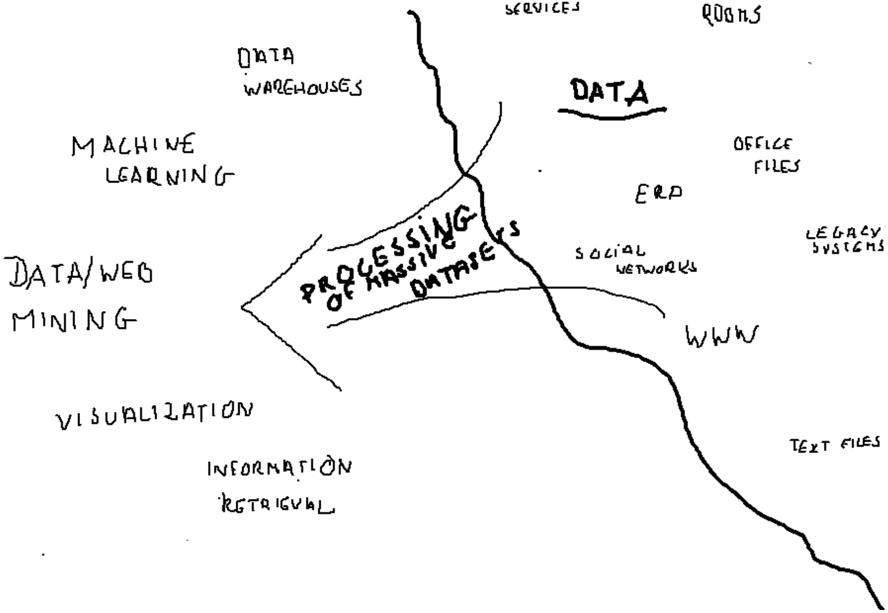
Introduction to Massive Datasets

Krzysztof Dembczyński

Intelligent Decision Support Systems Laboratory (IDSS)
Poznań University of Technology, Poland

Bachelor studies, eighth semester
Academic year 2018/19 (summer semester)

Goal: understanding data ...



SERVICES
RDBMS
DATA
WAREHOUSES
DATA
MACHINE
LEARNING
OFFICE
FILES
ERP
PROCESSING
OF MASSIVE
DATASETS
LEGACY
SYSTEMS
SOCIAL
NETWORKS
DATA/WEB
MINING
WWW
VISUALIZATION
TEXT FILES
INFORMATION
RETRIEVAL

Goal: . . . to make data analysis efficient.

Outline

Outline

- **Buzzwords**: Big Data, Data Science, Machine learning, NoSQL . . .

- **Buzzwords**: Big Data, Data Science, Machine learning, NoSQL ...
- **How Big Data Changes Everything:**
  - Several books showing the impact of Big Data revolution (e.g., **Disruptive Possibilities: How Big Data Changes Everything** by Jeffrey Needham).

- **Buzzwords**: Big Data, Data Science, Machine learning, NoSQL ...
- **How Big Data Changes Everything:**
  - Several books showing the impact of Big Data revolution (e.g., **Disruptive Possibilities: How Big Data Changes Everything** by Jeffrey Needham).
- **Computerworld (Jul 11, 2007)**:
  - *12 IT skills that employers can't say no to:*
    1) Machine learning
    ...

- **Buzzwords**: Big Data, Data Science, Machine learning, NoSQL ...
- **How Big Data Changes Everything:**
  - Several books showing the impact of Big Data revolution (e.g., **Disruptive Possibilities: How Big Data Changes Everything** by Jeffrey Needham).
- **Computerworld (Jul 11, 2007)**:
  - *12 IT skills that employers can't say no to:*
    1) Machine learning
    ...
- **Three priorities of Google announced at BoxDev 2015**:
  - Machine learning – speech recognition
  - Machine learning – image understanding
  - Machine learning – preference learning/personalization

- **Buzzwords**: Big Data, Data Science, Machine learning, NoSQL ...
- **How Big Data Changes Everything:**
  - Several books showing the impact of Big Data revolution (e.g., **Disruptive Possibilities: How Big Data Changes Everything** by Jeffrey Needham).
- **Computerworld (Jul 11, 2007)**:
  - *12 IT skills that employers can't say no to:*
    1) Machine learning
    ...
- **Three priorities of Google announced at BoxDev 2015**:
  - Machine learning – speech recognition
  - Machine learning – image understanding
  - Machine learning – preference learning/personalization
- **OpenAI** founded in 2015 as a non-profit artificial intelligence research company.

Data mining

- Data mining is the discovery of **models** for data, ...
- But what is a model?

if all you have is a hammer, everything looks like a nail

How to characterize your data?

- **Database programmer** usually writes:

```
select avg(column), std(column) from data
```

How to characterize your data?

- **Database programmer** usually writes:

```
select avg(column), std(column) from data
```

- **Statistician** might decide that the data comes from a Gaussian distribution and use a formula to compute the most likely parameters of this Gaussian: the mean and standard deviation.

How to characterize your data?

- **Database programmer** usually writes:

```
select avg(column), std(column) from data
```

- **Statistician** might decide that the data comes from a Gaussian distribution and use a formula to compute the most likely parameters of this Gaussian: the mean and standard deviation.
- **Machine learner** will use the data as training examples and apply a learning algorithm to get a model that predicts future data.

How to characterize your data?

- **Database programmer** usually writes:

```
select avg(column), std(column) from data
```

- **Statistician** might decide that the data comes from a Gaussian distribution and use a formula to compute the most likely parameters of this Gaussian: the mean and standard deviation.
- **Machine learner** will use the data as training examples and apply a learning algorithm to get a model that predicts future data.
- **Data miner** will discover the most frequent patterns.

They all want to understand data and use this knowledge for making better decisions

Data+ideas vs. statistics+algorithms

- About the Amazon's recommender system:

  > *It's often more important to creatively invent new data sources than to implement the latest academic variations on an algorithm.*

Data+ideas vs. statistics+algorithms

- About the Amazon's recommender system:

  > *It's often more important to creatively invent new data sources than to implement the latest academic variations on an algorithm.*

- WhizBang! Labs tried to use machine learning to locate people's resumes on the Web: the algorithm was not able to do better than procedures designed by hand, since a resume has a quite standard shape and sentences.

Data+computational power

- Object recognition in computer vision:

Data+computational power

- Object recognition in computer vision:
  - Scanning large databases **can perform better** than the best computer vision algorithms!

Data+computational power

- Object recognition in computer vision:
  - Scanning large databases **can perform better** than the best computer vision algorithms!
- Automatic translation

Data+computational power

- Object recognition in computer vision:
  - Scanning large databases **can perform better** than the best computer vision algorithms!
- Automatic translation
  - Statistical translation based on large corpora **outperforms** linguistic models!

Human computation

- CAPTCHA and reCAPTCHA
- ESP game
- Check a lecture given by Luis von Ahn:
  http://videolectures.net/iaai09_vonahn_hc/
- Amazon Mechanical Turk

Data+ideas vs. statistics+algorithms

Those who ignore Statistics are condemned to reinvent it.

Brad Efron

- In Statistics, a term **data mining** was originally referring to attempts to extract information that was not supported by the data.
- Bonferroni's Principle: "if you look in more places for interesting patterns than your amount of data will support, you are bound to find crap".
- Rhine paradox.

Data+ideas vs. statistics+algorithms

- The data mining algorithms can perform quite well!!!

Data+ideas vs. statistics+algorithms

- The data mining algorithms can perform quite well!!!
  - XBox Kinect: object tracking vs. pattern recognition (check: `http://videolectures.net/ecmlpkdd2011_bishop_embracing/`).

Data+ideas vs. statistics+algorithms

- The data mining algorithms can perform quite well!!!
  - XBox Kinect: object tracking vs. pattern recognition (check: `http://videolectures.net/ecmlpkdd2011_bishop_embracing/`).
  - Pattern finding: association rules.

Data+ideas vs. statistics+algorithms

- The data mining algorithms can perform quite well!!!
  - XBox Kinect: object tracking vs. pattern recognition (check: `http://videolectures.net/ecmlpkdd2011_bishop_embracing/`).
  - Pattern finding: association rules.
  - Netflix: recommender system.

Data+ideas vs. statistics+algorithms

- The data mining algorithms can perform quite well!!!
  - XBox Kinect: object tracking vs. pattern recognition (check: http://videolectures.net/ecmlpkdd2011_bishop_embracing/).
  - Pattern finding: association rules.
  - Netflix: recommender system.
  - Google and PageRank.

Data+ideas vs. statistics+algorithms

- The data mining algorithms can perform quite well!!!
  - XBox Kinect: object tracking vs. pattern recognition (check: `http://videolectures.net/ecmlpkdd2011_bishop_embracing/`).
  - Pattern finding: association rules.
  - Netflix: recommender system.
  - Google and PageRank.
  - Clustering of Cholera cases in 1854.

Data+ideas vs. statistics+algorithms

- The data mining algorithms can perform quite well!!!
  - XBox Kinect: object tracking vs. pattern recognition (check: `http://videolectures.net/ecmlpkdd2011_bishop_embracing/`).
  - Pattern finding: association rules.
  - Netflix: recommender system.
  - Google and PageRank.
  - Clustering of Cholera cases in 1854.
  - Win one of the Kaggle's competitions!!! `http://www.kaggle.com/`.

Data+ideas vs. statistics+algorithms

- The data mining algorithms can perform quite well!!!
  - XBox Kinect: object tracking vs. pattern recognition (check: `http://videolectures.net/ecmlpkdd2011_bishop_embracing/`).
  - Pattern finding: association rules.
  - Netflix: recommender system.
  - Google and PageRank.
  - Clustering of Cholera cases in 1854.
  - Win one of the Kaggle's competitions!!! `http://www.kaggle.com/`.
  - Autonomous cars.

Data+ideas vs. statistics+algorithms

- The data mining algorithms can perform quite well!!!
  - XBox Kinect: object tracking vs. pattern recognition (check: `http://videolectures.net/ecmlpkdd2011_bishop_embracing/`).
  - Pattern finding: association rules.
  - Netflix: recommender system.
  - Google and PageRank.
  - Clustering of Cholera cases in 1854.
  - Win one of the Kaggle's competitions!!! `http://www.kaggle.com/`.
  - Autonomous cars.
  - Deep learning.

Data+ideas vs. statistics+algorithms

- The data mining algorithms can perform quite well!!!
  - XBox Kinect: object tracking vs. pattern recognition (check: `http://videolectures.net/ecmlpkdd2011_bishop_embracing/`).
  - Pattern finding: association rules.
  - Netflix: recommender system.
  - Google and PageRank.
  - Clustering of Cholera cases in 1854.
  - Win one of the Kaggle's competitions!!! `http://www.kaggle.com/`.
  - Autonomous cars.
  - Deep learning.
  - And many others.

Data+ideas+computational power+statistics+algorithms

To be learned in the upcoming semester . . .

The aim and the scope of the course

- **Aim**: To get to know technologies and algorithms for processing massive datasets.

The aim and the scope of the course

- **Aim**: To get to know technologies and algorithms for processing massive datasets.
- **Scope**: We will learn how to organize, store, access, and process massive datasets:

The aim and the scope of the course

- **Aim**: To get to know technologies and algorithms for processing massive datasets.
- **Scope**: We will learn how to organize, store, access, and process massive datasets:
  - Write-once-read-many-times type of data stores,

The aim and the scope of the course

- **Aim**: To get to know technologies and algorithms for processing massive datasets.
- **Scope**: We will learn how to organize, store, access, and process massive datasets:
  - Write-once-read-many-times type of data stores,
  - Analytical databases and star schemas,

The aim and the scope of the course

- **Aim**: To get to know technologies and algorithms for processing massive datasets.
- **Scope**: We will learn how to organize, store, access, and process massive datasets:
  - Write-once-read-many-times type of data stores,
  - Analytical databases and star schemas,
  - Data structures and fast algorithms for processing massive datasets,

The aim and the scope of the course

- **Aim**: To get to know technologies and algorithms for processing massive datasets.
- **Scope**: We will learn how to organize, store, access, and process massive datasets:
  - Write-once-read-many-times type of data stores,
  - Analytical databases and star schemas,
  - Data structures and fast algorithms for processing massive datasets,
  - Distributed processing and MapReduce,

The aim and the scope of the course

- **Aim**: To get to know technologies and algorithms for processing massive datasets.
- **Scope**: We will learn how to organize, store, access, and process massive datasets:
  - Write-once-read-many-times type of data stores,
  - Analytical databases and star schemas,
  - Data structures and fast algorithms for processing massive datasets,
  - Distributed processing and MapReduce,
  - Approximate query processing,

The aim and the scope of the course

- **Aim**: To get to know technologies and algorithms for processing massive datasets.
- **Scope**: We will learn how to organize, store, access, and process massive datasets:
  - Write-once-read-many-times type of data stores,
  - Analytical databases and star schemas,
  - Data structures and fast algorithms for processing massive datasets,
  - Distributed processing and MapReduce,
  - Approximate query processing,
  - Nearest neighbor search.

The aim and the scope of the course

- **Aim**: To get to know technologies and algorithms for processing massive datasets.
- **Scope**: We will learn how to organize, store, access, and process massive datasets:
  - Write-once-read-many-times type of data stores,
  - Analytical databases and star schemas,
  - Data structures and fast algorithms for processing massive datasets,
  - Distributed processing and MapReduce,
  - Approximate query processing,
  - Nearest neighbor search.
- The course is based on the first 4 chapters of the **Mining of Massive Datasets** book: `http://www.mmds.org/`

Main information about the course

- Instructor:
  - dr hab. inż. Krzysztof Dembczyński (`kdembczynskicsputpoznanpl`)

Main information about the course

- Instructor:
  - dr hab. inż. Krzysztof Dembczyński (`kdembczynskicsputpoznanpl`)
- Website:
  - `www.cs.put.poznan.pl/kdembczynski/lectures/pmds-sc`

Main information about the course

- Instructor:
  - dr hab. inż. Krzysztof Dembczyński (`kdembczynskicsputpoznanpl`)
- Website:
  - `www.cs.put.poznan.pl/kdembczynski/lectures/pmds-sc`
- Office hours: Thursday, 10:00-12:00, room 2 CW (Institute of Computing Science).

Lectures

- Main topics of lectures:
  - Introduction
  - Processing of massive data sets
  - Distributed systems and MapReduce
  - MapReduce in Spark
  - Approximate query processing
  - Nearest neighbor search

Labs

- Strong connection between lectures and labs.
- Software: bash, Spark (Python, Java, Scala), programming language of your choice.
- List of tasks and exercises for each meeting (also homeworks).
- Mainly mini programming projects and short exercises.
- Main topics:
  - Data modeling, data transformation and processing
  - MapReduce in Spark
  - Approximate query processing
  - Finding similar items

Evaluation

- **Lecture**:

Test:	75 % of points	(min. 50%)
Labs:	25 % of points	(min. 50%)

- **Labs**:

  Regular exercises and home works: 100 % of points (min. 50%)

- **Scale**:

90 % of pts = 5.0	80 % of pts = 4.5	70 % of pts = 4.0
60 % of pts = 3.5	50 % of pts = 3.0	otherwise = 2.0

- **Bonus points for all**: up to 10 points.

Bibliography

- J. Leskovec, A. Rajaraman, and J. D. Ullman. *Mining of Massive Datasets*.
  Cambridge University Press, 2014
  `http://www.mmds.org`

- R. Kimball and M. Ross. *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling, 3rd Edition*.
  John Wiley & Sons, 2013

- H. Garcia-Molina, J. D. Ullman, and J. Widom. *Database Systems: The Complete Book. Second Edition*.
  Pearson Prentice Hall, 2009

Outline

Data is the new oil (?)

Database management system

- A database is a collection of information that exists over a long period of time.

Database management system

- A database is a collection of information that exists over a long period of time.
- A database management system (DBMS) is specialized software responsible for managing the database.

Database management system

- A database is a collection of information that exists over a long period of time.
- A database management system (DBMS) is specialized software responsible for managing the database.
- The DBMS is expected to:

Database management system

- A database is a collection of information that exists over a long period of time.
- A database management system (DBMS) is specialized software responsible for managing the database.
- The DBMS is expected to:
    - Allow users to create new databases and specify their schemas (logical structure of data),

Database management system

- A database is a collection of information that exists over a long period of time.
- A database management system (DBMS) is specialized software responsible for managing the database.
- The DBMS is expected to:
  - Allow users to create new databases and specify their schemas (logical structure of data),
  - Give users the ability of query the data and modify the data,

Database management system

- A database is a collection of information that exists over a long period of time.
- A database management system (DBMS) is specialized software responsible for managing the database.
- The DBMS is expected to:
  - Allow users to create new databases and specify their schemas (logical structure of data),
  - Give users the ability of query the data and modify the data,
  - Support the storage of very large amounts of data, allowing efficient access to data for queries and database modifications,

Database management system

- A database is a collection of information that exists over a long period of time.
- A database management system (DBMS) is specialized software responsible for managing the database.
- The DBMS is expected to:
  - Allow users to create new databases and specify their schemas (logical structure of data),
  - Give users the ability of query the data and modify the data,
  - Support the storage of very large amounts of data, allowing efficient access to data for queries and database modifications,
  - Enable durability, the recovery of the database in the face of failures,

Database management system

- A database is a collection of information that exists over a long period of time.
- A database management system (DBMS) is specialized software responsible for managing the database.
- The DBMS is expected to:
  - Allow users to create new databases and specify their schemas (logical structure of data),
  - Give users the ability of query the data and modify the data,
  - Support the storage of very large amounts of data, allowing efficient access to data for queries and database modifications,
  - Enable durability, the recovery of the database in the face of failures,
  - Control access to data from many users at once in isolation and ensure the actions on data to be performed completely.

Data models

- **Data model** is an abstract model that defines how data is represented and accessed.
  - **Logical data model** – from a user's point of view
  - **Physical data model** – from a computer's point of view.
- Data model defines:
  - Data objects and types, relationships between data objects, and constraints imposed on them.
  - Operations for defining, searching and updating data.

Approaches to data management

- File management system

Approaches to data management

- File management system
- Database management system

Approaches to data management

- File management system
- Database management system
  - Early database management systems (e.g. hierarchical or network data models)

Approaches to data management

- File management system
- Database management system
  - Early database management systems (e.g. hierarchical or network data models)
  - Relational database systems

Approaches to data management

- File management system
- Database management system
  - Early database management systems (e.g. hierarchical or network data models)
  - Relational database systems
  - Post-relational database systems

Approaches to data management

- File management system
- Database management system
  - Early database management systems (e.g. hierarchical or network data models)
  - Relational database systems
  - Post-relational database systems
  - Object-based database systems

Approaches to data management

- File management system
- Database management system
  - Early database management systems (e.g. hierarchical or network data models)
  - Relational database systems
  - Post-relational database systems
  - Object-based database systems
  - Multi-dimensional database systems

Approaches to data management

- File management system
- Database management system
  - Early database management systems (e.g. hierarchical or network data models)
  - Relational database systems
  - Post-relational database systems
  - Object-based database systems
  - Multi-dimensional database systems
- NoSQL and BigData

Approaches to data management

- File management system
- Database management system
  - Early database management systems (e.g. hierarchical or network data models)
  - Relational database systems
  - Post-relational database systems
  - Object-based database systems
  - Multi-dimensional database systems
- NoSQL and BigData
- NewSQL

Approaches to data management

- File management system
- Database management system
  - Early database management systems (e.g. hierarchical or network data models)
  - Relational database systems
  - Post-relational database systems
  - Object-based database systems
  - Multi-dimensional database systems
- NoSQL and BigData
- NewSQL
- **The choice of the approach strongly depends on a given application!**

What is NoSQL?

- Not every data management/analysis problem is best solved exclusively using a traditional relational DBMS

What is NoSQL?

- Not every data management/analysis problem is best solved exclusively using a traditional relational DBMS
- **No** means rather "Not only" and **SQL** states for "traditional relational DBMS".

What is NoSQL?

- Not every data management/analysis problem is best solved exclusively using a traditional relational DBMS
- **No** means rather "Not only" and **SQL** states for "traditional relational DBMS".
- NoSQL systems are alternative to traditional relational DBMS

What is NoSQL?

- Not every data management/analysis problem is best solved exclusively using a traditional relational DBMS
- **No** means rather "Not only" and **SQL** states for "traditional relational DBMS".
- NoSQL systems are alternative to traditional relational DBMS
  - Flexible schema (less restricted than typical RDBMS, but may not support join operations)

What is NoSQL?

- Not every data management/analysis problem is best solved exclusively using a traditional relational DBMS
- **No** means rather "Not only" and **SQL** states for "traditional relational DBMS".
- NoSQL systems are alternative to traditional relational DBMS
  - Flexible schema (less restricted than typical RDBMS, but may not support join operations)
  - Quicker/cheaper to set up

What is NoSQL?

- Not every data management/analysis problem is best solved exclusively using a traditional relational DBMS
- **No** means rather "Not only" and **SQL** states for "traditional relational DBMS".
- NoSQL systems are alternative to traditional relational DBMS
  - Flexible schema (less restricted than typical RDBMS, but may not support join operations)
  - Quicker/cheaper to set up
  - Massive scalability (scale-out instead of scale-up)

What is NoSQL?

- Not every data management/analysis problem is best solved exclusively using a traditional relational DBMS
- **No** means rather "Not only" and **SQL** states for "traditional relational DBMS".
- NoSQL systems are alternative to traditional relational DBMS
  - Flexible schema (less restricted than typical RDBMS, but may not support join operations)
  - Quicker/cheaper to set up
  - Massive scalability (scale-out instead of scale-up)
  - Relaxed consistency $\rightarrow$ higher performance and availability, but fewer guarantees (like ACID)

What is NoSQL?

- Not every data management/analysis problem is best solved exclusively using a traditional relational DBMS
- **No** means rather "Not only" and **SQL** states for "traditional relational DBMS".
- NoSQL systems are alternative to traditional relational DBMS
    - Flexible schema (less restricted than typical RDBMS, but may not support join operations)
    - Quicker/cheaper to set up
    - Massive scalability (scale-out instead of scale-up)
    - Relaxed consistency $\rightarrow$ higher performance and availability, but fewer guarantees (like ACID)
    - Not all operations supported (e.g., join operation)

What is NoSQL?

- Not every data management/analysis problem is best solved exclusively using a traditional relational DBMS
- **No** means rather "Not only" and **SQL** states for "traditional relational DBMS".
- NoSQL systems are alternative to traditional relational DBMS
  - Flexible schema (less restricted than typical RDBMS, but may not support join operations)
  - Quicker/cheaper to set up
  - Massive scalability (scale-out instead of scale-up)
  - Relaxed consistency $\rightarrow$ higher performance and availability, but fewer guarantees (like ACID)
  - Not all operations supported (e.g., join operation)
  - No declarative query language (requires more programming, but new paradigms like MapReduce appear)

NoSQL

- Different types of models:

NoSQL

- Different types of models:
  - MapReduce frameworks,

NoSQL

- Different types of models:
  - MapReduce frameworks,
  - key-values stores,

NoSQL

- Different types of models:
  - MapReduce frameworks,
  - key-values stores,
  - column stores and BigTable implementations,

NoSQL

- Different types of models:
  - MapReduce frameworks,
  - key-values stores,
  - column stores and BigTable implementations,
  - document-oriented databases,

NoSQL

- Different types of models:
  - MapReduce frameworks,
  - key-values stores,
  - column stores and BigTable implementations,
  - document-oriented databases,
  - graph database systems.

NoSQL

- Different types of models:
    - MapReduce frameworks,
    - key-values stores,
    - column stores and BigTable implementations,
    - document-oriented databases,
    - graph database systems.
- Designed for different purposes.

BigData – a lot of Vs[1]

- **Volume**: the quantity of generated and stored data.
- **Variety**: the type and nature of the data.
- **Velocity**: the speed at which the data is generated and processed.
- **Veracity**: the quality of captured data.

[1] https://en.wikipedia.org/wiki/Big_data

Two types of systems

- **Operational systems**:

Two types of systems

- **Operational systems**:
  - Support day-to-day operations of an organization,

Two types of systems

- **Operational systems**:
  - Support day-to-day operations of an organization,
  - Also referred to as **on-line transaction processing** (OLTP).

Two types of systems

- **Operational systems**:
  - Support day-to-day operations of an organization,
  - Also referred to as **on-line transaction processing** (OLTP).
  - **Main tasks**: processing of a huge number of concurrent transactions, and insuring data integrity.

Two types of systems

- **Operational systems**:
  - Support day-to-day operations of an organization,
  - Also referred to as **on-line transaction processing** (OLTP).
  - **Main tasks**: processing of a huge number of concurrent transactions, and insuring data integrity.
- **Analytical systems**:

Two types of systems

- **Operational systems**:
  - Support day-to-day operations of an organization,
  - Also referred to as **on-line transaction processing** (OLTP).
  - **Main tasks**: processing of a huge number of concurrent transactions, and insuring data integrity.
- **Analytical systems**:
  - support knowledge workers (e.g., manager, executive, analyst) in decision making,

Two types of systems

- **Operational systems**:
  - Support day-to-day operations of an organization,
  - Also referred to as **on-line transaction processing** (OLTP).
  - **Main tasks**: processing of a huge number of concurrent transactions, and insuring data integrity.
- **Analytical systems**:
  - support knowledge workers (e.g., manager, executive, analyst) in decision making,
  - Also referred to as **on-line analytical processing** (OLAP).

Two types of systems

- **Operational systems**:
  - Support day-to-day operations of an organization,
  - Also referred to as **on-line transaction processing** (OLTP).
  - **Main tasks**: processing of a huge number of concurrent transactions, and insuring data integrity.
- **Analytical systems**:
  - support knowledge workers (e.g., manager, executive, analyst) in decision making,
  - Also referred to as **on-line analytical processing** (OLAP).
  - **Main tasks**: effective processing of multidimensional queries concerning huge volumes of data.

Two types of systems

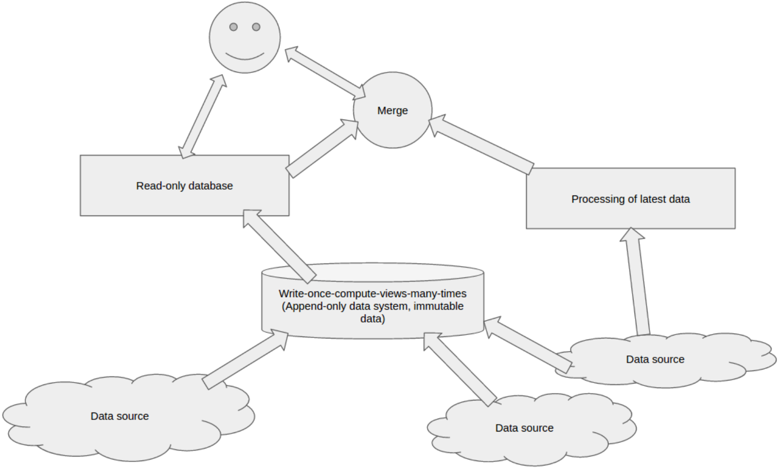
- **Operational systems**:
  - Support day-to-day operations of an organization,
  - Also referred to as **on-line transaction processing** (OLTP).
  - **Main tasks**: processing of a huge number of concurrent transactions, and insuring data integrity.
- **Analytical systems**:
  - support knowledge workers (e.g., manager, executive, analyst) in decision making,
  - Also referred to as **on-line analytical processing** (OLAP).
  - **Main tasks**: effective processing of multidimensional queries concerning huge volumes of data.
  - Database systems of a **write-once-read-many-times** type.

Outline

Analytical database systems

- Data warehouses,
- Business intelligence,
- Computational and analytical tools,
- Scientific databases,
- Analytics engines for large-scale data processing.

Analytical database systems



Analytical database systems

- The old and still good definition of the data warehouse:

Analytical database systems

- The old and still good definition of the data warehouse:
  - **Data warehouse** is defined as a subject-oriented, integrated, time-variant, and non-volatile collection of data in support of management's decision-making process.

Analytical database systems

- The old and still good definition of the data warehouse:
  - **Data warehouse** is defined as a subject-oriented, integrated, time-variant, and non-volatile collection of data in support of management's decision-making process.
    - **Subject oriented**: oriented to the major subject areas of the corporation that have been defined in the data model.

Analytical database systems

- The old and still good definition of the data warehouse:
  - **Data warehouse** is defined as a subject-oriented, integrated, time-variant, and non-volatile collection of data in support of management's decision-making process.
    - **Subject oriented**: oriented to the major subject areas of the corporation that have been defined in the data model.
    - **Integrated**: there is no consistency in encoding, naming conventions, etc., among different data sources that are heterogeneous data sources (when data is moved to the warehouse, it is converted).

Analytical database systems

- The old and still good definition of the data warehouse:
  - **Data warehouse** is defined as a subject-oriented, integrated, time-variant, and non-volatile collection of data in support of management's decision-making process.
    - **Subject oriented**: oriented to the major subject areas of the corporation that have been defined in the data model.
    - **Integrated**: there is no consistency in encoding, naming conventions, etc., among different data sources that are heterogeneous data sources (when data is moved to the warehouse, it is converted).
    - **Non-volatile**: warehouse data is loaded and accessed; update of data does not occur in the data warehouse environment.

Analytical database systems

- The old and still good definition of the data warehouse:
  - **Data warehouse** is defined as a subject-oriented, integrated, time-variant, and non-volatile collection of data in support of management's decision-making process.
    - **Subject oriented**: oriented to the major subject areas of the corporation that have been defined in the data model.
    - **Integrated**: there is no consistency in encoding, naming conventions, etc., among different data sources that are heterogeneous data sources (when data is moved to the warehouse, it is converted).
    - **Non-volatile**: warehouse data is loaded and accessed; update of data does not occur in the data warehouse environment.
    - **Time-variant**: the time horizon for the data warehouse is significantly longer than that of operational systems.

Life-cycle of analytical database systems

- Logical design of the database
- Design and implementation of ETL process
- Deployment of the system
- Optimization of the system
- Refreshing of the data

Logical design of the database

- University authorities decided to analyze teaching performance by using the data collected in databases owned by the university containing information about students, instructors, lectures, faculties, etc.
- They would like to get answers for the following queries:

Logical design of the database

- University authorities decided to analyze teaching performance by using the data collected in databases owned by the university containing information about students, instructors, lectures, faculties, etc.
- They would like to get answers for the following queries:
  - What is the average score of students over academic years?
  - What is the number of students over academic years?
  - What is the average score by faculties, instructors, etc.?
  - What is the distribution of students over faculties, semesters, etc.?
  - ...

Example

- An exemplary query could be the following:

```
SELECT Instructor, Academic_year, AVG(Grade)
FROM Data_Warehouse
GROUP BY Instructor, Academic_year
```

- And the result:

Academic_year	Name	AVG(Grade)
2013/14	Stefanowski	4.2
2014/15	Stefanowski	4.5
2013/14	Słowiński	4.1
2014/15	Słowiński	4.3
2014/15	Dembczyński	4.6

vs.

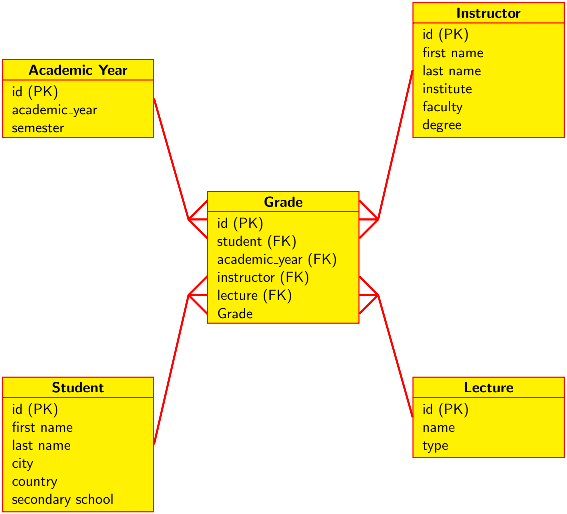
AVG(Grade)	Academic_year	
Name	2013/2014	2014/2015
Stefanowski	4.2	4.5
Słowiński	4.1	4.3
Dembczyński	4.7	4.6

Conceptual schemes of data warehouses

- Three main goals for logical design:
  - Simplicity:
    - Users should understand the design,
    - Data model should match users' conceptual model,
    - Queries should be easy and intuitive to write.
  - Expressiveness:
    - Include enough information to answer all important queries,
    - Include all relevant data (without irrelevant data).
  - Performance:
    - An efficient physical design should be possible to apply.

Star schema

- A single table in the middle connected to a number of dimension tables.



Star schema

- **Measures**, e.g. grades, price, quantity.

Star schema

- **Measures**, e.g. grades, price, quantity.
  - Measures should be aggregative.

Star schema

- **Measures**, e.g. grades, price, quantity.
  - Measures should be aggregative.
  - Measures depend on a set of dimensions, e.g. student grade depends on student, course, instructor, faculty, academic year, etc.

Star schema

- **Measures**, e.g. grades, price, quantity.
  - Measures should be aggregative.
  - Measures depend on a set of dimensions, e.g. student grade depends on student, course, instructor, faculty, academic year, etc.
- **Dimension tables**

Star schema

- **Measures**, e.g. grades, price, quantity.
  - Measures should be aggregative.
  - Measures depend on a set of dimensions, e.g. student grade depends on student, course, instructor, faculty, academic year, etc.
- **Dimension tables**
  - Represent information about dimensions (student, academic year, etc.).

Star schema

- **Measures**, e.g. grades, price, quantity.
  - Measures should be aggregative.
  - Measures depend on a set of dimensions, e.g. student grade depends on student, course, instructor, faculty, academic year, etc.
- **Dimension tables**
  - Represent information about dimensions (student, academic year, etc.).
  - Each dimension has a set of descriptive attributes.

Star schema

- **Measures**, e.g. grades, price, quantity.
  - Measures should be aggregative.
  - Measures depend on a set of dimensions, e.g. student grade depends on student, course, instructor, faculty, academic year, etc.
- **Dimension tables**
  - Represent information about dimensions (student, academic year, etc.).
  - Each dimension has a set of descriptive attributes.
  - The attributes of dimension tables are used for filtering and grouping.

Star schema

- **Measures**, e.g. grades, price, quantity.
  - Measures should be aggregative.
  - Measures depend on a set of dimensions, e.g. student grade depends on student, course, instructor, faculty, academic year, etc.
- **Dimension tables**
  - Represent information about dimensions (student, academic year, etc.).
  - Each dimension has a set of descriptive attributes.
  - The attributes of dimension tables are used for filtering and grouping.
- **Fact table**

Star schema

- **Measures**, e.g. grades, price, quantity.
  - Measures should be aggregative.
  - Measures depend on a set of dimensions, e.g. student grade depends on student, course, instructor, faculty, academic year, etc.
- **Dimension tables**
  - Represent information about dimensions (student, academic year, etc.).
  - Each dimension has a set of descriptive attributes.
  - The attributes of dimension tables are used for filtering and grouping.
- **Fact table**
  - Relates the dimensions to the measures.

Star schema

- **Measures**, e.g. grades, price, quantity.
  - Measures should be aggregative.
  - Measures depend on a set of dimensions, e.g. student grade depends on student, course, instructor, faculty, academic year, etc.
- **Dimension tables**
  - Represent information about dimensions (student, academic year, etc.).
  - Each dimension has a set of descriptive attributes.
  - The attributes of dimension tables are used for filtering and grouping.
- **Fact table**
  - Relates the dimensions to the measures.
  - Any new fact is added to the fact table.

Star schema

- **Measures**, e.g. grades, price, quantity.
  - Measures should be aggregative.
  - Measures depend on a set of dimensions, e.g. student grade depends on student, course, instructor, faculty, academic year, etc.
- **Dimension tables**
  - Represent information about dimensions (student, academic year, etc.).
  - Each dimension has a set of descriptive attributes.
  - The attributes of dimension tables are used for filtering and grouping.
- **Fact table**
  - Relates the dimensions to the measures.
  - Any new fact is added to the fact table.
  - The aggregated fact columns are the matter of the analysis.

Facts contain numbers, dimensions contain labels

- Fact table:
  - narrow,
  - big (many rows),
  - numeric (rows are described by numerical measures),
  - dynamic (growing over time).
- Dimension table
  - wide,
  - small (few rows),
  - descriptive (rows are described by descriptive attributes),
  - static.

Denormalization

- Denormalization is the process of attempting to optimize the performance of a database by adding redundant data or by grouping data.

Denormalization

- Denormalization is the process of attempting to optimize the performance of a database by adding redundant data or by grouping data.
- Denormalization helps cover up the inefficiencies inherent in relational database software.

Denormalization

- Denormalization is the process of attempting to optimize the performance of a database by adding redundant data or by grouping data.
- Denormalization helps cover up the inefficiencies inherent in relational database software.
- **Normalize until it hurts, denormalize until it works** :)

Denormalization

- Denormalization is the process of attempting to optimize the performance of a database by adding redundant data or by grouping data.
- Denormalization helps cover up the inefficiencies inherent in relational database software.
- **Normalize until it hurts, denormalize until it works** :)
- Star schema is a good trade-off between normalization and denormalization.

Multidimensional data model

- Retail sales data:

Location:Vancouver				
Time (quarters)	Items			
	TV	Computer	Phone	Security
Q1	605	825	14	400
Q2	680	952	31	512
Q3	812	1023	30	501
Q4	927	1038	38	580

Multidimensional data model

- Similar information for other cities:

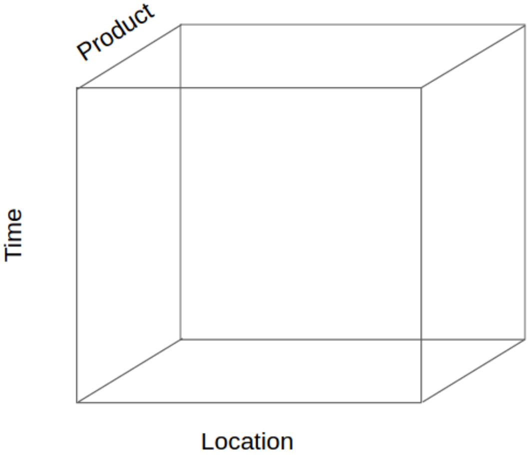
Location:Vancouver				
Time	Items			
(quarters)	TV	Computer	Phone	Security
Q1	605	825	14	400
Q2	680	952	31	512
Q3	812	1023	30	501
Q4	927	1038	38	580

Location:Toronto				
Time	Items			
(quarters)	TV	Computer	Phone	Security
Q1	1087	968	38	872
Q2	1130	1024	41	952
Q3	1034	1048	45	1002
Q4	1142	1091	52	984

Location:Chicago				
Time	Items			
(quarters)	TV	Computer	Phone	Security
Q1	854	882	89	623
Q2	943	890	64	698
Q3	1023	924	59	789
Q4	1129	992	63	870

Location:New York				
Time	Items			
(quarters)	TV	Computer	Phone	Security
Q1	818	746	43	591
Q2	894	769	52	682
Q3	940	795	58	728
Q4	978	864	59	784

Multidimensional cube



- More dimensions possible.

Different levels of aggregation

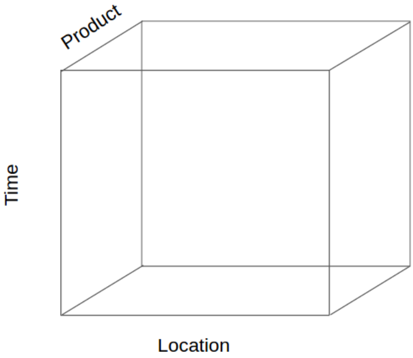
- Sales(time, product, *)

Time (quarters)	Items			
	TV	Computer	Phone	Security
Q1	3364	3421	184	2486
Q2	3647	3635	188	2817
Q3	3809	3790	186	3020
Q4	4176	3985	212	3218

- Sales(time, *, *); Sales(*, *, *)

Operators in multidimensional data model

- Roll up – summarize data along a dimension hierarchy.
- Drill down – go from higher level summary to lower level summary or detailed data.
- Slice and dice – corresponds to selection and projection.
- Pivot – reorient cube.
- Raking, Time functions, etc.



Exploring the cube

Time (quarters)	Items: TV	Computer	Phone	Security
Q1	3364	3421	184	2486
Q2	3647	3635	188	2817
Q3	3809	3790	186	3020
Q4	4176	3985	212	3218

$\Leftrightarrow$

Time		Items: TV	Computer	Phone	Security
Q1		3364	3421	184	2486
Q2		3647	3635	188	2817
Q3		3809	3790	186	3020
Q4	October	1172	960	105	1045
	November	1005	1340	45	987
	December	1999	1685	62	1186

ETL

- **ETL** = Extraction, Transformation, and Load

ETL

- **ETL** = Extraction, Transformation, and Load
  - **Extraction**: access heterogeneous data systems, identification of concepts and objects, change detection

ETL

- **ETL** = Extraction, Transformation, and Load
  - **Extraction**: access heterogeneous data systems, identification of concepts and objects, change detection
  - **Transformation**: cleansing of data, integration, transformation of a a useful format for analysis

ETL

- **ETL** = Extraction, Transformation, and Load
  - **Extraction**: access heterogeneous data systems, identification of concepts and objects, change detection
  - **Transformation**: cleansing of data, integration, transformation of a a useful format for analysis
  - **Load**: batch (bulk) load of data, checking integrity constraints, building of additional structures.

ETL

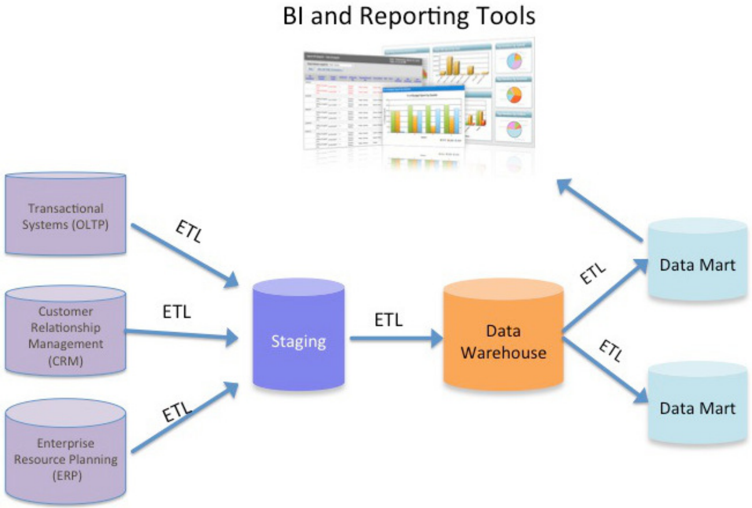
- **ETL** = Extraction, Transformation, and Load
  - **Extraction**: access heterogeneous data systems, identification of concepts and objects, change detection
  - **Transformation**: cleansing of data, integration, transformation of a a useful format for analysis
  - **Load**: batch (bulk) load of data, checking integrity constraints, building of additional structures.
- **Refreshment** of data warehouse.

ETL

- **ETL** = Extraction, Transformation, and Load
  - **Extraction**: access heterogeneous data systems, identification of concepts and objects, change detection
  - **Transformation**: cleansing of data, integration, transformation of a a useful format for analysis
  - **Load**: batch (bulk) load of data, checking integrity constraints, building of additional structures.
- **Refreshment** of data warehouse.
- Architecture of data warehousing:

  Data sources $\Rightarrow$ Data staging area $\Rightarrow$ Data warehouse

ETL



BI and Reporting Tools
Transactional Systems (OLTP)
Customer Relationship Management (CRM)
Enterprise Resource Planning (ERP)
ETL
ETL
ETL
Staging
ETL
Data Warehouse
ETL
ETL
Data Mart
Data Mart

Optimization of analytical systems

- Why analytical systems are so costly?

Optimization of analytical systems

- Why analytical systems are so costly?
  - An almost unconstrained number of possible queries.

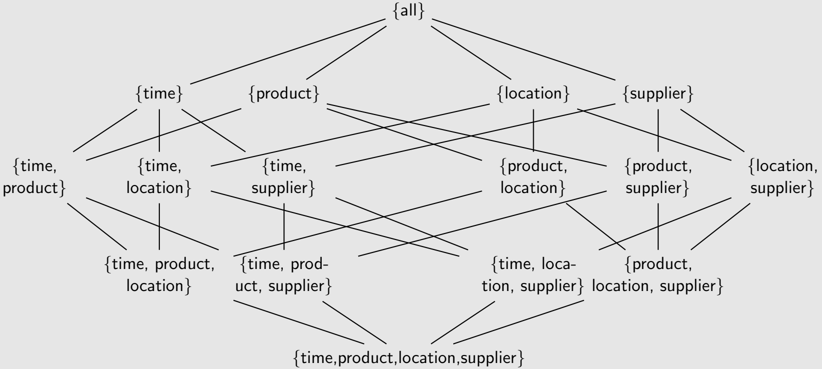
Optimization of analytical systems

- Why analytical systems are so costly?
  - An almost unconstrained number of possible queries.
  - Amount of data.

Lattice of cuboids

- Different degrees of summarizations are presented as a lattice of cuboids.

Example for dimensions: time, product, location, supplier



Using this structure, one can easily show roll up and drill down operations.

Total number of cuboids

- For an n-dimensional data cube, the total number of cuboids that can be generated is:
$$T = \prod_{i=1}^{n} (L_i + 1)\,,$$
where L_i is the number of levels associated with dimension i (excluding the virtual top level "all" since generalizing to "all" is equivalent to the removal of a dimension).
- For example, if the cube has 10 dimensions and each dimension has 4 levels, the total number of cuboids that can be generated will be:
$$T = 5^{10} = 9,8 \times 10^6\,.$$

Total number of cuboids

- **Example**: Consider a simple database with two dimensions:

Total number of cuboids

- **Example**: Consider a simple database with two dimensions:
  - Columns in `Date` dimension: `day`, `month`, `year`
  - Columns in `Localization` dimension: `street`, `city`, `country`.
  - Without any information about hierarchies, the number of all possible group-bys is

Total number of cuboids

- **Example**: Consider a simple database with two dimensions:
  - Columns in `Date` dimension: `day`, `month`, `year`
  - Columns in `Localization` dimension: `street`, `city`, `country`.
  - Without any information about hierarchies, the number of all possible group-bys is 2^6:

Total number of cuboids

- **Example**: Consider a simple database with two dimensions:
  - Columns in `Date` dimension: `day`, `month`, `year`
  - Columns in `Localization` dimension: `street`, `city`, `country`.
  - Without any information about hierarchies, the number of all possible group-bys is 2^6:

$\emptyset$		$\emptyset$
day		street
month		city
year		country
day, month	$\bowtie$	street, city
day, year		street, country
month, year		city, country
day, month, year		street, city, country

Total number of cuboids

- **Example**: Consider the same relations but with defined hierarchies:

Total number of cuboids

- **Example**: Consider the same relations but with defined hierarchies:
  - `day` $\rightarrow$ `month` $\rightarrow$ `year`
  - `street` $\rightarrow$ `city` $\rightarrow$ `country`

Total number of cuboids

- **Example**: Consider the same relations but with defined hierarchies:
  - `day` $\to$ `month` $\to$ year
  - `street` $\to$ `city` $\to$ `country`
  - Many combinations of columns can be excluded, e.g., `group by day, year, street, country`.
  - The number of group-bys is then

Total number of cuboids

- **Example**: Consider the same relations but with defined hierarchies:
  - `day` $\rightarrow$ `month` $\rightarrow$ year
  - `street` $\rightarrow$ `city` $\rightarrow$ `country`
  - Many combinations of columns can be excluded, e.g., `group by day, year, street, country`.
  - The number of group-bys is then 4^2:

Total number of cuboids

- **Example**: Consider the same relations but with defined hierarchies:
  - `day` $\rightarrow$ `month` $\rightarrow$ year
  - `street` $\rightarrow$ `city` $\rightarrow$ `country`
  - Many combinations of columns can be excluded, e.g., `group by day, year, street, country`.
  - The number of group-bys is then 4^2:

$\emptyset$		$\emptyset$
year	$\bowtie$	country
month, year		city, country
day, month, year		street, city, country

Outline

Summary

- Significant difference between operational and analytical systems.
- Different data models dedicated to particular applications.
- NoSQL = “Not only traditional relational DBMS.”
- OLAP vs. OLTP.
- Multidimensional data model.
- Star schema.
- ETL process.
- Computational challenges in analytical systems.