

Gulliver in the land of Data Modeling

Cesena, 21 Febbraio 2024

BIG Expertise

The Business Intelligence Group has been carrying out its research activity since 1997, mainly aiming at studying methodologies, techniques and technologies in the field of Data Analysis

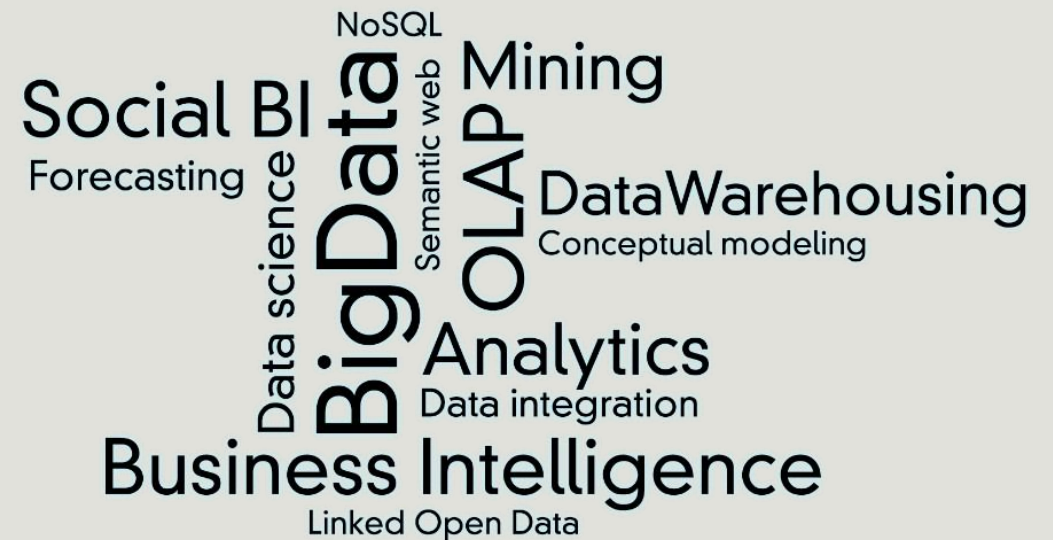
- Currently 5 researchers are involved

More in details:

- Business Intelligence
- Data Warehouse
- Simulations
- Pervasive BI
- Collaborative BI

Our current research topics are related to:

- Social BI
- Big Data & NOSql DBMS
- Semantic Data Warehousing
- Data mining



BIG Expertise

European funding

- *PANDA* (pattern management in DM)
- *ENPADASI* (EU Nutritional Phenotype Assessment and Data Sharing Initiative)
- *TOREADOR* (As-a-service Big Data Analytics)
- *WeLaser* (Laser-based Robotic Weeding)

Public funding

- *D2I* (integration and mining of heterogeneous DBs)
- *WISDOM* (ontology-enhanced web searching)
- *WebPoIEU* (Comparing Social Media and Political Participation across EU)
- *GenData2020* (data-centric genomic computing)
- *DyNamiTE* (Digital fightiNg Tax Evasion)
- *MO.RE.Farming* (Big Data for Precision Farming)
- *INNOFRUVE* (Ricerca industriale ed innovazione nel comparto ortofrutta)
- *AgroBigDataScience* (Big Data for Precision Farming)

Private funding (2015-2021)

- *Data Mining in the Fashion Field* with Valentino
- *Set-up of a Social Business Intelligence framework* with Amadori s.p.a.
- *Feasibility study for a Social Business Intelligence system* with DOXA
- *Anomaly detection in the gas network* with HERA spa
- *Harnessing Wellness Knowledge* with Technogym
- *Methodological and Scientific Support to several Public bodies* With Ministry of Justice, Economy and Finance
- *Vaccine monitoring* with Regione Veneto & ONIT
- *Intelligent Monitoring Systems for Critical Environments* with Leonardo-Finmeccanica
- *Data-driven budgetting* with Teddy
- *Digital Transformation* with BRT, PLT Energia

Dalla programmazione alla progettazione

Nel primo anno di studi vi siete concentrati sulla **programmazione** (imperativa): C, Java

... MA per realizzare applicazioni complesse sono necessarie competenze di progettazione, modellazione e astrazione

Progettazione: *il processo di definizione di architettura, componenti, moduli, interfacce e dati per un prodotto software, in modo che soddisfi determinati requisiti. E' fondamentale limitarsi agli aspetti salienti evitando i dettagli che saranno trattati in fase di programmazione*

Si può progettare un algoritmo, un'architettura HW e SW, un database, un processo,...

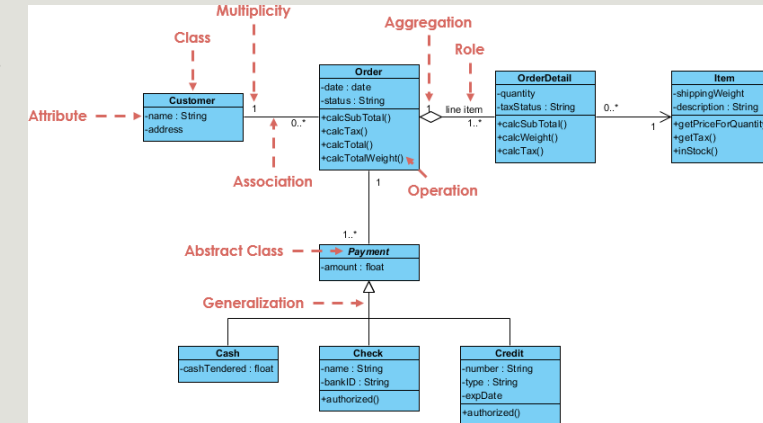
Per progettare è necessario sviluppare capacità di modellazione e astrazione che, partendo dalla descrizione di singoli casi reali, permettano di eliminare i dettagli dei singoli eventi e permettano di estrapolare i concetti salienti e le loro relazioni

Modellazione e Astrazione

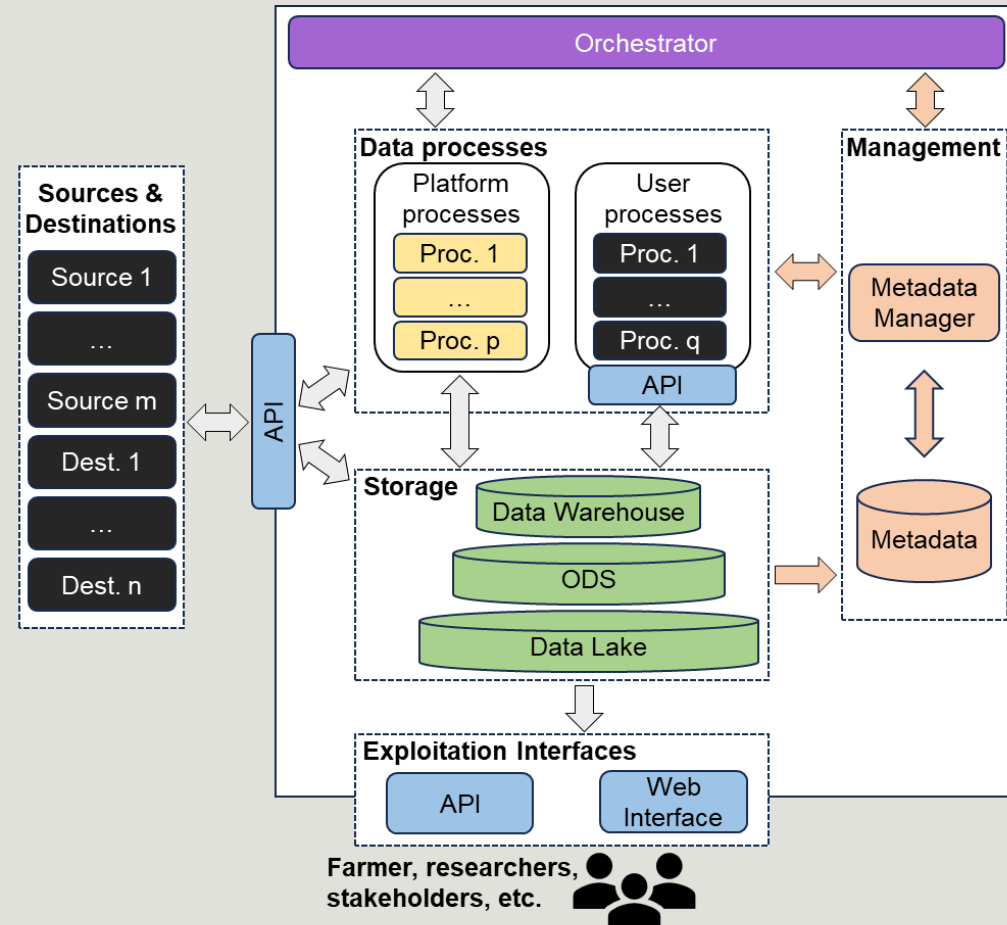
Ne avete avuto un'anteprima nell'ambito della progettazione a oggetti in cui avete usato il diagramma delle classi UML

L'astrazione utilizza diversi meccanismi quali:

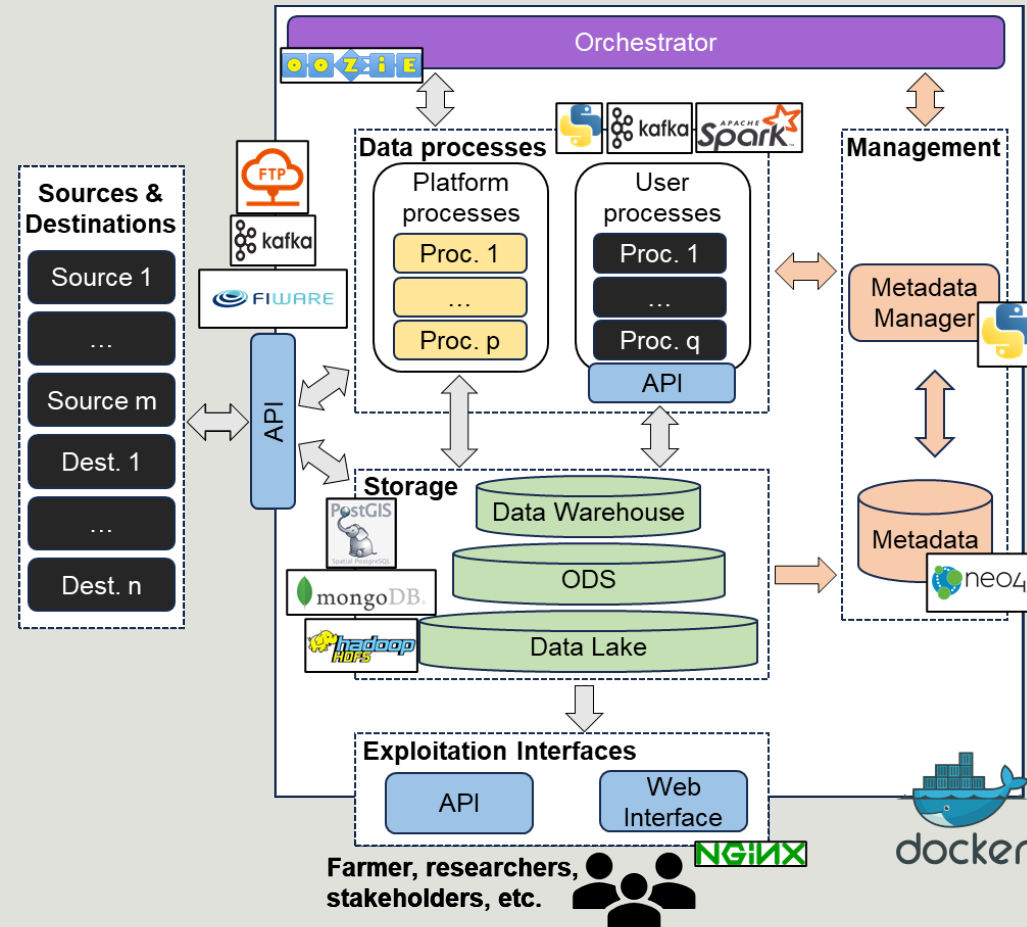
- **Classificazione**: si giunge al concetto astratto identificando le proprietà che accomunano le diverse istanze
- **Aggregazione**: si giunge al concetto astratto identificando le parti che lo compongono
- **Generalizzazione**: si giunge al concetto astratto come unione delle classi contenute nel concetto



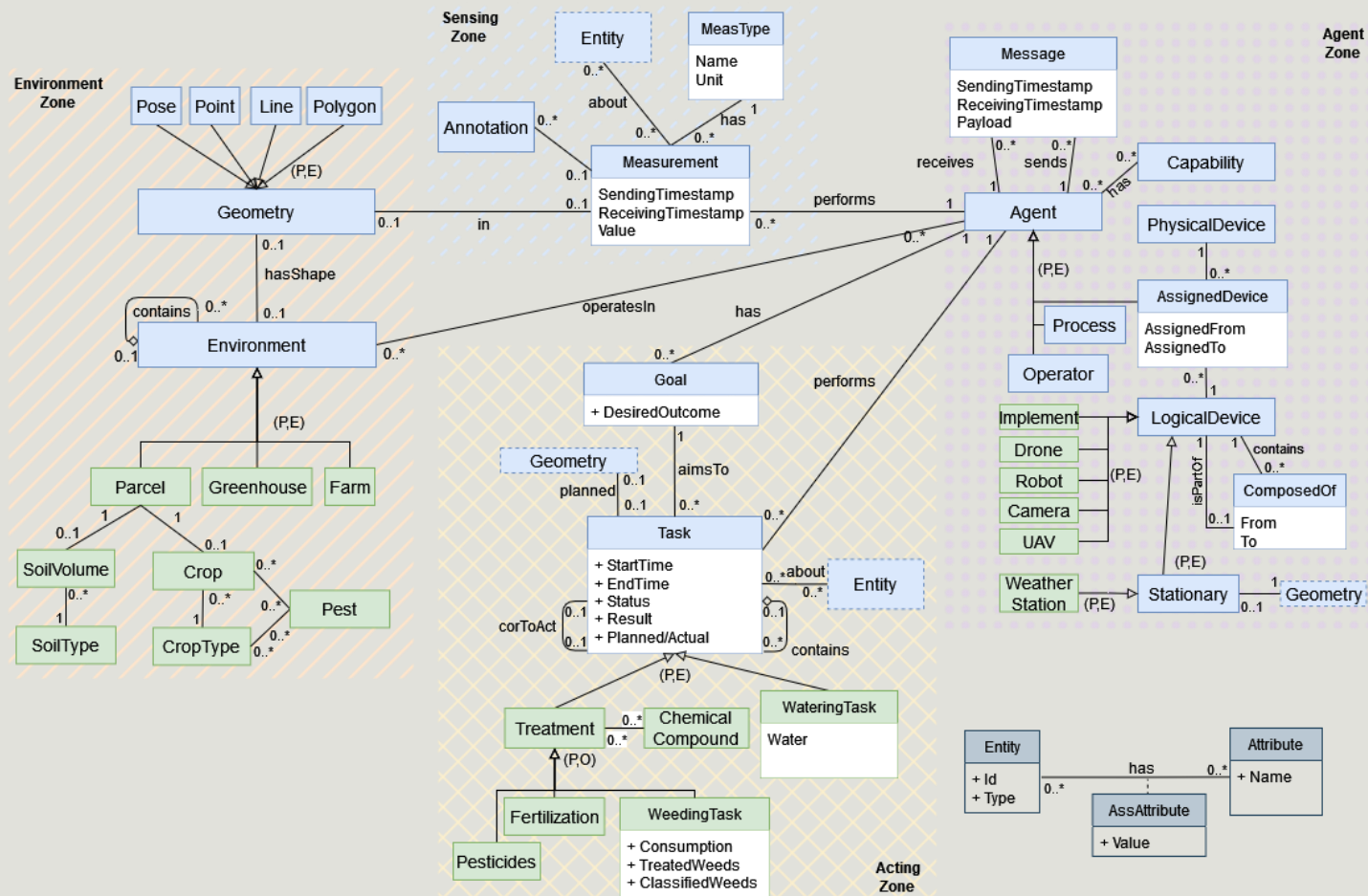
Dalla teoria alla pratica



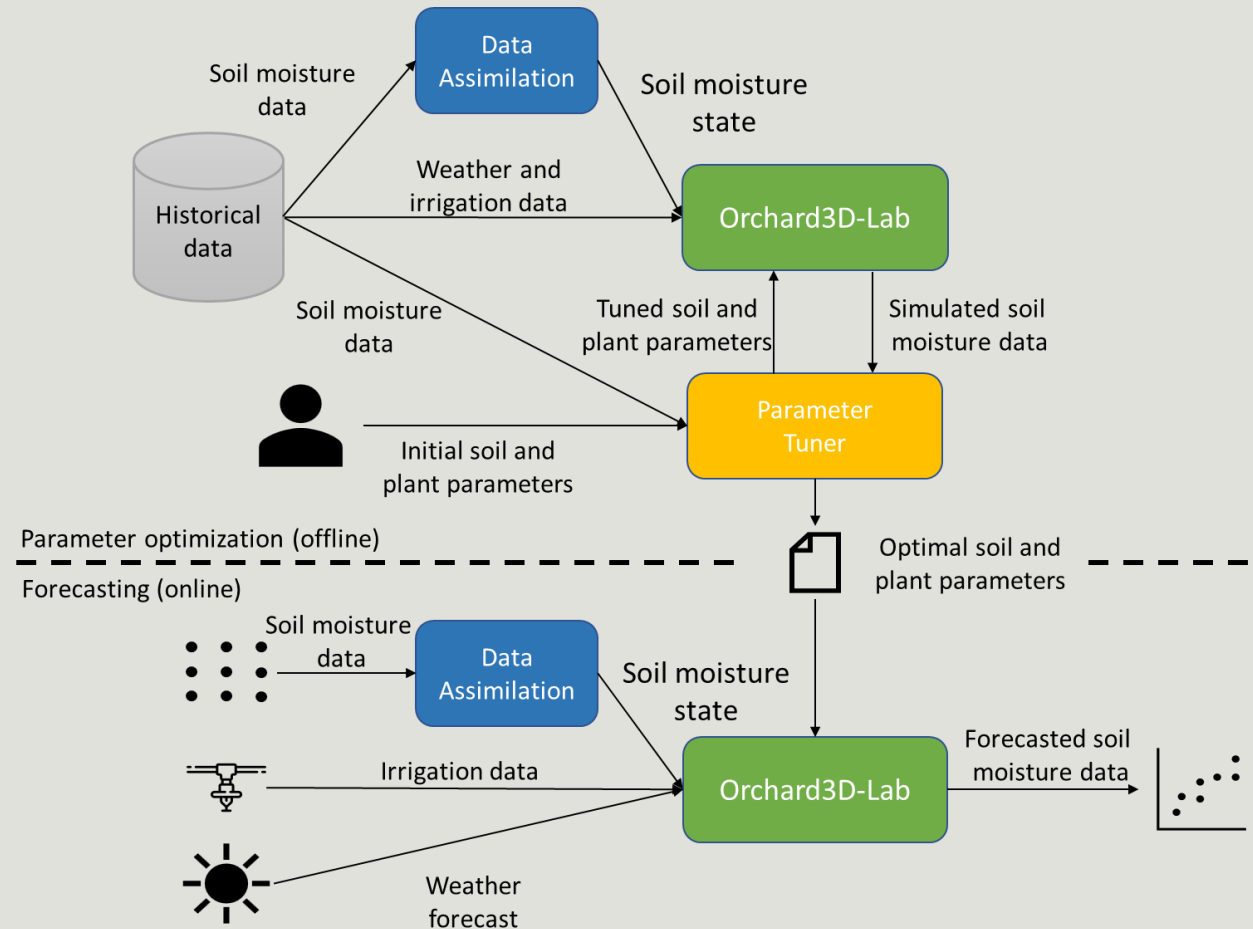
Dalla teoria alla pratica



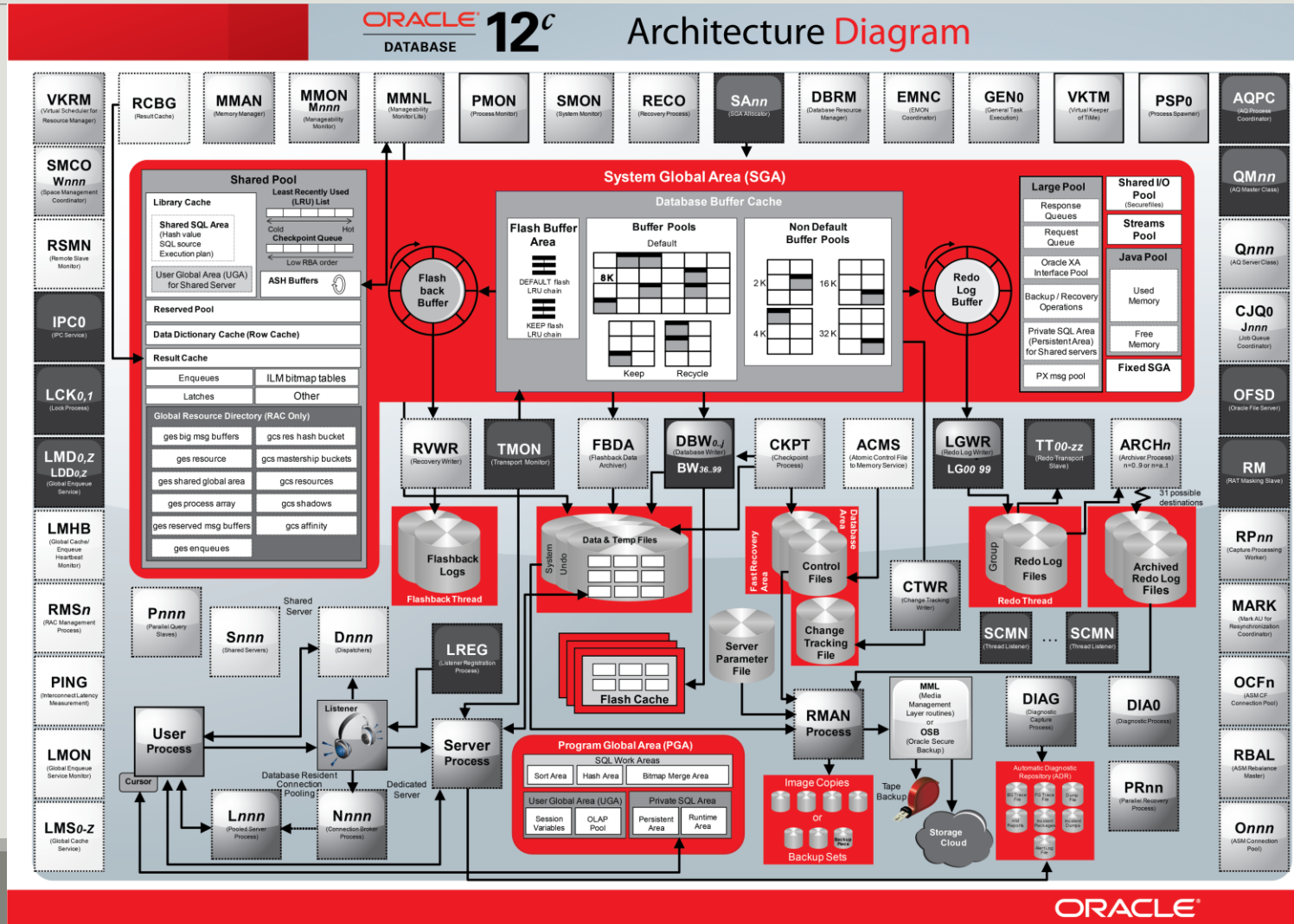
Dalla teoria alla pratica



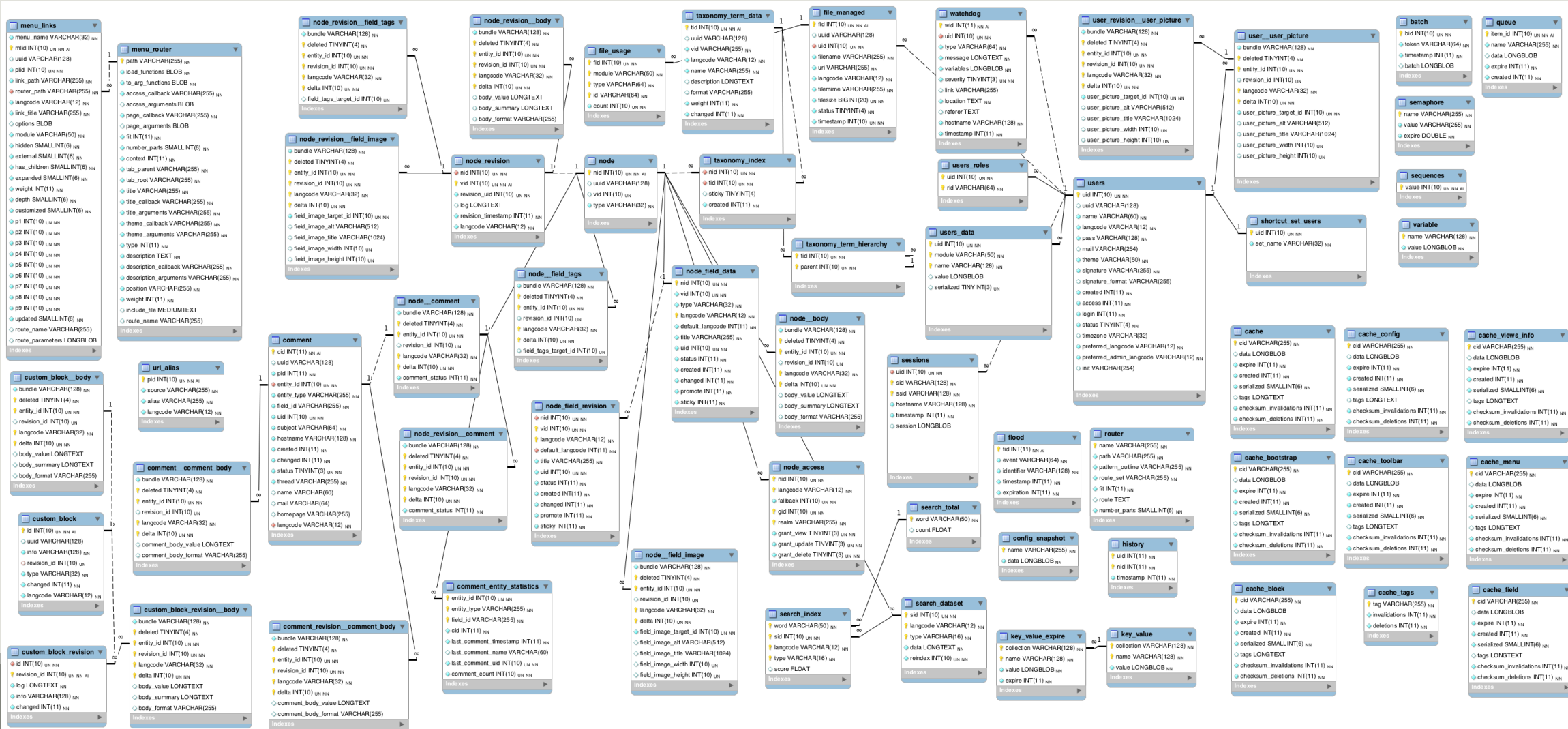
Dalla teoria alla pratica



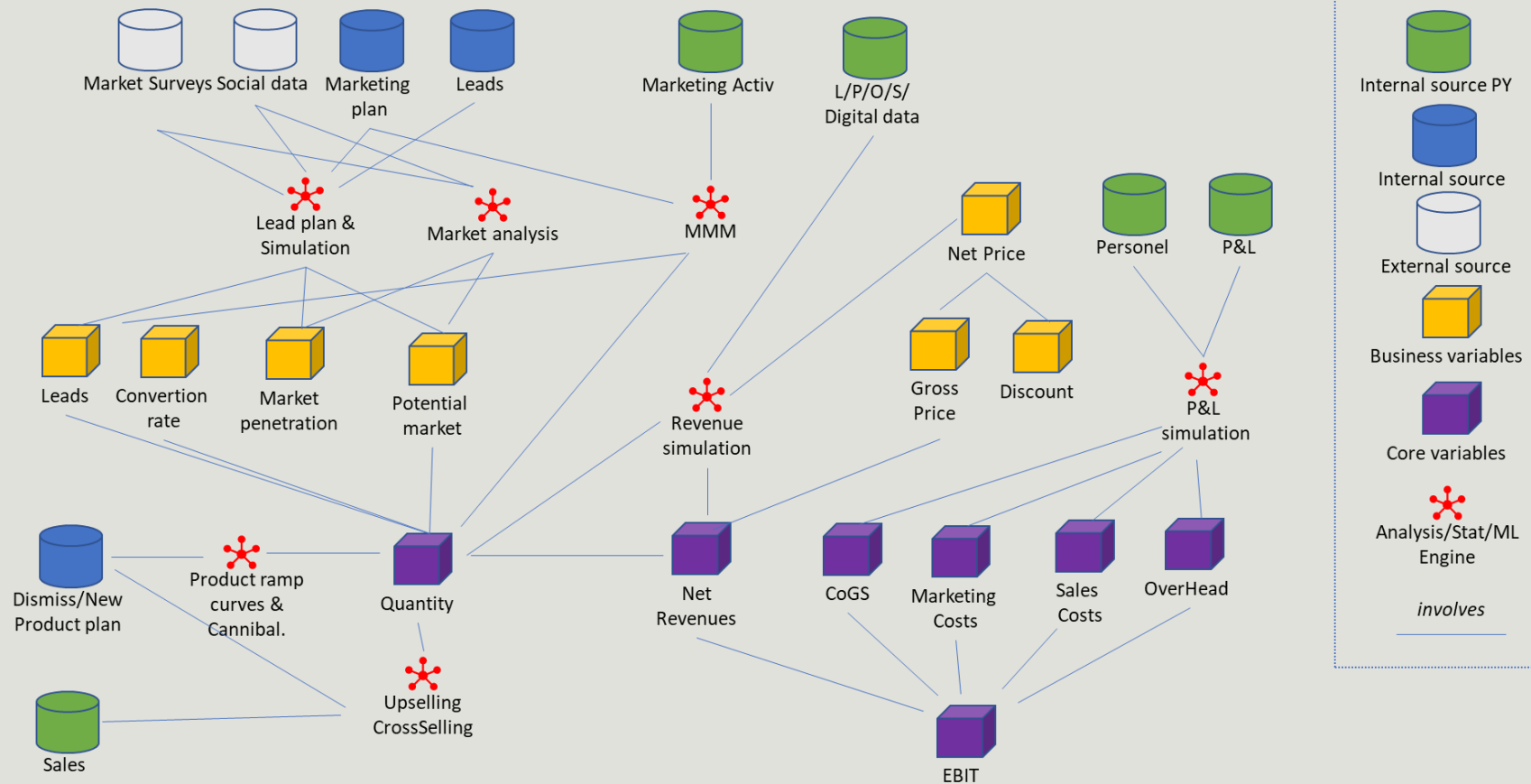
Dalla teoria alla pratica



Dalla teoria alla pratica



Dalla teoria alla pratica



Modellazione: una competenza fondamentale

Indipendentemente dal percorso lavorativo che seguirete la capacità di modellazione rappresenta

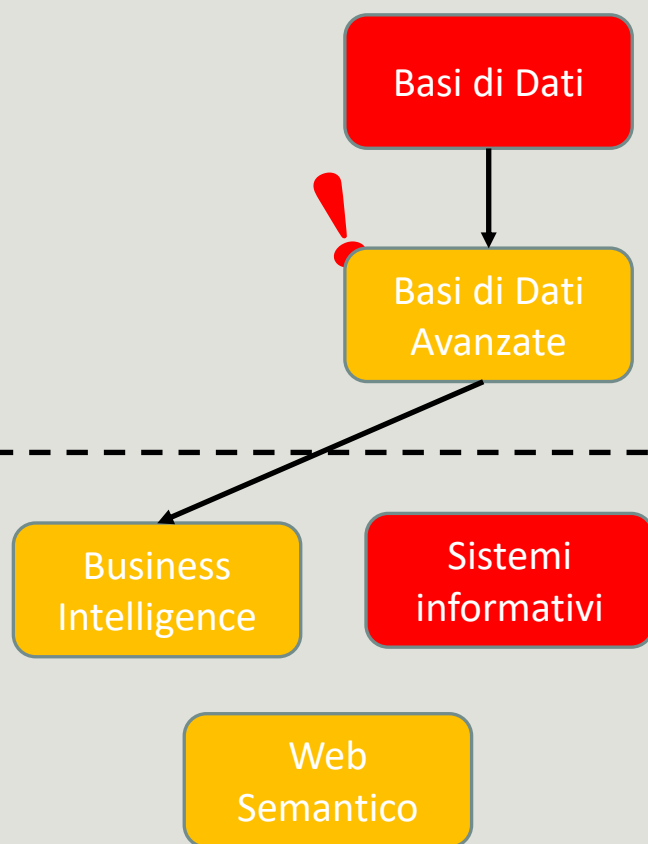
LA COMPETENZA FONDAMENTALE che distingue l'ingegnere informatico poiché permette di:

- Orientarsi in/comprendere problemi complessi
- Risolvere un problema complesso traducendolo in molti problemi semplici
- Permettere di comunicare e organizzare il lavoro

Progettazione e Modellazioni sono competenze difficili da acquisire:

- Richiedono capacità di astrazione
- Sono meno sintattiche della programmazione
- Richiedono la conoscenza del dominio applicativo
- Spesso non c'è sempre una sola soluzione corretta
- *Risultano cruciali solo quando i problemi sono complessi*
- *Danno soddisfazione solo quando applicate in contesti reali*

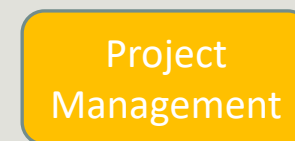
Orientiamoci!



Focus sui dati



Focus sul codice



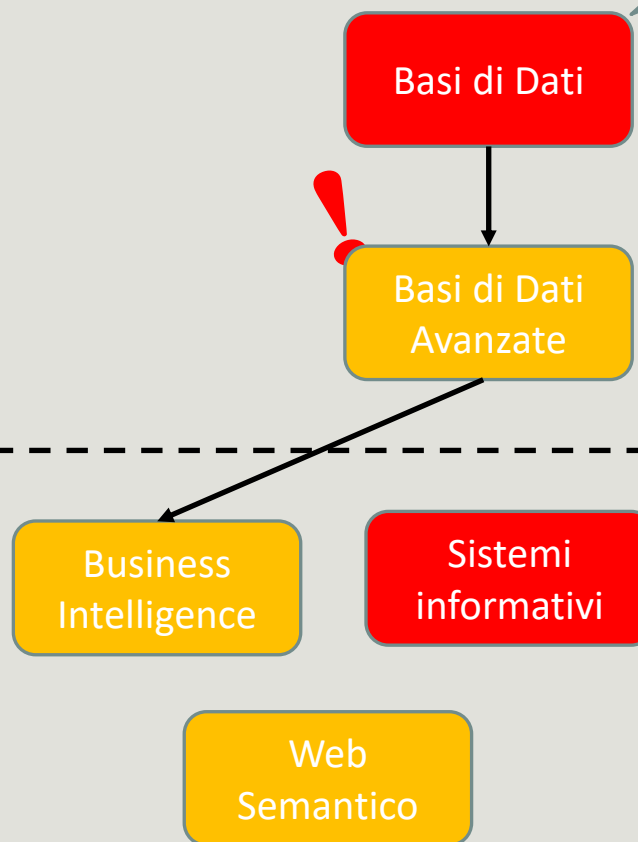
Focus sui progetti

Triennale

Magistrale

Orientiamoci!

Principi di modellazione e astrazione
Modellazione concettuale con formalismo ER
Progettazione dati relazionale



Focus sui dati

Ingegneria
del SW

Software
Architecture and
Data Platform

Software Process
Engineering

Focus sul codice

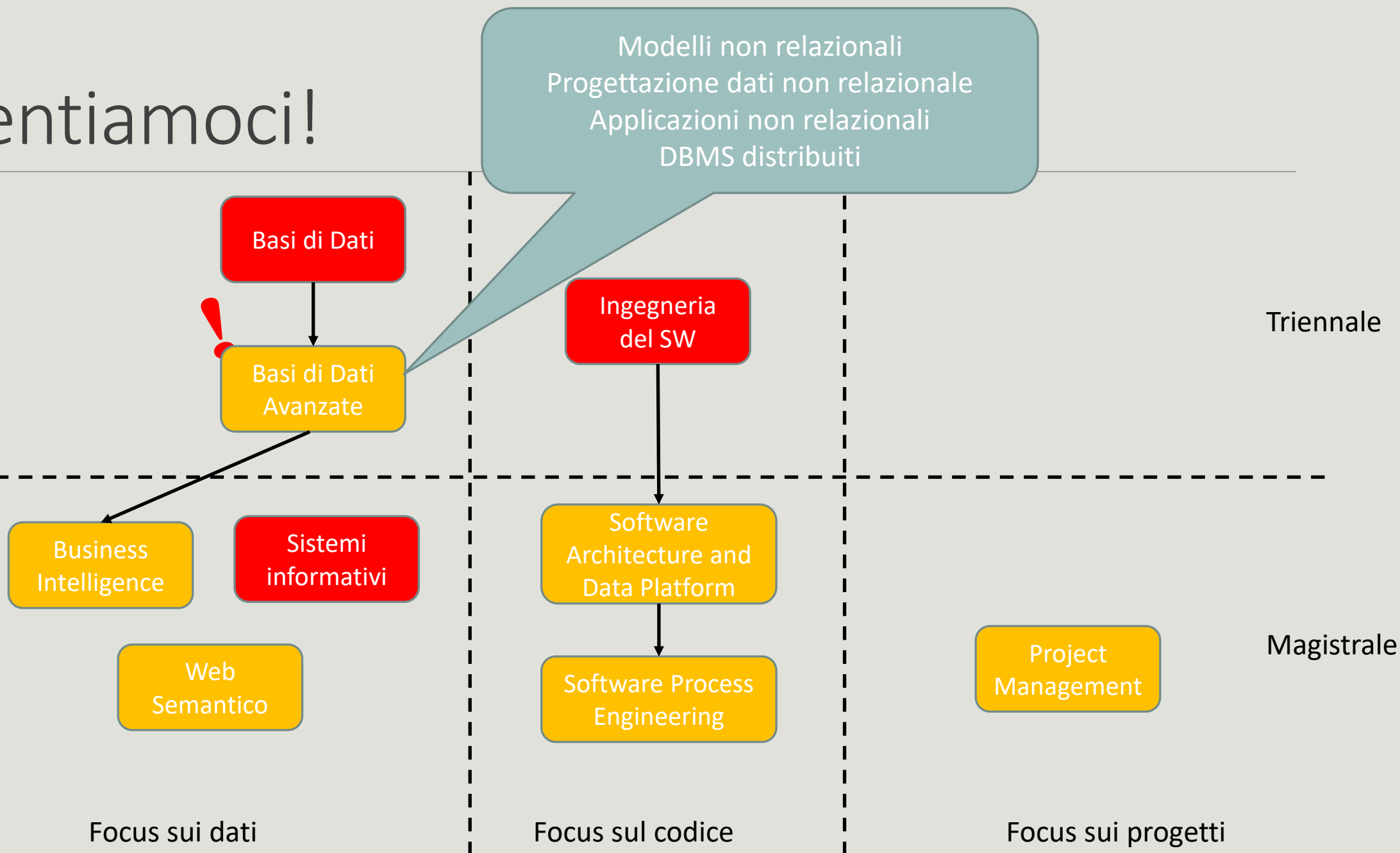
Project
Management

Focus sui progetti

Triennale

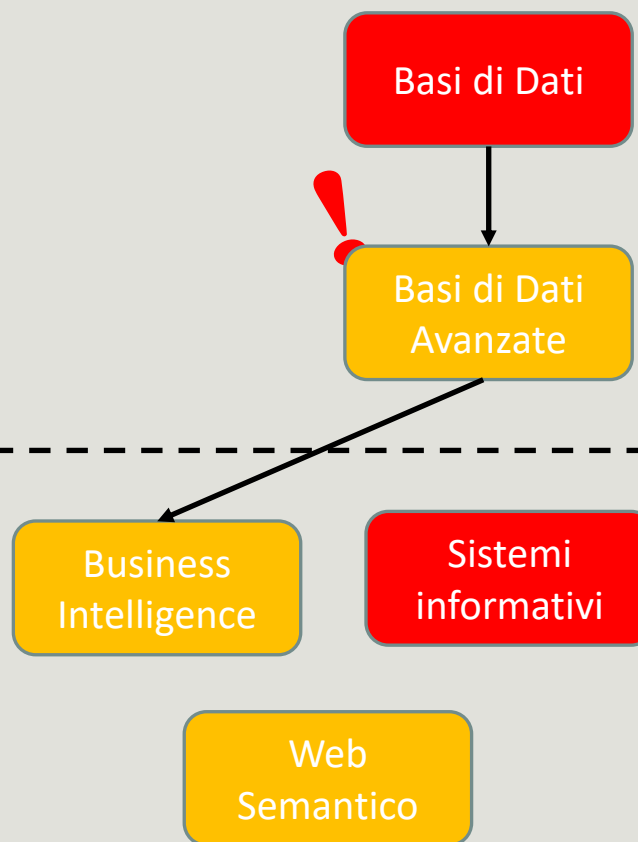
Magistrale

Orientiamoci!



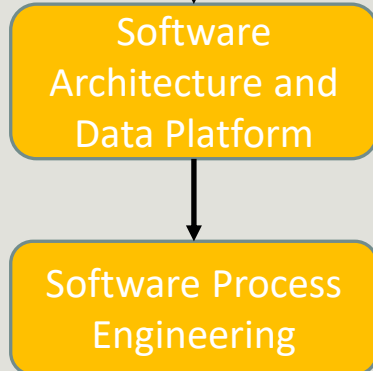
Orientiamoci!

Progettazione a oggetti
Modellazione concettuale con UML



Focus sui dati

Ingegneria
del SW



Focus sul codice

Project
Management

Focus sui progetti

Triennale

Magistrale

Sondaggio 1

Ti senti più:

- Programmatore
- Sistemista
- Progettista
- Project manager

Sondaggio 2

Dopo quanti anni di lavoro ritieni che avrai un ruolo di progettista?

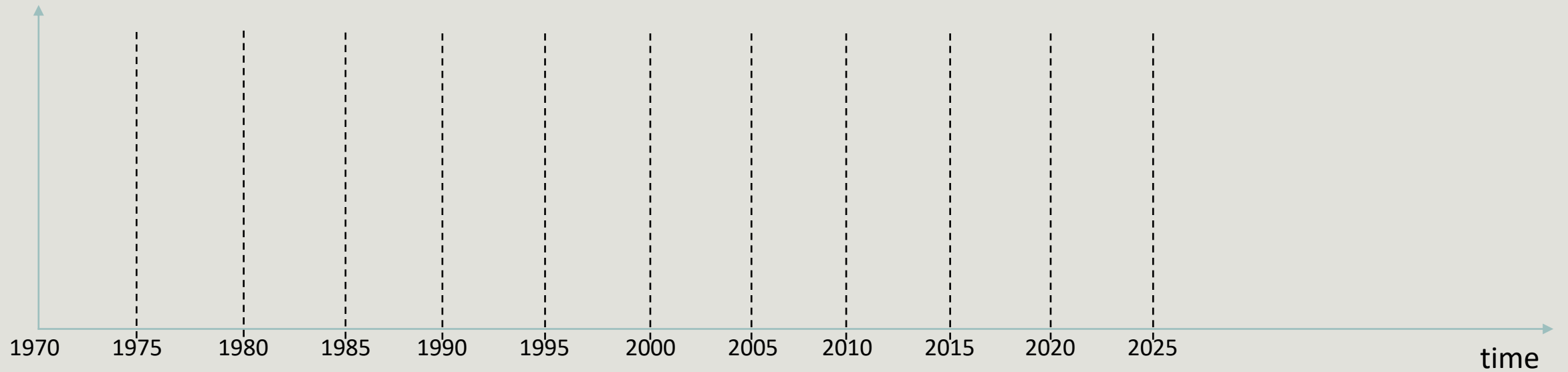
Dopo quanti anni di lavoro ritieni che avrai un ruolo di project manager?

I DBMS

Un DBMS è un sistema software che gestisce grandi quantità di dati persistenti e condivisi, e che offre supporto per almeno un modello dei dati

- La gestione di grandi quantità di dati richiede particolare attenzione ai problemi di efficienza (ottimizzazione delle richieste, ma non solo!)
- La persistenza e la condivisione richiedono che un DBMS fornisca meccanismi per garantire l'affidabilità dei dati (**fault tolerance**), per il **controllo degli accessi** e per il **controllo della concorrenza**
- Un **modello dei dati** consente agli utenti di disporre di un'astrazione di alto livello attraverso cui interagire con il DB.
- Diverse altre funzionalità vengono messe a disposizione per motivi di efficacia, ovvero per semplificare la descrizione delle informazioni, lo sviluppo delle applicazioni, l'amministrazione di un DB, ecc.

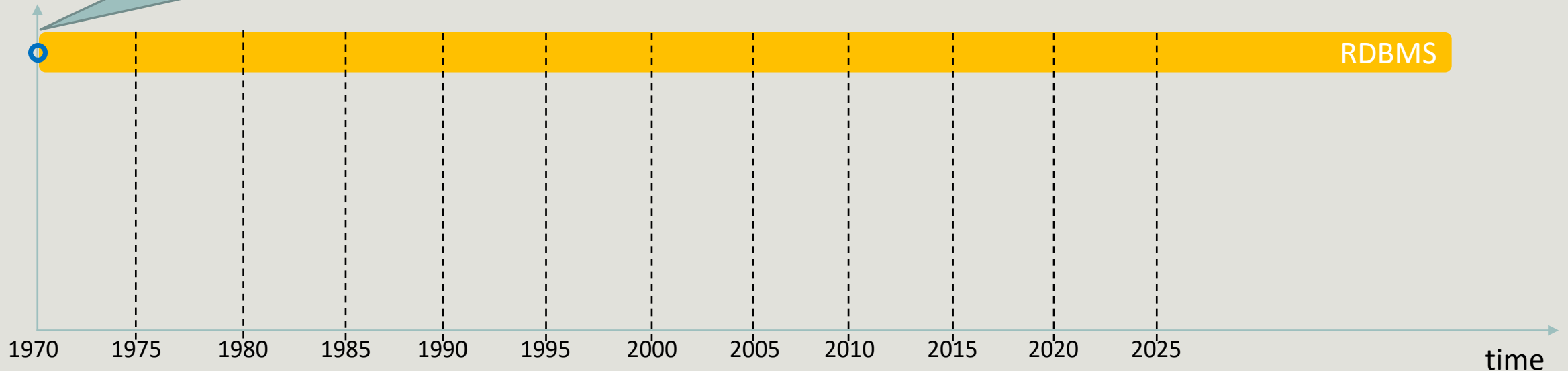
Una time line per i DBMS



Una time line per i DBMS

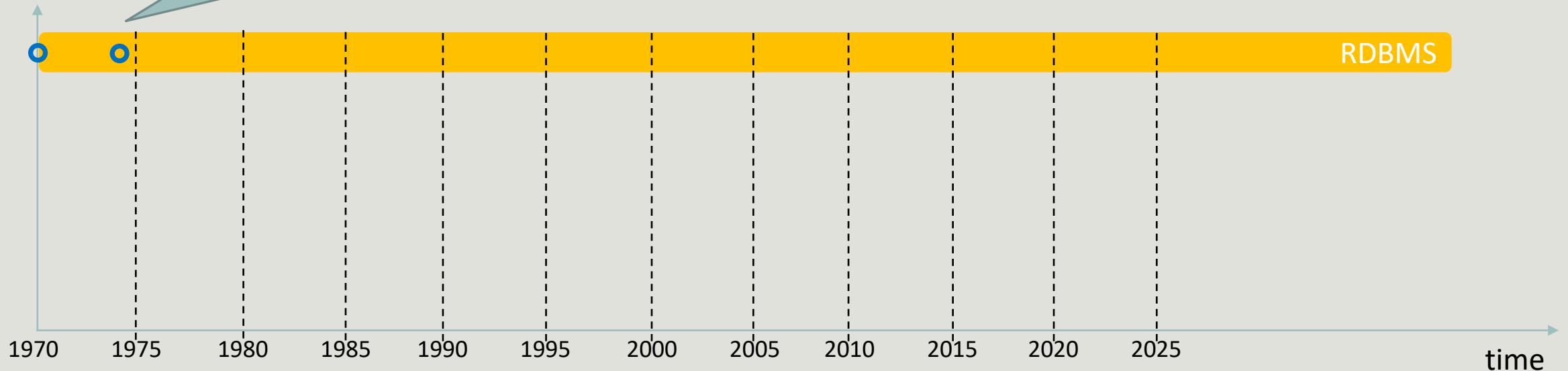
1970 E.F. Codd (IBM) pubblica il modello relazionale

"It provides a means of describing data with its natural structure only--that is, without superimposing any additional structure for machine representation purposes. Accordingly, it provides a basis for a high level data language which will yield maximal independence between programs on the one hand and machine representation on the other."(Codd 1970)



Una time line per i DBMS

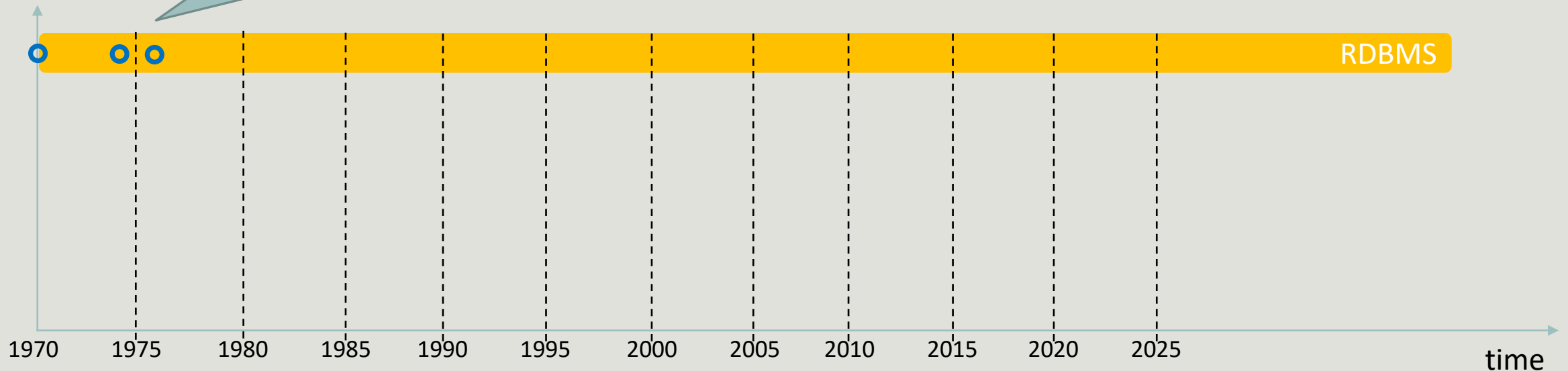
1974 Chamberlin (IBM) nasce il linguaggio SQL - Donald Chamberlin (IBM)
linguaggio standardizzato per database basati sul modello relazionale (RDBMS)



Una time line per i DBMS

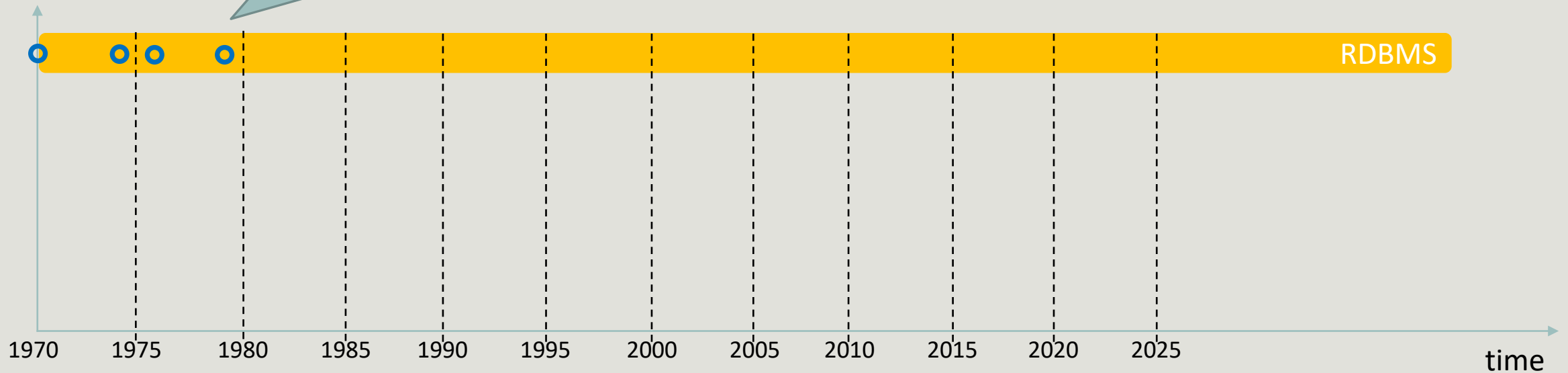
1976 Peter Chen pubblica il modello Entity Relationship

Un modello teorico per la rappresentazione concettuale e grafica dei dati a un alto livello di astrazione



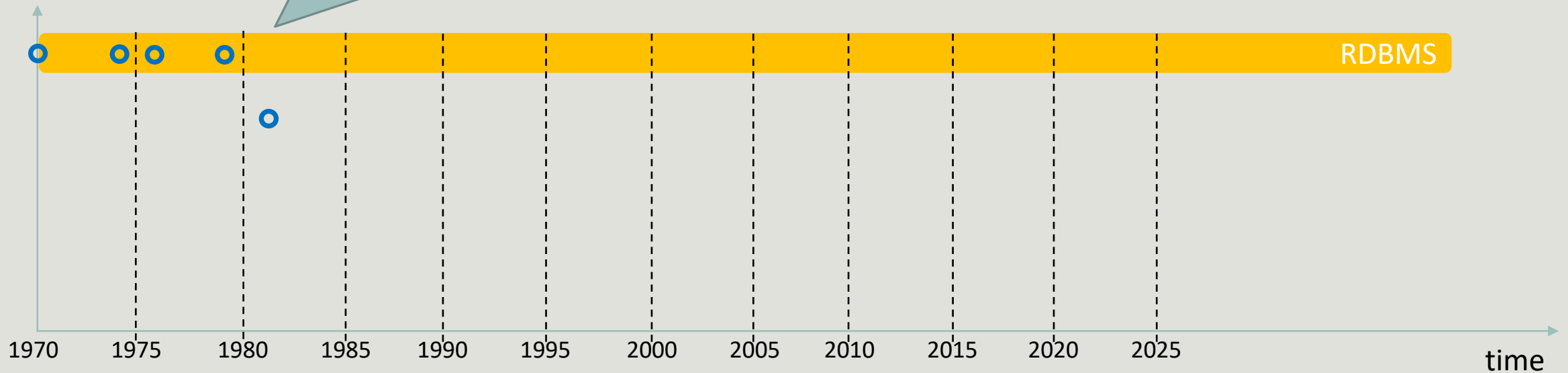
Una time line per i DBMS

1979 Larry Ellison lancia la prima release del DBMS relazione ORACLE



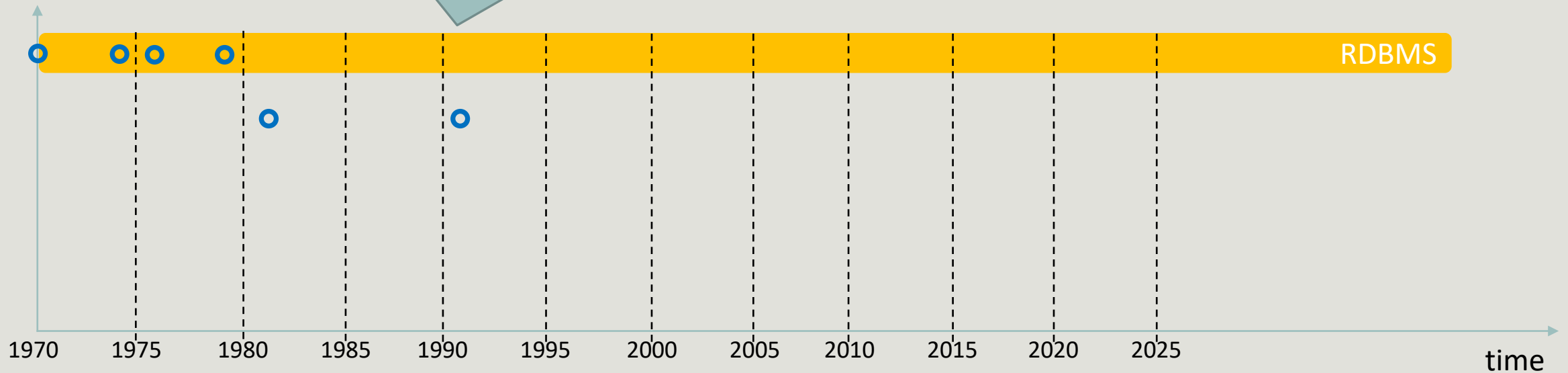
Una time line per i DBMS

1982 Definizione del protocollo TCP/IP e della parola "Internet".



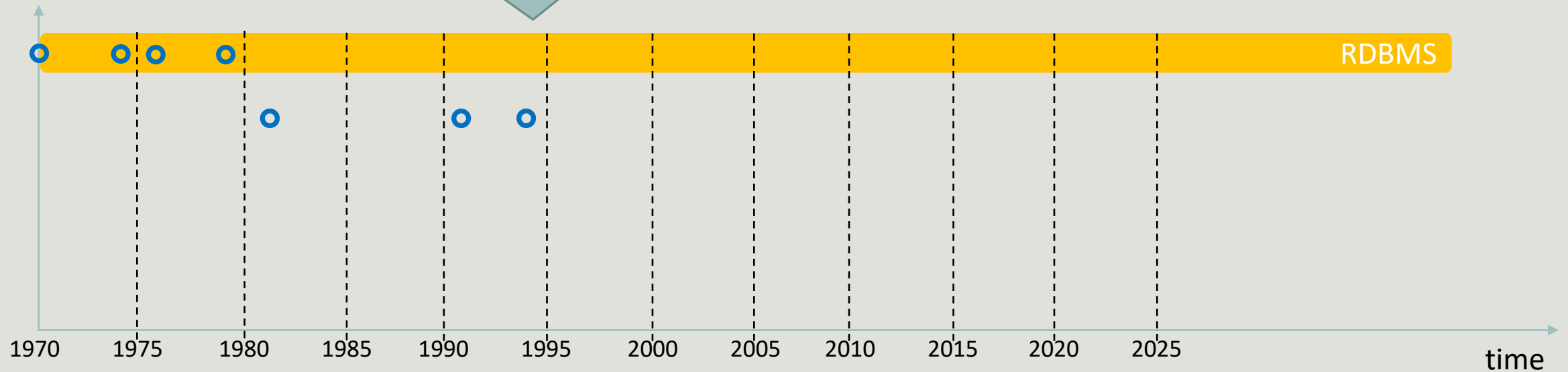
Una time line per i DBMS

1991 Il CERN annuncia la nascita di internet



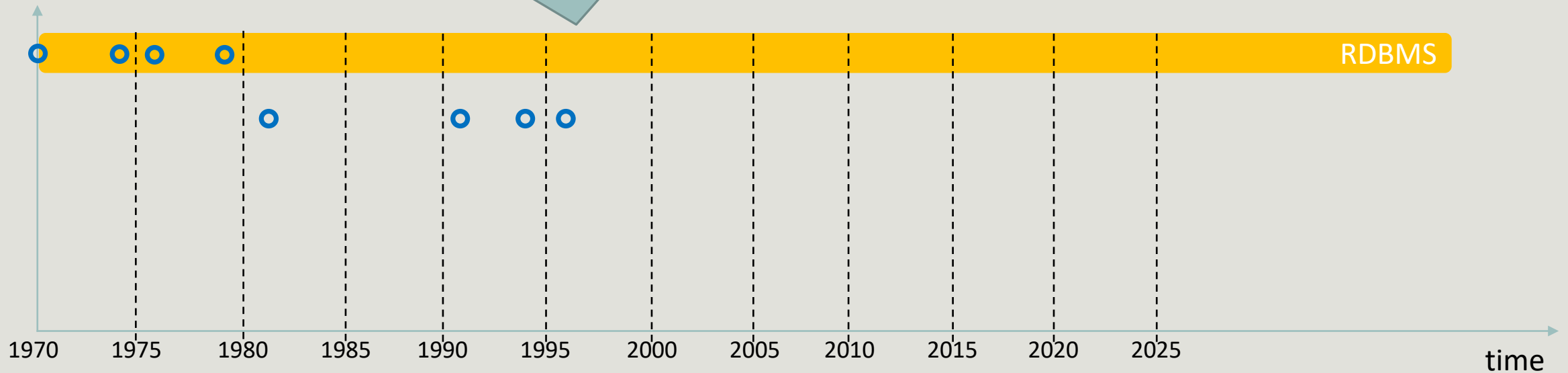
Una time line per i DBMS

1994 Jeff Bezos fonda Amazon



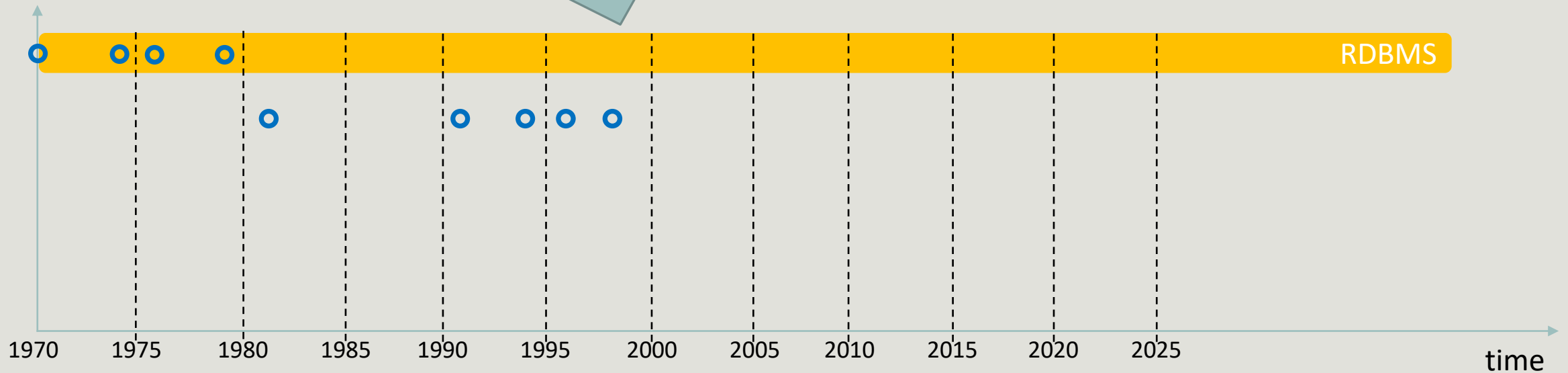
Una time line per i DBMS

1996 Sono connessi ad Internet 10 milioni di computer.



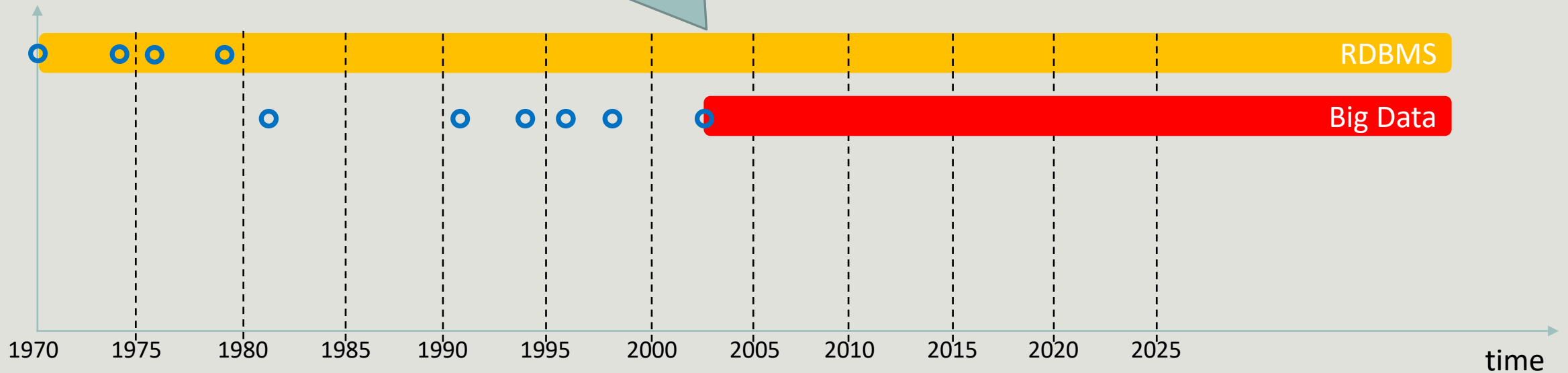
Una time line per i DBMS

1998 Larry Page e Sergey Brin fondano Google



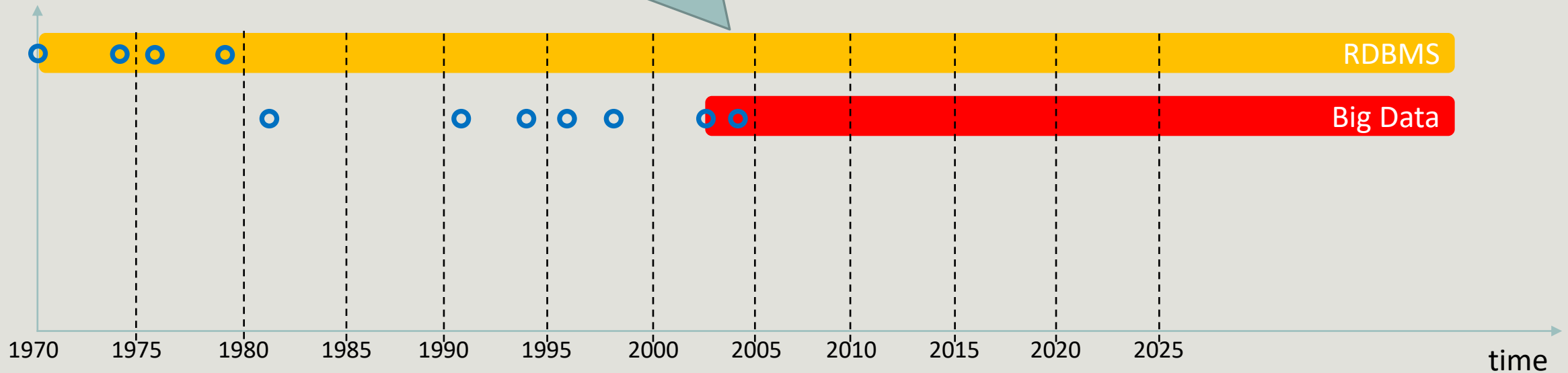
Una time line per i DBMS

2003 nasce il Google File System



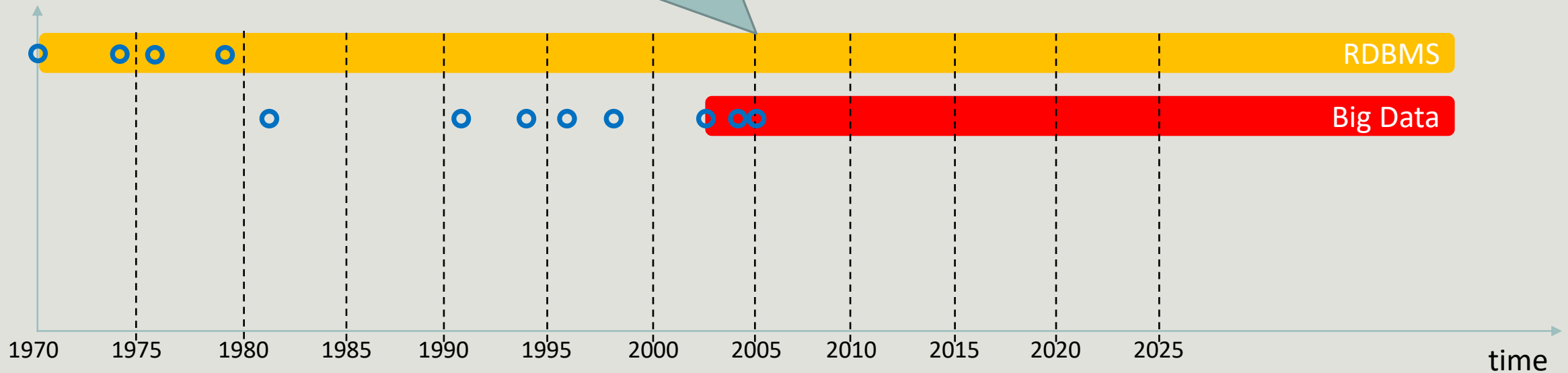
Una time line per i DBMS

2004 nasce Map Reduce (Google)
Linguaggio per il calcolo distribuito



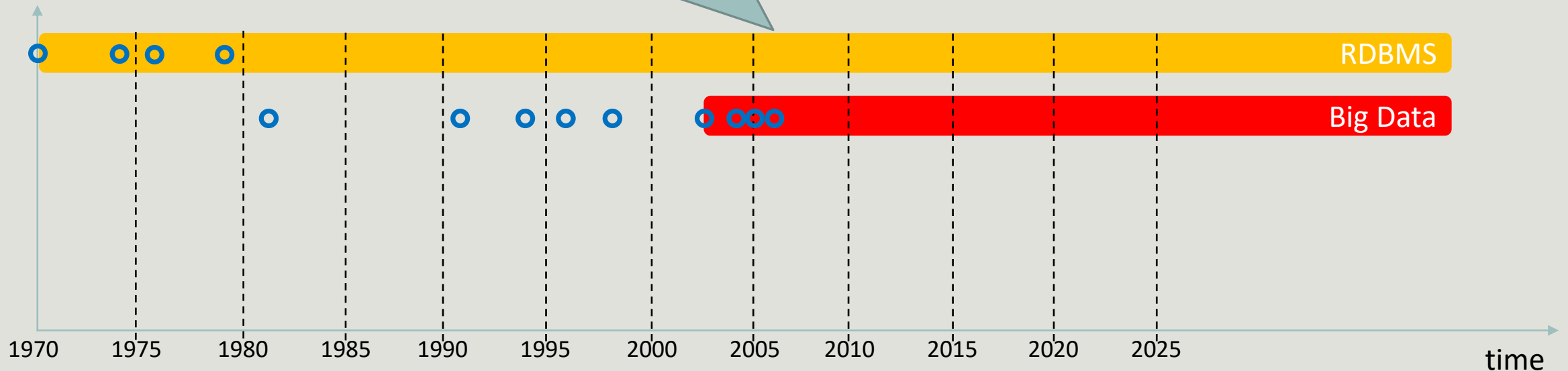
Una time line per i DBMS

2005 nasce Hadoop 1 (Apache)
basato sugli articoli di Google è un sistema Big Data open source a supporto del web crawler Nutch



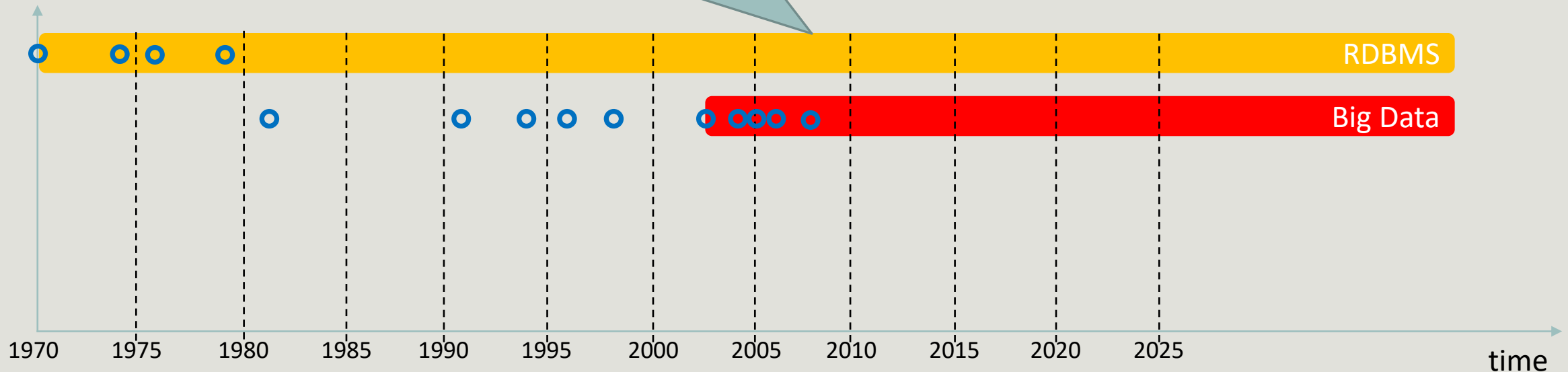
Una time line per i DBMS

2006 nascono i Web Services di Amazon (AWS)
la parola cloud computing (conciata nel 1997) è sulla bocca di tutti



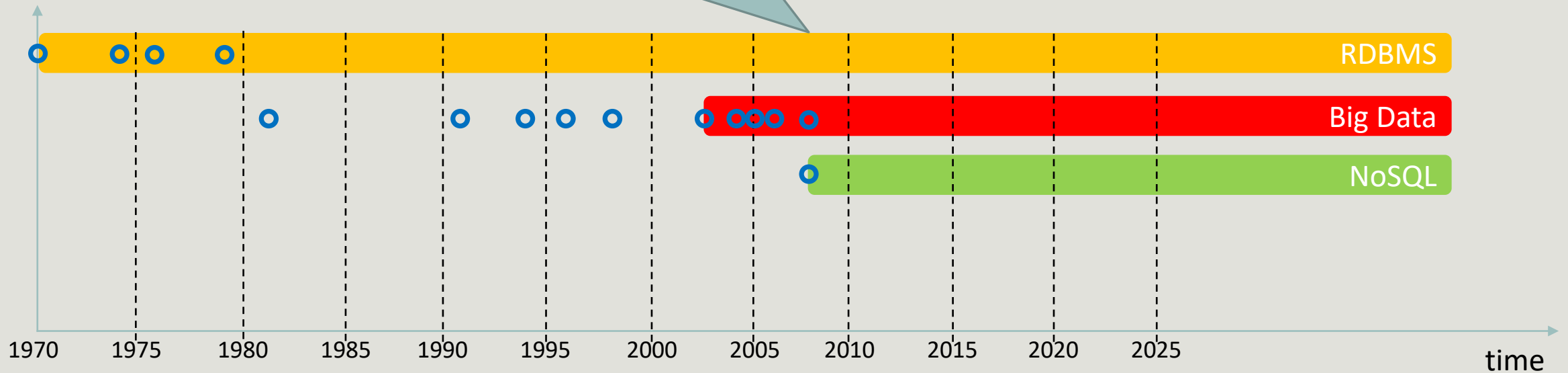
Una time line per i DBMS

2008 si afferma il ruolo di Data Science
DJ Patil afferma: *"Data Scientist is The Sexiest Job of the 21st Century"*



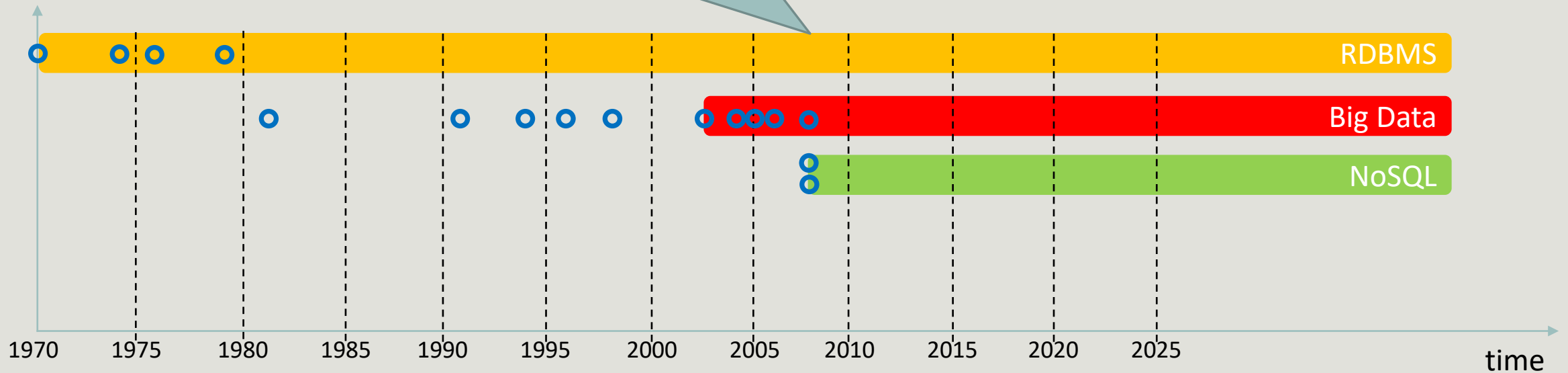
Una time line per i DBMS

2008 nasce Cassandra (Facebook)
DBMS non relazionale (wide column) distribuito e open source



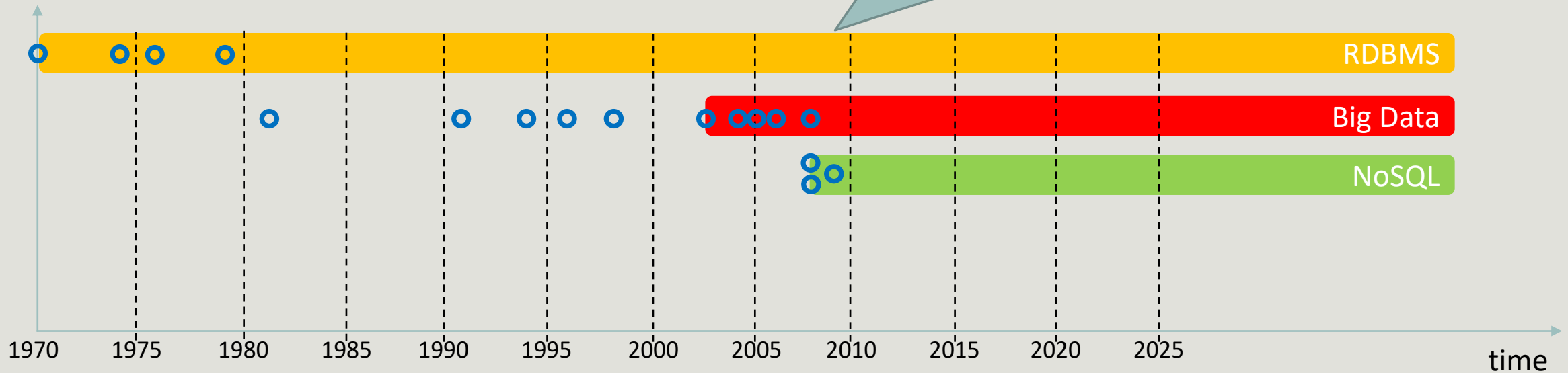
Una time line per i DBMS

2008 nasce Voldemort (Linkedin)
DBMS non relazionale (key-value) open source



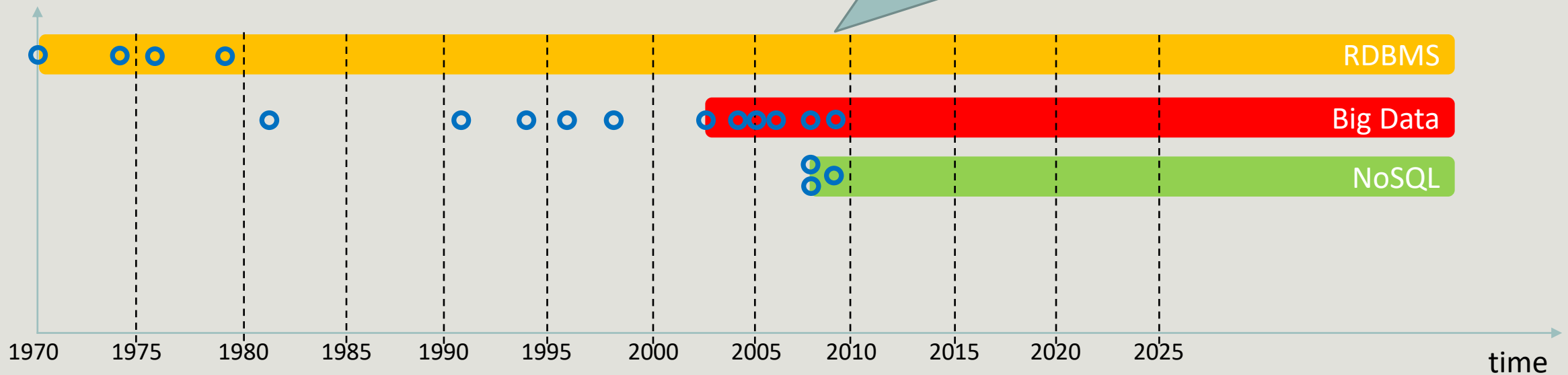
Una time line per i DBMS

2009 nasce MongoDB
DBMS non relazionale (document-based) open source



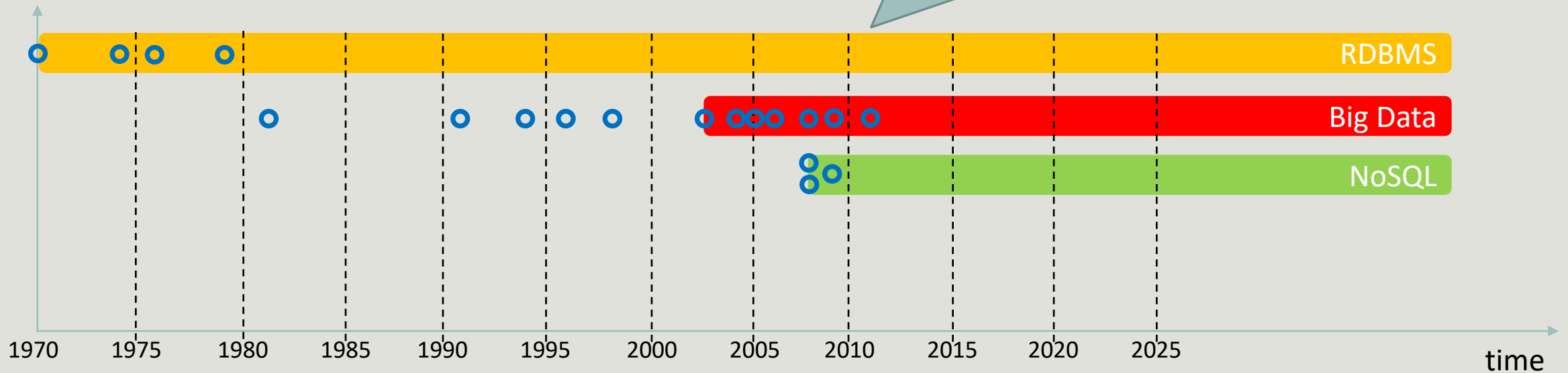
Una time line per i DBMS

2009 nasce l'IoT
secondo CISCO Internet Business Solutions Group sono connesse a Internet più cose che persone



Una time line per i DBMS

2011 nasce HIVE
DBMS relazionale su piattaforma Big Data



One size does not fit all!

- ❑ Per oltre 40 anni l'unico modello dati disponibile era quello relazionale
 - ✓ Applicazioni con necessità/caratteristiche diverse venivano implementate sullo stesso modello
 - ✓ Le performance erano limitate



- ❑ La necessità di supportare applicazioni con vincoli di performance stringenti ed enormi moli di dati ha portato alla nascita dei DBMS NoSQL
 - ✓ Ogni modello ha caratteristiche diverse, specifiche per carichi di lavoro

- ❑ Il progettista deve conoscere le caratteristiche e i principi di modellazione dei diversi modelli



One size does not fit all!

- ❑ Per oltre 40 anni l'unico modello dati disponibile era quello relazionale
 - ✓ Applicazioni con necessità/caratteristiche diverse venivano implementate sullo stesso modello

Modello	Descrizione	Casi d'uso	Applicazioni
Key-value	Associates any kind of value to a string	Dictionary, lookup table, cache, file and images storage	Web session profile, shopping cart, user preferences
Document	Stores hierarchical data in a tree-like structure	Documents, anything that fits into a hierarchical structure	Event log, CMS, blogging platform
Wide column	Stores sparse matrixes where a cell is identified by the row and column keys	Crawling, high-variability systems, sparse matrixes	Event log, CMS, blogging platform, GIS
Graph	Stores vertices and arches	Social network queries, inference, pattern matching	Social network, routing application, fraud detection

l'applicazione

modello

realtà

One size does not fit all!

- ❑ Per oltre 40 anni l'unico modello dati disponibile era quello relazionale
 - ✓ Applicazioni con necessità/caratteristiche diverse venivano implementate sullo stesso modello

Modello	Descrizione	Casi d'uso	Applicazioni
Key-value	Associates any kind of value to a string	Dictionary, lookup table, cache, file and images storage	Web session profile, shopping cart, user preferences
Document	Stores hierarchical data in a tree-like structure	Documents, anything that fits into a hierarchical structure	Event log, CMS, blogging platform
Wide column	Stores sparse matrixes where a cell is identified by the row and column keys	Crawling, high-variability systems, sparse matrixes	Event log, CMS, blogging platform, GIS
Graph	Stores vertices and arches	Social network queries, inference, pattern matching	Social network, routing application, fraud detection

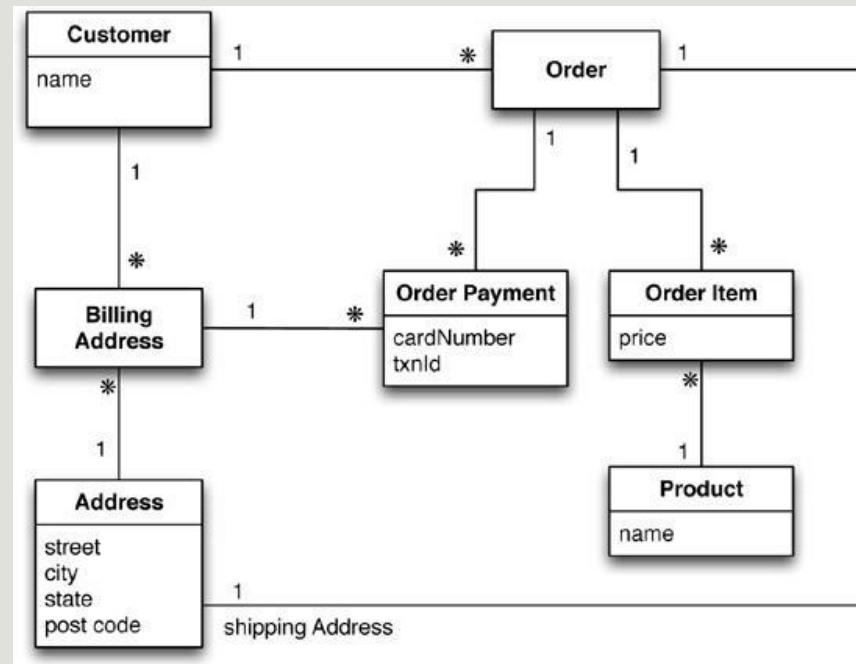
**Aggregate
Modelling**

l'applicazione

modello

realtà

Modello concettuale per un ecommerce



Modellazione relazionale

Customer	
Id	Name
1	Martin

Orders		
Id	CustomerId	ShippingAddressId
99	1	77

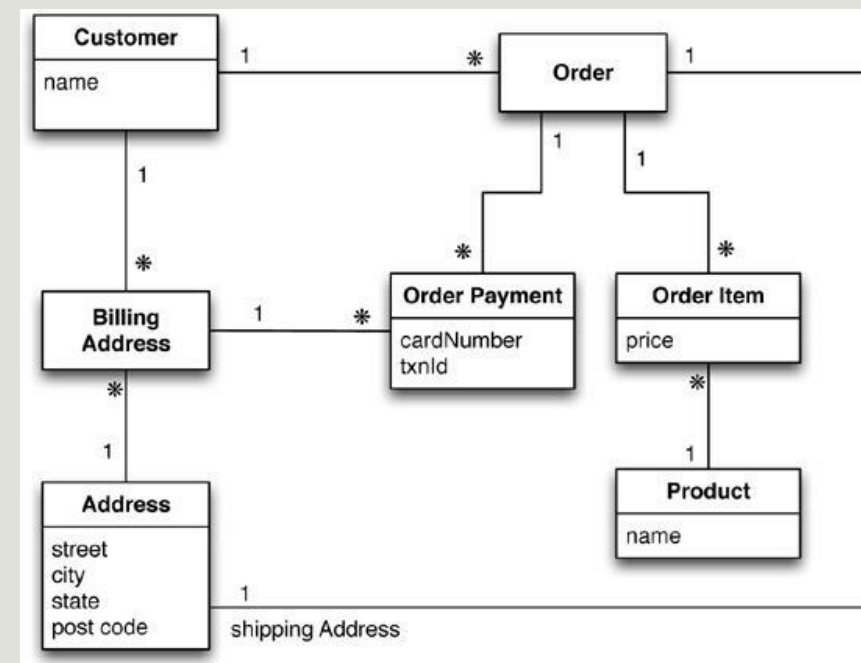
Product	
Id	Name
27	NoSQL Distilled

BillingAddress		
Id	CustomerId	AddressId
55	1	77

OrderItem			
Id	OrderId	ProductId	Price
100	99	27	32.45

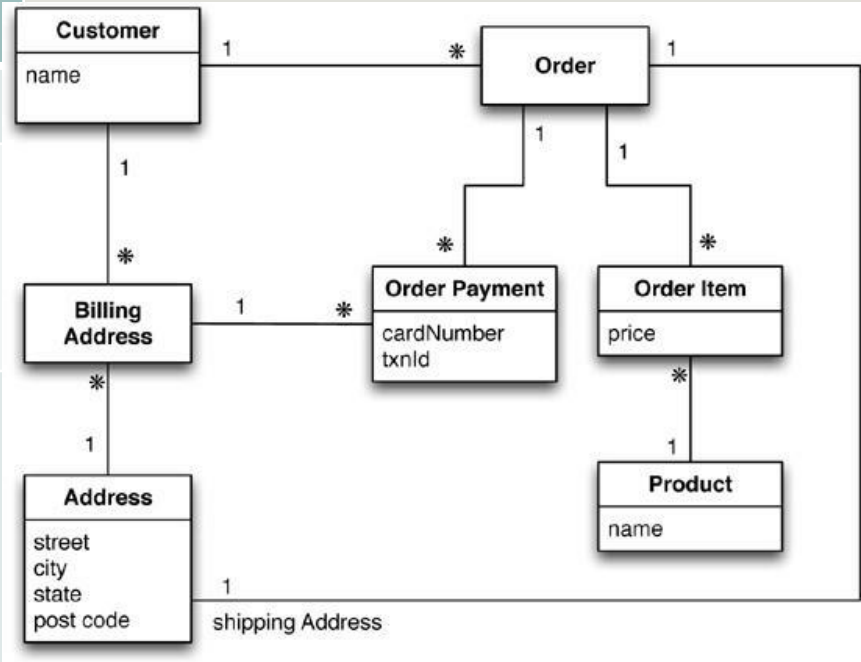
Address	
Id	City
77	Chicago

OrderPayment				
Id	OrderId	CardNumber	BillingAddressId	txnId
33	99	1000-1000	55	abelif879rft



Modellazione key-value

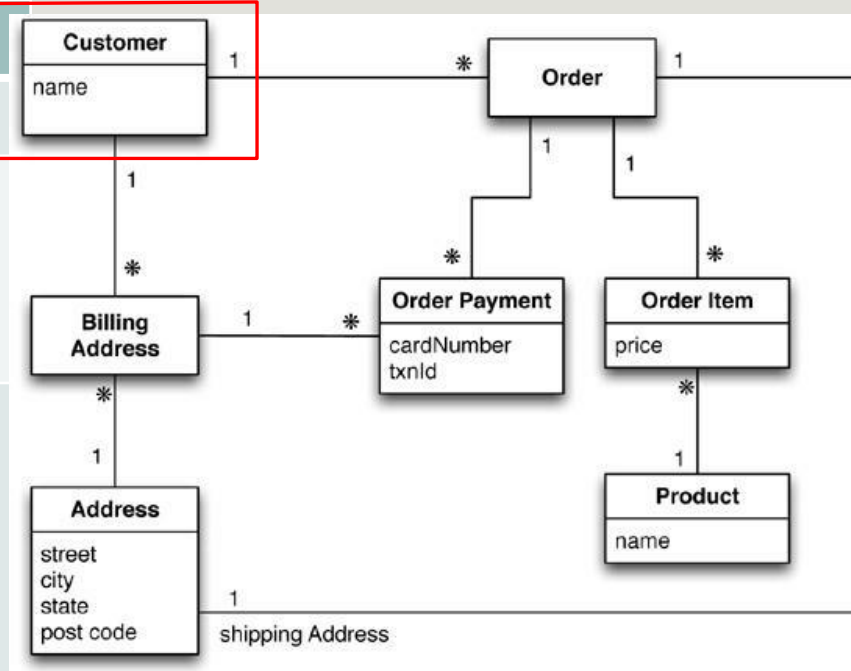
key	value
cust-1:name	Martin
cust-1:adrs	[{"street":"Adam", "city":"Chicago", "state":"illinois", "code":60007}, {"street":"9th", "city":"NewYork", "state":"NewYork", "code":10001}]
cust-1:order-99	<pre>{"orderpayments":[{"card":477, "billadrs":{"street":"Adam", "city":"Chicago", "state":"illinois", "code":60007}}, {"card":457, "billadrs":{"street":"9th", "city":"NewYork", "state":"NewYork", "code":10001}}], "products":[{"id":1, "name":"Cola", "price":12.4}, {"id":2, "name":"Fanta", "price":14.4}], "shipAdrs":{"street":"9th", "city":"NewYork", "state":"NewYork", "code":10001}}</pre>



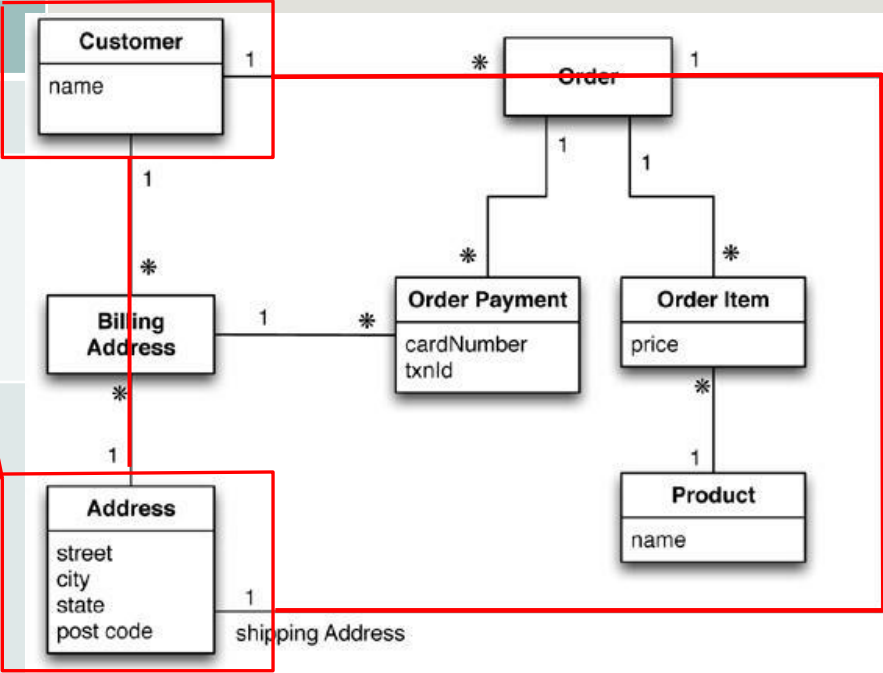
key	value
p-1:name	Cola
p-2:name	Fanta

Modellazione key-value

key	value
cust-1:name	Martin
cust-1:adrs	[{"street":"Adam", "city":"Chicago", "state":"illinois", "code":60007}, {"street":"9th", "city":"NewYork", "state":"NewYork", "code":10001}]
cust-1:order-99	<pre>{"orderpayments":[{"card":477, "billadrs":{"street":"Adam", "city":"Chicago", "state":"illinois", "code":60007}}, {"card":457, "billadrs":{"street":"9th", "city":"NewYork", "state":"NewYork", "code":10001}}], "products":[{"id":1, "name":"Cola", "price":12.4}, {"id":2, "name":"Fanta", "price":14.4}], "shipAdrs":{"street":"9th", "city":"NewYork", "state":"NewYork", "code":10001}}</pre>



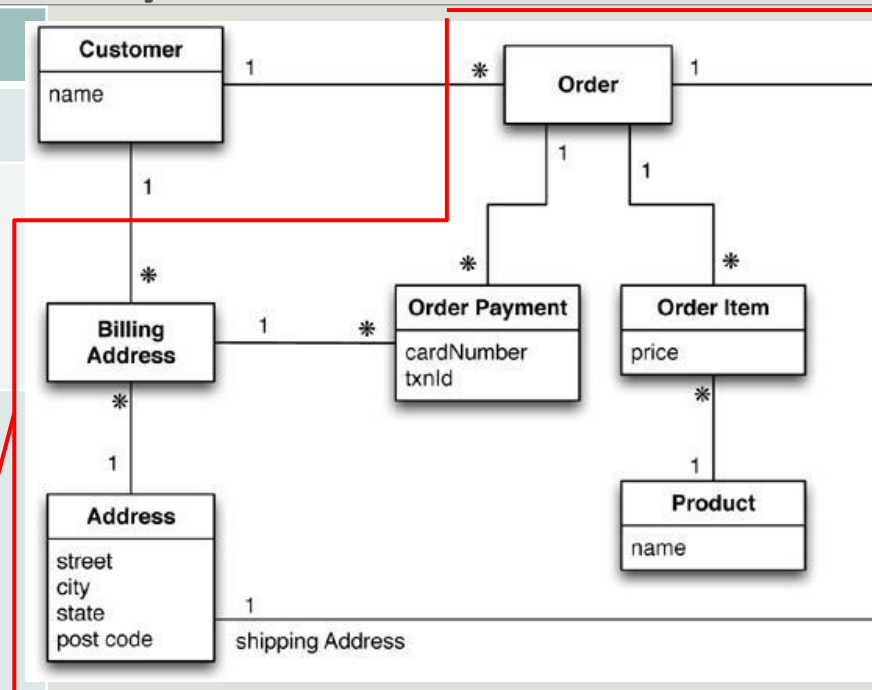
Modellazione key-value

key	value	
cust-1:name	Martin	
cust-1:adrs	[{"street":"Adam", "city":"Chicago", "state":"illinois", "code":60007}, {"street":"9th", "city":"NewYork", "state":"NewYork", "code":10001}]	
cust-1:order-99	{ "orderpayments": [{"card":477, "billadrs":{"street":"Adam", "city":"Chicago", "state":"illinois", "code":60007}}, {"card":457, "billadrs":{"street":"9th", "city":"NewYork", "state":"NewYork", "code":10001}}], "products": [{"id":1, "name":"Cola", "price":12.4}, {"id":2, "name":"Fanta", "price":14.4}], "shipAdrs":{"street":"9th", "city":"NewYork", "state":"NewYork", "code":10001} }	

key	value
p-1:name	Cola
p-2:name	Fanta

Modellazione key-value

key	value
cust-1:name	Martin
cust-1:adrs	[{"street":"Adam", "city":"Chicago", "state":"illinois", "code":60007}, {"street":"9th", "city":"NewYork", "state":"NewYork", "code":10001}]
cust-1:order-99	{"orderpayments":[{"card":477, "billadrs":{"street":"Adam", "city":"Chicago", "state":"illinois", "code":60007}}, {"card":457, "billadrs":{"street":"9th", "city":"NewYork", "state":"NewYork", "code":10001}}], "products":[{"id":1, "name":"Cola", "price":12.4}, {"id":2, "name":"Fanta", "price":14.4}], "shipAdrs":{"street":"9th", "city":"NewYork", "state":"NewYork", "code":10001} }



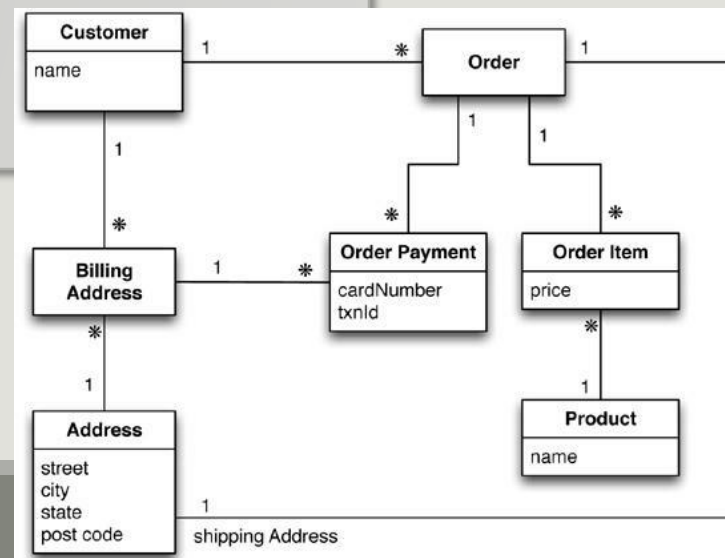
key	value
p-1:name	Cola
p-2:name	Fanta

Modellazione Document-based

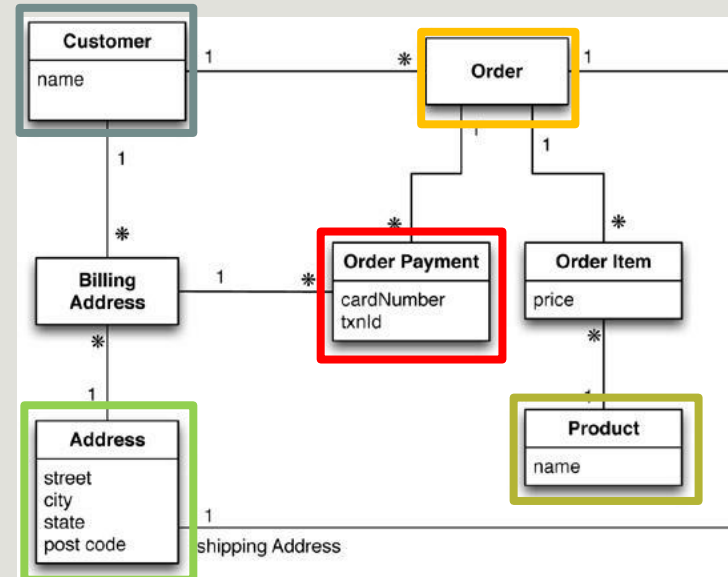
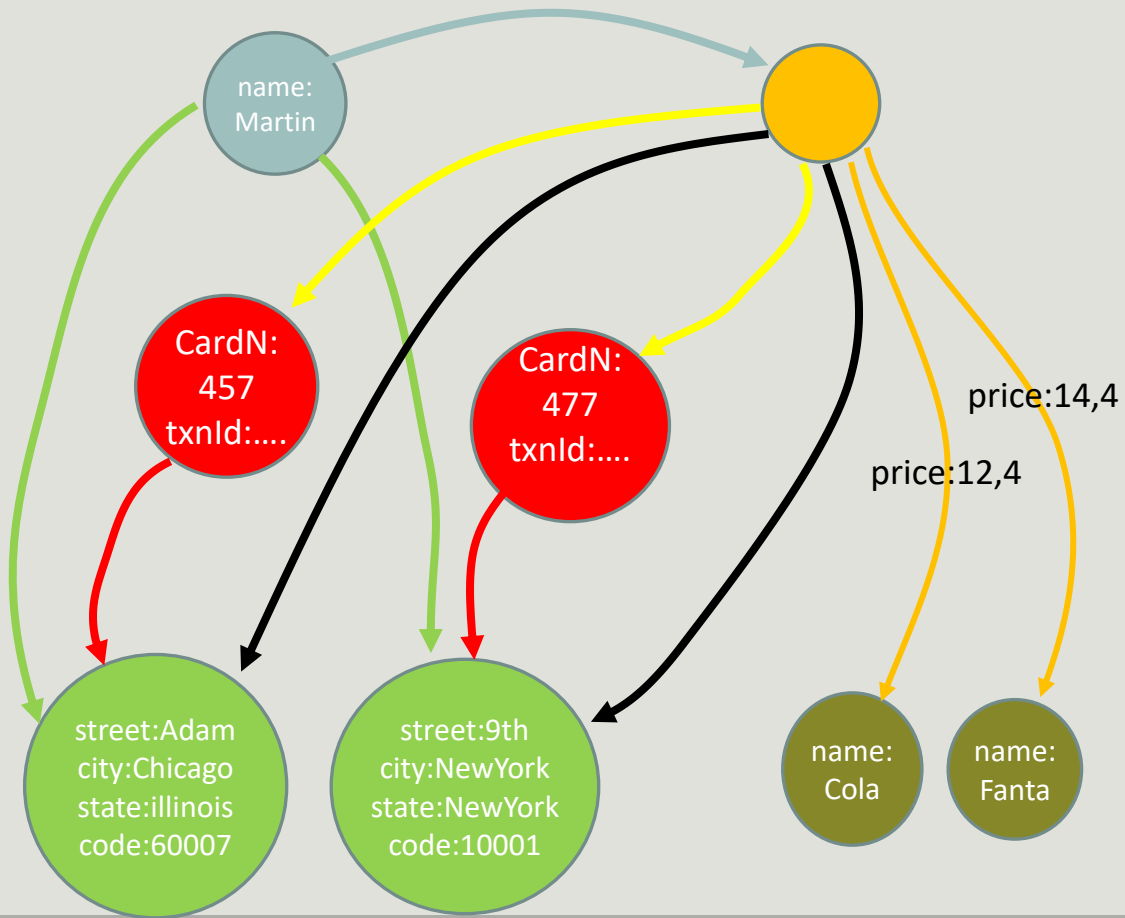
```
{  "_id": 1,
  "name": "Martin",
  "adrs": [ {"street": "Adam", "city": "Chicago", "state": "illinois", "code": 60007},
            {"street": "9th", "city": "NewYork", "state": "NewYork", "code": 10001}
        ]
}
```

```
{ "_id": 1,
  "customer": 1,
  "orderpayments": [ {"card": 477, "billadrs": {"street": "Adam", "city": "Chicago", "state": "illinois", "code": 60007}},
                     {"card": 457, "billadrs": {"street": "9th", "city": "NewYork", "state": "NewYork", "code": 10001}}
                    ],
  "products": [ {"id": 1, "name": "Cola", "price": 12.4},
                 {"id": 2, "name": "Fanta", "price": 14.4}
                ],
  "shipAdrs": {"street": "9th", "city": "NewYork", "state": "NewYork", "code": 10001}
}
```

```
{  "_id": 1,
    "name": "Cola",
    "price": 12.4
},
{ "_id": 2,
  "name": "Fanta",
  "price": 14.4
}
```



Modellazione a grafo



Una realtà frustrante

- ❑ Nella maggior parte delle aziende la **divisione informatica NON è tra le più importanti!**
- ❑ I progetti informatici sono finanziati solo se si riesce a esplicitarne il ritorno economico
 - ✓ Far approvare un progetto è più difficile che realizzarlo
- ❑ Le applicazioni informatiche sono spesso viste come un costo necessario, non come una opportunità



Ma le cose stanno cambiando!

La trasformazione digitale

- ❑ La DT mira a migliorare l'efficienza e l'efficacia delle aziende sfruttando le possibilità offerte dalle nuove tecnologie.
- ❑ Tutti i settori aziendali pubblici e privati saranno coinvolti in questa trasformazione anche se con tempi e modi diversi
- ❑ E' importante sperimentare e capire dove e quando digitalizzare
- ❑ La DT non è solo una questione tecnologica!
 - ✓ Richiede una strategia a lungo termine e un percorso a piccoli passi
 - ✓ Ha bisogno di cambiamenti nella mentalità delle persone e nella ricerca di talenti digitali



Il percorso di digital transformation

L'adozione delle tecnologie digitali è un percorso incrementale e raramente permette di saltare dei passaggi. Saltarli sarebbe *rischioso*, *costoso* e *inutile*

- Il personale non è pronto
 - ✓ Non ha il corretto mindset, non è disponibile al cambiamento, non percepisce il valore
- I dati non sono pronti
 - ✓ Dati di scarsa qualità e che non descrivono i processi a un sufficiente livello di dettaglio
- Le aziende non sono pronte
 - ✓ I processi aziendali non sono adeguati a sfruttare e reagire prontamente

Non fidatevi di consulenti e fornitori di software che vi propongono sistemi avanzati quando la vostra azienda opera ancora con fogli Excel o a malapena utilizza un sistema gestionale

Trasformare l'azienda in data-driven

Il termine *data-driven company* indica le aziende in cui le decisioni e i processi sono supportati dai dati

- Le decisioni si basano su una conoscenza quantitativa piuttosto che qualitativa
- Processi e Conoscenze sono un patrimonio dell'azienda e non vanno perduti se cambiano i manager

La differenza tra una decisione data-driven e una *buona* decisione è un *buon* manager

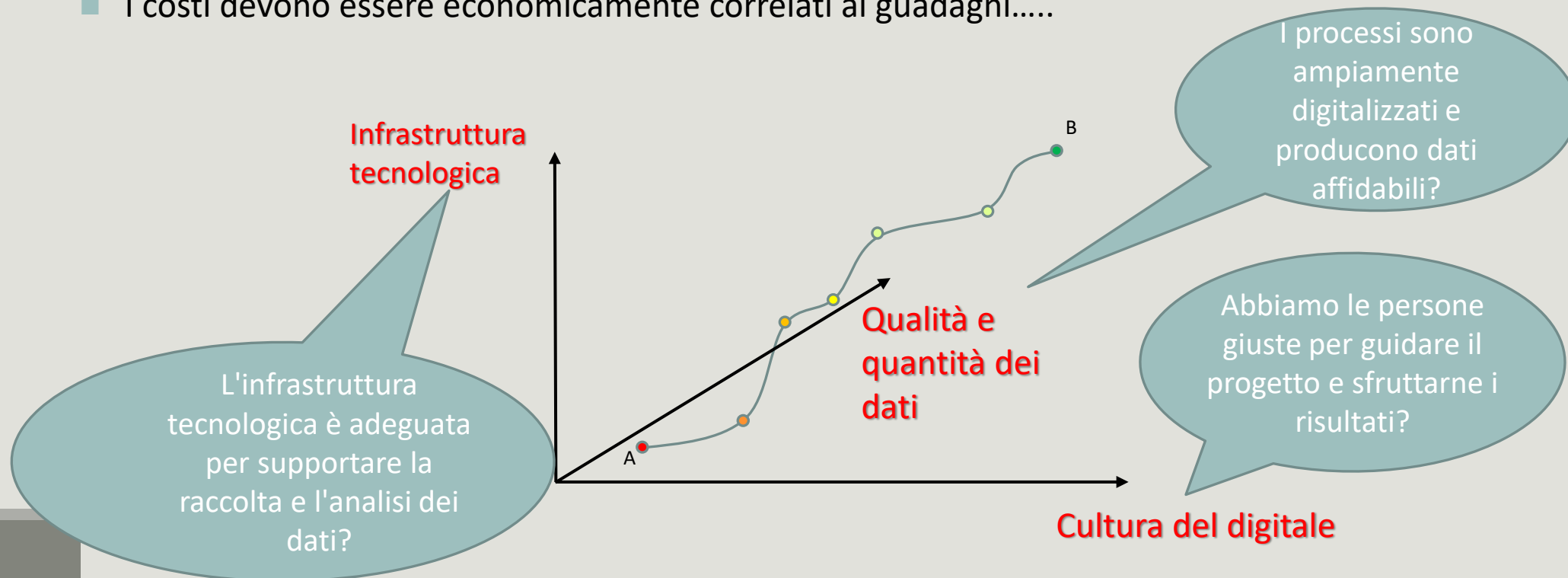
L'adozione di una mentalità basata sui dati va ben oltre l'adozione di una soluzione di business intelligence e comporta:

- ✓ Creare una cultura dei dati
- ✓ Cambia la mentalità dei manager
- ✓ Cambiare i processi
- ✓ Migliora la qualità di tutti i dati

Trasformare l'azienda in data-driven

La **Digitalizzazione** è un percorso che coinvolge tre dimensioni. Spostarsi tra A e B è un processo pluriennale fatto di obiettivi intermedi ognuno dei quali deve essere fattibile

- Risolve un problema aziendale e portare valore
- Può essere realizzato in un intervallo di tempo limitato (in genere meno di un anno)
- I costi devono essere economicamente correlati ai guadagni.....

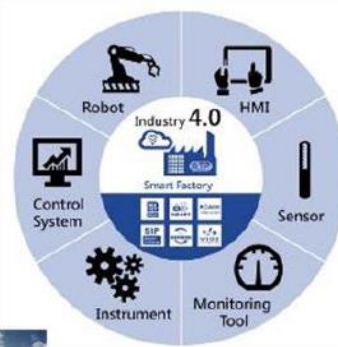
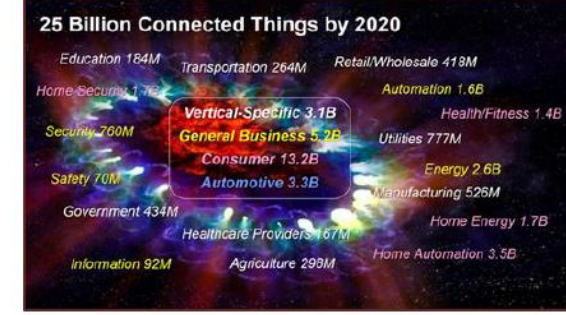
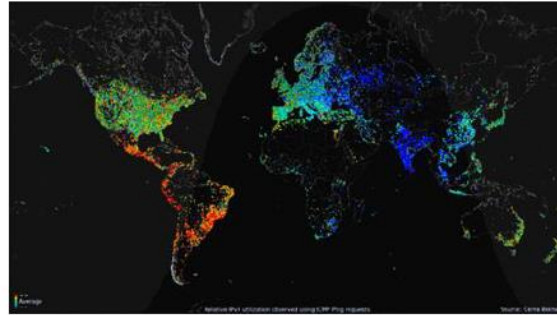


La Data Revolution

- ❑ I dati rappresentano il principale combustibile che alimenta la trasformazione digitale
- ❑ La digitalizzazione è iniziata negli anni '70s con la progressive diffusione dei calcolatori dando il via al processo di digitalizzazione dei processi e delle Informazioni che continua ad accelerare ancora oggi cambiando nome ma non obiettivo
 - ✓ Post-industrial society
 - ✓ Information technology revolution
 - ✓ Digital age
- ❑ Possiamo stimare l'inizio della **Digital Age** nel 2002, quando nel mondo sono state archiviate più informazioni digitali che analogiche. Alla fine degli anni '80 meno dell'1% delle informazioni era in formato digitale, mentre nel 2012 la percentuale era salita al 99% con un incremento annuo di circa il 30%, che porta ad un raddoppio delle informazioni conservate in meno di 3 anni.

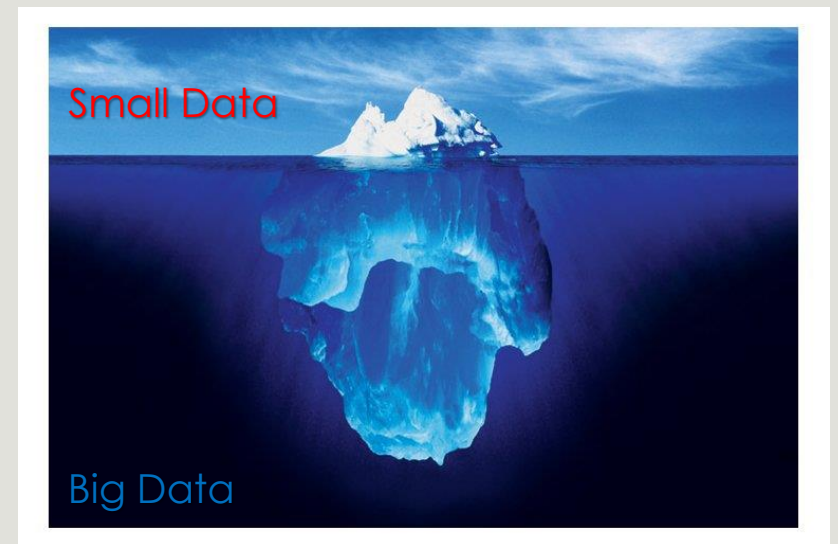
Chi produce i dati nella digital age?

- I sistemi informativi non sono più limitati ai dati prodotti dai processi aziendali ma vanno ripensati per permettere di sfruttare tutti i dati utili all'azienda e per poter supportare processi interni ed esterni

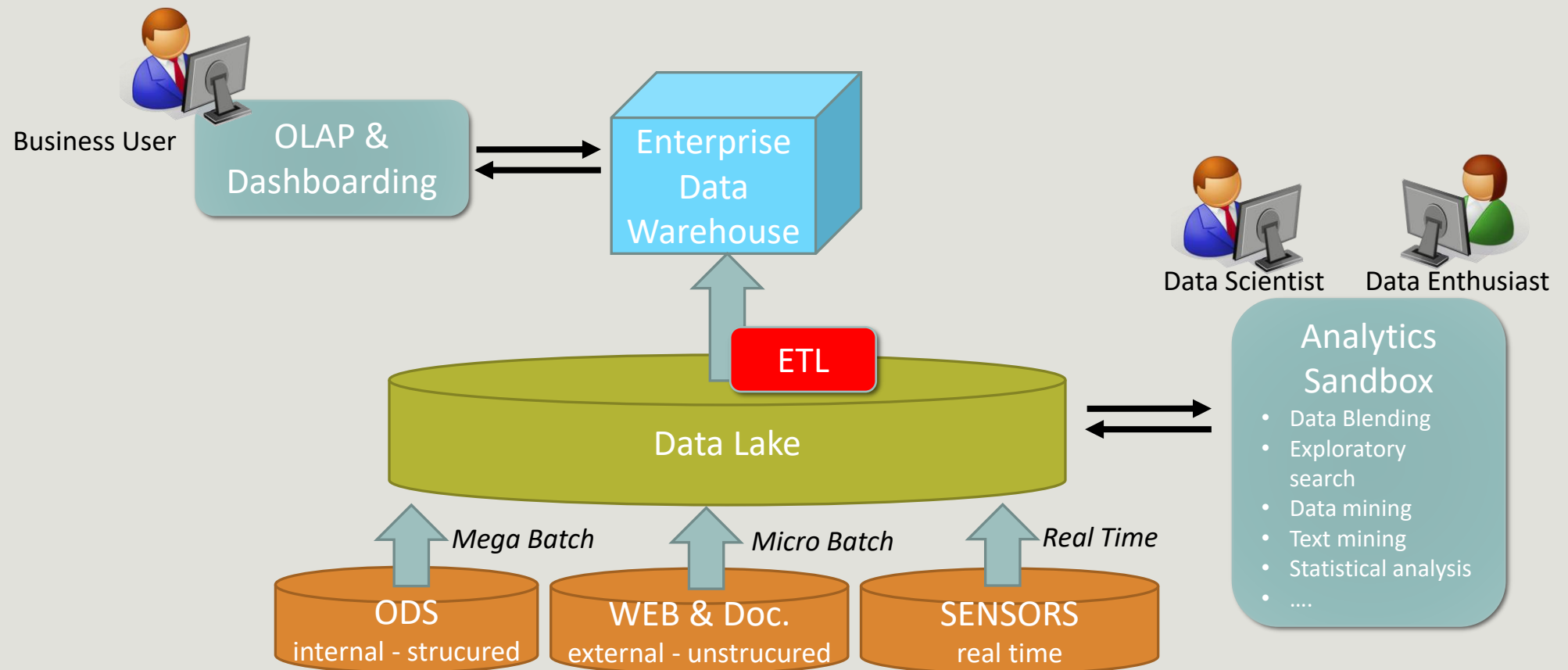


Big Data vs Small Data

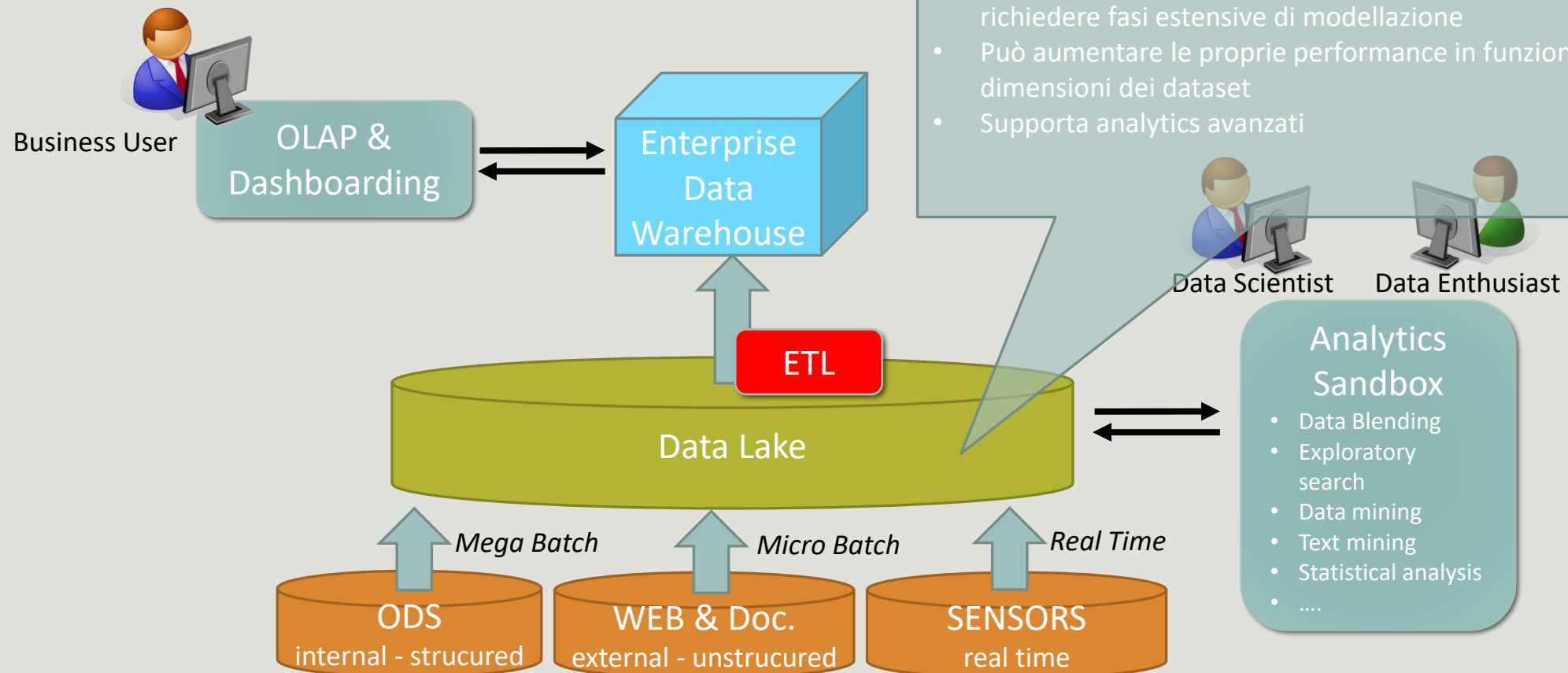
- ❑ La progressiva digitalizzazione di servizi e impianti genera una enorme massa di dati eterogenei e in tempo reale
- ❑ I Big Data devono essere trasformati in Small data affinché possano essere sfruttati ai fini decisionali
- ❑ Per gestire questa trasformazione occorrono
 - ✓ Tecnologia ad hoc (NO SQL DBMS)
 - ✓ Potenza di calcolo (cluster computing)
 - ✓ Sistemi automatizzati (Intelligenza artificiale)



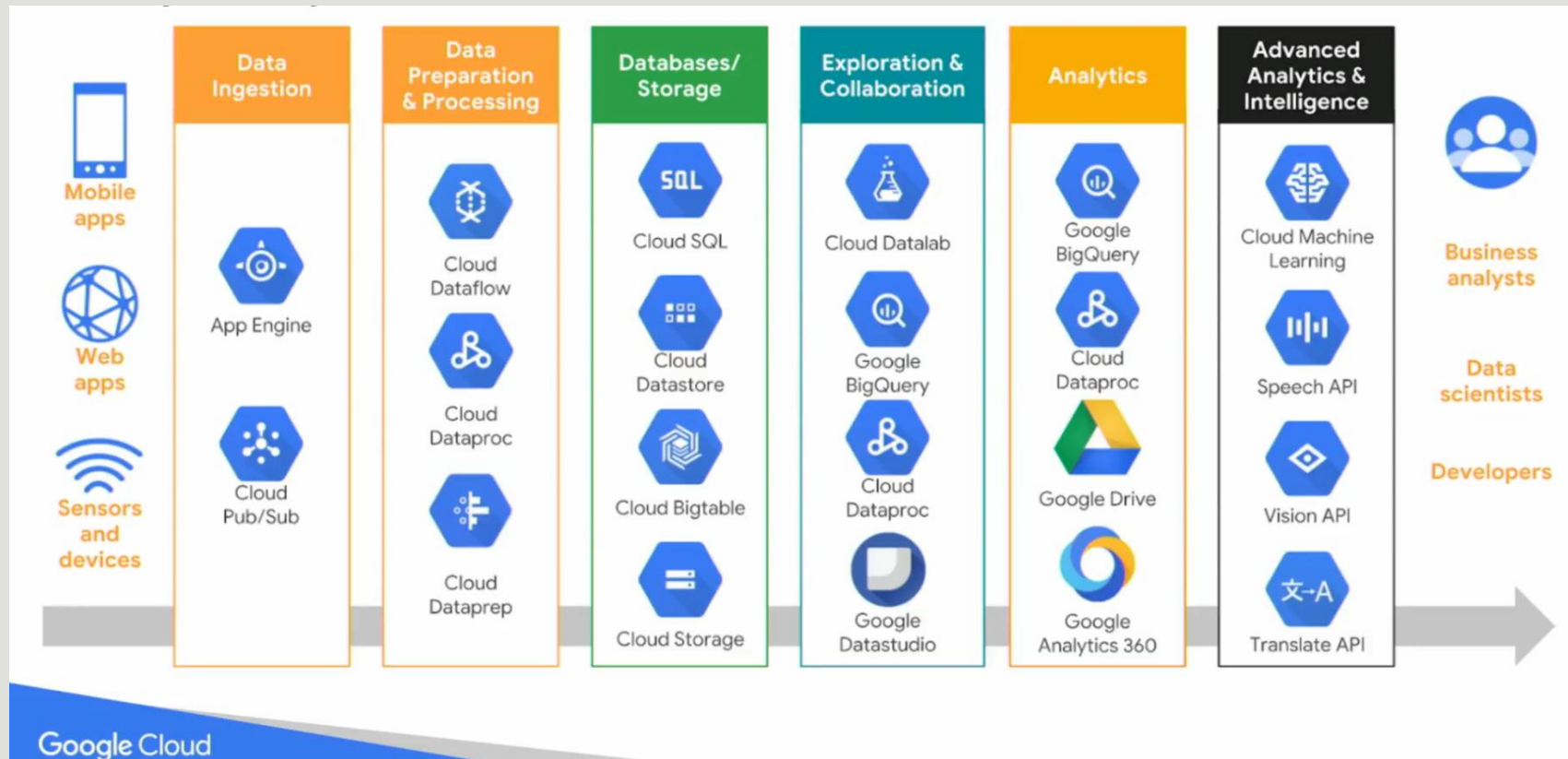
Un'architettura per i Big Data



Un'architettura per i Big Data



Oltre i Data Lake: le Data Platform



Considerazioni finali

Il principale carburante che alimenta la trasformazione digitale è la maggiore disponibilità di dati

Oggi, non esistono applicazioni informatiche che non siano data-intensive

- Digital Twins
- Data mesh
- AI

La capacità di modellare e di analizzare i dati rappresenta una competenza cruciale e qualificante

- Quando il dominio diventa complesso (è normale avere 100-200 relazioni in un applicativo)
- Quando i dati sono molti (1 M di tuple con interrogazioni che coinvolgono 5 relazioni)
- Quando le relazioni tra i dati sono complessi (DW, AI)



Salim Ismail
**Exponential
Organizations**

Con Michael S. Malone e Yuri van Geest
Prefazione e postfazione di Peter H. Diamandis
Presentazione di Fabio Troiani

Il futuro del business mondiale

Marsilio NODI