Marco Aldinucci

L'infrastruttura necessaria per creare interoperabilità tra pubbliche amministrazioni

Abstract: L'articolo affronta il tema dell'interoperabilità dal punto di vista informatico, ponendo l'accento sulle infrastrutture necessarie affinché la comunicazione tra sistemi informatici pubblici sia possibile. La struttura a silos su cui si basa il sistema informativo della pubblica amministrazione italiana risulta inadeguato all'approccio della *big data analysis* che, a contrario, richiede la piena comunicabilità tra sistemi informativi affinché il reperimento dei dati su cui condurre sperimentazioni sia quanto più facile e mirato.

1. *Introduzione*

Con piacere intervengo in questo convengo sulla pubblica amministrazione con i BigData, la cui natura multidisciplinare arriva ad abbracciare la scienza dell'informazione e dei dati che è la materia della mia ricerca. Un convegno con apertura coraggiosa e necessaria, visto la crescente interdipendenza degli aspetti etici, giuridici e tecnologici nella gestione dei dati.

In particolare vorrei parlare delle banche dati della Pubblica Amministrazione e della loro integrazione. Le banche dati sono tradizionalmente organizzate secondo un modello detto "datasilos": insiemi di dati contenuti in diversi database ognuno dei quali organizzato secondo uno *schema*. Lo schema deve essere definito al momento della progettazione del database prima di iniziare a memorizzare i dati e definisce l'organizzazione logica dei dati in esso contenuti. Immaginiamo come semplice caso uno schema "scuola" con due tabelle: alunni e classi; i campi di ogni tabella hanno degli attributi, per esempio nome, cognome, matricola per gli alunni e anno e sezione per le classi. Lo schema definisce anche le relazioni fra alunni e classi, per esempio il fatto che ogni alunno deve appartenere a una sola classe ma che ogni classe deve avere almeno un certo numero di alunni. In definitiva lo schema descrive la conoscenza ex-ante che il progettista del database ha dei dati e delle loro correlazioni.

Uno dei fattori di complessità nell'incrociare dati di database diversi è proprio l'assenza di un unico schema. Incrociare i dati significa sostan-

zialmente creare un nuovo database e quindi un nuovo schema che li unisca. Un'operazione che deve essere ripetuta ad ogni nuovo incrocio di dati con una crescita esponenziale della complessità e del costo. Gli aspetti tecnici non sono l'unico problema. Giuridicamente, in accordo alla General Data Protection Regulation (GDPR), i dati possono avere vincoli di localizzazione, vincoli temporali, vincoli di accesso. La copia dei dati è un'operazione che può facilmente indurre violazioni di questi vincoli; una copia dei dati potrebbe rimanere accessibile anche dopo la data ultima di validità delle informazioni. Ritornerò su questo aspetto.

Un altro motivo per cui l'approccio datasilos è poco adatto alla ricerca di nuove informazioni – come correlazioni fra alcune categorie di fornitori e di bandi pubblici – è il fatto che i database esprimono in modo efficiente solo le relazioni definite dallo schema al momento della progettazione e quindi sono intrinsecamente poco adeguati alla ricerca di nuove relazioni fra i dati.

Oggi in Italia sono presenti oltre diecimila datacenter con diversi database della pubblica amministrazione: i dati comunali (fino ad arrivare al singolo PC nel comune di montagna), regionali, dell'anagrafe, degli ospedali, delle targhe automobilistiche, delle imposte, etc. Incrociare i dati è praticamente impossibile.

2. *L'approccio BigData*

Allora come si va oltre? Un paradigma di successo per incrociare i dati è il cosiddetto approccio BigData, che si basa sul modello "datalake". Un datalake, figurativamente un lago di dati, è un insieme di dati eterogenei (testo, immagini, video, etc.) che si distingue da un datasilos proprio per l'assenza di uno schema definito ex-ante; un lago che riceve dati da diversi immissari, cioè sorgenti di dati come dati dell'anagrafe, dati provenienti dai social network, dall'agenzia delle entrate. In un datalake è possibile cercare correlazioni non originariamente previste o prevedibili. La ricerca di queste correlazioni è dominio della cosiddetta "data science", che è oggi supportata da strumenti software per "BigData analytics" molto maturi e disponibili su diverse piattaforme on-premise e cloud commerciali (Microsoft Azure, Amazon AWS) e opensource (OpenStack). Esattamente in questo punto l'Intelligenza Artificiale entra in gioco. L'insieme delle tecniche legate all'Intelligenza Artificiale altro non sono che metodi per analizzare dati al fine di segmentarli, ridurli in cluster, classificarli, etc. Da questo punto di vista l'unica differenza fra i metodi statistici tradizionali e

quelli dell'Intelligenza Artificiale è che questi ultimi "imparano" a filtrare i dati guardando molti esempi, senza regole precostituite. Così come i bambini imparano a riconoscere gli oggetti o gli odori.

Per tornare all'analogia del lago, i "data scientists" sono proprio i professionisti che "pescano" informazioni dal lago di dati. Ovviamente la qualità del pescato dipende fortemente dalla qualità dei dati che arrivano nel lago in termini di pulizia, completezza, copertura, veracità e anche dalla abilità dei data scientists nell'individuare strumenti di pesca (cioè di analisi) sufficientemente selettivi per uno specifico obiettivo. Sono professionisti con un forte background di statistica e informatica in grado di valutare per ogni singolo caso se e quali metodi utilizzare scegliendo in un insieme di strumenti che va da quelli statistici tradizionali (Business Intelligence) a quelli dell'Intelligenza Artificiale. Chiaramente non esiste un metodo buono per tutti gli scopi, la scelta è complessa e per questo i data scientists sono oggi fra i professionisti più richiesti e pagati del mercato professionale tecnologico. L'università di Torino ha da qualche anno attivato un corso di dottorato innovativo in data science supportato da diversi dipartimenti, fra cui il dipartimento di informatica, matematica e economia.

3. *Datasilos e Datalake nella pubblica amministrazione*

Parrebbe quindi che il paradigma datalake sia la soluzione adatta per l'organizzazione dei dati della Pubblica Amministrazione. Purtroppo la risposta non è così semplice. Un datalake è tecnicamente un sistema OLAP (On Line Analytics Processing) che è adatto a cercare nuove relazioni fra i dati ma è totalmente inadeguato a mantenere la consistenza e la correttezza dei dati nel tempo per cui i tradizionali datasilos (tecnicamente detti OLTP - On Line Transaction Processing) basati su schemi sono ancora l'unica soluzione. Di fatto oggi è necessario utilizzare insieme datasilos e datalake. Per questo il problema relativo alla costruzione di sistemi di immagazzinamento e analisi che siano veramente GDPR compliant va oltre il singolo database o datacenter ma richiede una visione che abbraccia una rete di sistemi connessi e cooperanti. Da questo punto di vista, le attuali procedure di certificazione di sicurezza o di rispetto della GDPR penso assolvano più alla necessità di creare opportunità di business per le società di consulenza che altro. Faccio un paio di esempi relativi alla GDPR.

Il primo esempio riguarda lo scopo per cui i dati vengono raccolti, che in alcuni casi deve essere esplicitamente dichiarato (come per i dati

sanitari). La frontiera della ricerca medica si muove su un territorio sconosciuto e cerca relazioni ignote cercando correlazioni fra diversi fonti di conoscenza: cartelle cliniche, immagini TAC, DNA/RNA. Il possibile uso della scoperte è spesso non definito a priori. Vogliamo contare i prodotti che oggi funzionano bene o rispondono a uno scopo completamente diverso da quello per cui erano stati pensati?

Alla fine dell'Ottocento, ad Atlanta, il farmacista John Stith Pemberton, riprendendo la formula del "Vin Mariani", una miscela di vino e foglie di coca creata dal farmacista corso Angelo Mariani, inventò la Coca-Cola: un elisir contro mal di testa e stanchezza. Sempre alla fine dell'Ottocento, all'Ospedale Saint Mary di Londra, il chimico Alder Wright, ottiene, da una sintesi chimica della morfina, l'eroina. Tuttavia la scoperta non è ritenuta interessante. Una ventina di anni dopo l'eroina è nuovamente sintetizzata dal dottor Felix Hoffmann, ricercatore della casa farmaceutica Bayer. La Bayer commercializza da subito l'eroina come medicinale per il trattamento della tosse, dei problemi respiratori e per combattere la dipendenza dalla morfina.

Gli esempi sono deliberatamente riferiti a un'epoca in cui la conoscenza non veniva estratta dai dati, come invece avviene per i BigData. Oggi la situazione è completamente diversa e la frontiera della ricerca medicina si usa tecniche di BigData e Intelligenza Artificiale e quindi dipende direttamente dalla GDPR e dalle leggi che disciplinano l'uso dei dati. A questo riguardo a Torino abbiamo da poco attivato un nuovo trial clinico che si svolge presso l'Ospedale San Luigi come collaborazione fra dipartimento di radiomica e informatica. È uno studio sul polmone finalizzato ad eliminare molte delle biopsie oggi in uso nella pratica clinica utilizzando tecniche di Intelligenza Artificiale per prevedere le varianti genetiche a partire da immagini TAC dei pazienti osservati. Uno studio assolutamente nuovo di cui neanche noi riusciamo a immaginare esattamente tutti i contorni, che potrebbero essere anche fuori dall'alveo dello scopo per cui i dati del trial clinico sono stati originariamente raccolti.

Un secondo esempio riguarda le reti di sistemi, che abbiamo visto essere davvero necessarie a mantenere e analizzare i dati. Per funzionare le reti devono copiare i dati sistemi datasilos/OLTP, finalizzati alla archiviazione (come quello di ANAC o le cartelle cliniche dell'ospedale) a sistemi datalake/OLAP, finalizzati all'analisi. Secondo la GDPR i dati possono essere vincolati all'uso all'interno di una finestra temporale e di un luogo geografico specifici. Quindi da un lato la copia dei dati da sistemi OLTP a OLAP è necessaria per l'efficacia dell'analisi, dall'altro ogni copia è un pericolo per l'integrità dei vincoli perché è tecnicamente difficilissimo se

non impossibile tracciare la vita di tutte le copie dei dati e renderle inaccessibili quando il dato originale arriva alla sua scadenza. E inoltre, cosa ne sarebbe dei processi di analisi che dipendono da quei dati? La catena di conseguenze è ad oggi ignota e personalmente non vedo nessun modo serio per utilizzare i dati per estrarre informazioni e rispettare alcune delle prescrizioni GDPR. In linea di principio l'insieme delle tecniche conosciute come "distributed ledger" (blockchain) potrebbero essere una soluzione, ma questo argomento è fuori dagli scopi di questa relazione. A questo proposito mi fa piacere ricordare che grazie alla Professoressa Racca, l'Università di Torino ha appena stipulato un accordo con ANAC e Team Digitale della Presidenza del Consiglio per poter accedere al database dei bandi pubblici e studiare possibili irregolarità all'interno degli stessi. A questo fine copieremo i dati ANAC nel sistema HPC4AI/OLAP.

4. *Le prospettive future*

Credo quindi di poter sostenere che il futuro prossimo sia ricco di sfide per gli informatici, per i giuristi, ma anche e sopratutto per i team multidisciplinari. Ricordo una scuola estiva, nel 2003, in cui ero giovane dottorando, sul tema del fraud detection mediante tecniche di data mining. Era già allora tecnicamente possibile, dal punto di vista strettamente algoritmico, cercare correlazioni per esempio fra stile di vita e dichiarazione dei redditi ma non c'era la capacità tecnica di organizzare i dati in datalake per poterli incrociare facilmente e non c'era neanche forse la volontà di farlo. Volontà che può arrivare solo attraverso la profonda comprensione degli aspetti giuridici ma anche organizzativi e tecnici da parte dei team a supporto dei legislatori e dei manager della Pubblica Amministrazione. Purtroppo questa è solo una condizione necessaria ma non sufficiente alla modernizzazione del sistema paese che rimane sempre appeso alle buone intenzioni senza riuscire ad andare oltre.

Negli ultimi anni tuttavia questa volontà pare emergere. Guardiamo con interesse al Team Digitale della Presidenza del Consiglio che aveva lo scopo dichiarato di ridurre il numero di datacenter della Pubblica Amministrazione da molte migliaia a una decina di datacenter di importanza nazionale, centralizzando alcuni importanti database, come per esempio l'anagrafe. E proprio l'anagrafe e l'esperienza della città di Torino nella integrazione dalla anagrafe locale (gestita da CSI) a nazionale necessaria per la transizione alla carta di identità elettronica ci suggerisce che i processi di consolidamento sono necessari ma complessi sia tecnicamente

che politicamente. Il consolidamento dei datacenter di fatto aggredisce piccole o grandi rendite di posizione che negli anni si sono incrostate e l'energia e la competenza necessaria per rimuoverle è spesso molto più grande di quanto possa inizialmente apparire.

E a questo proposito, secondo me, un solo team radicato a Roma è un approccio al problema "donchisciottesco". Specialmente perché il lavoro necessita specifiche competenze e diffuse collaborazioni su tutto il territorio nazionale.

Sulle competenze l'America della Silicon Valley ha costruito il proprio monopolio. Per citare Mario Tchou che a metà degli anni cinquanta era il giovane ingegnere responsabile dei progetti più innovativi di Olivetti: "Perché le cose nuove si fanno solo con i giovani. Solo i giovani ci si buttano dentro con entusiasmo, e collaborano in armonia senza personalismi e senza gli ostacoli derivanti da una mentalità consuetudinaria". Tchou aveva guidato il gruppo di lavoro per progettare il primo calcolatore interamente italiano: la Calcolatrice Elettronica Pisana (CEP). Adriano Olivetti fu fra i primi al mondo a investire sulle competenze informatiche; lo fece a Pisa perché l'Università aveva accumulato un ingente capitale per sviluppare un acceleratore di particelle e fu Enrico Fermi a suggerire al rettore dell'Università di Pisa di usare i soldi per fare qualcosa di nuovo: una calcolatrice elettronica. Un'esperienza da cui poi nacque nel 1969 a Pisa il primo dipartimento di Informatica italiano (dove anche io ho studiato). Queste stesse competenze portarono il gruppo di Tchou a sviluppare la macchina "Elea 9003", il primo computer della storia interamente realizzato con componenti a stato solido (senza valvole) e che di fatto ha aperto la frontiera della miniaturizzazione e della velocità dei calcolatori elettronici.

In ambito scientifico, sui temi come BigData e Intelligenza Artificiale noi informatici dobbiamo essere estremamente attenti a non diventare esclusivamente dei selezionatori di giovani talenti da spedire in America. Per questo voglio ricordare HCP4AI, un nuovo centro di competenza sui temi del calcolo ad alte prestazioni, del BigData e l'Intelligenza Artificiale che l'Università e il Politecnico di Torino (sotto il mio coordinamento). HPC4AI è stato recentemente finanziato con 4.5M€ (50% Regione Piemonte, 50% Università e Politecnico) con l'obiettivo di creare un grande laboratorio federato dove poter sperimentare nuovi sistemi e tecniche per l'analisi dei dati in diversi ambiti, dalla biologia alla medicina alla giurisprudenza. Il recente accordo fra Università di Torino e ANAC per l'analisi dei bandi pubblici si appoggia a questa infrastruttura di ricerca.

Da questo scopo primario derivano diversi altri obiettivi, fra i quali:

- Sviluppare e mantenere le competenze su questi temi. Perché le competenze sono la base della catena del valore del trasferimento tecnologico e dell'innovazione.
- Sviluppare gli strumenti necessari alla transizione digitale e l'analisi dei dati. Perché se anche si compisse la transizione digitale della Pubblica Amministrazione ma questa avvenisse utilizzando esclusivamente tecnologia di proprietà di multinazionali ci sarebbe poco da gioire. Il funzionamento del nostro paese, dall'anagrafe alla raccolta delle tasse, dipenderebbe in modo esclusivo da società estere. Un lock-in tecnologico che potrebbe essere molto costoso e lesivo della sovranità nazionale (Italiana e Europea).
- Supportare la transizione al digitale della Pubblica Amministrazione e il consolidamento delle banche dati sono nuovi processi, sviluppati e testati localmente e poi integrati su scala nazionale. Perché l'idea che un solo team, a Roma, da solo e contro tutti possa davvero digitalizzare tutta l'Italia è piuttosto ingenua. Anche senza considerare gli aspetti politici, questo team avrebbe bisogno di centinaia o migliaia di persone (ne ha una trentina adesso) che dovrebbero integrare sistemi regionali e comunali sparsi in tutto il territorio, generare una quantità di processi e proprietà intellettuale che oggi solo i centri partecipati delle università possono garantire. In Piemonte, oltre HPC4AI che è specificamente caratterizzato sulla ricerca scientifica, si può menzionare il CSI che oggi gestisce la gran parte dei dati regionali e comunali e che - almeno secondo me - dovrebbe essere una componente di un sistema federato nazionale per la Pubblica Amministrazione.