Business Intelligence

Data Mining and
Data Warehousing

Dominik Ślęzak

slezak@infobright.com

www.infobright.com

Research Interests

Data Warehouses, Knowledge Discovery, Rough Sets

Machine Intelligence, Hybrid Information Technologies

Business Intelligence, Multimedia / Medical Databases

Scientific Degrees

PhD in Computer Science, Warsaw University, Poland
Faculty of Mathematics, Informatics and Mechanics

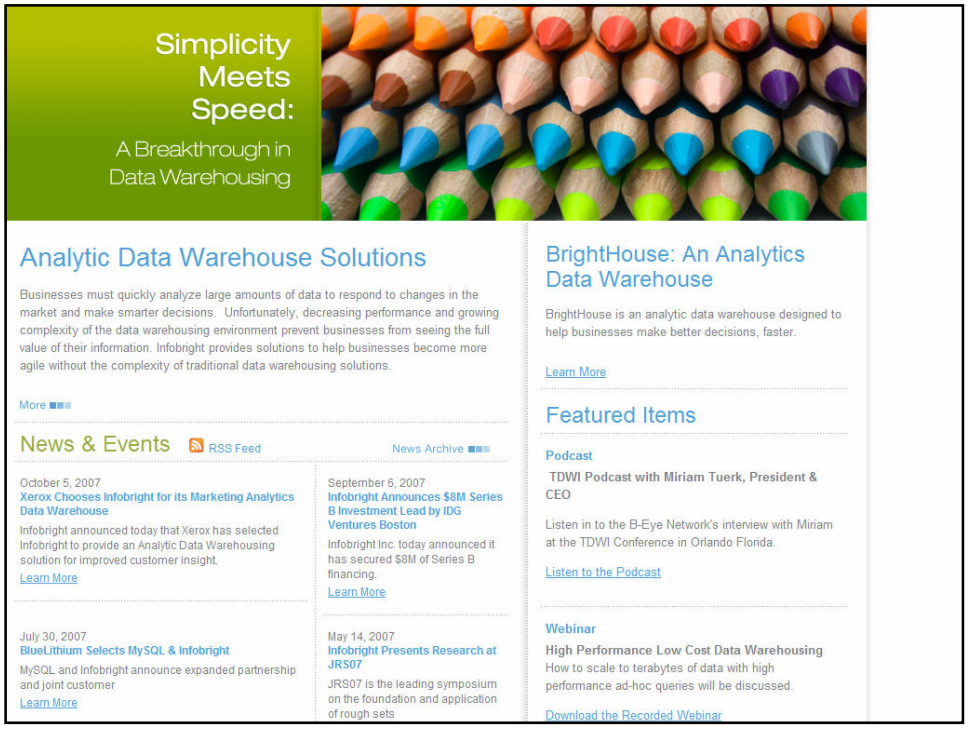
MSc/BSc in Mathematics, Warsaw University, Poland
Faculty of Mathematics, Informatics and Mechanics

Academy Record

2008 to date	York University, Canada Adjunct Professor
2008 to date	McMaster University, Canada Adjunct Professor
2006 to date	University of Regina, Canada Adjunct Professor
2004 to date	PJIIT, Warsaw, Poland Adjunct Professor
2003 – 2006	University of Regina, Canada Assistant Professor
2002 – 2004	PJIIT, Warsaw, Poland Assistant Professor
1995 – 2002	PJIIT, Warsaw, Poland Teaching Assistant

Industry Record

2006 to date	Infobright Inc., Canada/Poland Chief Scientist (& Co-Founder)
1999 to date	QED Software Inc., Poland President



Course Contents

- Business Intelligence
- Data Warehouses

  Monday, 18:00 – 20:30
- Data Mining

  Wednesday, 18:00 – 20:30
- Presentations

  Saturday, 9:00 – 11:30 (++)

Wikipedia (50 Years of BI!!!!!)

The term **business intelligence** (**BI**) dates to 1958. It refers to technologies, applications, and practices for the collection, integration, analysis, and presentation of business information and also sometimes to the information itself. The purpose of business intelligence is to support better business decision making. D. J. Power explains in *A Brief History of Decision Support Systems*:

> **BI describes a set of concepts and methods to improve business decision making by using fact-based support systems. BI is sometimes used interchangeably with briefing books, report and query tools and executive information systems. BI systems are data-driven DSS.**

BI systems provide historical, current, and predictive views of business operations, most often using data that has been gathered into a data warehouse or a data mart and occasionally working from operational data. Software elements support reporting, interactive "slice-and-dice" pivot-table analyses, visualization, and statistical data mining. Applications tackle sales, production, financial, and many other sources of business data for purposes that include, notably, business performance management.

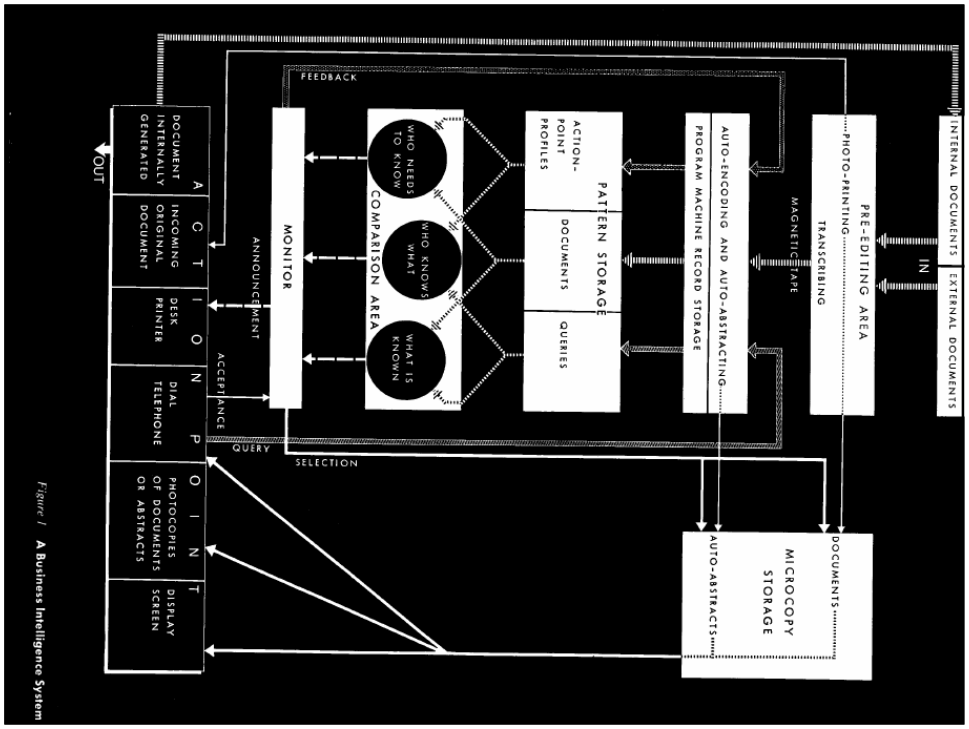


Figure 1 **A Business Intelligence System**

BI, Data Warehousing & Mining

(…) Business intelligence is a term commonly associated with data warehousing. In fact, many of the tool vendors position their products as business intelligence software rather than data warehousing software. There are other occasions where the two terms are used interchangeably. So, exactly what is business intelligence?

> Business intelligence usually refers to the information that is available for the enterprise to make decisions on. A data warehousing (or data mart) system is the backend, or the infrastructural, component for achieving business intelligence. Business intelligence also includes the insight gained from doing data mining analysis, as well as unstructured data (thus the need of content management systems). For our purposes here, we will discuss business intelligence in the context of using a data warehouse infrastructure. (…)

b•eye Business Intelligence Network™ The Global Vision for BI and Beyond

www.b-eye-network.com

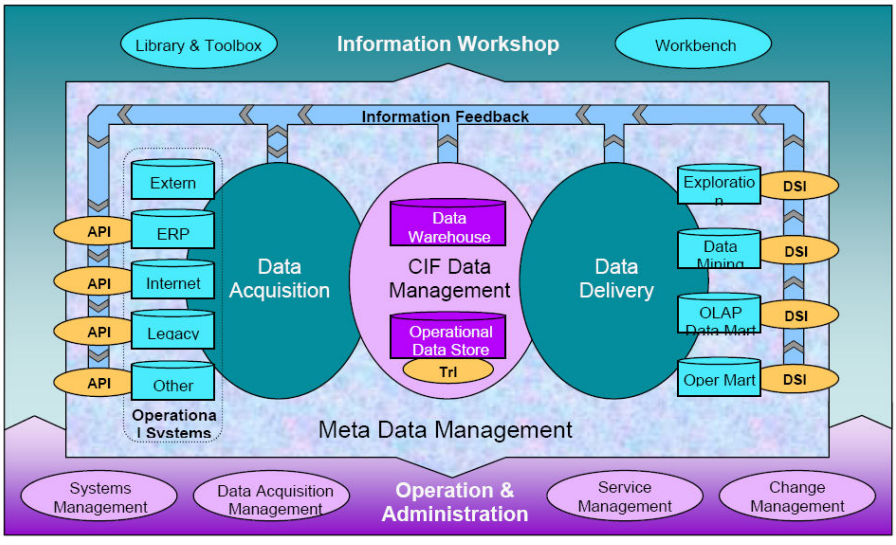
Three Categories of BI

Strategic BI provides performance metrics to management and executives. It is generally called performance management (PM). PM might be preceded by a C for corporate performance management, an E for enterprise performance management or a B for business performance management.

Tactical BI, called traditional and/or analytical, is the application of BI tools to analyze business trends, frequently comparing a specific metric (such as sales or expenses) to the same metric from a previous month or year. In most companies, there are usually a few analysts in each department who use online analytical processing (OLAP) and ad hoc query to perform this task. To date, BI tools are mostly used to analyze historical business data to discover trends or anomalies that need attention.

Operational BI delivers information to the point of business - the front lines of a business where information is used as part of an operational process. For example, when a person calls a toll-free number to speak to a customer service representative about his or her telephone bill, that representative will most likely be looking at a report about the caller's previous billing history and payment record.

Corporate Information Factory



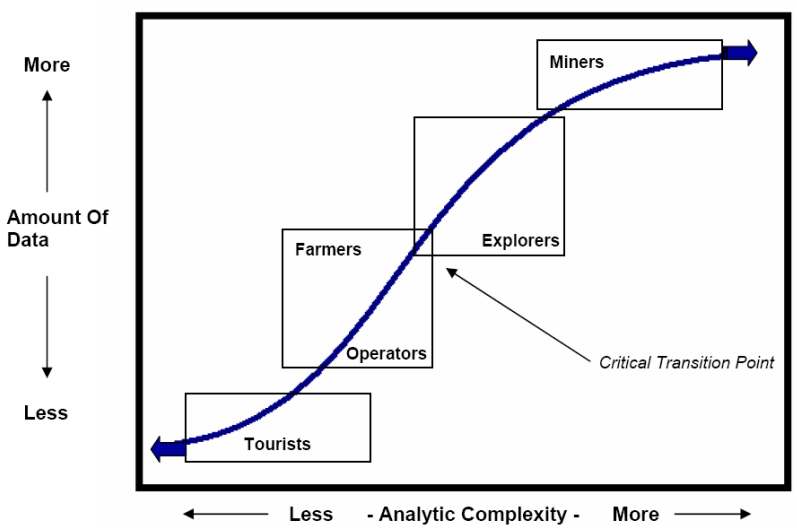
Five Communities that use the CIF

- **Farmers** usually come from the management or business planning groups. Typical farmers may be the financial analysts responsible for reporting on revenues and costs. They may be sales and product analysts determining how well a product is selling in some part of the world. They maybe the people who track campaigns or promotions from week to week or they may be the analysts who monitor the budget vs. actual reports.
- **Tourists** generally come from the executives or from very technical resources. Executives are among the most critical users not just because of their need to get information directly from the CIF – they can get information through other means. They are critical because their satisfaction – or dissatisfaction – can significantly impact the future viability of the CIF.

Five Communities that use the CIF

- **Operators** are the most common set of users of the CIF. Sometimes operators make special requests for information, but most of the time they need current detailed information on a scheduled basis, so they rely heavily on standardized queries. As information storage has increased via computers, more and more operators have adopted the view that all the information they need is in the system somewhere. If only they could get someone to give them access to it – and now! As a result, many of today's operators must continually struggle to consolidate and evaluate current information from disparate sources. The good news is that an alternative exists for effectively satisfying operators' needs for current, detailed, enterprise-wide, consolidated tactical information – the Operational Data Store.

The BI User Curve



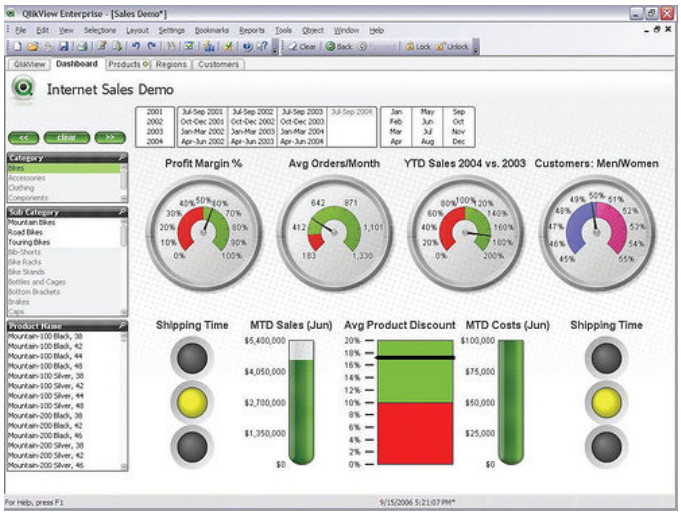
Five Communities that use the CIF

- **Explorers** are the original corporate "out of the box" thinkers. They are individuals who do not look at the life and commerce of the corporation in the standard ways. In some cases these insights are very valuable...
- **Miners** are equipping themselves with tools to find rare and valuable nuggets of information from mountains of data found in data warehouses and data mining data marts. Once settled with appropriate tools, they are patient in their search, determined to find the few insights that can make their organizations rich. They analyze the data to make sure they get meaningful correlations. Miners are thorough. If the data looks odd for some reason, they check it out, since data outliers might indeed be just what they're looking for (finding an instance of fraud, for example).

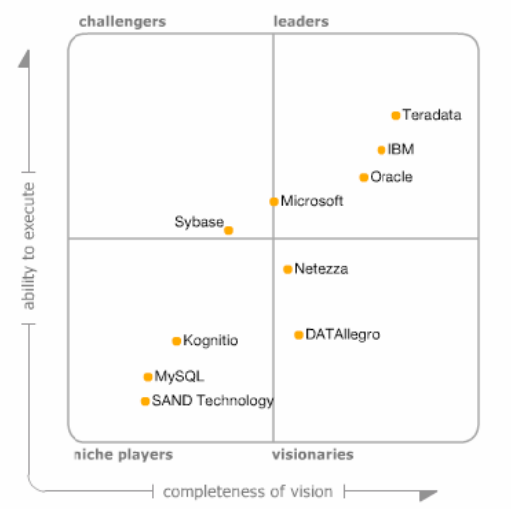
Presentation Topics

For Saturday, January 12

BI Tools and Rules



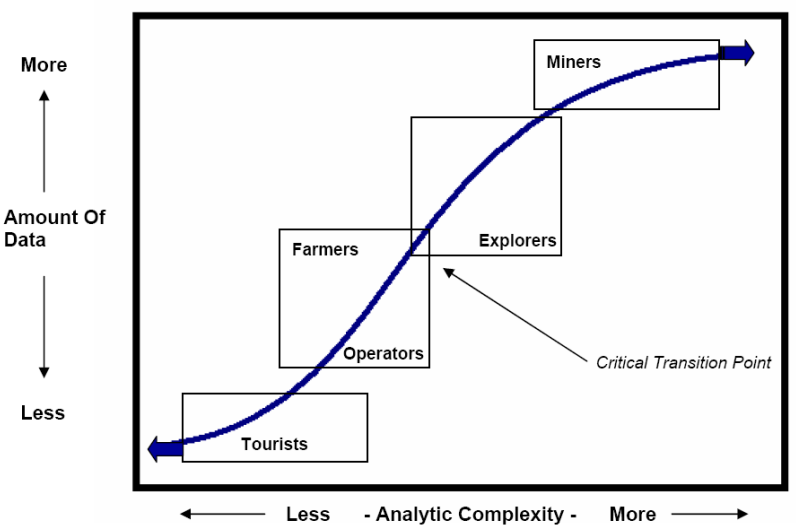
Gartner Magic Quadrants



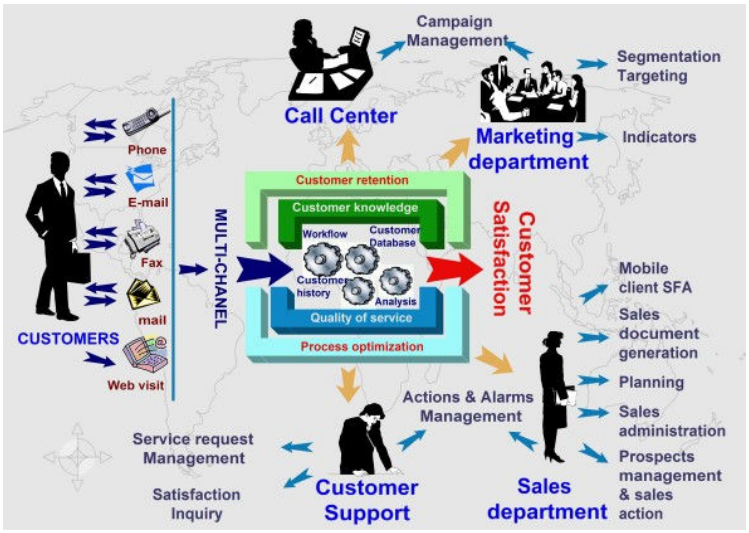
- Gartner's Magic Quadrant for Data Warehouse Database Management Systems
- MySQL is an Open Source Based Organization

Open Source in BI / DM / DW

Mixed Workload Challenge



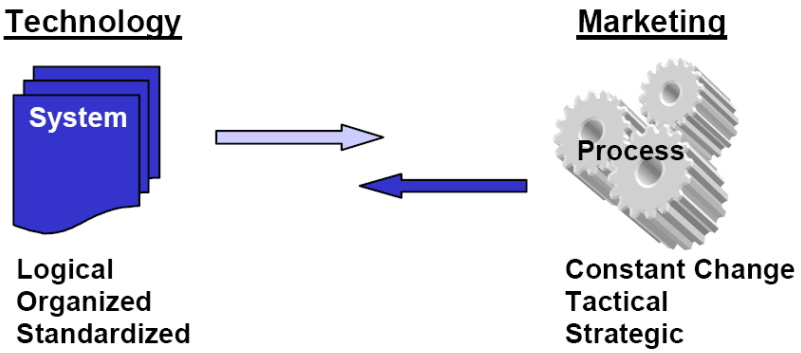
CRM from BI Perspective



Technology-Marketing Misalignment for CRM

- (…) CRM software was seen as a ‘silver bullet’ solution that would change the way companies were organized and run to focus on the customer.
- CRM software systems, when deployed, instead resulted in massive data overload that companies were not prepared to handle. In the worst cases, no attention was paid to helping business users take advantage of the systems, and the software was simply put on the shelf.
- The vaunted ‘closed loop’ promise really became a closed loop breakdown in software-only CRM implementations. (…)

Technology-Marketing Misalignment



Technology-Marketing Misalignment

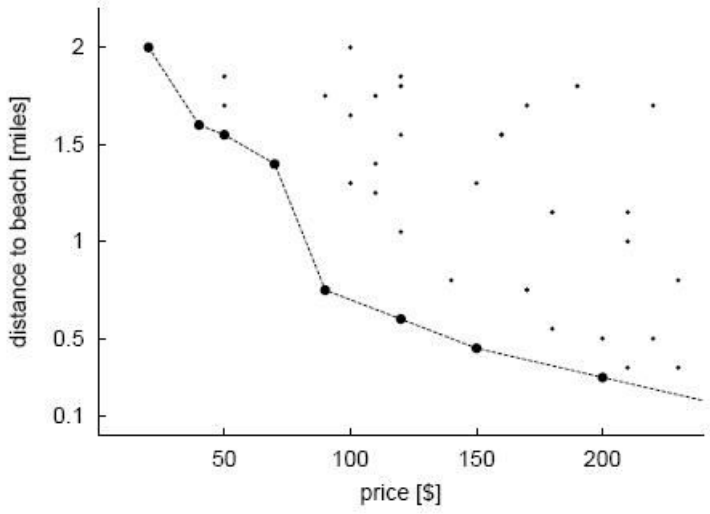
- The questions the IT professional tends to ask:
  - What storage, management, and access systems are most efficient and effective?
  - What tools, programming, and protocols need to be in place to keep the system running (nearly) all the time?
  - What analytic or viewing tools allow the best 'cut' at the data in my system without disrupting everything or slowing the system to a halt?
- The questions marketers want to answer:
  - What is the right balance of marketing spending?
  - What mix of tactics and communication channels deliver the highest return?
  - Why does my customer choose my competitor's product over mine?

Five Communities – Marketing

- Farmers: a sizeable group focused mostly on monitoring and keeping track of the dynamic activities of marketing. This includes: tracking (campaigns, advertising, brand / customer loyalty,) and receiving reports from explorers.
- Tourists: are rare in marketing, as most analysis filters down to the marketing farmers, who then condense reports to the chief executive of marketing.
- Operators: these are the brand and marketing managers, who attempt to track marketing allocations, pay back on campaigns, and success rates of advertising. Their analysis is very current, and if they have a fault it is in not using technology successfully to track, trend, and evaluate short, medium, and long-term results. This aspect of marketing is one of the least developed from a technology and BI perspective.

- Explorers: in marketing, due to the nature of its filtering and creative processes, exploration and discovery is a major part of BI. And the community is extended from internal practitioners to outside consultants, agencies, and specialists. Many of the explorers' roles have evolved over the years into specific, non-technology based activities, such as: focus group testing of new product introductions, or surveys of customer attitudes.
- Miners: as marketing databases have grown, and the introduction of e-commerce has taken hold, miners perform a particularly specialized role in marketing. Their link in the process is segmenting, classifying, and using predictive modeling to assist marketer's efforts. They are also, however, the most out of touch with marketer's daily concerns. Two primary functions they carry are to answer or address marketing explorer's questions and to provide support to direct marketer's targeting.

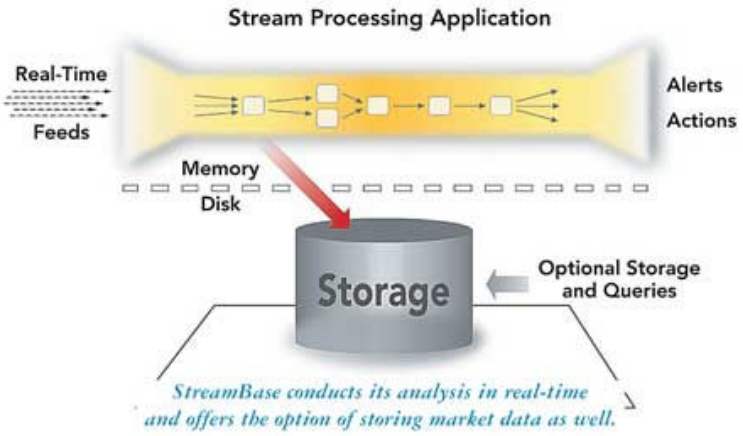
New Queries for Analytics / BI



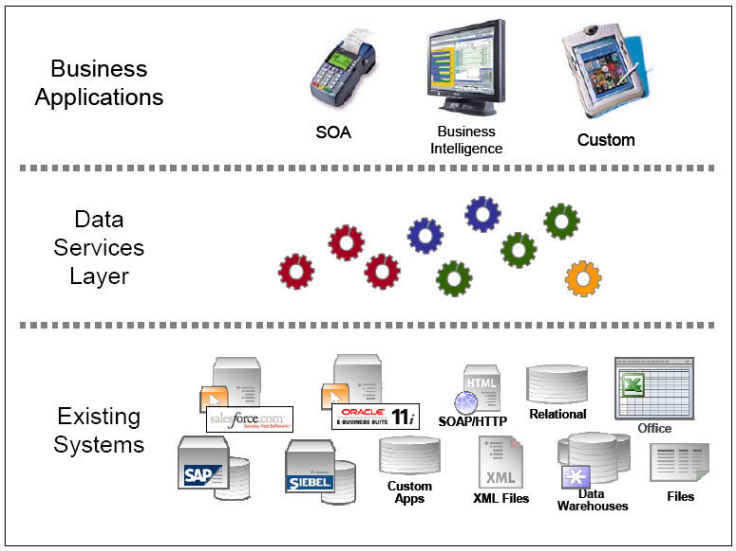
New Queries for Analytics / BI

(...) Forecasting future events based on historic data is useful in many domains like system management, adaptive query processing, environmental monitoring, and financial planning. We describe the Fa system where users and applications can pose declarative forecasting queries – both one-time queries and continuous queries – and get forecasts in real-time along with accuracy estimates. Fa supports efficient algorithms to generate execution plans automatically for forecasting queries from a novel plan space comprising operators for transforming data, learning statistical models from data, and doing inference using the learned models. (...)

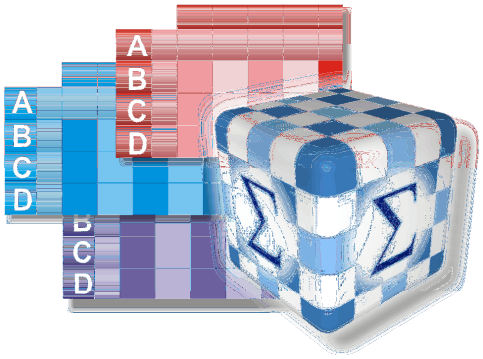
Data Stream Processing



Between DW and BI



OLAP – Extensions and Mining

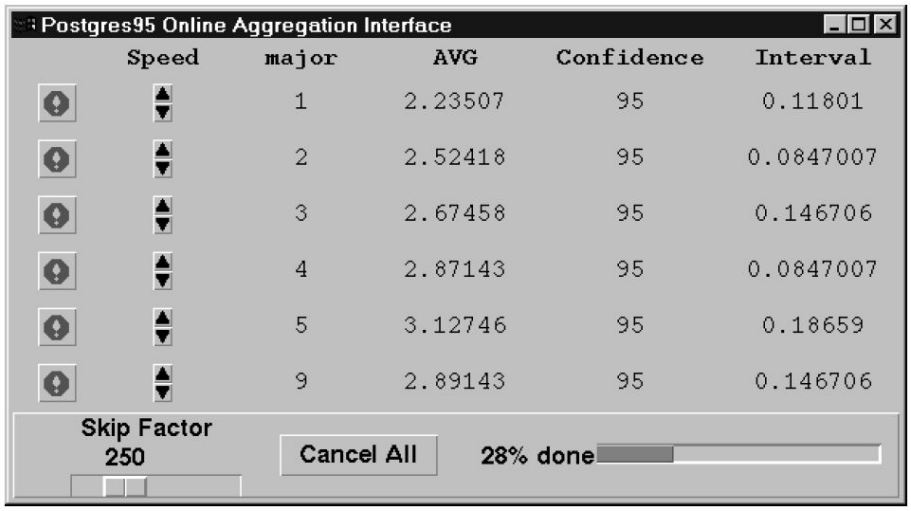


(…) OLAP cannot help marketers distinguish or understand the 'who, what, and why' of their customers beyond simple dimensional representations of historical data (…)

OLAP Rules (E. F. Codd, 1993)

- **Multidimensional conceptual view.** OLAP operates with CUBEs of data that represent multidimensional construct of data. Event though the name implies three dimensional data, the number of possible dimensions is practically unlimited.
- **Transparency.** OLAP systems should be part of an open system that supports heterogeneous data sources.
- **Accessibility.** The OLAP should present the user with a single logical schema of the data.
- **Consistent reporting performance.** Performance should not degrade as the number of dimensions in the model increases.
- **Client/server architecture.** Should be based on open, modular systems.
- **Generic dimensionality.** Not limited to 3-D and not biased toward any particular dimension. A function applied to one dimension should also be able to be applied to another.
- **Dynamic sparse-matrix handling.** Related both to the idea of nulls in relational databases and to the notion of compressing large files, a sparse matrix is one in which not every cell contains data. OLAP systems should accommodate varying storage and data-handling options.
- **Multiuser support.** OLAP systems should support more than one user at the time.
- **Unrestricted cross-dimensional operations.** Similar to rule of generic dimensionality; all dimensions are created equal, and operations across data dimensions should not restrict relationships between cells.
- **Intuitive data manipulation.** Ideally, users shouldn't have to use menus or perform complex multiple-step operations when an intuitive drag-and-drop action will do.
- **Flexible reporting.** Save a tree. Users should be able to print just what they need, and any changes to the underlying financial model should be automatically reflected in reports.
- **Unlimited dimensional and aggregation levels.** The OLAP cube can be built with unlimited dimensions, and aggregation of the contained data also does not have practical limits.

Interaction and Visualization



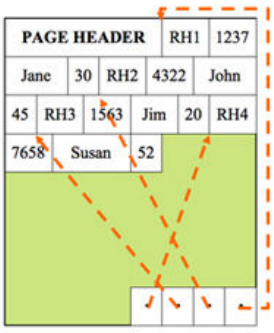
MPP Architectures in DW

Column Oriented Databases

Standard record layout for row stores

Employee Table

RID	SSN	Name	Age
1	1237	Jane	30
2	4322	John	45
3	1563	Jim	20
4	7658	Susan	52
5	2534	Leon	43
6	8791	Dan	37



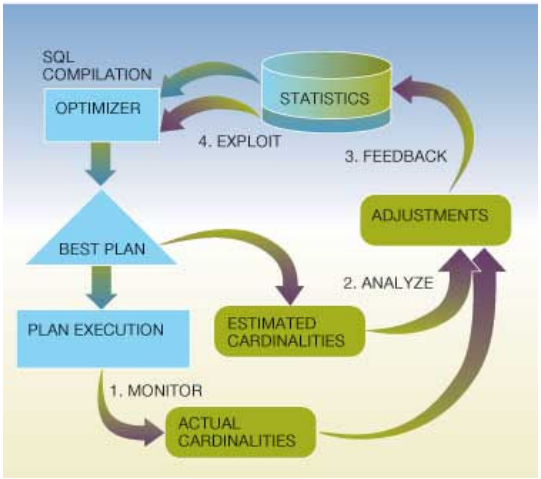
- Records are stored sequentially
- Offsets to start of each record at end of page

Decision Support Benchmarks

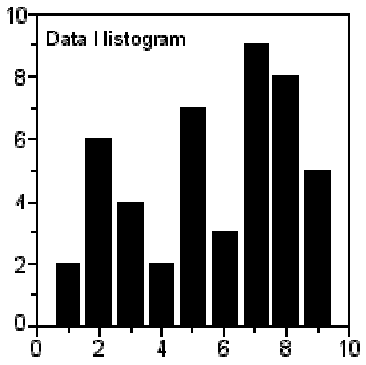
TPC-H → TPC-DS

TPC Transaction Processing Performance Council

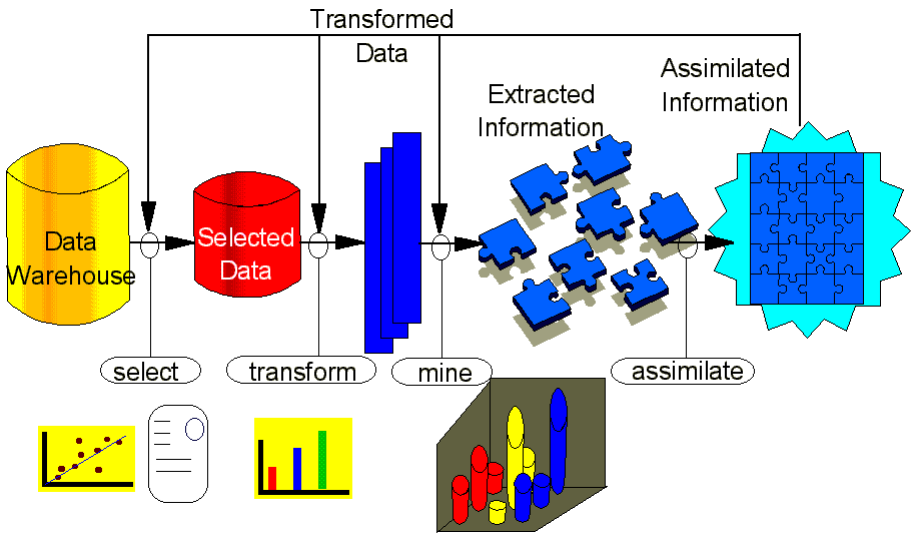
Query Optimization



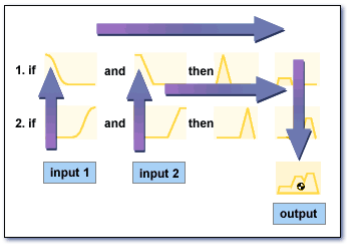
Histograms in DW

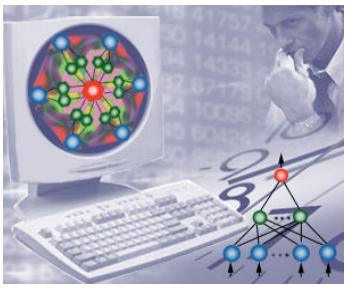


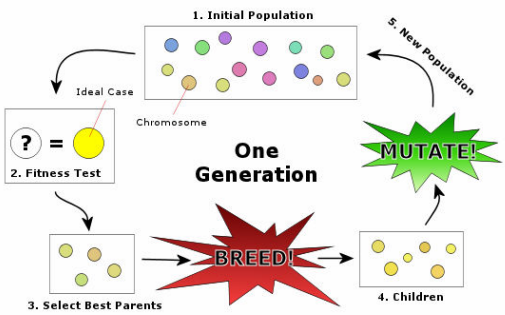
Data Mining for Databases



Artificial Intelligence for BI / DW







DZIĘKUJĘ !!!

slezak@infobright.com