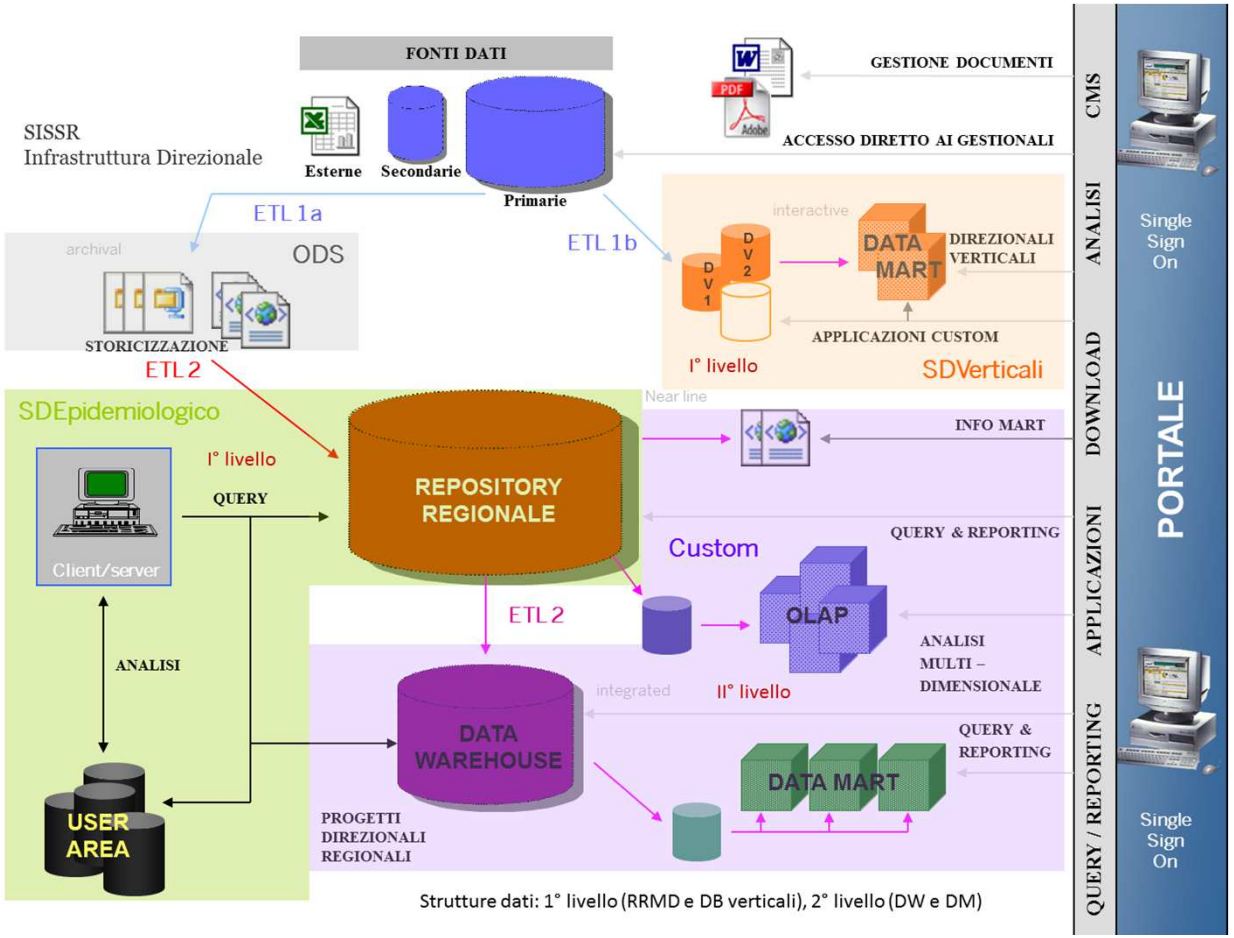
Uso delle fonti di **dati sanitari** correnti **per finalità epidemiologiche**

Il percorso dei dati all'interno di un sistema integrato

Pierantonio Romor

ISS 3-5 Aprile 2013

Oggetto: Il percorso dei dati all'interno di un sistema integrato



Serve una infrastruttura

Uso delle fonti di **dati sanitari** correnti per finalità epidemiologiche

Il percorso dei dati all'interno di un sistema integrato

Introduzione: i dati e il sistema

Laboratorio dei Sistemi di Babele

pierantonio.romor@insiel.it

Sistema informatici

Gestiscono dati (?) in forma nativa mediante funzioni di:

- inserimento,
- modifica,
- cancellazione.

Forniscono informazioni (?) mediante funzioni di:

- interrogazione.

Ma dove finiscono?

Sanità

Cloud & big data

Una **realtà** :

1. Complessa in cui si predilige l'interrogazione.
2. Dispone di soluzioni HW (cpu e storage) performanti.
3. Interessa principalmente i dati non strutturati (web e social network).
4. Non coinvolge,attualmente, i sistemi «operazionali» sanitari (sistemi chiusi)
5. Anche se…esiste una tendenza ad utilizzare tali soluzioni (portali dei servizi per il cittadino e le «business analytics») per condividere e fornire «rapidamente» e senza «conoscenze» le informazioni ai destinatari.

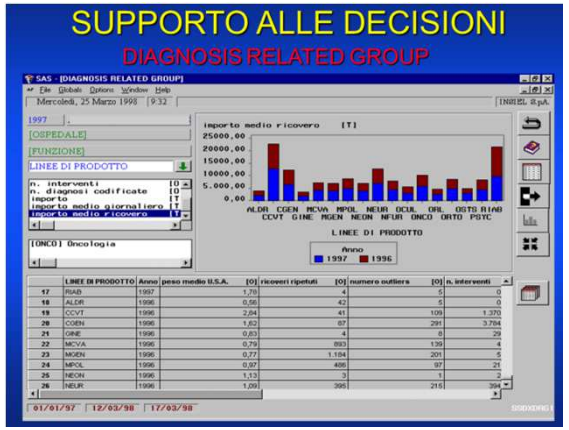
La risposta di ieri (?) dei sistemi informatici alle richieste di dati

Si utilizzavano tecniche non organizzate dai sistemi operazionali:

- **Reportistiche**
  - ❖On line
  - ❖Batch
- **Flussi** ad hoc trattati successivamente con software di «office automation»

Si utilizzavano interfacce «custom» simili ai sistemi transazionali:

- **Decision Support Sistem** (1990)



L'evoluzione: prevedere ambienti distinti

OLTP Systems are used to *"run"* a business

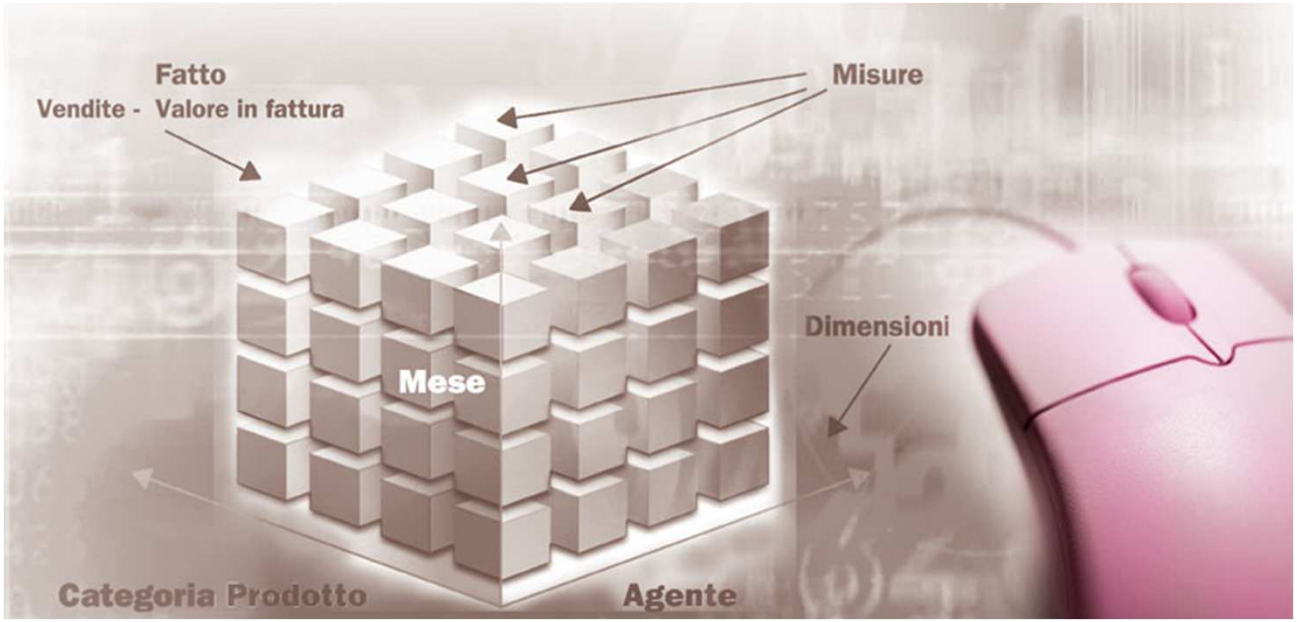
≠

The OLAP Systems helps to *"optimize"* the business

I data warehouse

1990 Bill Inmon pubblica «Building the DW»

- In OLAP il primo nato è il cubo con le sue declinazioni (data mart) e caratteristiche:

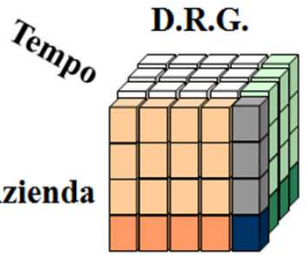


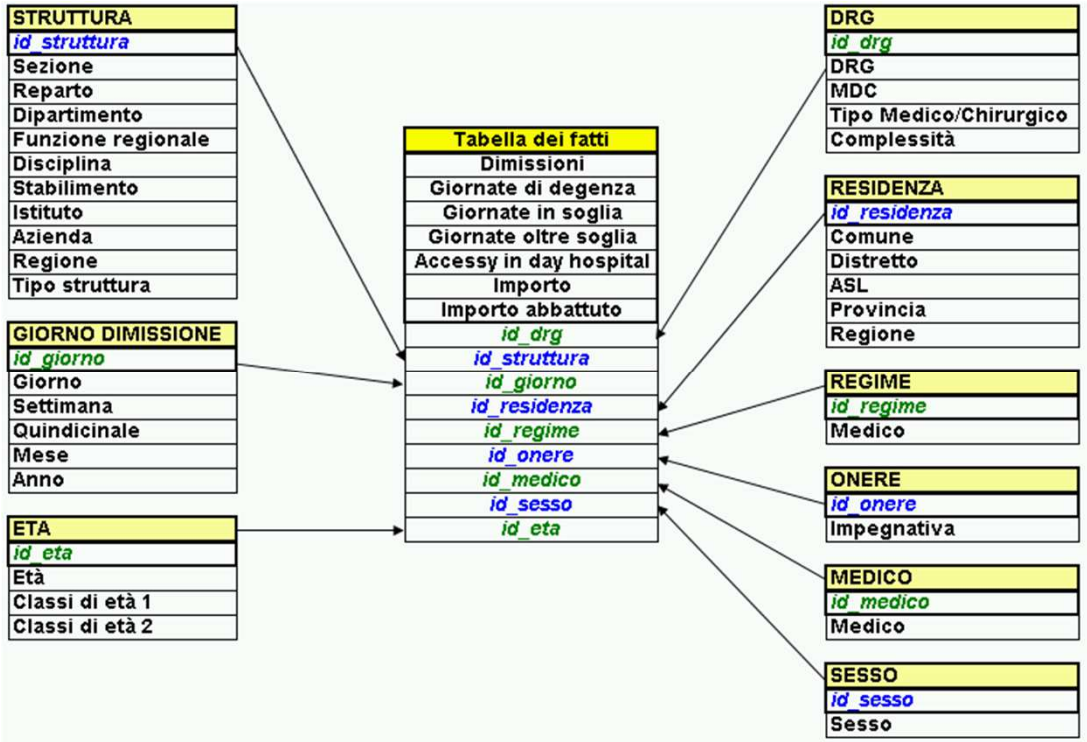
…..su sistemi amministrativo-contabili

Traslato in sanità

FUNZIONI

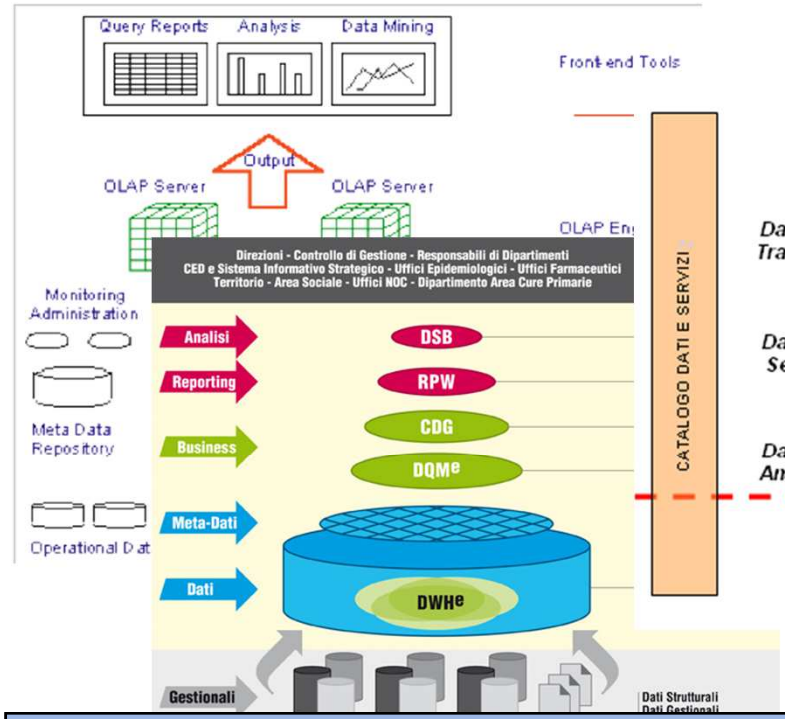
- Pivot
  - Orienta la dimensione
- Roll-up &Drill-down
  - Navigazione nella cella
- Slice & Dice
  - Navigare nella dimensione

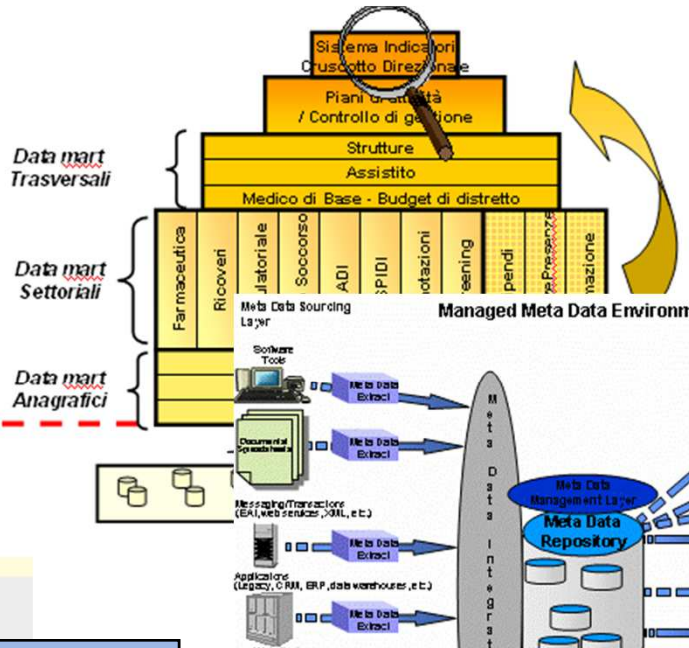


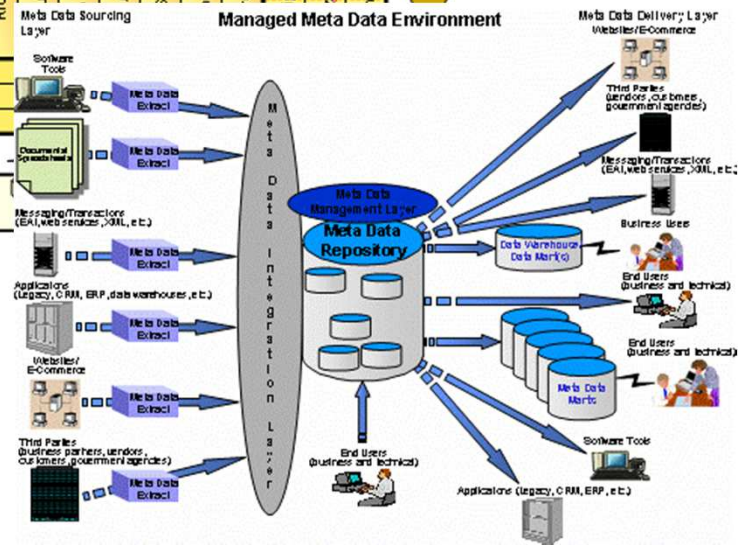


Il CUBO è nato come proposta tecnologica per superare la staticità dei report (uso amministrativo).

Dai cubi…ad un sistema:
Business Intelligence







Ogni realtà ha il suo modello
di B.I.(«Babele» Intelligence)

La complessità è di sistema

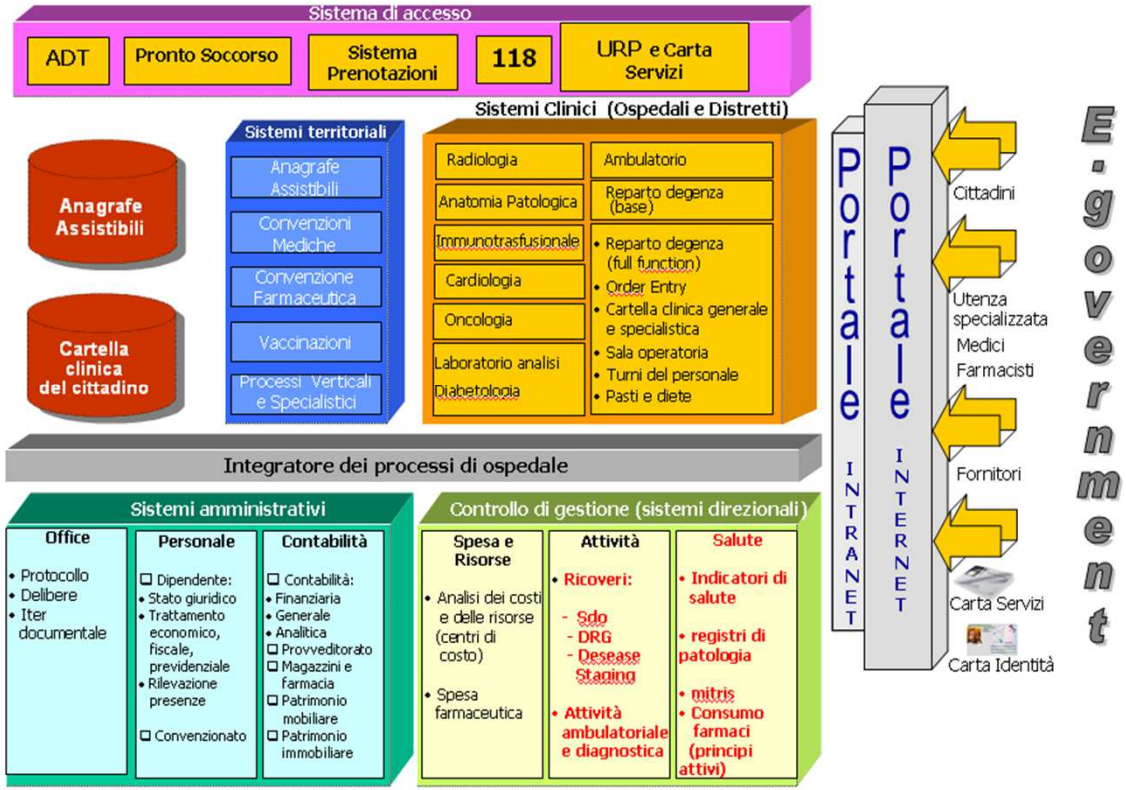
Sistemi operazionali (OLTP)

- Sistema contabile (COAN,COGE,COFI,.)
- Sistema amministrativo (es. Cardionet)
- Sistema clinico (es. Cardionet)

Sistemi direzionali

(OLAP, BSC, GIS, Analytics,…)

- ✓ Controllo di gestione
- ✓ Epidemiologia
- ✓ Monitoraggio
- ✓ Pianificazione



Utilizzare il S.I.S.

Il sistema socio-sanitario è costituito da un insieme complesso di applicazioni e per utilizzarlo richiede conoscenze su:

- Organizzazione del processo (dipendente dalla singola organizzazione)
- Dominio applicativo (competenze su diversi livelli)
- Accessibilità ai dati (scarichi, viste, estrazioni)
- Omogeneità informativa (es. S.I. Clinici, Laboratori)
- Storicizzazione (disponibilità in linea)
- Ridondanza (minimum data set)
- Adozione Sistemi di classificazioni (uniformità e completezza)
- Presenza di dati semi-strutturati o destrutturati (es. referti)

Principali attori del processo di **trasformazione** del DATO in INFORMAZIONE

- **Operatore**
  - generatore del dato, responsabile della qualità
- **Tecnico sistema informatico operazionale**
  - «Ricevitore» analogico -> digitale, responsabile automatizzazione requisiti utente
  - manutentore del sistema, custode del d… …el valore semantico (relazioni)
- **Mediatori e Trasformatori**
  - Tecnico di data wareh…
  - Analista e data … …ser (epidemiologo e statistico)
- **Fruitore fin…**
  - In… …te (pianificatore / controller)

Processo di trasformazione **alla base dei sistemi di data warehousing**

SISTEMI OPERAZIONALI

SISTEMI DIREZIONALI

La NON soluzione ….

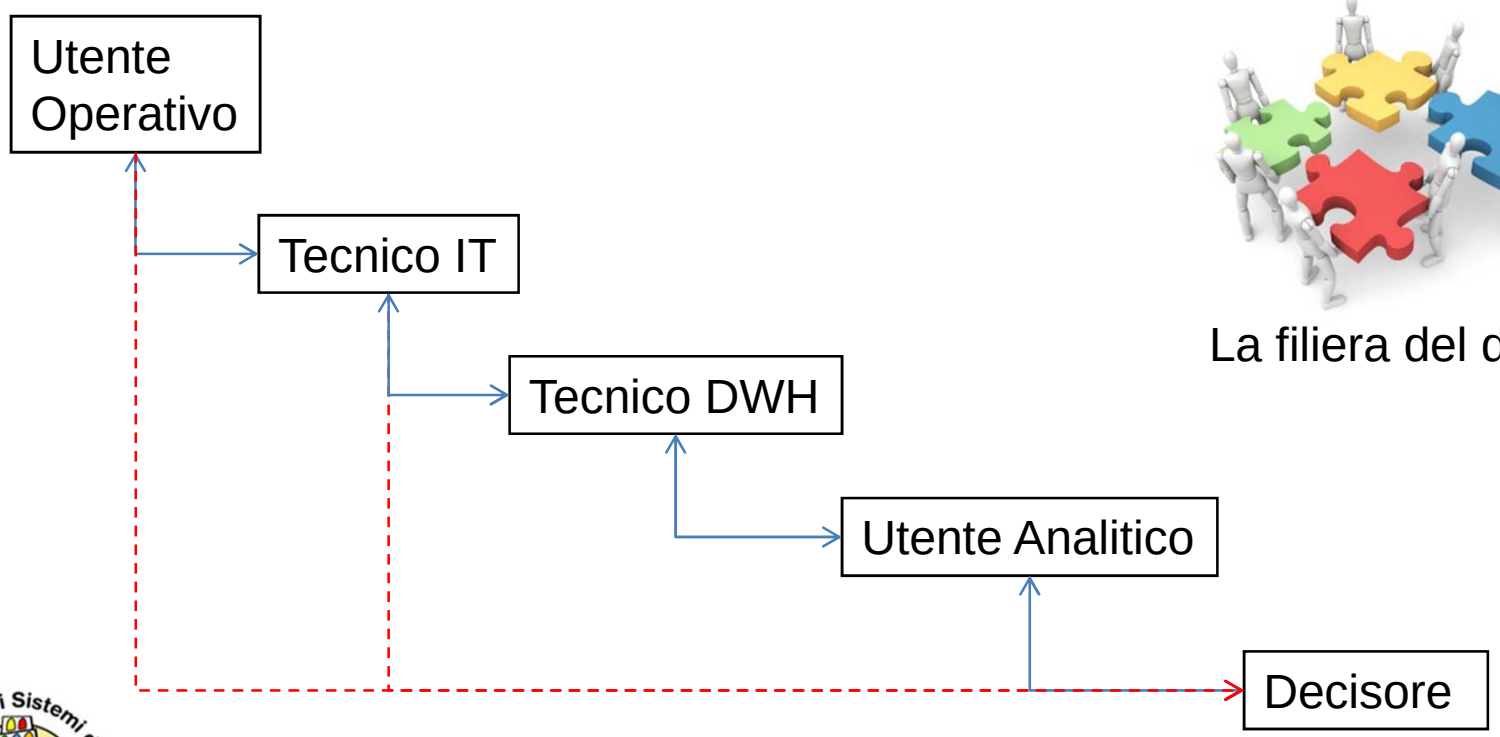
Mi scusi, mi sa dire dove sono?

Certo, lei è su una mongolfiera a 9 metri di altezza da me.

La soluzione: **il progetto** :

PER RILASCIARE INFORMAZIONI CONDIVISE

Il processo di trasformazione dei dati in informazioni significative NON può risolversi in una richiesta di dati, ma si declina su diversi ruoli.



La filiera del dato

Gli stakeholder del dato sanitario

Generano e modificano i dati

(sistemi transazionali)

Clinici

Enti di Ricerca

Controllo di Gestione

Epidemiologi

Decisori

Istituzioni

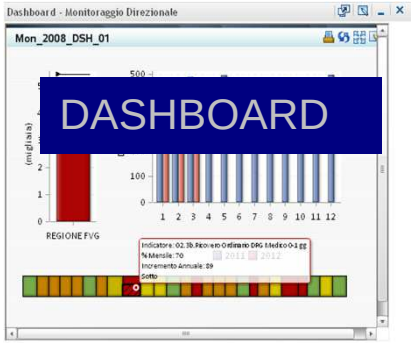
Il percorso dei dati

Il percorso di cura

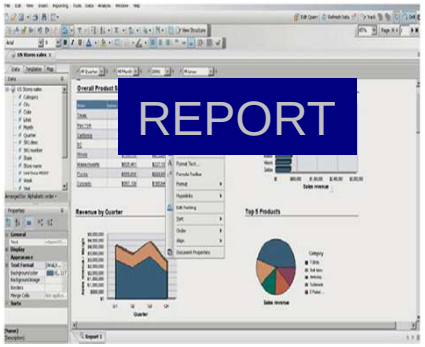
Leggono (?) ed interpretano il dato

(con strumenti informatici diversi)

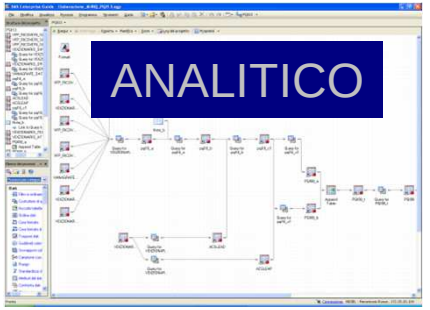
Necessario progettare una infrastruttura comune per utilizzatori diversi

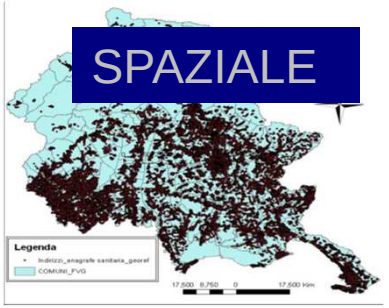


- L'utilizzo del dato ha sfumature diverse in base al contesto applicativo
- Fondamentale mettere in comunicazione gli attori del sistema



- **E' necessaria un'architettura adeguata: modello integrato distribuito e manutenibile.**





Laboratorio dei Sistemi di Babele

Uso delle fonti di **dati sanitari** correnti per finalità epidemiologiche

Il percorso dei dati all'interno di un sistema integrato

I sistemi integrati

pierantonio.romor@insiel.it

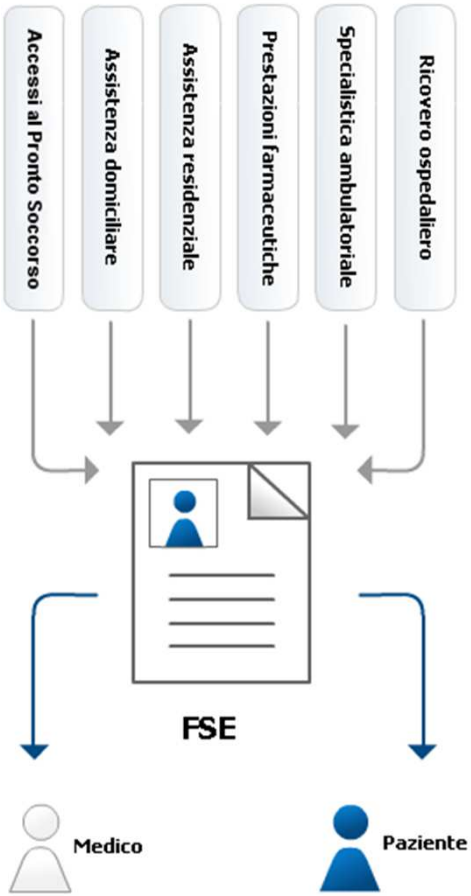
Architetture per il S.I. integrato

- **Virtuale:** viene definita solo la meta conoscenza necessaria a ottenere le informazioni sullo schema globale. Queste saranno create solo quando richieste mediante interrogazioni eseguite sugli schemi locali. Questa soluzione è quella maggiormente utilizzata nei sistemi OLTP.

- **Materializzato:** i dati vengono trasformati e memorizzati in versione duplicata. Questa soluzione viene utilizzata per esempio nei sistemi OLAP.

Un esempio di Sistema Integrato Sanitario Soluzione virtuale: Fascicolo e Dossier

- Definizione: l'insieme di dati e documenti digitali di tipo sanitario e sociosanitario generati dagli eventi clinici di ogni assistito, presenti e passati, che ha come scopo principale quello di agevolare l'assistenza al paziente anche quando lo stesso si affida alle cure di specialisti diversi.
- Obiettivo: sviluppare strumenti innovativi che, mediante l'utilizzo dell'informatica, possono assicurare una tempestiva disponibilità di informazioni ai diversi professionisti sanitari, allo scopo di rendere le cure più tempestive e di garantire la migliore **continuità assistenziale.**
- Requisito: essere una base informativa consistente (OLTP)



http://www.salute.gov.it/eHealth/paginaInternaEHealth.jsp

Caratteristiche principali FSE e DS

- Ha un orizzonte temporale che copre l'intera vita del paziente.
- E' alimentato in maniera continuativa dai soggetti che prendono in cura l'assistito nell'ambito dei servizi socio-sanitari.
- Rende disponibile la storia clinica del paziente a tutti gli attori coinvolti.
- Importante supporto all'emergenza/urgenza.
- Supporto per la continuità delle cure.
- Permette di condividere tra gli operatori le informazioni amministrative.
- Richiede il consenso dell'assistito

Il consenso

- A seguito dell'entrata in vigore della prima normativa sulla Privacy, è stato rilevato il consenso cosiddetto "generico", per tutti i trattamenti effettuati in maniera cosiddetta "tradizionale".
- Per il FSE e DS si rileva un ulteriore consenso che riferisce alla gestione degli stessi dati, ma informatizzati e messi a disposizione dei diversi professionisti sanitari (selezionabili) che possono così avere accesso alla storia clinica (tutta o in parte, per quali finalità).

Attualmente si chiede il consenso:

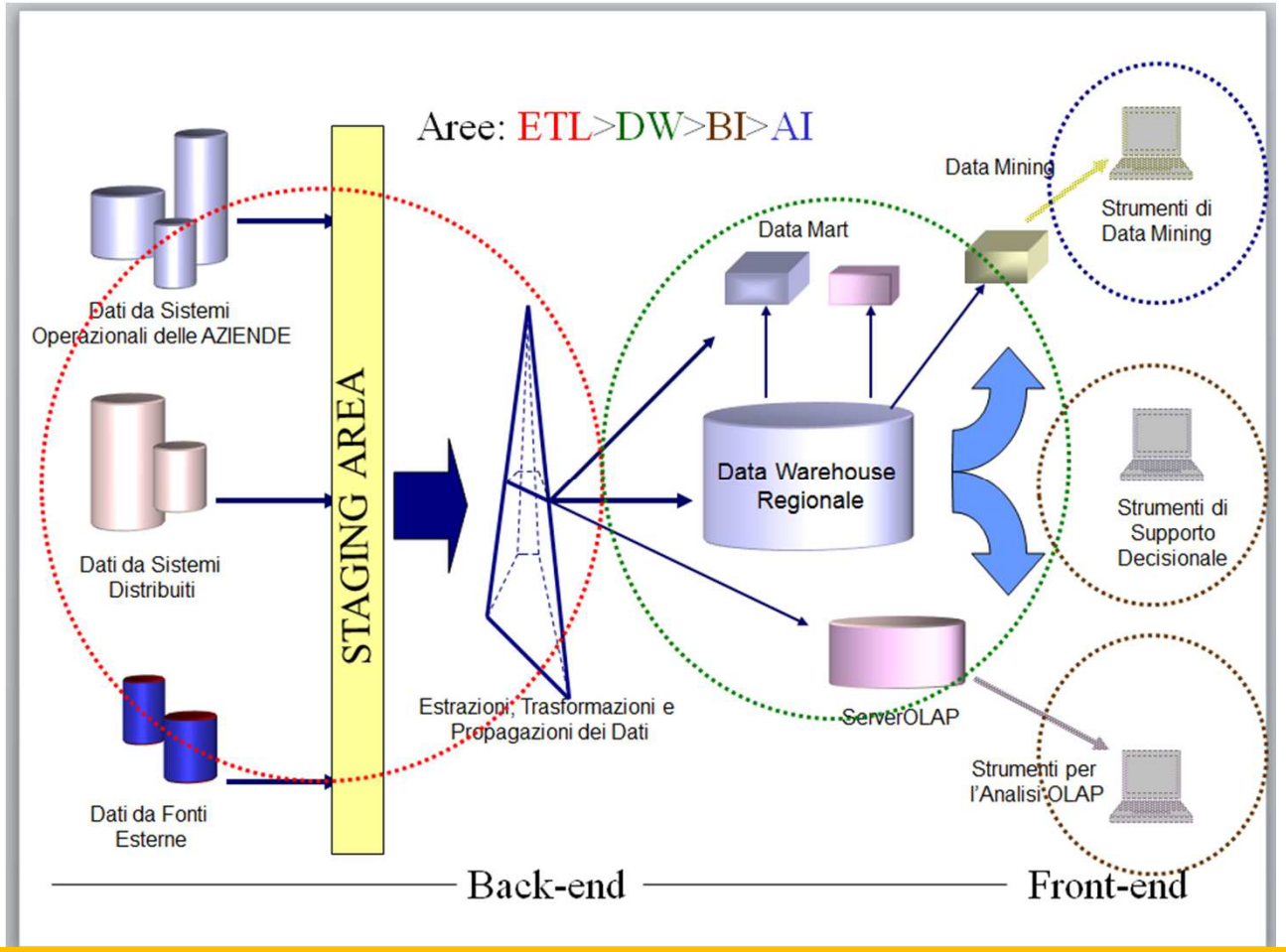
1. Alla gestione di base dei dati INFORMATIZZATI sensibili e personali ma non clinici (registrato in anagrafe)
2. Alla gestione di base dei dati nella struttura sanitaria (registrato in anagrafe)
3. Alla gestione dei dati INFORMATIZZATI (registrato in GECO):
   - nella struttura (con o senza pregresso)
   - al di fuori della struttura (con o senza pregresso), visibili a:
     - MMG e PLS;
     - Strutture SSN;
     - Strutture Sanitarie Private
   - nella ricetta elettronica
   - per scopi di ricerca clinica, epidemiologica e formazione

Caratteristiche della soluzione virtuale

- Livello integrato, accessibile, condiviso di fruibilità del dato al massimo dettaglio informativo.
- In consultazione.
- Tecnologia OLTP.
- Non arricchito, dato in formato originale.
- Finalizzato agli obiettivi di cura (da cui le informazioni selezionate).
- Focalizzato sul soggetto 
- Basato obbligatoriamente sul consenso.

Un modello generico di Sistema Integrato Sanitario
Soluzione materializzata: B.I. & A.I.

- ✓ Complesso
- ✓ Funzionalmente integrato
- ✓ Vendor dependent
- ✓ Multi user
- ✓ Pervasivo
- ✓ OLAP (Query & Reporting)
- ✓ Analitico
- ✓ Autoreferenziale



Sistemi chiusi che comunicano con l'esterno mediante flussi dati

Caratteristiche della soluzione materializzata

- Livello integrato, accessibile, condiviso a diversa granularità informativa
- In consultazione
- Tecnologia OLAP
- Arricchito da processi E.T.L.
- Finalizzato all'analisi
- Focalizzato su analisi aggregate
- Fruibile in maniera anonima
- Progettato per le analisi

Sistema «enterprise» che si appoggia a substrati tecnologici per la gestione delle profilature, a livello di:

- **Funzioni e ruoli**
- **Fonti dati**

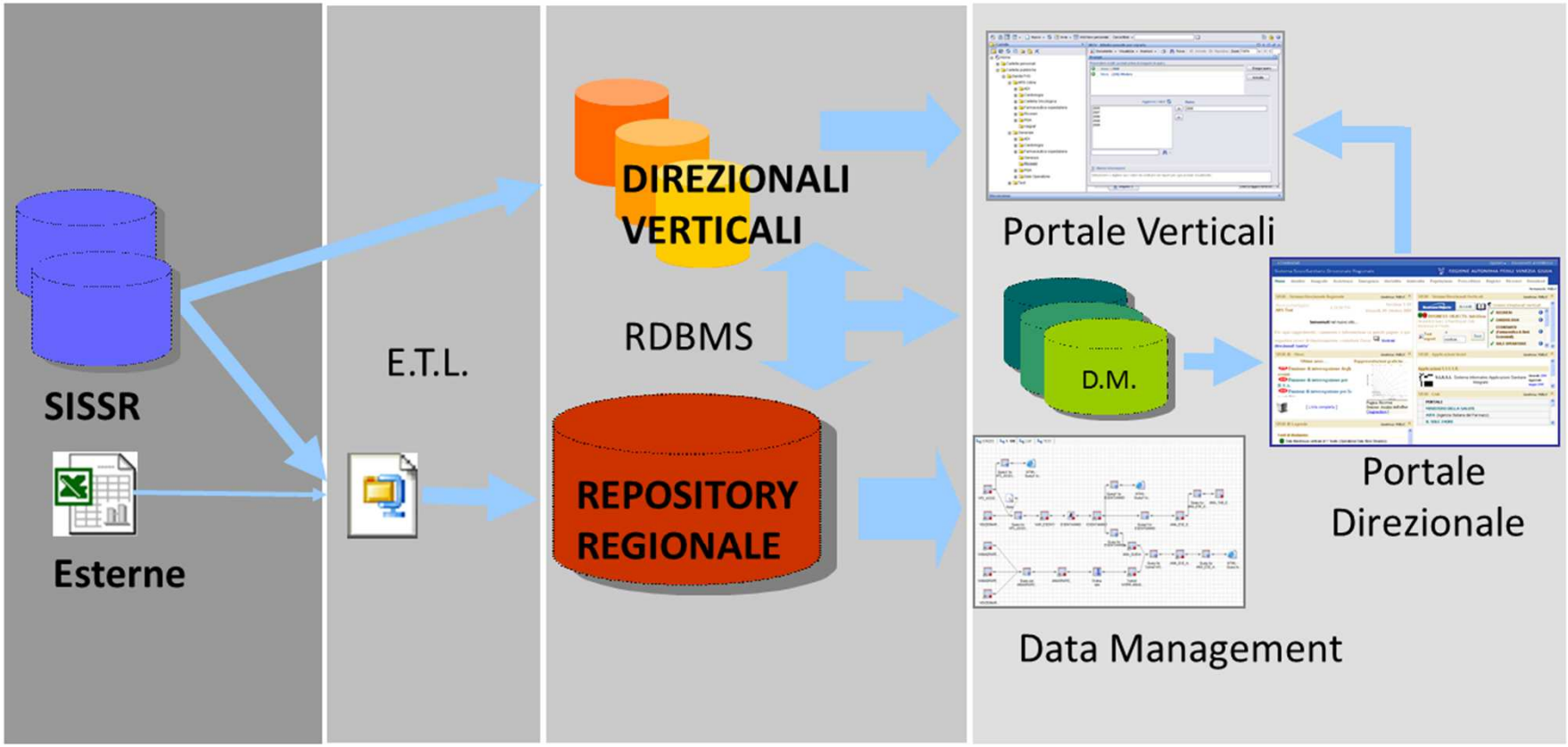
Uso delle fonti di dati sanitari correnti per finalità epidemiologiche

Il percorso dei dati all'interno di un sistema integrato

La soluzione tecnologica

pierantonio.romor@insiel.it

Il percorso dei dati all'interno di un sistema integrato



Fasi del processo di acquisizione

- La fase di selezione di una fonte è il punto fondamentale del processo e viene effettuato in base a criteri di eleggibilità concordati con l'utilizzatore finale.
- La fase di acquisizione (una tantum) richiede un'analisi congiunta (IT, esperti dominio, epidemiologi) del sistema operazionale con analisi del processo di acquisizione del dato.
- Le fasi di ETL e definizione delle strutture target sono eseguite da personale tecnico, che garantisce la presenza continua ed integrata della fonte nel tempo.
- Fase di CQ con implementazione dei test di coerenza (con modellazione in serie storica) al fine di segnalare possibili incompletezze di caricamento.

Conoscere il dato

Conoscere la provenienza e quindi i criteri di produzione

ci permette di capire e selezionare.

Prima di immettere sul mercato i dati:

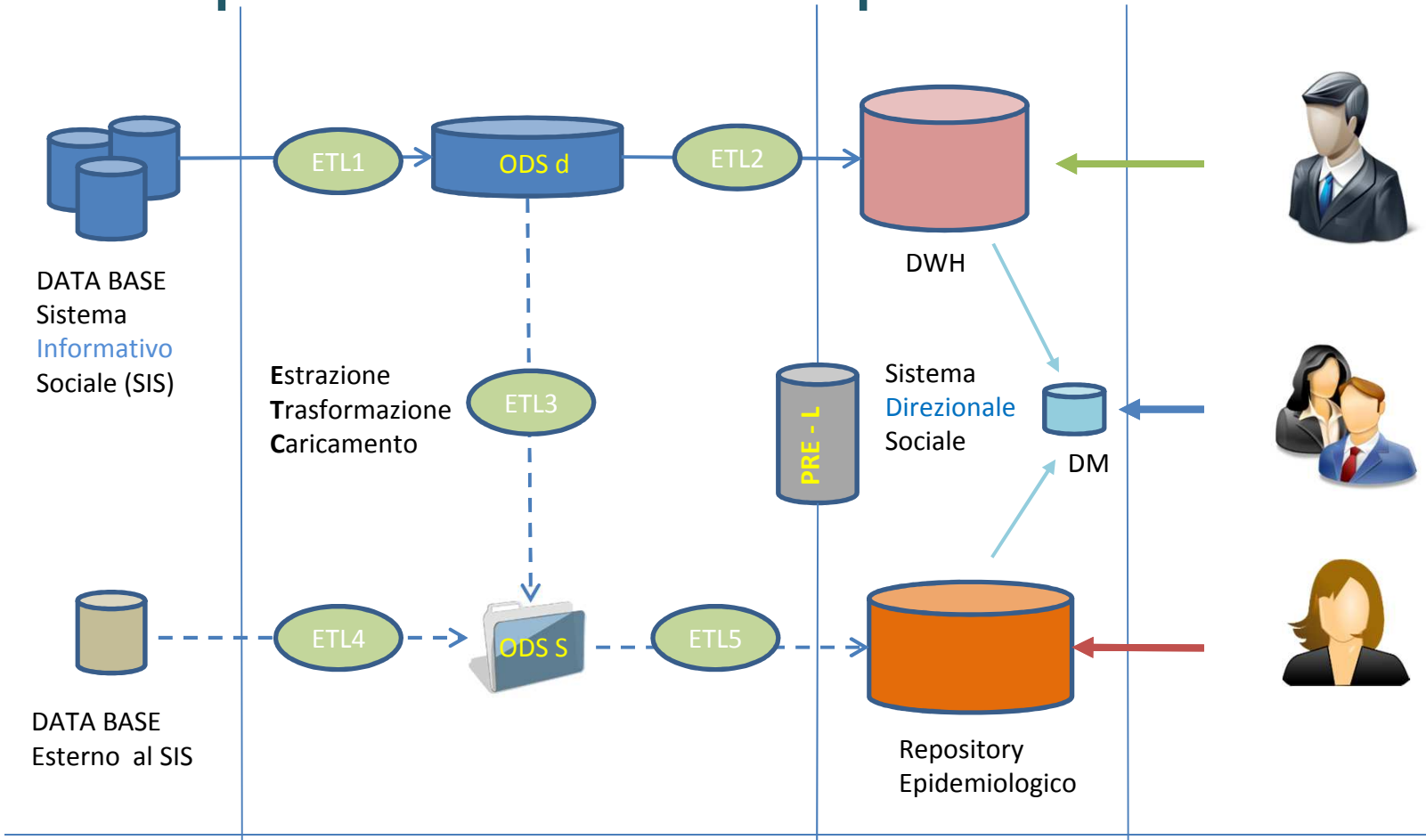
1. Esaminare il sistema di produzione (data profiling)
2. Effettuare controlli a posteriori (data quality)

Accessibilità ai dati

Tutti abbiamo sperimentato diverse modalità con cui richiedere un «accesso» ai dati:

- Accesso diretto o su copia (non mediato, richiede competenze, media condivisione, complesso, on line e dinamico)
- Vista (mediato dal richiedente, favorisce lo sviluppo, alta condivisione, strutturato , dinamico)
- Flusso (non sempre mediato, uso di flussi esistenti, bassa condivisione, non favorisce lo sviluppo del sistema, off line e statico)

Esempio di modello implementativo



Staging area: l'operational data store (ODS)

- Rappresenta un costrutto operativo che comporta l'immagazzinamento e la classificazione di una gran quantità di dati, in forma elementare, di agevole lettura e memorizzazione.
- Si configura principalmente come un'area tecnica, dove vengono consolidati, nelle fasi che precedono le attività di "cleaning", trasformazione e alimentazione, gli output dei processi di estrazione dei dati operazionali.
- I dati della staging area possono essere strutturati in DBMS oppure mantenuti sotto forma di file sequenziali e costituiscono l'ambito di disponibilità informativa statica (off-line) e dinamica (on-line).

Investimento fondamentale per la continuità informativa

Modelli di data warehousing applicati all'epidemiologia

Il modello infrastrutturale di riferimento proposto (R.E.R.) è composto da 3 livelli logici :

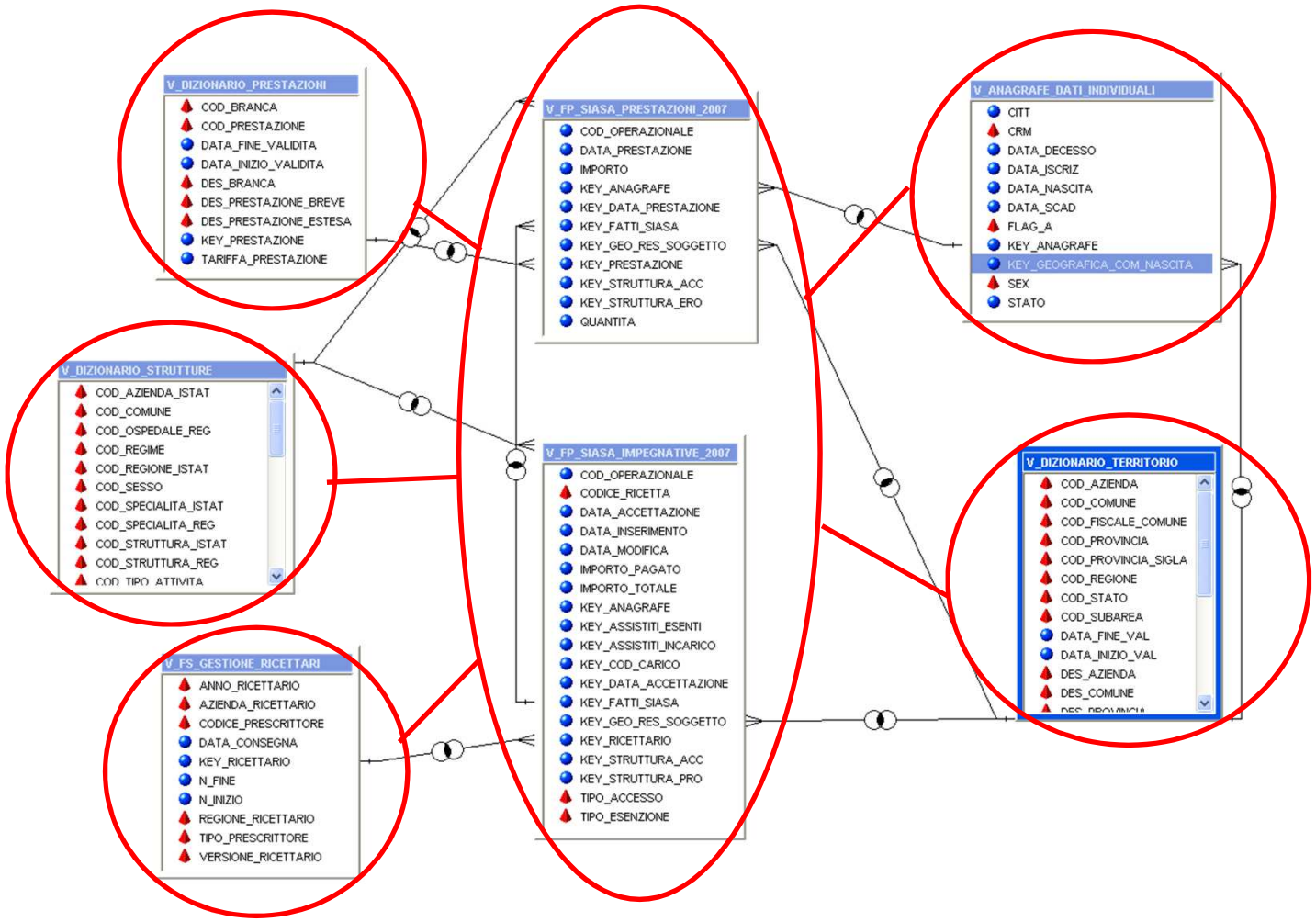
- livello dell'alimentazione e dei dati riconciliati,il repository regionale di microdati (RRMD).
- livello del Warehouse (infrastrutture dati derivate da algoritmi)
- livello dei Data Mart (strutture dinamiche per attività analitiche o di pubblicazione)

Il repository regionale di microdati

- Modello a costellazione di schemi dati
- Schemi di dati con elementi comuni: FP, FS, DIZ.
- Chiave anagrafica unica in forma di chiave surrogata, ri-generata ad ogni caricamento.
- Dizionari unici…
- Sistema con profilatura personalizzata.
- Accessibile in rete mediante tool di data management.
- Attività di analisi e propagazione all'interno dell'infrastruttura.

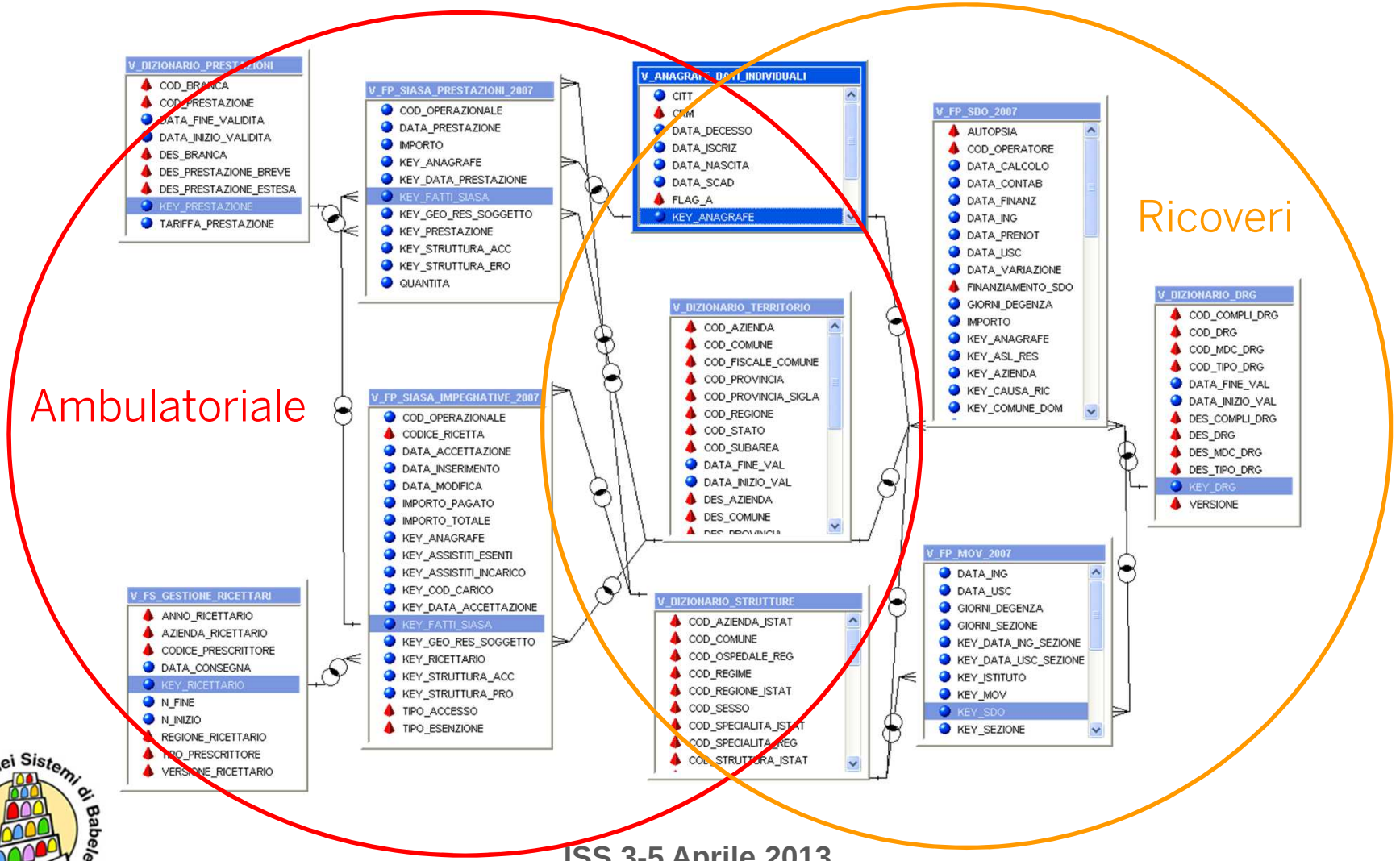
Strutture dati del RRMD

Esempio di schema a stella a livello di Repository



Strutture dati del RRMD

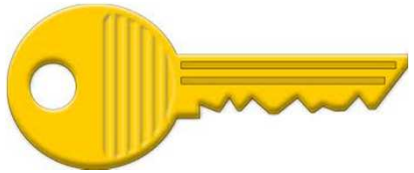
Esempio di relazioni tra tabelle e fonti a livello di Repository



La chiave surrogata (trimestrale)

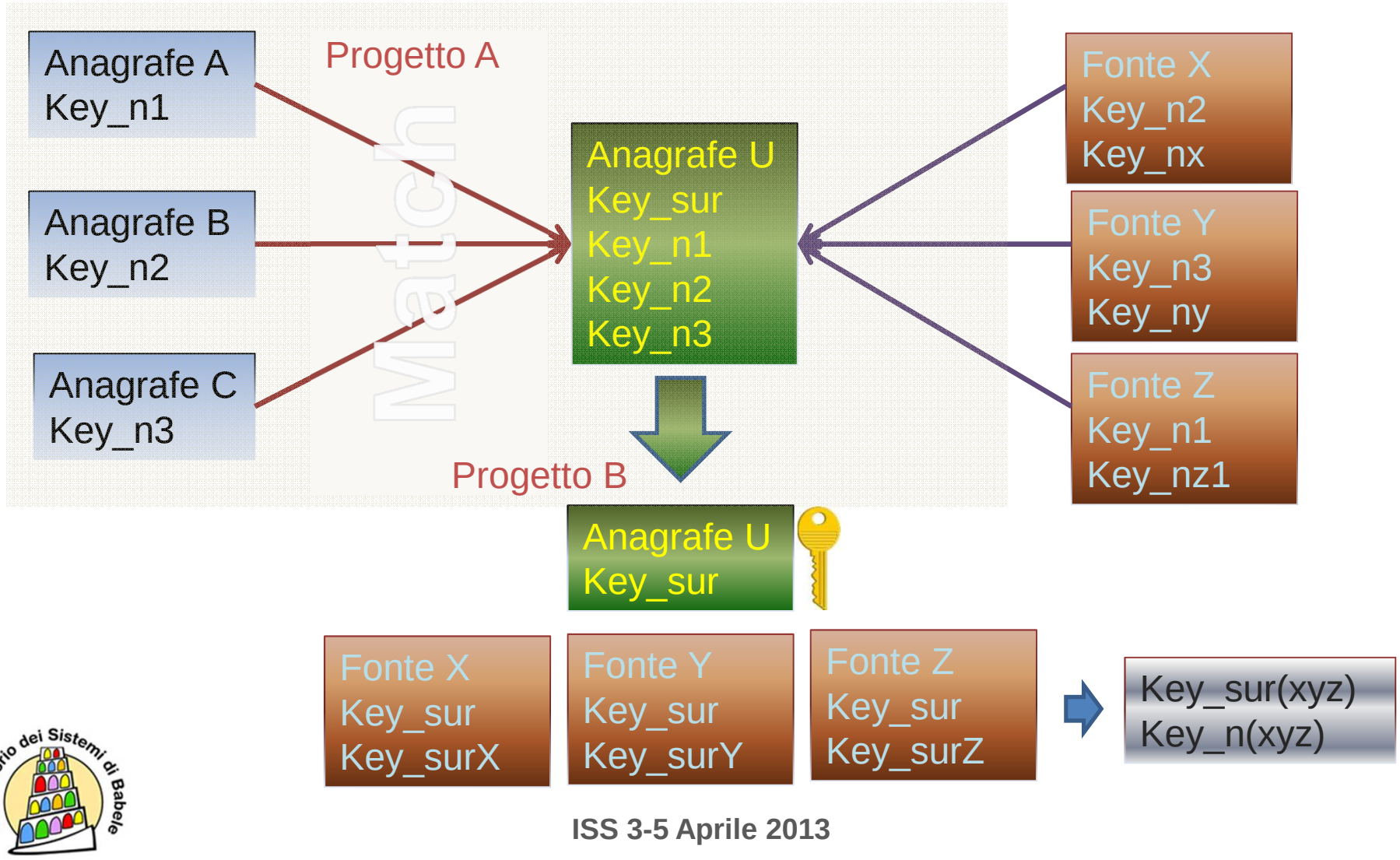
Fasi per la costruzione di una chiave anagrafica comune:

- Ordinamento su chiave naturale o sequenziale provvisoria (N1)
- Generazione numero casuale (N2)
- Ordinamento della sequenza casuale (N2)
- Generazione del nuovo numero sequenziale (N3)



Per le fonti si utilizza una chiave surrogata non casuale

Fonte dati = anagrafica + eventi

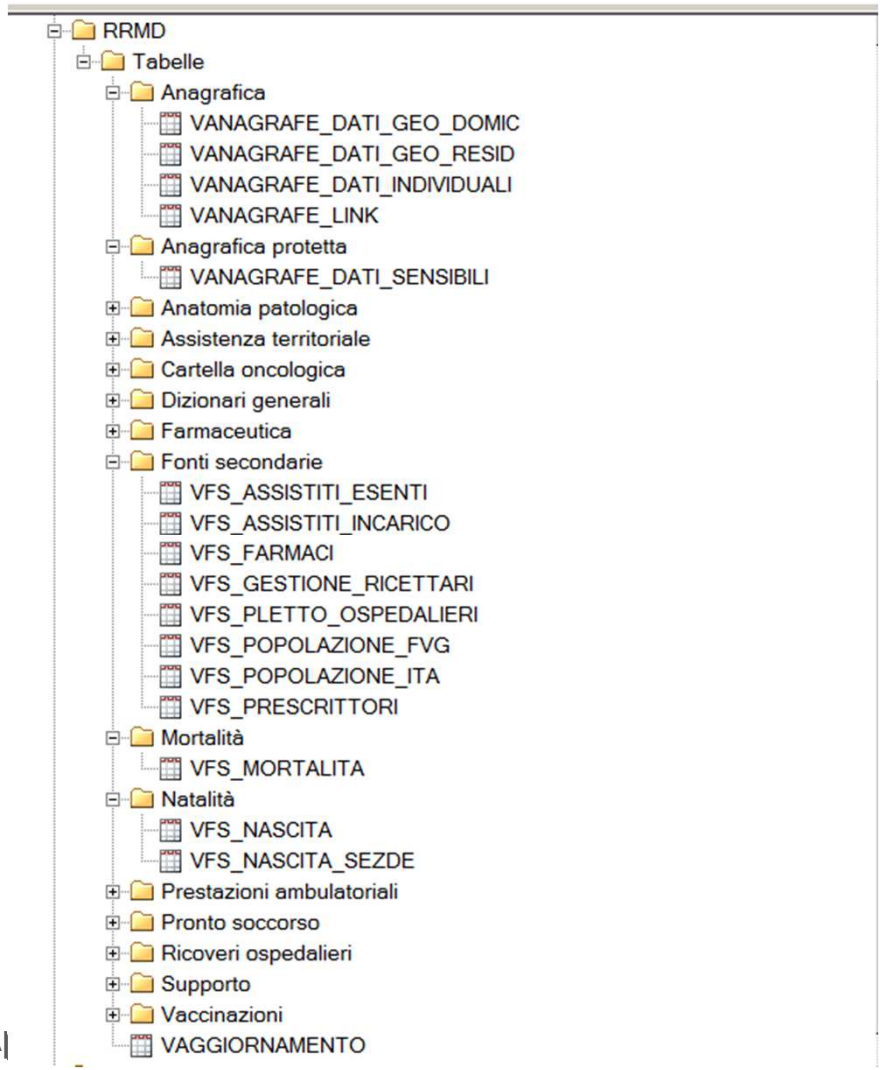


Il trattamento dei dati.
Tecniche di profilatura ed anonimizzazione

Le funzionalità disponibili sono:

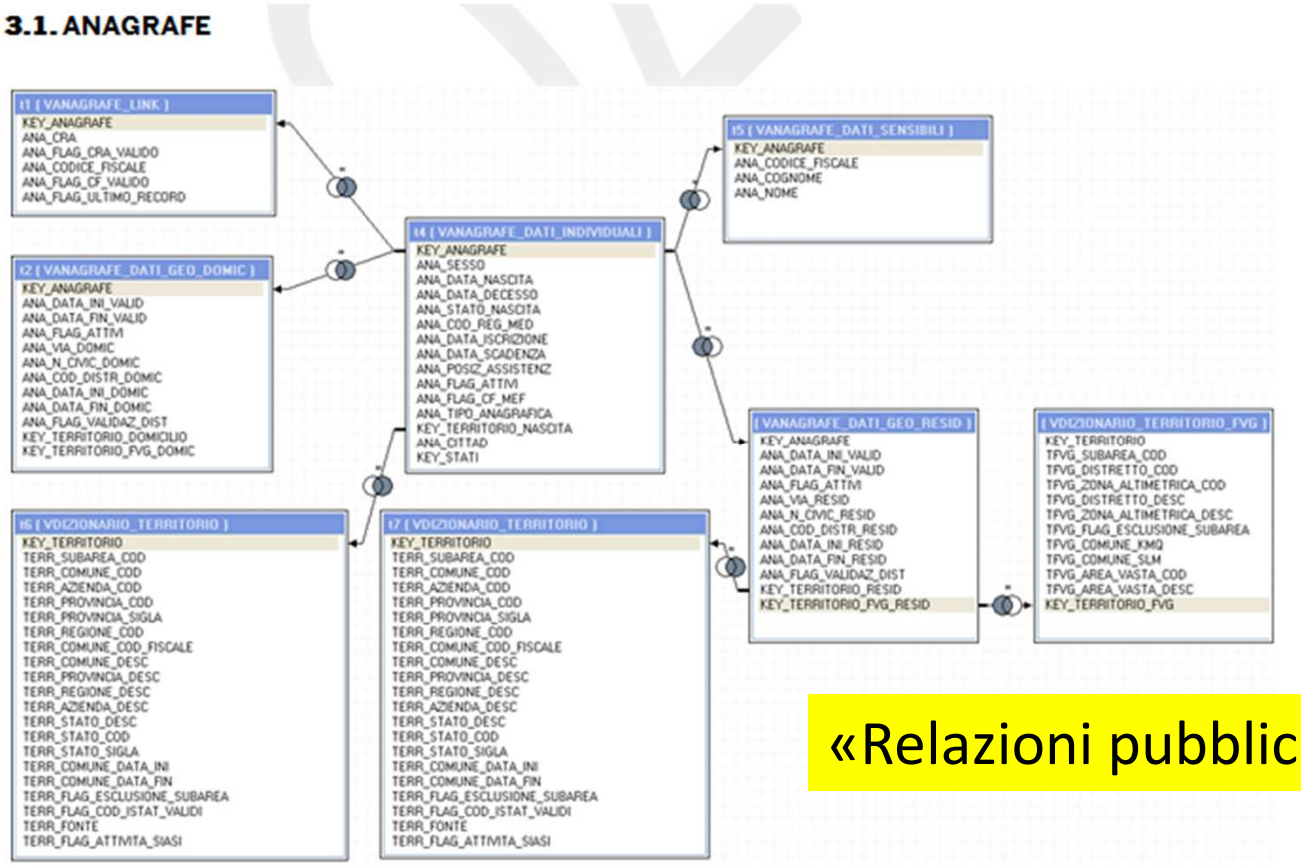
- Accesso completo anonimo
- Accesso parziale (residenza o struttura) o totale nominativo
- Profilato per fonte primaria

Nome	Descrizione
DIZIONARI SAS	Dizionari SAS di supporto
FLUSSI_SCREENING	
REGISTRO_DIABETE	Registro Regionale Diabete
REGISTRO_DIALISI	Registro Regionale Dialisi
REGISTRO_INCIDENTI	Registro Regionale Incidenti Stradali
REGISTRO_MALRARE	Registro Regionale Malattie Rare
REGISTRO_TUMORI	Registro Regionale Tumori
RRMD_SANITA_FVG	Repository Epidemiologico
SCREENING_CERVICE	DWH Programma Screening Cervice
SCREENING_COLONRETTO	DWH Programma Screening Colon-Retto
SCREENING_MAMMOGRAFICO	DWH Programma Screening Mammogr...
WORK	



Il Manuale operativo utente

Necessario un documento di ausilio all'utente finale con rappresentazione degli schemi base.



«Relazioni pubbliche»

RRMD - Aperture

- (IN) Utilizzare i dati presenti nel RRMD per effettuare integrazione con dati locali, sfruttando direttamente le risorse dell'infrastruttuta (utenza privilegiata) o in link-service (CED), su aree di work.
- (OUT) Distribuire basi informative anonime una tantum (coorti) a diverso livello di granularità.
- Aggiornamento e adeguamento nel tempo a seguito dell'evoluzione dei sistemi (es. SEI)
- Acquisizione dati non strutturati, coordinate spaziali
- Soluzione unica e condivisa per integrazione con altri sistemi analitici (GIS e reti neurali) tramite aree comuni di scambio (egtask e geotask).

I DWH – Registri di patologia

Dati strutturati

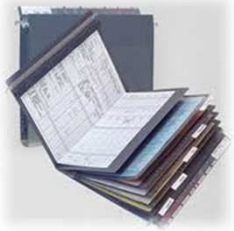
Tumori
Incidenti Stradali

R.R.M.D.

Diabete

Dialisi

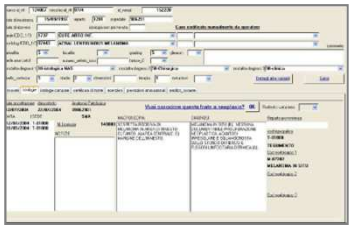
Registri di Patologia

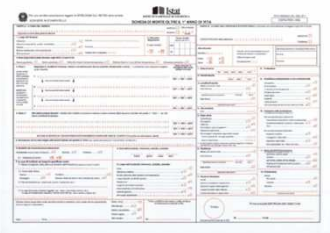


Documenti Sanitari

Malattie Rare

Res. Batteriche



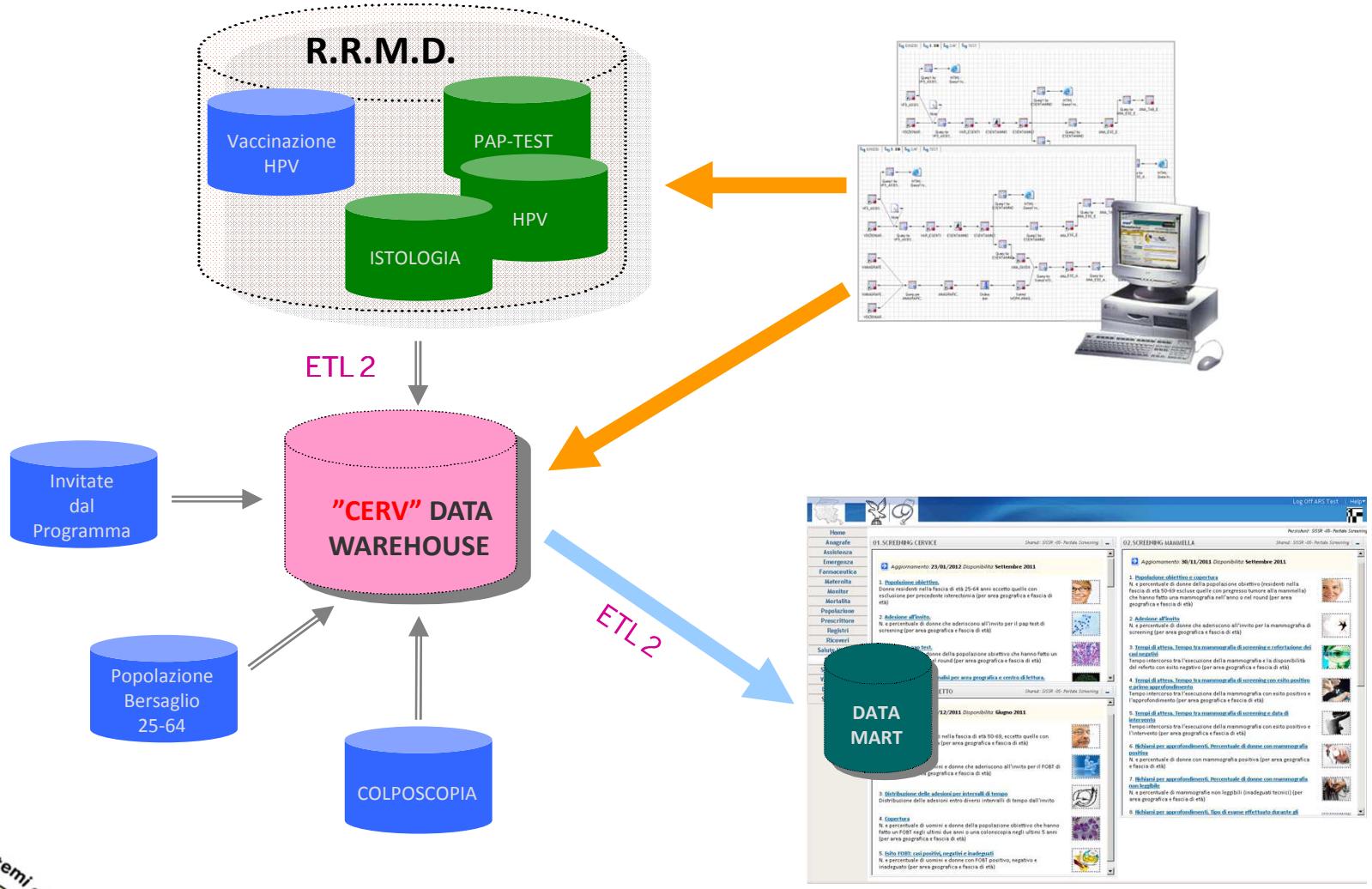


Cause
di Morte

Soluzioni in relazione a:
qualità, tempi e costi

Laboratorio dei Sistemi di Babele

I DWH – Programmi di screening



Attenzione all'infrastruttura

Buona parte di ciò che è stato realizzato ha già subito versionamenti, non infrastrutturali, di contenuto.

Tutte le risorse informatiche vengono **centralizzate** a favore del centro e della periferia.

Centralizzare significa, **ridurre la complessità infrastrutturale** minimizzando, o eliminando completamente, ogni ridondanza hardware e di gestione/manutenzione, **liberando buona parte del budget IT** per ulteriori investimenti.

Uso delle fonti di dati sanitari correnti per finalità epidemiologiche

Il percorso dei dati all'interno di un sistema integrato

Un caso d'uso

pierantonio.romor@insiel.it

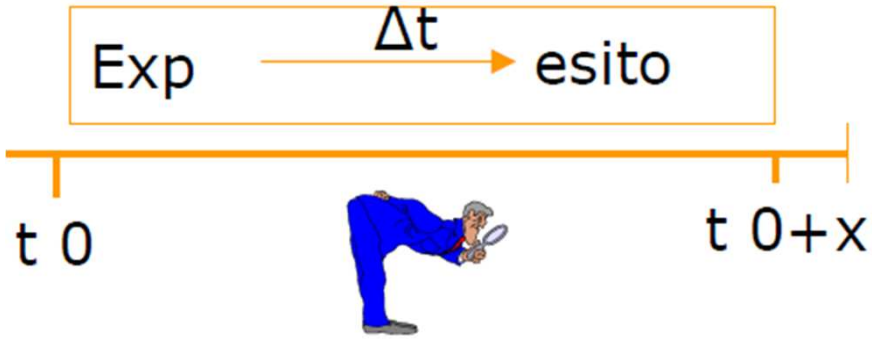
Gli studi di coorte

Alcuni svantaggi:

- Necessità di arruolare un numero elevato di soggetti (da seguire nel tempo).
- Spesso di lunga durata, organizzativamente difficile (tempi lunghi e costi elevati).
- Richiesta consenso.

Distorsioni più frequenti:

- Cambiamenti nel tempo delle metodologie di rilevamento.



Gli svantaggi e le distorsioni possono essere superati dai Sistemi Analitici Integrati

Approccio standardizzato per la generazione di coorti

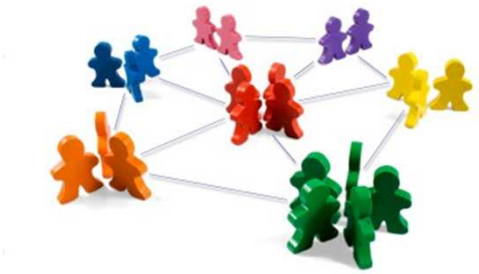
Un caso d'uso

- Coorte (chiusa) «post bellica dei sopravvissuti» nati prima del 01.01.1946 e vivi al 01.01.2000.
- Coorte (aperta) di tutti i nuovi nati a partire dal 01.01.1989 individuati tramite i certificati di assistenza al parto (CEDAP).

Residenti in un'area

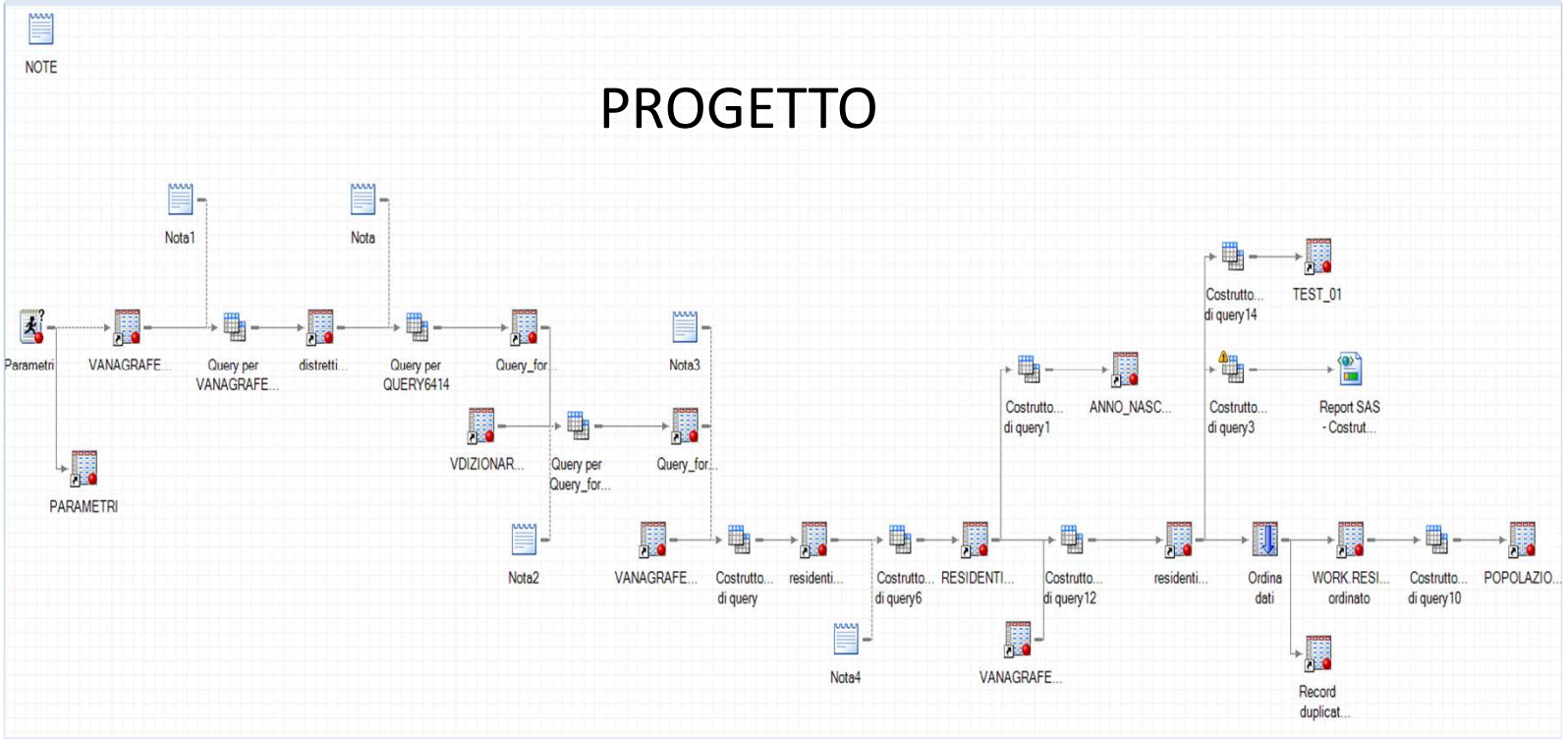
La progettazione e distribuzione del lavoro

- Fasi del progetto:
  - Definizione ed implementazione del protocollo di estrazione
  - Individuazione ed estrazione delle variabili indipendenti
  - Individuazione delle fonti e delle variabili dipendenti
- Standard:
  - Repository Regionale di MicroDati
  - Tool di data management ed analisi statistica
- Cooperazione applicativa
  - Condivisione progetto a livello di:
    - Regione
    - Aziende Sanitarie Territoriali
    - Istituti di ricerca interni
  - Condivisione dati con:
    - Istituti di ricerca esterni



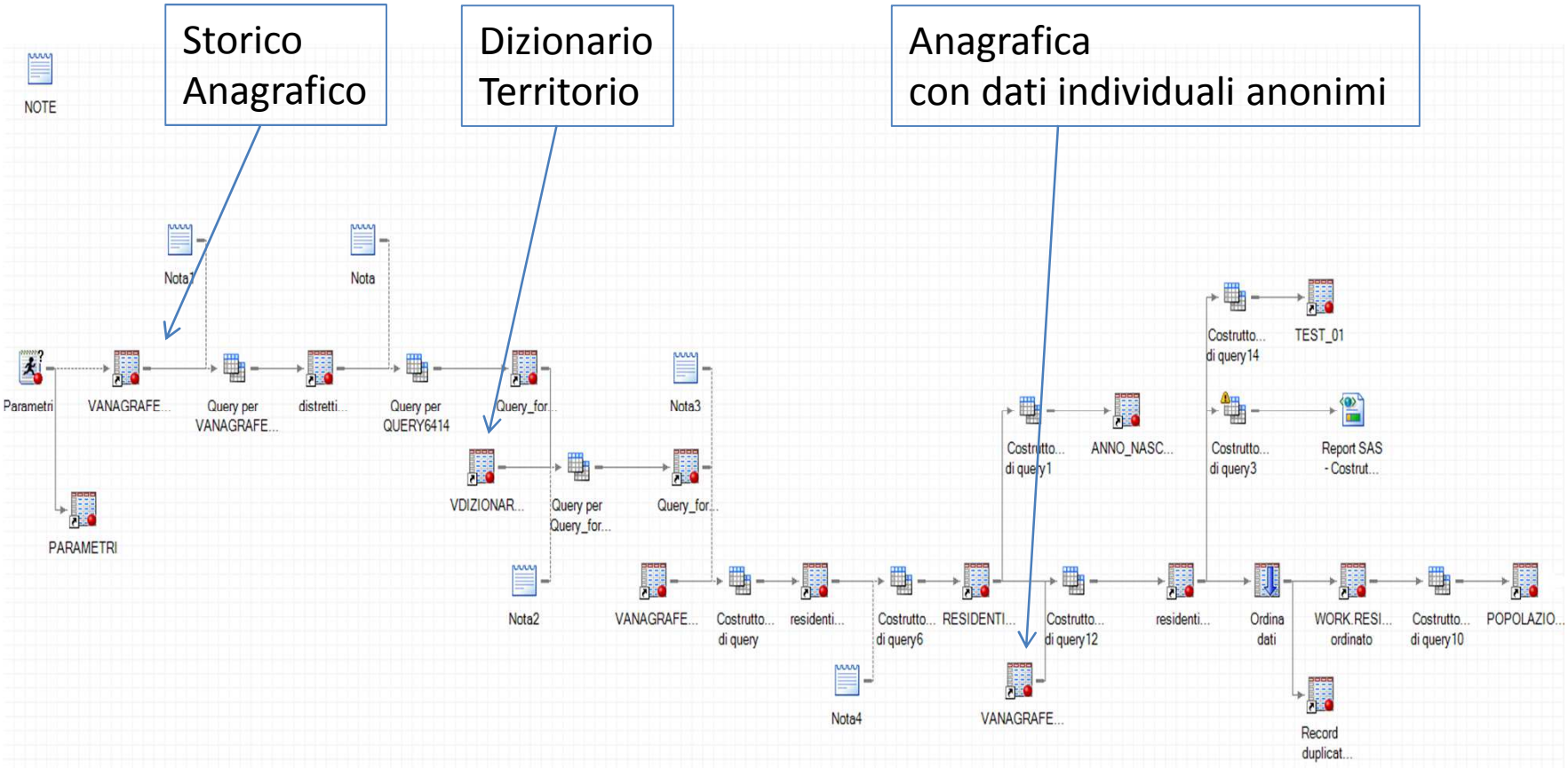
L’attività di data management.

Coorte «nuovi nati dal 1989» | Coorte «post bellica dei sopravvissuti»



Acquisizione della coorte è la **componente variabile** di progetto

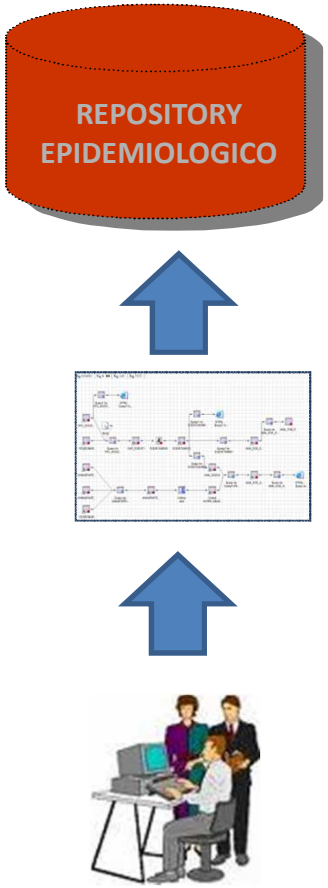
La scomposizione di un progetto



Estrazione residenti, in carico al SISSR e attivi al 2000	Selezione data nascita	Output

Il network

- La realizzazione del progetto di estrazione di coorti può essere distribuito su diversi utilizzatori competenti sul «Repository» e sullo strumento di «data management» indipendentemente dal profilo di accesso.
- Obiettivo: collaborazione, condivisione e riproducibilità dei processi.



Sistema auto-incrementale

Il rilascio in produzione del progetto (il cui risultato è la generazione di un set di chiavi anagrafiche) non genera un processo esterno al sistema ma riporta la coorte nel ciclo produttivo dell'infrastruttura dati.

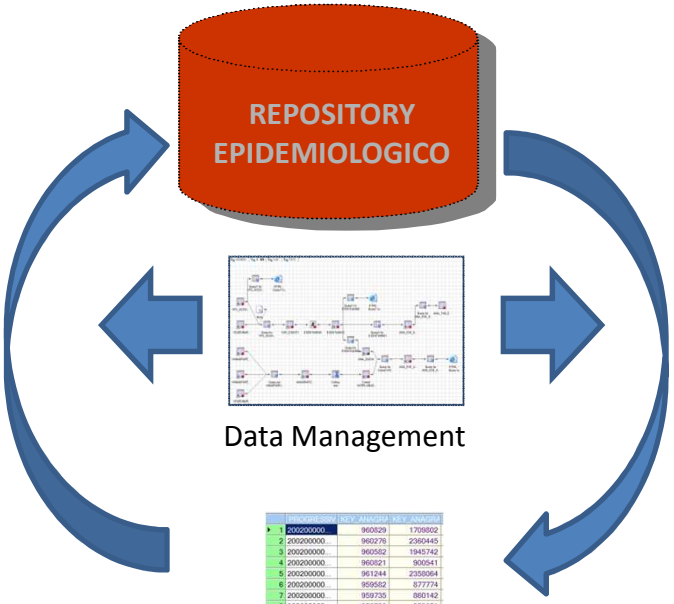
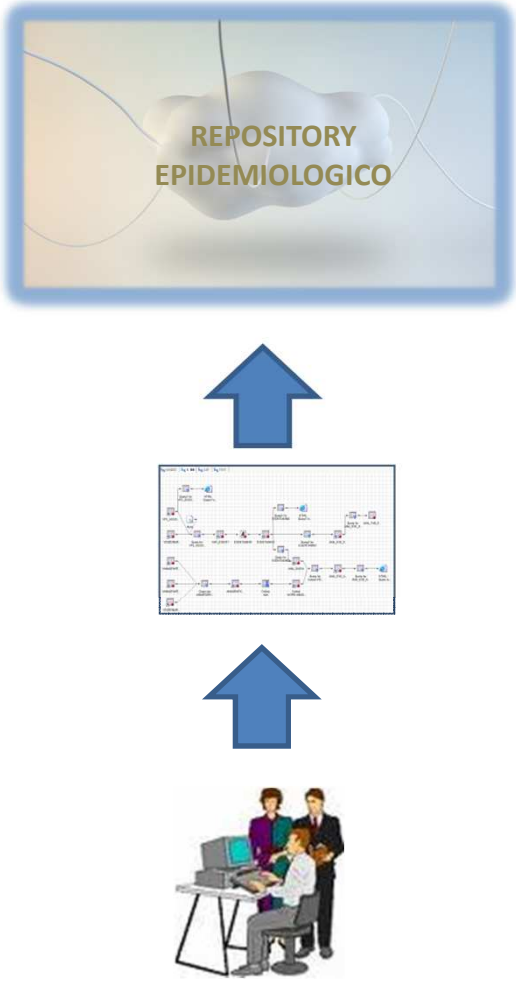


Tabella
COORTI

Il sistema distributivo

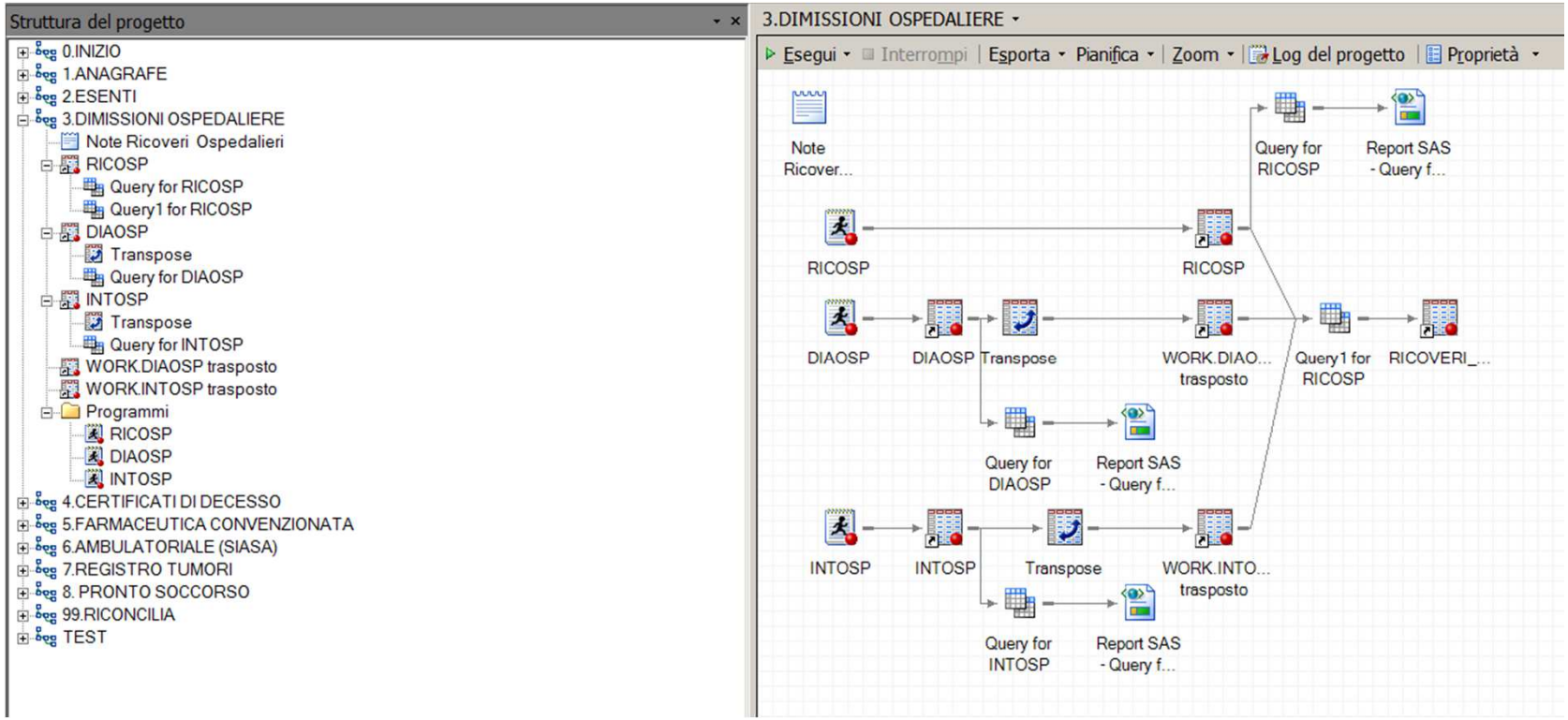
La pubblicazione di una tabella di coorti nell'infrastruttura consente ulteriori vantaggi:

- Utilizzo delle coorti rilasciate per le analisi on-line (interrogazione estesa del sistema)
- Il monitoraggio dell'arruolamento, trasversale alle coorti per la condivisione delle informazioni comuni extra sistema.
- La condivisione delle coorti nel network regionale.
- La predisposizione di flussi ad hoc per collaborazioni esterni mediante processi standardizzati di estrazione dati.



Predisposizione flussi ad hoc

N coorti -> 1 progetto



Infrastruttura aperta e collaborativa