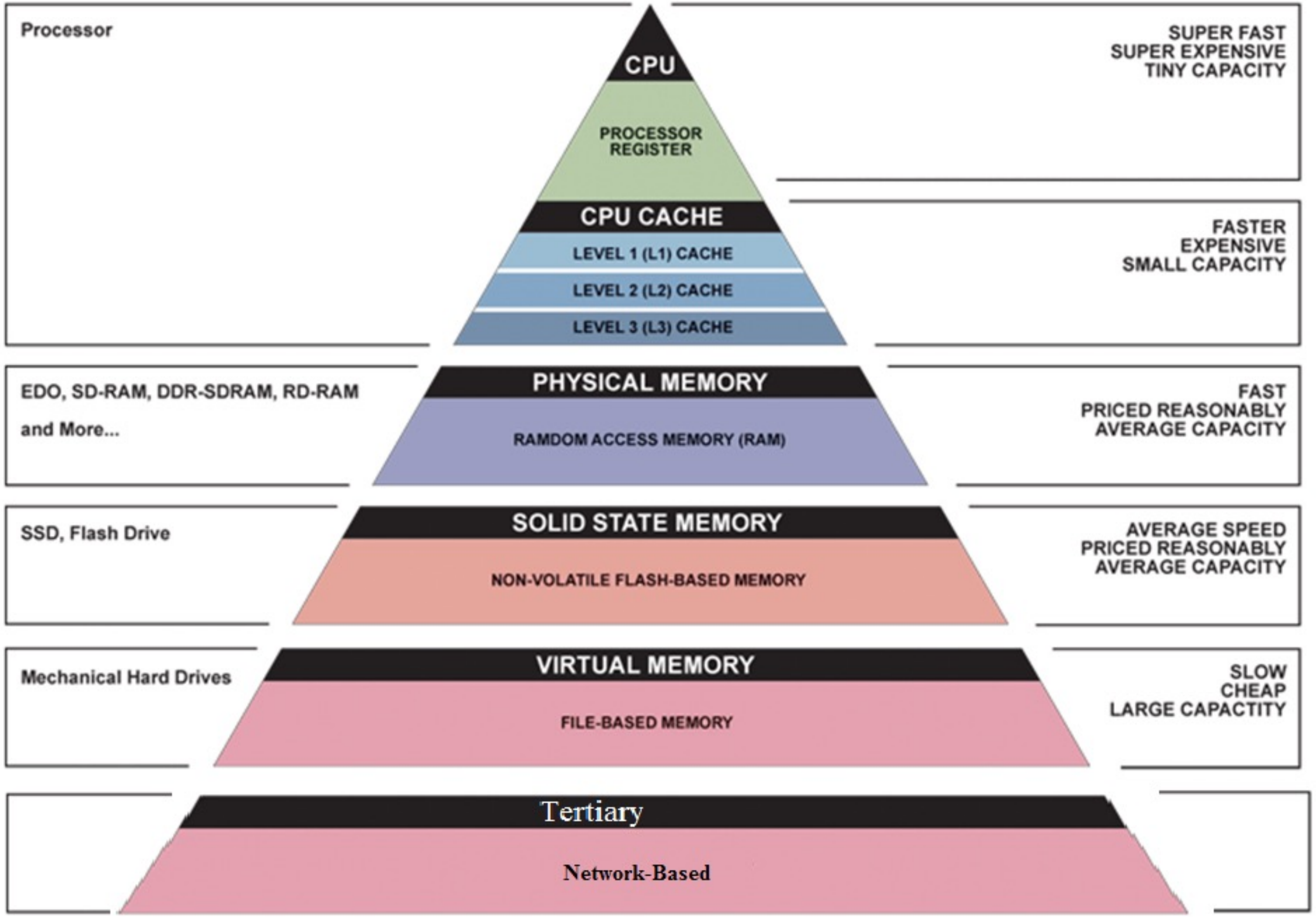
Introduction to Operating Systems and SQL for Data Science

Lecture 6 – Introduction to database management systems (DBMS)

Previous lecture

- Blocks & Files
- File structure – sequential & linked list
- File implementation – FAT & inode
- Folders
- File linking

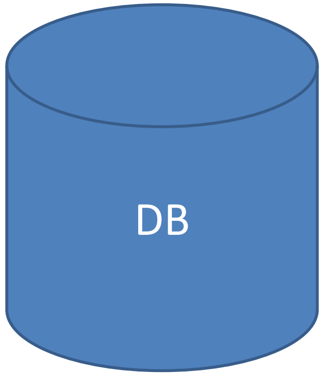
Levels of the Memory Hierarchy



Reichman University

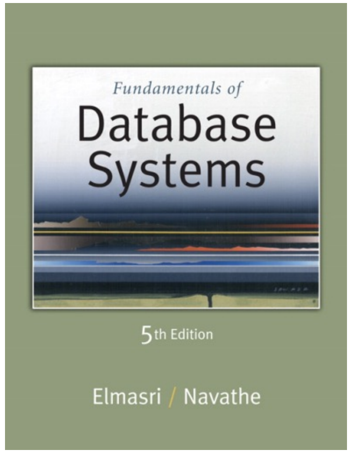
Database definition

Database is a collection of interrelated data which can be recorded and have implicit meaning.



Reading

Elmasri R. & Navathe S. Fundamentals of Database Systems (6th Ed.), Addison-Wesley, 2010.



Fundamentals of

Database Systems

5th Edition

Elmasri / Navathe

File system management

- At first, most users use the FSM of the OS.
- The data was store in files. For example, a student file or an academic staff file.
- Every file contains records, each record contains data for a single unit (e.g., a student).
- Each record could posses more than one attributes. For example: student name, student address, etc.

FSM – disadvantages (partial list)

1. Clumsy access to data:
   a. For instance, fetch “Avi Cohen” record from the Student file.

   b. This operation need to write a code that search the entire file.

   c. If we want to fetch all the student that are live in TLV we need to write another code.

   d. High decency in programmers, they become the “bottle neck” of developing analytics.

FSM – disadvantages (partial list)

2. No connection between different data:
   a. For example, a university data management. In this case, we expect for serval files such: Students, Courses, Lectures, Grades.

   b. There is no connection between the data that stored in different files. For instance, there is no prevention of adding a student in the grade file, which is no longer studying.

   c. Hence, the connection needs to be managed in a software layer and not in the data files.

FSM – disadvantages (partial list)

3. Can’t handle multiply requests

As we seen, the file system(FS) locked the file when reading it. Thus, two users can’t access the same file together. For instance, the grade file.

4. Hard to maintain changes, between files.
5. Duplicate record.

And many more…

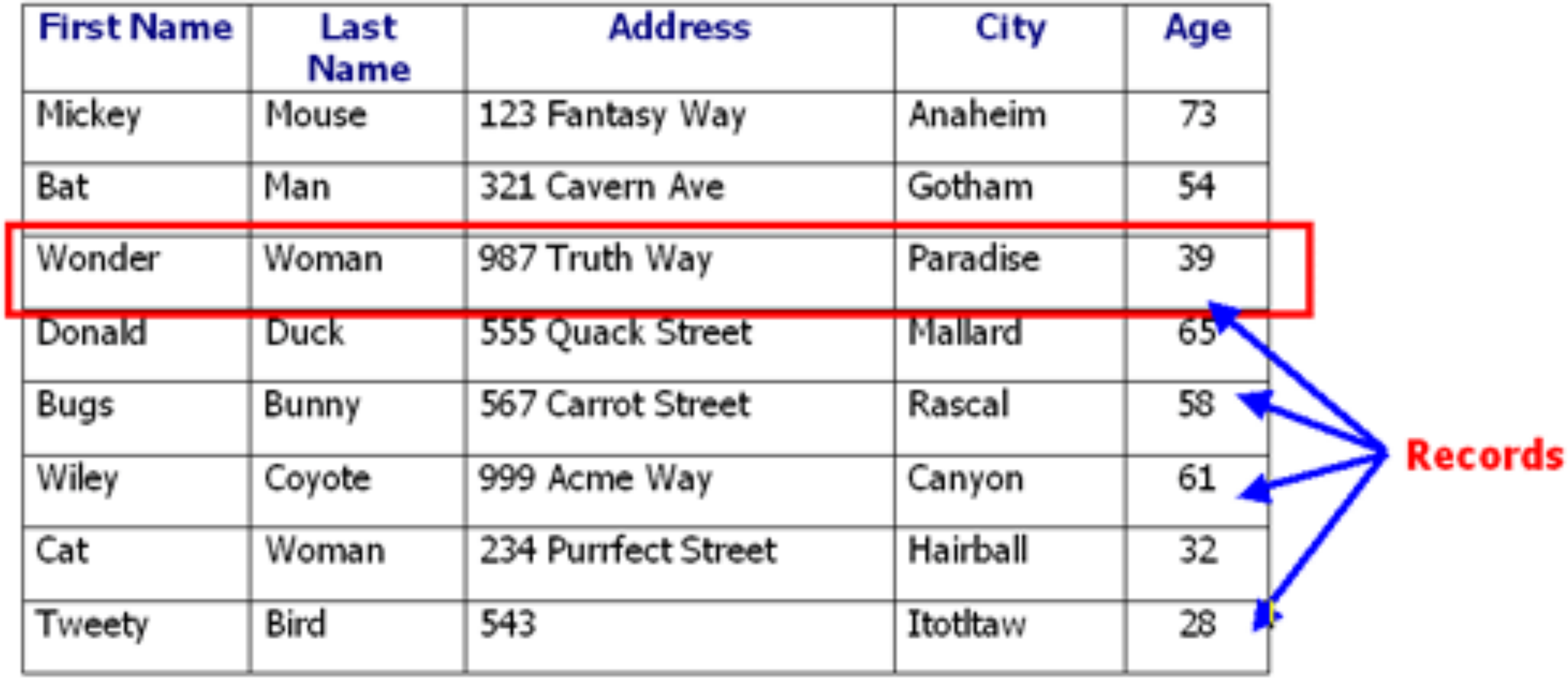
DBMS

Database management system

Definition: Database management system (DBMS) is a software system that support multiple users that support storage and access to a big amount of data in an efficient, easy and safe manner.

Record

Definition: A record is all the data or information about one entity.



First Name	Last Name	Address	City	Age
Mickey	Mouse	123 Fantasy Way	Anaheim	73
Bat	Man	321 Cavern Ave	Gotham	54
Wonder	Woman	987 Truth Way	Paradise	39
Donald	Duck	555 Quack Street	Mallard	65
Bugs	Bunny	567 Carrot Street	Rascal	58
Wiley	Coyote	999 Acme Way	Canyon	61
Cat	Woman	234 Purrfect Street	Hairball	32
Tweety	Bird	543	Itotltaw	28

Records

Huge data

The database need to save a huge number of records.

“Israel example” communication companies have new three milliards records that store the conversions with clients.

“Bezeq law” enforce companies to save records for seven years back.

Conclusion – Bezeq database need to store about 252 milliards records

What about AT&T?

The amount of data in the world

Total amount of data is measured today in units of zettabytes (10^{21} bytes = 1 billion terabytes).

World’s data more than doubling every two years.

Big data - Technologies being applied to big data include massively parallel processing (MPP) databases, distributed databases and cloud computing platforms,

Multiple users

The number of users or processes that might access the database in the same time and even to the same record.

Multiple users - example

Jane @ ATM1: withdraw $100 from account #55

```
;get balance from database
if balance > 100 then
;balance := balance - 100
;dispense cash
;put new balance into database
```

John @ ATM2: withdraw $50 from account #55

```
;get balance from database
if balance > 50 then
;balance := balance - 50
;dispense cash
;put new balance into database
```

Initial balance = 400
Final balance = ??

Safe

1. From a bug in the system:

   Computers are always break down – what happen if in time that a user is pick up money from an ATM, the database shutting down? For this, scenario we need the ability to go “back in time” – Rollback.

2. From malicious users:

   Hackers penetrating the database and pull or damage the data. Thus, we need authentication mechanism that only authorized users can access the data; and monitoring system to monitor any suspicious attempts.

Easy

1. Simple commands:
   a) To declare a database.
   b) To handle records – add, update, remove.
   c) Fetch records.
   d) Handle access.

We will learn SQL (Structured Query Language) language that help us achieve all of this.

SQL language is the dominate language in the DMBS market.

Efficient

- The operations needs to happen fast.

- For instance, a user pick up money from an ATM don’t need to wait more than a few seconds.

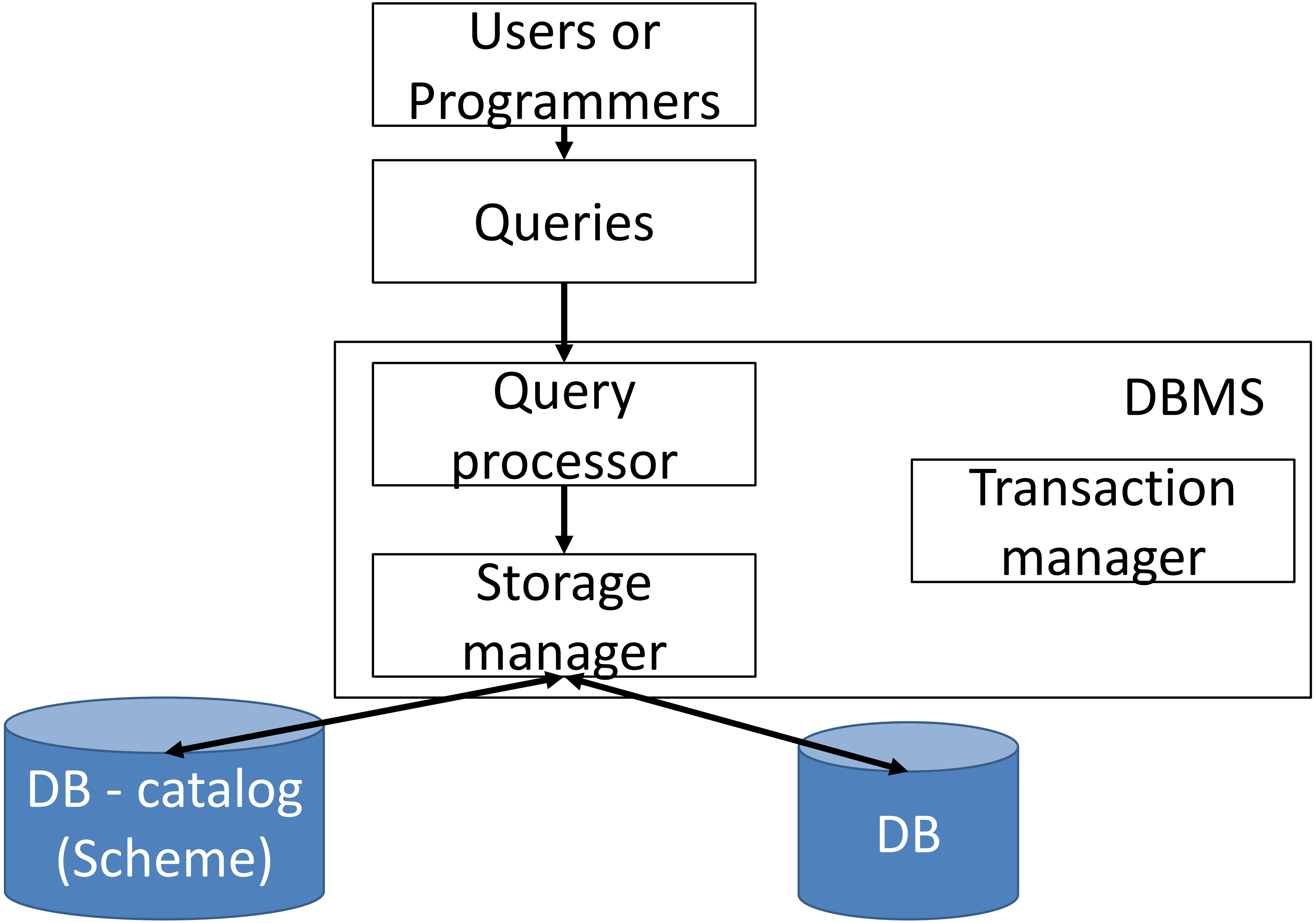
Software

Don’t forget, a DMBS is a software that like any other software needs to install. There are DBMS for pc, workstations and central computers.

Simple operations on DBMS

- **Define** – define and create appropriate settings of the components and type of the data. The structure and the dependences/rules/principles that apply on the data. The scheme of the data base.
- **Create** – to make a first save of the data in the DBMS.
- **Manipulate** – to make operations like push/pull/update records in work time – with the help of queries.

DBMS environment



Reichman University

How users interact with the DBMS

- Users are accessing the DBMS with query language or code executions.

- The differences between the database to data catalog is:
  - The databases stores the data.
  - The catalog stores the setting of the database, aka database schema like:
    - Files structure
    - Columns type
    - Connections between records/files
    - Metadata – data on the database

DMBS architecture

- Query processor – translate the queries from an high language a lower language that the machine understands.

- Storage manager – works with the OS and update the files in the disk.

- Transaction manager – responsible on the queries order. Like the example of the ATM.

How shall we handle transactions?

- Atomicity – all of the transaction runs or not.

For example, if we move money from one bank account to another, we can’t remove the money from the first account without adding it to the second one.

- Consistency – The data needs to be consistent. For example, we can’t have two people in the same seat in a flight.

How shall we handle transactions?

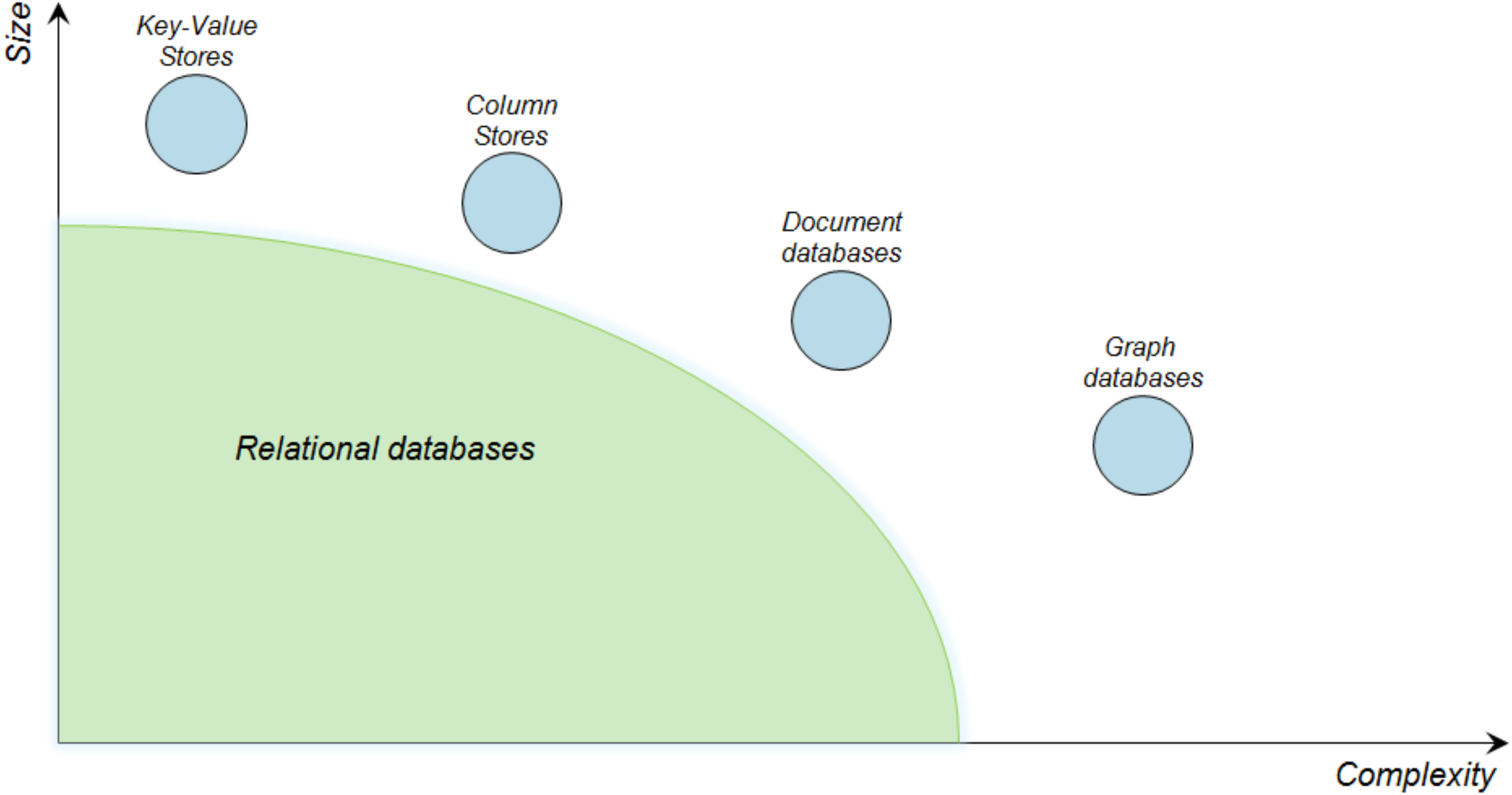
- Isolation – if two transaction are running together, they need to be isolated, so they won’t interfere each other.

- Durability – if the transaction ended successfully, then the change needs to reflect in the DBMS even if its shutdown from any reason.

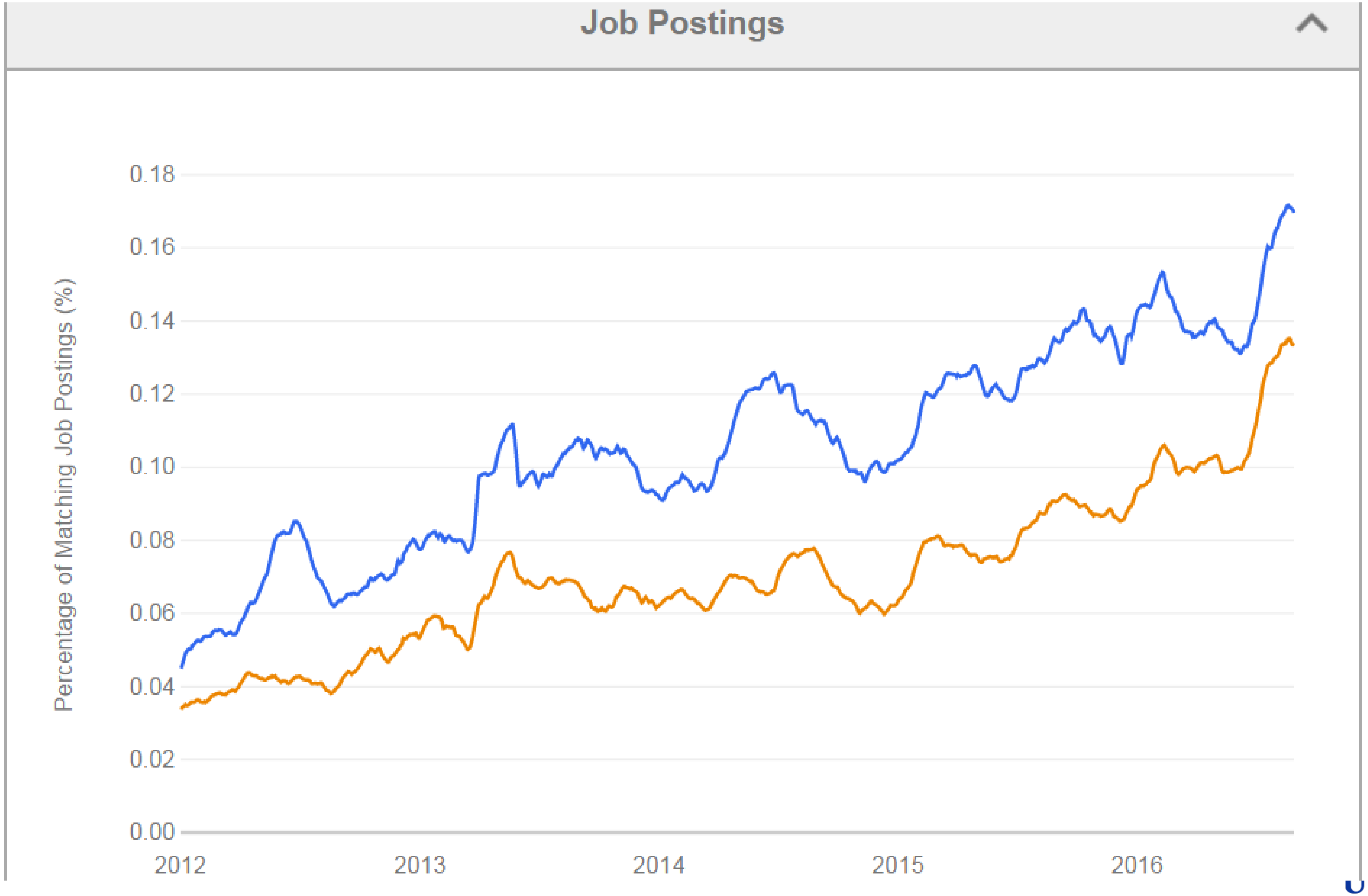
How do we choose a DMBS?

- Data model:
  - Tabular model (still relevant in today’s market)
  - hierarchical model
  - Object oriented model
  - NOSQL models
- Number of users
- Data storage
- Response time
- Hardware and OS
- Price

Different databases

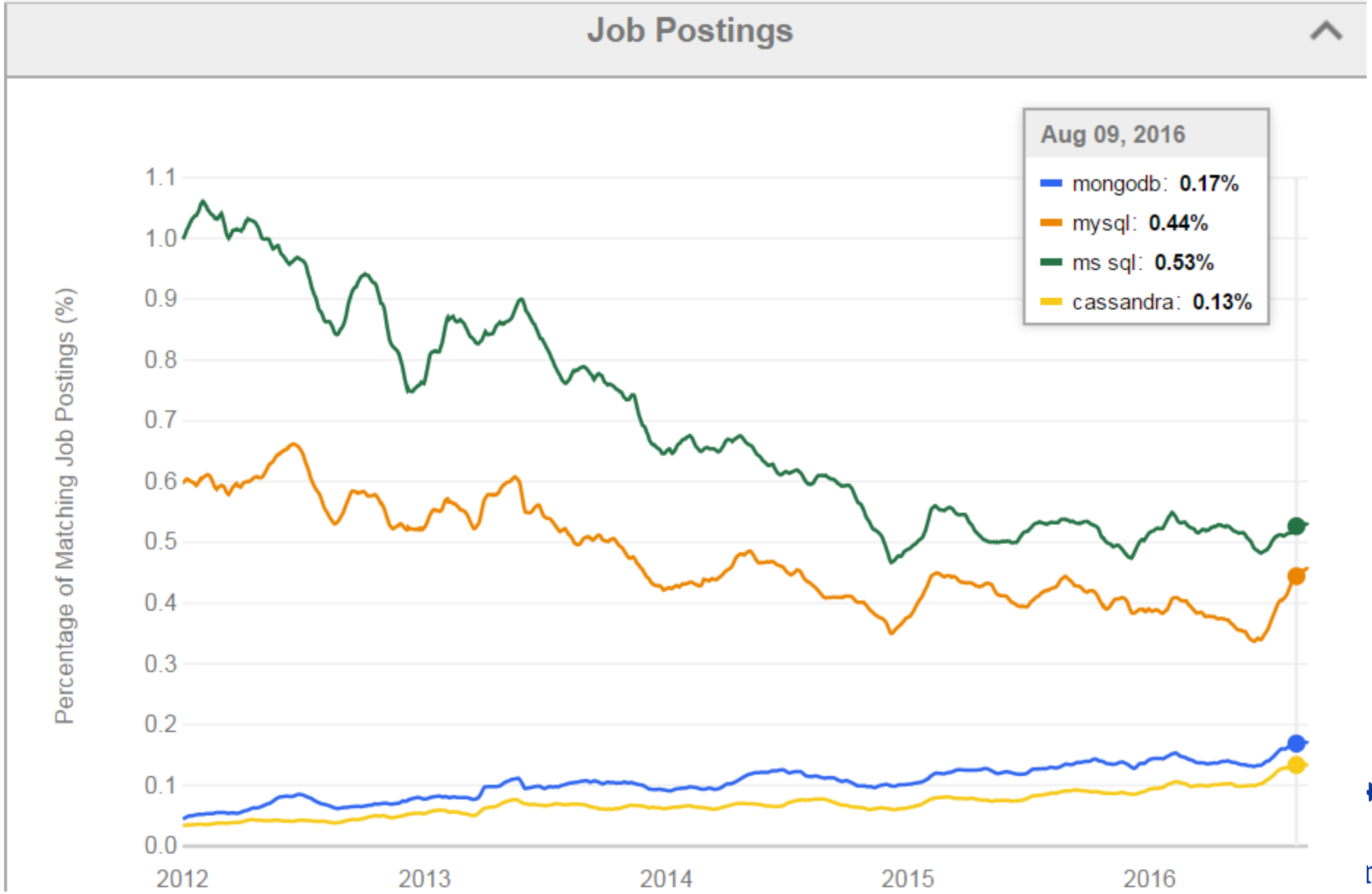


DB trends

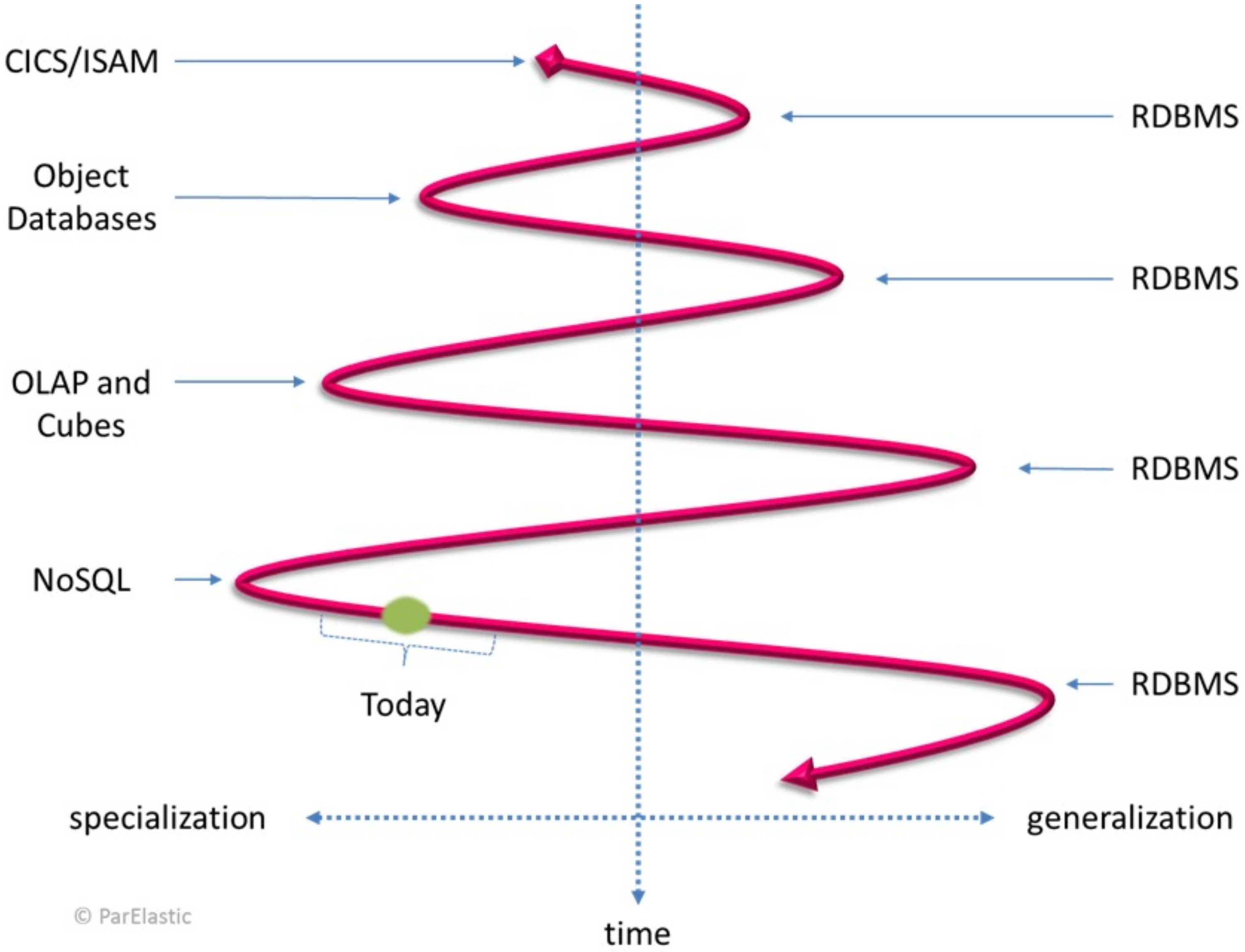


Reichman University

DB trends


Job Postings
Aug 09, 2016
mongodb: 0.17%
mysql: 0.44%
ms sql: 0.53%
cassandra: 0.13%
Percentage of Matching Job Postings (%)
1.1
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
2012
2013
2014
2015
2016

DB trends



© ParElastic

The main steps in open a database

1. Understanding the needs and requirements of the system.

2. Conceptual design
   - This stage represent the needs from the database.
   - The purpose of this step is to organize the data and the relationships between them the different tables in a way that be understood by a systems analyst.
   - This model is devoid of constraints arising from a particular DBMS system, hardware constraint or any other technological constraint.
   - We will learn conceptual design using the Entity Relationship Model (ERM).
   - We will also get to know the object-oriented (OO) model that is like the model known from programming languages.

The main steps in open a database

3. Logical design:

- Translate the conceptual model to this model.
- The logical model is design for a specific DBMS.
- The structure that we have in this stage called the database schema.
- We will learn the relational model.
- Most of the modern databases are based on the relational model.

The main steps in open a database

4. Physical design:

- Translate the database scheme to a physical structure of the database. We can build the logical model in many physical structures.

- The physical structure used for:
  - Shorten the response time
  - Take advantage of the storage volume.

- The physical design is taking into consideration, metrices like:
  - How many storage we need?
  - How to partition the disk?
  - Which index should we define? Indexes improve how we pull data, but overboard the time of updates.

The main steps in open a database

4. Implement applications that use the DB.

5. Preparation for production/usage.

6. DB maintenance, like backups in parallel to the day-to-day usage.

Important to remember!

The process is an iterative process, so moving to the next stage could affect previous stages.

Databases roles

- System analyst:
  - Responsible of stages - 1, 2.
  - Participate in stage 3.

- Software engineering:
  - Responsible of stages - 3, 4.

- DBA:
  - Responsible of stages – 4,6,7.

- Users:
  - Naïve / parametric users – use the database from an application, while the DB is transparent to them

Operation types in DBMS

- DDL – Data Definition Language
  - Operations to define the database scheme.
  - This is intended to a DBA or database designer – the one that builds the database of the application.

Operation types in DBMS

- DML – Data Manipulation Language
  - Operations on the data
  - Perform search / retrieve operations and update operations (add; update; delete).
  - Intend for programmers and “advance” users.
  - We sometimes distinguish between:
    - High level DML like SQL
    - Low level DML like for programmers to the specific host (aka host language)
    - Usually, we won’t distinguish between the DDL and the DML. And the DBMS do it for us.

Stored procedure

- Allow us to implement algorithms with DML.
- Contain control flow rules like: IF, WHILE, LOOP, REPEAT, and CASE statements.

Database system	Implementation language
DB2	SQL PL or Java
Microsoft SQL Server	Transact-SQL and various .NET Framework languages
MySQL	own stored procedures, closely adhering to SQL:2003 standard.
Oracle	PL/SQL or Java
PostgreSQL	PL/pgSQL, can also use own function languages such as pl/perl or pl/php
Sybase ASE	Transact-SQL

DBMS system utilities

- Loading – loads data files (e.g., csv, json, xml), support data conversions.
- Backup – backup files on a different disks or tape disk.
- File Reorganization – reorganize the files of the DB.
- Performance monitoring – checking and monitoring the system, bottle necks in preforming.

Different ways of implementing DBMS

Hierarchical; Network ; Object orientated

Multiusers

Single user

Distributed

Centralized

Homogenous

Heterogenous