Data-Intensive Computing, CRS4

Caccia al valore nei Big Data

Luca Pireddu

23 gennaio 2020

Mi presento...

- Ricercatore al CRS4 dal 2009
- Gruppo *Calcolo distribuito*, settore *Data-intensive computing*
- Nato e cresciuto in Canada (vicino a Toronto)
- Studiato informatica alla Laurentian University e University of Alberta; dottorato all'Università di Cagliari
- Dal 2011 lavoro su problemi che richiedono calcolo a larga scala

CRS4 – Chi siamo?

- Centro di ricerca multidisciplinare
  - Non a scopo di lucro; società della Regione
  - Sede principale a Pula; sede secondaria a Cagliari
  - Operativo dal 1992; Staff di circa 130 persone
- Attività prevalentemente orientate verso problemi di ricerca in informatica applicata a vari contesti
- Competitivo a livello internazionale
  - Gran parte delle attività autofinanziate attraverso progetti di ricerca europei, nazionali o regionali

Risorse e infrastrutture

- Infrastruttura di calcolo – centinaia di nodi, petabyte di storage
- Connesso ad alta velocità alla rete nazionale GARR
- Uno dei più grandi laboratori di sequenziamento genomico in Italia

Risorse che permettono al CRS4 di supportare importanti progetti di ricerca

Risorse e infrastrutture

- Infrastruttura di calcolo – centinaia di nodi, petabyte di storage
- Connesso ad alta velocità alla rete nazionale GARR
- Uno dei più grandi laboratori di sequenziamento genomico in Italia

Risorse che permettono al CRS4 di supportare importanti progetti di ricerca

- in particolare ricerca che presentano problematiche di tipo “Big Data”

Index

Cosa sono i Big Data?

Cosa vuol dire “Big Data”?

Gartner la definisce come:

> *high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing*

Insiemi di dati troppo grossi, complessi, o generati da processi troppo “rapidi” per essere trattati con metodi convenzionali.

- ✗ Workstation
- ✗ Pennina USB
- ✗ Excel
- ✗ Scripting, calcolo multi-core

Le quattro V

Quanto sono “Big” i miei Big Data?

Uno dei metodi standard per caratterizzare le collezioni di “Big Data”:

Volume: la quantità di dati (*data stanziali*)

Variety: le forme di dato – strutturate e non (e.g., testo, immagini), provenienti da diverse fonti

Velocity: la velocità alla quale i dati vengono generati e devono essere gestiti (*dati in movimento*)

Veracity: il livello di rumore o di errori

Le quattro V: esempio dalla genomica

Alcuni anni fa, in uno studio su una malattia rara sono stati raccolti i genomi e altri dati fenotipici di 1200 persone:

Volume	1200 persone $\times$ 250 GB = 300 TB
Variety	DNA, variabili numeriche e booleane
Velocity	i sequenziatori generavano circa 700 MB/minuto
Veracity	P(error) $\approx$ 1%

Le quattro V: esempio dalla genomica

Alcuni anni fa, in uno studio su una malattia rara sono stati raccolti i genomi e altri dati fenotipici di 1200 persone:

Volume	1200 persone $\times$ 250 GB = 300 TB
Variety	DNA, variabili numeriche e booleane
Velocity	i sequenziatori generavano circa 700 MB/minuto
Veracity	P(error) $\approx$ 1%

Le "quattro V" ci permettono di capire meglio il problema e identificarne i requisiti

Esempi di Big Data

Alcuni esempi famosi di problemi per cui vengono gestiti Big Data:

- Google ads, Visa (fino a 1,5 milioni di transazioni al minuto!)
  - Hanno fatto scalare le loro operazioni a grandi dimensioni
- Large Hadron Collider, Airbus A380 ($\approx$ 10k sensori per ala)
  - Analizzano processi e macchinari nel dettaglio
- La guida autonoma di Tesla
  - Hanno automatizzato (+ o -) un'operazione complessa (attraverso l'intelligenza artificiale)

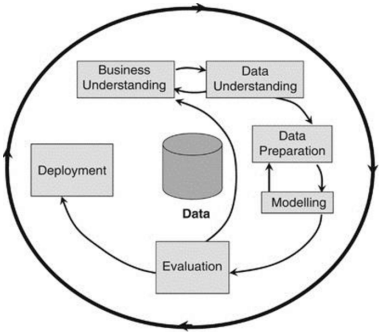
Cosa fare coi Big Data?

L'obiettivo…

In generale, puntiamo ad osservare un processo o fenomeno per:

- capirlo e/o monitorarlo
- influenzarlo, ottimizzarlo o predirne l'esito

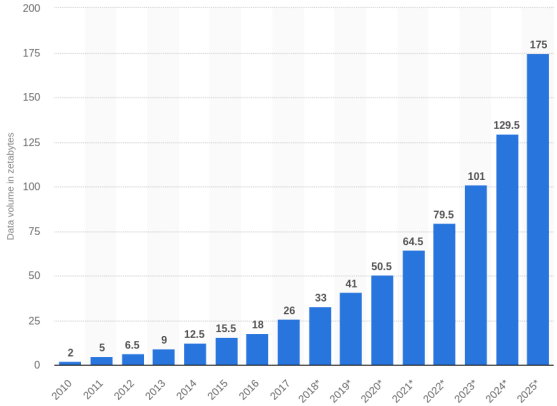
- Per esempio:
  - Un processo biologico o fisico
  - Un processo manufatturiero
  - Un processo psicologico
- Formiamo il nostro modello del processo dalle nostre osservazioni (i dati)
- Aggiorniamo il modello con la frequenza possibile/necessaria



Cross-Industry Standard Process for Data Mining (CRISP-DM)

Crescita dei dati creati

La quantità di dati generati globalmente è in crescita esponenziale.



Previsione volume (ZB) totale di dati creati per anno (fonte: Statista/IDC)

Zettabytes (ZB) = 10^{21} bytes = 1 milione petabyte = 1 milione di milioni di gigabyte

Crescita dei dati creati

I dati sono generati da sensori sempre più pervasivi; per esempio:

- smart-phone, smart-watch, PC
- Dispositivi IoT – nelle nostre case, fabbriche, città, addosso a noi
- Automobili con autopilota
- Dispositivi medici digitali
- Dispositivi automatici/robotici per applicazioni industriali
  - lettori, attuatori, braccia robotiche…
- Satelliti

Potenziale opportunità

La crescente disponibilità di dati offre nuove potenziali opportunità

Dal punto di vista economico…

> *Worldwide Big Data market revenues for software and services are projected to increase from $42B in 2018 to $103B in 2027 (CAGR of 10.48%) (Forbes)*

> *AI augmentation will create $2.9 trillion of business value and 6.2 billion hours of increased worker productivity in 2021 (Gartner)*

E dal punto di vista sociale…

> *P. Ström, et al. Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study. In The Lancet.*

Fattori abilitanti

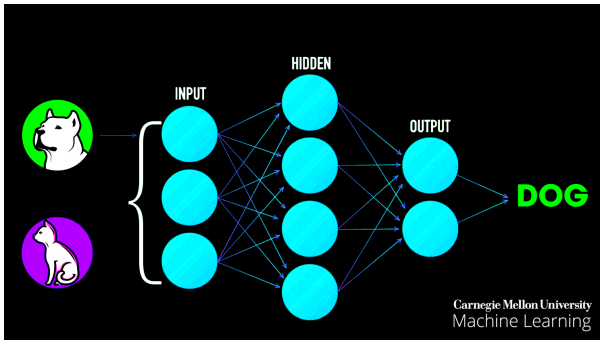
La valorizzazione di collezioni di dati è favorita da una serie di fattori abilitanti:

- Metodi e algoritmi
- Tecnologie e piattaforme
- Infrastruttura di calcolo

Nell'arco dell'ultima decade ci sono stati importanti progressi nello stato dell'arte e nelle possibilità di accesso a queste risorse.

Intelligenza artificiale e Big Data

- Le tecniche di intelligenza artificiale moderne sono legate strettamente ai Big Data
  - in primo piano, Deep Learning e più in generale Machine Learning
- Si tratta di tecniche per creare modelli matematici che catturano le relazioni "nascoste" nei dati



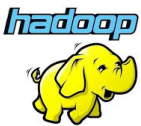
Intelligenza artificiale e Big Data

- In linea di massima, più dati = miglior modello
- Il problema di creare modelli di intelligenza artificiale sofisticati è un problema Big Data
  - E.g., modelli per il riconoscimento di oggetti generici da immagini addestrati su decine di milioni di immagini

Progressi negli algoritmi di Deep Learning in combinazione con tecniche Big Data ha permesso il grande sviluppo di applicazioni in questo ambito

Tecnologie e Piattaforme abilitanti

- Esistono una serie di piattaform open source per gestire e trattare Big Data
- Alcuni esempi…



Perché open source?

Le piattaforme in uso comune sono sviluppate da più aziende private. Perché?

- collaborare sulle basi
- competere sugli aspetti più specifici dei casi d'uso

Piattaforme di data analytics

Tema comune

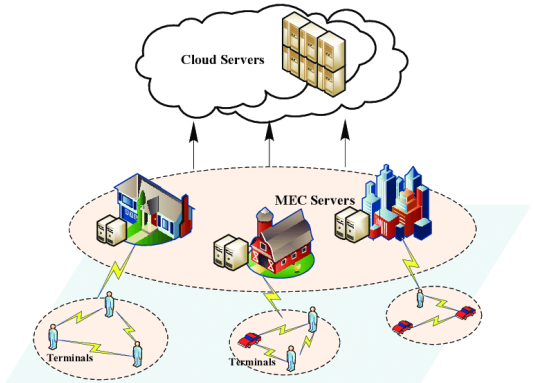
Scalabilità attraverso la distribuzione del lavoro su molti nodi di calcolo

- Le piattaforme astraggono le funzionalità necessarie per distribuire il lavoro
- Offrono un modello di programmazione in cui implementare il proprio algoritmo
- E.g., Hadoop MapReduce (la prima piattaforma): 2 funzioni
  - $map(x) \rightarrow y$: trasforma oggetto x in un oggetto y
  - $reduce(ys) \rightarrow z$: aggrega una collezione di y in un nuovo valore

Edge computing

Architettura dove spostiamo parte dell'"intelligenza" alle estremità della rete

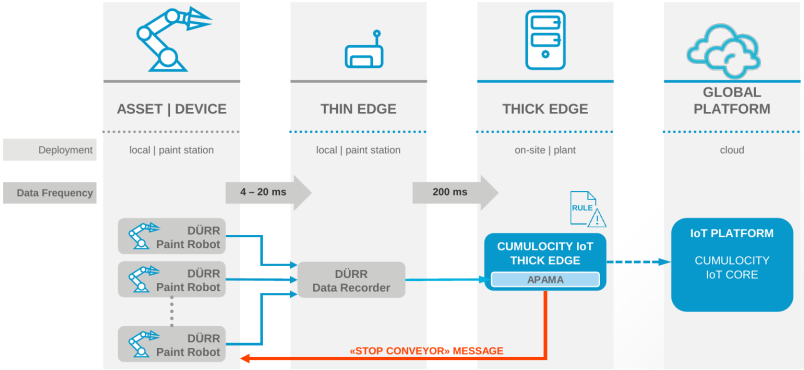
- Necessario per:
  - applicazioni a bassa latenza
  - ridurre la quantità di dati da spedire in centrale
- Sistema gerarchico di elaborazione



Credits: Yuan, et al., 2018

Edge computing: esempio industriale

- Monitoraggio in tempo reale di operazioni di verniciatura automatica
  - con blocco produzione in caso di problemi
- Accumulo e analisi di dati aggregati nella piattaforma centrale



Credits: Software AG

Infrastruttura di calcolo abilitanti

- Big Data implica l'utilizzo di risorse di calcolo rilevanti
- L'acquisto di queste risorse richiederebbe una spesa/investimento anticipato importante
  - Richiede anche le competenze per la loro gestione
- Per molti utilizzare risorse dal cloud può essere un'opzione migliore

Cloud computing

Cloud computing = risorse di calcolo a consumo

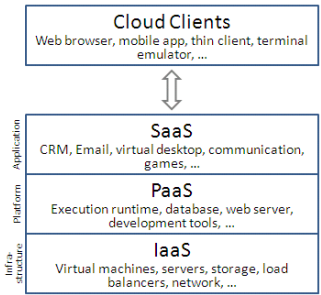
- Il cloud computing formalizza l'interfaccia tra utilizzatore e infrastruttura di calcolo
- Rende possibile affidare l'implementazione e la gestione dell'infrastruttura è affidata a specialisti
  - verso un venditore esterno
  - oppure verso un altro gruppo interno all'organizzazione (cloud privato)
- L'infrastruttura è relativamente generica: non è specializzata per applicazioni
- Modello a consumo permette di acquisire accesso temporaneo a risorse di calcolo senza investimento anticipato

Cloud computing: modello as a Service

- Modello *as a Service*: risorse gestite dal fornitore
- Infrastructure as a Service: ore macchina, spazio di storage, connessione di rete…

Ma anche **servizi più ad alto livello**:

- Platform as a Service
- Service as a Service
- Function as a Service



Modelli di servizio di un Cloud (img: Wikipedia)

Esempi servizi cloud commerciali

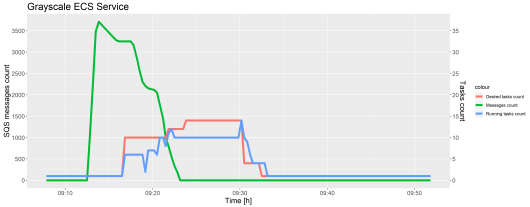
- AWS Rekognition per il riconoscimento di immagini
- AWS Elastic MapReduce PaaS
- Azure Databricks (Spark PaaS)
- Google Kubernetes Engine

AWS Rekognition

Cloud computing: Infrastruttura

- Tratto distintivo del “cloud”: API programmabile
- Permette di creare software per automatizzare la configurazione del sistema
  - importante per gestire in maniera efficiente livelli variabili di *velocity*
  - come dicono a Netflix, “automate everything”!

E.g., *Autoscaling*
Modifichiamo dinamicamente le risorse dedicate ad un compito in base alla richiesta



Credits: pgs-soft.com

Big Data al CRS4

Big Data al CRS4

- Uno dei maggiori focus delle attività del CRS4 è studiare e applicare i metodi per estrarre informazioni da dati
- Nel settore *Data-intensive Computing* ci focalizziamo in particolare sul trattamento di grosse quantità di dati
- Due progetti attivi al momento sono particolarmente rilevanti a questo ambito:
  - TDM: Tessuto Digitale Metropolitano
  - DeepHealth: Deep-Learning and HPC to Boost Biomedical Applications for Health

Tessuto Digitale Metropolitano

Obiettivo

Studiare metodi e tecnologie per migliorare la consapevolezza dei consumi energetici e limitare i rischi meteorologici.

- Soggetti attuatori: CRS4 e Univ. di Cagliari
- Finanziamento: POR FESR 2014-2020
- Durata: 48 mesi (fine a giugno 2021)

CRS4 IDEAS BECOME LIFE

UNIONE EUROPEA
Fondo europeo di sviluppo regionale

REPUBBLICA ITALIANA

REGIONE AUTÒNOMA DE SARDIGNA
REGIONE AUTONOMA DELLA SARDEGNA

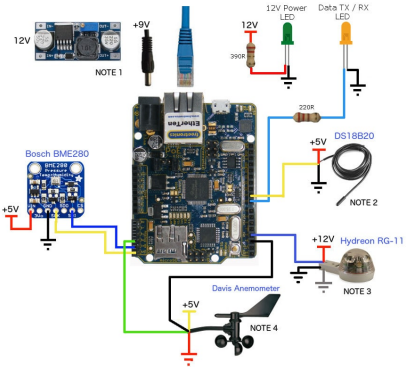
POR FESR
SARDEGNA 2014-2020

SARDEGNA RICERCHE

Tessuto Digitale Metropolitano

Più in concreto,

- Abbiamo creato dei dispositivi edge con connessi sensori meteo/ambientali e sensori di consumo elettrico
- Stiamo distribuendo i dispositivi al Comune di Cagliari e a volontari privati per installarli in vari spazi nell'Area Metropolitana
- I dispositivi raccolgono e preprocessano i dati dei sensori; ogni pochi minuti inviano alla piattaforma di analitica



TDM edge device

TDM: Data integration

Nella piattaforma integriamo anche altri dati georeferenziati di vario tipo, e.g.:

- Radar meteorologico dell'Univ. di Cagliari
- Meteo mosaico della Protezione Civile
- Dati satellitari (Copernicus)
- Esiti di simulazioni meteo

TDM: piattaforma di analitica

Interrogazioni omogenee su volumi spazio-temporali

- Dati visti come serie temporali multi-dimensionali georefenziate
- Possibilità di integrare dati eterogenei
  - e.g., sensori, immagini satellitari, simulazioni atmosferiche
- Scalabile: può gestire migliaia di interrogazioni REST al secondo
- Sfrutta lo stato dell'arte in technologie cloud-native
  - Kubernetes, OpenStack Cloud, Apache Kafka,…


```
r = requests.get(
    'http://tdmq:8000/api/v0.0/sensors/{}/timeseries'.format(code),
    params=args)
```

Progetto DeepHealth

Obiettivo

1. Creare una libreria per deep learning e computer vision che funzioni in maniera trasparente su risorse di calcolo eterogenee distribuite
2. Dimostrarne l'efficacia in ambito sanitario.

- Soggetti attuatori: CRS4 e altri 21 partner da 9 paesi europei
  - UP Valencia, Barcelona Supercomputing Center, Philips, Everis, Thales, …
- Finanziamento: H2020
- Durata: 36 mesi (fine a dicembre 2021)

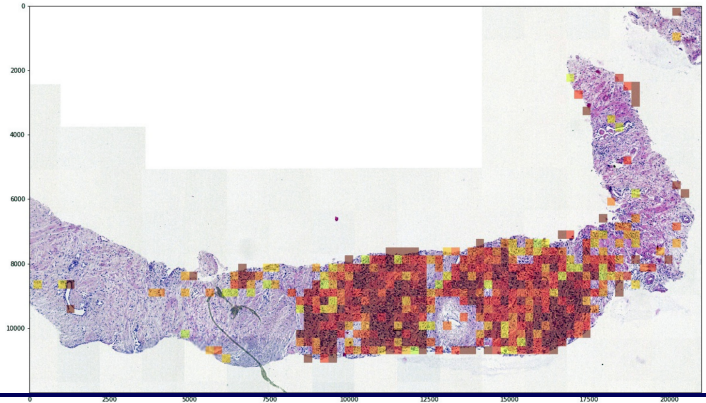
DEEPHEALTH

- Si lavora verso il completamento di un toolkit open source con due componenti principali:
  - EDDLL: European Distributed Deep Learning Library
  - ECVL: European Computer Vision Library
- Le librerie saranno integrate in sette diverse piattaforme biomediche
- Ne sarà provata l'efficacia in 14 diversi casi d'uso

DeepHealth: Caso d'uso CRS4

- Oltre al contributo tecnico, il CRS4 sta collaborando ad un caso d'uso nel campo della patologia digitale

Identificazione e classificazione automatica di tessuto tumorale prostatico



DeepHealth: Caso d'uso CRS4

- Dataset di circa 18 TB di immagini da microscopia ad alta risoluzione
- Addestriamo modelli Deep Learning per predire `Prob(tumore)` per ogni quadretto dell'immagine (256x256 pixel)
- Il modello di predittivo cosí generato verrà integrato nella nostra piattaforma di patologia digitale
- Verrà provato in un contesto clinico in un processo di "active learning"
  - Gli utenti (i medici) forniranno feedback riguardo le predizioni che verranno integrate nel processo di addestramento.

Conclusioni

Ethical and privacy implications

- Le possibilità dei mezzi a disposizione in questo ambito rendono possibile ledere la privacy delle persone
- Il potenziale lucro motiva servizi di dubbia morale
- È relativamente facile indurre persone ignare a consegnare informazioni apparentemente anonime, o sbagliare la procedure di anonimizzazione

Un caso recente

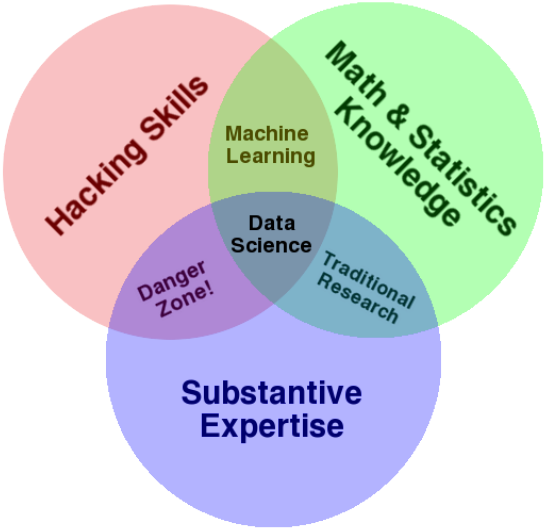
S.A. Thompson and C. Warzel, *Twelve Million Phones, One Dataset, Zero Privacy*, New York Times (2019)

- Con la sola localizzazione del cellulare delle persone si possono inferire, abitudini, tragitti, relazioni, e altro

Where to go from here?

- Sperimentare coi servizi cloud e con piccoli dataset è veramente accessibile
- Le varie Paas e SaaS permettono di ridurre lo sforzo necessario per prototipare idee
  - Sono in genere anche ben documentati
- Molti dati disponibili gratuitamente
  - e.g., smart cities, satelliti Copernicus, repository di dati scientifici

Data science skills



Competenze necessarie nel campo del data science (credits: Drew Conway)

Conclusioni

- Abbiamo visto cosa vuol dire “Big Data” e come inquadrare un problema in questo ambito
- La tecnologia e i metodi a disposizione rendono possibile osservare in dettaglio processi di varia natura
  - Per monitorare e prendere in tempi brevi decisioni supportate dai dati
  - Per accumulare serie storiche per migliorare o prevedere i comportamento del processo
- Nell’industria aver implementato questo tipo di approccio rappresenta ora un vantaggio strategico
  - Ci si può aspettare che nei prossimi anni diverrà la norma

Conclusioni

- I potenziali benefici sono anche sociali, attraverso innumerevoli possibili applicazioni
  - ambito sanitario, urbano, etc.
- La tecnologia usata in maniera inappropriata può essere usata anche per ledere la nostra privacy

Come si collabora con il CRS4?

- Collaborazioni aperte
- Partecipazione a bandi di progetti di ricerca
- Erogazione di servizi e consulenze

Parco Tecnologico della Sardegna
Loc. Piscina Manna, 09050 Pula (CA)
www.crs4.it