Tutorship per la preparazione ai concorsi della Banca d'Italia.

Tecnologie di data management:

- Big Data
- Data Governance
- Analisi dati avanzata

CIDA ACADEMY
Formazione specialistica del Sindirettivo Banca Centrale

CIDA
MANAGER E ALTE PROFESSIONALITÀ PER L'ITALIA

Agenda

- Big Data: i driver
- Approccio DWH vs Data Lake
- Data Governance: le sfide
- Analisi dati avanzate: gli strumenti per i data scientist e gli analisti

Big Data:
i driver

CIDA ACADEMY
Formazione specialistica del Sindirettivo Banca Centrale

Crescente disponibilità di diversi tipi di dati

- fonti di dati granulari strutturati o non strutturati
  dati amministrativi, dati web, notizie e piattaforme per blog, dati di pagamento, dati testuali
- estrazioni di dati relativi a temi riguardanti imprese, famiglie, finanza, mercato del lavoro o pubbliche amministrazioni
  dati provenienti da Google, Twitter, notizie Factiva feed finanziari, dati sui brevetti europei

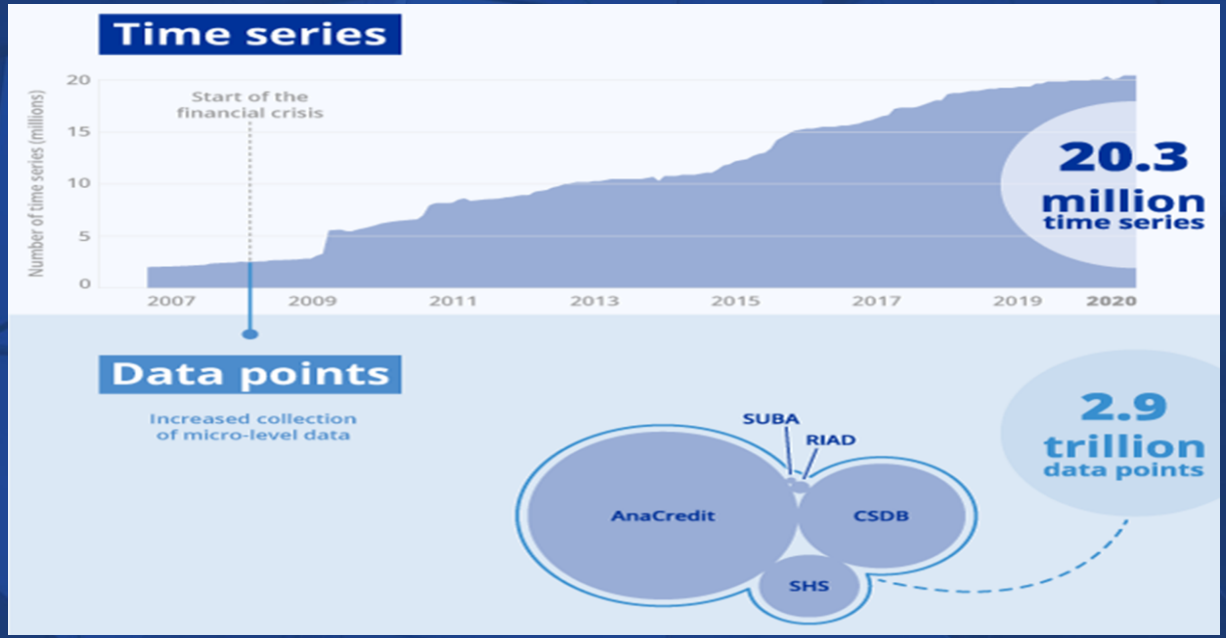
Big Data:
i driver

CIDA ACADEMY
Formazione specialistica del Sindirettivo Banca Centrale

Rapida espansione di tecniche di modellizzazione derivanti dall'apprendimento automatico o statistico

- machine learning (deep learning) per le previsioni economiche
  random forest, tecniche di bagging e bootstrapping
- regressione con tecniche di regolarizzazione
  modellizzazione con molte covarianti o forti non linearità
- text mining per classificazione degli articoli e sentiment analysis

Dati al servizio della politica monetaria e della stabilitá del mercato

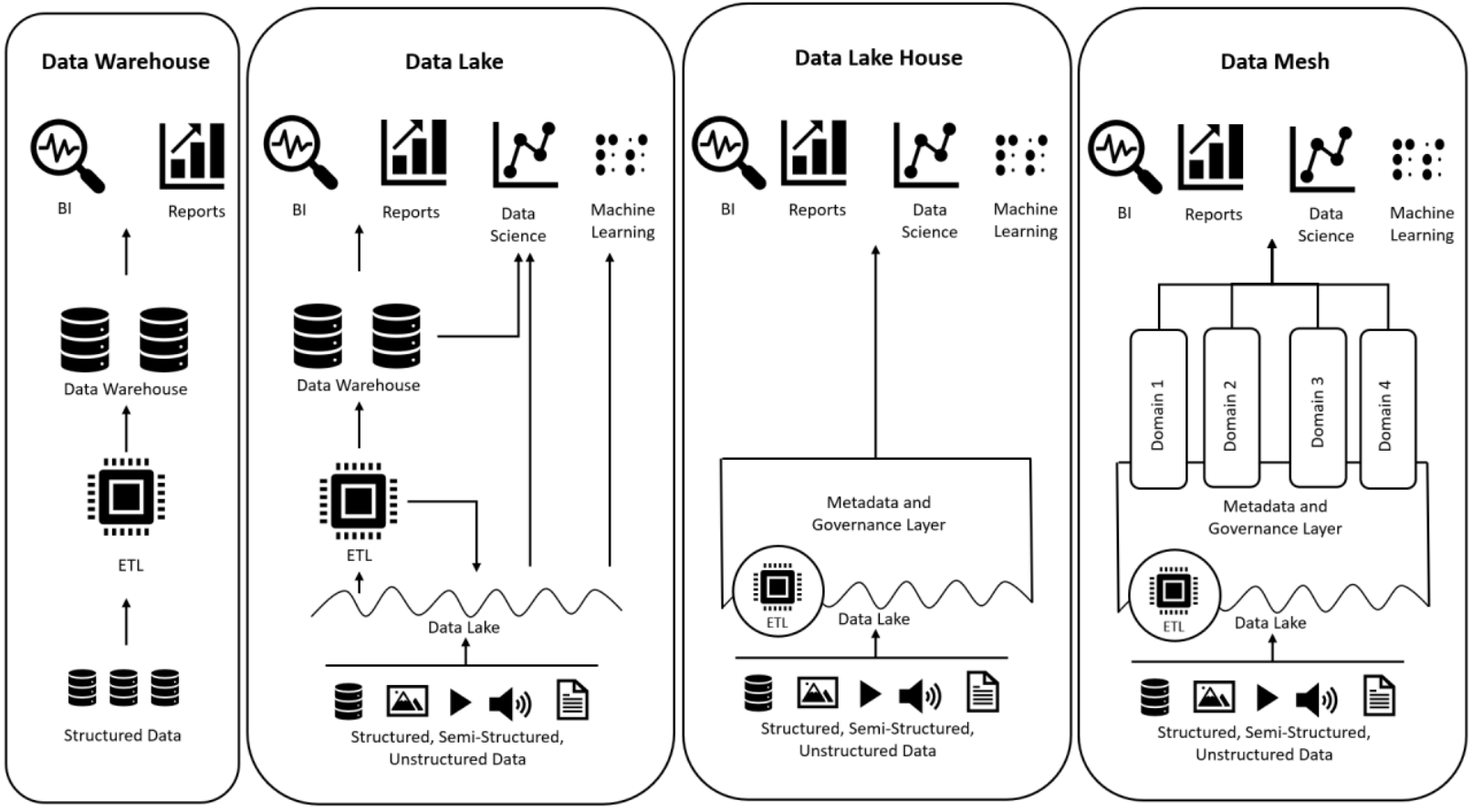


Source: Don't take it for granted: the value of high-quality data and statistics for the ECB

Data Warehouse VS Data Lake

1	Un approccio diverso ai dati	**Conosco il dato, lo modello, lo memorizzo, lo elaboro** **VS** **Memorizzo il dato, lo conosco, lo elaboro, lo modello**
2	Tecniche e strumenti innovativi	**Devo effettuare interrogazioni predefinite** **VS** **Interpreto i dati e derivo conoscenza**
3	Il superamento dei silos informativi	**Devo conoscere l'applicazione di origine** **VS** **Cerco un dato, lo trovo, lo utilizzo**

L'Evoluzione continua...



Source: [Data Warehouse vs. Data Lake vs. Data Lakehouse vs. Data Mesh]

Data Governance:
disciplinare la raccolta, l'accesso, la conservazione e l'utilizzo dei dati

CIDA ACADEMY
Formazione specialistica del Sindirettivo Banca Centrale

Le sfide

- Integrare e gestire il patrimonio informativo fornendo un punto unico di accesso ai dati
- Gestire il ciclo di vita e la qualità del dato
- Garantire l'accesso ai dati con il dovuto *need-to-know*
- Evitare la creazione di copie

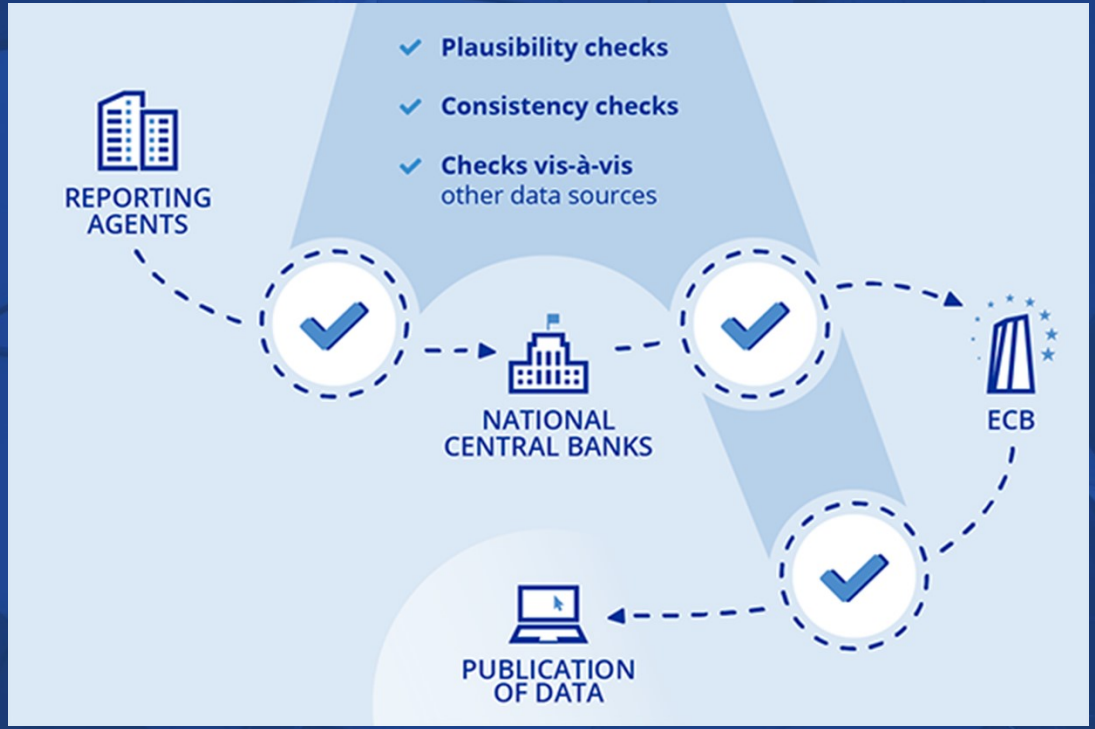
Analisi dati avanzate:
fornire gli strumenti più adatti in funzione dell'obiettivo

CIDA ACADEMY
Formazione specialistica del Sindirettivo Banca Centrale

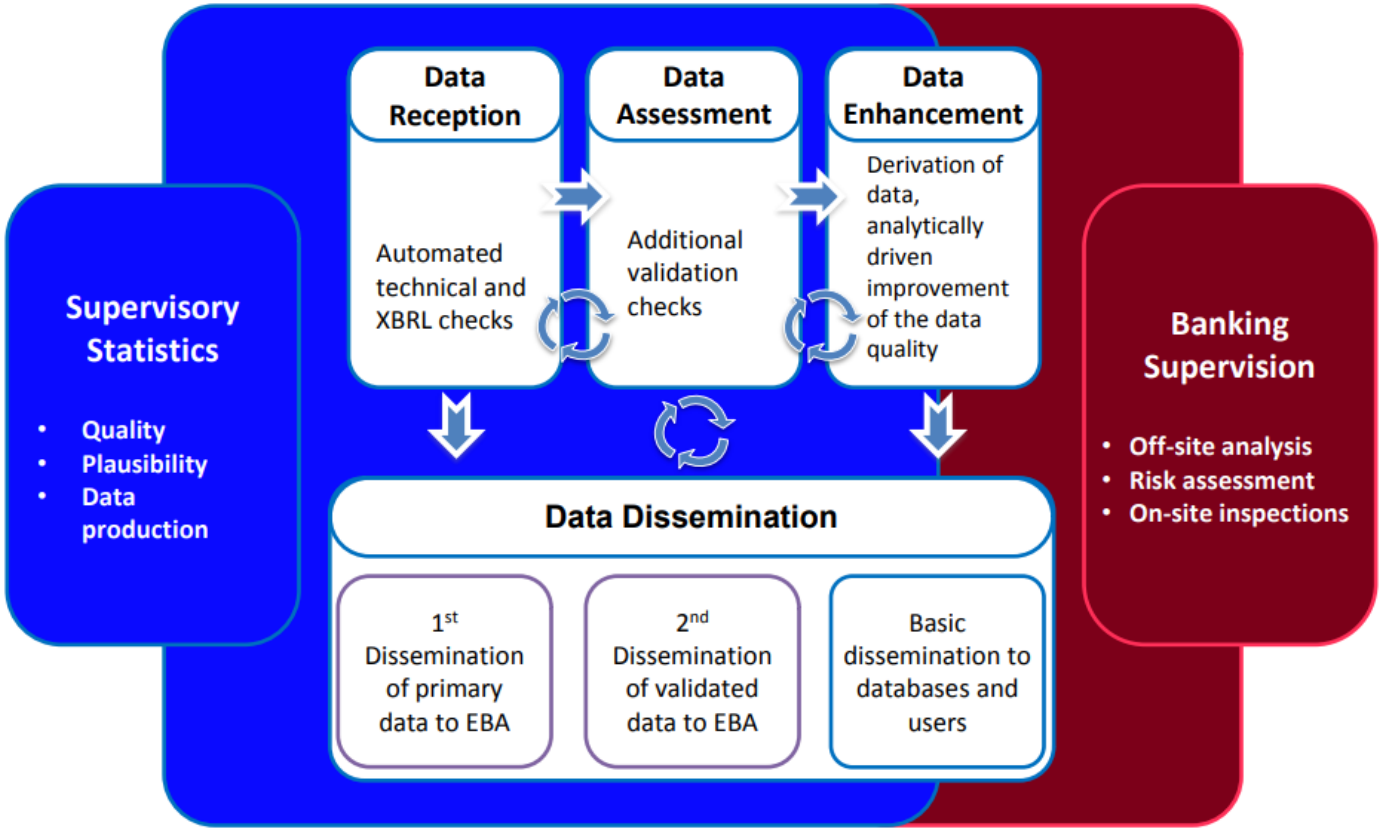
Il toolset dei *data scientist* e degli analisti

- procedure *self-service* e aree di lavoro dove autonomamente caricare dati e definire le relative strutture per l'interrogazione
- accedere ai dati di produzione senza impatto sulle applicazioni
- utilizzo di hardware specializzato
- piattaforme per analisi programmatiche interattive e visuali
- strumenti di collaborazione senza frizioni

I dati nel contesto europeo:
Una collaborazione continua



Un esempio di DataSet Europeo … SUBA



Source: www.eurofiling.info/201411/presentations/20141125SupervisoryBankingDataSystemGiancarloPellizzari.pdf

Il Futuro: A European data Platform for banking system

European

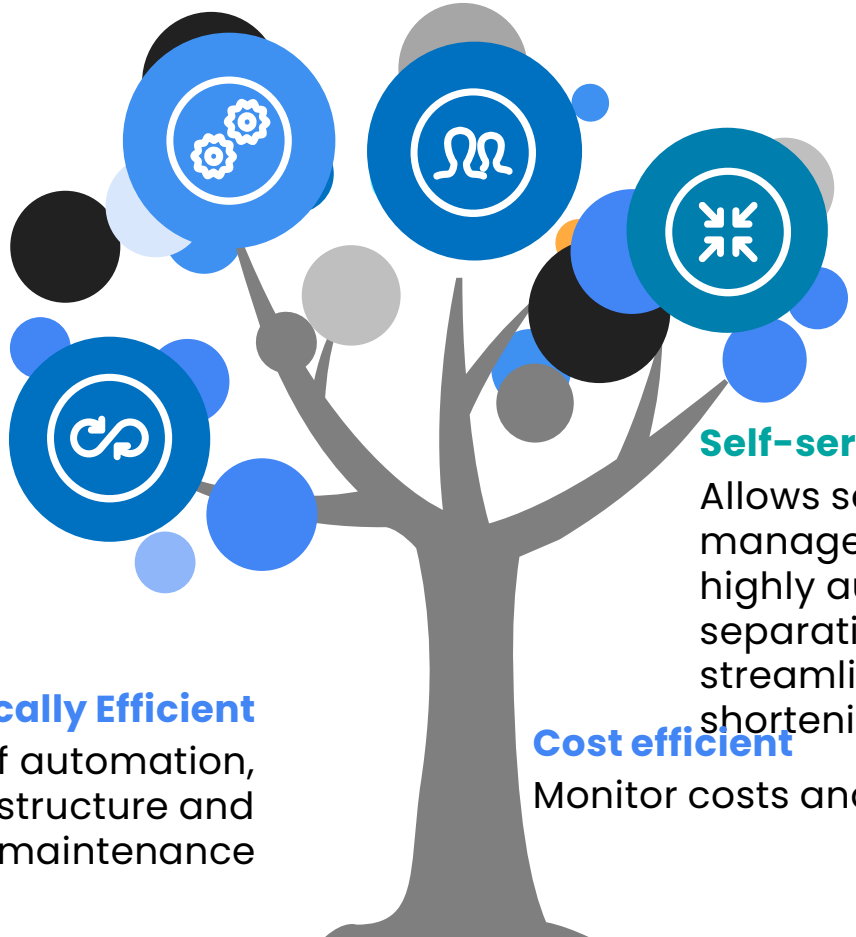
Ability to access and integrate data across various data sources in the Eurosystem

Multitenancy

Shared tooling and infrastructure while allowing for separation of concerns and independence

Technically Efficient

High degree of automation, shared infrastructure and maintenance



Future proof - innovative

Modern infrastructure and tools for all users within ESCB/SSM, keep up to date with the newest innovations including AI

Self-service

Allows setup, operation and management of data pipelines in a highly automated manner, improving separation of IT and business concerns, streamlining implementation, and shortening time-to-market

Cost efficient

Monitor costs and pay per use

CIDA ACADEMY
Formazione specialistica del Sindirettivo Banca Centrale

Grazie dell'Attenzione

Domande?

Data Design Patterns	Data Warehouse	Data Lake	Data Lakehouse	Data Mesh
What it is	A subject-oriented data architecture that integrates detailed data in a consistent way while maintaining a non-volatile history of it.	A set of long-term data containers for managing and refining raw data, using low-cost object storage often delivered from the cloud.	A combination of a data warehouse and a data lake.	A domain-driven data design pattern divided either logically or physically among the teams working in those domains.
Benefits	Generates actionable insights (e.g., in dashboards) from huge amounts of curated data, including the creation of predictive analytics and dashboards that drive operational actions. It aggregates data from all enterprise sources in a central location with consistent governance and supports sandboxes for new idea testing.	Captures previously discarded “dark data” to drive innovation later on and stores data as-is without having to structure it first. The lake also allows insights to be efficiently captured by AI and machine learning services analysing raw information.	Enables an enterprise to systematically extract insights in the mode of a data warehouse — via SQL, machine learning, or any other process — while taking advantage of the vast scale and low costs of a data lake.	Data mesh allows for autonomous active management of data by the teams closest to it and permits increased agility because there’s no central bottleneck. Each team can create its own data products.
Limitations	Not ideal for big data use cases that require the storage and extraction of value from large amounts of raw data, such as that created by IoT devices and web and mobile sources.	Relatively few off-the-shelf tools are available for data lakes, which necessitates significant experience with open-source software. There’s also a high risk of silos due to limited governance, and there can be great difficulty balancing issues between security and ease of access.	Limited agility in adding new features because everything is centralized and monolithic. Data engineers end up spending a lot of their time cleaning up data from teams that have limited incentive to ensure their information is accurate as it goes in.	It’s a relatively new architecture that enterprises are still working out. Performance and governance may suffer because users need to go over the network every time to access different data. Without cross-domain governance and semantic linking of data, it can become very siloed and yield disappointing results.