Two-Phase Data Warehouse Optimized for Data Mining

Balázs Rácz
Csaba István Sidló
András Lukács
András A. Benczúr

Data Mining and Web Search Research Group

Computer and Automation Research Institute
Hungarian Academy of Sciences (MTA SZTAKI)

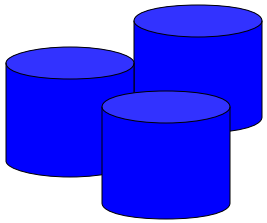
http://datamining.sztaki.hu/

Business Intelligence for the Real Time Enterprise (BIRTE), Seoul, 11. 09. 2006

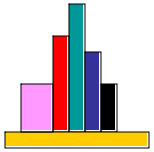
Outline

- Demands and challenges
- Related data warehouse techniques
- The two-phase architecture
- The second phase component
- Data model and data mining framework
- Case studies and measurements

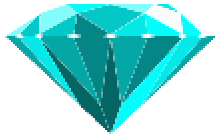
BI and knowledge discovery



data sources



patterns



knowledge

Data storage and management

- long term storage

Data manipulation, statistical queries

Data mining platform

- custom tasks
- quick development/reconfiguration

Basic DBMS functionalities

- long range
- changing
- high dimensional

DM visualization specialized to user needs

Data analysis know-how

- e.g. web/telco usage, churn, user groups, …

Demands in data processing

Long term storage

- growing data volume
- cost of storage

Proper coupling between data warehouses and data mining tools

- optimized data access
- effectively implemented data mining tools

Data mining query language

- help and guide the knowledge discovery process
- flexibility, fast deployment
- code reusing

Related data warehouse techniques

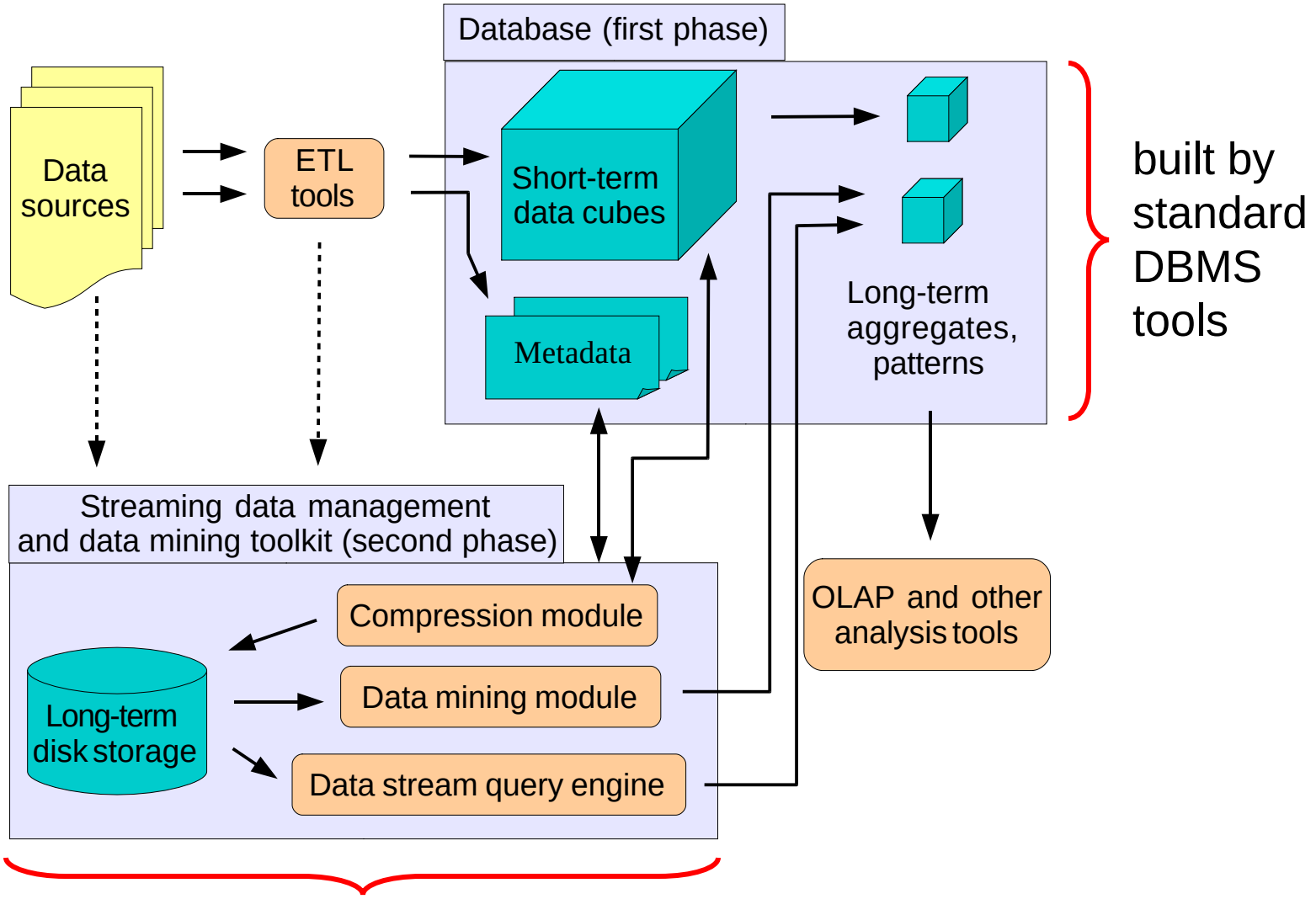
Read/cost optimized databases

- column-oriented databases (C-store), column wise compression
- nearline data warehouses (Sybase IQ, Sand/DNA)

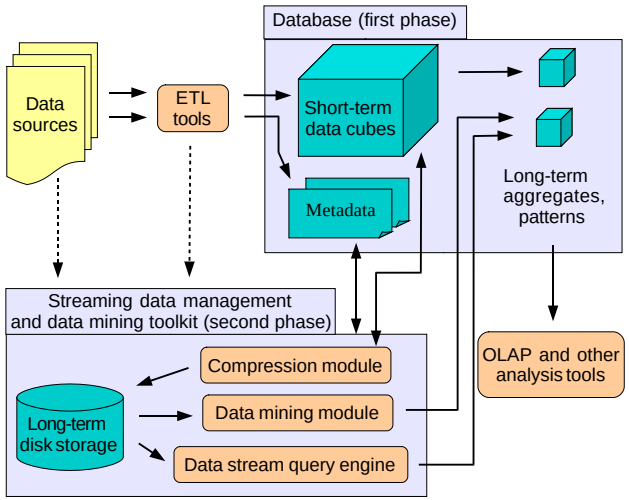
Coupling with data mining systems

- tight – DM is integrated into the DBMS
- semi-tight – interfaced extension of the SQL
- loose – separate DM system

The two-phase architecture



The second phase



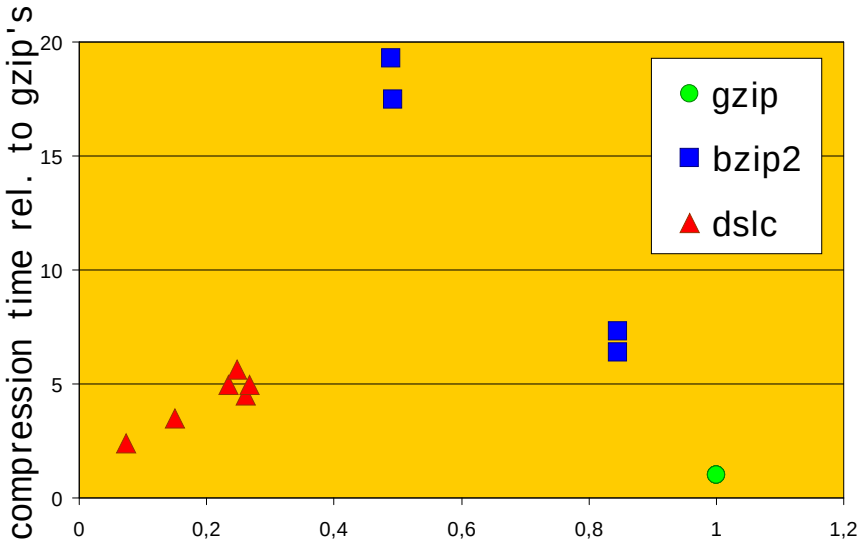
Effective storage

- compression by columns
- large compression rate
- row-wise storage
- fast read-only access
- a new semantic compression method

Data stream approach

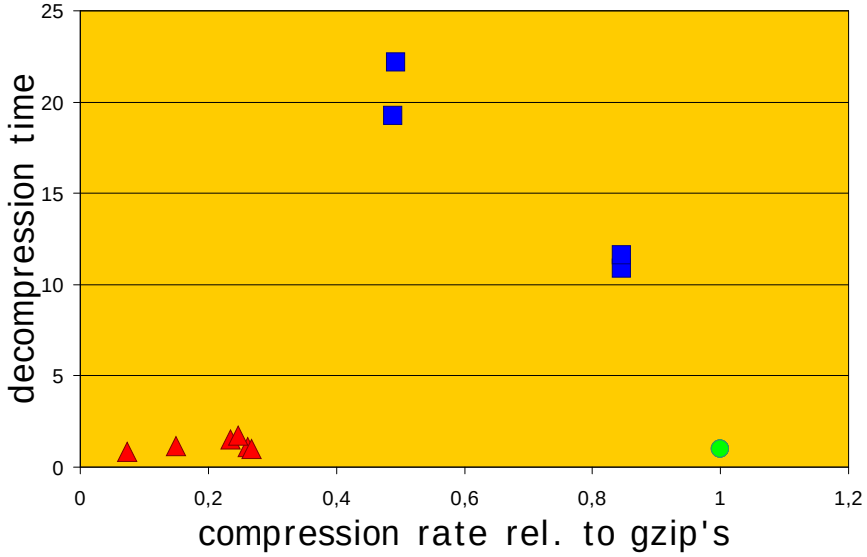
- optimize data mining access to very large data sets
- (block) sequential access, typically full scans of data
- fit for data mining algorithms: frequent itemset mining, partitioning clustering, Bayes classification, decision trees

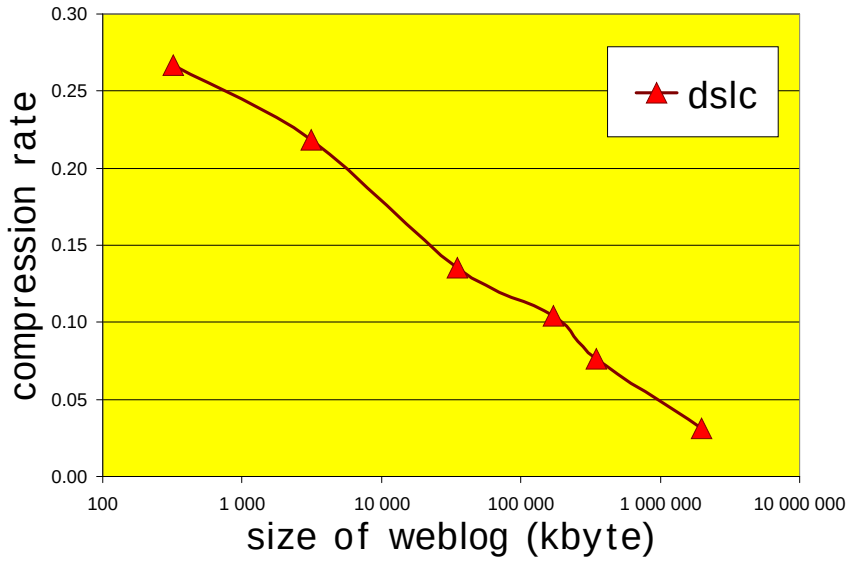
New compression method for log data



Examples of compressions

data type	original size (GB)	compr. size (MB)	compr. rate (%)
weblog	1.89	57.03	2.95%
PIX log	0.92	46.70	4.97%

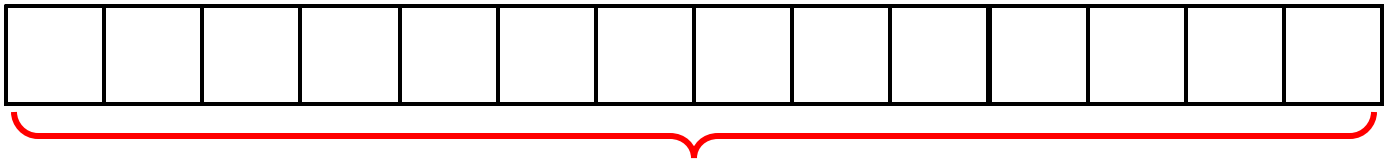




Data models

relational data model

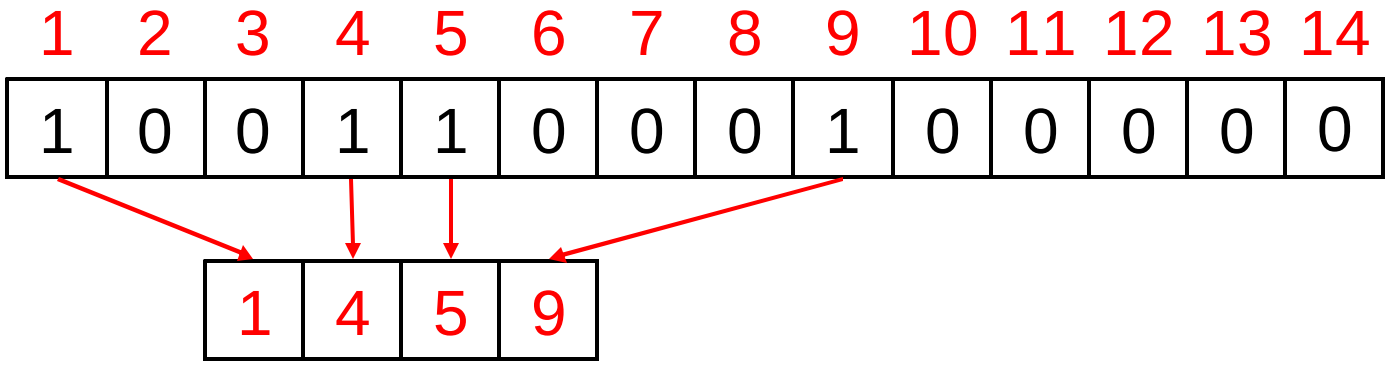
a record



n-tuple

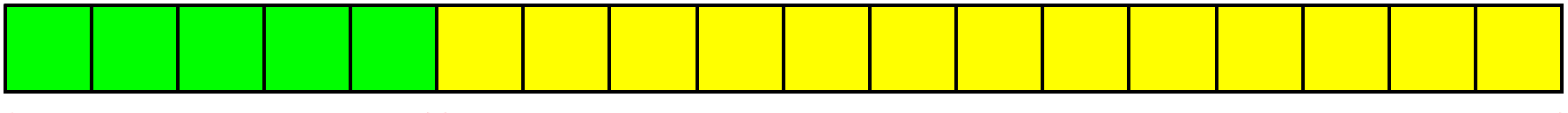
sparse matrix data model

a row of a binary matrix



sparse format of the row

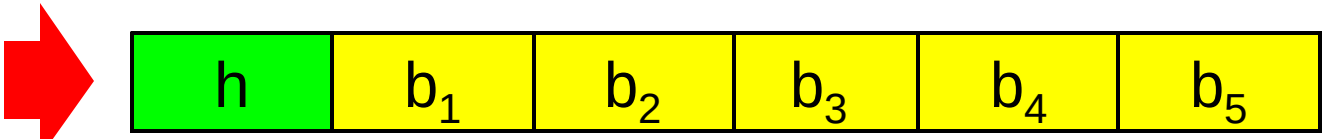
The data model of the second phase



k header

n-k body

h	b_1
h	b_2
h	b_3
h	b_4
h	b_5



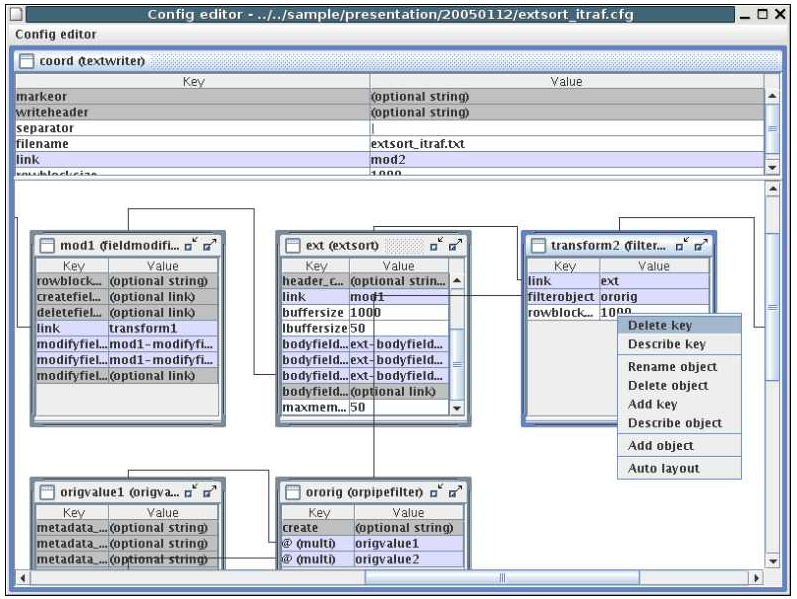
bodies with a same value in header

A common generalization of the *relational* (k=0 or k=n) and the *sparse matrix* (*h* for the rows, *b* for the nonzero columns) data model

DM framework and its query language

Data mining toolkit

- flexible modular architecture
- standardized streaming interface between modules
- pre- and postprocessing, data transforming and mining
- variety of data mining algorithms implemented
- so far 200+ modules (more than 100k lines C++ code)
- new modules can be added

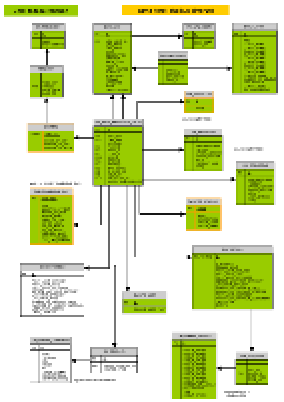


DM query language

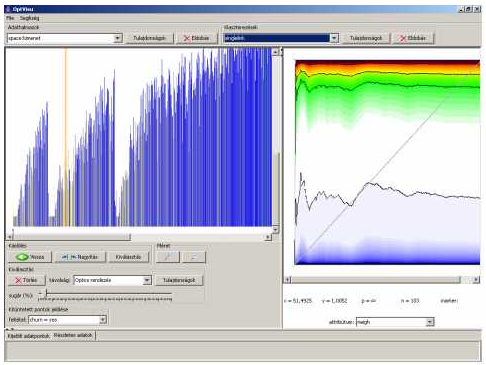
- order and configuration of the necessary modules should be given
- graphical application builder

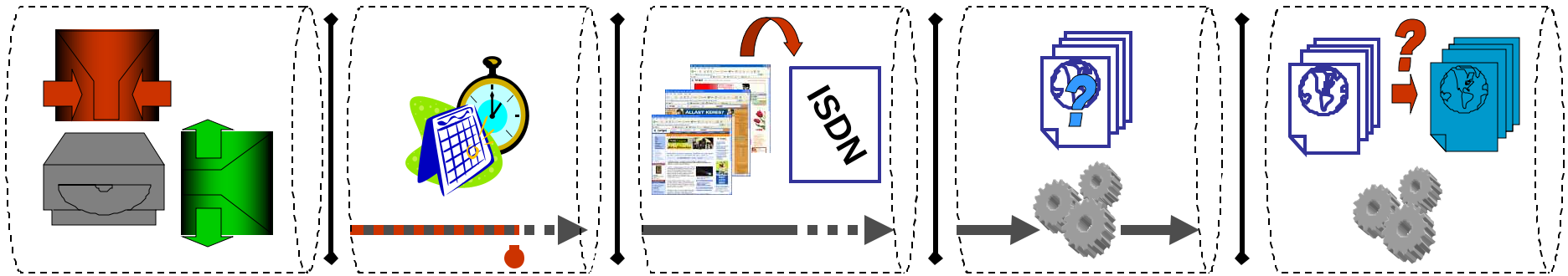
Example of a DM pipeline (telecom. data)

transactional & customer data from DW



results refilled into data tables





compression for internal aggregates

prefix, private, company

users with several numbers, user groups

churn classifier, clustering

cluster info refeed as input

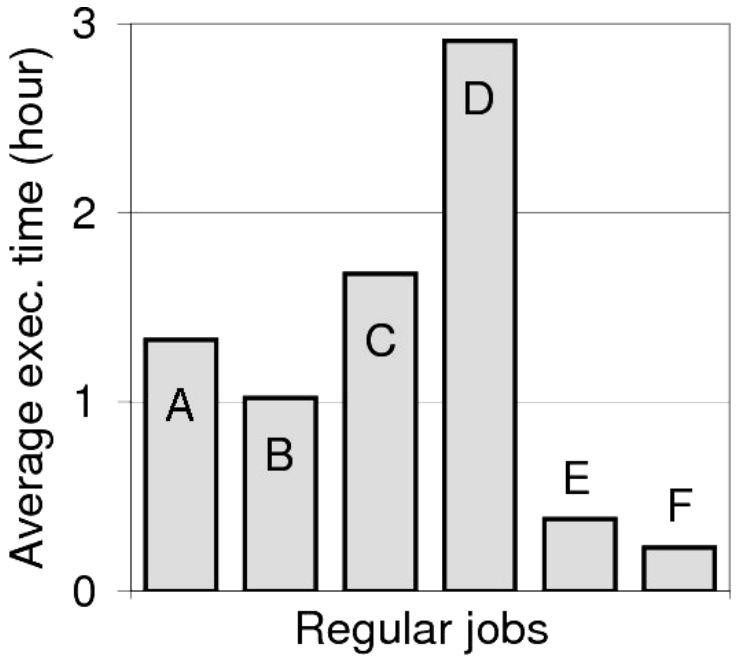
A case study: T-Online Hungary

- Field of operation: online content provider
- Business needs:
  - Long-term log storage with access to analysis
  - Custom reports to management and editors
- State before:
  - Logs kept on tapes, never read back
  - Previous data warehouse projects failed due to data volume
  - MOLAP technology failed on dimensionality
- Data Size: 6.5M HTML hits/day, TB+ log/month

Space requirements of one month web log

Storing method	Size on disk
Compressed (bzip) raw log files	180.7 GB
Compressed (bzip) preprocessed log files	17.1 GB
Standard DB table	44.9 GB
Compressed DB table	39.1 GB
Second-phase compressed storage	1.9 GB

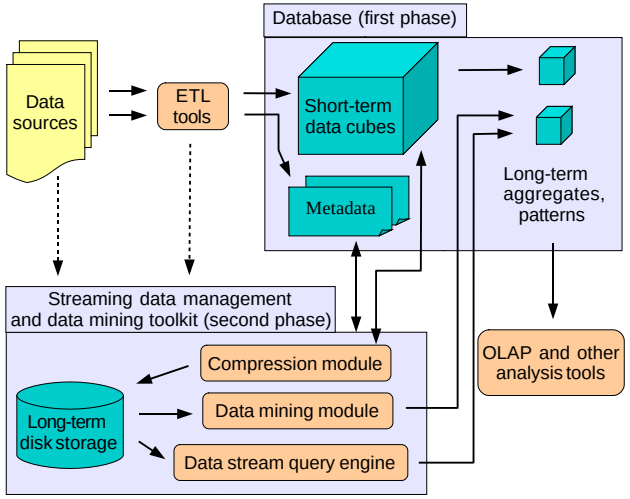
Average execution times of regular jobs



- A - daily preprocessing of data sources
- B - daily dimension and hierarchy update
- C - daily fact table update
- D - computing database aggregates daily
- E - daily compression for the second phase
- F - simple queries against one month data in second phase

Short summary of the two phase

first phase	*second phase*
shorter time storage	longer time storage (archive)
relational data model	stream data model
handling dimension tables	prepared for data mining
built by standard DBMS tools	new compression method and data mining framework



Thanks

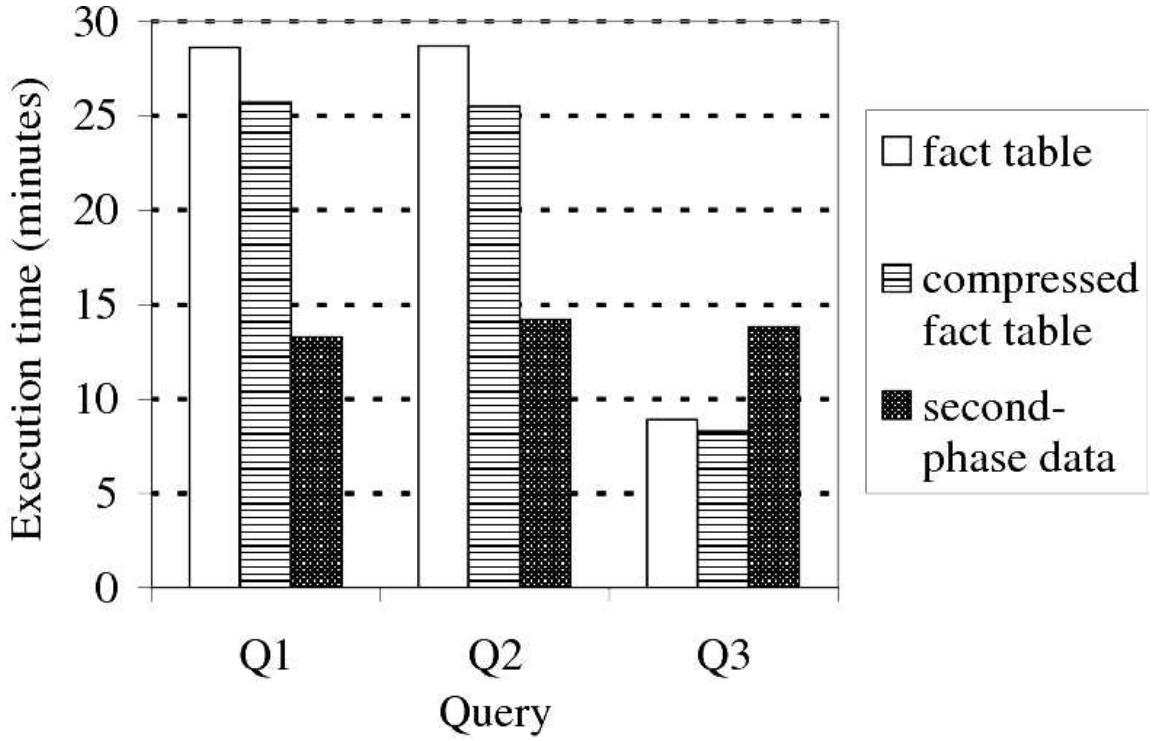
- Katalin Hum and László Lukács
- T-Online Hungary Inc.
- Hungarian National Office for Research and Technology (NKTH)
  - *T-mining* GVOP-3.1.1-2004/-05-0054/3.0
  - *Data Riddle* NKFP-2/0017/2002
- Hungarian Scientific Research Fund (OTKA)
  - T042706
- Inter-University Center for Telecommunications and Informatics, Budapest (ETIK)

Compression of PIX router log

- 10^3-10^5 records/sec
- Goal: preservation for off-line security
- Exemplar PIX log of 100 minutes router activity
  - 5.2 million records
  - 940 MB in its raw form
- Results:
  - 46.7 MB compressed log
  - 4.96% compression ratio
  - 5 min parsing from text +
    2 min compression on a standard P4

Execution times for reference SQL queries

Q1	select sum(PAGE_ID) from FACT_PAGE_IMPRESSION where DATE_KEY between 20060101 and 20060131
Q2	select count(*) from FACT_PAGE_IMPRESSION where DATE_KEY between 20060101 and 20060131 and HTTP_STATUS_CODE = 200
Q3	select count(distinct USER_ID) from FACT_PAGE_IMPRESSION where DATE_KEY between 20060116 and 20060122



Frequent itemset mining (ACCIDENTS dataset)