Data Analysis & Visualization

Introduzione ai Big Data

Ing. Giulio Destri

Dr. Ing. Giulio Destri, Ph.D.

Professore a contratto di Sistemi Informativi @Università di Parma dal 2003

Digital Transformation Advisor, Business Coach, Trainer, Innovation Manager @LINDA

Esaminatore ISO27021 e UNI11506-11621 BA (EPBA) @Intertek

Membro commissione UNI/CT 526 @UNINFO e coordinatore commissione ICT Ordine Ingegneri di Cremona

Blogger @6MEMES di MAPS

Certificazioni: ISO27001LA , ISO27021, ITILv3, COBIT-2019, SCRUM Master, EPBA, NLP Coach, NLP AMP

https://www.linkedin.com/in/giuliodestri

https://www.lindaconsulting.it/

giulio.destri@unipr.it

twitter.com/GiulioDestri

Argomenti

- Dalla Business Intelligence ai Big Data
- Volume: l'irresistibile crescita dei dati
- Varietà: forme differenti per i dati
- Veridicità: la qualità dei dati
- Velocità dei dati
- Validità: i dati e il contesto
- Volatilità: quanto "durano" i dati?
- Visualizzazione: come rappresentare i dati?
- Valore: che cosa ottenere dai dati?
- Fare analisi dei dati
- Strumenti per Big Data
- DBMS NoSQL
- Big Data Analytics

Business Intelligence e Big Data

Dai dati alla conoscenza: la piramide DIKW

Analisi dei dati: business intelligence

- Insieme di applicazioni e tecnologie per l'analisi dei dati e l'estrazione di informazioni da essi
- Comprende:
  - DSS (Decision Support System)
  - Query e Report
  - OLAP (Online Analytical Processing)
  - Analisi statistiche
  - Modelli previsionali
  - Data Mining
  - Data Warehouse e Data Mart

Ma che succede quando…?

1. Le moli di dati da esaminare crescono a dismisura?
2. I dati sono di tipo molto diverso fra loro (es. immagini, registrazioni audio, testo, numeri…)
3. I tempi di elaborazione sono limitati rispetto alla mole di dati da elaborare

Big data: definizione

1. Big data: sono dati che superano i limiti degli strumenti DBMS tradizionali
2. Big data: sono anche le tecnologie finalizzate ad estrarre da essi conoscenze e valore.
3. Big data: l'analisi di quantità incredibilmente grandi di informazioni.

Tipologie di «Big data»

1. **Dati non strutturati**: tipicamente provenienti da Social Media, sono post/testi, tweet, immagini, file audio, video (con loro metadati…)
2. **Dati semi-strutturati**: csv, json, tracciati dati vari
3. **Dati strutturati**: conformi e/o provenienti da DB relazionali

Big data: il modello delle 5 V

Dati «molto grandi e/o complessi», quindi

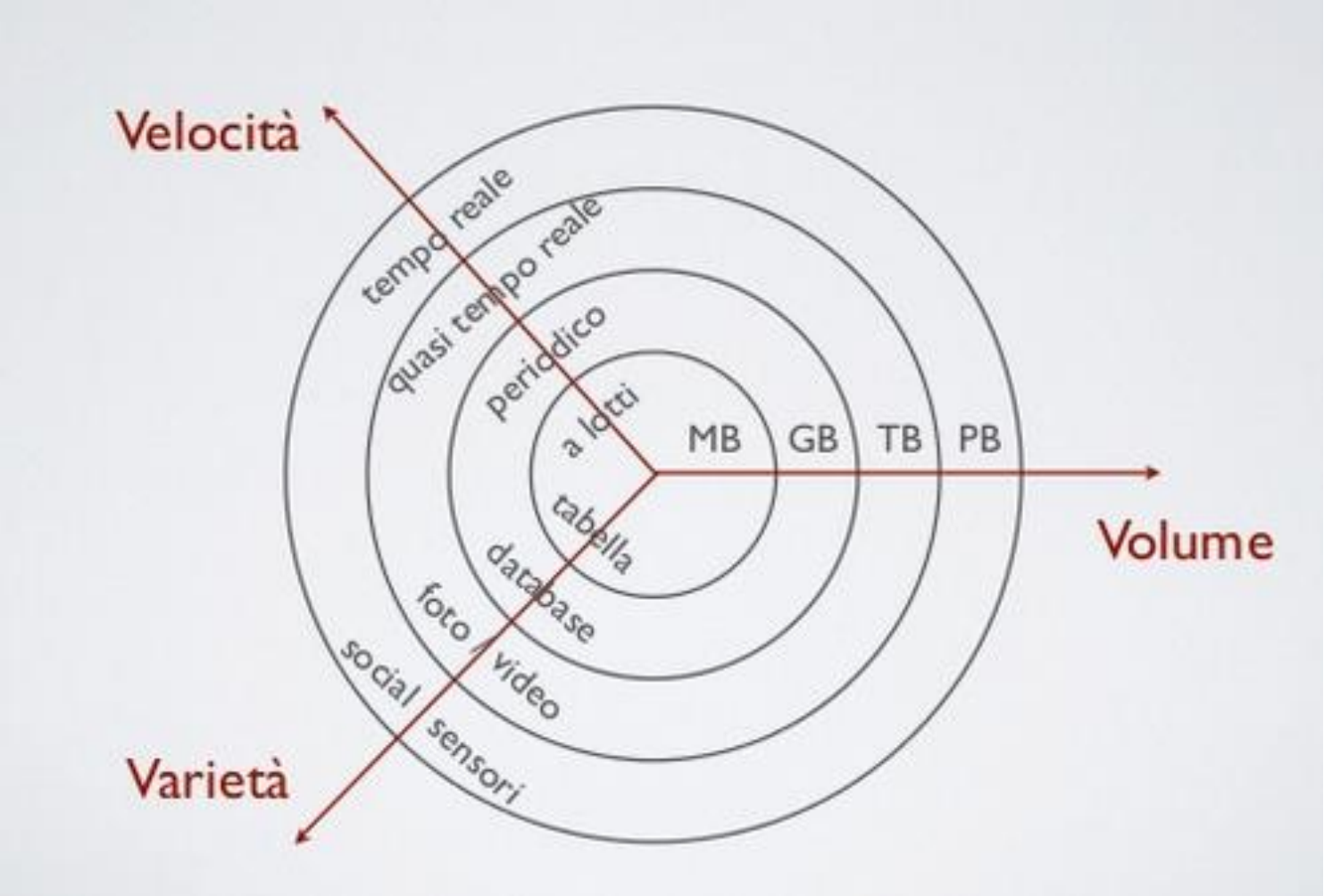
1. **Volume**: quantità di dati generati per unità di tempo
2. **Varietà**: differenti tipologie di dati generati, memorizzati, utilizzati
3. **Velocità**: rapidità di elaborazione o trasmissione necessaria
4. **Veridicità**: qualità di ingresso a sistemi di analisi
5. **Valore**: capacità di ottenere valore

Effetti delle 5 V

- Molti sistemi DBMS «tradizionali» non sono in grado di trattare volumi così grandi
- I dati possono essere molto variegati (es. foto crocchie e dati fabbricazione…)
- I dati possono non essere strutturati adeguatamente e costruire un ETL per adattarli potrebbe essere estremamente costoso…

- Nuovo approccio alla elaborazione

Estremizzazione delle prime 3 V



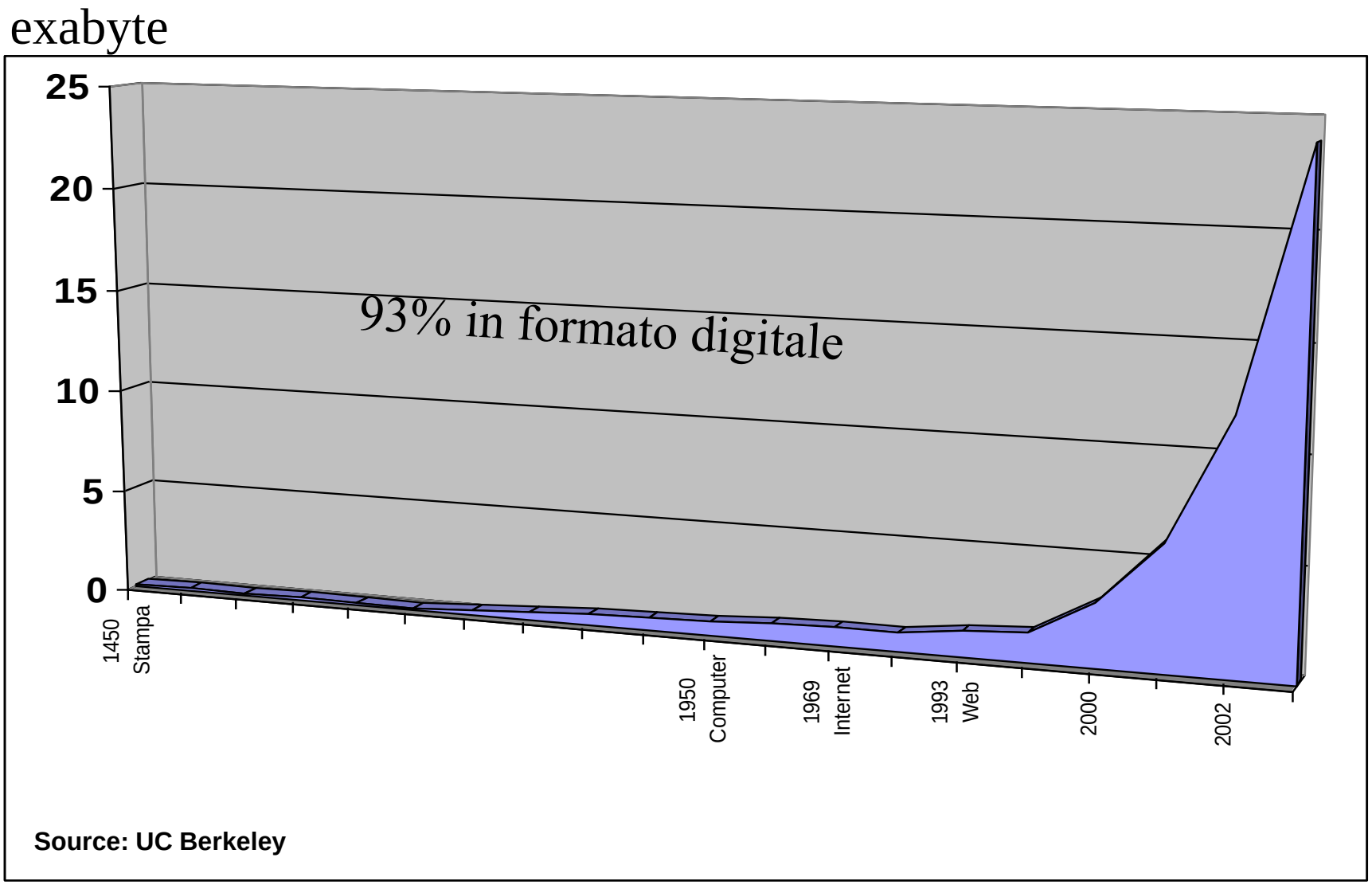
Big data: il modello delle 8 V

1. Volume
2. Varietà (Variety)
3. Velocità (Velocity)
4. Veridicità/Affidabilità (Veracity)
5. Validità (per il contesto)(Validity)
6. Volatilità/durevolezza (Volatility)
7. Visualizzazione (Visualization)
8. Valore (Value)

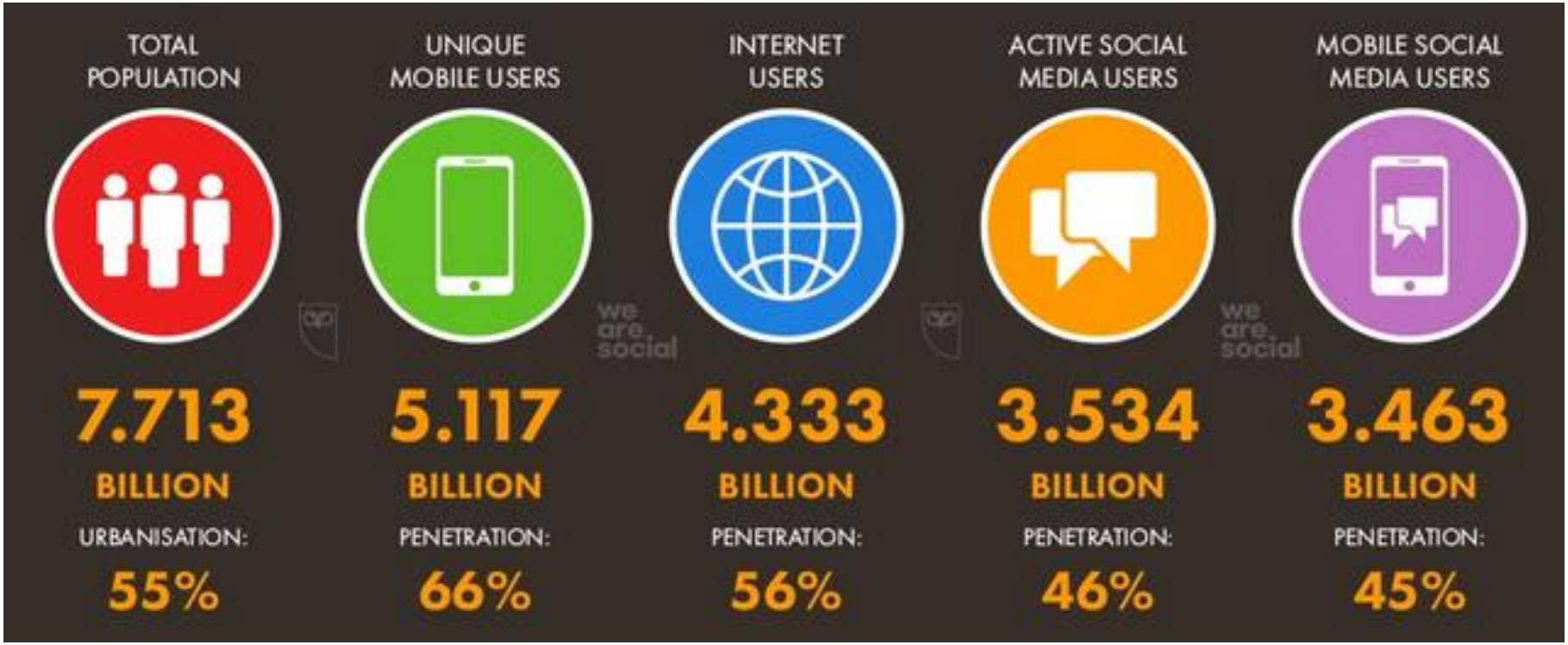
Volume: L'irresistibile aumento dei dati

L'Internet-minute nel 2018

La crescita dei dati: nel 2006…

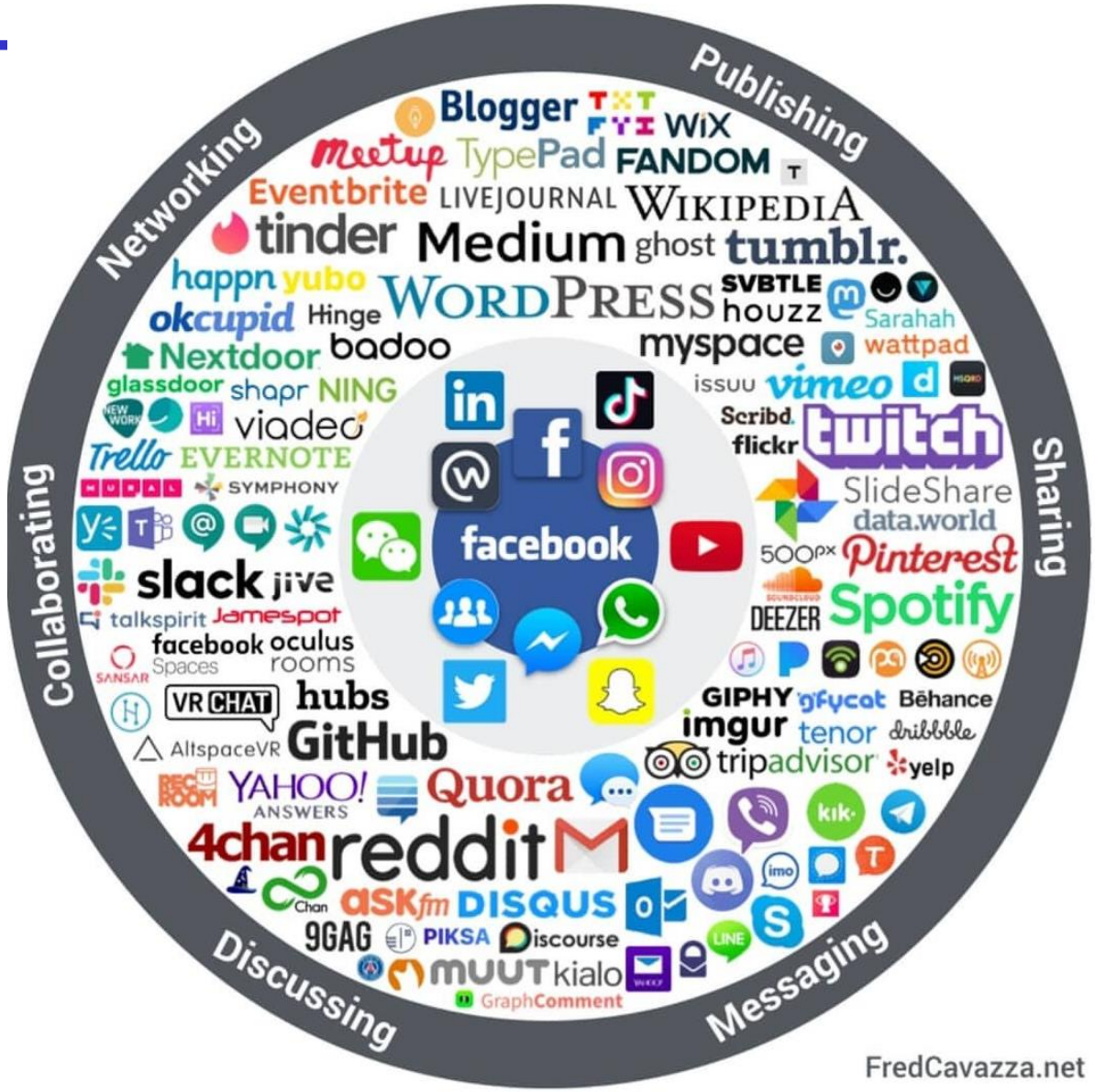


Il mondo nel luglio 2019



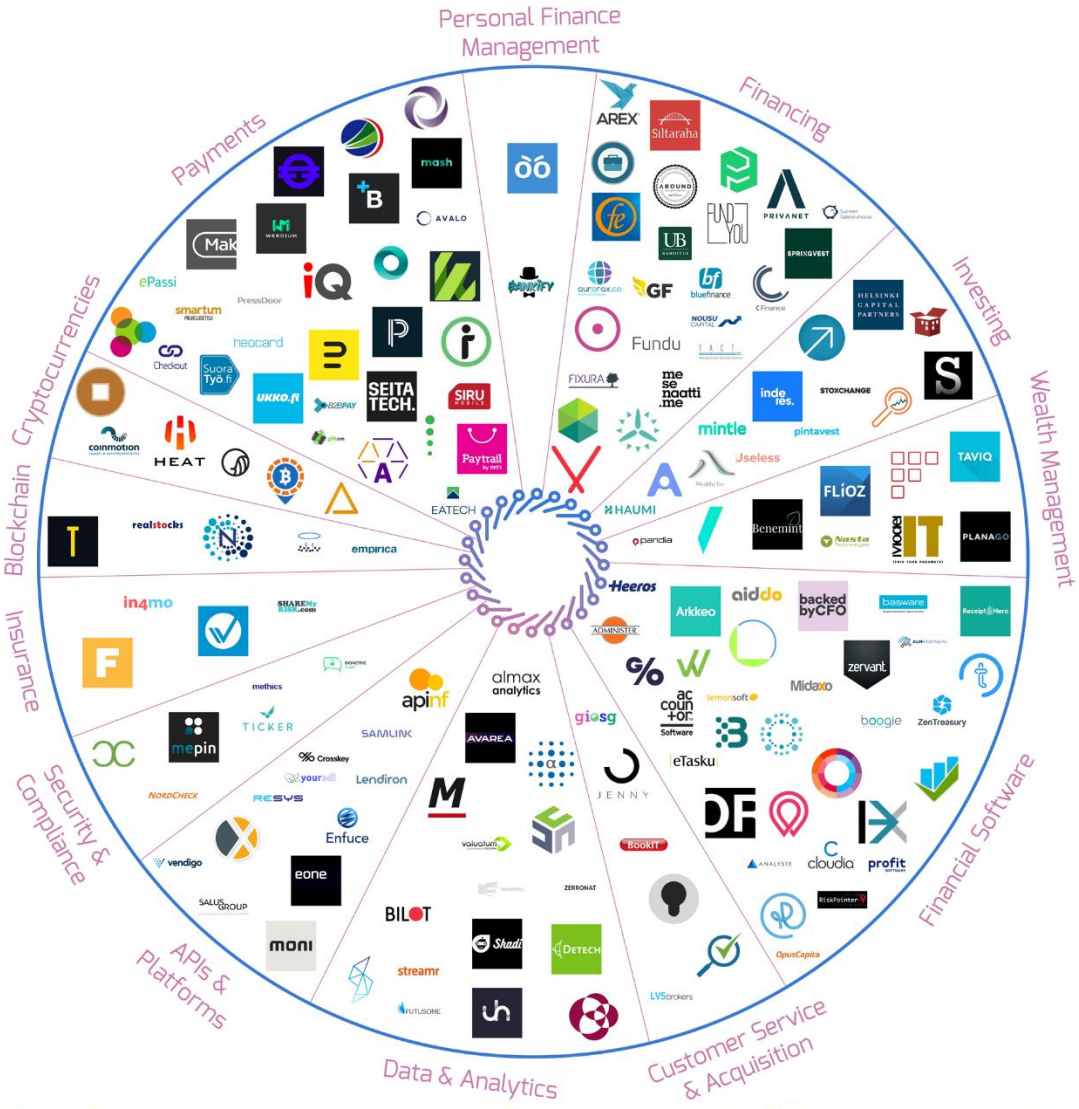
Fonte: Hootsuite – We are social

L’universo dei social media nel 2019



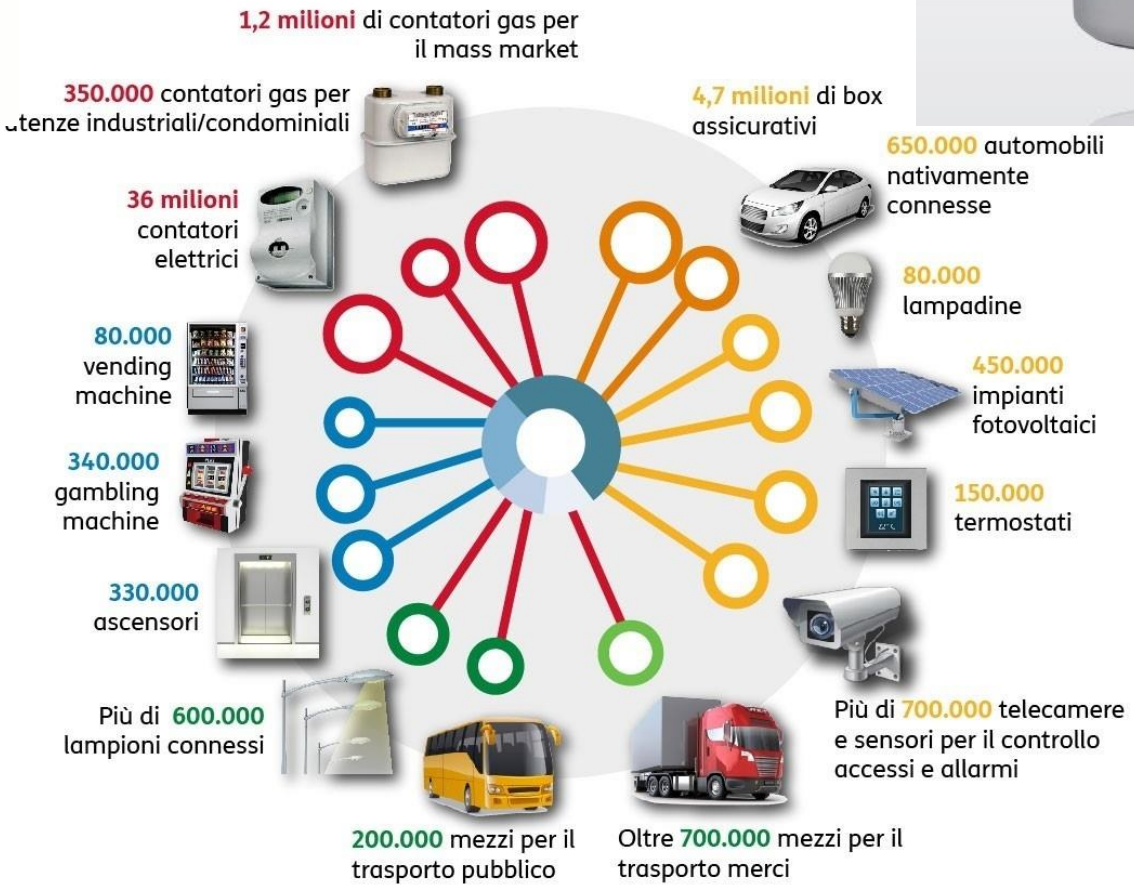
FredCavazza.net

L’universo dei servizi nel 2019



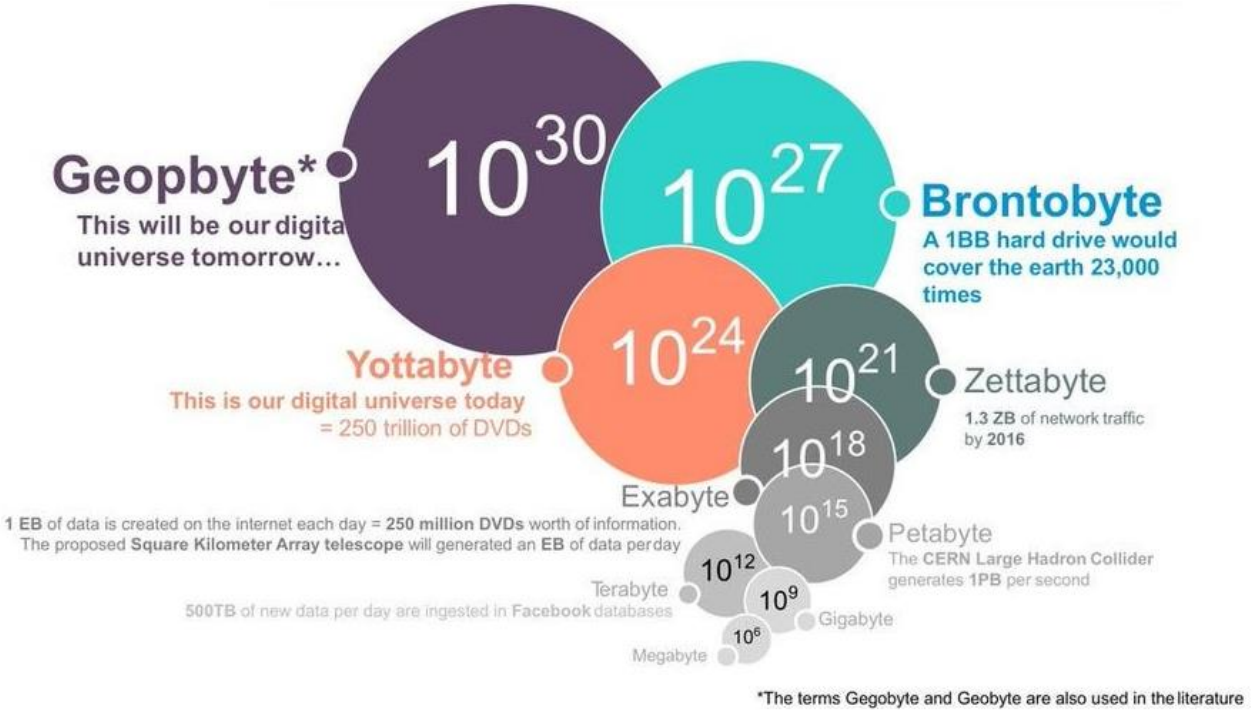
Smart Meter

Dati da TIM, 2017



I numeri dei dati nel 2019

- Megabyte (10^6)
- Gigabyte (10^9)
- Terabyte (10^12)
- Petabyte (10^15)
- Exabyte (10^18)
- Zettabyte (10^21)
- Yottabyte (10^24)
- Brontobyte (10^27)
- Geopbyte (10^30)



Fonte: Simon Kuestenmacher

Varietà:
Forme differenti per i dati

Varietà

- Tipologie
  - Testi, in tantissime forme
  - Numeri
  - Immagini
  - Filmati
  - Audio
  - Dati strutturati

Varietà

- Significati
  - Dati da sistemi gestionali / ERP
  - Dati da e-commerce
  - Dati da IoT «civili»
  - Dati da Smart Car
  - Dati da IIoT
  - Dati da Social Media
  - Dati ambientali
  - Dati da pre-elaborazioni

Veridicità: La qualità dei dati

Veridicità (Veracity)

- Origine: da dove vengono i dati?
- Autenticità: sono quelli inviati?
- Attendibilità: sono attendibili?
- Completezza: sono completi?
- Integrità: sono integri?

Velocità dei dati

Velocità

- Velocità di generazione dei dati
- Velocità di analisi / elaborazione dei dati
- Frequenza di generazione
- Frequenza di analisi / elaborazione

Validità: I dati ed il contesto

Validità dei dati per un contesto

- Che obiettivo ho per l'analisi?
- Quali dati mi sono utili?
- Come li raccolgo / da dove li raccolgo?
- I dati di cui dispongo sono validi per il mio contesto e scopo?

Volatilità: Quanto "durano" i dati?

Volatilità dei dati

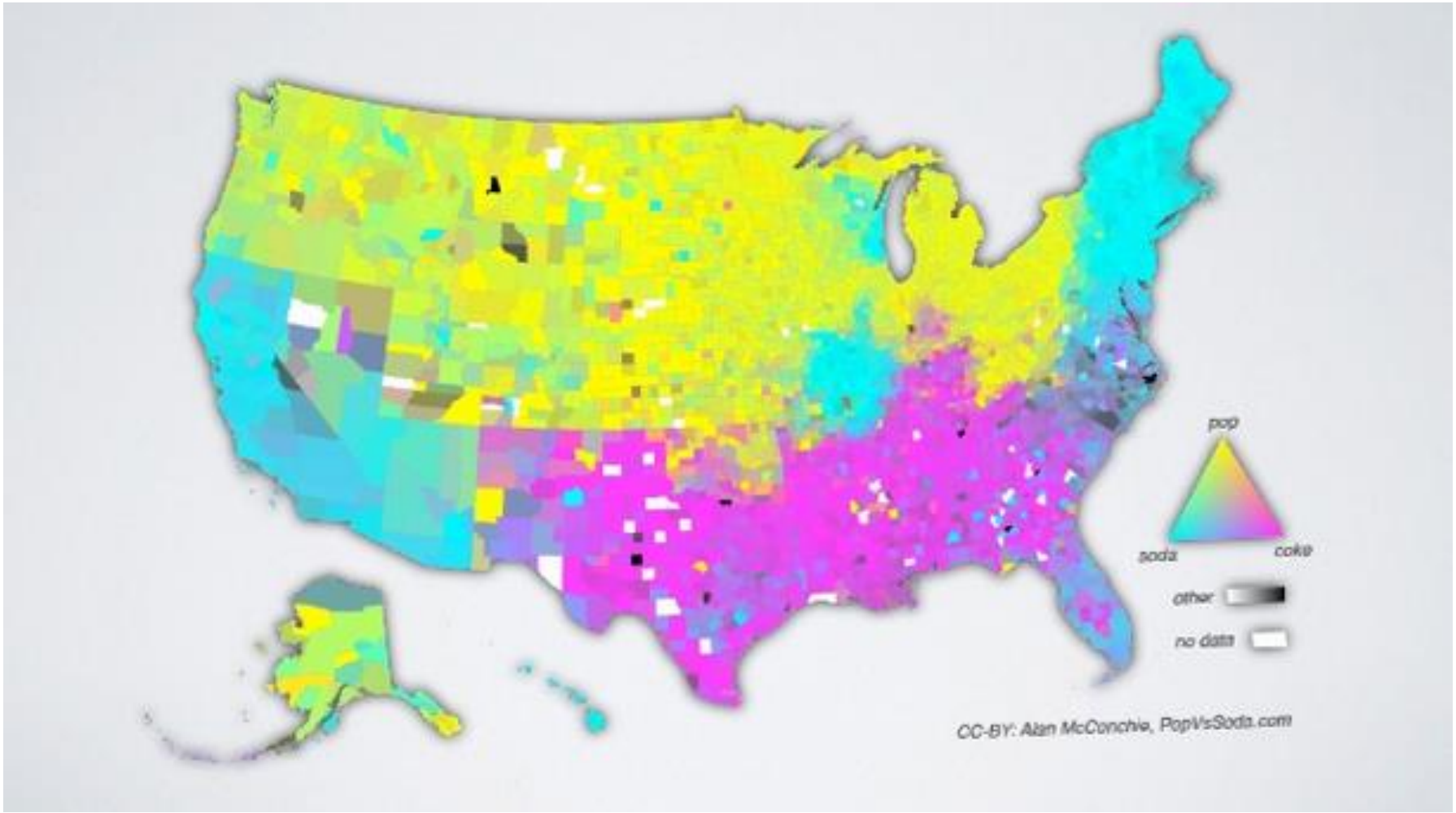
- I dati hanno una «scadenza»?
- Devono essere elaborati entro un limite temporale?
- Devono essere conservati dopo elaborazioni riduttive?

Visualizzazione: Come rappresentare i dati?

Visualizzazione

- I dati vengono elaborati per trarne informazione e conoscenza
- Come deve essere rappresentata questa nuova risorsa?

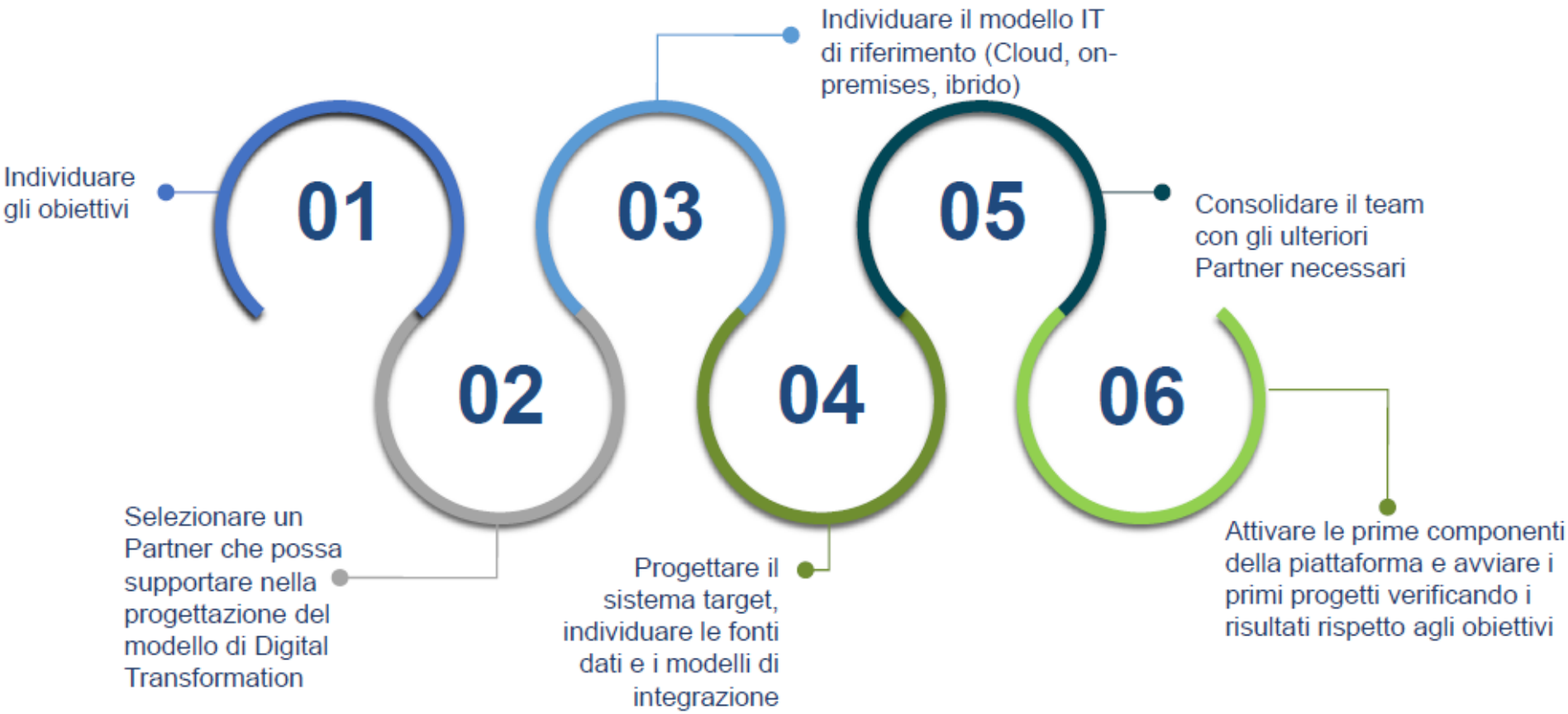
Esempio di Visualizzazione



Valore: che cosa ottenere dai dati?

Cosa conosciamo?

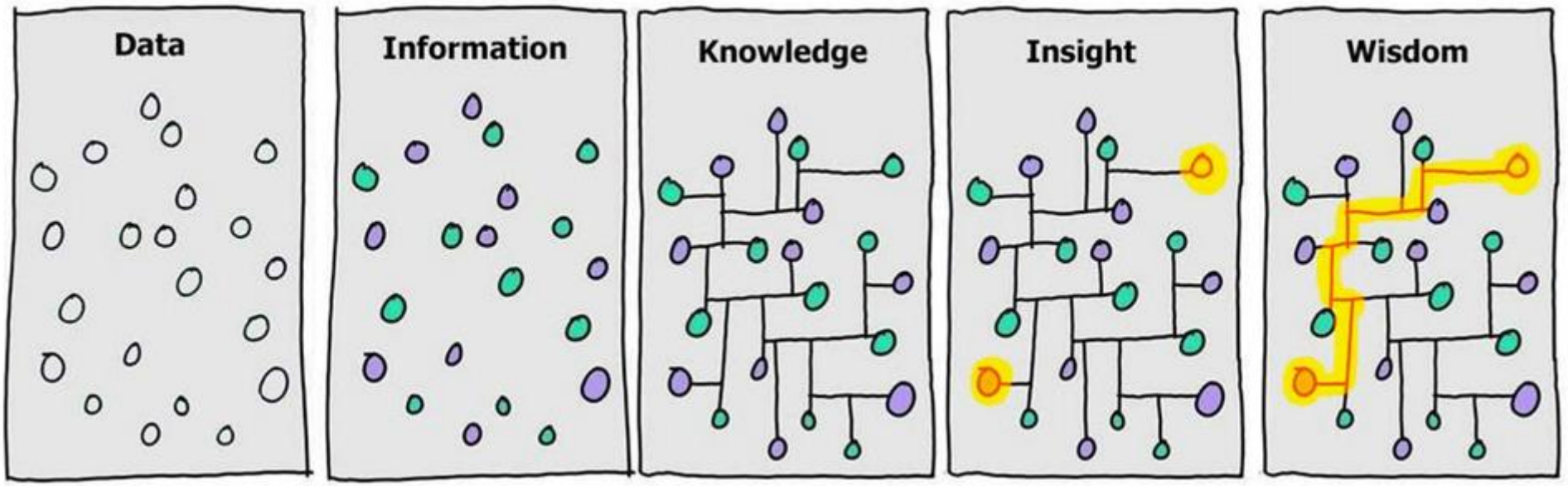
Un progetto coi Big Data…



Per ottenere un valore…

- Cosa vogliamo ottenere: avere chiaro l'obiettivo
- Gli strumenti possono essere i più diversi…

L'evoluzione del ciclo DIKV: information continuum



- Dati
- Informazione
- Conoscenza
- Intuizione
- Consapevolezza
- Saggezza

Fare analisi dei dati

Big Data Analytics

- **Descrittivi**: spiegano eventi avvenuti nel passato;
- **Diagnostici**: spiegano il perchè un evento si è verificato;
- **Predittivi**: quello di maggiore valore per le aziende, analizza i dati per prevedere quello che potrebbe succedere in futuro;
- **Prescrittivi**: analizza i dati per prendere una decisione di business, ad esempio dove inserire un annuncio pubblicitario per avere un bacino di ascolto più ampio, quale strada prendere per evitare il traffico.

Da campioni a tutti

- Le tecniche di analisi statistiche usate sempre sino al 2012 sono basate sui campioni
- Campionamento uniforme, evitare errori di polarizzazione campioni ecc…
- Ma quando il campione è il 50%, il 70% o il 100% dei dati, come si deve agire?

- Ecco l'azione del Data Scientist

Da campioni a tutti

- E’ sempre conveniente?
- Dipende dal contesto
- In taluni casi il volume (e la conseguente necessità di storage ed altre componenti) rende non conveniente questo approccio

Strumenti per Big Data

Necessità per Big Data

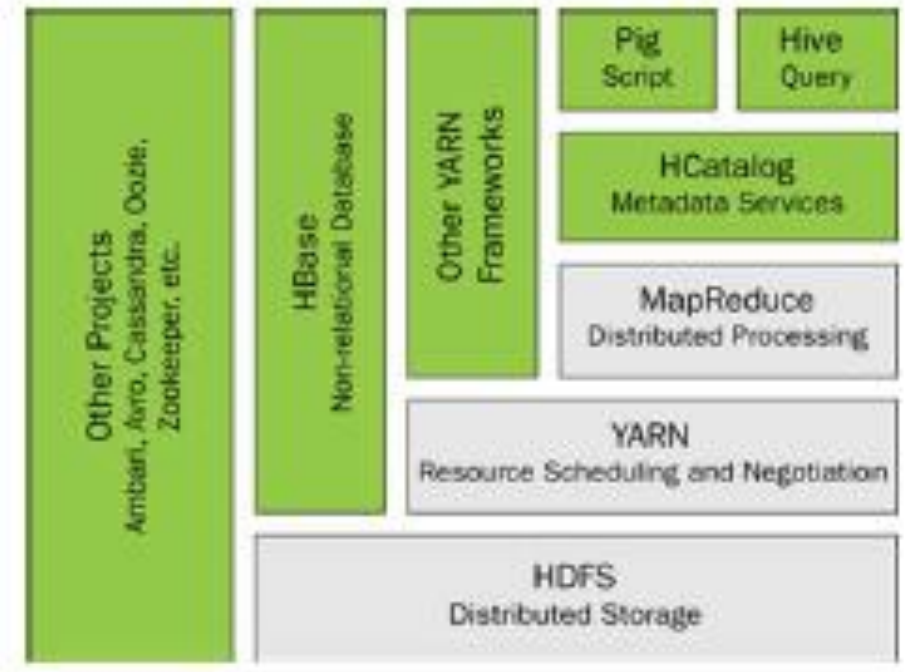
- Gestione Storage (con volumi…)
- Gestione strutture dati
- Gestione di elaborazioni distribuite (infrastruttura)
- Estrazione di caratteristiche (aggregazioni…)
- Sistemi automatici
- Sistemi interattivi (linguaggi di interrogazione)
- Presentazione di risultati…

Strumenti per Big Data

- Apache Hadoop è una delle prime piattaforme open-source nate per gestire l'archiviazione e l'analisi di grandi quantità di dati
- permette di lavorare con dataset dell'ordine dei Petabyte all'interno di un ambiente distribuito di cluster di macchine «comuni»

Hadoop Ecosystem

- Insieme di strumenti da affiancare al «DBMS» Hadoop



Hadoop: HDFS

- L'Hadoop Distributed File System (in sigla HDFS) è un file system distribuito, portabile e scalabile scritto in Java per il framework Hadoop.
- Un cluster in Hadoop tipicamente possiede uno o più name node (su cui risiedono i metadati dei file) e un insieme di data node (su cui risiedono, in blocchi di dimensione fissa, i file dell'HDFS).

Hadoop: HDFS

- I formati più usati per i file su HDFS sono Comma-separated values, Apache Avro, Apache ORC e Apache Parquet.
- Hadoop supporta anche:
- Amazon S3 file system;
- Azure data lake store;

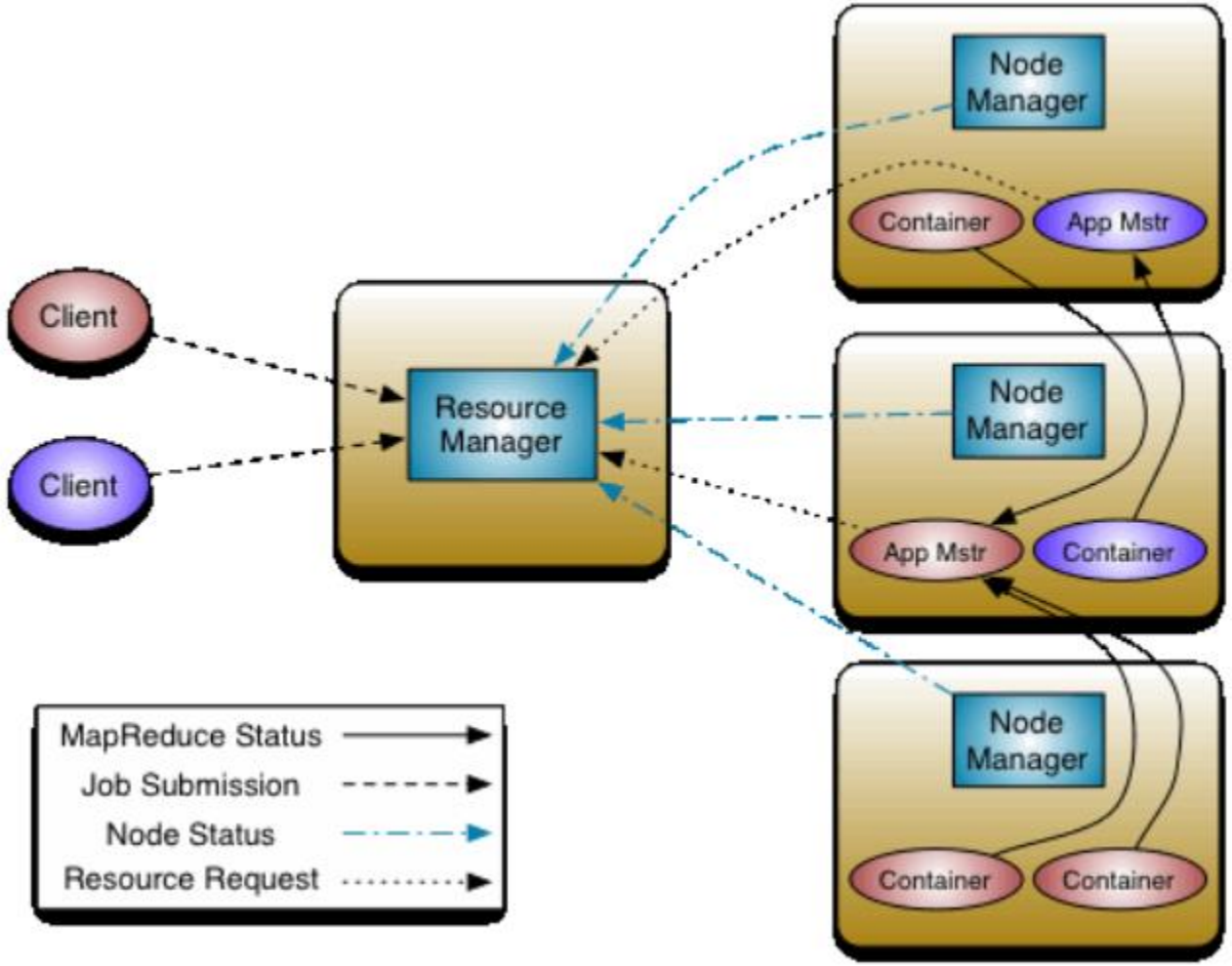
Hadoop: HDFS

- Hadoop può lavorare direttamente con qualsiasi file system distribuito che possa essere montato da un sistema operativo sottostante semplicemente usando un URL del tipo 'file://'.

Hadoop: YARN

- Yet Another Resource Negotiator
- E’ uno schedulatore distribuito di elaborazioni
- Si compone di un Resource Manager principale e di più Application Manager secondari

Hadoop: YARN

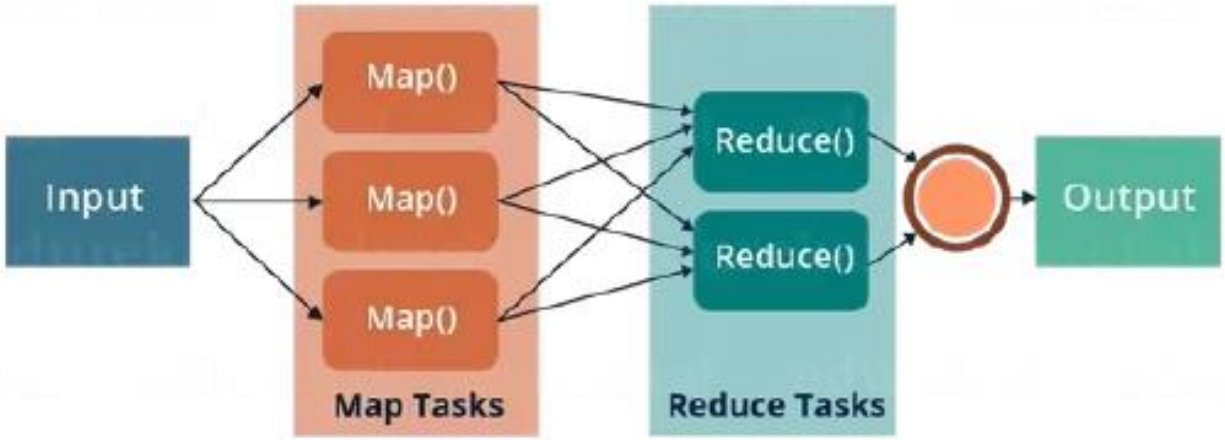


Hadoop: MapReduce

- E’ il cuore di Hadoop
- Opera basandosi su HDFS e YARN
- Elabora i dati in parallelo fra i nodi entro il cluster su cui Hadoop opera
- I dati vengono suddivisi fra i nodi che li processeranno in modo autonomo uno dall’altro
- I dati vengono elaborati sul nodo ove risiedono (in HDFS)

Hadoop: MapReduce

- Algoritmo basato su 2 task
  - MAP: trasformazione degli input in coppie chiave-valore
  - REDUCE: generazione del task di elaborazione su un insieme parziale di dati



MapReduce vs RDBMS

	RDBMS	MapReduce
Dati	GB	PB
Tipologie di dati	Strutturati	Semi-strutturati / Non Strutturati
Accesso	Interattivo e Batch	Batch
Modifiche	Molteplici read & write	unica write (append) e molteplici read
Transazioni	ACID	—
Struttura	Schema-on-write	Schema-on-read
Integrità	Alta	Bassa
Scalabilità	Non lineare	Lineare

Hadoop: MapReduce

- E’ il cuore di Hadoop
- Opera basandosi su HDFS e YARN
- Elabora i dati in parallelo fra i nodi entro il cluster su cui Hadoop opera
- I dati vengono suddivisi fra i nodi che li processeranno in modo autonomo uno dall’altro
- I dati vengono elaborati sul nodo ove risiedono (in HDFS)

Apache PIG

- E’ una piattaforma per l'analisi di grandi set di dati (accoppiata a Hadoop)
- Consiste in un linguaggio di alto livello per esprimere i programmi di analisi dei dati, insieme all'infrastruttura per la valutazione di questi programmi.
- La proprietà saliente dei programmi Pig è che la loro struttura è suscettibile di sostanziale parallelizzazione, che a sua volta consente loro di gestire set di dati molto grandi.

Apache PIG

- Al momento, il livello di infrastruttura di Pig è costituito da un compilatore che produce sequenze di programmi Map-Reduce
- Il livello linguistico di Pig è attualmente costituito da un linguaggio testuale chiamato Pig Latin, che ha le seguenti proprietà chiave:
  - Facilità di programmazione
  - Opportunità di ottimizzazione
  - Estensibilità

Apache PIG

Facilità di programmazione

- È banale ottenere l'esecuzione parallela di compiti di analisi dei dati semplici, "imbarazzanti parallelamente".
- Compiti complessi composti da più trasformazioni di dati correlate sono esplicitamente codificati come sequenze di flussi di dati, rendendoli facili da scrivere, comprendere e mantenere.

Apache PIG

Opportunità di ottimizzazione

Il modo in cui le attività sono codificate consente al sistema di ottimizzare automaticamente l'esecuzione, consentendo all'utente di concentrarsi sulla semantica piuttosto che sull'efficienza.

Estensibilità.

Gli utenti possono creare le proprie funzioni per eseguire elaborazioni speciali.

Apache HIVE

- E’ un software di data warehouse
- Semplifica la lettura, la scrittura e la gestione di grandi set di dati che risiedono nella memoria distribuita tramite SQL.
- La struttura può essere proiettata su dati già archiviati.
- Uno strumento da riga di comando e un driver JDBC vengono forniti per connettere gli utenti a Hive.

DBMS NoSQL

L’approccio Not Only SQL

- Usare anche DBMS non relazionali
- Rinunciare, quando necessario, a usare tutte le caratteristiche degli RDBMS (ACID in primis)

I DBMS NoSQL

- Database orientati al documento
- Basi di dati a grafo
- Chiave/valore archiviato su disco
- Chiave valore con cache in RAM
- Altri chiave/valore
- Basi di dati a oggetti

MongoDB

- DBMS non relazionale, orientato ai documenti
- Struttura basata su documenti in stile JSON con schema dinamico
- MongoDB chiama il formato BSON
- Alte performance
- Già usato in applicazioni commerciali

Big Data Analytics

Strumenti per Big Data Analytics

Automated Analytics

Capacità di prendere decisioni in modo autonomo

la combinazione di AI, Deep Learning e interfacce utente (ad esempio, riconoscimento vocale) sta rendendo sempre più vicina l'automatizzazione di una serie di task lavorativi che fino a qualche tempo fa si pensava fosse impossibile realizzare con una "macchina".

Big Data => Cloud?

- La quantità di potenza elaborativa e di storage necessario potrebbe essere proibitiva rispetto ad una elaborazione in loco
- Soprattutto le elaborazioni possono non essere continue
- La potenza di calcolo on-demand => cloud
- Storage => Storage in cloud

Sommario

- Dalla Business Intelligence ai Big Data
- Volume: l'irresistibile crescita dei dati
- Varietà: forme differenti per i dati
- Veridicità: la qualità dei dati
- Velocità dei dati
- Validità: i dati e il contesto
- Volatilità: quanto "durano" i dati?
- Visualizzazione: come rappresentare i dati?
- Valore: che cosa ottenere dai dati?
- Fare analisi dei dati
- Strumenti per Big Data
- DBMS NoSQL
- Big Data Analytics