Search and Data Mining Techniques

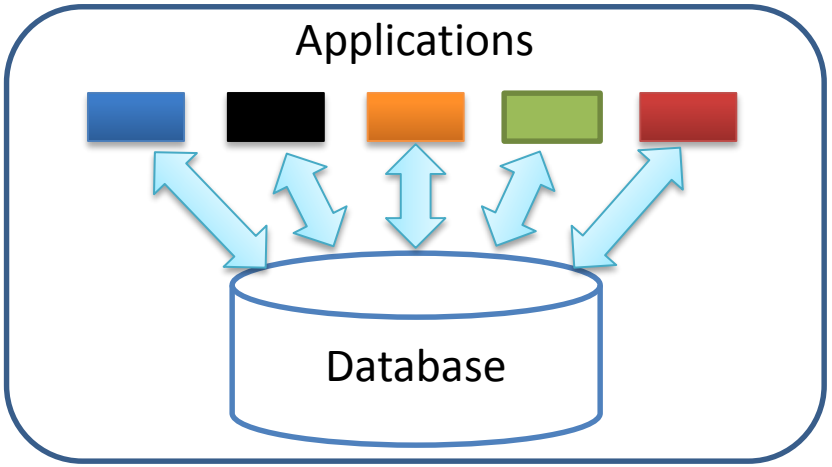
OLAP

Anna Yarygina

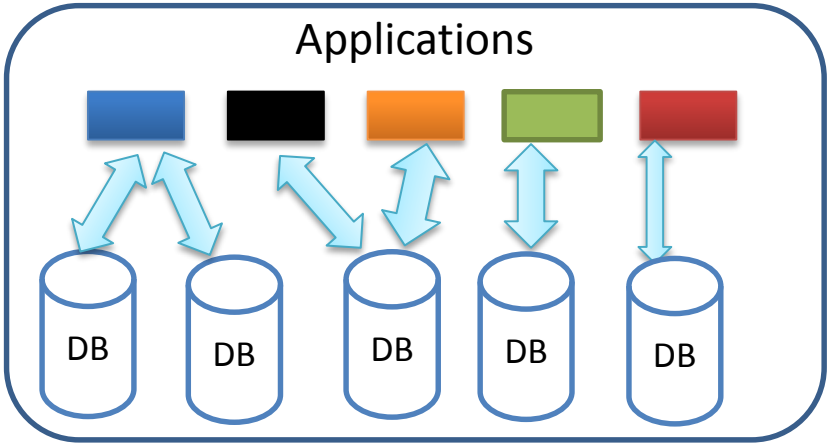
Boris Novikov

The Database: Shared Data Store?

- A dream from database textbooks: Sharing data between applications
- This NEVER happened.

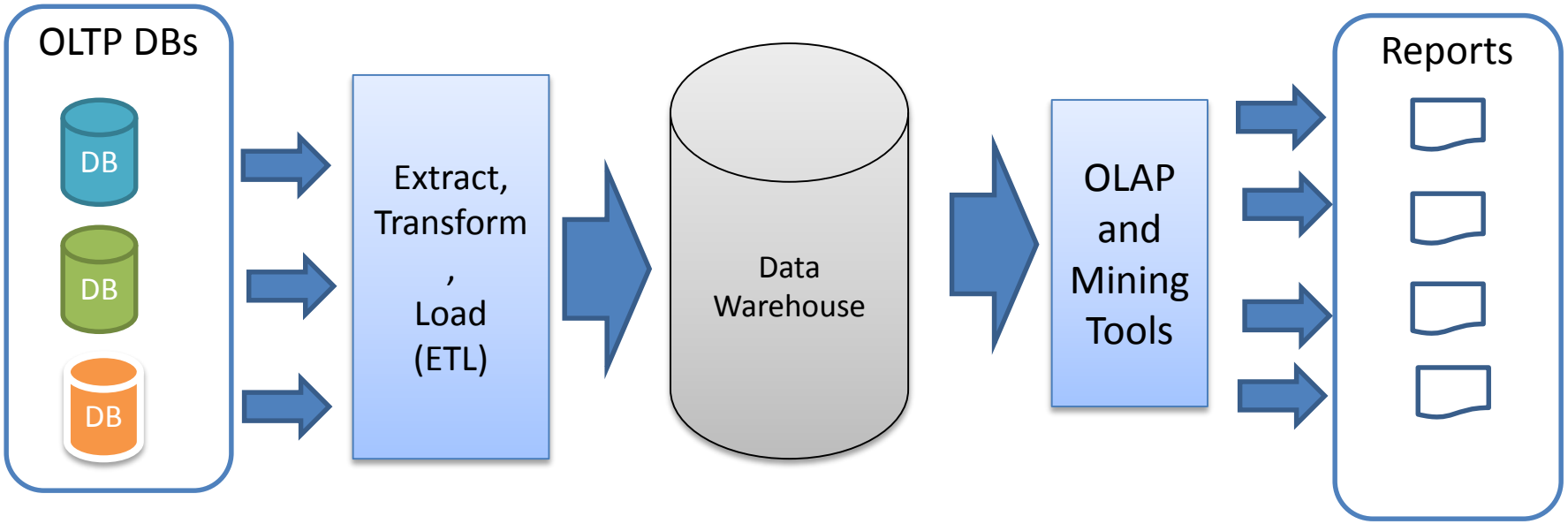


- In reality, a databse is created for small number of related applications
- An application can access multiple databases



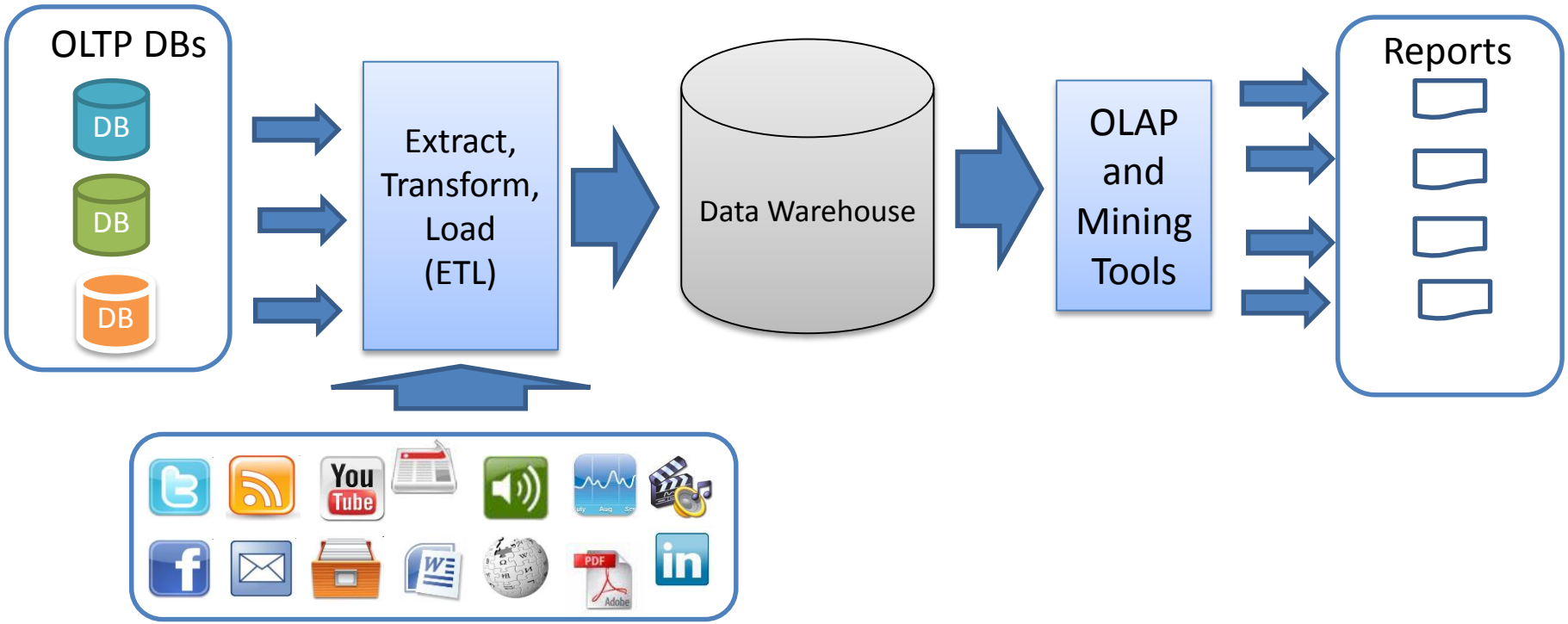
Building Analytical Reports

- Create single point of truth in DW with ETL process
- DW is needed for performance reasons



Including External Resources

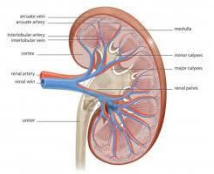
- External resources include Web, search, social networks, text streams, sensor streams, etc.

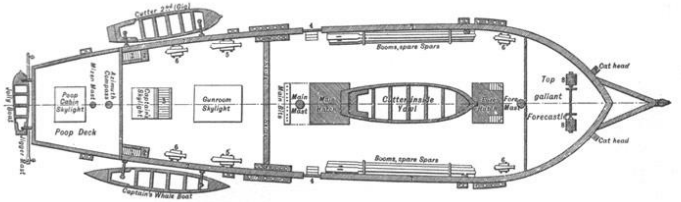


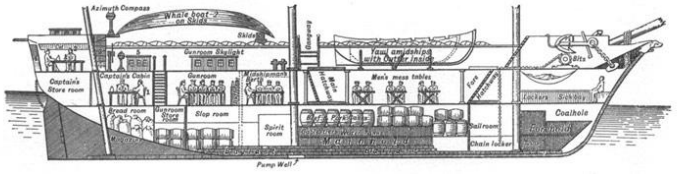
Analytical Reports

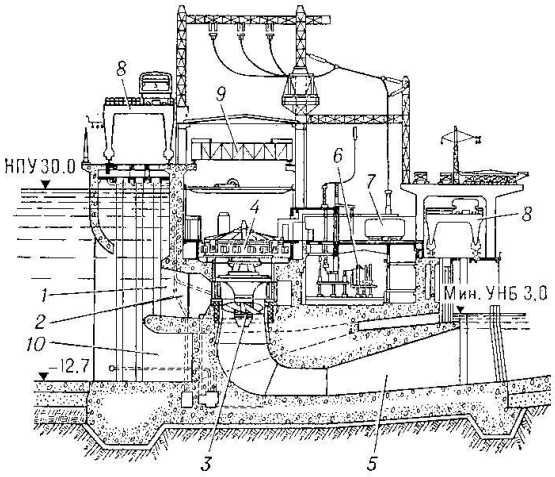
- Results of any analytical task are represented with printed or interactive reports
- A report provides compact representation of certain essential properties of large volume of data
- Typically contains statistical characterizations such as sum, average, quantity, or ranges for specific groups of data items

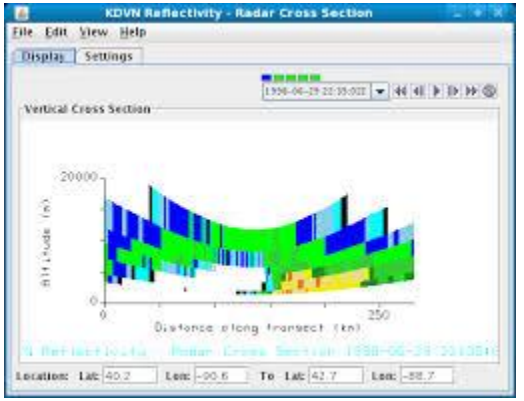
Cross-Sections

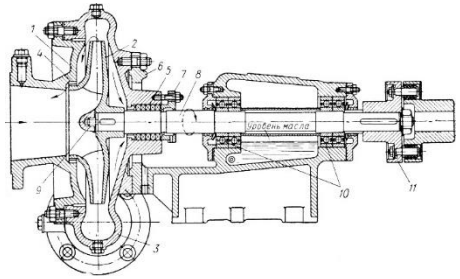


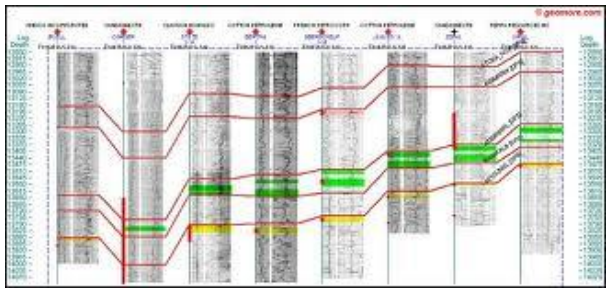








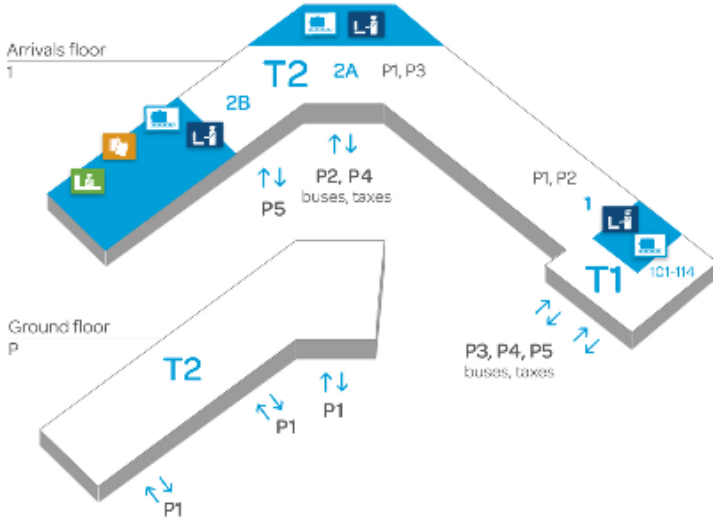




http://www.math.spbu.ru/user/boris_novikov/courses/c-sdmt-index.shtml

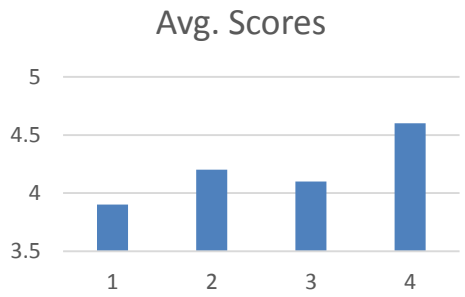
Slices


M/S Baltic Princess
TALLINK Cruise
www.tallink.com
KAT
1 INFO
2 SUN DISCO
3 STARLIGHT PALACE
4 BALTIC PUB
5 KIDS HARBOUR
6 SAILOR'S PUB
7 VIDEO GAMES
8 CIGAR CLUB
9 MARTINI BAR
10 SAUNA
11 GOURMET BALTIC PRINCESS
12 RUSSIAN À LA CARTE KATARINA
13 GRILL HOUSE
14 BUFFET TALLIK
15 TANGO LOUNGE
16 CAFETERIA
17 CONFERENCE
18 COSMETIC BOUTIQUE
19 BOUTIQUE
20 SUPERMARKET

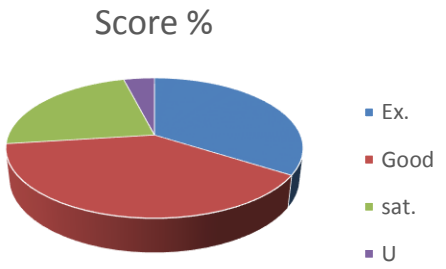

Arrivals floor
1
T2
2A
P1, P3
2B
P2, P4
buses, taxes
P5
P1, P2
1
T1
101-114
P3, P4, P5
buses, taxes
Ground floor
P
T2
P1
P1
P1

Back to Business: Simple Cross-Sections

Avg. Scores	F	M	J	S
	3.9	4.2	4.1	4.6



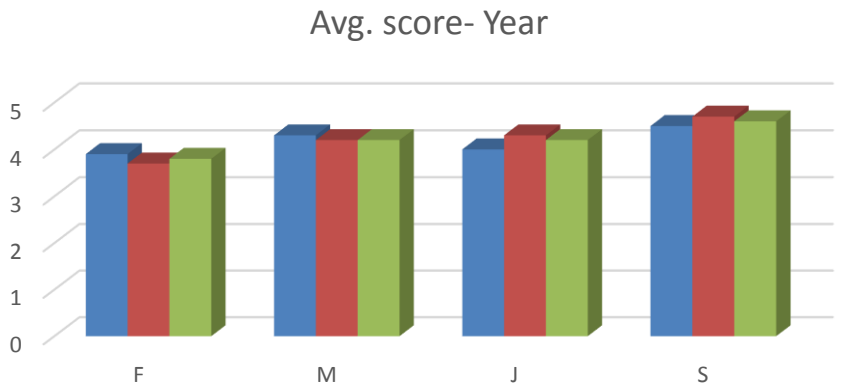
Score %	Ex	G	S	U
	34	39	23	4



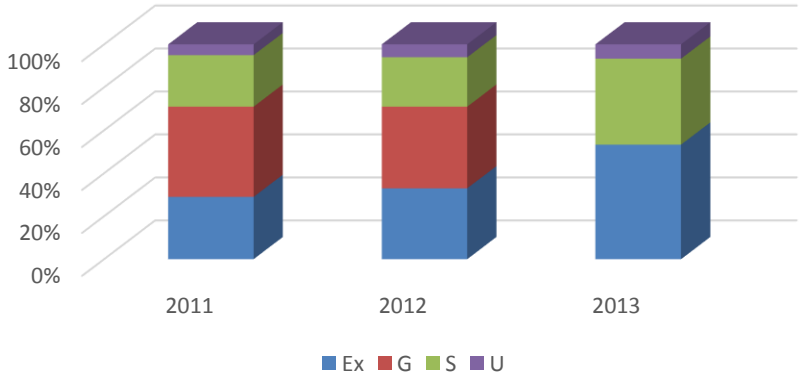
http://www.math.spbu.ru/user/boris_novikov/courses/c-sdmt-index.shtml

Two-Dimensional Cross-Sections

Avg. Scores	F	M	J	S
3013	3.9	4.2	4.1	4.6
2012	3.8	4.1	4.2	4.7
2011	3.9	4.3	4.3	4.5



%	Ex	G	S	U
2011	29	42	24	5
2012	33	38	23	6
2013	32	0	24	4



http://www.math.spbu.ru/user/boris_novikov/courses/c-sdmt-index.shtml

Operations on Reports

- Slice and dice
  - Choose specific value or range in certain dimensions

	Ex	G	S	U
2011	29	42	24	5
2012	33	38	23	6
2013	32	0	24	4

- Drill down
  - Find out how a data value was calculated from more detailed information
- Drill up
  - Find an impact of a value on upper levels

Avg. Scores	F	M	J	S
	3.9	4.2	4.1	4.6
AIS				4.9
CS				4.4
P				4.3
SE				4.8

Data Warehouse

- Used for reporting purposes
- Batch update only
- Most of the time read-only access
- Each query retrieves large data volumes
- Data storage redundancy (aggregated, non-normalized data)

Multi-dimensional Data Model

- Facts
  - References to dimensions
  - Values of measures
- Dimensions
  - Categories
  - Attributes

Hierarchies

- Location
  - Zip
  - City
  - Region
- Product
  - Model
  - Category
- Time
  - Hour
  - Day
  - Month
  - Quarter
  - Year
  - What about weeks?

An Example

- Facts: Exams, measures:
  - Score
- Dimensions
  - Time
    - Date, semester, year
  - Course
    - Type, title, tecturer, dept
  - Student
    - Id, name, year (1-4), group, dept

Mapping to Tables

Fact table (Exams)	Time_id	Course_id	Student_id	Score

Stud_id	Name	Year	Group#	Dept

Course_id	type	Lecturer	dept

Time_id	Day	Semester	Year	Month

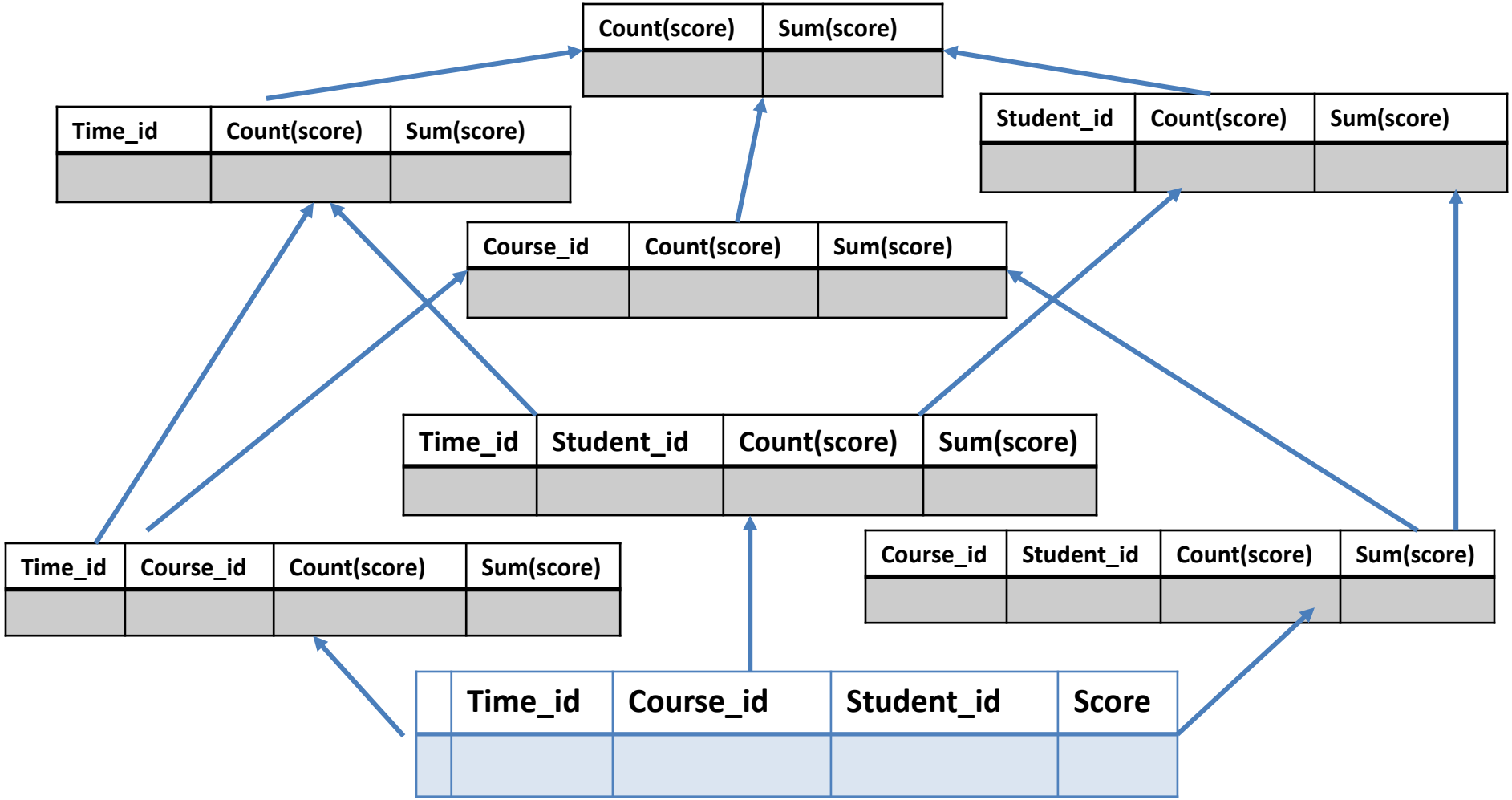
Extracting Data for Reports

```
select
        stud.year,
        avg(score)
from  exams
    right join students using stud_id
    riht join time_dim using time_id
    right join course using course_id
Where
                time_dim.year = 2013  --- slice or dice
Group by
        stud.year       -- cross-section
;
```

Cubes

- Typically any report would require full scan of a huge fact table
- A cube is a set of pre-calculated partial aggregates for certain combinations of dimansions
- Reports are built from pre-calculated aggregates, rather than fact measures.

Building a Cube



Pivot Tables

A	B	C
Russia	red	1
Russia	blue	1
Russia	white	1
France	red	1
France	blue	1
France	white	1
Netherlands	red	1
Netherlands	blue	1
Netherlands	white	1
UK	red	1
UK	blue	1
UK	white	1
Finland	blue	1
Finland	white	1
Ukraine	blue	1
Ukraine	yellow	1
Estonia	blue	1
Estonia	Black	1
Germany	yellow	1
Germany	Black	1
Germany	red	1
Sweden	blue	1
Sweden	yellow	1
Switzerland	red	1
Switzerland	white	1

```
select *  from (
   select         a,       case
      when b in ('red', 'blue', 'white')
then b
      else 'other‘       end as b
   from t_pivot)  p
       pivot (count(*) for b in (
       'red', 'blue', 'white', 'other'));
```

A	'red'	'blue'	'white'	'other'
France	1	1	1	0
Finland	0	1	1	0
Estonia	0	1	0	1
Germany	1	0	0	2
Russia	1	1	1	0
UK	1	1	1	0
Sweden	0	1	0	1
Switzerland	1	0	1	0
Netherlands	1	1	1	0
Ukraine	0	1	0	1

SQL Analytical Extensions

- Window Queries
- Additional aggregate functions

ETL

- Filer source data
- Consolidate sources
- Joins
- Load into data warehouse
- Build cubes

Under the Hood

- Bitmap Indexes
  - Efficient for dimensions with low number of values
- Column-oriented stores
  - Fast scans
  - Compression
  - Low number of attributes per query
- Parallelism
  - ETL
  - Scans

http://www.math.spbu.ru/user/boris_novikov/courses/c-sdmt-index.shtml