

Big Data Analytics in Cloud

Donato Malerba
Università degli Studi di Bari Aldo Moro
CINI - Big Data Lab

donato.malerba@uniba.it

Consortium
GARR

Workshop, Roma, 6 Aprile 2017



Articolazione della presentazione

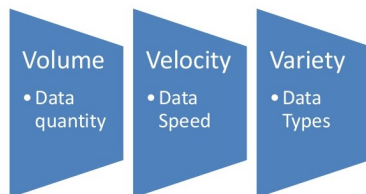
- Big Data: cosa sono
- Big Data: una prospettiva europea
- Big Data Analytics: workflow
- Big Data Analytics: potenzialità e sfide per il cloud computing
 - Località dei dati
 - Portabilità
 - Real-time interaction
- Big-data-as-a-Service: il cloud computing ispira
 - Progetto Toreador



2

Caratteristiche dei Big Data

- Le principali sono volume, velocità e varietà



3

Attenzione verso il tema

- Le conclusioni del Consiglio Europeo di Ottobre 2013 concentrano l'attenzione sulla economia digitale, sull'innovazione e sui servizi come fattori propulsivi della crescita economica e aumento dell'occupazione.
- La UE è chiamata a fornire le giuste condizioni strutturali per un singolo mercato per **big data** e **cloud computing**.

[Towards a thriving data-driven economy, 2014](#)

[Building a European data economy, 2017](#)



4

Come mai?

- Si prende atto che c'è una nuova rivoluzione industriale guidata dai dati, che sono accumulati in quantità crescente in modo esponenziale e ...
- ... spronano la produzione di nuovi prodotti e servizi, di nuovi processi aziendali, ma anche di nuove metodologie scientifiche
- **Data economy**: valore assoluto (e %) del bilancio UE
 - 2015: 272 miliardi di € (1,87%)
 - 2020: 643 miliardi di € (3,17%)



5



La Commissione Europea preme sui governi nazionali

"The European digital economy has been slow in embracing the data revolution compared to the USA and also lacks comparable industrial capability."

Nelle Kroes
(European Commissioner for the Digital Agenda)
"Data is not scary, or intrusive. With the right legal protection and anonymisation tools data is the fuel which lays the foundation of a new economy. Giving every kind of organisation the building blocks to boost productivity and performance, from farm to factory, from the lab to the shop floor, that's what Europe needs."



6



Quali sono gli ostacoli?

- Carenza di coordinamento di azioni intraprese da diversi paesi
- Infrastrutture insufficienti
- Ridotte opportunità di finanziamento
- Carenza di data scientist
- Un contesto legale frammentato e troppo complesso



7



Cosa è stato fatto dalla Commissione Europea?

- Ha attivato un partenariato pubblico-privato con l'industria europea (e con i centri di ricerca) finalizzato a finanziare bandi per innovazione e ricerca sul tema Big Data
- La parte privata è la [Big Data Value Association](#) (BDVA)
- Il finanziamento promesso è > 500M €
- Il ritorno atteso dalla Commissione è di 2000M €



8

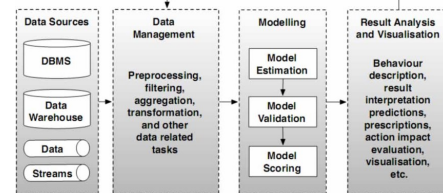


Cosa devono fare il mondo industriale e quello della ricerca?

- Promuovere la formazione di figure professionali specifiche
- Rimuovere gli ostacoli tecnici !
- Al di là della popolarità dell'argomento, lo sviluppo di reali soluzioni basate su Big Data Analytics resta un'impresa complessa e impegnativa.

Workflow per Big Data Analytics

- I dati sono usati per costruire modelli che vanno validati prima di essere utilizzati operativamente.



Workflow per Big Data Analytics

- Le principali tipologie di analitiche sono tre:
 - **Descrittive**: usano dati storici per identificare regolarità nei dati (pattern);
 - **Predittive**: usano dati storici e correnti per fare predizioni di valori per variabili dipendenti;
 - **Prescrittive**: usano dati storici per prescrivere delle azioni di cui valutano gli impatti.

Cosa frena?

- A causa di tutto questo, gli strumenti richiesti per memorizzare, ritrovare e analizzare i big data devono essere particolarmente raffinati,
 - Il che significa che richiedono costosi investimenti
 - Infrastrutture tecnologiche
 - Organizzative
- È per questo che il cloud computing offre soluzioni interessanti allo sviluppo della big data analytics.

Cloud computing: quale ruolo?

- Fra le caratteristiche essenziali del cloud computing si annoverano:
 - **Elasticità**: le richieste di risorse aggiuntive sono autogestite e automatiche in relazione alla richiesta.
 - **Misurazione (trasparente) delle risorse impiegate**, che consente un servizio di tipo use-and-pay (o pay-as-you-go).

(il cloud è il "quinto servizio" che si aggiunge a elettricità, acqua, gas, e telefonia)



13



Cloud computing: quale ruolo?

- Queste due caratteristiche sono essenziali per inquadrare la Big Data Analytics in una visione industriale in cui:
 - Le richieste di risorse per le fasi di *data management* e di *modeling* possono variare notevolmente e in modo dinamico in base alle necessità;
 - Il costo del servizio di Big Data Analytics deve essere basato sul modello pay-as-you-go.



14



Le sfide: data management

- Le soluzioni di analytics su cloud devono tener conto dei diversi modelli di deployment cloud adottati:
 - **Privato**: gestito dall'organizzazione (o da terzi per conto dell'organizzazione).
 - Ottimizzazioni più semplici ai fini delle analytics, potendo intervenire sulla struttura;
 - **Pubblico**: disponibile via Internet, con qualità di servizio (privacy, sicurezza, disponibilità) fissata da contratto.

Maggiore economicità delle analytics



15



Le sfide: data management

- **Ibrido**: combina entrambi i cloud (pubblico e privato).
 - Le applicazioni analitiche possono essere sviluppate per ambienti privati (> sicurezza), ma all'occorrenza si possono usare le risorse illimitate di un cloud pubblico (> elasticità);



16



Le sfide: data management

- Per quanto concerne la disponibilità di dati e modelli, questi potranno essere pubblici o privati, consentendo diverse configurazioni:

Dati \ Modelli	pubblico	privato
pubblico		
privato		

- La possibilità di costruire facilmente soluzioni portabili su diversi modelli di deployment cloud è una sfida aperta.



17



Le sfide: data storage

- Un aspetto chiave nelle prestazioni delle analytics su Big Data è la **località dei dati**.
- Il volume dei dati rende proibitivo il trasferimento dei dati per elaborarli.
- Ma questa è proprio l'opzione preferita dei sistemi HPC: portare i (relativamente) pochi dati dove c'è potenza di calcolo.

La computazione va spostata dove sono i dati!



18



Le sfide: data storage

- In effetti il modello MapReduce affermatosi per la Big Data Analytics sfrutta questo principio.
- Hadoop, una implementazione MapReduce open source, permette di creare cluster che usano lo Hadoop Distributed File System (HDFS) per partizionare e replicare i data set ai nodi dove saranno consumati dai *mapper*.
- Tuttavia *la virtualizzazione del cloud rende difficile il tener traccia della località dei dati*.



19



Le sfide: data storage

Altra sfida:


- Come memorizzare l'informazione in modo che essa possa essere facilmente migrata/portata fra data center / cloud provider?*
- Ci sono diverse proposte, ma nessuna ancora si è affermata.




20




Le sfide: costruzione e scoring del modello

- Le capacità di memorizzazione sono importanti, ma è altrettanto importante costruire il modello e utilizzarlo in fase operativa.
- Diverse soluzioni:
 - Zementis: strumenti di data analysis 


Modelli di deployment\servizio	Infrastructure-as-a-Service	Software-as-a-Service
pubblico		
privato		

 UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO 21


Le sfide: costruzione e scoring del modello

- Le capacità di memorizzazione sono importanti, ma è altrettanto importante costruire il modello e utilizzarlo in fase operativa.
- Diverse soluzioni:
 - Google Prediction API : strumenti per la creazione, condivisione e uso di modelli


Modelli di deployment\servizio	Infrastructure-as-a-Service	Software-as-a-Service
pubblico		
privato		

 UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO 22

Le sfide: costruzione e scoring del modello


- Le capacità di memorizzazione sono importanti, ma è altrettanto importante costruire il modello e utilizzarlo in fase operativa.
- Diverse soluzioni:
 - Apache Mahout : strumenti per la costruzione di librerie di algoritmi scalabili di machine learning

Modelli di deployment\servizio	Infrastructure-as-a-Service	Software-as-a-Service
pubblico		
privato		

 UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO 23

Le sfide: costruzione e scoring del modello

- La sfida è quella di sviluppare tecniche che siano in grado di sfruttare l'elasticità e la scalabilità dei sistemi cloud
- Si può anche prefigurare una soluzione **"prediction and analytics-as-a-service"** (o "big data analytics-as-a-service") in cui diversi provider competono per costi e prestazioni dei servizi di big data analytics.

 UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO 24

Le sfide: visualizzazione

- Un tema a parte è quello della visualizzazione, che completa i progetti di Big Data Analytics.
- Idealmente gli utenti vorrebbero visualizzare i dati elaborati nel cloud avendo la stessa esperienza come se fossero disponibili localmente.
- In molti scenari, la visualizzazione e l'interazione utente su cloud è ostacolata dal collo di bottiglia della comunicazione su rete.



25



Le sfide: visualizzazione

- Le piattaforme cloud assomigliano a uno scenario di batch job
- Gli ambienti Cloud non supportano adeguatamente la visualizzazione in tempo reale
- L'analista non può essere messo nel ciclo, riducendo la possibilità di abbattere i tempi per l'estrazione di modelli utili dai dati.



26



Il progetto Toreador

- TOREADOR (www.toreador-project.eu) è uno dei principali progetti di ricerca finanziati dalla Commissione Europea nell'ambito della call Big Data Research del programma H2020.
- Coordinato dal Laboratorio Nazionale Big Data del CINI
- Principal Investigator: *Prof. Ernesto Damiani*.
- Partecipano SAS, Atos, Engineering che fanno parte della Big Data Value Association.



27



Il progetto Toreador

- **Obiettivo primario:** aiutare le organizzazioni e le PMI europee che desiderano eseguire computazioni di analitiche Big Data, diminuendo l'investimento iniziale e il livello di competenza tecnologica necessari.
- **Come raggiungerlo?**



28



Il progetto Toreador

- Sviluppando **modelli ontologici** indipendenti dalla tecnologia per rappresentare
- (i) le entità (dati, algoritmi, metriche) coinvolte nelle computazioni di analitiche su Big Data e
- (ii) le attività che tali computazioni richiedono: l'acquisizione, memorizzazione e preparazione dei dati, le analitiche da calcolare, la distribuzione e parallelizzazione dei calcoli, la presentazione e interpretazione dei risultati.



29



Il progetto Toreador

- **Metodologia:** accompagnare il committente di un'analitica su Big Data attraverso tre fasi:
- *Individuazione* degli obiettivi (funzionali e di prestazioni) della computazione da eseguire (**modello dichiarativo**)
- *Derivazione* assistita di una descrizione della computazione indipendente dalla tecnologia (**modello procedurale**).
- *Traduzione* semi-automatica del modello procedurale in una procedura operativa (**modello di messa in opera**) eseguibile nell'ambiente di calcolo desiderato (sistema ICT del committente o in **cloud**)



30



Il progetto Toreador

- Un elemento importante della prima fase della metodologia TOREADOR è la **gestione dei conflitti** tra scelte diverse nei modelli delle varie fasi (ad esempio, tra la rappresentazione del dato e l'accuratezza dei risultati dell'analitica).



31



Il progetto Toreador

- La metodologia TOREADOR comprende la modellazione dello **status giuridico** del dato (dato personale/non personale, consenso, scopo).
- TOREADOR ha affidato a un partner del consorzio, lo studio legale internazionale **Bird&Bird**, uno studio delle **casistiche** che permettono di modellare lo stato giuridico del dato alla luce della recente **normativa europea**.



32



Il progetto



- Duplice scopo della modellazione dello stato giuridico del dato:
 1. aiutare il committente di una analitica Big Data a specificare obiettivi di anonimizzazione adeguati allo status giuridico dei dati su cui opera, nel contesto dell'analitica che vuole eseguire;
 2. includere automaticamente nei modelli procedurali che descrivono la computazione dell'analitica, ed in particolare nel modello della fase di preparazione del dato, le necessarie procedure di anonimizzazione (ad esempio, hashing non reversibile).



33



Il progetto



- La metodologia TOREADOR individua poi, in sede di risoluzione dei conflitti tra gli obiettivi delle varie fasi, quali sono le analitiche compatibili con le procedure di anonimizzazione da adottare e gli obiettivi di accuratezza raggiungibili.



34



Il progetto



- Oltre ad essere traducibile nella messa in opera di una computazione eseguibile, il modello TOREADOR che specifica lo status giuridico assegnato ai vari dati e le corrispondenti procedure di anonimizzazione da adottare è utilizzabile ai fini contrattuali (ad esempio, per specificare le condizioni di fornitura dell'analitica) e per ispezione/validazione da parte di terzi.
- Questo consente di definire un Privacy Seal per modelli, che permetterebbe di certificare il modello e poi di individuare la tecnologia che lo realizza.



35



Il progetto



- In questo modo si intende anche affrontare il problema della privacy che spesso impedisce l'uso di sistemi cloud pubblici in applicazioni di big data analytics.



36

