# Corsi di preparazione concorsi interni - ruolo tecnico

Metodologie e strumenti per l'analisi dei dati – Advanced Analytics


# Big Data & Data Science

## Business intelligence vs. advanced analytics

| | BUSINESS INTELLIGENCE | ADVANCED ANALYTICS |
|---|---|---|
| **Answers the questions:** | ▪ What happened?<br>▪ When?<br>▪ Who?<br>▪ How many? | ▪ Why did it happen?<br>▪ Will it happen again?<br>▪ What will happen if we change *x*?<br>▪ What else does the data tell us that we never thought to ask? |
| **Includes:** | ▪ Reporting (KPIs, metrics)<br>▪ Automated monitoring and alerting (thresholds)<br>▪ Dashboards<br>▪ Scorecards<br>▪ OLAP (cubes, slice and dice, drilling)<br>▪ Ad hoc query<br>▪ Operational and real-time BI | ▪ Statistical or quantitative analysis<br>▪ Data mining<br>▪ Predictive modeling<br>▪ Multivariate testing<br>▪ Big data analytics<br>▪ Text analytics |

# The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015
**4.4 MILLION IT JOBS**
will be created globally to support big data, with 1.9 million in the United States

## Volume
### SCALE OF DATA

**40 ZETTABYTES**
[ 43 TRILLION GIGABYTES ]
of data will be created by 2020, an increase of 300 times from 2005


It's estimated that
**2.5 QUINTILLION BYTES**
[ 2.3 TRILLION GIGABYTES ]
of data are created each day

**6 BILLION PEOPLE**
have cell phones


Most companies in the U.S. have at least
**100 TERABYTES**
[ 100,000 GIGABYTES ]
of data stored

## Variety
### DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be
**150 EXABYTES**
[ 161 BILLION GIGABYTES ]

By 2014, it's anticipated there will be
**420 MILLION WEARABLE, WIRELESS HEALTH MONITORS**

**4 BILLION+ HOURS OF VIDEO**
are watched on YouTube each month

**30 BILLION PIECES OF CONTENT**
are shared on Facebook every month

**400 MILLION TWEETS**
are sent per day by about 200 million monthly active users

## Velocity
### ANALYSIS OF STREAMING DATA

The New York Stock Exchange captures
**1 TB OF TRADE INFORMATION**
during each trading session

Modern cars have close to
**100 SENSORS**
that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be
**18.9 BILLION NETWORK CONNECTIONS**
– almost 2.5 connections per person on earth

## Veracity
### UNCERTAINTY OF DATA

**1 IN 3 BUSINESS LEADERS**
don't trust the information they use to make decisions

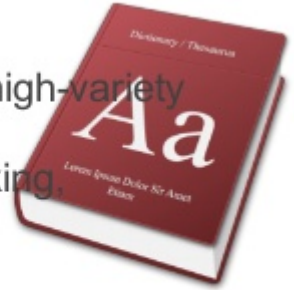
Poor data quality costs the US economy around
**$3.1 TRILLION A YEAR**


in one survey were unsure of how much of their data was inaccurate

**Sources:** McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

IBM

# Defining Big Data




- **Gartner:** High-volume, high-velocity, and/or high-variety information assets that require new forms of processing to enable enhanced decision-making, insight discovery and process optimization.
- **IBM**: Datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze.
- **NY Times:** Shorthand for advancing trends in technology that open the door to a new approach to understanding the world and making decisions.
- **McKinsey:** Large pools of data that can be brought together and analyzed to discern patterns and make better decisions

## Artificial Intelligence

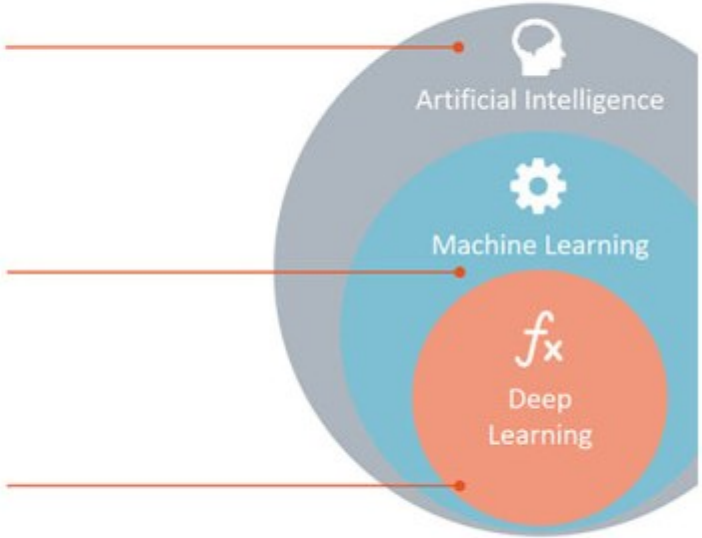
Any technique which enables computers to mimic human behavior.

## Machine Learning

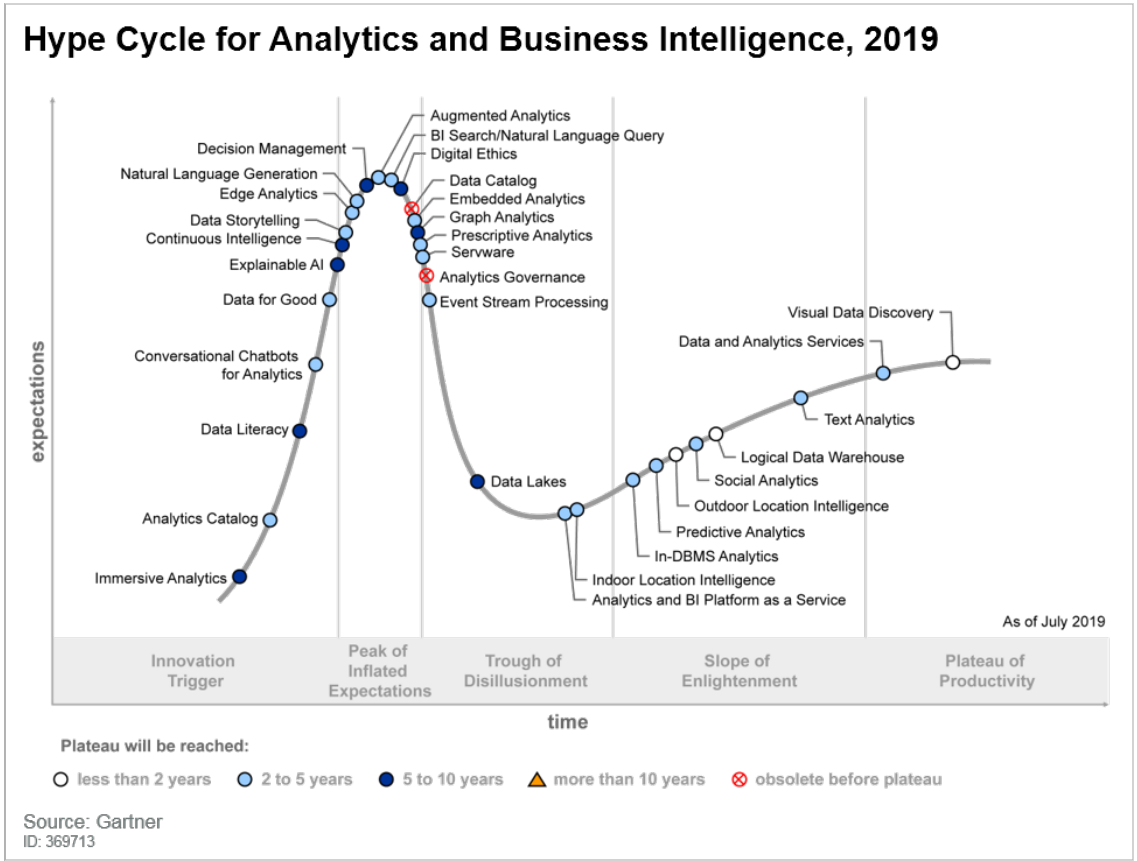
Subset of AI techniques which use statistical methods to enable machines to improve with experiences.

## Deep Learning

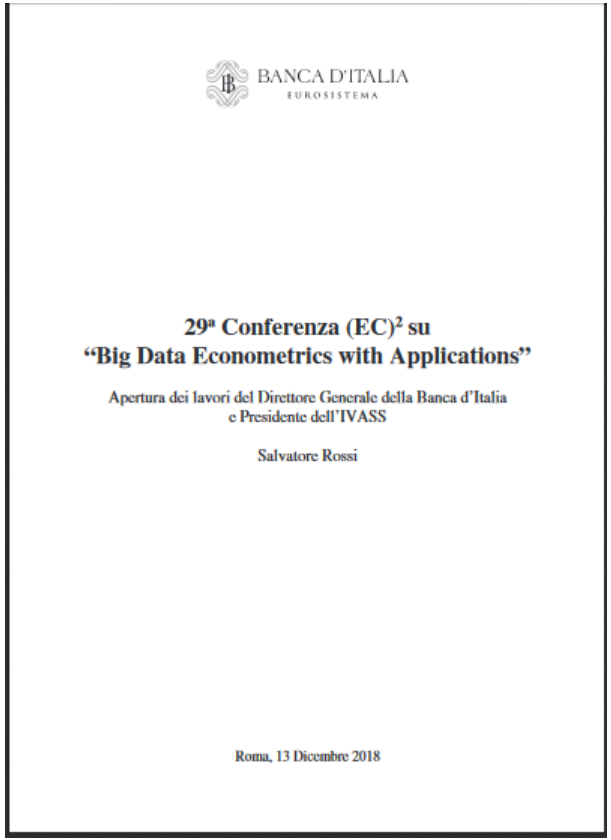
Subset of ML which make the computation of multi-layer neural networks feasible.



# Hype Cycle Analytics & BI



# Big Data in Banca d'Italia



BANCA D'ITALIA
EUROSISTEMA

**29ª Conferenza (EC)² su**
**"Big Data Econometrics with Applications"**

Apertura dei lavori del Direttore Generale della Banca d'Italia
e Presidente dell'IVASS

Salvatore Rossi

Roma, 13 Dicembre 2018

Appare sempre più chiaro, tuttavia, che i *big data* potrebbero segnare la fine dell'econometria così come la conosciamo. La raccolta di dati provenienti dai *social media* ha generato basi di dati sui comportamenti e sulle interazioni degli agenti economici che, ancorché non strutturate, presentano un'ampiezza e una complessità senza precedenti. Tutto ciò si sta rivelando una miniera d'oro in termini di informazione economica.

La Banca d'Italia, in ragione della molteplicità delle funzioni che svolge, segue con estrema attenzione l'evoluzione delle piattaforme *big data.* Queste saranno oggetto di un seminario organizzato in collaborazione con la Banca dei Regolamenti Internazionali il prossimo 15 gennaio qui a Roma.

Negli ultimi anni abbiamo fatto passi avanti nel nostro percorso di apprendimento. Abbiamo coniugato competenze economiche, econometriche, statistiche e informatiche per lavorare con dati in continua crescita in termini di volumi, eterogeneità e velocità. Li abbiamo utilizzati per stimare la disoccupazione e l'inflazione, per migliorare le nostre previsioni economiche, per misurare il clima di fiducia di consumatori e imprese.

Il Sole 24 ORE
27-MAR-2018
Pagina 3
Foglio 1

**L'authority.** Il vice dg Fabio Panetta: sui social per monitorare le attese di inflazione

# Bankitalia vigila anche su Twitter: «Faro sulla fiducia dei depositanti»

**Davide Colombo**
ROMA

Alla strumentazione tradizionale utilizzata da Bankitalia per misurare le aspettative di inflazione o valutare la fiducia dei risparmiatori si aggiungono ora i Big data e il monitoraggio diretto dei social network. Lo ha annunciato ieri il vice direttore generale, Fabio Panetta, nel suo intervento di saluto anche per analisi di impatto delle politiche economiche o per (qui il riferimento di Panetta è stato all'enorme volume di dati granulari raccolto dal Ssm sui singoli prestiti degli istituti oppure sulla rendicontazione statistica del mercato monetario giornaliero da parte del Sistema europeo di banche centrali e sui depositi commerciali previsti dal regolamento europeo sulle infrastrutture di mercato).

**IL TREND**
La Banca d'Italia ha costituito un proprio team sui Big data che include statistici, economisti e informatici per affrontare analisi innovative



# La Banca d'Italia studia i nostri tweet per prevedere l'inflazione

di **Pietro Deragni**
26 MAR, 2018

Bankitalia ha un team digitale che monitora i social media per prevedere gli effetti sulla fiducia dei consumatori e sui prezzi degli immobili. Analisi anche sulla stabilità finanziaria

CORRIERE DELLA SERA

*Il commento*

# Banca d'Italia suona la sveglia su Big Tech e Big Data

di **Daniele Manca**

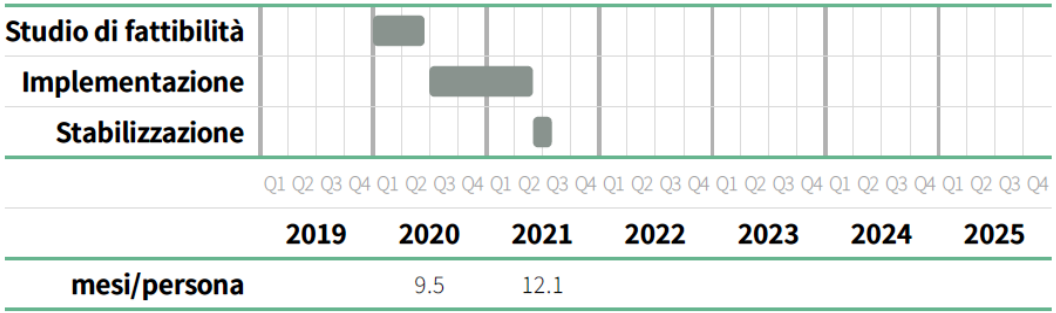
Siamo abituati a parlare dell'Italia come un Paese immobile. E sicuramente lo siamo in molti settori. Ma a giudicare dal workshop organizzato dalla Banca d'Italia sui Big Data, qualcosa si sta muovendo sul fronte dell'innovazione. Di modificando e influenzando i comportamenti. Il 40% delle persone nel nostro Paese hanno accesso al conto bancario, e molti ormai acquistano, attraverso lo smartphone. Utilizzando i dati che noi depositiamo sulle varie piattaforme e tramite l'intelligenza artificiale, le big tech sono in grado di analizzare e indicare comportamenti e aspettative dei consumatori. Cosa che le banche dovranno imparare in fretta a fare. Basta guardare a come si sono mossi i nuovi attori nel sistema dei pagamenti da PayPal alle nuove nate Satispay ma

# SUPTECH - ANALISI ESPOSTI

Progetto 908-P0691 Classe 2 - Priorità 804

Dipartimento Tutela della clientela ed educazione finanziaria
SERVIZIO TUTELA INDIVIDUALE DEI CLIENTI

Realizzazione di un'applicazione, denominata EspTech, a sostegno delle attività di ricerca e indagine conoscitiva sugli esposti privatistici pervenuti alla Banca e di identificazione precoce di fenomeni emergenti d'interesse per la vigilanza.




| | |
|---|---|
| **mesi/persona** | 9.5 12.1 |

L'iniziativa soddisfa esigenze nell'ambito **FinTech**.

**Studio**: dal 1 gennaio 2020 al 15 giugno 2020
**Realizzazione**: dal 1 luglio 2020 al 31 luglio 2021

GPDP | GARANTE PER LA PROTEZIONE DEI DATI PERSONALI

**Parere alla Banca d'Italia sullo schema di regolamento concernente il trattamento dei dati personali effettuato nell'ambito della gestione degli esposti - 24 febbraio 2022 [9751895]**

**VEDI ANCHE** Newsletter del 14 marzo 2022

[doc. web n. 9751895]

**Parere alla Banca d'Italia sullo schema di regolamento concernente il trattamento dei dati personali effettuato nell'ambito della gestione degli esposti - 24 febbraio 2022**

# ALCUNE INIZIATIVE IN CORSO

| | |
|---|---|
| Migrazione delle basi dati CERVED/CEBIL e CSDB - (studio di progetto in corso) | Archiviazione delle basi dati CEBIL e CSDB su piattaforma *Big Data* |
| Creazione di un archivio INPS - (realizzazione in corso) | Archiviazione dell'archivio INPS su piattaforma *Big Data* |
| Caricamento dati PATSTAT su piattaforma *Big Data* - (realizzazione in corso) | Archiviazione su piattaforma *Big Data* (*Data Science Lab*) dei dati sui brevetti europei del database PATSTAT |
| Realizzazione piattaforma tecnologica per *Artificial Intelligence* e *Machine Learning* - (realizzazione in corso) | Introduzione di *tools* per *data science* |
| Introduzione strumenti di *graph analysis* in RADAR - (realizzazione in corso) | Soluzioni innovative di *graph analysis* e di analisi visuale |
| Progetto di sfruttamento del patrimonio informativo della Tesoreria - (realizzazione in corso) | Utilizzo del *Data Lake* come piattaforma di archiviazione dei dati e di sfruttamento con tecniche avanzate di data science |

# DIPARTIMENTO VIGILANZA BANCARIA E FINANZIARIA

## SERVIZIO SUPERVISIONE BANCARIA 1



## SERVIZIO ISPETTORATO VIGILANZA



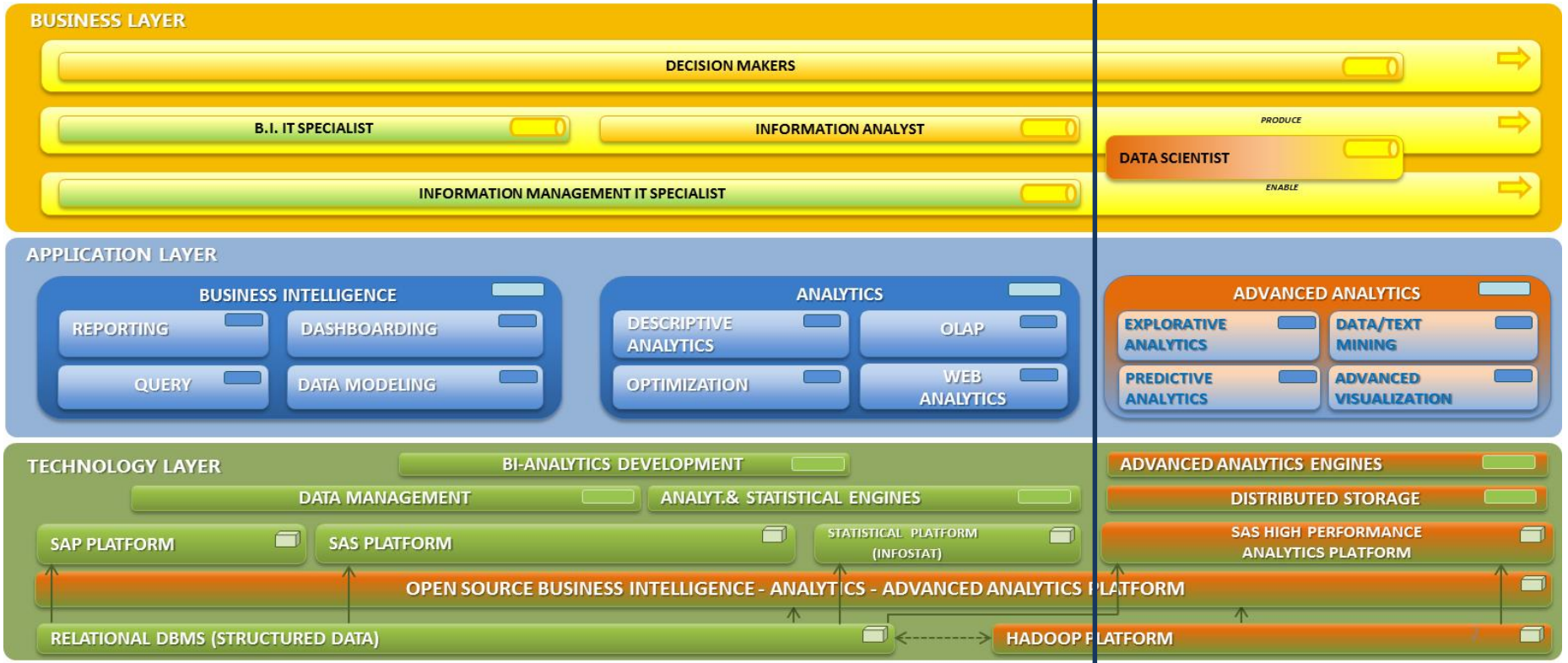
## SERVIZIO SUPERVISIONE BANCARIA 2

Sperimentazioni di **tecnologie di frontiera:**

- Sperimentazione di algoritmi e modelli di *machine learning*
- Studio e applicazione delle tecnologie di *automated reasoning* e dei *knowledge graph*
- Sostegno allo sviluppo di servizi in ambito finanziario ad alto contenuto tecnologico

- Realizzazione di un prototipo per l'euro digitale, nell'ambito della fase di *investigation* sulla CBDC europea avviata e coordinata dalla BCE
- Sperimentazione di soluzioni DLT per offrire servizi condivisi per la produzione di dati statistici a livello europeo (iniziativa IReF)
- Sperimentazione di soluzioni per l'interoperabilità di TIPS con altre infrastrutture di mercato (TIPS-HashLink, Buna, progetto Nexus)

# Enterprise Architecture View



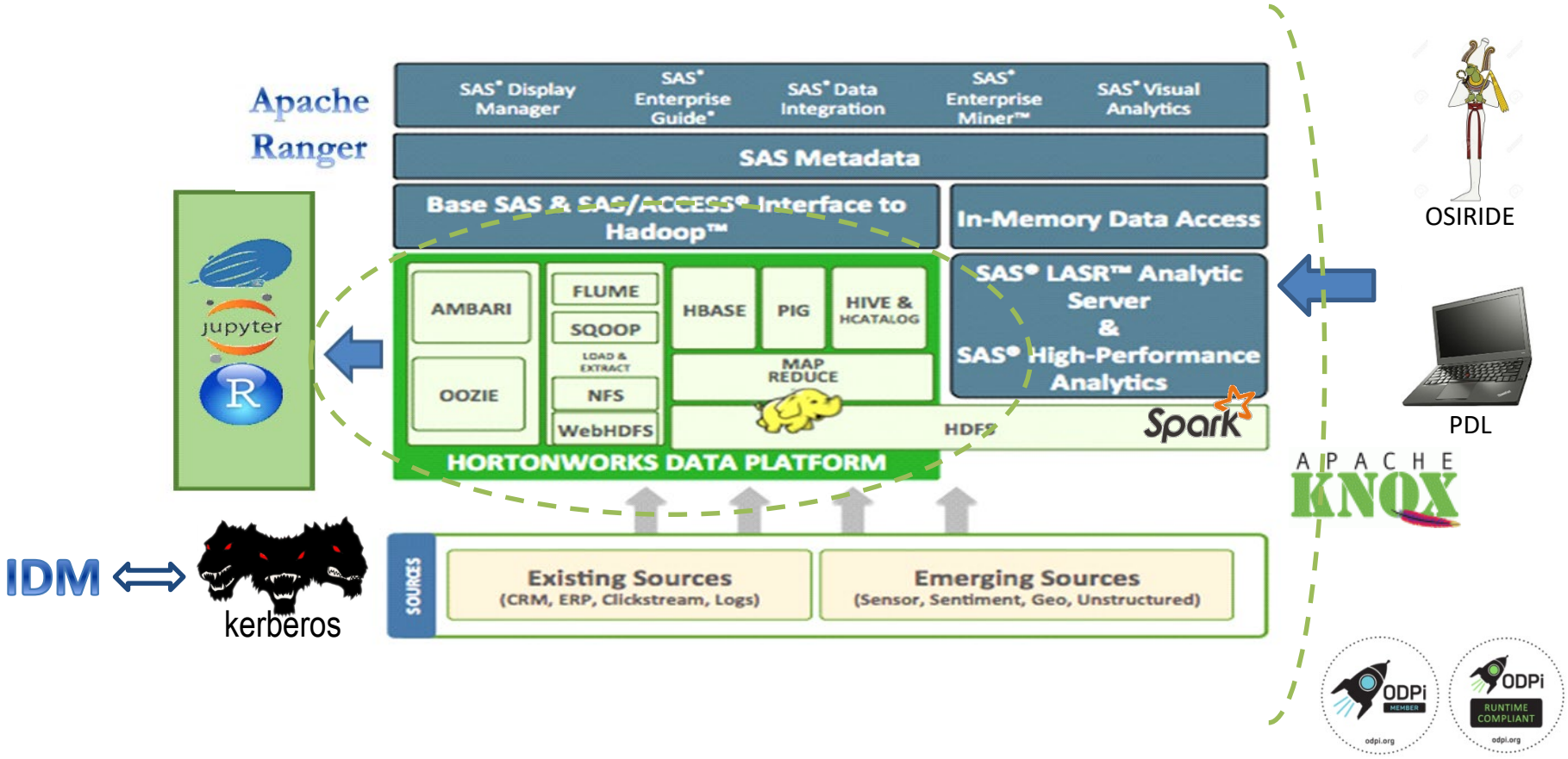
# Big Data Platforms

*"Big data platform generally consists of big data storage, servers, database, big data management, business intelligence and other big data management utilities. It also* ***supports custom development, querying*** *and integration with other systems. The primary benefit behind a big data platform is to reduce the complexity of multiple vendors/solutions into a one* ***cohesive solution****. Big data platform are also delivered through cloud where the provider provides an all inclusive big data solutions and services."* [source: Techopedia]


## Benefici delle piattaforme BD

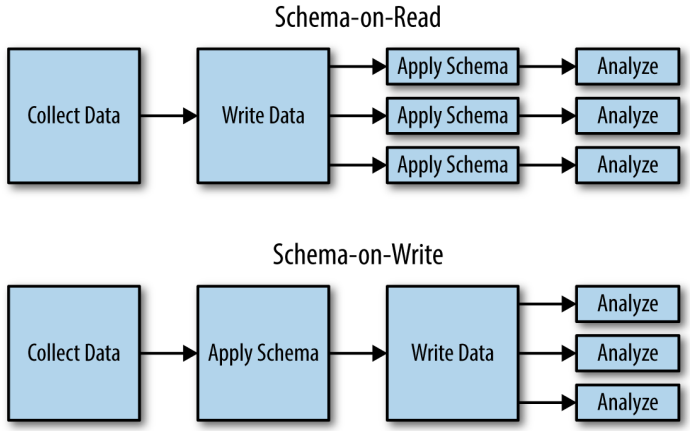
- Gestire volumi elevati di dati a costi ridotti e con alte prestazioni (date lake, data staging, archiving, streaming, EDW extension, etc. )
- Superare le limitazioni prestazionali dei db relazionali
- Abilitare analisi e applicazioni di nuova tipologia (p.e. text analytics, machine learning)

# PIATTAFORMA INTEGRATA + Sicurezza

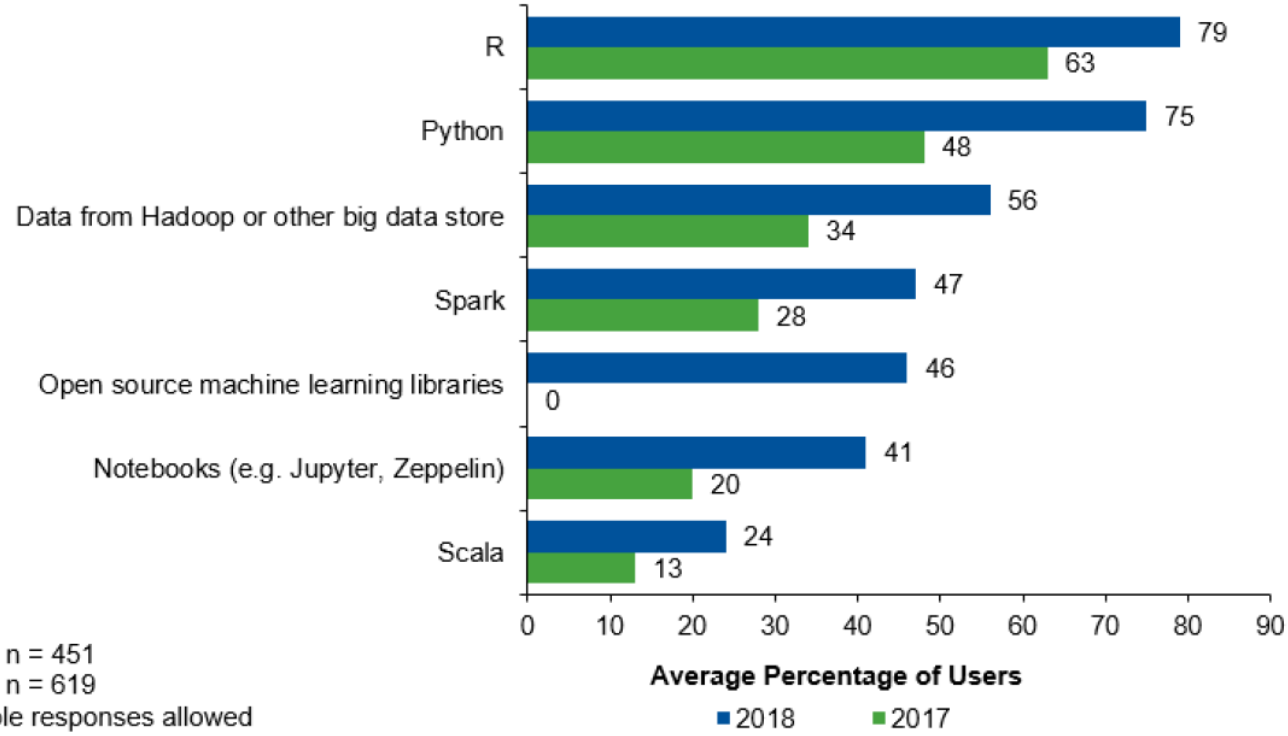


# Schema on-read

| Schema On-Write | Schema On-Read |
|---|---|
| Structured Data | Un-Structured Data |
| OLTP & OLAP systems | Big Data & NoSQL systems |
| Heavy ETL tool's role in staging and moving data | Very low ETL tool involvement |
| Modification of data model is costly | Schema is just a structured file, can be switched dynamically |
| Works well in small data sets. | Ideal for large volumes of data |
| User knows exactly what he/she is looking for | User is exploring the data without predefined questions |



# Open Source Is Here to Stay — Technologies Adopted by Data Science Teams




2018: n = 451
2017: n = 619
Multiple responses allowed

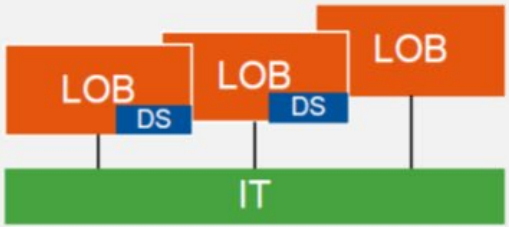



Gartner

# Data Science Deconstructed

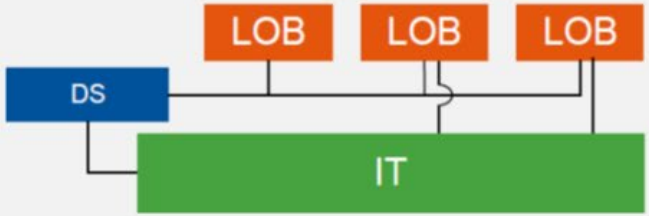
# Where to Put the Data Scientists?

Data Scientists @ LOBs

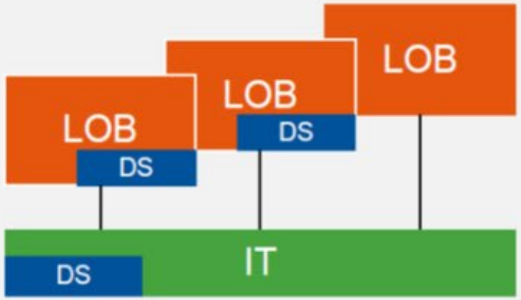



Data Scientists @ IT


Data Scientists as Separate BU

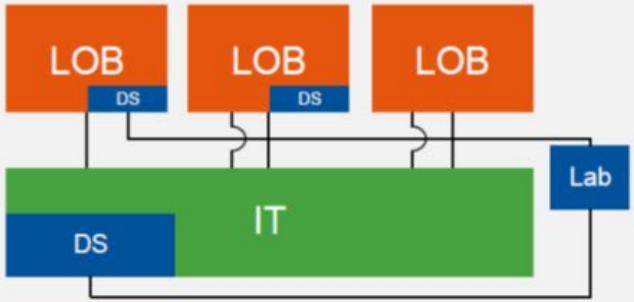



Scattered Experts




- Business Intimacy
- Knowledge Sharing
- Agility
- Cross-Functional View
- Proximity to Process and Data

Data Science Lab/CoE




Source: "Organizational Principles for Placing Data Science and Machine Learning Teams," (G00325989)
DS = Data Science; LOB = Line of Business




Gartner

# Taming the «Wild West» of Big Data

# Issues

- **Data scientists waste time on DevOps work.** Data scientists are precious, highly paid people, yet they often must spend 25% of their time dealing with DevOps tasks like installing packages and moving files between machines.
- **Data scientists waste time duplicating effort and reinventing the wheel**. Beyond individual data scientists wasting time on DevOps, entire teams can waste time pursuing projects that reinvent the wheel or don't build upon past organizational knowledge, because that past work was siloed and undiscoverable.
- **Important business processes become dependent on unreliable infrastructure.** Data scientists will often set up scheduled jobs to run on their own local machines, or operate shared servers as "lab" or "dev" machines.
- **Compute costs can become excessive and uncontrolled.** Unlike BI, data science involves computationally intensive techniques, which demand high-powered machines and specialized resources like GPUs. Especially in a cloud environment, data scientists in the wild west can unintentionally burn thousands of dollars a month by leaving expensive machines running unnecessarily.
- **High-value intellectual property is improperly secured.** Predictive models and analyses can encapsulate insights key to competitive advantage, and that work is often scattered throughout network drives, wikis, or Sharepoint sites.

Source: *DominoDataLab.com*

# Solution

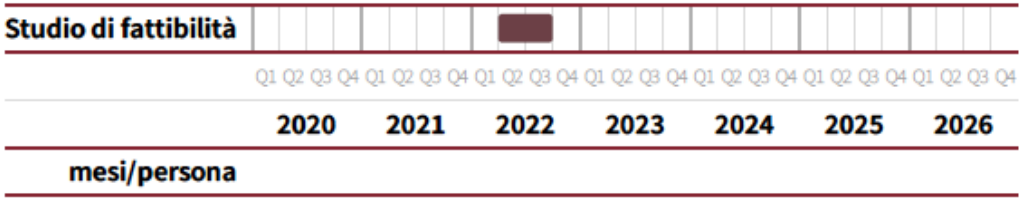
- **Self-service infrastructure**, so data scientists can do exploratory data analysis and model development without configuring and using their own compute resources. The data science platform encompasses compute resources—as well as the languages, packages and tools necessary for modern data science work—with controls and reporting around resource usage to administer or attribute costs.
- **Ways to deploy, productionize or operationalize finished models**, instead of driving data scientists to set up shadow systems. This includes deploying models to power scheduled jobs, reports, APIs or dashboards in one place. The data science platform also provides a consistent baseline of non-functional requirements (security, HA, etc.) and a catalog that offers transparency into assets and utilization across the enterprise.
- **Governance, collaboration and knowledge management** around all the artifacts created in the process of the research and deployment work described above.

# REALIZZAZIONE DI UN ENTERPRISE DATA LAKE ORIENTATO AI DATI STATISTICI

Progetto 908-P0875 Classe 2 - Priorità 688

Dipartimento Economia e statistica
SERVIZIO RILEVAZIONI ED ELABORAZIONI STATISTICHE

Realizzazione di un "Enterprise Data Lake" orientato ai dati statistici che consenta l'archiviazione di basi dati di grandi dimensioni anche con caratteristiche eterogenee (es. dati strutturati, testi, social data) e che faciliti le attività di integrazione e sfruttamento da parte degli utenti tramite metodi e strumenti tipici della data science.




L'iniziativa è stata **inserita** per la prima volta in Portafoglio.

**Studio**: dal 1 aprile 2022 al 30 settembre 2022

→ Studio di fattibilità Data lake
→ GdL principi di Data Governance

# TAKEAWAYS

- Differenza tra applicazioni transazionali e informative
- La funzione informatica presidia non solo le basi dati operative e le relative applicazioni, ma anche gli sfruttamenti diretti da parte dell'utente, attraverso la creazione di DW e reportistica predefinita e la disponibilità agli utenti di ambienti evoluti di sfruttamento.
- Il perimetro dei dati oggetto di analisi e la tipologia di analisi si ampliano costantemente e rapidamente → Big Data, AI/ML, ecc.
- Per evitare il «wild west» dei Big Data, sono necessari nuovi strumenti tecnologici e nuovi presidi organizzativi, di sicurezza e di governance. → nuove iniziative ML framework e Data Lake