Large-Scale and Multi-Structured Databases

Introduction to the Course

Academic Year 2022-2023

Prof Pietro Ducange

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

UNIVERSITÀ DI PISA

CROSSLAB Innovation for industry 4.0

Who is Talking to You?

Pietro Ducange

- Born in Apulia, South of Italy
- Master Degree in Computer Engineering in 2005, University of Pisa
- PhD in Information Engineering in 2009, University of Pisa
- Post-doc Researcher 2009-2014, University of Pisa
- Associate Professor 2014-2019, eCampus University
- Associate Professor 2019-on going, University of Pisa

Pietro's Research Activity

Main Research Topic:

- Big Data Mining and Analytics
- Text Analysis
- Explainable Artificial Intelligence

Member of:
Cloud Computing, Big Data and Cyber Security Lab@DII:
https://crosslab.dii.unipi.it/cloud-computing-big-data-cybersecurity-lab

Publication Records:
https://scholar.google.it/citations?user=HCgZqXEAAAAJ&hl=it

The Course

Large Scale and Multi-Structured Databases

9 CFU-> 90 Hours

Program Degrees:

- M.Sc. in Artificial Intelligence and Data Engineering (1-2 Year)
- M.Sc. in Computer Engineering (1 Year)

Syllabus

Introduction and Motivations: Introduction to the Course, The Big Data Era, The Database Revolutions

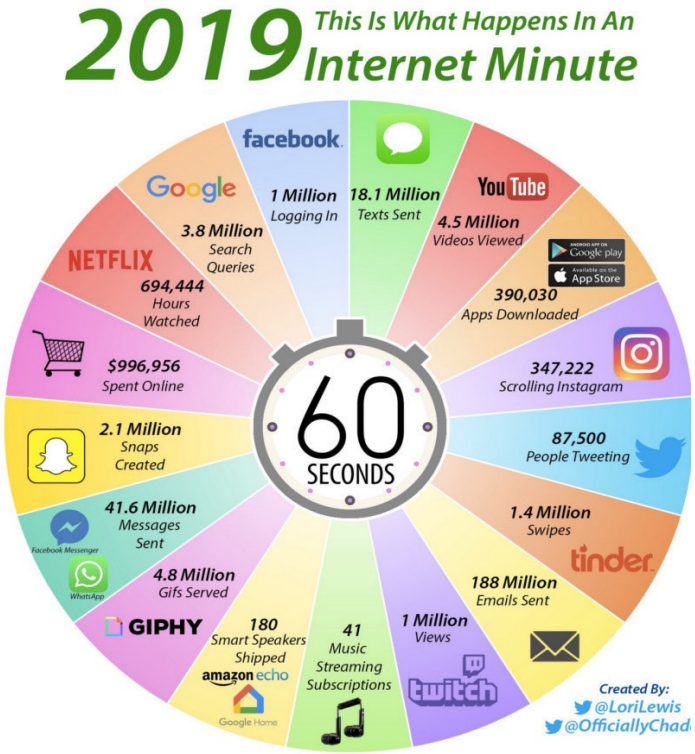
Fundamentals and properties of the NoSQL databases: ACID vs BASE properties, The Cap Theorem, Scalability, Sharding, Replication, Consistency

Architectures of NOSQL databases: Document Databases, Key-values Databases, Column Databases, Graph Databases

Recaps: Recap of Software Engineering and Java (basics, connections and queries towards SQL databases).

Modern Infrastructures for NoSQL Databases: LevelDB/REDIS, MongoDB, Neo4J (Installation, configuration, CRUD operations, main queries)

The Big Data Era



DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

UNIVERSITÀ DI PISA

The Data Base Revolutions

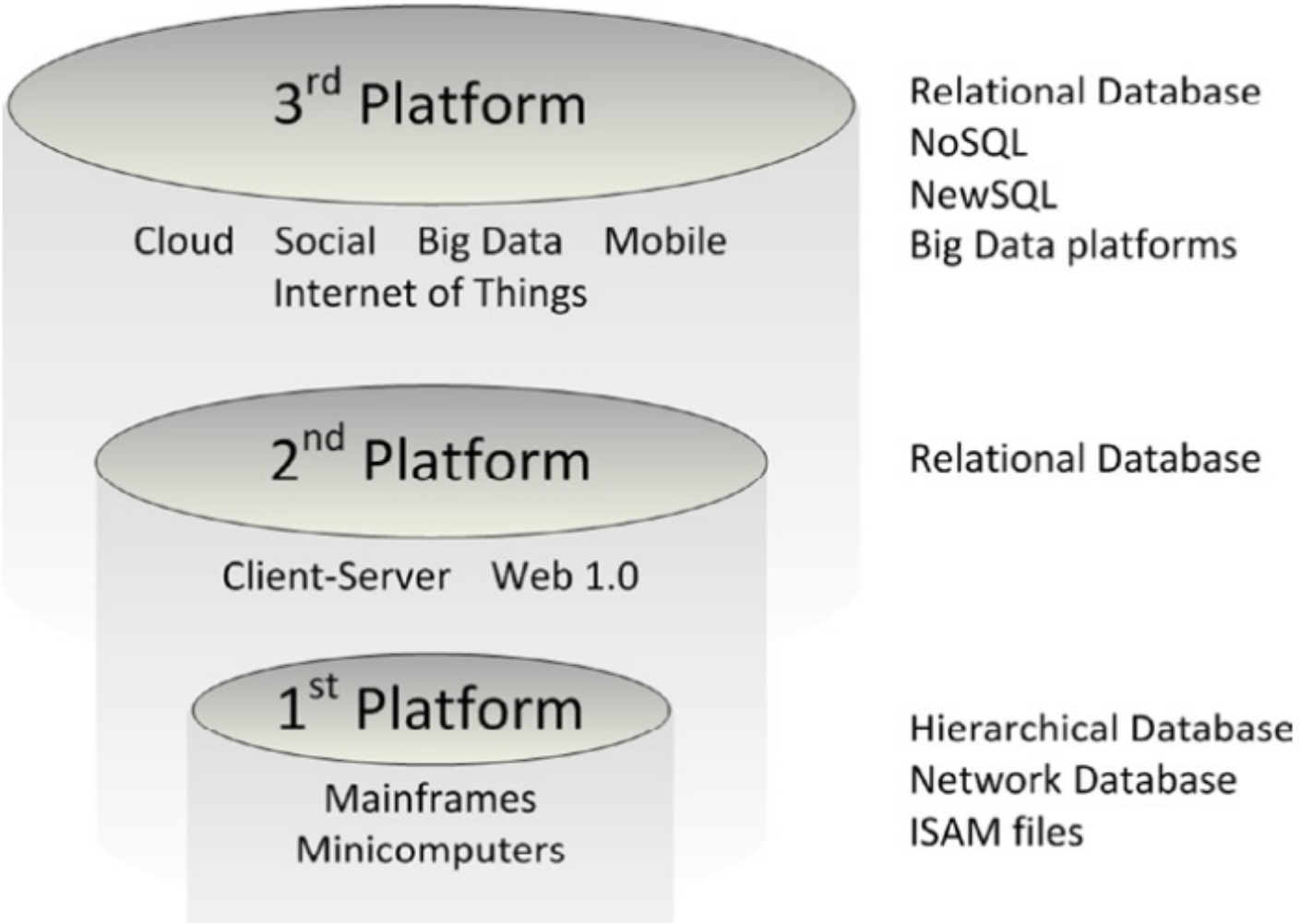


Image extracted from "Guy Harrison, Next Generation Databases, Apress, 2015"

ACID vs BASE

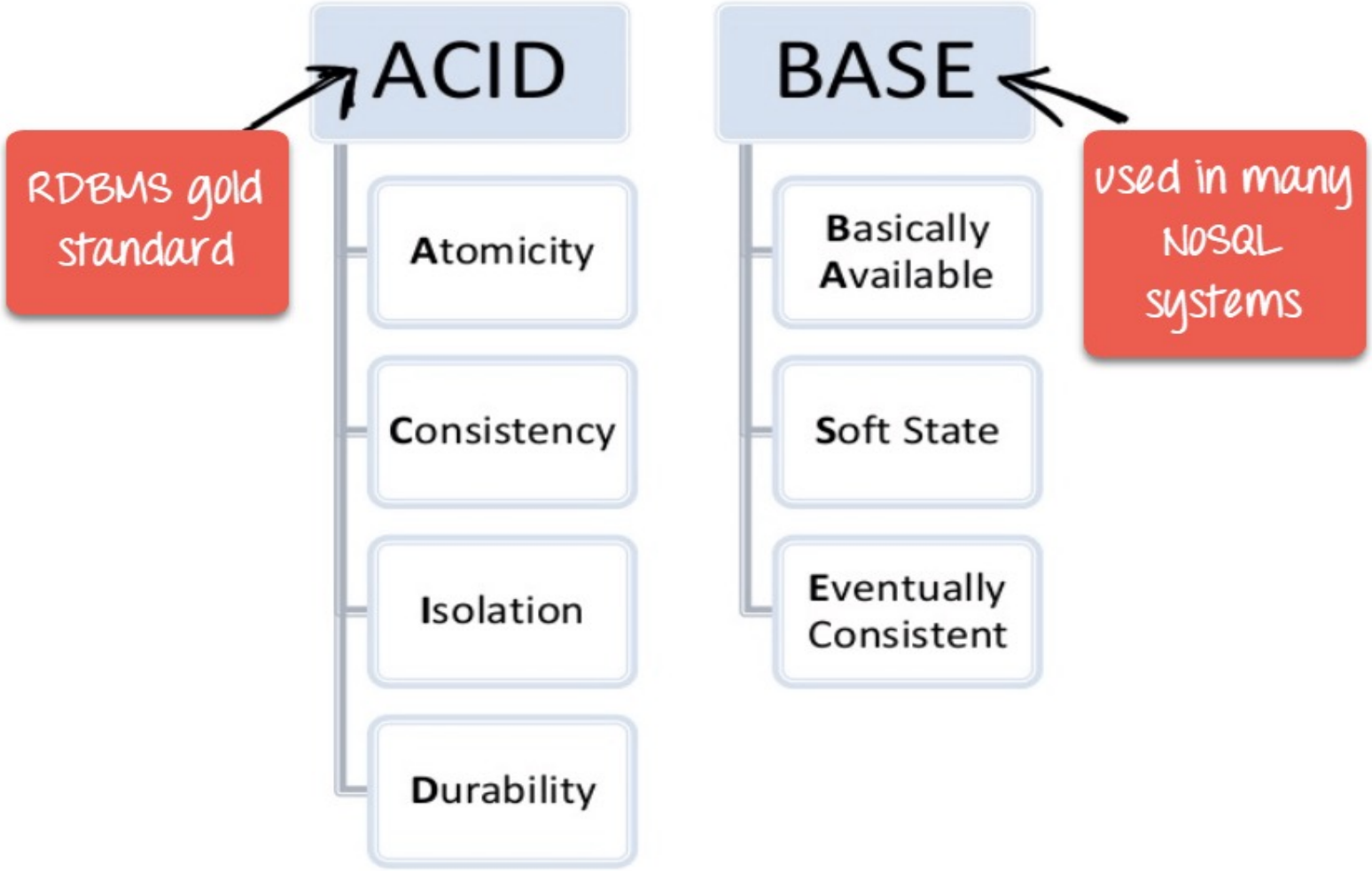


Image extracted from: https://www.guru99.com/sql-vs-nosql.html

Key-Value Databases

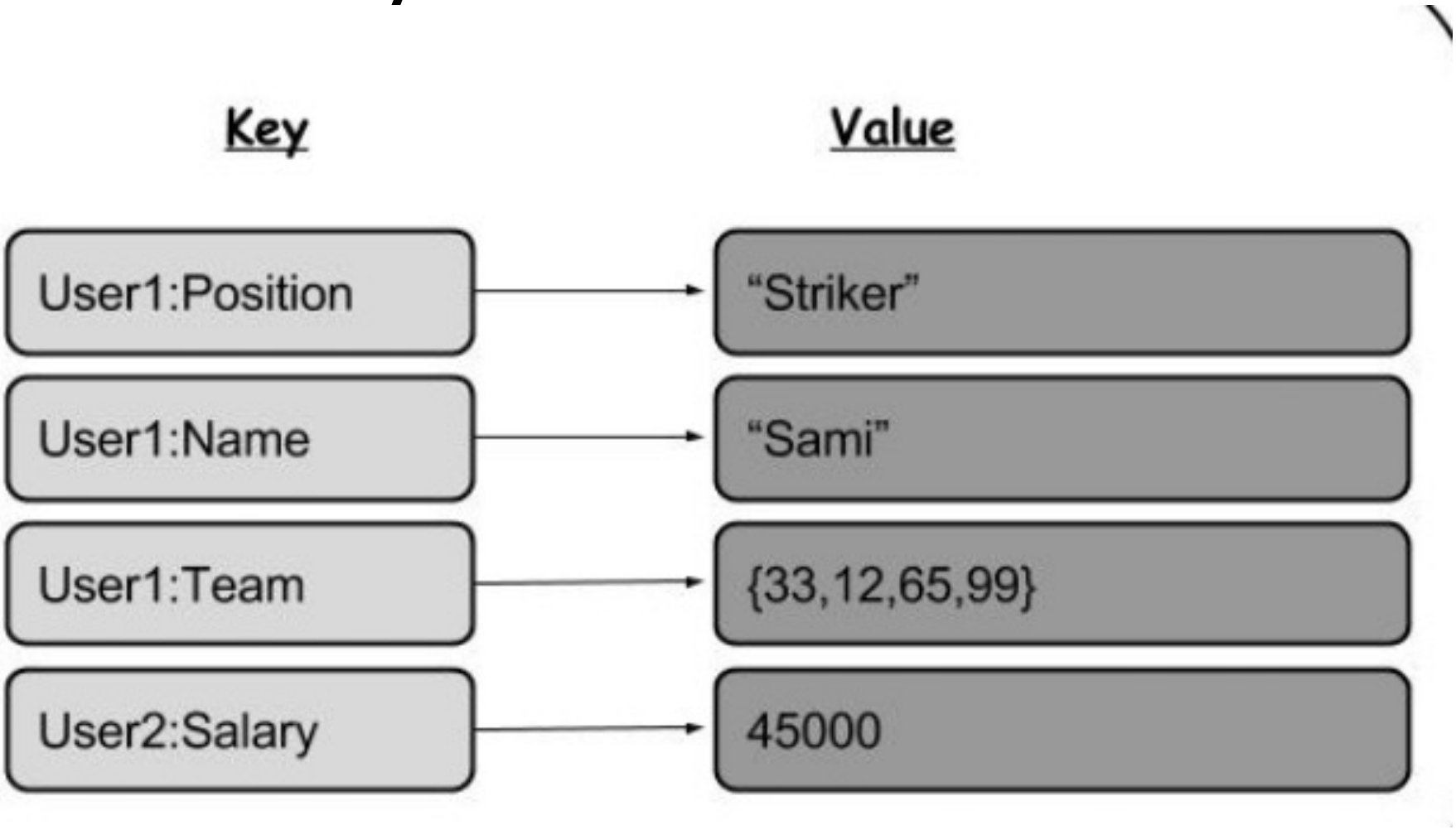
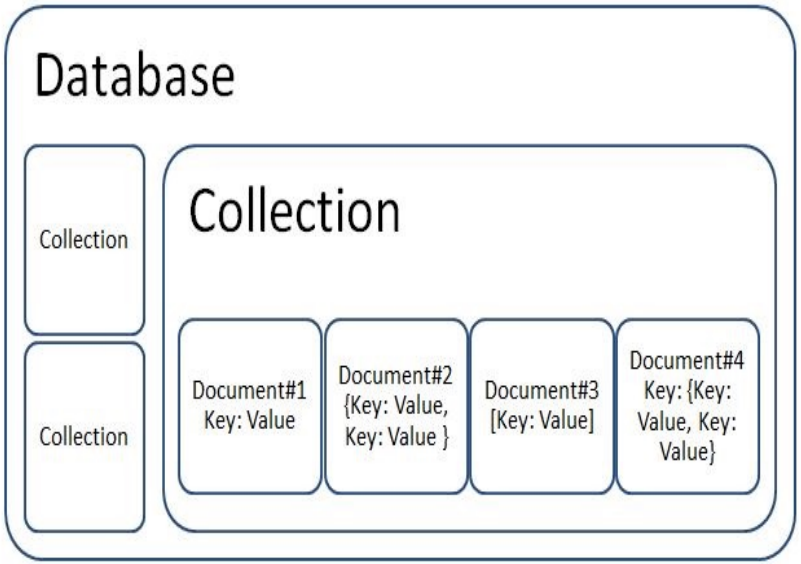
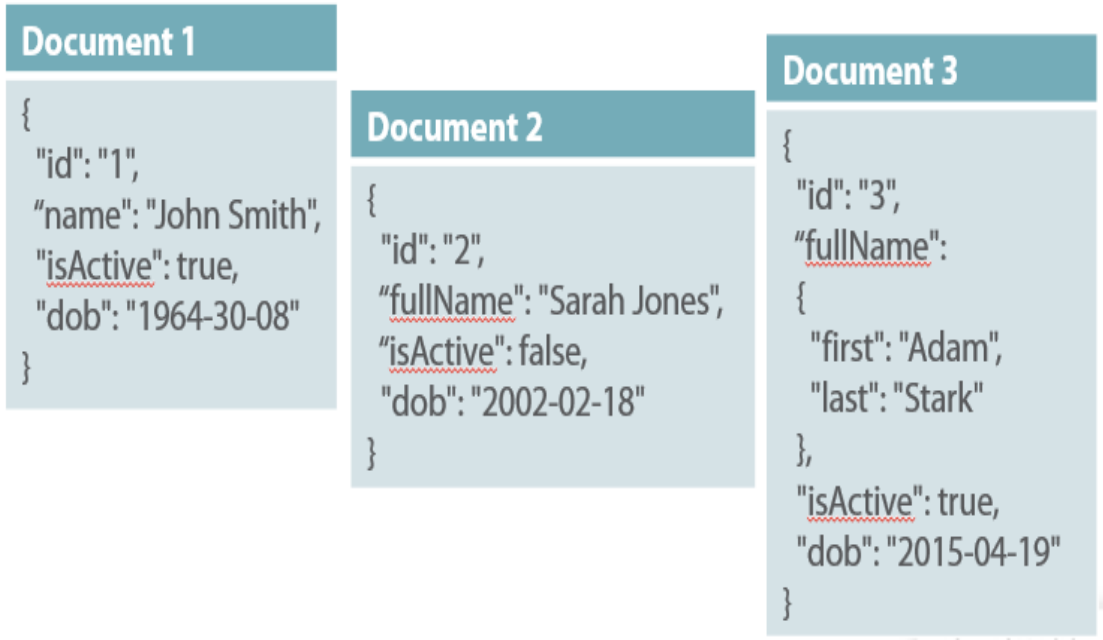


*Image extracted from:*https://www.researchgate.net/figure/Key-value-NoSQL-Database_fig1_332188615

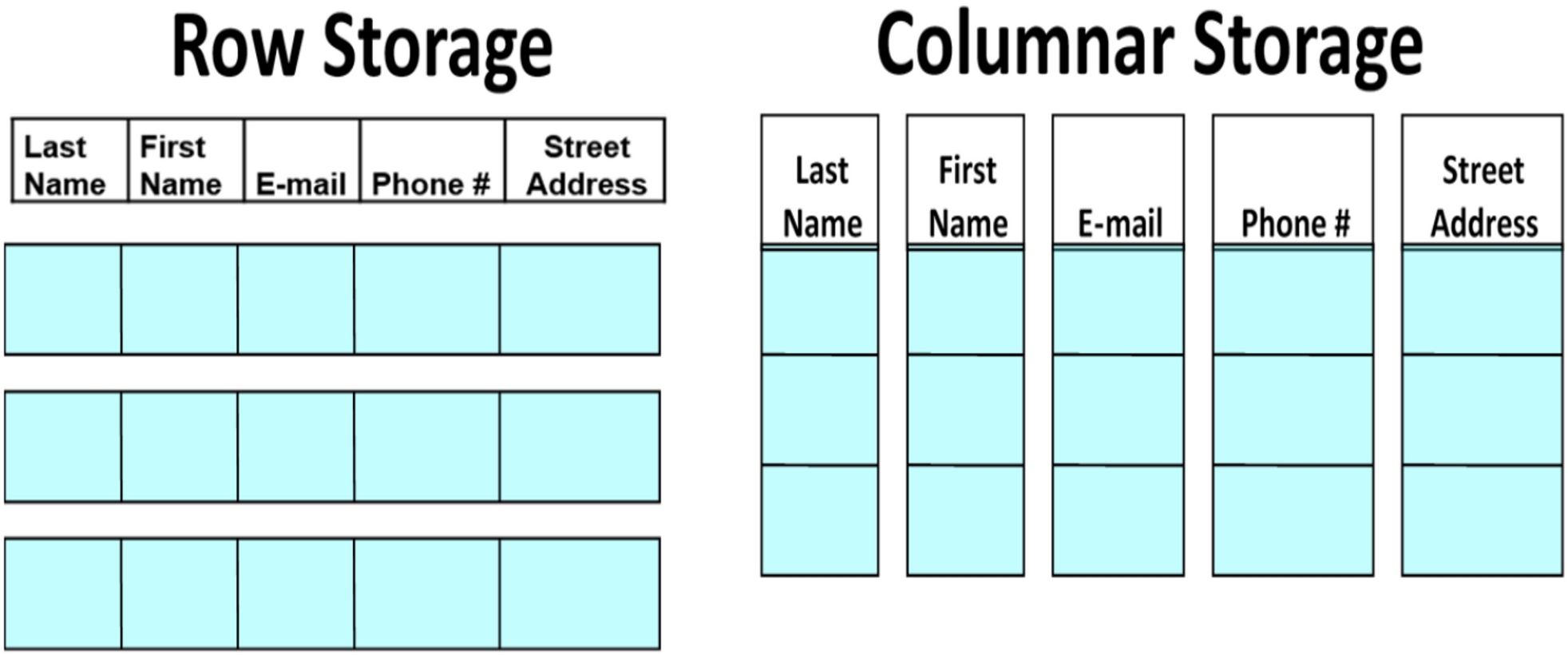
Document Databases





Images extracted from: https://dzone.com/articles/a-primer-on-open-source-nosql-databases
https://lennilobel.wordpress.com/2015/06/01/relational-databases-vs-nosql-document-databases/

Column Databases



Graph Databases

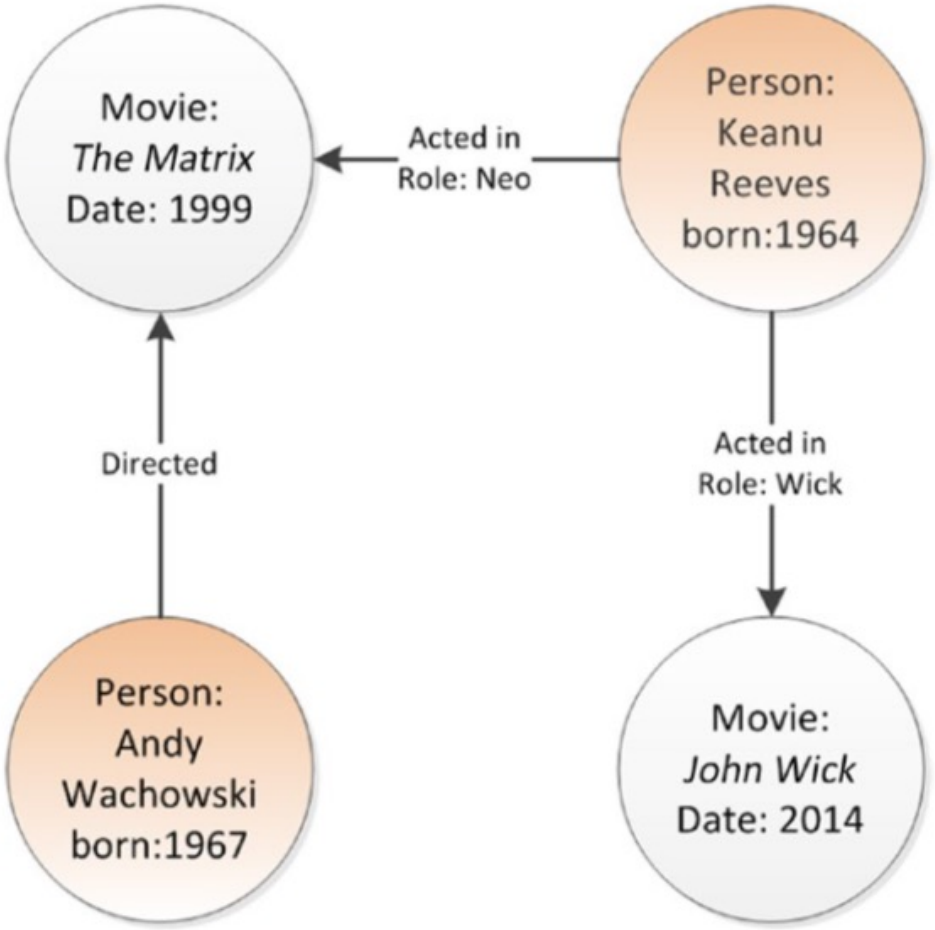


Image extracted from "Guy Harrison, Next Generation Databases, Apress, 2015"

Recaps Software Engineering

- Functional and non-functional requirements
- Use cases definitions
- UML Diagrams
- Some exercises on designing a complete application

Recaps on Java

- Hello World! In Java using an IDE (Eclipse/IntelliJ)
- Some Java Programming Exercises
- Connection to MySQL server: JDBC (using Maven for handling dependencies)

Modern NoSQL Infrastructures

Type	Example
Key-Value Store	redis, riak, amazon DynamoDB
Wide Column Store	HBASE, cassandra
Document Store	mongoDB, CouchDB relax
Graph Store	Neo4j, InfiniteGraph The Distributed Graph Database

Learning Outcomes: *Knowledge*

At the end of the course:

- The student will have acquired knowledge about the ***tools*** and ***methodologies*** for the design of ***non-relational databases***.
- The student will acquire knowledge about the ***architectures***, ***performances*** and ***costs*** of modern infrastructures for the management of complex data.
- The student will be able to correctly ***set up*** a ***project*** for the management of multi-structured and large data, integrating it into a ***real computer application*** and choosing in an appropriate manner the design and implementation strategies.

Assessment Criteria of Knowledge

Group activities will be proposed to assess theoretical and practical knowledge.

Group activities will be proposed to the ***working groups*** with the objective of:

- ***deepening*** of theoretical and technical issues
- ***implementing*** of technical projects

Periodic classroom discussions between the teacher and the group of students developing the above activities will be organized.

Skills

At the end of the course the student will be able to:

- ***Design*** a non-relational database based on the ***requirements*** (functional and non-functional) of a specific ***application***.
- Use modern ***technological infrastructures*** for the management of non-relational databases (LevelDB, Redis, MongoDB, Neo4j, etc.)

Assessment Criteria of Skills

During lab class:

- The student will be shown how to ***install*** and ***configure*** some of the modern technological infrastructures for the management of non-relational databases.
- ***Practical activities*** will be proposed for the creation, management and querying of different non-relational databases.
- Group activities will be proposed for the ***in-depth study*** of technical issues and for the implementation of educational projects.

Prerequisites

- Programming in ***JAVA*** (including the use of an IDE)
- Design and query of ***relational databases***
- Basics of ***Software Engineering*** (including realization of UML Diagrams)
- Basics of ***Unix-Based*** Operating Systems

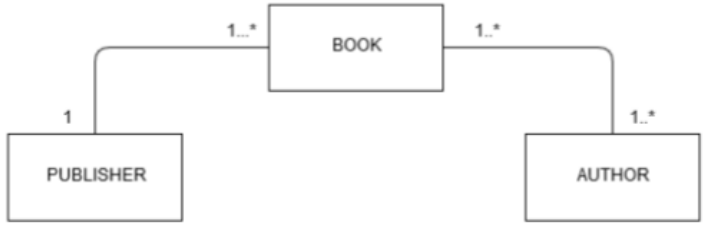
About Software Engineering Skills

1.1 Functional requirements

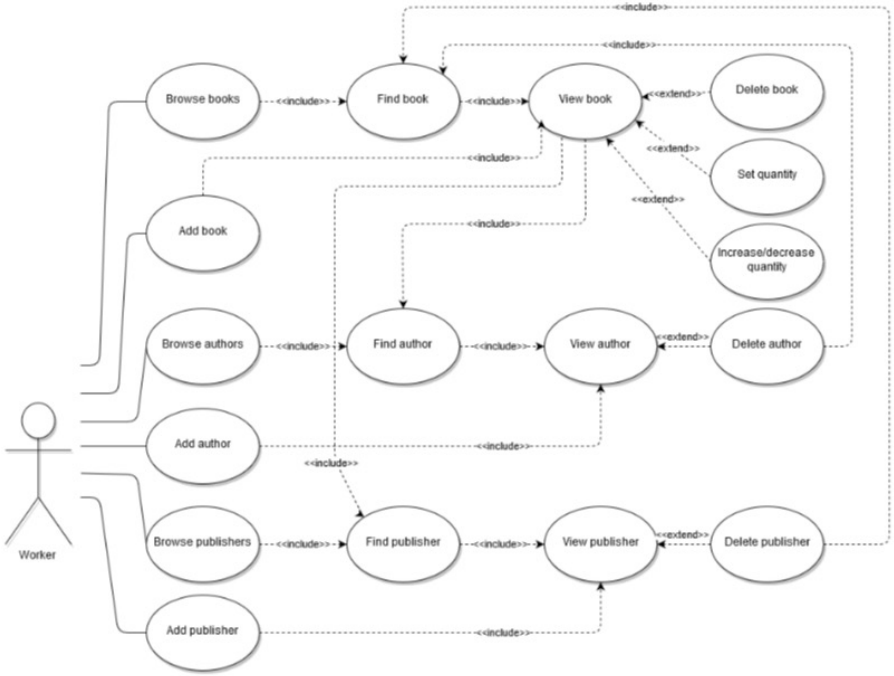
- The system has to allow the user to add and delete a book and modify its quantity;
- The system has to allow the user to add and delete a publisher;
- The system has to allow the user to add and delete an author.

1.2 Non functional requirements

- Performance: The software has to be fast to avoid delays when a customer asks for a book;
- Integrity: The data integrity is crucial to avoid to give wrong information to the customers;
- Usability: The application must be user friendly and intuitive to be easily used by the workers.



UML Analysis Class Diagrams



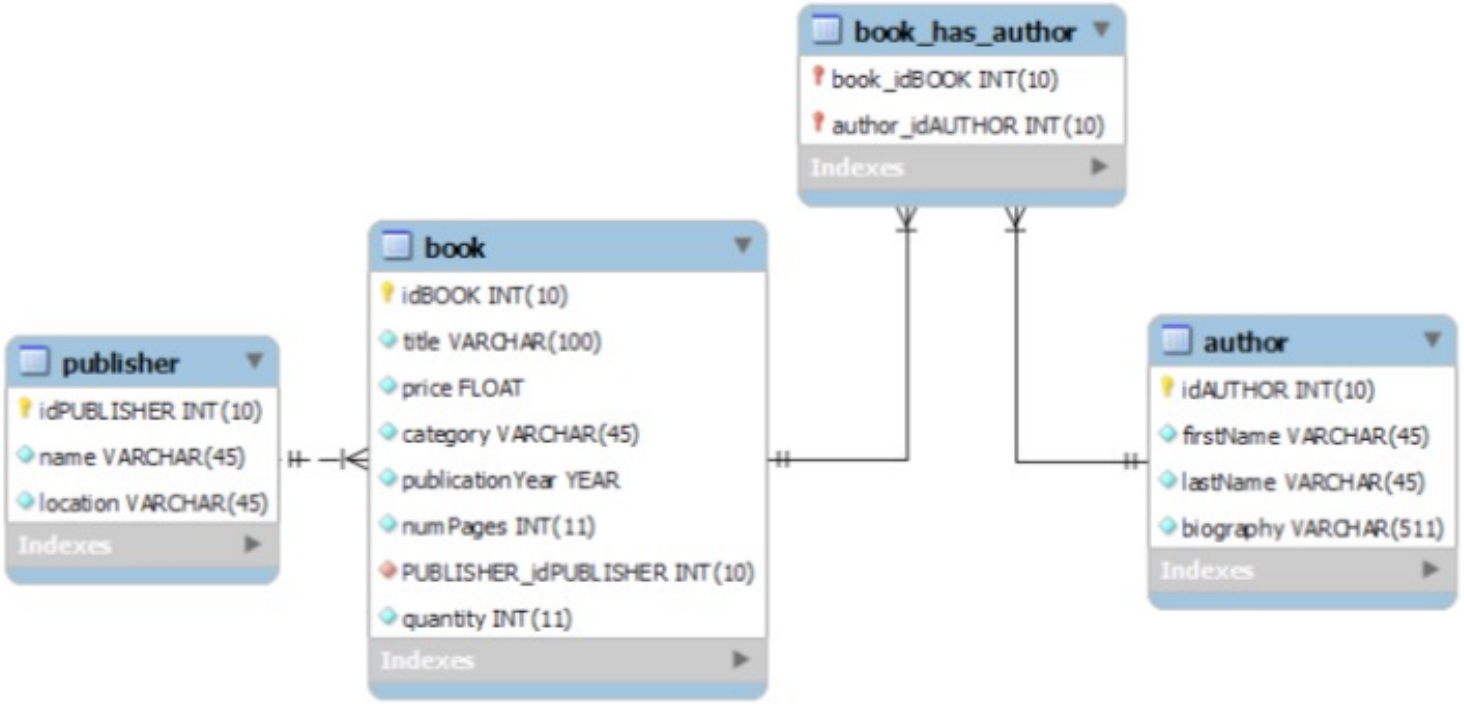
UML Use Case Diagrams

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

UNIVERSITÀ DI PISA

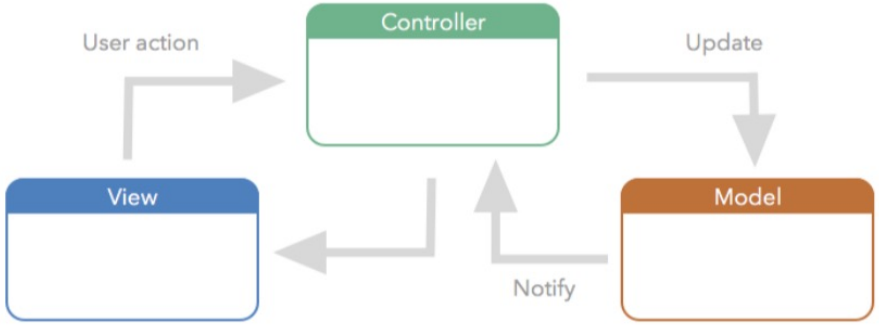
CROSSLAB Innovation for industry 4.0

About Relational Databases

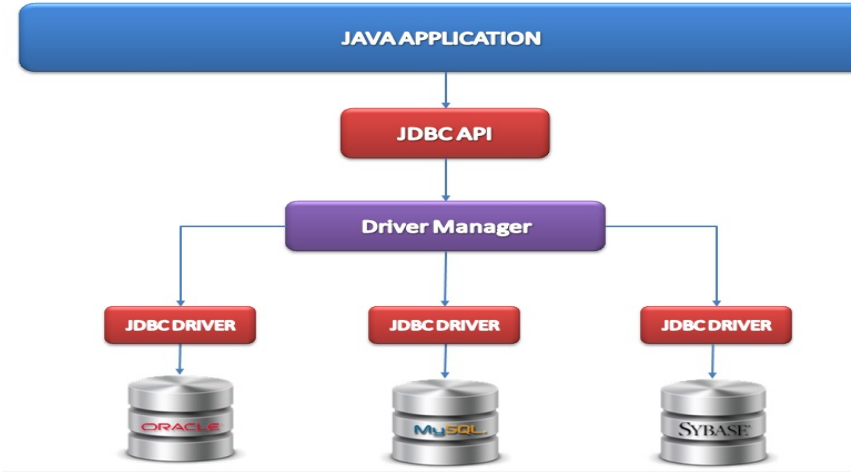


E-R Diagram

About (JAVA) Skills



MVC Model



Client-Server Application Interacting with Data Bases

Teaching Method

The course will be held in ***face to face!***

The teacher will provide in advance (hopefully) the slides used during the lessons (with suggestions to the book chapter to be read).

Video recordings of the last-year classes may be provided.

The course will be held entirely in ***English***.

Tutoring hours (two hours per week) will be provided each ***Monday (REMOTE MODE)*** (17:00-19:00) ***by request***.

Book your meeting (MANDATORY) with teacher here: shorturl.at/fgjFL

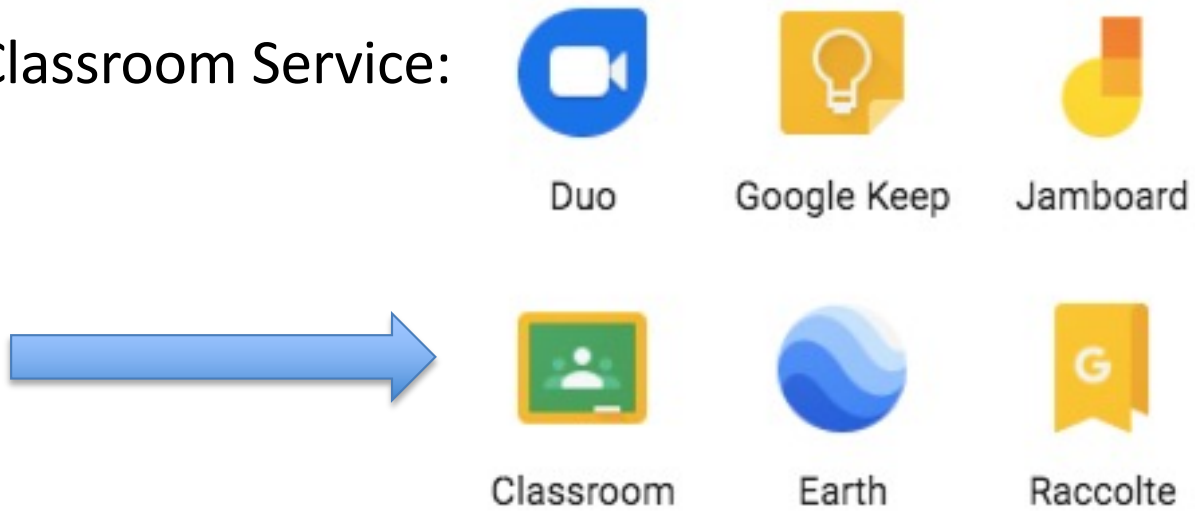
The teacher will be available after the streaming class for a ***Q&A session.***

The E-learning Platform

We will exploit the Google GSuite service for the activities related to the course (materials, tests, projects).

Each student can login to the service with his/her own UNIPI credentials (check details here https://start.unipi.it/en_GB/gsuite/).

Once logged in, select the Classroom Service:



From the + button Join a class (specify Class Code ***xu7bnya***).

Studying Materials

- Slides provided by the Teacher (almost self contained)
- Video of the 2022-2023 Classes
- Scientific articles provided by the teacher
- Official Documentation of the NoSQL DBMSs.
- Recommended Books:
  - “Guy Harrison, Next Generation Databases, Apress, 2015”
  - “Dan Sullivan, NoSQL For Mere Mortals, Addison-Wesley, 2015”
  - “Andreas Meier, Michael Kaufmann , SQL & NoSQL databases : models, languages, consistency options and architectures for big data management, 2019”

Check available books at: https://onesearch.unipi.it

THE HISTORY OF THIS COURSE…

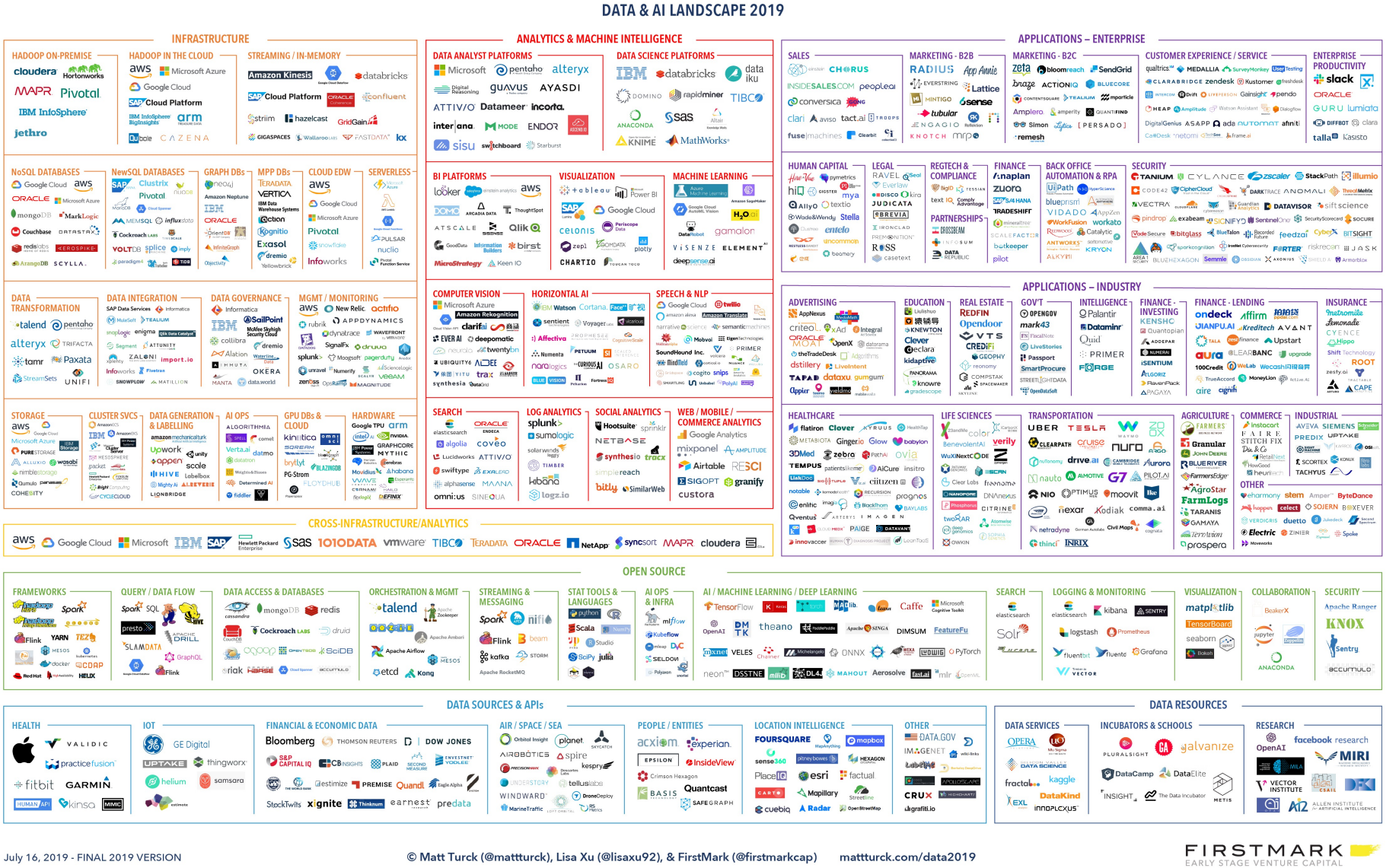
Emerging Paradigms and Technologies (EPT)

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

IN SUPREMÆ DIGNITATIS 1343
UNIVERSITÀ DI PISA

CROSSLAB
Innovation for industry 4.0

Issues with EPT



July 16, 2019 - FINAL 2019 VERSION

© Matt Turck (@mattturck), Lisa Xu (@lisaxu92), & FirstMark (@firstmarkcap)

mattturck.com/data2019

FIRSTMARK
EARLY STAGE VENTURE CAPITAL

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

UNIVERSITÀ DI PISA

CROSSLAB
Innovation for industry 4.0

Teaching EPT: Problem Based Learning

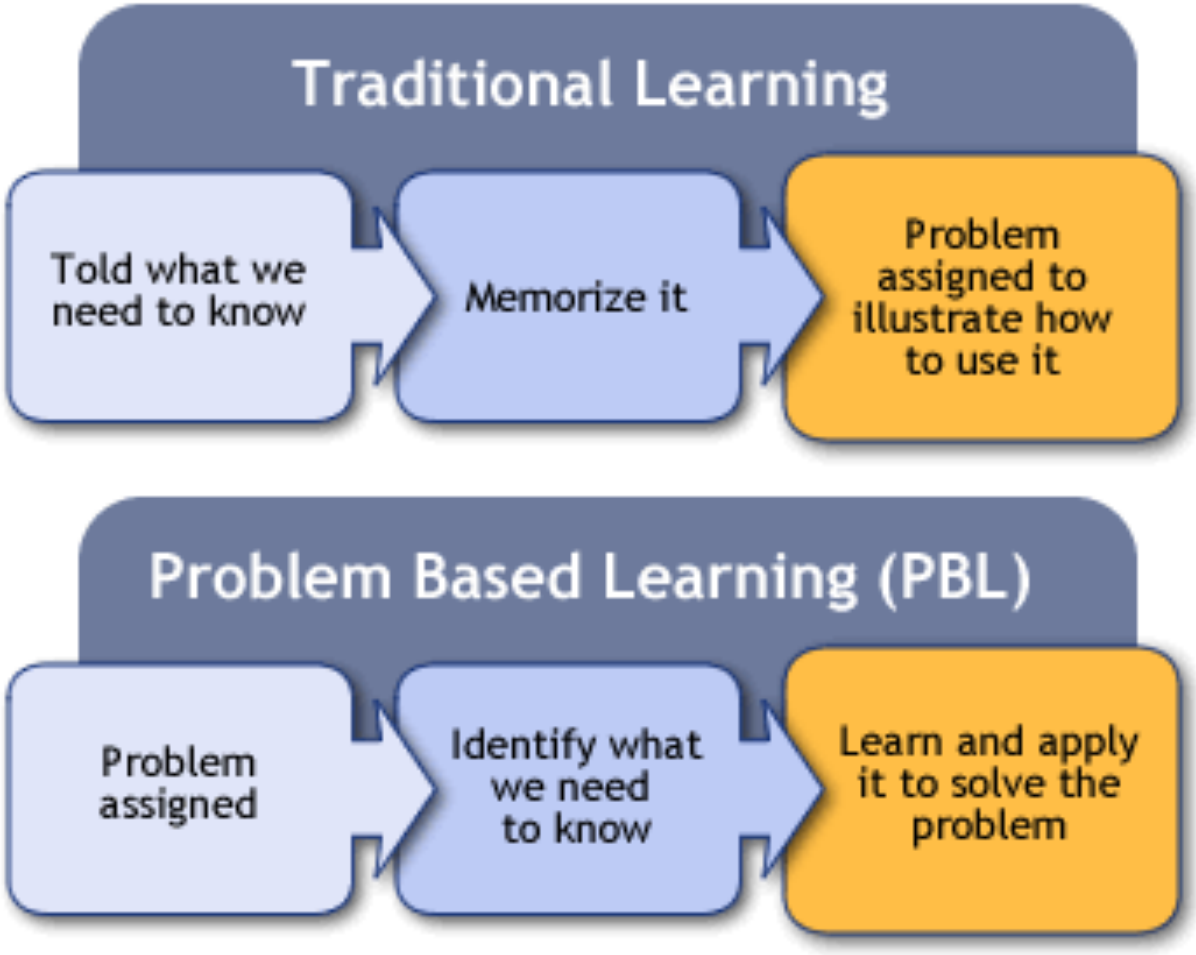


Image extracted from: https://gianfrancomarini.blogspot.com/2017/01/metodologie-didattiche-apprendimento.html

Teaching EPT: Blended Learning

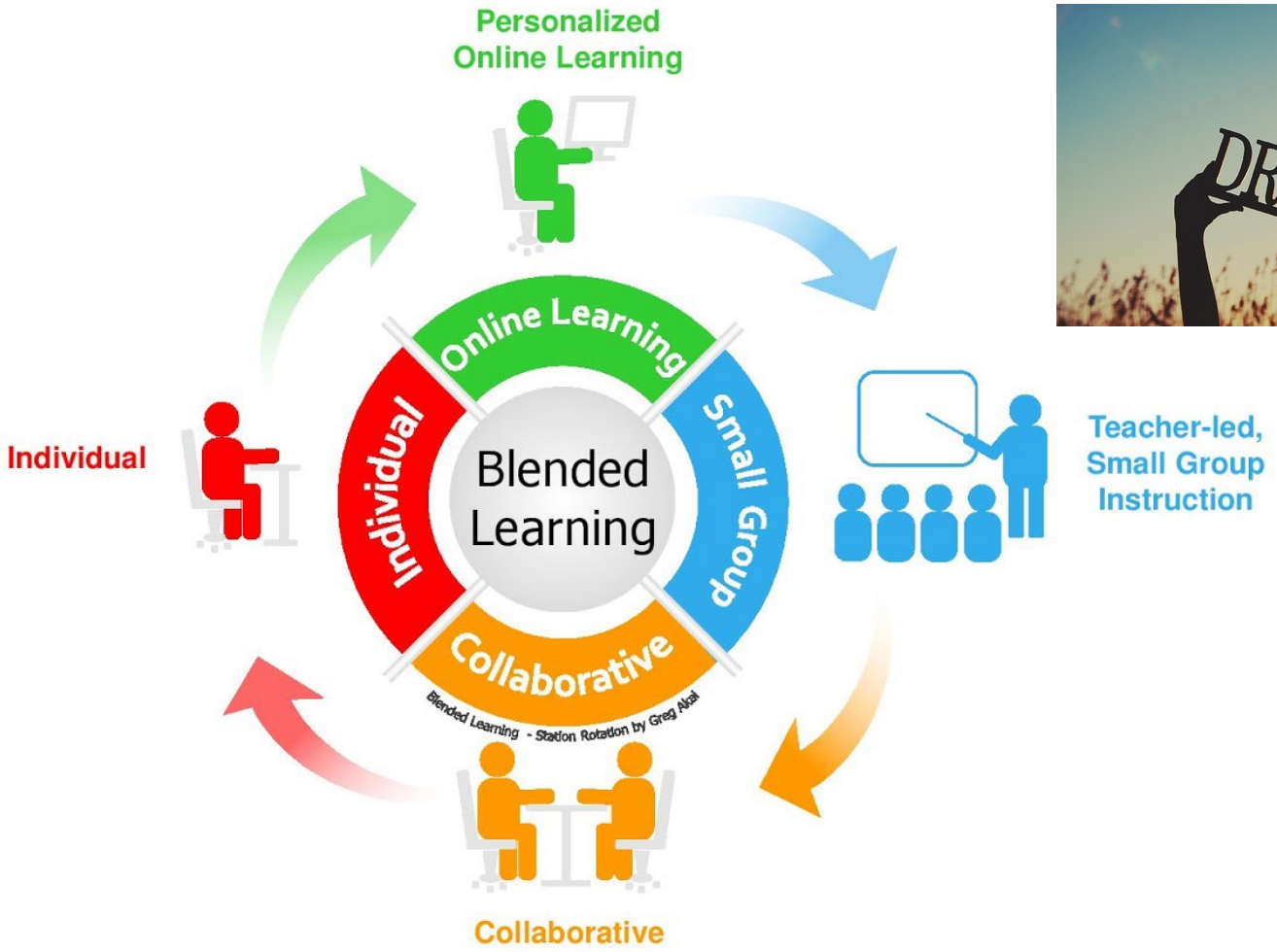


Image extracted from: http://esheninger.blogspot.com/2019/03/the-pedagogy-of-blended-learning.html

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

UNIVERSITÀ DI PISA

CROSSLAB Innovation for industry 4.0

2019-2020 A.Y. Experience

Course: *Large-Scale and Multi-Structured Databases*

Master Program: *Artificial Intelligence and Data Engineering*

Where: *University of Pisa*

Delivery mode: In presence with a Soft-Blended Environment online from September to December 2019

2020-2021 A.Y. Experience

Courses:
Large-Scale and Multi-Structured Databases
Data Mining and Machine Learning

Master Program: *Artificial Intelligence and Data Engineering*

Where: *University of Pisa*

Delivery mode: Full online from September to December 2020

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

UNIVERSITÀ DI PISA

CROSSLAB Innovation for industry 4.0

2020-2021 A.Y. Experience
Basic Structure of the Courses

- *Theoretical Classes*
- *Practical Classes*
- *Developing of a Group Project*

2020-2021 A.Y. Experience
The Proposed Approach

- Cooperative Group Activities
- Students were randomly divided into groups of 6/7 members (one coordinator)
- Each practical class, in which a cooperative group activity was assigned to each group, lasted 2 hours

Structure of the Practical Classes

- During the ***first 20/30 minutes***, usually the instructor introduced some ***new features*** of a specific tool or framework and showed some example of ***practical applications*** of methods and algorithms introduced during the theoretical classes.

- The ***outline*** of the ***activity*** was presented to the groups, including its ***objectives*** and expected ***results***.

- The ***students*** were allowed to ***collaborate*** in the groups for developing a solution of the activity. To this aim, each group created its own MS Teams channel.

- One or two groups were invited to ***share*** with the classroom the proposed solution for the activity.

- A final ***discussion*** with the classroom was ***moderated*** by the instructor.

2020-2021 A.Y. Experience Discussions

- During each cooperative activity, instructors perceived ***a high level of engagement*** of the students.
- First of all, the number of student attending the practical classes was ***more or less constant.***
- Their ***participation*** was very ***active*** and students requested frequently the support of the instructors while developing the activities in groups.
- Almost all the groups were able, in the different meetings scheduled for the practical classes, to ***share*** their solutions and to ***stimulate*** a discussion in the classroom.

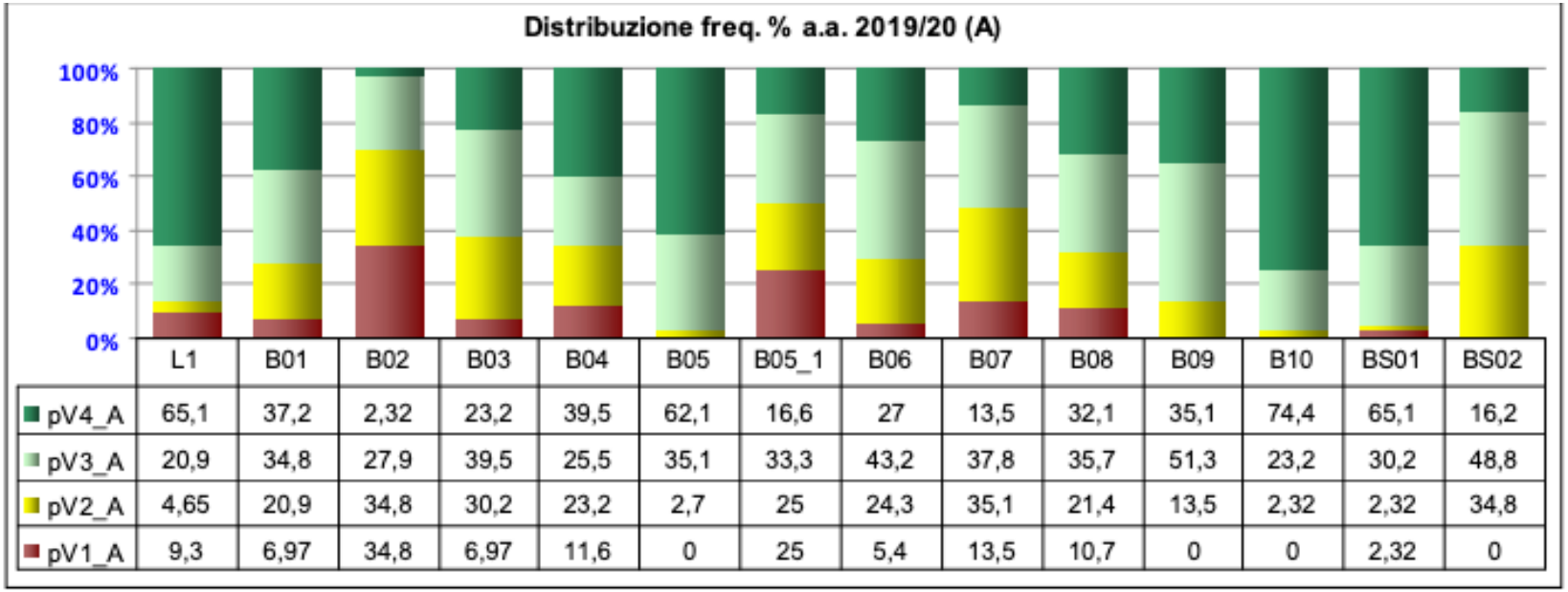
Positive Effects of the Cooperative Activities

- ***All the students*** were requested to develop a teaching project for each of the course.

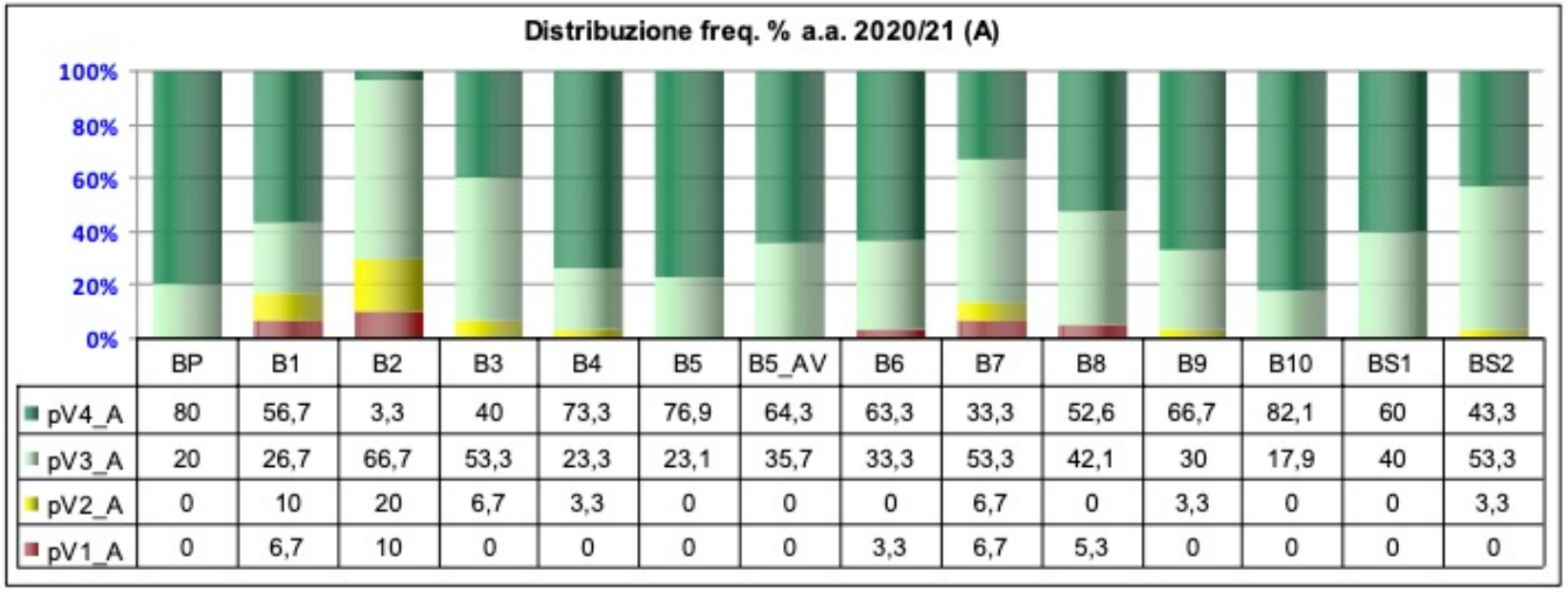
- The ***percentage*** of students attending the course who ***passed the exam*** after the first exam session was equal to ***85.42%.***

- ***Last year (2019)***, when the course were delivered in face-to-face mode without considering cooperative group activities, this percentage were equal to ***45.8%.***

2020-2021 A.Y. Experience: Course Evaluation



	L1	B01	B02	B03	B04	B05	B05_1	B06	B07	B08	B09	B10	BS01	BS02
pV4_A	65,1	37,2	2,32	23,2	39,5	62,1	16,6	27	13,5	32,1	35,1	74,4	65,1	16,2
pV3_A	20,9	34,8	27,9	39,5	25,5	35,1	33,3	43,2	37,8	35,7	51,3	23,2	30,2	48,8
pV2_A	4,65	20,9	34,8	30,2	23,2	2,7	25	24,3	35,1	21,4	13,5	2,32	2,32	34,8
pV1_A	9,3	6,97	34,8	6,97	11,6	0	25	5,4	13,5	10,7	0	0	2,32	0



	BP	B1	B2	B3	B4	B5	B5_AV	B6	B7	B8	B9	B10	BS1	BS2
pV4_A	80	56,7	3,3	40	73,3	76,9	64,3	63,3	33,3	52,6	66,7	82,1	60	43,3
pV3_A	20	26,7	66,7	53,3	23,3	23,1	35,7	33,3	53,3	42,1	30	17,9	40	53,3
pV2_A	0	10	20	6,7	3,3	0	0	0	6,7	0	3,3	0	0	3,3
pV1_A	0	6,7	10	0	0	0	0	3,3	6,7	5,3	0	0	0	0

2021-2022 A.Y. Experience ?????Hybrid Mode?????

Students in Presence

Traditional Seminar-Like Classes for the theoretical topics and for introduction to the main features of NoSQL Infrastructures.

Guided hands-on applications during the labs.

Q&A Sessions

Meetings with the Instructors

Cooperative Group Activities???

Students at Home

Watching Videos of the theoretical topics and for introduction to the main features of NoSQL Infrastructures.

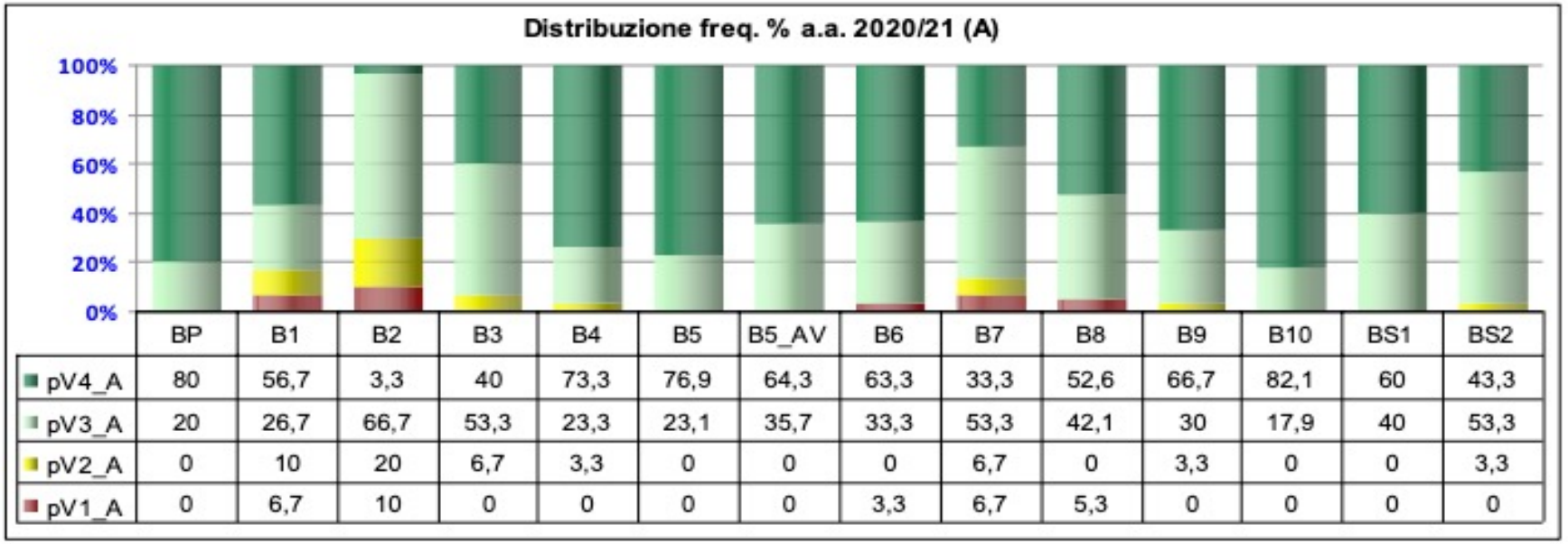
Guided hands-on applications during the labs in streaming.

Q&A Sessions

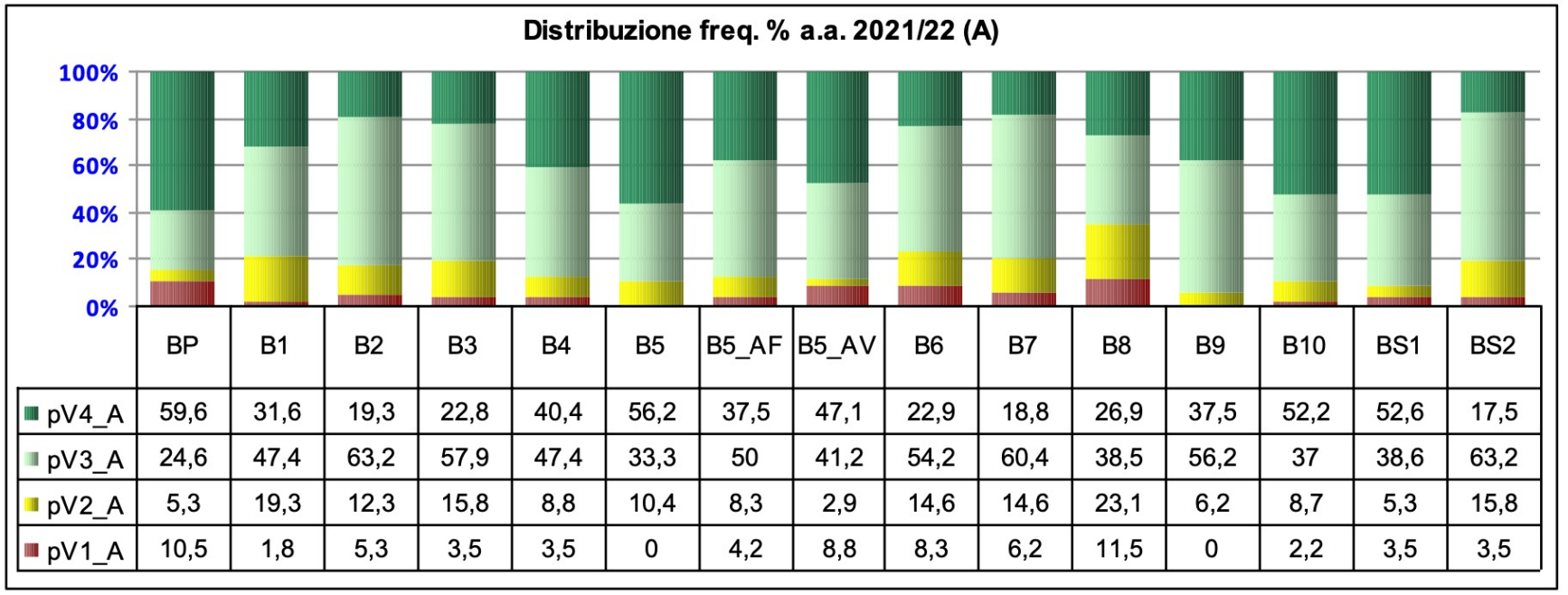
Meetings with the Instructors

Cooperative Group Activities???

2021-2022 A.Y. Experience: Course Evaluation



	BP	B1	B2	B3	B4	B5	B5_AV	B6	B7	B8	B9	B10	BS1	BS2
pV4_A	80	56,7	3,3	40	73,3	76,9	64,3	63,3	33,3	52,6	66,7	82,1	60	43,3
pV3_A	20	26,7	66,7	53,3	23,3	23,1	35,7	33,3	53,3	42,1	30	17,9	40	53,3
pV2_A	0	10	20	6,7	3,3	0	0	0	6,7	0	3,3	0	0	3,3
pV1_A	0	6,7	10	0	0	0	0	3,3	6,7	5,3	0	0	0	0



	BP	B1	B2	B3	B4	B5	B5_AF	B5_AV	B6	B7	B8	B9	B10	BS1	BS2
pV4_A	59,6	31,6	19,3	22,8	40,4	56,2	37,5	47,1	22,9	18,8	26,9	37,5	52,2	52,6	17,5
pV3_A	24,6	47,4	63,2	57,9	47,4	33,3	50	41,2	54,2	60,4	38,5	56,2	37	38,6	63,2
pV2_A	5,3	19,3	12,3	15,8	8,8	10,4	8,3	2,9	14,6	14,6	23,1	6,2	8,7	5,3	15,8
pV1_A	10,5	1,8	5,3	3,5	3,5	0	4,2	8,8	8,3	6,2	11,5	0	2,2	3,5	3,5

2021-2022 A.Y. Experience: Exam Results

Year	*Code*	*Course*	*Average Mark*	*#Students*
2022	883II	LARGE-SCALE AND MULTI-STRUCTURED DATABASES	27,00	92
2021	883II	LARGE-SCALE AND MULTI-STRUCTURED DATABASES	27,82	57
2020	883II	LARGE-SCALE AND MULTI-STRUCTURED DATABASES	28,58	83

- The ***percentage*** of project discussed during the first exam session was equal to 86.6% (30 groups in total, 26 discussions)

- The ***percentage*** of students attending the course who ***passed the exam*** after the first exam session was equal to ***81.5%.*** (90 students attending the course)

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

UNIVERSITÀ DI PISA

CROSSLAB Innovation for industry 4.0

2022-2023 Course Implementation

From Now till the mid or end of November 2022

- Classes for introducing the main NoSQL architecture and strategies.
- Recaps of Software Engineering and Java (with exercises)
- Examples of Applications based on NoSQL DB architectures
- Seminars on Best Practices in Developing Java Application (to be confirmed)
- Introduction to the main features of NoSQL Infrastructures (LevelDB, MongoDB, Neo4j)

2021-2022 Course Implementation

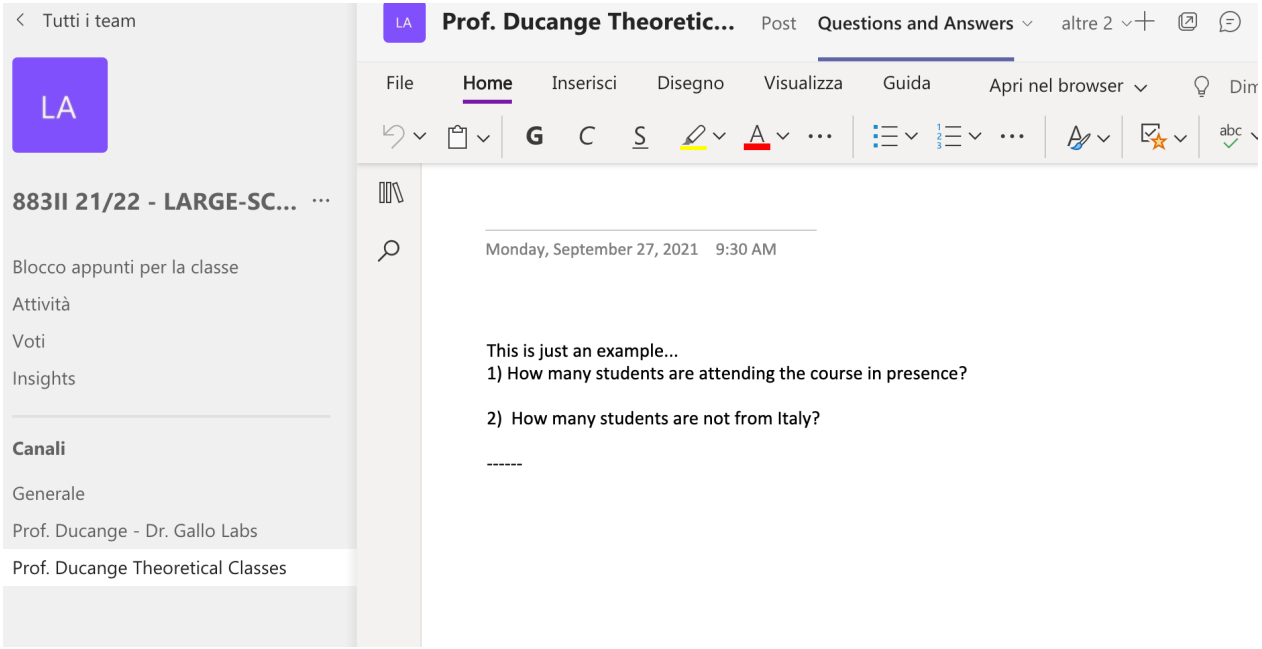
Starting from the mid or the end of November 2022

- Introduction to the Project and final Student Group Definition
- Project development

Questions and Answer Sessions and FAQs

Q&A sessions the beginning and at the end of each class.

Students are invited to write their questions also in the NOTE FAQs sheet of “Generale” Channel of the Teams (at any moment).



The teacher will always answer to the questions, even asynchronously.

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

UNIVERSITÀ DI PISA

CROSSLAB Innovation for industry 4.0

COOPERATIVE GROUP ACTIVITIES

Instructions (I)

1. Students will autonomously create groups (they will also randomly choose a ***group-coordinato***r)
2. ***Each member*** of the group will ***read the text*** of the activity.
3. A ***brain storming*** stage will be supervised by the coordinator.
4. If possible, the activities will be divided into ***sub tasks***.
7. Each sub task (or the entire activity) will be ***assigned*** to ***each members*** or to a ***subset of the members***.
8. Once all the sub ***tasks*** will be ***concluded***, each sub task will be ***presented*** to the group.
9. After a ***discussion***, a final work will be produced ***merging*** and ***refining*** all the contributions.

Instructions (II)

7. After a ***discussion***, a final work will be produced ***merging*** and ***refining*** all the contributions.
8. The instructor will ***randomly select*** the activity of some groups. The ***coordinator*** of each selected group, will the asked to ***present*** the performed activity.
9. The ***instructor will upload*** a proposal of ***solution*** of the activity and each group will be invited to make a ***self-correction of the activity***.
10. If required, each group can ask the instructor for ***clarification***.

The Enrolment form

Please fill and submit the following form:

https://forms.gle/yCxAwd1CoLaRBSTy7

It’s mandatory for attending labs and for the cooperative group activities

About the exam

- Discussion of the project (50%)
- Written test on the theoretical parts of the course (50%): three open questions, 30 minutes.

About the Project

Design and develop of an Application interacting with NoSQL Databases

- Start with an idea and a draft of application requirements, use case diagrams and data enties (quick discussion in the classroom during the course or during meeting slots, ***to be approved by the teacher)***
- Refine requirements and use cases
- Define data entities and relationships by means of UML analysis class diagrams
- Provide the design of the Data Base (at least two of the main non relational models must be considered when defining the requirements. ***Document DB model is mandatory***)
- Define the main queries on the Data Base(s)
- Implement and deploy the application, hopefully on the Virtual Lab
- Test the application (provide a user interface, ***a GUI is not necessary***).
- Write a complete documentation resuming the above items and including a quick guide of the application.

Some Advices for the Project

- ***Attend carefully*** the lessons during the first 6-8 weeks, especially pay attention to the examples of application discussed by the teacher (often the best projects of past students).

- ***Deepen*** your ***skills*** in Java programming and Software Engineering and make the suggested exercises.

- ***Do not expect*** to receive a ***full coverage*** of all the aspects that may be involved by your projects. ***Spend time*** to check for updates, solutions and news ***by your own***.

- ***Ask support*** to the teacher whenever required.

- Come to the meeting with the teacher for asking for additional clarifications, explanations, advices, and resolving doubts.

Rules for the Project

- The project must be developed ***in groups*** (for special cases, talk with the teacher)
- ***No reviews*** are allowed, students will receive a number of examples of past projects.
- The project must be ***discussed before the*** written test (the date will be fixed by the teacher)
- The ***final documentation*** must be submitted to the teacher ***in advance*** (deadline will be fixed some days before the discussion)
- Avoid to ***involve*** the teacher for ***solving problems*** among group members.

Rules for Project Evaluation

Overall Project Evaluation:

- 25 % for the Idea, requirements definitions, the entity-relationship model (and UML diagrams)
- 40% for DB design and query definition
- 25% for the implementation (***not mandatory***)
- 10% for the clarity of the overall documentation

The ***individual assessment*** will depend on the overall evaluation of the project and the answers given by the specific student during the ***project discussions***.

The Self-Assessment Survey

Each student will be required to fill a survey for ***self-assessing*** their technical and theoretical skills.
https://forms.gle/amQYuzb17nZEmuei6

The results of the survey will be used for ***better focusing the teaching activities***.

The data collected will be used only for ***statistical***, ***teaching*** and ***research*** purposes.

Publications (if any) will only report analysis that will use aggregated and anonymized data .

The ***data will not be transferred*** to third parties or to user profiling companies that may use them for commercial purposes.

If personal data (such as Name, Surname and e-mail address) will be provided, the owner may at ***any time request to delete the data*** and not to use them for further analysis.

Contacts

Prof. Ducange office is located at:

Dipartimento di Ingegneria dell'Informazione, University of Pisa.

Office Address: 1, Largo Lucio Lazzarino, I-56100, Pisa (ITALY)

Room: 4-029

Telephone: +39 050 2217684

EMAIL: pietro.ducange_at_unipi.it

Web: https://sites.google.com/site/ducangepietro

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

UNIVERSITÀ DI PISA

CROSSLAB Innovation for industry 4.0