Data stream نیمسال اول ۹۴/۹/۱۴

data warehousing

Chapter 3: Data Warehousing and OLAP Technology: An Overview

- What is a data warehouse?
- Data warehouse architecture معماری آن
- From data warehousing to data mining

What is Data Warehouse?

- Defined in many different ways, but not rigorously.
  - A decision support database that is maintained separately from the organization's operational database
  - Support information processing by providing a solid platform of consolidated, historical data for analysis.
- "A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision-making process."—W. H. Inmon
- Data warehousing:
  - The process of constructing and using data warehouses

Data Warehouse—Subject-Oriented

- Organized around major subjects, such as customer, product, sales
- Focusing on the modeling and analysis of data for decision makers, not on daily operations or transaction processing
- Provide a simple and concise view around particular subject issues by excluding data that are not useful in the decision support process

Data Warehouse—Integrated

- Constructed by integrating multiple, heterogeneous data sources
  - relational databases, flat files, on-line transaction records
- Data cleaning and data integration techniques are applied.
  - Ensure consistency in naming conventions, encoding structures, attribute measures, etc. among different data sources
    - E.g., Hotel price: currency, tax, breakfast covered, etc.
  - When data is moved to the warehouse, it is converted.

Random Sampling نمونه‌گیری هدف اینکه از داده‌ها نمونه می‌گیریم نمی‌دانیم قرار است
داده‌ها چقدر است. باید دنبال روشی باشیم که اطلاعات حل مسأله از دست ندهیم.
یکی از این روش‌ها Reservoir Sampling است.

Data Warehouse—Time Variant

- The time horizon for the data warehouse is significantly longer than that of operational systems
  - Operational database: current value data
  - Data warehouse data: provide information from a historical perspective (e.g., past 5-10 years)
- Every key structure in the data warehouse
  - Contains an element of time, explicitly or implicitly
  - But the key of operational data may or may not contain "time element"

Data Warehouse—Nonvolatile

- A physically separate store of data transformed from the operational environment
- Operational update of data does not occur in the data warehouse environment
  - Does not require transaction processing, recovery, and concurrency control mechanisms
  - Requires only two operations in data accessing:
    - *initial loading of data* and *access of data*

حافظه استفاده شود که متفاوت است و جواب درستی می توان به این
کوئری ها داد. (ابزار است). می توان کاهش حجم انجام داد.

سه challange اصلی این است که برای هر داده ای باید یک روش بکار برد. یعنی له
شکاف ما با جریان داده است. مثل در امنیت شبکه که مقدار زیادی داده ؟
باید در طول زمان بررسی کنیم که ساده نیست.
روش های استفاده شده تقریبی است و دقت پایین است.

روش های استفاده شده برای حل این مسائل random Sampling است یعنی
داده پیوسته را disortype کنیم. هر بیس عرض برابر کنیم که حجم داده ها کم شود.
هیستوگرام سبک کردن داده ها که بر اساس تعداد رها از آنها را حدس بزنیم.
شاید داده ها را از یک پنجره رد کنیم و جای آنها را threshold کنیم. بخواهیم داده های پرت
و نویز را حذف کنیم. برای سرعت بالاتر از randomize algorithm استفاده می کنیم.

Data Warehouse vs. Heterogeneous DBMS

- Traditional heterogeneous DB integration: A query driven approach
  - Build wrappers/mediators on top of heterogeneous databases
  - When a query is posed to a client site, a meta-dictionary is used to translate the query into queries appropriate for individual heterogeneous sites involved, and the results are integrated into a global answer set
  - Complex information filtering, compete for resources
- Data warehouse: update-driven, high performance
  - Information from heterogeneous sources is integrated in advance and stored in warehouses for direct query and analysis

Data Warehouse vs. Operational DBMS

- OLTP (on-line transaction processing)
  - Major task of traditional relational DBMS
  - Day-to-day operations: purchasing, inventory, banking, manufacturing, payroll, registration, accounting, etc.
- OLAP (on-line analytical processing)
  - Major task of data warehouse system
  - Data analysis and decision making
- Distinct features (OLTP vs. OLAP):
  - User and system orientation: customer vs. market
  - Data contents: current, detailed vs. historical, consolidated
  - Database design: ER + application vs. star + subject
  - View: current, local vs. evolutionary, integrated
  - Access patterns: update vs. read-only but complex queries

معماری DSMS :

یک سری داده داریم که در طول زمان تولید می‌شوند مثل جریان باید هندلش را پیش‌بینی کنیم

آیا کنترل است ، برآورد است ، پیش‌بینی است؟ پیش‌بینی هم چیز مهمی است

بر اساس هدفی که داریم مرتب‌سازی را انجام می‌دهیم چیزهایی را که لازم داریم آن‌ها را

ذخیره کنیم نه همه آن‌ها را (مثل مقدار مشخص)

چند تا challenge برای Stream Data Processing داریم :

داده‌ها یک بار هستند (که رد می‌شوند) داده‌ها پیوسته هستند نمی‌دانیم با چه

سرعتی باید کار کنیم . به دلیل اینکه هستند سریع تولید می‌شوند و رد می‌شوند

کوئری‌ها معمولاً پیوسته است . پیچیده است . کار ما چندبعدی است .

کوئری‌ها از قبل تعریف شده یا (adhoc) باشند یعنی به دلیل زمانی (مثلاً ...)

حداقل جوابی را برآورد می‌کنیم و یک برآورد خوب به دست می‌آوریم . بدین هم دلیل ...

OLTP vs. OLAP

	OLTP	OLAP
users	clerk, IT professional	knowledge worker
function	day to day operations	decision support
DB design	application-oriented	subject-oriented
data	current, up-to-date detailed, flat relational isolated	historical, summarized, multidimensional integrated, consolidated
usage	repetitive	ad-hoc
access	read/write index/hash on prim. key	lots of scans
unit of work	short, simple transaction	complex query
# records accessed	tens	millions
#users	thousands	hundreds
DB size	100MB-GB	100GB-TB
metric	transaction throughput	query throughput, response

Why Separate Data Warehouse?

- High performance for both systems
  - DBMS— tuned for OLTP: access methods, indexing, concurrency control, recovery
  - Warehouse—tuned for OLAP: complex OLAP queries, multidimensional view, consolidation
- Different functions and different data:
  - missing data: Decision support requires historical data which operational DBs do not typically maintain
  - data consolidation: DS requires consolidation (aggregation, summarization) of data from heterogeneous sources
  - data quality: different sources typically use inconsistent data representations, codes and formats which have to be reconciled
- Note: There are more and more systems which perform OLAP analysis directly on relational databases

فرق DBMS با DSMS (Data stream management system

کار با DSMS سخت‌تر از DBMS است. کار کردن با داده‌هایی که به زمان وابسته است دشوارتر است.

random access : داده‌های که در DBMS هستند بصورت تصادفی جمع آوری شده‌اند و ما قبلاً روی آنها عرضه برداشت می‌کنیم اما داده‌های DSMS چون به زمان وابسته است بصورت دنباله است و هم چنان نظم برای دسترسی به آنها باید تلاش کرد

در DBMS با داده‌های جاری کار می‌کنیم اما در DSMS داده‌های تاریخی نقش مهمی دارند. در DBMS نیازی به Real time بودن نداریم اما DSMS real time است. update کردن در DBMS نیازی نیست اما در DSMS حجم داده‌ها زیاد است که مرتب زیاد می‌شوند. در DBMS داده‌ها پایدار است اما در DSMS نه

در DBMS داده‌ها دقیق است اما در DSMS دقیق نیست

Chapter 3: Data Warehousing and OLAP Technology: An Overview

- What is a data warehouse?
- Data warehouse architecture
- From data warehousing to data mining

Design of Data Warehouse: A Business Analysis Framework

- Four views regarding the design of a data warehouse
  - Top-down view
    - allows selection of the relevant information necessary for the data warehouse
  - Data source view
    - exposes the information being captured, stored, and managed by operational systems
  - Data warehouse view
    - consists of fact tables and dimension tables
  - Business query view
    - sees the perspectives of data in the warehouse from the view of end-user

DBMS سیستم داده‌های ثبت و ضبط شده،

تفاوت اصلی جریان داده و dbms هم است

مشخصه‌های جریان داده‌ها پیوسته است مثل مقدار سنجی که لحظه به لحظه گزارش

می‌دهند معمولاً بصورت real time تولید می‌شوند یا براساس مثل داده‌های نرخ بیکاری

در ماههای مختلف است.

ممکن است آموزشی برای مرتب‌سازی مقادیر در نظر بگیریم

مثلاً 5 داده در مورد میزان برق مصرفی جمع آوری می‌کنیم و خود داده‌ها را

ذخیره نمی‌کنیم و مقادیر خلاصه شده مثل متوسط آنها را ذخیره می‌کنیم.

مشکل اینست داده‌ها حاوی بصورت 1 متغیره نباشند multiple چند متغیره

باشند. مثل میزان برق مصرفی براساس جنسیت بودن تا [illegible] خانواده اندازه‌گیری

کردم که جمع آن مصرف برق [illegible] از [illegible] تشکیل شده (چند متغیره)

چرا استفاده‌ها [illegible] دارد؟ مدت زمان مکالمات مصرف کارت بانکی ترافیک شبکه

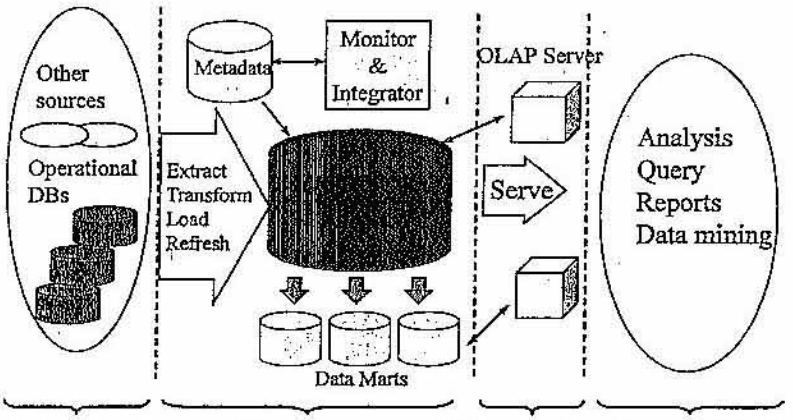
قیمت لحظه‌ای مصرف برق [illegible] مانیتورینگ در [illegible] هدف برای این است که شناسایی

روابط و الگوهای حجم داده‌ها در [illegible] داده‌های حجیم [illegible]

Data Warehouse Design Process

- Top-down, bottom-up approaches or a combination of both
  - Top-down: Starts with overall design and planning (mature)
  - Bottom-up: Starts with experiments and prototypes (rapid)
- From software engineering point of view
  - Waterfall: structured and systematic analysis at each step before proceeding to the next
  - Spiral: rapid generation of increasingly functional systems, short turn around time, quick turn around
- Typical data warehouse design process
  - Choose a business process to model, e.g., orders, invoices, etc.
  - Choose the *grain* (*atomic level of data*) of the business process
  - Choose the dimensions that will apply to each fact table record
  - Choose the measure that will populate each fact table record

Data Warehouse: A Multi-Tiered Architecture

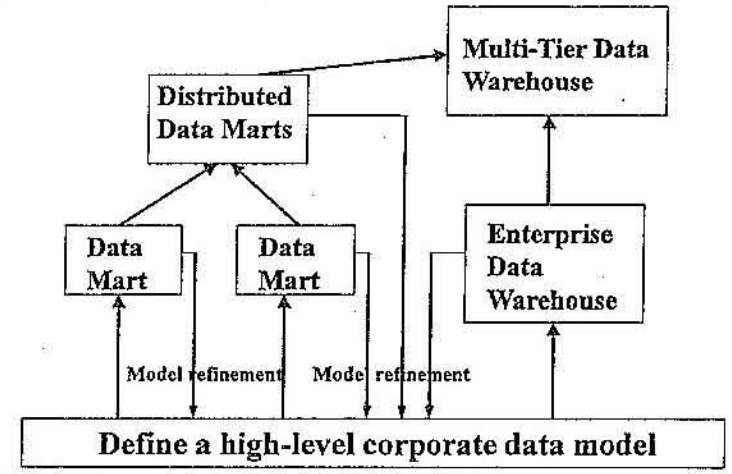


(سایتی است)

Three Data Warehouse Models

- Enterprise warehouse
  - collects all of the information about subjects spanning the entire organization
- Data Mart
  - a subset of corporate-wide data that is of value to a specific groups of users. Its scope is confined to specific, selected groups, such as marketing data mart
    - Independent vs. dependent (directly from warehouse) data mart
- Virtual warehouse
  - A set of views over operational databases
  - Only some of the possible summary views may be materialized

Data Warehouse Development: A Recommended Approach



جریان داده‌ها : مثل محاسبه نرخ بیکاری کشور، محاسبه جریان آنی اطلاعات مقدار بیکاری کشور در روز، ماه، هفته می‌دهد.

جریان داده‌ها می‌تواند بصورت پیوسته تولید شوند مثل مصرف برق هر خانه‌ای که میزان آنرا اندازه‌گیری کرد

در داده‌های stream دنبال changePoint detection هستیم. (نقاط تغییر)

داده‌های stream خیلی سریع تولید می‌شوند و حجم آنها زیاد است

و سرعت پردازش نیز می‌تواند خیلی سریع تولید می‌شوند

Data Warehouse Back-End Tools and Utilities

- Data extraction
  - get data from multiple, heterogeneous, and external sources
- Data cleaning
  - detect errors in the data and rectify them when possible
- Data transformation
  - convert data from legacy or host format to warehouse format
- Load
  - sort, summarize, consolidate, compute views, check integrity, and build indicies and partitions
- Refresh
  - propagate the updates from the data sources to the warehouse

Metadata Repository

- Meta data is the data defining warehouse objects. It stores:
- Description of the structure of the data warehouse
  - schema, view, dimensions, hierarchies, derived data defn, data mart locations and contents
- Operational meta-data
  - data lineage (history of migrated data and transformation path), currency of data (active, archived, or purged), monitoring information (warehouse usage statistics, error reports, audit trails)
- The algorithms used for summarization
- The mapping from operational environment to the data warehouse
- Data related to system performance
  - warehouse schema, view and derived data definitions
- Business data
  - business terms and definitions, ownership of data, charging policies

داده‌های طولی Longitudinal Data یا stream است. داده‌های پزشکی جمع‌آوری می‌شوند مثل فشار خون یک بیمار بصورت روزانه. مشخصه بارز داده‌های طولی داده‌های stream سری‌های زمانی این است که این داده‌ها مستقل از هم نیستند استقلال فرض اساسی در تحلیل آماری است. مثل داده‌هایی که در طول تولید، تولید می‌شود مثل وزن متوالی رب. داده‌هایی که در طول زمان تولید می‌شوند و به هم وابسته هستند

داده‌هایی که زمان برایشان مهم است Sequential Data می‌گویند.

اما درجه حرارت شب و روز با هم متفاوت است چون زمان مهم است.

گاهی بعد زمان از یک بعد بیشتر باشد مثلاً که با ماشین کار می‌کردیم ممکن بود چند بعدی باشد مثل قد وزن یک آدم، یا اندازه‌گیری ۱۰۰۰۰ ژن (یک مشاهده با ۱۰۰۰ متغیر)

اگر در مکان و زمان مختلف داده‌ای اندازه‌گیری شود به آن Spatial-Temporal می‌گویند در این حالت به عنوان ابعاد مثل ژن با ماشین برخورد می‌کنیم یک زمانی داریم که شده سه بعدی یک بعد زمان دو بعد دیگر مکان است. وقتی زمان از یک به دو بعد تبدیل پیدا می‌کند فرمول‌ها عوض می‌شود. در ابعاد بالا مثل آدم و ۱۰۰۰۰ ژن، رفتار فرمول‌ها هم عوض می‌شود.

OLAP Server Architectures

- Relational OLAP (ROLAP)
  - Use relational or extended-relational DBMS to store and manage warehouse data and OLAP middle ware
  - Include optimization of DBMS backend, implementation of aggregation navigation logic, and additional tools and services
  - Greater scalability
- Multidimensional OLAP (MOLAP)
  - Sparse array-based multidimensional storage engine
  - Fast indexing to pre-computed summarized data
- Hybrid OLAP (HOLAP) (e.g., Microsoft SQLServer)
  - Flexibility, e.g., low level: relational, high-level: array
- Specialized SQL servers (e.g., Redbricks)
  - Specialized support for SQL queries over star/snowflake schemas

Chapter 3: Data Warehousing and OLAP Technology: An Overview

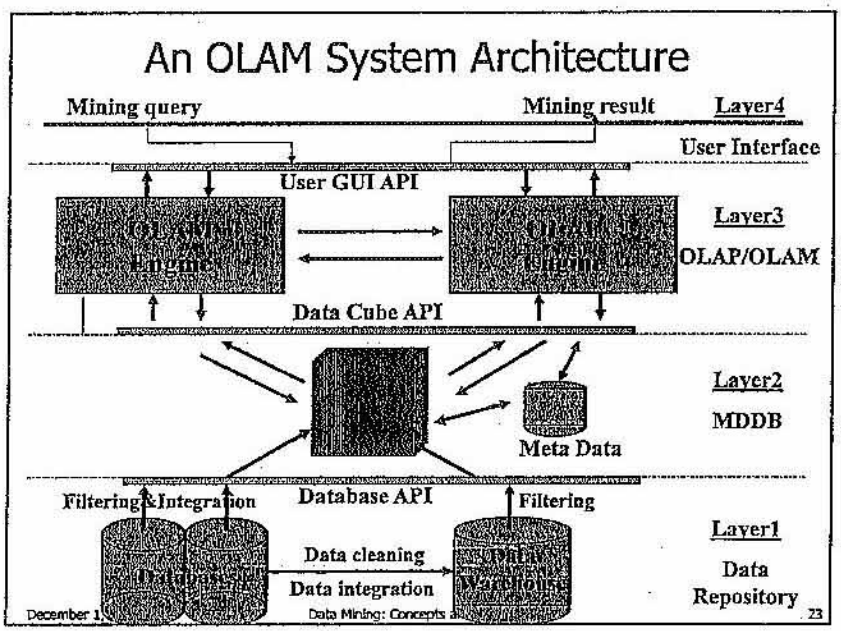
- What is a data warehouse?
- Data warehouse architecture
- From data warehousing to data mining

Data Warehouse Usage

- Three kinds of data warehouse applications
  - Information processing
    - supports querying, basic statistical analysis, and reporting using crosstabs, tables, charts and graphs
  - Analytical processing
    - multidimensional analysis of data warehouse data
    - supports basic OLAP operations, slice-dice, drilling, pivoting
  - Data mining
    - knowledge discovery from hidden patterns
    - supports associations, constructing analytical models, performing classification and prediction, and presenting the mining results using visualization tools

From On-Line Analytical Processing (OLAP) to On Line Analytical Mining (OLAM)

- Why online analytical mining?
  - High quality of data in data warehouses
    - DW contains integrated, consistent, cleaned data
  - Available information processing structure surrounding data warehouses
    - ODBC, OLEDB, Web accessing, service facilities, reporting and OLAP tools
  - OLAP-based exploratory data analysis
    - Mining with drilling, dicing, pivoting, etc.
  - On-line selection of data mining functions
    - Integration and swapping of multiple mining functions, algorithms, and tasks



Chapter 3: Data Warehousing and OLAP Technology: An Overview

- What is a data warehouse?
- A multi-dimensional data model
- Data warehouse architecture
- Data warehouse implementation
- From data warehousing to data mining
- Summary

Summary: Data Warehouse and OLAP Technology

- Why data warehousing?
- Data warehouse architecture
- From OLAP to OLAM (on-line analytical mining)

References (I)

- S. Agarwal, R. Agrawal, P. M. Deshpande, A. Gupta, J. F. Naughton, R. Ramakrishnan, and S. Sarawagi. On the computation of multidimensional aggregates. VLDB'96
- D. Agrawal, A. E. Abbadi, A. Singh, and T. Yurek. Efficient view maintenance in data warehouses. SIGMOD'97
- R. Agrawal, A. Gupta, and S. Sarawagi. Modeling multidimensional databases. ICDE'97
- S. Chaudhuri and U. Dayal. An overview of data warehousing and OLAP technology. ACM SIGMOD Record, 26:65-74, 1997
- E. F. Codd, S. B. Codd, and C. T. Salley. Beyond decision support. Computer World, 27, July 1993.
- J. Gray, et al. Data cube: A relational aggregation operator generalizing group-by, cross-tab and sub-totals. Data Mining and Knowledge Discovery, 1:29-54, 1997.
- A. Gupta and I. S. Mumick. Materialized Views: Techniques, Implementations, and Applications. MIT Press, 1999.
- J. Han. Towards on-line analytical mining in large databases. *ACM SIGMOD Record*, 27:97-107, 1998.
- V. Harinarayan, A. Rajaraman, and J. D. Ullman. Implementing data cubes efficiently. SIGMOD'96

References (II)

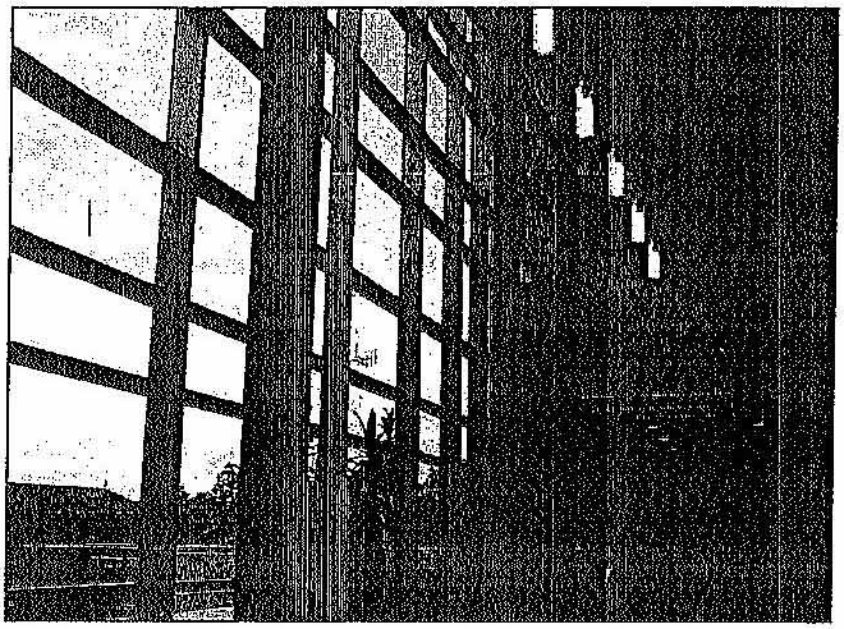
- C. Imhoff, N. Galemmo, and J. G. Geiger. Mastering Data Warehouse Design: Relational and Dimensional Techniques. John Wiley, 2003
- W. H. Inmon. Building the Data Warehouse. John Wiley, 1996
- R. Kimball and M. Ross. The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling. 2ed. John Wiley, 2002
- P. O'Neil and D. Quass. Improved query performance with variant indexes. SIGMOD'97
- Microsoft. OLEDB for OLAP programmer's reference version 1.0. In http://www.microsoft.com/data/oledb/olap, 1998
- A. Shoshani. OLAP and statistical databases: Similarities and differences. PODS'00.
- S. Sarawagi and M. Stonebraker. Efficient organization of large multidimensional arrays. ICDE'94
- OLAP council. MDAPI specification version 2.0. In http://www.olapcouncil.org/research/apily.htm, 1998
- E. Thomsen. OLAP Solutions: Building Multidimensional Information Systems. John Wiley, 1997
- P. Valduriez. Join indices. ACM Trans. Database Systems, 12:218-246, 1987.
- J. Widom. Research problems in data warehousing. CIKM'95.

Data Mining: Concepts and Techniques

— Chapter 8 —

8.1. Mining data streams

Jiawei Han and Micheline Kamber
Department of Computer Science
University of Illinois at Urbana-Champaign
www.cs.uiuc.edu/~hanj

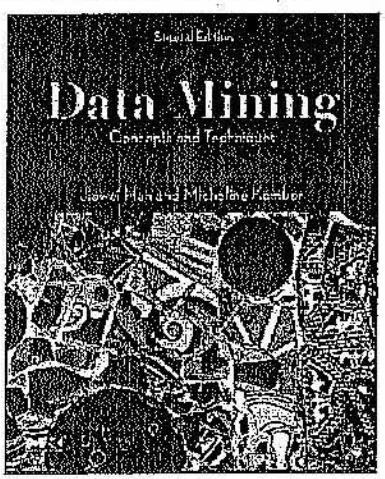


Data and Information Systems (DAIS:) Course Structures at CS/UIUC

- Three streams: Database, data mining and text information systems
- Database Systems:
  - Database mgmt systems (CS411: Fall and Spring)
  - Advanced database systems (CS511: Fall)
  - Web information systems (Kevin Chang)
  - Information integration (An-Hai Doan)
- Data mining
  - Intro. to data mining (CS412: Han—Fall)
  - Data mining: Principles and algorithms (CS512: Han—Spring)
  - Seminar: Advanced Topics in Data mining (CS591Han—Fall and Spring)
- Text information systems and Bioinformatics
  - Text information system (CS410Zhai)
  - Introduction to BioInformatics (CS598Sinha, CS498Zhai)

Data Mining: Concepts and Techniques, 2ed. 2006

- Seven chapters (Chapters 1-7) are covered in the Fall semester
- Four chapters (Chapters 8-11) are covered in the Spring semester



Coverage of CS412@UIUC (Intro. to Data Warehousing and Data Mining)

1. Introduction
2. Data Preprocessing
3. Data Warehouse and OLAP Technology: An Introduction
4. Advanced Data Cube Technology and Data Generalization
5. Mining Frequent Patterns, Association and Correlations
6. Classification and Prediction
7. Cluster Analysis

Coverage of CS512@UIUC (Data Mining: Principles and Algorithms)

8. Mining stream, time-series, and sequence data
   - Mining data streams
   - Mining time-series data
   - Mining sequence patterns in transactional databases
   - Mining sequence patterns in biological data
9. Graph mining, social network analysis, and multi-relational data mining
   - Graph mining
   - Social network analysis
   - Multi-relational data mining
10. Mining Object, Spatial, Multimedia, Text and Web data
    - Mining object data
    - Spatial and spatiotemporal data mining
    - Multimedia data mining
    - Text mining
    - Web mining
11. Applications and trends of data mining
    - Data mining applications
    - Data mining products and research prototypes
    - Additional themes on data mining
    - Social impacts of data mining
    - Trends in data mining

Chapter 8. Mining Stream, Time-Series, and Sequence Data

- **Mining data streams**
- **Mining time-series data**
- **Mining sequence patterns in transactional databases**
- **Mining sequence patterns in biological data**

Mining Data Streams

- What is stream data? Why Stream Data Systems?
- Stream data management systems: Issues and solutions
- Stream data cube and multidimensional OLAP analysis
- Stream frequent pattern analysis
- Stream classification
- Stream cluster analysis
- Research issues

ما یک سری داده خام داریم که باید اطلاعات را از آنها بکشیم. بعد دیتا اطلاعاتی است

که براساس هدفی که می خواهیم انجام دهیم در دسترس ما هستند یعنی یک سری اطلاعات

که می توانند بصورت اطلاعات پیشین باشد که براساس آن بیاییم داده های مورد نیاز را

تلفیق دهیم در DataWarehouse ارسال دهیم.

OLAP مرحله بعد Data Storage است و یک سمت نرم افزاری است مثل

Sql Server . بعد از آن تحلیل داده کاوی را انجام می دهیم.

Characteristics of Data Streams

- Data Streams
  - Data streams—continuous, ordered, changing, fast, huge amount
  - Traditional DBMS—data stored in finite, persistent data sets
- Characteristics
  - Huge volumes of continuous data, possibly infinite
  - Fast changing and requires fast, real-time response
  - Data stream captures nicely our data processing needs of today
  - Random access is expensive—single scan algorithm (*can only have one look*)
  - Store only the summary of the data seen thus far
  - Most stream data are at pretty low-level or multi-dimensional in nature, needs multi-level and multi-dimensional processing

Stream Data Applications

- Telecommunication calling records
- Business: credit card transaction flows
- Network monitoring and traffic engineering
- Financial market: stock exchange
- Engineering & industrial processes: power supply & manufacturing
- Sensor, monitoring & surveillance: video streams, RFIDs
- Security monitoring
- Web logs and Web page click streams
- Massive data sets (even saved but random access is too expensive)

برای انجام سفارش Data Warehouse با یک شرکت مذاکره کردم.

DBMS versus DSMS

- Persistent relations
- One-time queries
- Random access
- "Unbounded" disk store
- Only current state matters
- No real-time services
- Relatively low update rate
- Data at any granularity
- Assume precise data
- Access plan determined by query processor, physical DB design

- Transient streams
- Continuous queries
- Sequential access
- Bounded main memory
- Historical data is important
- Real-time requirements
- Possibly multi-GB arrival rate
- Data at fine granularity
- Data stale/imprecise
- Unpredictable/variable data arrival and characteristics

Ack. From Motwani's PODS tutorial slides

Mining Data Streams

- What is stream data? Why Stream Data Systems?
- Stream data management systems: Issues and solutions
- Stream data cube and multidimensional OLAP analysis
- Stream frequent pattern analysis
- Stream classification
- Stream cluster analysis
- Research issues

OLTP برای مشتری است OLAP برای مارکت است

OLTP داده‌ی در جریان است با جزئیات. OLAP بر اساس داده‌ی تاریخی است.

برای هر کدام از سیستم‌ها یک کاربری با امکانی می‌خواهیم برای دسترسی؛ اندیس گذاری، بر حسب کنترل بازیابی. معماری ما باید درست باشد. مثل مقایسه سازی Yahoo با google

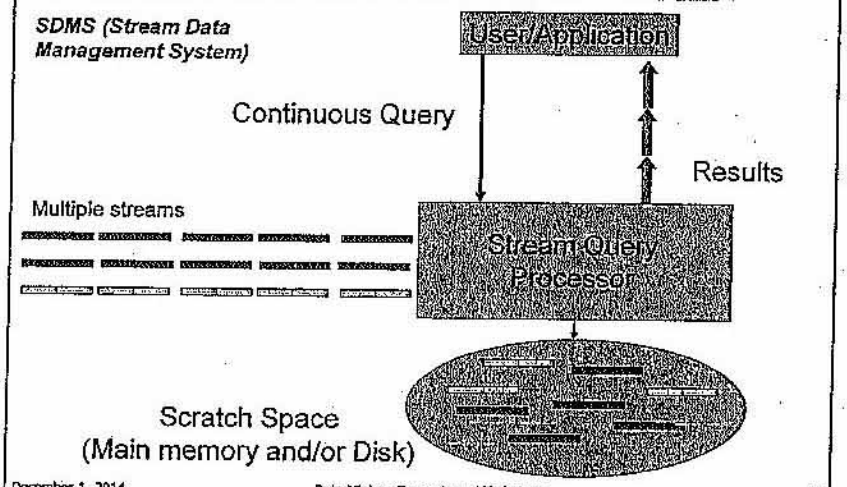
ما برای عمل روی داده‌های مختلف بکار می‌بریم برای تشکیل Datawarehouse.

missing Data داده‌ی گم شده مستقل پاسخ ندادن به یک سؤال در پرسش نامه

با EM الگوریتم می‌توانیم داده‌های گم شده را impute کنیم.

کیفیت داده‌ها مهم است اینکه چطور جمع آوری شدند بصورت درست یا ساختگی یا تقلبی یا اینکه با دقت جمع آوری شده.

Architecture: Stream Query Processing



Challenges of Stream Data Processing

- Multiple, continuous, rapid, time-varying, ordered streams
- Main memory computations
- Queries are often continuous
  - Evaluated continuously as stream data arrives
  - Answer updated over time
- Queries are often complex
  - Beyond element-at-a-time processing
  - Beyond stream-at-a-time processing
  - Beyond relational queries (scientific, data mining, OLAP)
- Multi-level/multi-dimensional processing and data mining
  - Most stream data are at low-level or multi-dimensional in nature

فرق Datawarehouse با DBMS :

oltp جاهایی استفاده می‌شود که با Transaction کار داریم مثل عملیات بانکی، خرید
و... یک ارتباط مستقیم با DBMS دارد

olap : از داده‌ها استفاده می‌کنیم بتوانیم آمار را تحلیل کنیم و بعد از آن تصمیم‌گیری را انجام دهیم

هدف اصلی Datawarehouse ← olap است.

oltp

Processing Stream Queries

- Query types
  - One-time query vs. continuous query (being evaluated continuously as stream continues to arrive)
  - Predefined query vs. ad-hoc query (issued on-line)
- Unbounded memory requirements
  - For real-time response, main memory algorithm should be used
  - Memory requirement is unbounded if one will join future tuples
- Approximate query answering
  - With bounded memory, it is not always possible to produce exact answers
  - High-quality approximate answers are desired
  - Data reduction and synopsis construction methods
    - Sketches, random sampling, histograms, wavelets, etc.

Methodologies for Stream Data Processing

- Major challenges
  - Keep track of a large universe, e.g., pairs of IP address, not ages
- Methodology
  - Synopses (trade-off between accuracy and storage)
  - Use *synopsis data structure*, much smaller ($O(\log^k N)$ space) than their base data set ($O(N)$ space)
  - Compute an *approximate answer* within a *small error range* (factor ε of the actual answer)
- Major methods
  - Random sampling
  - Histograms
  - Sliding windows
  - Multi-resolution model
  - Sketches
  - Radomized algorithms

Stream Data Processing Methods (1)

- Random sampling (but without knowing the total length in advance)
  - *Reservoir sampling*: maintain a set of *s* candidates in the reservoir, which form a true random sample of the element seen so far in the stream. As the data stream flow, every new element has a certain probability (*s*/N) of replacing an old element in the reservoir.
- Sliding windows
  - Make decisions based only on *recent data* of sliding window size *w*
  - An element arriving at time *t* expires at time *t* + *w*
- Histograms
  - Approximate the frequency distribution of element values in a stream
  - Partition data into a set of contiguous buckets
  - Equal-width (equal value range for buckets) vs. V-optimal (minimizing frequency variance within each bucket)
- Multi-resolution models
  - Popular models: balanced binary trees, micro-clusters, and wavelets

Stream Data Processing Methods (2)

- Sketches
  - Histograms and wavelets require multi-passes over the data but sketches can operate in a single pass
  - Frequency moments of a stream A = {a_1, ..., a_N}, F_k: $F_k = \sum_{i=1}^{v} m_i^k$
    where v: the universe or domain size, m_i the frequency of i in the sequence
    - Given N elts and v values, sketches can approximate F_0, F_1, F_2 in O(log v + log N) space
- Randomized algorithms
  - Monte Carlo algorithm: bound on running time but may not return correct result
  - Chebyshev's inequality: $P(|X - \mu| > k) \leq \frac{\sigma^2}{k^2}$
    - Let X be a random variable with mean μ and standard deviation σ
  - Chernoff bound: $P[X < (1+\delta)\mu] < e^{-\mu\delta^2/4}$
    - Let X be the sum of independent Poisson trials X_1, ..., X_n, δ in (0, 1]
    - The probability decreases expoentially as we move from the mean