3.13 Research on DW, OLAP and Mining

Leticia I. Gomez (Instituto Tecnológico de Buenos Aires, AR)

We discuss different challenges in Business Intelligence. Traditional Data Warehouse and OLAP tools store and manipulate strings, numbers and timestamp data types in a native way. Incorporating new data types in DW enriches the capability of analysis. For instance, spatio-temporal data recorded by GPS and RFID sensors can be used for discovering trajectories hidden patterns. In addition, as new types of data are incorporated into DW and Mining, novel techniques have to be proposed related to physical and logical design of data as well as new forms of human interaction. We showed briefly the idea of some related projects.

3.14 Sequential OLAP

Christian Koncilia (Universität Klagenfurt, AT)

Although Business Intelligence is a still a hot topic in computer science, little effort has been put on how to analyze sequences in data warehouses.

In my talk I presented a novel approach which enables the user to analyze sequential data within a standard OLAP environment. This approach consists of two core functions (SOLAP() and RTA() where RTA is short for "Real Time Axis") which both return a standard OLAP cube. Hence, the user can still use all the well-known OLAP operations like drill-down, roll-up, slice, etc. to analyze the cube. Furthermore, this approach may be applied to all data warehousing solutions.

3.15 Data warehouse mining: what about mining query logs?

Patrick Marcel (Université de Tours – Blois, FR)

This presentation gives a brief overview of the researches lead at the computer science lab of Université François Rabelais Tours on mining OLAP query logs.

Two examples are presented. In the first one, a joint work with Università di Bologna, association rules are extracted from a user's query log for presonalising the user's queries on the fly.

In the second one, a multi user log is analysed to recommend queries for supporting discovery driven analyses.

Some research perspectives around OLAP log mining are underlined.

3.16 Open Business Intelligence

Jose-Norberto Mazon (Univ. of Alicante, ES)

Citizens demand an increasingly transparent behavior of public institutions.

Importantly, transparency implies that public data should be available with the aim of providing the greatest benefit to the wider society through an active participation of citizens. Therefore, public data should be freely available to be easily used, reused and redistributed by anyone, i.e. open data. Consequently, open data are generally shared as a raw data which, unfortunately, prevents non-expert citizens from analyzing them to acquire actionable information. Mechanisms that allow citizens to analyze and understand open data in a user-friendly manner are thus highly required.

To this aim, the concept of Open Business Intelligence (OpenBI) is introduced. OpenBI facilitates non-expert users to (i) analyze and visualize open data, thus generating actionable information by means of reporting, OLAP analysis, dashboards or data mining; and to (ii) share the new acquired information as open data to be reused by anyone. As a consequence, OpenBI requires the development of systematic approaches for guiding non-expert users in obtaining and sharing the most reliable knowledge from the available open data. Finally, it is worth noting that OpenBI has to deal with several research challenges: the extraction and integration of heterogeneous open data sources by non-expert users (as citizens), tackling data quality issues in a user-friendly manner, intuitive data visualization, smooth integration of unstructured data, and so on.

3.17 Mining Association Rules from Data Cubes

Rokia Missaoui (Université du Québec en Outaouais, CA)

Substantial work has been conducted in knowledge discovery from data warehouses. This includes (but is not limited to) outlier detection in multidimensional data, cubegrade generation, constrained gradient analysis, association rule mining, and discovery-driven examination of cubes. To the best of our knowledge, Kamber et al. were the first researchers who addressed the issue of mining association rules (ARs) from multidimensional data. They introduced the concept of metarule-guided mining which consists in using rule templates defined by users in order to guide the mining process.

The objective of the present talk is to show that triadic concept analysis can be used as a formal data mining framework to generate clusters (closed trisets) and triadic association rules from data cubes in an efficient and meaningful way. The merits of triadic association rules over dyadic ones lie in the fact that they represent patterns in a more compact and meaningful way than association rules extracted from one of the possible flattened forms of the data cube.

3.18 BI Research Overview Gong Show

Torben B. Pedersen (Aalborg University, DK)

This talk will introduce myself and my research within BI. I have worked on real- time DW and BI, spatio-temporal DWs, ETL frameworks and testing, compressed bitmaps, On-demand integration of cubes and XML, multidimensional schema discovery, contextualized warehouses, semantic web warehousing, and mining of data cubes.

3.19 Efficient estimation of Joint Queries from Multiple OLAP Databases

Elaheh Pourabbas (National Research Council – Rome, IT)

In this study, we investigate the problem of estimation of a target database from summary databases derived from a base data cube. We show that such estimates can be derived by choosing a primary database with the desired target measure but not the desired dimensions, and use a proxy database to estimate the results. The technique we use is linear indirect estimation, commonly used for statistical estimation. We examine two obvious computational methods for computing such a target database, called the "Full cross product" (F) and the "Pre-aggregation" (P) methods. We study the accuracy and computational cost of these methods. While the F method provides a more accurate estimate, it is more expensive computationally than P. Our contribution is in proposing a third new method, called the "Partial Pre- aggregation" method (PP), which is significantly less expensive than F, but is just as accurate. We prove formally that the PP method yields the same results as the F method, and provide analytical and experimental results on the accuracy and computational benefits of the PP method. Then, we consider the problem of how to select a primary and a proxy database given that there are multiple primary databases available with the same measure and multiple proxy databases with the desired target dimensions in order to get the most accurate estimation results.

3.20 Spatial and Spatio-Temporal Data Warehousing

Markus Schneider (University of Florida – Gainesville, US)

Traditional data warehousing and OLAP deal with decision support for alphanumerical data. However, new emerging applications produce new kind of data categories like spatial, spatiotemporal, and biological data. The question is how these data can be be made available and leveraged for decision support. The nature of these new data categories is that they lead to complex (variable-length, multi-structured, and hierarchical) data objects and that aggregate operations on them are currently not supported. Spatial data warehousing and Spatial OLAP

(SOLAP) are promising extensions with respect to spatial decision support. A spatial data warehouse is a full-fledged data warehouse with additional capabilities to store, retrieve, query, and analyze spatial data represented as values of spatial data types. In addition, spatial data collections can be spatial connectivity structures like spatial partitions (maps) and spatial graphs (spatial networks) that satisfy certain topological constraints. Spatial OLAP operations are full-fledged OLAP operations that include spatial objects as values of spatial data types in dimensions and measures and that make use of spatial aggregation operations like spatial union, spatial intersection, convex hull, centroid, nearest neighbor, and contour. These features require a basic logical data model for formally defining and integrating needed spatial data warehouse and OLAP concepts. Since spatial data types are abstract data types that hide the internal representation of geometries, the architecture and infrastructure of existing data warehousing and OLAP technology can be used. Extensibility mechanisms at all levels of the architecture of a data warehouse system and an OLAP system are needed to be able to integrate new data categories and aggregation operations. At the conceptual level, a user-centric modeling strategy is needed that abstracts from the logical level and emphasizes user considerations. Unfortunately, a standardized conceptual model for multidimensional data modeling as well as a standardized query interface for data warehouses do not exist. Our main idea is to take the cube metaphor literally as a conceptual user model and to design cube types as abstract data types. An analysis or query language then provides high-level OLAP operations like roll-up, drill-down, slice, etc.

3.21 IBM Smart Analytics Optimizer - Technical Introduction

Knut Stolze (IBM Deutschland – Böblingen, DE)

The IBM Smart Analytics Optimizer (ISAO) is a new product designed to enhance and improve IBM's relational database system DB2 for z/OS for data warehouse and BI applications. ISAO builds on several new technologies to dramatically improve query performance for long-running queries. All data is stored in main memory and advanced processor features like SIMD operations are exploited. The presentation gave an overview on the integration aspects for ISAO into DB2 for z/OS. The architecture and data flows for query processing and data loading were explained on a high-level as well.

3.22 Business Intelligence 2.0: A general overview

Juan Trujillo (Univ. of Alicante, ES)

Business Intelligence (BI) solutions allow decision makers to query, understand, and analyze their business data in order to make better decisions.

However, as the technology and society evolve, faster and better informed decisions are required. Nowadays, it is not enough to use only the information from the own organization and making isolated decisions, but rather requiring also to include information present in the web like opinions or information about competitors, while using collective intelligence, collaborating through social networks, and supporting the BI system with cloud computing. In response to this situation, a vision of a new BI 2.0, based on the evolution of the web and the emerging technologies, arises. However, different researchers differ in their vision of this BI evolution. In this talk, we provide an overview of the aspects proposed to be included in BI 2.0. We describe which success factors and technologies have motivated each aspect. Finally, we review how tool developers are including these new features in the next generation of BI solutions.

3.23 Data Warehouse design and consistency through trace metamodels on a hybrid approach

Juan Trujillo (Univ. of Alicante, ES)

Data warehouses (DW) integrate several heterogeneous data sources in multidimensional structures in support of the devision-making process. The complexity of the DW development process requires to follow a methodological approach in order to be successful. Mainly three approaches have been proposed to tackle this problem in a similar way to software product development, bottom-up, top-down, and an hybrid approach. The hybrid approach, being the one with the most advantages, still suffers from a drawback, requirements and data sources must be reconciliated, acommodating either one or both, to a new DW model. In this process, relationships between requirements, data sources and conceptual elements is lost, since no traceability is included. In turn, this hurts the requirements validation, lowering the user satisfiability (since the validation is not done until the DW is implemented), makes the derivation of the platform specific models incomplete (since currently data types are not included at the conceptual level in DW models), and increases the complexity of the design extraction, transformation and load processes (since the initial relationships between the source tables and the target ones could have been used but was lost). In this paper, we review the proposals for traceability in the Requirements Engineering (RE) and Model Driven Development (MDD) fields and propose a metamodel and a set of Query/View/Extraction (QVT) transformations to include traceability in the DW field, solving the aforementioned problems and enabling impact change analysis in an easy way, increasing the user satisfaction.

3.24 Research in Data Warehousing, OLAP & Beyond

Alejandro Vaisman (Université Libre de Bruxelles, BE)

In this presentation I summarize my research in the field of Data Ware- housing and OLAP, and present current and future research directions I am aimed at following. With respect to the former, I briefly discuss my work on

1. Dimension updates in OLAP;
2. Temporal OLAP;
3. GIS and OLAP integration;
4. Trajectory Mining.

Regarding future research directions, I will comment my work on Spatio-Temporal SOLAP (specifically the temporal extension of the Piet data model), the extension of SOLAP to support Raster data, and my work on RDF and the semantic web, in particular, the need to support semantic web BI.

3.25 Near-Real-Time & Evolving ETL

Panos Vassiliadis (University of Ioannina, GR)

This presentation refers to two important topics for the area of ETL: near real time ETL and evolution of ETL scenarios. The talk discusses the area and then moves on to the presenter's high-level point of view on how the problems should be handled. In the end, there is a quick discussion of other alternatives, too.

3.26 Business Intelligence as a Service, in the Cloud

Gottfried Vossen (Universität Münster, DE)

Business intelligence (BI) has long been provided by tools and software systems that needed local installation. More recently, efforts have successfully been launched to provide BI as a service of the Internet, thereby making BI amenable to cloud computing. We briefly review why this is relevant, what potential drawbacks and problem areas are, and which questions seem most relevant to consider.

3.27 Some not fully solved problems in data warehouse research

Robert Wrembel (Poznan University of Technology, PL)

Despite over 20 years of conducted research in the data warehousing area, there exist several either only partially solved or unsolved research problems. Some of them, which are being researched at the Institute of Computing Science at the Poznan University of Technoloyg, are outline in this presentation. These problems include: (1) handling the evolution of ETL, (2) bitmap compression algorithms on GPUs, and (3) indexing dimension data. The first problem is caused by the evolution of external data sources used for delivering data into a data warehouse. These sources change not only its content but also its structures. Structural changes impact all the layers in a data warehouse architecture. One of the layers includes ETL. In the project that we conduct, we try to develop a framework for semi-automatic or automatic (if possible) adjustment of the ETL layer to structural changes in data sources. The second problem concerns the application of Graphic Processing Units to data processing.

The processing power of GPUs and massive processing parallelism that can be achieved offers a promising framework. In our work, we port various bitmap index compression algorithms to the GPU platform. The third problem concerns the development of efficient indexing techniques for dimensions. In our work we exploit the fact that the most frequent queries in a data warehouse join fact a table with the Time dimension and the fact that dimensions have often a hierarchical structure. So far, we have proposed an index, called Time-HOBI that eliminates the need for joining a fact table with its Time dimension.

3.28 What is Spatio-Temporal Data Warehousing?

Esteban Zimányi (Université Libre de Bruxelles, BE)

In the last years, extending OLAP (On-Line Analytical Processing) systems with spatial and temporal features has attracted the attention of the GIS (Geographic Information Systems) and database communities. However, there is no a commonly agreed definition of what is a spatio-temporal data warehouse and what functionality such a data warehouse should support.

Further, the solutions proposed in the literature vary considerably in the kind of data that can be represented as well as the kind of queries that can be expressed. In this presentation I present a conceptual framework for defining spatio-temporal data warehouses using an extensible data type system. This is based on a taxonomy of different classes of queries of increasing expressive power.

4 Working Groups

4.1 Situational BI

Stefano Rizzi (University of Bologna, IT)

When decisions have to be made quickly and under uncertainty in complex and dynamic environments, the selection of an action plan requires external business variables coupled with variables on company performance. In situational BI, data from a company's data warehouse are spontaneously correlated with "external" information sources that may come from the corporate intranet, be acquired from some external vendor, or be derived from the internet.

4.2 Situational BI Revisited

Stefano Rizzi (University of Bologna, IT)

The research group in charge of the situational BI topic describes a typical interaction scenario and the related research challenges.

4.3 On-demand self-service BI

Torben B. Pedersen (Aalborg University, DK)

This talk summarizes the work of the Dagstuhl group on on-demand self-service BI. It answers the five questions put forward by the organizers and presents the paper skeleton of the paper to be written by the group.

4.4 Research Problems in Data Warehousing

Markus Schneider (University of Florida – Gainesville, US)

Data warehousing and online analytical processing (OLAP) have produced fundamental and important technologies for the design, management, and use of information systems for decision support. This is manifested by their acceptance in many industrial and administrative organizations that have successfully leveraged these technologies to improve their business processes and their decision making efforts. However, these technologies tend to be largely system-centric, that is, system aspects at the logical and physical levels are emphasized and be visible to the user. Users like managers or system analysts are usually supported by sophisticated GUI tools.

But these tools have a logical model like a snowflake schema or a constellation schema as their basis which make it difficult to understand the data. Therefore, a user-centric modeling strategy is needed at the conceptual level that focuses on an appropriate user view on data warehouses and OLAP operations. In general, a consensus has to be achieved for a standardized conceptual model for multidimensional data. The modeling framework to be developed (and consequently the whole data warehouse architecture as well as the OLAP operations) should be extensible so that new kinds of data categories like spatial, spatiotemporal, and biological data can be be made available and used for decision support. The nature of these new data categories is that they lead to complex (variable-length, multi-structured, and hierarchical) data objects and aggregate operations on them that are currently not supported. Finally, a standardized query interface is needed both at the conceptual level and at the logical level.

4.5 Query Processing in Data Warehouses

Anisoara Nica (Sybase, An SAP Company, CA)

Joint work of Nica, Anisoara; Eavis, Todd; Fekete, David; Graefe, Goetz; Neumann, Thomas; Petrov, Ilia; Pourabbas, Elaheh; Stolze, Knut

This talk is the first presentation of the Dagstuhl group on "Query Processing in Data Warehouses". It answers the five questions put forward by the organizers and summarizes the most relevant research problems in query processing in the data warehouse environment. This includes query processing in the elastic cloud as well as parallel query processing.

4.6 Query Processing in the Elastic Cloud

Anisoara Nica (Sybase, An SAP Company, CA)

Joint work of Eavis, Todd; Fekete, David; Graefe, Goetz; Neumann, Thomas; Nica, Anisoara; Petrov, Ilia; Pourabbas, Elaheh; Stolze, Knut

This talk is the second presentation of the Dagstuhl group on "Query Processing in Data Warehouses". It revises the answers to the five questions put forward by the organizers, and it addresses one of the relevant research issues discussed by the group, namely query processing in the elastic cloud. A central promise of cloud services is elastic, on- demand provisioning. For data-intensive services such as data management, growing and shrinking the set of nodes implies copying data to nodes with temporary membership in a service. At best, a node might retain its copy of the data while it provides another service; at worst, a node that rejoins the database service requires a new copy of data. Many solutions have been proposed to the elasticity for cloud services, but few address the problems raised by "sometimes available copy" nodes. The talk proposes a new approach to elasticity for the environments where sometimes available copies can be used by permanent nodes for query processing.

4.7 Knowledge Discovery and Management in Data warehouses

Rokia Missaoui (Université du Québec en Outaouais, CA)

The presentation gives the preliminary view of the KDMD group about integrating data, information, and knowledge in data warehouses in a user-centric manner in order to help the end-user get better benefits from data by allowing (i) a unified and meaningful representation of multidimensional data as well as knowledge patterns, and (ii) advanced query mechanisms and guidance to get targeted information and knowledge through information retrieval and data mining techniques.

4.8 Knowledge Discovery and Management in Data warehouses: Final report

Rokia Missaoui (Université du Québec en Outaouais, CA)

The presentation focuses on some of the seven specific objectives stated in the Wednesday talk by the KDMD group. In particular, a user-centric architecture is presented with some illustrations ofadvanced OLAP query processing such as cooperative and intensional query management using both data and knowledge.

4.9 Right time BI framework

Robert Wrembel (Poznan University of Technology, PL)

Business Intelligence (BI) has grown to a powerful method for supporting strategic decisions in companies. Data warehouses are the most promising software tool support in that area, providing services to structure, index and query large amounts of business data to find patterns and business opportunities. In traditional data warehousing, large amounts of data are handled by bulk-loading them through and ETL process into the data warehouse's database. This process typically has good throughput at the cost of very high latencies between the data occurring and becoming visible to a business user. Our approach aims at shortening this delay tremendously, without a perceivable impact on throughput. We advocate a two-folded approach: first, we introduce a fast track data path, which uses techniques from event-based systems to analyze incoming data on the fly and raise an alarm in case of a certain pattern appearing. Second, we suggest on-demand view updates, where incoming data is stored in an unstructured way and only structured and indexed upon a query requesting these data. This allows us to avoid handling data that is never queried and keep each loading cycle as lightweight as possible.

5 Open Problems

We present next the position statement of the five groups.

5.1 Real-Time Data Warehouses and Business Intelligence

In a typical BI infrastructure, data, extracted from operational data sources, is transformed and cleansed and subsequently loaded into a data warehouse where it can be queried for reporting purposes. ETL, the process of extraction, transformation, and loading, is a periodic process that may involve an elaborate and rather established software ecosystem. Typically, the actual ETL process is executed on a nightly basis, i.e., a full day's worth of data is processed and loaded during off-hours. Depending on the resources available and the nature of the data and the reporting, ETL may also be performed more frequently, e.g., on an hourly basis.

It is desirable to reduce this delay further and ideally provide reports and business insights at real-time. However, this requires overcoming throughput bottlenecks and improving latency throughout the ETL process. Instead of attempting to incrementally improve this situation, we propose a radically different approach: leveraging a data warehouse's capability to directly import raw, unprocessed records, we defer the transformation and cleaning of data until needed by pending reports. At that time, the database's own processing mechanisms can be deployed to process the data on-demand.

This technique ensures that resources are utilized optimally rather than spent speculatively on processing, potentially irrelevant, data in bulk. Besides excluding irrelevant data from processing all together, a more likely scenario is the case where different types of reports are run at different times of the day, week, or month and require different categories of source data. Again, using an on-demand approach helps optimize resource utilization and improves data freshness substantially.

In addition to running periodic reports, an important component of modern BI architectures are interactive elements such as alerting when detecting outliers or encountering exceptional situations during data processing. While highly efficient, our on-demand approach naturally lacks the continuous processing and thus 'surveying' of data of the traditional technique. In order to capture such events, we augment our on-demand approach with an active component that performs light-weight data screening independent of the ELT processing and may be integrated with a BI dashboard.

Besides outlining an overall architecture, we also developed a roadmap for implementing a complete prototype using conventional database technology in the form of hierarchical materialized views.

5.2 Spatio-Temporal Data Warehouses

Modern organizations nowadays need to use OLAP analytical capabilities together with geographical information. In this direction, SOLAP (standing for Spatial OLAP) aims at exploring spatial data by drilling on maps, in the same way as OLAP operates over tables and charts. However, SOLAP only accounts for discrete spatial data, where spatial objects are represented as geometries. More sophisticated GIS-based decision support systems are increasingly being needed, able to handle more complex types of data, like continuous fields (from now on, fields). Fields describe physical phenomena that change continuously in time and/or space, like temperature, land elevation, land use and population density. They are perceived as having a value at each point in a continuous N-dimensional spatial and/or spatio-temporal domain.

In real-world practice, scientists and practitioners register the values of a field by taking samples at (generally) fixed locations, and inferring the values at other points in space using some interpolation method. Thus, fields can be described by a function that indicates the distribution of the phenomena or feature of interest. The most popular discrete representation for fields is the raster model, where the 2D space is divided into regular squares. The raster model is frequently used for representing soil type, temperature, among other physical phenomena. Other representations have also been proposed, like the Voronoi diagrams or TIN, the latter usually employed to represent an 'Altitude' field.

For adding spatial information to OLAP tools, many models have been proposed, at the conceptual and logical levels. We find this unnecessary from the user's point of view. We believe that a user of a spatio-temporal enabled OLAP system would like to find the data cube at the conceptual level, independently of the kind of underlying data. Such a model would allow to analyze any type of spatial data, continuous (independently of the underlying representation) or discrete, together with typical alphanumerical discrete OLAP data, using only the classic OLAP operators, like Roll-up, Drill-down, and/or Drill-across. To achieve this, at the logical and physical levels we need different mechanisms to manage these different kinds of data and data representations. That means, the final user only sees the typical OLAP operators at the user level. At the logical level operations that allow to manage different kinds of spatial data are needed, independently of their physical representation. Finally, at the physical level, we need a model that allows managing continuous and discrete data, and a collection of operators over this model. For continuous data, it is at this level where we need to care about raster, Voronoi, and/or TIN representations.

5.3 Situational Business Intelligence

Well-informed and effective decisions often require a tight relationship to be established between the variables on company performance stored in the DW and other data that are not even resident in the company information system. These valuable data may be related for instance to the market, to competitors, or to potential customers; they are sometimes called situational data because they have a narrow focus on a specific business problem and, often, a short lifespan for a small group of users with a unique set of needs. In some cases, situational data can be retrieved (for free or on a fee) in a semi-structured form by accessing registered data providers. In other cases, situational data are chaotically scattered across several, heterogeneous, unstructured sources available on the Web (e.g., opinions expressed by users on social networks, ratings of products on portals, etc.). As such they tend to be highly dynamic in contrast with common enterprise data, which are used to address a large set of business problems and impose a slow and careful management.

The capability of incorporating situational data into the decisional process gives rise to a new class of applications, that in the context of BI 2.0 are often labeled as situational BI or on-demand BI or even collaborative BI. Here we prefer to use the term self-service BI, to emphasize that the search, extraction, and integration of situational data should be accomplished in near-realtime by business users through a continuous interaction with the application, without any mediation or intervention by analysts, designers, or programmers. As also emphasized by Gartner Inc., self-service BI appears to be the big wave in BI for 2011; the key idea is to let end users navigate information in a "surf and save" mode, meaning that data can be stored for reuse or sharing. Among the main applications for self-service BI, we mention brand positioning, pricing, competitor monitoring, policy planning, risk management; the domains involved range from retail, telco, and entertainment to finance and public services such as health and transportation. In all these domains, the strategic information needs are currently not fully satisfied by traditional BI approaches and could be positively affected by including situational data.

5.4 Query Processing in Data Warehouses: Elasticity in Cloud Databases

A central promise of cloud services is elastic, on-demand provisioning. For data-intensive services such as data management, growing and shrinking the set of cloud nodes implies copying data to nodes with temporary membership in a service. At best, a node might retain its copy of the data while it provides another service; at worst, a node that rejoins the database service requires a new copy of data. Many solutions have been proposed to the elasticity for cloud services, but few address the problems raised by "sometimes available copy" (SAC) nodes with on-demand, incremental updates. We believe that SAC nodes should become up-to-date and useful for query processing incrementally by key range. On-demand, based on the queries being evaluated, additional key ranges for a SAC node become up-to-date, with overall update performance comparable to a traditional high-availability strategy that carries the entire dataset forward, until eventually the entire dataset becomes up-to-date and useful for query processing. The maintenance scheme combines log-based replication and updates with sorting data in transit. Our proposed solution relies on techniques from partitioned B-trees, adaptive merging, deferred maintenance of secondary indexes and of materialized views, and query optimization using materialized views. The paper introduces a family of maintenance strategies for temporarily available copies, the space of possible query execution plans and their cost functions, and query processing techniques for this type of elastic environments.

5.5 Knowledge Extraction and Management in Data Warehouses

The objective of this research is to define techniques towards integrating data, information, and knowledge in data warehouses in a user-centric manner in order to help the end user get better benefits from data by allowing (i) a unified and meaningful representation of multidimensional data as well as knowledge patterns, and (ii) advanced query mechanisms and guidance to provide targeted information and knowledge through information retrieval and data mining techniques.

In this work, we first outline the importance of knowledge (e.g., existing ontology and knowledge discovered from data) in data warehouse management, propose a user-centric architecture for OLAP query processing and then define new solutions towards intensional and cooperative query answering using knowledge and exploiting the peculiarities of data warehouses. Such work aims to help a user (who is not necessarily familiar with a query language or aware about the detailed structure and content of the data warehouse) get a richer and more complete and even compact answer to his/her possibly incomplete or vague OLAP query. For example, an executive may ask for the top sales in a given time period, and the answer could be the concise and semantically rich one: the top sales concern almost all sport products bought by customers from the West Coast.

Participants

- Alberto Abello
Universitat Politècnica de Catalunya – Barcelona, ES
- Marie-Aude Aufaure
Ecole Centrale – Paris, FR
- Cristina Ciferri
University of Sao Paolo, BR
- Ricardo Ciferri
University of Sao Carlos, BR
- Alfredo Cuzzocrea
ICAR-CNR – Rende, IT
- Maria Luisa Damiani
Universitá di Milano, IT
- Jérome Darmont
Université Lumière Lyon 2, FR
- Todd Eavis
Concordia Univ. – Montreal, CA
- Lorena Etcheverry
Universidad de la Republica – Montevideo, UY
- Cécile Favre
Université Lumière Lyon 2, FR
- David Fekete
Universität Münster, DE
- Tobias Freudenreich
TU Darmstadt, DE
- Pedro Furtado
University of Coimbra, PT
- Matteo Golfarelli
Università di Bologna, IT
- Leticia I. Gómez
Instituto Tecnológico de Buenos Aires, AR
- Goetz Graefe
HP Labs – Madison, US
- Christian Koncilia
Universität Klagenfurt, AT
- Patrick Marcel
Université de Tours – Blois, FR
- Jose-Norberto Mazón
Univ. of Alicante, ES
- Rokia Missaoui
Université du Québec en Outaouais, CA
- Felix Naumann
Hasso-Plattner-Institut – Potsdam, DE
- Thomas Neumann
TU München, DE
- Anisoara Nica
Sybase, An SAP Company, CA
- Torben B. Pedersen
Aalborg University, DK
- Ilia Petrov
TU Darmstadt, DE
- Elaheh Pourabbas
National Research Council – Rome, IT
- Stefano Rizzi
University of Bologna, IT
- Markus Schneider
University of Florida – Gainesville, US
- Knut Stolze
IBM Deutschland – Böblingen, DE
- Maik Thiele
TU Dresden, DE
- Juan Trujillo
Univ. of Alicante, ES
- Alejandro Vaisman
Université Libre de Bruxelles, BE
- Panos Vassiliadis
University of Ioannina, GR
- Gottfried Vossen
Universität Münster, DE
- Florian M. Waas
EMC – San Mateo, US
- Robert Wrembel
Poznan Univ. of Technology, PL
- Esteban Zimányi
Université Libre de Bruxelles, BE