

DMIF, Università di Udine

---

# Tecnologie Digitali per il Cibo e la Ristorazione

*Introduzione ai Big Data*

Andrea Brunello

[andrea.brunello@uniud.it](mailto:andrea.brunello@uniud.it)

A.A. 2021–2022



# Outline

1 Big Data

2 Casi d'uso

3 Tecnologie

# Big Data

# Big Data

## Cosa sono i Big Data? - Le '3V'

I big data trascendono le limitazioni dei classici database, e sono caratterizzati da una o più delle seguenti proprietà:

- *Volume*
- *Varietà*
- *Velocità*

Nel 2001, Doug Laney, analista di Gartner, introduce le '3V' dei big data nella pubblicazione "*3D data management: Controlling data volume, variety and velocity*"

Superquark sui Big Data:

[https://www.youtube.com/watch?v=A2pUx5B\\_\\_C4A](https://www.youtube.com/watch?v=A2pUx5B__C4A)

# Volume

Con il termine *Volume* si intende la dimensione dei dati che vengono memorizzati, che possono derivare da azioni svolte da esseri umani o essere generati da delle macchine.

Talvolta, i dati possono essere così voluminosi da non poter essere memorizzati nella loro interezza; si rendono necessarie operazioni di compressione/trasformazione online, svolte a mano a mano che i dati vengono ricevuti (es., dati scientifici).

A volte, i dati possono essere memorizzati all'interno di tradizionali basi di dati relazionali, mentre altre volte ciò risulta essere troppo oneroso dal punto di vista computazionale  
~~ soluzioni NoSQL, Hadoop.

# Volume (Ordini di grandezza)

**IBM 350 disk storage  
(1956, 3.75 MB)**

**Walmart's DW  
(1992, 1 TB)**

**1 year of CERN's  
LHC data (15 PB)**

Quantities of bytes							
Common prefix				Binary prefix			
Name	Symbol	Decimal SI	Binary JEDEC	Name	Symbol	Binary IEC	
kilobyte	KB/kB	$10^3$	$2^{10}$	kibibyte	KiB	$2^{10}$	
megabyte	MB	$10^6$	$2^{20}$	mebibyte	MiB	$2^{20}$	
gigabyte	GB	$10^9$	$2^{30}$	gibibyte	GiB	$2^{30}$	
terabyte	TB	$10^{12}$	$2^{40}$	tebibyte	TiB	$2^{40}$	
petabyte	PB	$10^{15}$	$2^{50}$	pebibyte	PiB	$2^{50}$	
exabyte	EB	$10^{18}$	$2^{60}$	exbibyte	EiB	$2^{60}$	
zettabyte	ZB	$10^{21}$	$2^{70}$	zebibyte	ZiB	$2^{70}$	
yottabyte	YB	$10^{24}$	$2^{80}$	yobibyte	YiB	$2^{80}$	

*Trivia: nel 2016, l'intero volume dei dati presenti su Internet è stato stimato essere di 1.3 ZB.*

Differenze fra formati e l'assenza di una struttura comune sono caratteristiche tipiche dei big data.

Ciò è naturale, in quanto i dati possono provenire da fonti molto eterogenee.

Si pensi ai dati generati dagli utenti (come i log della navigazione web, le foto caricate sui social network, ...), e ai dati generati dalle macchine (telemetrie, log di server, ...).

L'eterogeneità dei formati e delle strutture rende difficile processare e memorizzare i big data utilizzando strumenti tradizionali.

## Velocità

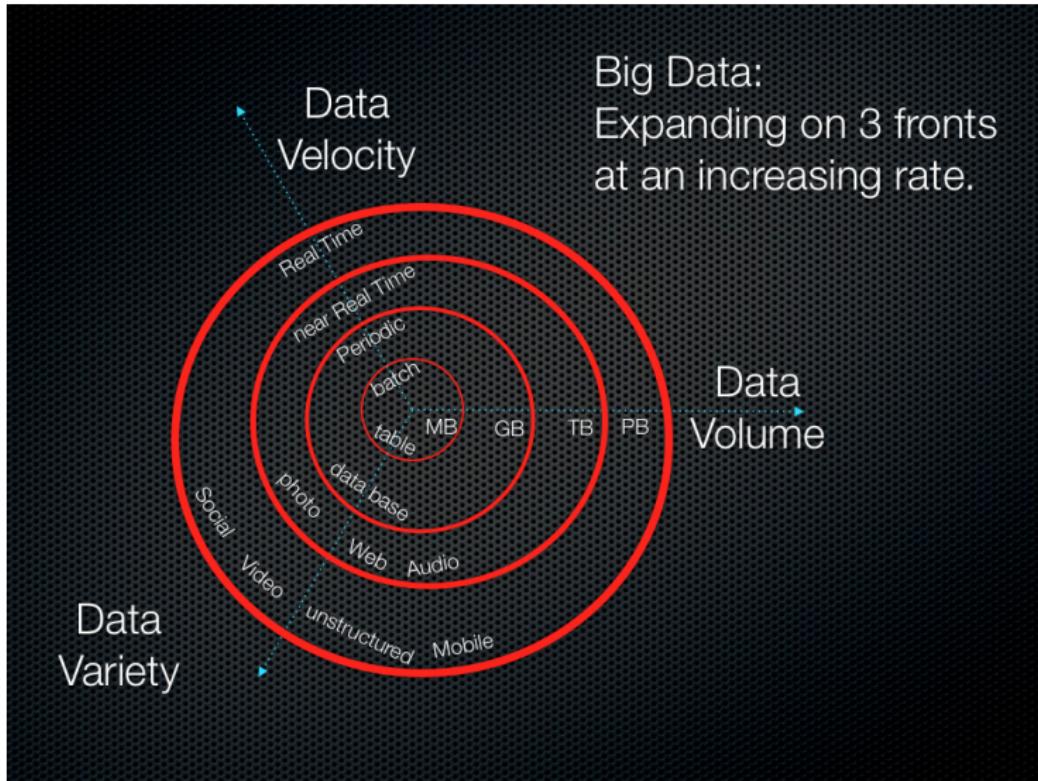
I dati acquisiti da sensori o strumenti scientifici possono essere trasmessi a velocità estremamente elevate.

Certi dati devono essere memorizzati nell'istante stesso in cui arrivano, in quanto sono volatili (es., stream di dati, log).

Per le aziende che si affidano a dati acquisiti velocemente è anche importante sfruttare/analizzare tali dati nel modo più rapido possibile.

*"Just in its 1st phase, the SKA telescope will produce some 160 TB of raw data per second that the supercomputers will need to handle."*  
<https://www.skatelescope.org/frequently-asked-questions/>

# Le '3V' originali



- **Volume:** dimensione dei dati
- **Varietà:** eterogeneità in formato e struttura
- **Velocità:** es., stream di dati
- **Variabilità:** cambiamenti nelle altre 'V'
- **Valore:** che può essere estratto/generato dai dati
- **Verità:** affidabilità dei dati

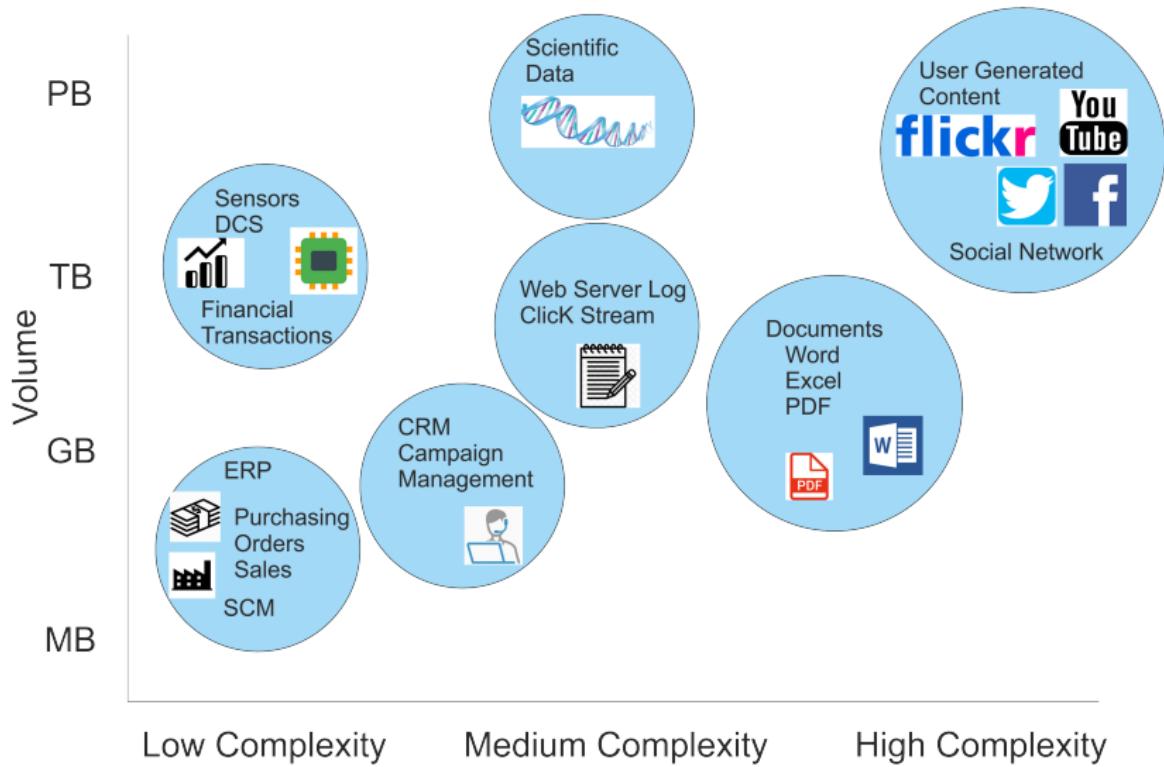
# Big Data

## E ancora più 'V'...

- **7 Vs:** <https://impact.com/marketing-intelligence/7-vs-big-data/>
- **10 Vs:** [tdwi.org/articles/2017/02/08/10-vs-of-big-data.aspx](https://tdwi.org/articles/2017/02/08/10-vs-of-big-data.aspx)
- **17 Vs:** <https://www.irjet.net/archives/V4/i9/IRJET-V4I957.pdf>
- **42 Vs:** <https://www.kdnuggets.com/2017/04/42-vs-big-data-data-science.html>

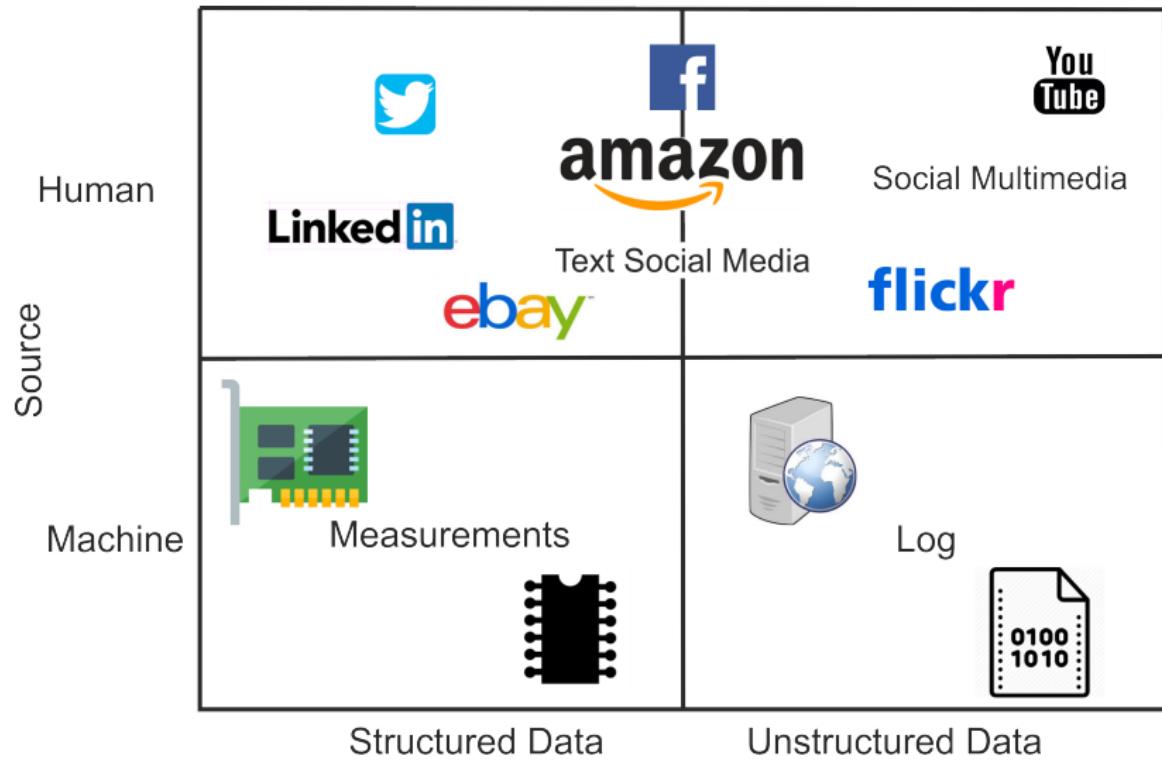
# Big Data

## Classificazione secondo volume e complessità



# Big Data

## Classificazione secondo livello di strutturazione



I Big Data comportano sia delle sfide che delle opportunità:

- *business*: i big data offrono alle aziende l'opportunità di sviluppare nuovi modelli di business o di ottenere vantaggi rispetto alla concorrenza
- *tecnologia*: le caratteristiche peculiari dei big data richiedono delle soluzioni adeguate
- *aspetti finanziari*: diversi casi d'uso mostrano che lo sfruttamento dei big data può portare a benefici economici. Molti altri si sono rivelati clamorosi fallimenti. A tal fine è importante valutare attentamente i costi legati alla loro gestione.
- *aspetti legali*: talvolta l'uso dei big data può comportare problematiche dal punto di vista della proprietà dei dati e il loro sfruttamento (es., scandalo Cambridge Analytica)

# Aspetti critici riguardanti i Big Data

## Qualità

La qualità dei big data riguarda diversi aspetti:

- *Completezza*: tutti i dati necessari alla descrizione di un'entità/transazione/evento sono presenti (es., non ci sono campi mancanti)
- *Consistenza*: assenza di informazioni in conflitto nei dati, anche rispetto alle regole del dominio
- *Accuratezza*: i dati rappresentano fedelmente l'informazione reale
- *Assenza di duplicazioni*: assenza di ridondanza in istanze, tuple, campi, tavelle, ...
- *Integrità*: rispetto ai vincoli propri di un DBMS (tipi dei dati, chiavi primarie, chiavi esterne, altro)

# Aspetti critici riguardanti i Big Data

## Fonti d'errore

I dati possono essere influenzati da diverse fonti d'errore:

- errori dovuti all'inserimento/compilazione manuale di campi
- errori dovuti a basi di dati mal progettate
- errori dovuti ad un errato trattamento dei dati (es., problemi nel processo di importazione)

Il processo di *data quality* si occupa di determinare quali dati offrono un livello accettabile di qualità, e quali no.

Se le analisi o le predizioni sono basate su dati di bassa qualità, i risultati saranno probabilmente errati o inaccurati (*garbage in = garbage out*).

# Aspetti critici riguardanti i Big Data

## Privacy e proprietà dei dati

La privacy e, soprattutto, la proprietà sono aspetti strettamente collegati alle possibilità di utilizzo dei dati:

- Il Web, con moltissimi *contenuti generati dagli utenti*, è una miniera di comportamenti, preferenze e pensieri personali. Dai social network possono essere estratte opinioni politiche, sessuali o religiose
- I dati riservati, come le cartelle cliniche dei pazienti, sollevano preoccupazioni sulla sicurezza: sono abbastanza al sicuro da possibili attacchi informatici?
- È impossibile non lasciare tracce elettroniche dei propri *movimenti* tramite: telefonate, carte di credito, dispositivi GPS, foto georeferenziate, ...

# Big Data - Casi d'uso

L'RFID (*Radio Frequency Identification*) è una tecnologia che consente l'identificazione di oggetti, animali, o persone.

Un *tag* RFID identifica univocamente l'oggetto al quale è collegato, e può essere letto da remoto attraverso le onde radio.

Applicazioni tipiche includono: gestione del magazzino, tracciamento, logistica, passaporti, sistemi anti-rapina.

Un singolo tag RFID contiene una quantità d'informazione limitata; considerando che nel mondo vi sono miliardi di tag RFID, essi possono essere considerati un buon esempio di big data, dato il volume complessivo dell'informazione generata.



I dati dal Web hanno un ruolo centrale nel dominio dei big data e sono caratterizzati da volume, varietà e velocità. Essi possono riguardare:

- Pagine HTML (in ogni lingua)
- Tweet
- Contenuto dei social network (Facebook, LinkedIn, ecc.)
- Post su forum e blog
- Documenti di vario formato: XML, PDF, Word, Excel, ecc.

In una banca di medie dimensioni si possono verificare facilmente diversi milioni di transazioni al giorno. Trattare ed analizzare solamente dati aggregati (es., su base mensile) può portare a una potenziale perdita di informazioni utili:

- la dimensione dei dati grezzi limita l'applicabilità di tecniche di analisi che possano essere impiegate in un tempo ragionevole
- sfruttando tecnologie adatte alla gestione dei big data (Hadoop MapReduce, Spark) è possibile sviluppare analisi riguardanti il delineamento di campagne di marketing, l'identificazione di anomalie di lavorazione, predizione di crediti insoluti, ...

L'Industria 4.0 è un ambito molto popolare al giorno d'oggi. È un processo che ha come obiettivo finale il realizzare una fabbrica (quasi) completamente automatizzata e interconnessa.

Ad esempio, considerando dati generati da sensori:

- possono essere utilizzati per svolgere monitoraggio real time, ad esempio ai fini della manutenzione predittiva
- strumenti specifici per la gestione e l'analisi di stream molto rapidi di dati possono rendersi necessari (es., Apache Flume)

# Big Data - Casi d'uso

## Internet Of Things

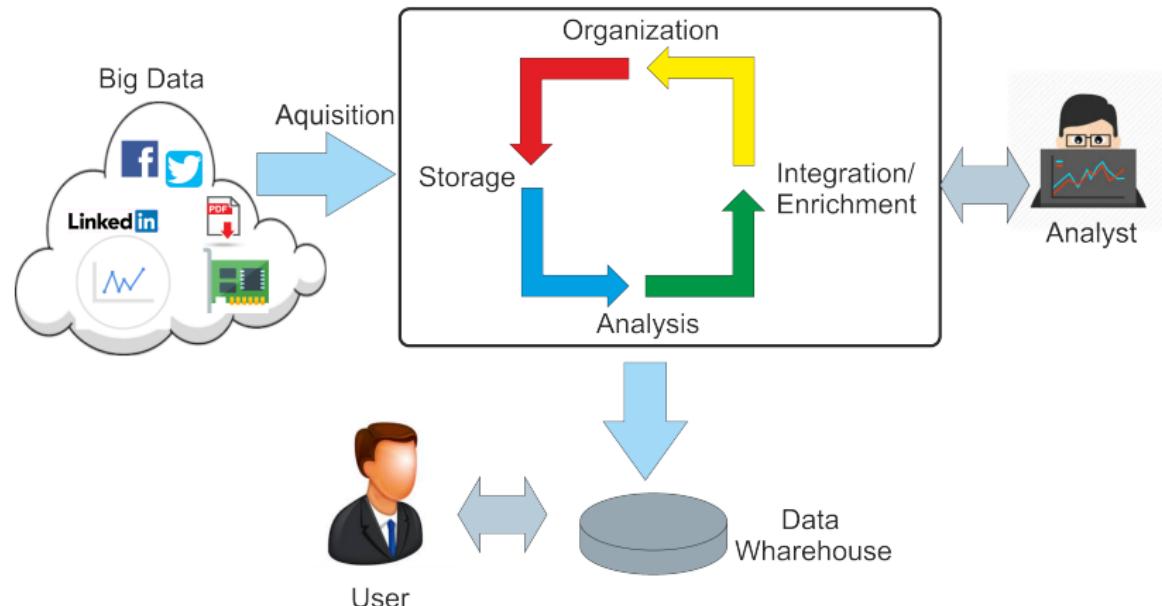
L'Internet Of Things (IoT) riguarda l'equipaggiamento di oggetti d'uso comune con sensori e funzionalità di connessione. In tal modo, tali oggetti diventano delle sorgenti di dati.

- automobili, elettrodomestici, sensori indossabili, ...
- i dati generati possono essere utilizzati per realizzare applicazioni nell'ambito della sorveglianza, manutenzione predittiva, sanità, ... caso d'uso dell'orto digitale che abbiamo visto nell'introduzione al corso

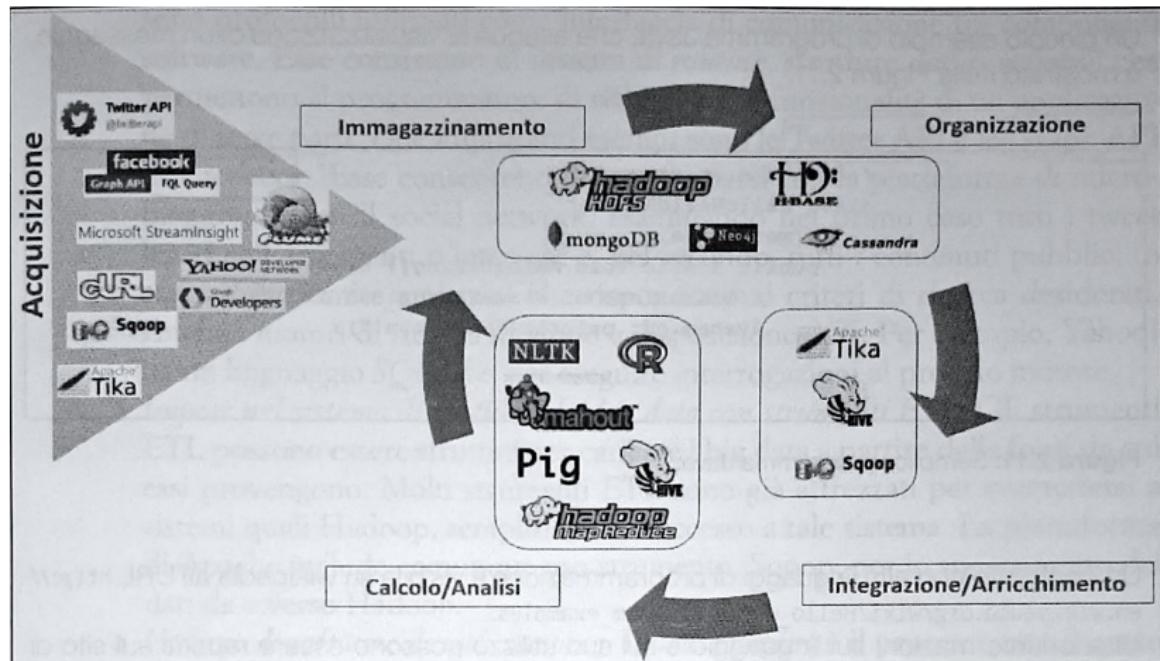


# **Big Data - Tecnologie**

## Il ciclo di vita dei big data



# Strumenti per i Big Data - Hadoop



# Che cos'è Hadoop?

Hadoop è una piattaforma open-source progettata per supportare il calcolo distribuito in modo affidabile e scalabile.

Hadoop è stato sviluppato da Doug Cutting e Mike Cafarella nel 2005 per risolvere un problema di scalabilità di un crawler open-source (Nutch).

La prima release è del 2008, come progetto Apache. Oggi è una collezione di progetti che fanno capo ad una stessa infrastruttura per il calcolo distribuito.

Il suo principale punto di forza è la capacità di sfruttare *commodity hardware* (economico) per gestire la scalabilità.

# Prima di Hadoop

Il processamento di enormi quantità di dati era effettuato per mezzo di soluzioni *High Performance Computing* (HPC) e *Grid Computing*.

HPC distribuisce il carico di lavoro su più nodi appartenenti ad un cluster; ogni nodo dialoga con un unico file system distribuito, connesso in rete.

Se il compito da svolgere richiede principalmente risorse di calcolo, il sistema funziona ottimamente. Se invece si rendono necessarie molte operazioni di lettura e scrittura di dati, la necessità di accesso al file system distribuito può facilmente portare ad un collo di bottiglia.

Sono richieste funzionalità di rete avanzate, veloci (e costose), come *Infiniband*, per ridurre quanto più possibile il tempo di trasferimento dati e la latenza.

# Vantaggi di Hadoop

Hadoop è più semplice da utilizzare rispetto a soluzioni HPC in quanto possiede librerie di più alto livello (e in quanto è ormai utilizzato da molti utenti).

Il criterio di partizionamento dei dati sui diversi nodi è cruciale per evitare costosi trasferimenti di rete (principio di *data locality*).

Hadoop è affidabile: è progettato per utilizzare *commodity hardware*; ha la capacità di gestire fallimenti hardware in maniera automatica.

La scalabilità è facile ed economica, basta aggiungere nodi al cluster; non c'è bisogno di hardware costoso e appositamente progettato.

# Cosa significa il nome Hadoop?

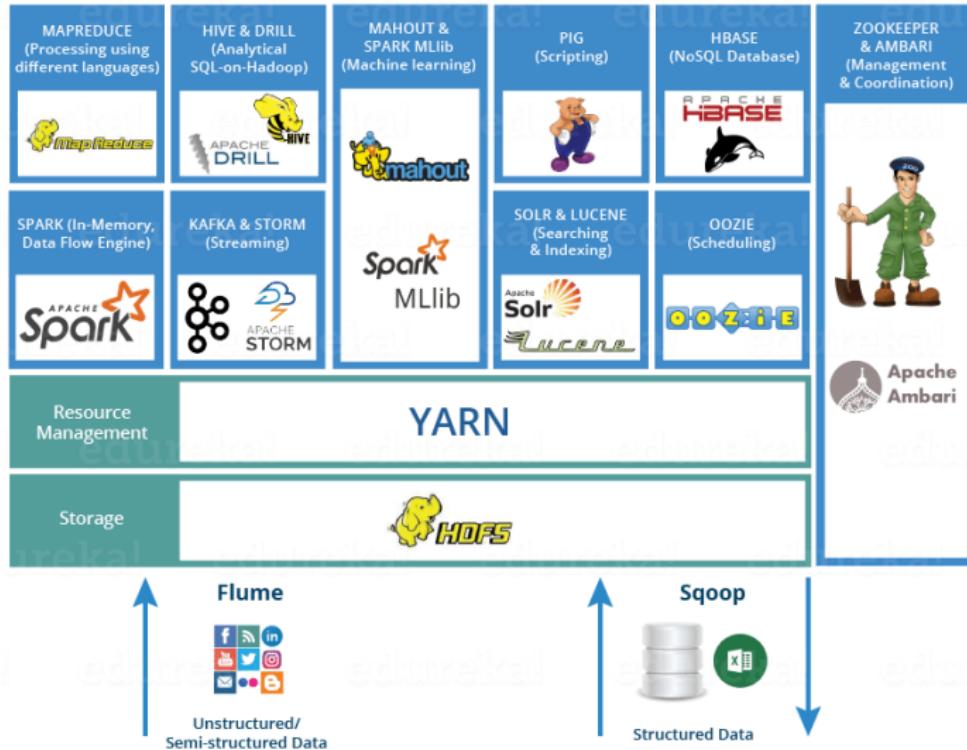


Doug Cutting

# Componenti principali di Hadoop

- *Hadoop common*: strato software che supporta gli altri moduli, mettendo a disposizione librerie e utilità.
- *HDFS*: file system distribuito che memorizza i dati su macchine di tipo commodity. Fornisce un modo efficace per accedere ai dati, garantendo la loro ridondanza per far fronte ai guasti. Qualsiasi formato di file è supportato, strutturato o meno.
- *YARN*: (Yet Another Resource Negotiator) piattaforma per la gestione delle risorse di elaborazione di un cluster, e della pianificazione delle operazioni lanciate degli utenti.
- *MapReduce*: sistema di calcolo parallelo in grado di trattare enormi quantità di dati, seguendo il principio del *divide et impera*.

# Hadoop - Ecosistema esteso





# Apache Sqoop

Il nome sta per *SQL to Hadoop*.

Strumento progettato per trasferire efficacemente grosse quantità di dati fra Apache Hadoop e basi di dati relazionali.

Supporta la lettura incrementale di una tabella relazionale e la scrittura verso HDFS, Hive, o HBase.

# Apache Flume, Storm, and Kafka

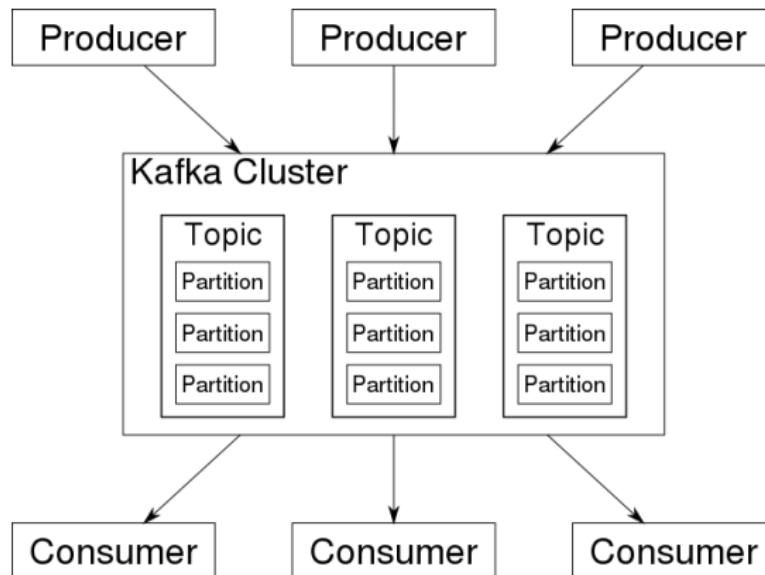
Consentono tutti di gestire stream di dati.

*Flume* è specificamente rivolto al trasferimento di dati non strutturati o semi-strutturati verso Hadoop (HDFS, Hive, HBase), in particolare dati da log.

*Kafka* è uno strumento più generico. Adotta un sistema di messaggistica distribuito in cui gli *editor* scrivono dati su *topic* e i *subscriber* leggono da tali topic, fornendo una piattaforma unificata, a bassa latenza e ad alta efficienza per la gestione dei feed di dati in tempo reale. È particolarmente indicato quando la destinazione dei dati non è (solo) Hadoop.

*Storm* è un sistema per il calcolo distribuito real-time che non viene utilizzato solo per lo streaming, ma include anche altre funzionalità come analisi in tempo reale, computazione continua, ...

# Apache Kafka



# HBase, Hive, and Drill

*HBase* è un database distribuito column-oriented, modellato su Google Bigtable e scritto in Java.

*Hive* è una soluzione per il data warehousing. Mette a disposizione HiveQL, un linguaggio simile a SQL che permette di eseguire query che sfruttano MapReduce, in modo totalmente trasparente.

*Drill* è un motore di query SQL schema-free. Permette di eseguire query SQL verso diversi database noSQL e file su disco (come file .xls).

# Mahout e Spark MLlib

Entrambe le librerie offrono delle implementazioni scalabili e distribuite di algoritmi per il machine learning.

*Mahout* include algoritmi per lo svolgimento di diversi task come classificazione, clustering, selezione degli attributi. Originariamente basato su MapReduce, oggi fa principalmente uso di Spark.

*Spark MLlib* è una libreria di machine learning basata su Spark. Include tutti gli algoritmi di machine learning più popolari, come random forest, K-means, LDA, ...

## Oozie, Solr e Lucene

*Oozie* consente la schedulazione di job Hadoop. Combina più job in modo sequenziale in una singola unità di lavoro.

*Solr* e *Lucene* sono delle librerie che mettono a disposizione motori di ricerca. Le loro funzionalità includono la ricerca full-text, l'indicizzazione in tempo reale, l'integrazione con basi di dati (relazionali e non), e la gestione di documenti complessi (es., Word, PDF).

# Spark

Spark è un sistema open-source per il calcolo parallelo, come Mapreduce.

A differenza di MapReduce, che deve effettuare letture e scritture da disco durante i processi di calcolo, Spark può svolgere tutte le operazioni in memoria.

Di conseguenza, gli sviluppatori affermano che Spark è in grado di eseguire programmi fino a 100 volte più velocemente di MapReduce, rendendolo adatto per il calcolo in tempo reale.

Tuttavia, Spark richiede molta memoria per caricare i processi. Al contrario, sfruttando il disco, MapReduce è in grado di lavorare con insiemi di dati molto più grandi di Spark.

# Ambari and Zookeeper

Ambari and Zookeeper forniscono funzionalità di amministrazione.

*Ambari* permette il provisioning, la gestione e il monitoraggio dei cluster Hadoop.

*Zookeeper* è un servizio per sistemi distribuiti che offre un archivio chiave-valore, utilizzato per fornire un servizio di configurazione distribuita.

# References

A. Pavlo and M. Aslett, *What's Really New with NewSQL?*, SIGMOD Record, June 2016 (Vol. 45, No. 2)

A. Rezzani *Big data. Architettura, tecnologie e metodi per l'utilizzo di grandi basi di dati*, Maggioli Ed. 2013

A. Rezzani *Big data Analytics*, Maggioli Ed. 2017