BIG DATA & OPEN DATA

Whitehall Reply guida i propri Clienti nel percorso di pubblicazione dei dati, portandoli al massimo livello di interoperabilità, attraverso la loro trasformazione in formato “Linked Open Data”. Sulla base della nostra esperienza, l’approccio progettuale si basa su una metodologia iterativa strutturata per il contesto degli Open Data attraverso: Data Sources Identification; Data Analysis & Quality Review; Data Transformation & Dataset creation; Dataset Transfer.

BIG DATA & OPEN DATA

Nell'ultimo decennio la diffusione di nuove tecnologie e di device mobili, ha prodotto una maggiore interconnessione ed una conseguente crescita esponenziale di dati da gestire. Il fenomeno dei cosiddetti **Big Data** è in continua evoluzione grazie al propagarsi di tecnologie e processi per la loro gestione.

La crescita dei Big Data è sempre maggiore, soprattutto grazie all'evoluzione delle tecnologie e dei processi per la loro gestione.

Oggi, trattare dati richiede nuovi strumenti, metodologie e algoritmi attraverso cui svolgere analisi ed elaborazioni cruciali per migliorare servizi, risultati, processi e decisioni.

I dati devono essere gestiti al meglio, allo scopo di ampliare prospettive di business e relazioni. Inoltre è necessario fornire, nel settore pubblico e privato, servizi sempre più efficienti anche e soprattutto in relazione alle esigenze e alle specifiche esperienze degli utenti finali.

Per sfruttare al meglio tali nuove opportunità, tuttavia, è necessario limitare le restrizioni imposte sulle modalità di trattamento dei dati attraverso l'utilizzo degli **Open Data**, così da aumentare la pletora di soggetti che possono riutilizzare i dati per produrre valore economico ed

efficienza, ma anche per promuovere i principi basilari dell'Open Government, come trasparenza, partecipazione e collaborazione.

Per Whitehall Reply questa è la direzione da seguire, affinché i dati divengano anche in Italia un driver di opportunità per costruire una rete di conoscenza basata sui dati condivisi.

BIG DATA

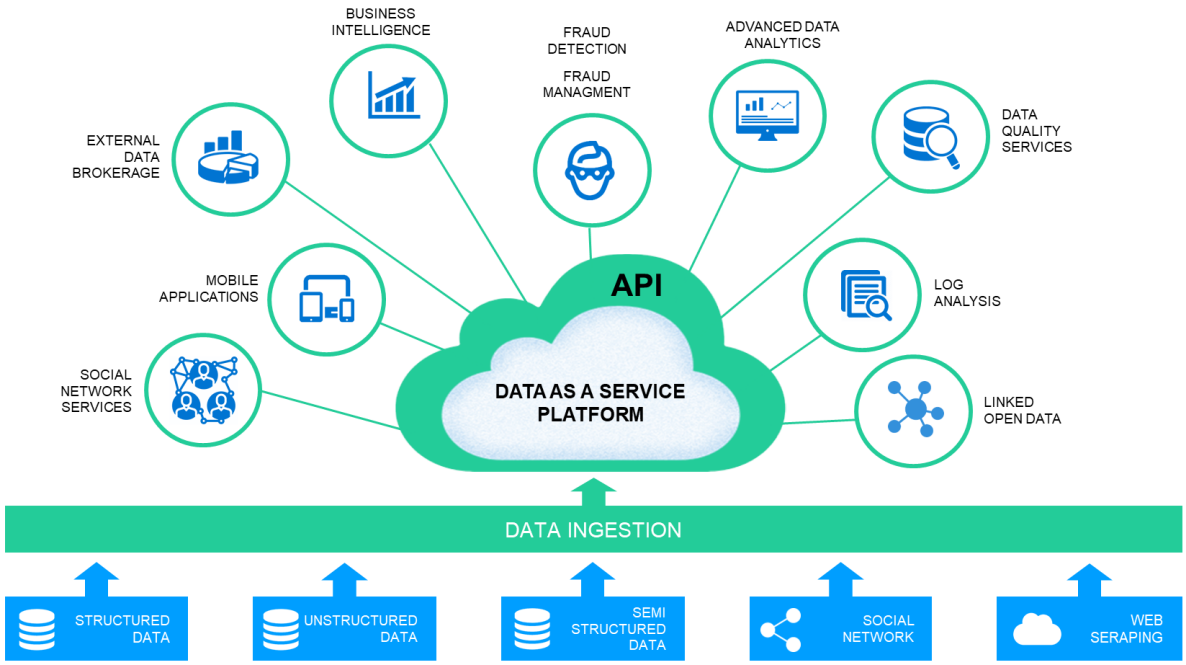
Con il termine **Big Data** si identifica un'enorme mole di dati in termini di **volume, varietà** e **velocità** di produzione che richiede tecnologie e metodologie differenti da quelle tradizionali allo scopo di processare, estrarre informazioni, analizzare e mettere in relazione una gigantesca quantità di dati eterogenei in maniera più efficiente e rapida possibile, che siano essi strutturati, semi-strutturati o non strutturati. Altra caratteristica molto importante è l'**attendibilità** dei dati che vengono lavorati o prodotti allo scopo di non generare report o prendere decisioni che in futuro potrebbero rivelarsi errate.

Le tecnologie *Big Data* permettono di superare numerosi limiti in diversi ambiti. Ad esempio, è difficile scalare i database tradizionali per la gestione di una quantità ingente di dati in continua crescita. Quello che si fa, solitamente, è scalare verticalmente, aumentando le risorse di calcolo. Una modalità che, però, risulta essere molto onerosa sia in termini di costi che di risultati: dopo una serie di analisi, infatti, si arriva ad una "cima" che non è più possibile scalare verticalmente. La soluzione che si propone nelle tecnologie *Big Data* sono sistemi distribuiti che hanno la capacità di scalare orizzontalmente. La scalabilità orizzontale consiste nell'aggiunta, quando necessario, di nodi all'architettura insieme a quelli già esistenti. Le tecnologie che vengono utilizzate, come *Hadoop*, permettono queste operazioni in maniera rapida e trasparente.

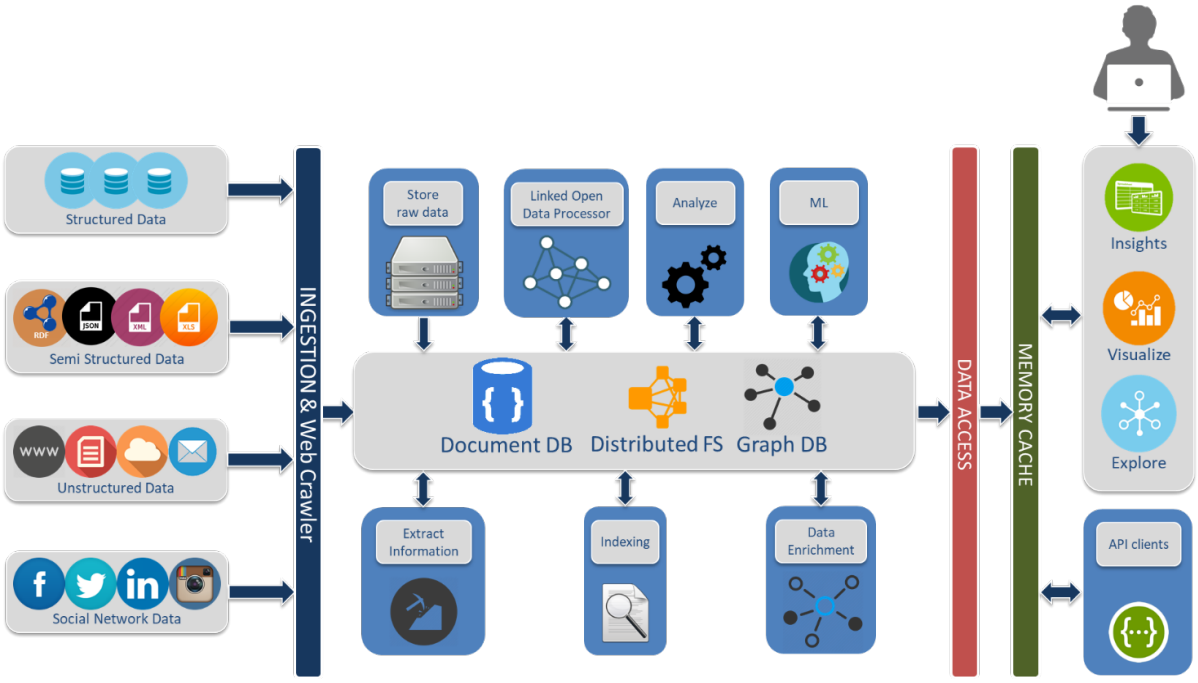
Un altro limite è la gestione dei dati

eterogenei e non strutturati, molto difficili da gestire con un database relazionale: anche in questo caso ci vengono in aiuto le tecnologie *Big Data* e *NoSQL* che permettono di avere database schemaless. in una grande mole di dati. Ma gestire la navigazione di relazioni tra diverse entità, con tecnologie tradizionali, risulta essere troppo oneroso. Per questa ragione, sono nati database a grafo che permettono di effettuare tale navigazione grazie al loro modello dati.

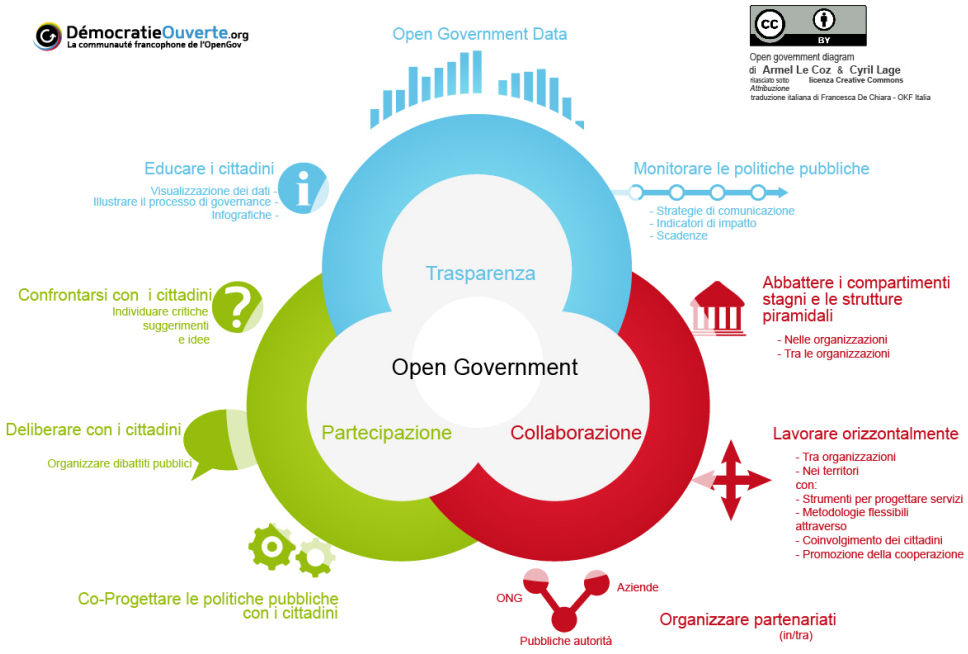
Altra particolarità fondamentale dei Big Data è il paradigma dello schema "on read" rispetto allo schema "on write". Quest'ultimo consiste nel creare una definizione dello schema dei dati direttamente sul database o data warehouse che si utilizza, per questo, diverse operazioni risultano pesanti. Un esempio, ci potrebbe essere fornito dalla modifica dello schema in un secondo momento, oppure dalla definizione di trasformazioni quanto più complesse per adattare i dati di input al database. Al contrario lo schema "on read" consiste nel definire lo schema in lettura del dato precedentemente importato. Ciò permette di importare i dati RAW e utilizzare lo schema che meglio si crede a seconda dell'analisi che si deve effettuare, senza preoccuparsi troppo degli aggiornamenti dello schema (ad esempio, aggiungere più colonne). Whitehall Reply ha una strategia ben definita sulla gestione dei dati e punta ad importare e gestire numerosi dati eterogenei da moltissime fonti dati,



fruendoli successivamente tramite API. Questo ha portato alla creazione di una vasta piattaforma Data As A Service. I dati esposti vengono o possono essere utilizzati per numerosi casi d'uso come ad esempio su device mobile, creazione di advanced data analytics, servizi di data quality, applicazione di algoritmi di ML (come ad esempio la fraud detection) e così via. La piattaforma è stata realizzata dopo anni di studio e lavoro tramite la creazione di un framework, sviluppato interamente da Whitehall Reply.
Il framework architetturale sviluppato è una soluzione, interamente basata sulle tecnologie Open Source più innovative in ambito Big Data e annoverate come "leader" nei report di autorevoli società di market e tech research come *Gartner.* Il framework, interamente modulare, è stato sviluppato sfruttando le potenzialità del file System distribuito Hadoop, e dei database NoSQL MongoDB e Neo4J, standard de-facto per lo sviluppo di applicazioni in ambito Big Data. Tale soluzione consente di memorizzare e gestire differenti tipologie di dati (interni ed esterni) in un unico punto centralizzato ovvero il **Data Lake**. Inoltre, il framework fornisce agli utenti le più avanzate tecniche di **Data Discovery** and **Analytics** mediante la navigazione di Applicazioni Web dinamiche sviluppate con l'utilizzo di framework evoluti di **Data Visulization** o mediante l'esposizione di API seguendo lo standard **OpenAPI.**



OPEN DATA



Gli Open Data sono dati liberamente accessibili, utilizzabili, modificabili e ricondivisibili da ogni cittadino, con eventuali deboli restrizioni, come dover citare la fonte e dover mantenere il dato aperto. Il concetto di **Open Government**, alla base degli Open Data, spinge per aprire quanto più possibile i dati pubblici, allo scopo di:

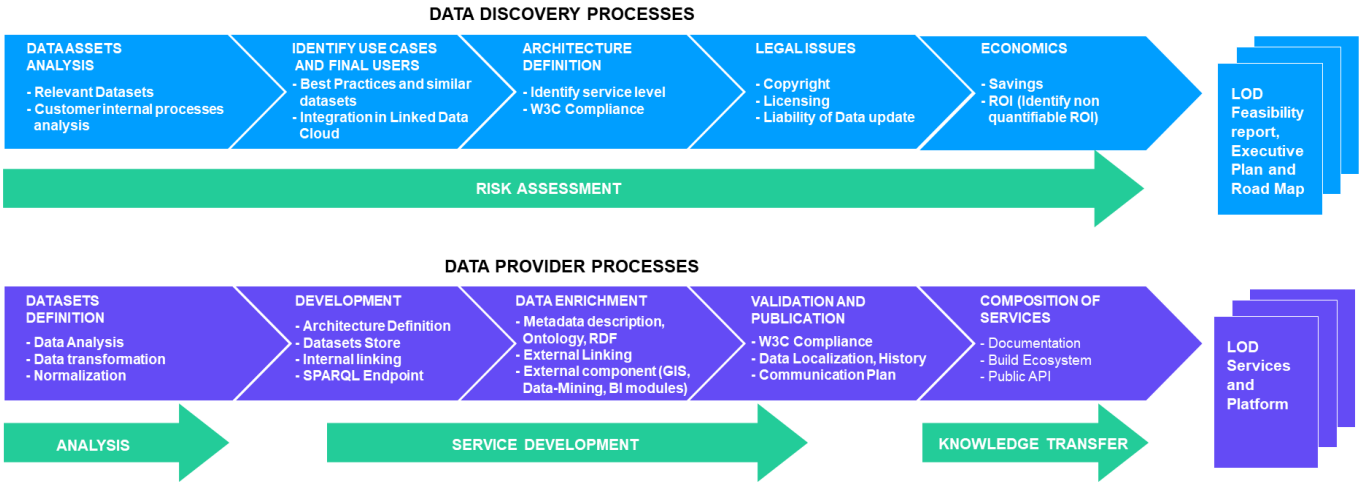
- rendere l'amministrazione trasparente;
- incentivare la partecipazione del cittadino alla vita politica;
- creare una rete collaborativa e partecipata. Il tema relativo alla pubblicazione dei dati da parte della pubblica amministrazione è, ormai da diversi anni, oggetto di numerosi direttive (ad esempio il Decreto Trasparenza del 2013). Tali orientamenti sono in continua evoluzione: in un primo momento si limitavano all'obbligo di pubblicazione di dati da parte della Pubblica Amministrazione, oggi mirano a fornire indicazioni precise riguardo il formato di pubblicazione, in modo da poter garantire un livello di interoperabilità sempre maggiore. I dati devono quindi essere pubblicati strutturati, in formato aperto e con opportuna metadatazione e semantica, secondo i principi del Web 3.0 (in particolare del **Web Semantico**).

L'**AgID** (Agenzia per l'Italia Digitale) è molto attiva su questo tema e ha fornito diverse linee guida per seguire la Pubblica Amministrazione in questo processo di pubblicazione dei dati. L'obiettivo di Whitehall Reply è quello di guidare i propri Clienti (in particolare la Pubblica Amministrazione) nel percorso di pubblicazione dei dati, portandoli al massimo livello di interoperabilità, attraverso la loro trasformazione in formato Linked Open Data, e permettendo ai clienti stessi di diventare parte del database globale formato dai Linked Open Data già pubblicati.

METODOLOGIA ODW (OPEN DATA WHITEHALL)

Whitehall Reply garantisce la trasformazione dell'innovazione dalla fase di sperimentazione a quella di mercato, assicurando la valorizzazione di idee ed iniziative migliori grazie ad un approccio governato e strutturato. Tutti i progetti sono accomunati dalla metodologia ODW (Open Data Whitehall) applicata per la modellazione, produzione e pubblicazione di Open Data in diversi formati, con un focus su Open Data di livello cinque (Linked Open Data). La metodologia ideata da Whitehall Reply, sulla base delle competenze acquisite, è rappresentata nella figura in calce.
ODW si caratterizza per un forte processo di analisi dei dati e dell'architettura, conforme agli standard internazionali del **W3C**, che meglio si presta per il rilascio dei dati. In questa fase è svolta un'attenta analisi della domanda dei dati da parte degli utenti e dei requisiti normativi per l'apertura. Dall'altro lato, partendo dai risultati dell'analisi, la metodologia si connota per lo sviluppo di una soluzione completa e sostenibile che consenta di produrre dati di tipo aperto in ogni formato neutro dal punto di vista tecnologico, con particolare attenzione al modello **RDF** e dei Linked Open Data. Il nostro approccio progettuale si basa su una metodologia iterativa strutturata, sulla base della nostra esperienza, proprio per il contesto degli Open Data:

- Data Sources Identification;
- Data Analysis & Quality Review;
- Data Transformation & Dataset creation;
- Dataset Transfer



www.reply.com/whitehall-reply/it/