Studio Dell'Incidenza Del Cancro Sulla Popolazione Statunitense Negli Anni 2000 – 2010

Saccotelli Leonardo Matricola: 646932

Troia Gianmarco Matricola: 647815

Abstract

"Il cancro dal 2014 è ufficialmente la prima causa di morte in 22 stati degli USA. Il sorpasso sulle malattie cardiovascolari, che detenevano da decenni questo primato, è stato di recente comunicato dai **Centers for Disease Control and Prevention *(CDC)*.***"* (MONTEBELLI, 2016, para. 1)

In questo caso di studio si è voluto analizzare l'evoluzione dei tassi di mortalità di varie forme tumorali stimando come questi abbiamo avuto ripercussione sulla popolazione.

L'obiettivo principale è quello di valutare come negli anni 2000 – 2010 nelle 3141 contee degli Stati Uniti 31 diverse forme tumorali abbiano subito variazioni in termini di tassi di mortalità. Il tutto realizzando un sistema Data Warehouse a supporto dello scopo.

Attraverso vari report realizzati utilizzando lo strumento Power B.I. sono stati resi disponibili una serie di grafici che mettono in evidenza vari aspetti dell'analisi che si è voluta condurre. In particolare, sarà possibile mostrare con chiarezza come l'andamento dei decessi sia variato nel corso degli anni a seconda del sesso, della contea e del tipo di tumore. Infine, sarà possibile evidenziare quali tumori hanno rappresentino la principale causa di morte nel periodo di tempo preso in considerazione.

Keywords: Tumore, Tassi Di Mortalità, Data Warehouse, Power B.I.

---

1 Introduzione

Il lavoro svolto è stato possibile utilizzando i dati relativi ai tassi di mortalità forniti dall' **Istitute for Health Metrics and Evaluation** (**IHME**).

"L'Institute for Health Metrics and Evaluation (IHME) è un istituto di ricerca che lavora nell'area delle statistiche sanitarie globali e della valutazione dell'impatto presso l'Università di Washington a Seattle." (*Institute for Health Metrics and Evaluation*, n.d., para. 1)

IHME conduce ricerche e forma scienziati, responsabili delle politiche riguardanti concetti, metodi e strumenti relativi alle metriche sanitarie. La sua missione include giudicare l'efficacia e l'efficienza delle iniziative sanitarie e dei sistemi sanitari nazionali.

L'IHME raccoglie dati relativi alla salute e sviluppa strumenti analitici per tenere traccia delle tendenze in termini di mortalità, malattie e fattori di rischio.

Valuta interventi quali vaccini, politiche di controllo della malaria, screening del cancro e cure alla nascita.

Per consentire ai ricercatori di replicare il lavoro dell'IHME e promuovere nuove ricerche, l'IHME ha creato il **Global Health Data Exchange** (**GHDx**) in cui i metodi e i risultati sono catalogati e liberamente accessibili.

Come riportato nell'articolo di Montebelli, oggi il cancro è sicuramente una delle principali cause di morte non solo negli Stati Uniti ma in tutto il mondo.

Abbiamo scelto di svolgere questo caso di studio in quanto riteniamo interessante analizzare l'andamento e l'incidenza di questa malattia nel corso degli anni.

In particolare, ci aspettiamo di riscontrare nel corso degli anni una riduzione della mortalità di queste malattie correlata, con buona probabilità, ad una maggiore prevenzione e ad un miglioramento delle terapie adottate.

2 Letteratura a supporto del caso di studio

Possiamo definire la "Business intelligence" come:

> *"Un insieme di strumenti e procedure che consentono a un'azienda di trasformare i propri dati di business in informazioni utili al processo decisionale, da rendere disponibili alla persona giusta e nel formato idoneo. Le informazioni così ottenute sono utilizzate dai decisori aziendali per definire e supportare le strategie di business, così da operare decisioni consapevoli e informate con l'obiettivo di trarre vantaggi competitivi, migliorare le prestazioni operative e la profittabilità e, più in generale, creare valore per l'azienda."* (Golfarelli, Matteo; Rizzi, 2006)

Attraverso la business intelligence si riesce quindi a convertire dati in informazioni e successivamente informazioni in conoscenza.

Negli ultimi decenni la diffusione dell'informatica è stata parte fondamentale per la crescita delle imprese. I vari sistemi informativi sono parte integrante delle realtà aziendali e ne supportano i processi. Qui una breve distinzione tra le tipologie di sistemi.

Sistemi OLTP (OnLine Transaction Processing) impiegano tecniche software utilizzate per la gestione di applicazioni orientate alle transazioni. In questi sistemi vengono memorizzati dati per produrre informazioni.

Si dividono in:

1) **TPS** (Transaction Processing System)
   I sistemi di elaborazione delle transazioni (TPS) sono i sistemi informativi che servono il livello operativo dell'azienda. Le risorse di un TPS devono supportare le attività di base dell'azienda, ovvero registrare fatti, eventi, entità che si generano durante la "quotidiana vita aziendale". I sistemi di elaborazione delle transazioni sono fondamentali per un'azienda: costituiscono i principali produttori di dati per gli altri tipi di sistemi informativi. I sistemi TPS memorizzano sia dati strutturati all'interno dei data base sia dati non strutturati all'interno di file gestiti dal file system

del sistema operativo su cui sono memorizzati. L'obiettivo ultimo del sistema TPS è quello di fornire informazioni attraverso l'esecuzione di query SQL.

2) **MIS** (Management Information System)
   L'obiettivo di tali sistemi è quello di fornire al management informazioni tempestive, affidabili, standardizzate e routinarie (attraverso strumenti quali budget, report e statistiche) al fine di agevolare l'assunzione di decisioni ripetitive.
   I sistemi di gestione delle informazioni (MIS) forniscono ai middle manager dell'azienda un accesso online alle prestazioni correnti. Queste informazioni sono prodotte attraverso query, sempre in SQL, che operano sugli stessi data base in cui sono memorizzati i dati prodotti dai sistemi TPS. Non introducono dati ma accedono al DataBase in sola lettura.

Sistemi OLAP (OnLine Analysis Processing) forniscono un'analisi per il supporto alle decisioni e per produrre conoscenza.

Si dividono in:

1) **DSS** (Decision Support System)
   I DSS forniscono supporto ai manager e a tutti coloro che devono prendere decisioni strategico/operative di fronte a problemi poco strutturati o non strutturati; essi consentono di effettuare analisi ad-hoc sui dati in tempi brevi e in modo flessibile producendo, quella conoscenza che serve a supportare e migliorare in termini di efficacia il processo decisionale. I sistemi DSS sono costituiti da software che elaborano dati memorizzati in repository diversi, denominati Data Warehouse.
2) **ESS** (Executive Support System)
   I sistemi ESS applicano tecniche statistiche ai dati memorizzati nel Data Warehouse per prevedere il futuro e pianificare al meglio lo sviluppo aziendale. Gli ESS sono progettati per incorporare i dati legati ad eventi esterni. Essi filtrano, comprimono, estraggono e individuano i dati critici.
   I sistemi ESS sono costituiti da software che insistono sullo stesso Data Warehouse progettato per i DSS ed utilizzano tecniche di statistica avanzata per fare previsioni.

È chiaro come le due tipologie siano differenti, i sistemi OLTP hanno un numero alto di utenti, le query sono congelate nel software che accede al DataBase. Il DataBase infatti è al centro del sistema, e per essere corretto ha molte tabelle normalizzate in maniera tale da eliminare la ridondanza del dato. I sistemi OLAP invece hanno un numero di utenti ristretto e specifico, servono per produrre conoscenza a partire dai dati e fare previsioni aziendali, si basano su un sistema chiamato Data Warehouse.

2.1 Architettura di un data warehouse

"Un data warehouse (DW) è una base di dati utilizzata per il supporto alle decisioni e possiede per questo alcune caratteristiche peculiari che in parte la distinguono rispetto a una base di dati dedicata allo svolgimento di operazioni OLTP." (Paolo Atzeni et al., 2002)

Di seguito riportiamo alcuni dei requisiti che caratterizzano un data warehouse:

- È una base di dati integrata in quanto i dati di interesse provengono da diverse sorgenti informative preesistenti e questo richiede un'attività propedeutica di riconciliazione delle eterogeneità.
- Contiene informazioni di carattere storico/temporale poiché in un DW è di interesse l'evoluzione storica delle informazioni.
- Contiene tipicamente dati in forma aggregata.
- Ha un'estensione autonoma essendo fisicamente separato dalle sorgenti informative.
- È una base di dati fuori linea in quanto i meccanismi di importazione dei dati sono normalmente di tipo asincrono e periodico, in modo da non penalizzare le prestazioni delle sorgenti di dati.

Per soddisfare i seguenti requisiti un'architettura di DW comprende i seguenti componenti.

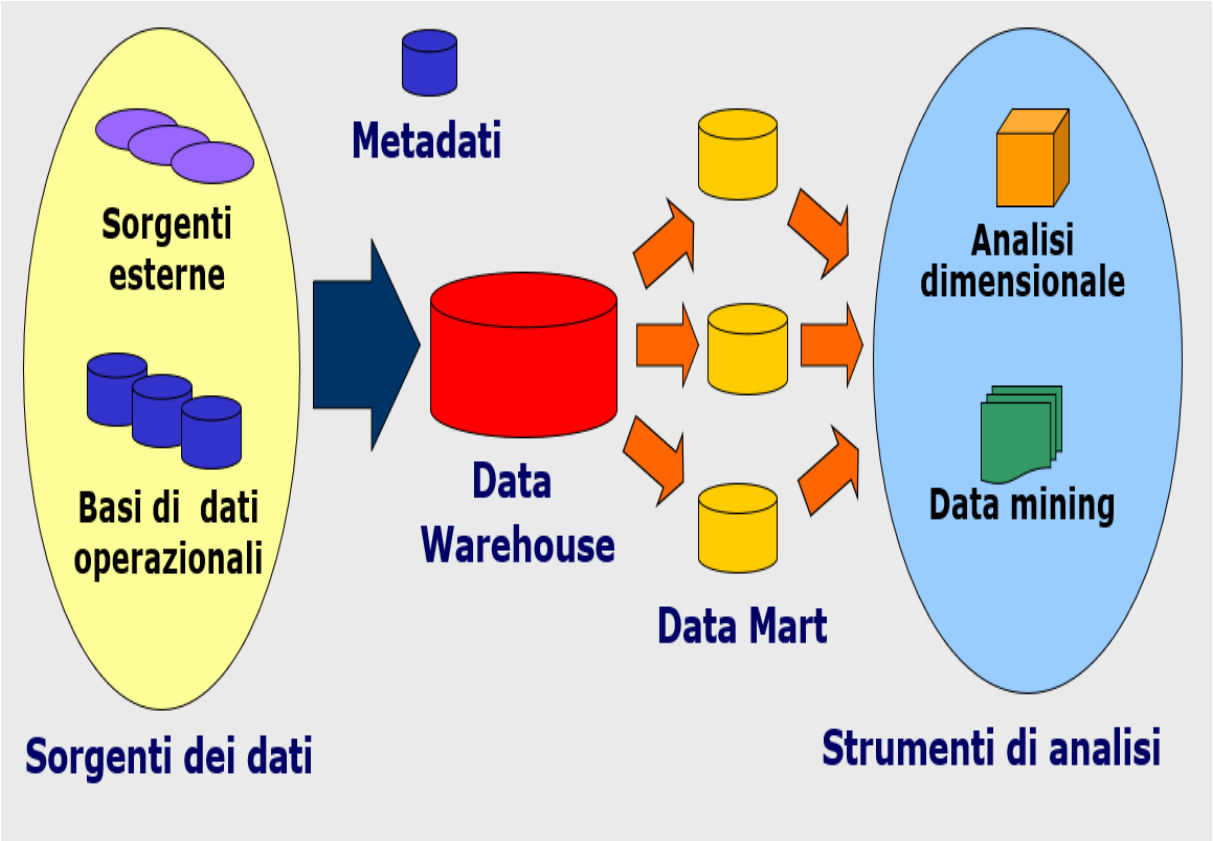

Figura 2.1 Architettura del Data Warehouse

- Le sorgenti o data source. I dati vengono estratti da uno o più sistemi preesistenti, nel sistema informativo aziendale, per la gestione operativa quotidiana oppure esterni a esso ma comunque accessibili.
- Il data warehouse server. È il sistema dedicato alla gestione del DW e può basarsi su diverse tecnologie (ROLAP o MOLAP). I dati vengono memorizzati in opportune strutture fisiche e realizza in modo efficiente query complesse. Inoltre, consente operazioni speciali che verranno illustrate successivamente.
- Un sistema di alimentazione. Consiste di una serie di strumenti detti di ETL (Extract, Transform, Load) che svolgono le varie attività di base. Le procedure E.T.L. vengono eseguite in modalità batch, cioè senza intervento umano, e se ci dovessero essere errori durante la procedura dovrebbero essere comunicati attraverso un apposito sistema di alert. Le attività svolte sono le seguenti.
  - L'estrazione dei dati dalle sorgenti. Dopo il popolamento iniziale, il processo di estrazione è tipicamente incrementale.

- La pulizia dei dati (data cleaning). Lo scopo è quello di analizzare la correttezza dei dati prima dell'inserimento nel DW. Si procede ad eliminare dati palesemente scorretti applicando controlli ad hoc sui singoli data source, oppure correggendo errori e inconsistenze nei dati estratti.
- La trasformazione dei dati. In questa fase vengono svolte conversioni, trasformazioni di formato, associazioni tra campi equivalenti di sorgenti diverse, operazioni di denormalizzazione, di ordinamento e di aggregazione.
- Il caricamento dei dati nel DW. Il caricamento può avvenire attraverso un refresh dove riscriviamo interamente i dati, o un update dove solo i cambiamenti nei dati sorgente sono caricati.

- Alcuni strumenti di analisi. Questi strumenti consentono di effettuare analisi dei dati usufruendo dei servizi offerti dal DW server e offrono interfacce amichevoli in grado di presentare, in forma adeguata, i risultati delle analisi.

Infine, può essere presente un livello fisico intermedio tra sorgenti e DW (detto staging area) nel quale vengono memorizzati i dati dopo la fase di estrazione, pulizia e integrazione ma prima del caricamento dei dati nel DW.

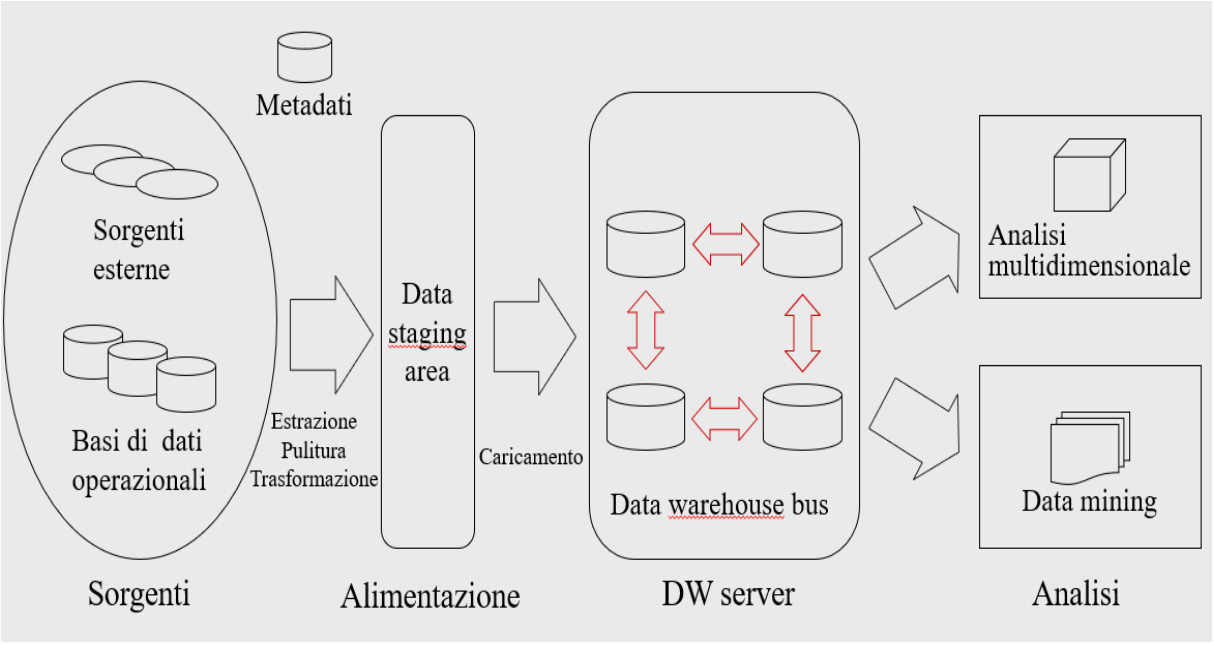

Figura 2.2 Architettura del Data Warehouse con stagin area

2.2 Rappresentazione multidimensionale dei dati

Il modello dei dati multidimensionale è basato su tre concetti principali: il fatto, la misura e la dimensione.

Possiamo definire un "fatto" come

> *"Un concetto del sistema informativo aziendale (o più precisamente, della relativa realtà di interesse) sul quale ha senso svolgere un processo di analisi orientato al supporto alle decisioni."* (Paolo Atzeni et al., 2002)

Possiamo definire una "misura" come

> *"Una proprietà atomica di un fatto che intendiamo analizzare (tipicamente un attributo numerico o un conteggio delle sue istanze)."* (Paolo Atzeni et al., 2002)

Possiamo definire infine una "dimensione" come

> *"Una particolare prospettiva lungo la quale l'analisi di un fatto può essere effettuata."* (Paolo Atzeni et al., 2002)

Esiste una naturale rappresentazione grafica nella quale le istanze di un fatto sono rappresentate da cubi multidimensionali costituiti da elementi atomici detti celle.

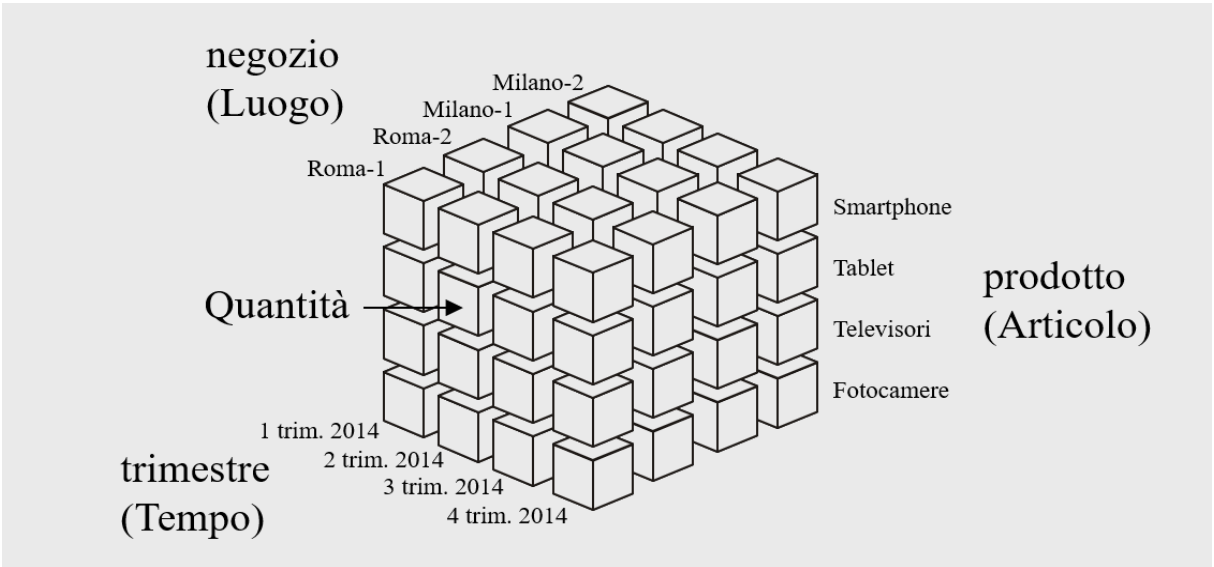

Figura 2.3 Rappresentazione multidimensionale di un Data Warehouse

Vengono definite pertanto alcune operazioni di analisi che si esprimono come operazioni sui cubi. Si tratti quindi di operazioni che si applicano a cubi multidimensionali e restituiscono nuovi cubi. Le operazioni più note sono:

- Slice-and-dice, che consiste nella semplice selezione di un sottoinsieme delle celle di un cubo.
- Roll-up, che consiste in un'aggregazione dei dati di un cubo seguito dall'applicazione di una funzione aggregativa (tipicamente la somma).
- Drill-down, che è l'operazione inversa del roll-up. Consente quindi di aggiungere dettaglio a un cubo disaggregandolo lungo una o più dimensioni.

2.3 Realizzazione di un data warehouse

Per la realizzazione di un data warehouse si contrappongono tre diverse soluzioni.

- La prima consiste nel memorizzare i dati direttamente in forma multidimensionale, tramite speciali strutture dati tipicamente proprietarie. Sistemi di questo tipo si dicono MOLAP (Multidimensional OLAP).
- La seconda consiste nell'uso della tecnologia relazionale adattata ed estesa. Sistemi di questo tipo si dicono ROLAP (Relational OLAP).
- Una terza soluzione prevede una combinazione delle tecniche precedenti. In questo caso si parla di Hybrid OLAP.

In particolare, in una realizzazione ROLAP i dati di un fatto multidimensionale sono organizzati secondo una semplice struttura relazionale, detta anche schema dimensionale oppure schema a stella. Lo schema a stella ha una struttura molto semplice ed è composta da:

- Una relazione principale, detta tabella dei fatti, che memorizza le istanze di un fatto;
- Varie relazioni ausiliarie chiamate tabelle dimensione che memorizzano i membri delle dimensioni associate al fatto;

- Un insieme di vincoli di integrità referenziale ognuno dei quali collega un attributo della tabella dei fatti a una tabella dimensione.

Lo schema a stella possiede le seguenti caratteristiche.

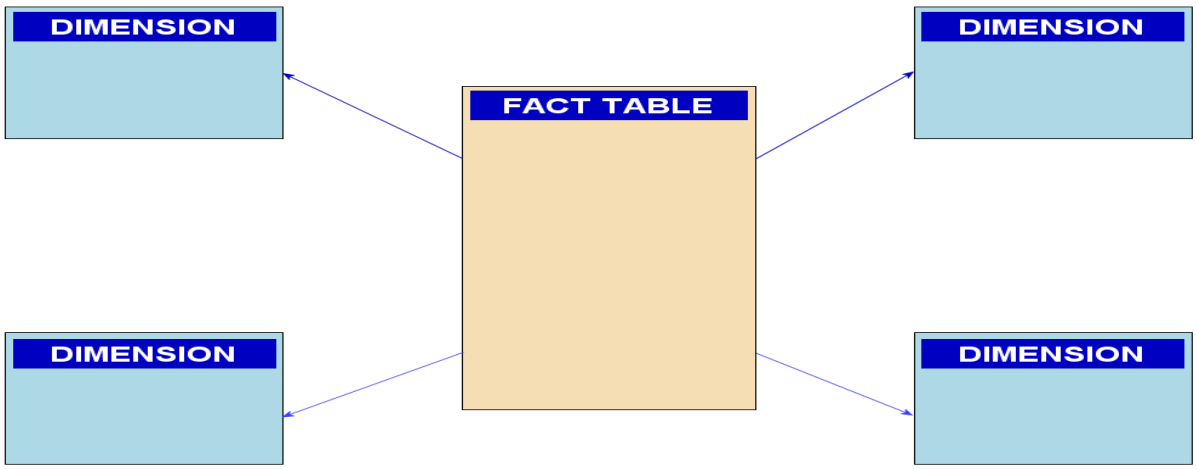

Figura 2.4 Schema a stella

La tabella dei fatti ha una chiave composta da attributi che sono riferimenti alle chiavi di tabelle dimensione; gli attributi rappresentano le misure del fatto e sono in generale numerici; infine, soddisfa la forma normale di Boyce-Codd.

Le tabelle dimensione hanno una chiave semplice (un solo attributo); gli attributi rappresentano i livelli della dimensione e sono tipicamente testuali e descrittivi; generalmente sono denormalizzate per motivi di efficienza (pur generando una certa ridondanza).

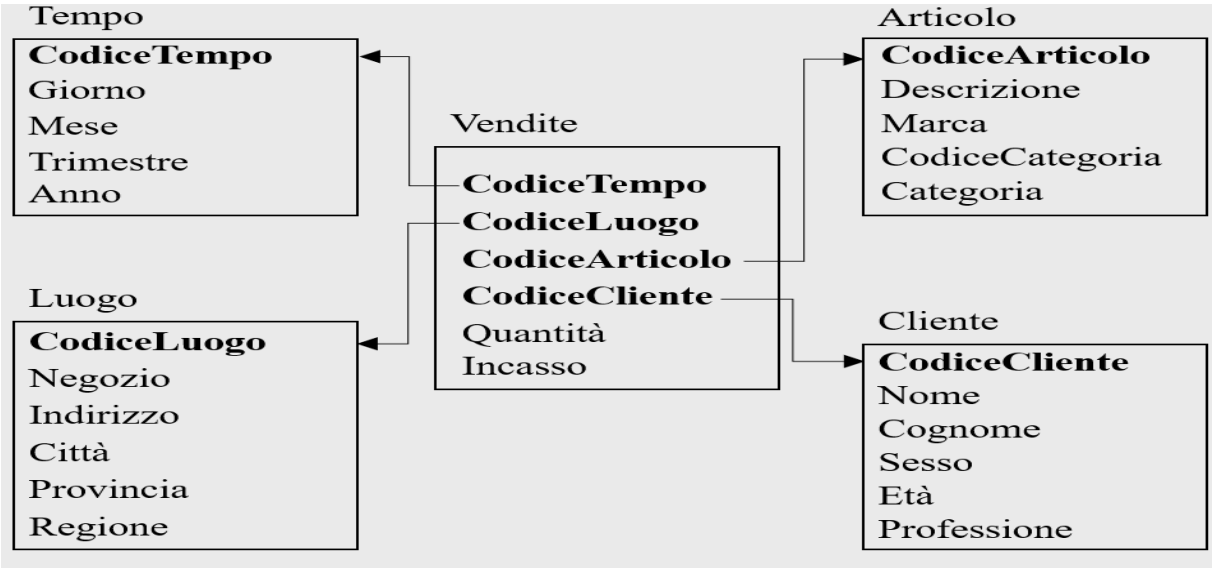

Figura 2.5 Esempio di schema a stella

Nel caso in cui si decide di normalizzare uno schema a stella per ridurre la ridondanza degli schemi dimensionali si ottiene uno schema detto schema a fiocco di neve.

In genere è sconsigliato procedere a normalizzazioni troppo spinte degli schemi a stella perché generalmente il beneficio che si ottiene in termini di riduzione di spazio non compensa il degrado delle prestazioni che le operazioni di join necessarie per ricostruire le dimensioni possono generare.

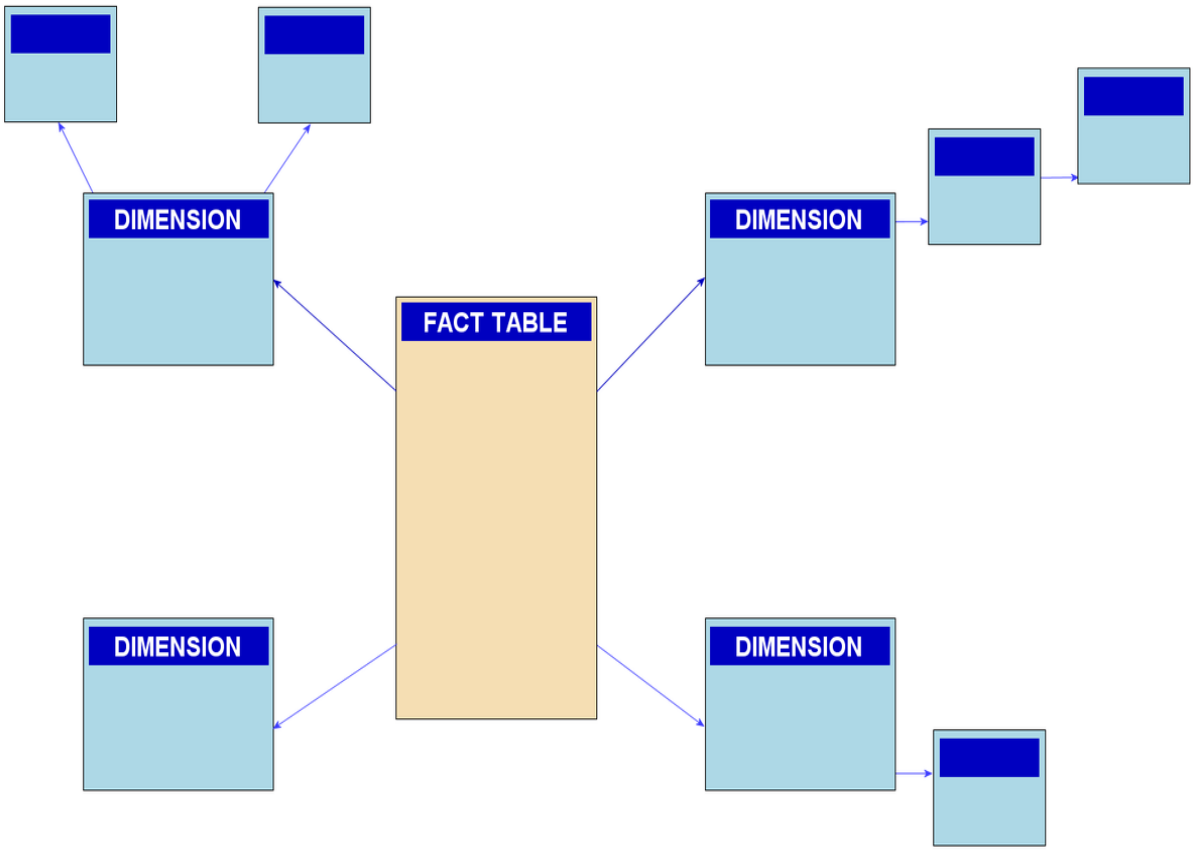

Figura 2.6 Schema a fiocco di neve

2.4 Progettazione di un data warehouse: una metodologia generale

Una metodologia generale per la costruzione e l'uso di data warehouse si articola nelle seguenti fasi principali.

1) Pianificazione, in cui vengono definiti gli obiettivi, i costi e i possibili benefici dalla realizzazione di un DW.

2) Raccolta e analisi dei requisiti, in cui si definiscono i requisiti di analisi che il DW dovrà soddisfare sulla base delle sorgenti di dati a disposizione.
3) Progettazione, che si divide in tecnologica, dei dati e delle applicazioni. Queste tre attività richiedono un forte coordinamento a causa delle forti dipendenze che esistono tra di esse.
4) Validazione e avviamento, che consiste nella realizzazione del DW secondo la struttura e le caratteristiche definite nella fase di progettazione e nella verifica del corretto funzionamento del sistema.
5) Manutenzione e crescita, in cui si eseguono periodicamente le operazioni necessarie alla sua alimentazione con nuovi dati proveniente dalle sorgenti.

Come avviene in tutti i processi di sviluppo del software, è anche presente un'attività di gestione che si svolge nell'intero ciclo di vita del Data Warehouse ed è finalizzata a monitorare lo sviluppo del progetto, curare la comunicazione tra i partecipanti e verificare il rispetto delle tempistiche e il raggiungimento degli obiettivi prefissati.

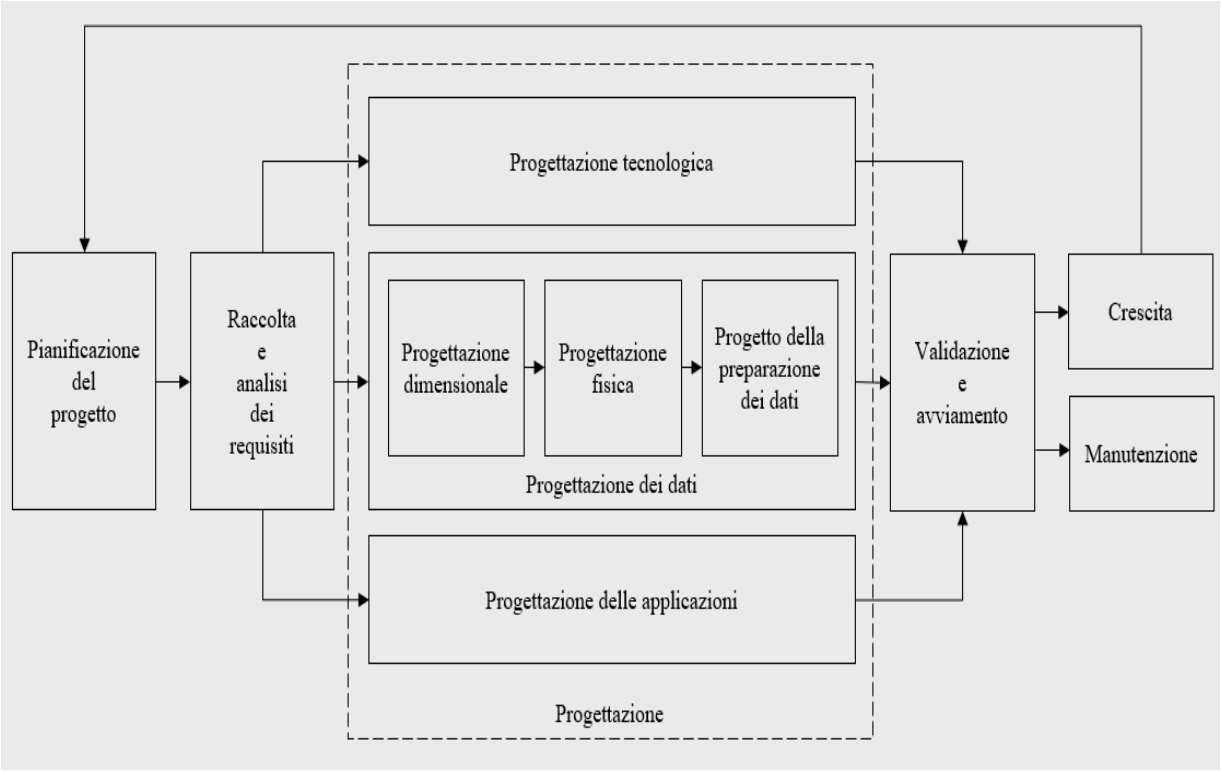

Figura 2.7 Metodologia di sviluppo di un Data Warehouse

3 Definizione del processo di data warehousing

3.1 Sorgenti

I dati utilizzati nel caso di studio sono contenuti in file csv accessibili direttamente dal GHDx. In particolare, verranno utilizzati 50 diversi file csv, uno per ogni stato degli Stati Uniti per un totale di 1.968.934 tuple.

Ai fini dell'analisi si è considerato un intervallo di tempo compreso fra l'anno 2000 e l'anno 2010. Inoltre, in ogni file rappresentante un singolo stato, sono riportate le contee che lo costituiscono. Sono state scelte 31 diverse forme tumorali, per ciascuna delle quali è noto il tasso di mortalità. Infine, per ogni anno si tiene traccia della popolazione delle singole contee divise per sesso.

Di seguito riportiamo un'istanza del dataset a titolo di esempio.

FIPS	51
NameState	Virginia
FIPS_County	51087
Name_County	Henrico County
Year_ID	2000
ID_Tumor	411
Type_Tumor	Esophageal cancer
Sex_ID	1
Sex_name	Male
Mortality	9.67209455168102
Population	123594

Tabella 3.1 Istanza del file Florida_Mortality_2000_2010.csv

- “FIPS” rappresenta il codice identificativo dello stato.
- “NameState” rappresenta il nome dello stato.
- “FIPS_County” rappresenta il codice identificativo della contea.
- “Name_County” rappresenta il nome della contea.
- “Year_ID” rappresenta l’anno di riferimento.
- “ID_Tumor” rappresenta il codice identificativo di un tumore.
- “Type_Tumor” rappresenta il nome di un tumore.
- “Sex_ID” rappresenta il codice identificativo del sesso.
- “Sex_name” rappresenta il genere.
- “Mortality” rappresenta il tasso di mortalità di un tumore in un anno specifico in una contea per un certo sesso per 100.000 abitanti.
- “Population” rappresenta la popolazione della contea in un certo anno per un certo sesso.

3.2 Ricognizione degli archivi

I dati estratti dal database sono sottoposti a un processo detto ricognizione degli archivi, volto a esaminare gli schemi del database e individuare nuove associazioni tra le entità, dipendenze funzionali precedentemente tralasciate oppure errori di vario tipo (dati mancanti, dati duplicati, inconsistenza).

Nel nostro caso sono state individuate le entità State, County, Sex, Tumor e Years con i relativi attributi. Inoltre, nel dataset sono presenti i tassi di mortalità e la popolazione. I tassi di mortalità dipendendo dalla contea, dall’anno, dal sesso e dal tumore. Pertanto, note queste informazioni posso determinare il tasso di mortalità. Per quanto riguarda la popolazione, questa dipende dalla contea, dall’anno e dal sesso. In un primo momento è stata individuata una sola entità contenente le informazioni sui tassi di mortalità e sulla popolazione. Un’analisi più attenta ci ha permesso di individuare due dipendenze funzionali:

1) County, Sex, Tumor, Years → Mortality
2) County, Sex, Years → Population

Mentre nel primo caso abbiamo una dipendenza funzionale dalla PK, quindi non genera anomalie, nel secondo caso abbiamo una dipendenza funzionale generata da un sottoinsieme della PK e pertanto porta ad anomalie.

È stato opportunamente realizzato un processo di normalizzazione che ha condotto alla separazione dell'entità iniziale in due entità separate Mortality e Population.

Per quanto riguarda le associazioni, sono state individuate le seguenti:

- Associazione tra l'entità State e l'entità County.
- Associazione tra l'entità Mortality e le entità County, Sex, Years, Tumor
- Associazione tra l'entità Population e le entità County, Sex, Years.

Al termine del processo di ricognizione e normalizzazione degli schemi, il database relazionale costruito a partire dal dataset è il seguente.

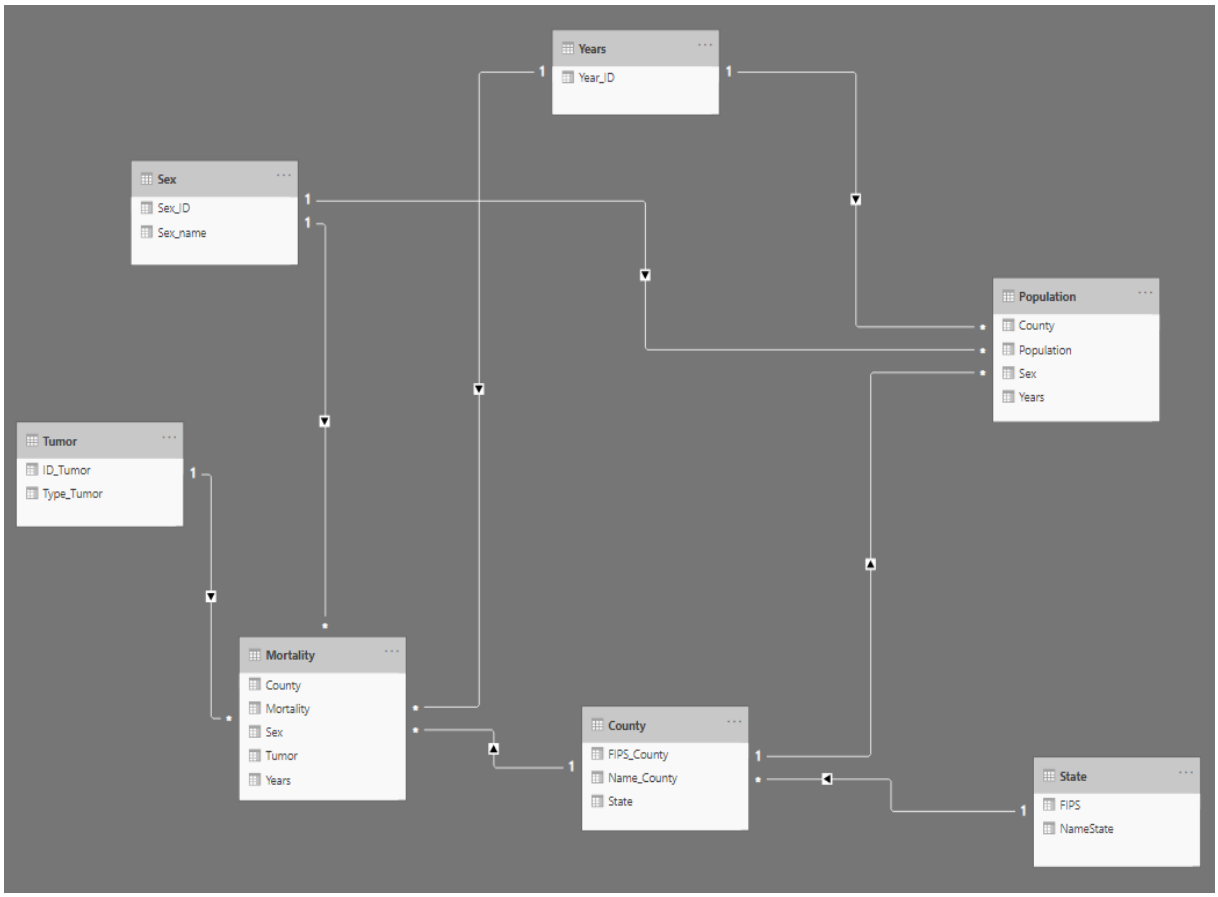

Figura 3.1 Modello Logico Relazionale del dataset iniziale

3.3 Estrazione

Attraverso uno script in Java i dati sono estratti dal dataset e sottoposti ad un controllo preliminare che mira ad individuare eventuali incoerenze nella struttura dello schema. Eventuali violazioni impediranno il caricamento del dataset.

3.4 Pulizia

In questa fase si procede a sanificare e riparare i dati contenuti nelle tuple laddove fosse possibile, e a rimuovere le tuple laddove non sia possibile ripararle. In particolare, le operazioni eseguite sono le seguenti:

- Rimozione di tuple vuote.
- Rimozione di tuple qualora uno o più campi fossero vuoti, in quanto non utili al fine dell'analisi.
- Lo standard CSV prevede che stringhe contenenti virgole debbano essere racchiuse tra doppi apici in modo tale da non interpretarle come separatori di campi; procediamo a verificare la presenza di virgole interne alla stringa e a rimuoverle.
- Conversione del campo mortality da tipo stringa a tipo numerico; verifica che il campo mortality contenga effettivamente un valore numerico maggiore di 0.
- Conversione del campo population da tipo stringa a tipo numerico; verifica che il campo population contenga effettivamente un valore numerico maggiore di 0.
- Verifica che il campo Year_ID sia composto da 4 cifre; se il controllo è superato, il campo è convertito da tipo stringa a tipo numerico; verifica che il campo Year_ID contenga un valore compreso tra 2000 e 2010.
- Per il campo NameState si procede a rimuovere eventuali spazi e segni di punteggiatura; verifica che la stringa sia effettivamente uno stato americano; in caso di esito positivo verifico la corretta corrispondenza tra il nome dello stato e l'identificativo associato.
- Per il campo Name_County si procede a rimuovere eventuali spazi e segni di punteggiatura; verifica che la contea appartenga allo stato

americano presente sulla stessa tupla; in caso di esito positivo verifico la corretta corrispondenza tra il nome della contea e l'identificativo associato.

- Per il campo Sex_name si procede a rimuovere eventuali spazi e segni di punteggiatura; verifica che il campo contenga le stringhe "Male" o "Female"; se l'esito è positivo si procede ad uniformare il campo Sex_ID assegnando a "Male" l'identificativo 1 e a "Female" l'identificativo 2.
- Per il campo Type_Tumor si procede a rimuovere eventuali spazi e segni di punteggiatura; verifica che il campo contenga uno dei 31 tumore scelti al fine dell'analisi. Se il tumore è valido si verifica in un primo momento la corretta validità del tumore per il sesso specificato sulla tupla; successivamente si uniforma l'identificativo attraverso un ID incrementale.

3.5 Trasformazione

Nella fase di trasformazione sono state eseguite le seguenti operazioni:

- Denormalizzazione: tipicamente le tabelle del DW sono denormalizzate. Pertanto, in questa fase si procede a denormalizzare varie tabelle presenti nel database relazionale. Infatti:
  - La tabella County e la tabella State vengono denormalizzate costituendo un'unica tabella Locality.
  - La tabella Mortality e la tabella Population vengono denormalizzate aggregando il tutto in una sola tabella Mortality.
- Matching: nel nostro DB relazionale non sono presenti in tabelle diverse campi uguali, pertanto in questa fase non abbiamo nulla da eseguire.
- Selezione: dopo aver denormalizzato le tabelle State e County si procede ad eliminare il campo FIPS (State) e il campo State(FK in County) utilizzando come PK della relazione il campo FIPS_County.
- Aggregazione:
  - I campi StateName e Name_County vengono aggregati in un solo campo Locality_Name.

- Viene aggiunto nella tabella Mortality un nuovo campo (calcolato)

  $$Deaths_{estimate} = \frac{(Mortality \times Population)}{100000}$$

Infine, per una maggiore comprensione i seguenti campi sono stati opportunamente rinominati:

- Il campo ID_Tumor e il campo Type_Tumor viene rinominato rispettivamente in Tumor_ID e Tumor_Name.
- Il campo Year_ID viene rinominato in Years_ID.
- Il campo Population viene rinominato in Population_Estimate.
- Il campo Mortality viene rinominato in Mortality_Rate.

Il processo di realizzazione dello schema a stella è stato:

1) Selezione del processo aziendale da analizzare;
2) Dichiarazione della sua grana (atomicità);
3) Scelta delle dimensioni;
4) Identificazione dei fatti numerici che popoleranno ogni riga della tabella dei fatti.

Nel nostro caso di studio le scelte adottate sono le seguenti:

1) Il processo aziendale individuato per l'analisi OLAP è stimare il numero dei decessi divisi per sesso legati alle varie forme tumorali nel corso dei vari anni nelle diverse contee americane.
2) La grana dichiarata è il singolo tasso di mortalità del tumore.
3) Per la realizzazione dello schema a stella sono state individuate quattro dimensioni (punti di vista), ciascuna collegata (tramite chiave esterna) alla tabella dei fatti. Le tabelle delle dimensioni sono: Locality, Sex, Years, Tumor.

4) Nella tabella dei fatti (Mortality) le misure oggetto di analisi sono: Mortality_Rate, Population_Estimate, Deaths_Estimate.

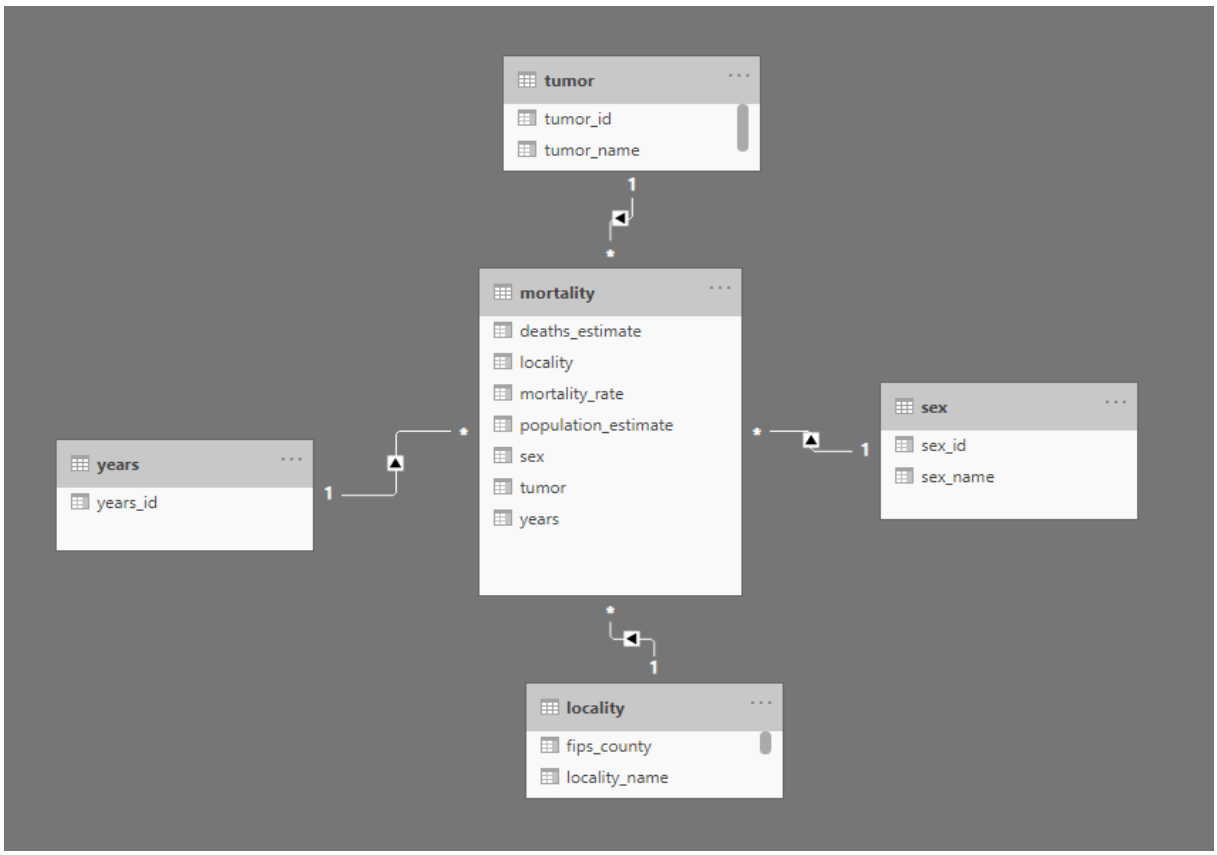

Figura 3.2 Schema a stella finale

Di seguito riportiamo le relazioni tra la tabella dei fatti e le tabelle delle dimensioni:

- Relazione 1-N tra il campo years_id della tabella “years” e il campo years della tabella “mortality”.
- Relazione 1-N tra il campo tumor_id della tabella “tumor” e il campo tumor della tabella “mortality”.
- Relazione 1-N tra il campo sex_id della tabella “sex” e il campo sex della tabella “mortality”.
- Relazione 1-N tra il campo fips_county della tabella “locality” e il campo locality della tabella “mortality”.

4 Tool utilizzato per le analisi OLAP

Il tool utilizzato per eseguire le analisi OLAP è stato Power BI. Power BI è una raccolta di servizi software, app e connettori che interagiscono per trasformare le origini dei dati non correlate in un insieme di informazioni coerenti, visivamente accattivanti e interattive. I dati potrebbero essere un foglio di calcolo di Excel o una raccolta di data warehouse basati sul cloud o ibridi locali. Power BI consente di connettersi facilmente alle origini dati, visualizzare e scoprire le informazioni importanti e condividerle con tutti gli utenti o con quelli necessari. Power BI è costituito da:

- Un'applicazione desktop Windows denominata Power BI Desktop.
- Un servizio SaaS online denominato servizio Power BI.
- App per dispositivi mobili Power BI per dispositivi Windows, iOS e Android.

Con Power BI Desktop, è possibile:

1) Connettersi ai dati, incluse più origini dati.
2) Eseguire il data shaping con query che creano modelli di dati utili e accattivanti.
3) Usare i modelli di dati per creare visualizzazioni e report.
4) Condividere i file di report che altri utenti possono usare, ampliare e condividere. È possibile condividere i file con estensione pbix di Power BI Desktop come qualsiasi altro file.

Power BI Desktop integra le tecnologie collaudate di Microsoft Query Engine, modellazione dei dati e visualizzazione. Analisti di dati e altri utenti possono creare raccolte di query, connessioni dati, modelli e report e condividerli facilmente. Grazie alla combinazione di Power BI Desktop con il servizio Power BI, è possibile modellare, creare, condividere ed estendere con maggiore facilità nuove informazioni dettagliate ricavate dai dati.

Power BI Desktop consente di connettersi a molti tipi diversi di dati, tra cui origini dati di base come un file di Microsoft Excel o servizi online.

Per connettersi ai dati, dalla barra multifunzione Home selezionare Get Data.

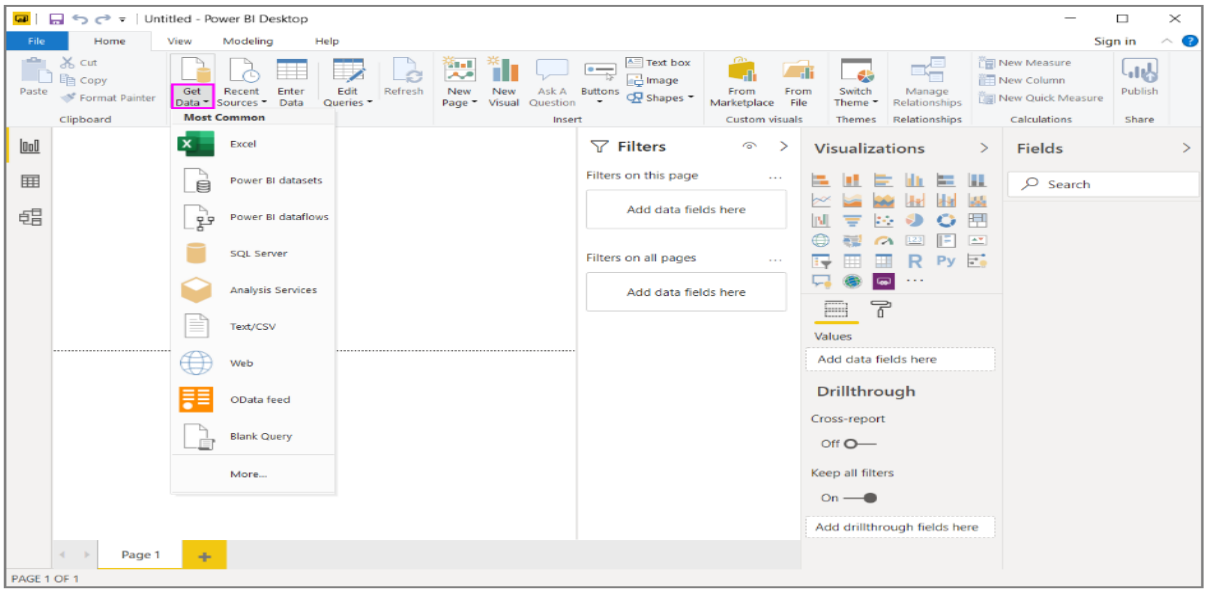

Figura 4.1 Interfaccia Power B.I. per connettersi a un'origine dati

Lungo il lato sinistro della schermata di Power BI Desktop sono presenti icone per le tre viste di Power BI Desktop: dall'alto verso il basso Report, Dati e Relazioni. Nella vista Report di Power BI Desktop è possibile creare visualizzazioni e report. La vista Report include sei aree principali:

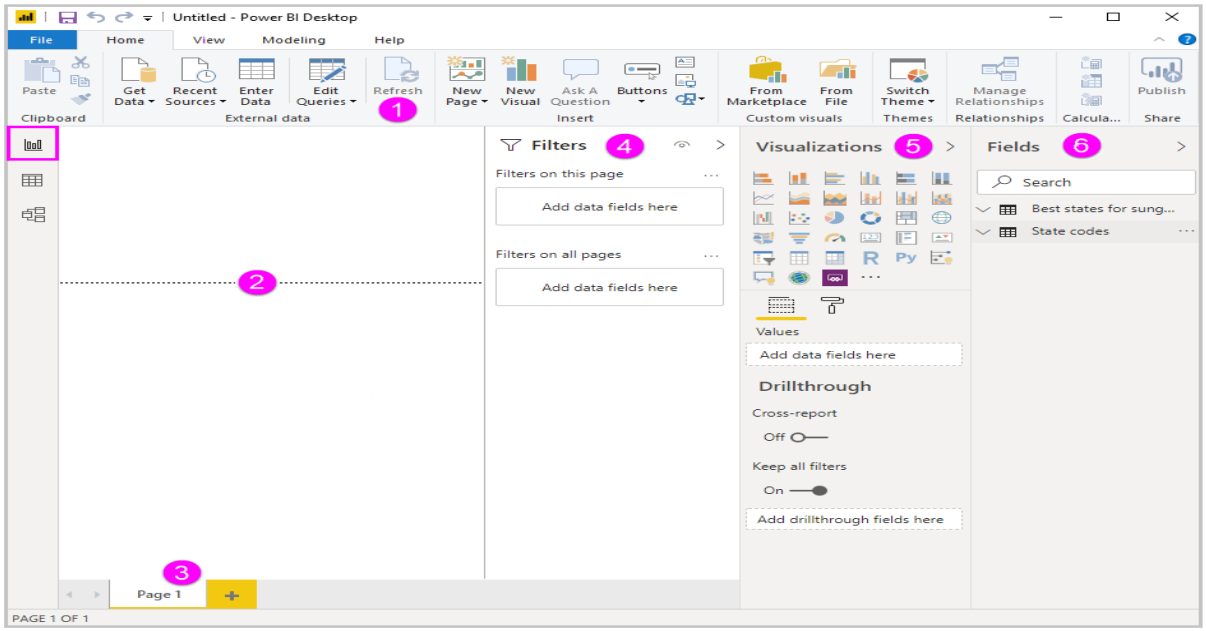

Figura 4.2 Sezione di Power B.I. per la creazione dei grafici

1) La barra multifunzione nella parte superiore della vista, che include le attività comuni associate ai report e alle visualizzazioni.

2) L'area di disegno nella parte centrale, dove vengono create e disposte le visualizzazioni.
3) La scheda delle pagine nella parte inferiore, che consente di selezionare o aggiungere pagina di report.
4) Il riquadro Filtri, in cui è possibile filtrare le visualizzazioni dei dati.
5) Il riquadro Visualizzazioni, in cui è possibile aggiungere, modificare o personalizzare le visualizzazioni e applicare il drill-through.
6) Il riquadro Campi, che mostra i campi disponibili nelle query. È possibile trascinare questi campi nell'area di disegno, nel riquadro Filtri o nel riquadro Visualizzazioni per creare o modificare le visualizzazioni.

La visualizzazione Modello mostra tutte le tabelle, le colonne e le relazioni presenti nel modello.

Selezionare l'icona Modello accanto al lato della finestra per attivare una visualizzazione del modello esistente.

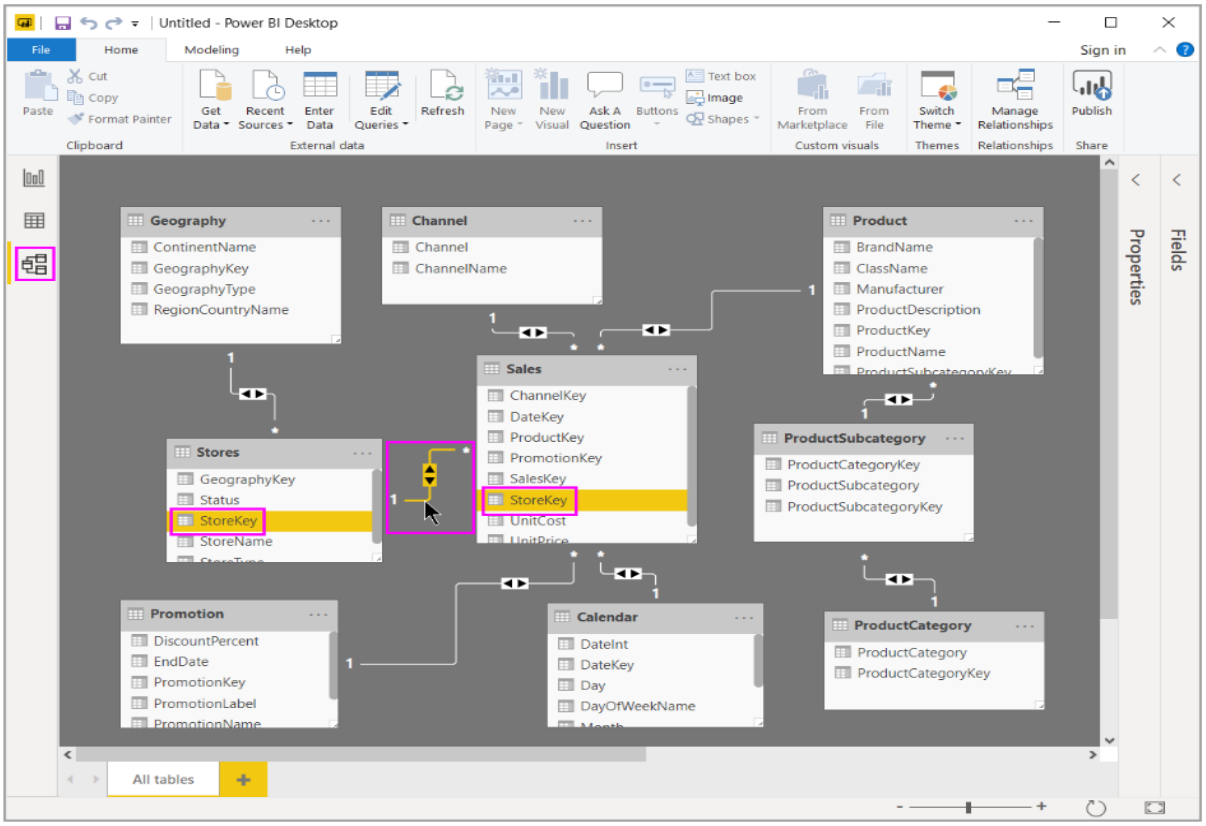

Figura 4.3 Sezione Modello che mostra le relazioni e le associazioni tra esse

5 Esperienza

Nella sezione Report del tool Power B.I. è possibile costruire grafici per creare business intelligence, quindi formulare query grafiche e completamente navigabili. Per realizzare correttamente query complesse è stato necessario definire le seguenti misure calcolate:

- Grouping_tumor_by_sex – che raggruppa i tumori in base al sesso in cui si manifestano.
- Grouping_county_by_state_name – che permette di raggruppare le contee in base allo stato di appartenenza.
- Deaths_estimate_year_2000 – che permette di sommare i decessi stimati avvenuti nell'anno 2000.
- Deaths_estimate_year_2010 - che permette di sommare i decessi stimati avvenuti nell'anno 2010.
- Mortality_rate_year_2000 che permette di raggruppare i tassi di mortalità relativi all'anno 2000.
- Mortality_rate_year_2010 – che permette di raggruppare i tassi di mortalità relativi all'anno 2010.
- Sum_population_estimate - che permette di determinare il totale della popolazione negli anni presi in analisi.
- Living_population_estimate – che permette di determinare il totale della popolazione vivente negli anni presi in analisi.
- Percentage_change_deaths_estimate_year_2000_2010 – che permette di determinare la variazione percentuale relativa ai decessi stimati nell'intervallo 2000 – 2010.
- Percentage_change_mortality_rate_year_2000_2010 – che permette di determinare la variazione percentuale relativa ai tassi di mortalità nell'intervallo 2000 – 2010.

Vengono riportate alcune delle interrogazioni che è possibile effettuare:

- A quanto ammontano i decessi stimati divisi per sesso in un dato anno, in un dato stato, per un data forma tumorale?

- Come si sono evoluti i tassi di mortalità nelle varie contee nel corso degli anni per un data forma tumorale, per un dato sesso e per un dato stato?
- A quanto ammontano i decessi stimati divisi per sesso nel corso degli anni rispetto ad una data forma tumorale?
- Come sono concentrati i decessi stimati nelle varie contee in un dato anno? Dato uno stato e un anno, quale è la contea che ha registrato il più alto numero di decessi?
- Data una forma tumorale e una contea, in che modo i tassi di mortalità hanno subito una variazione percentuale nel decennio 2000-2010?
- Dato uno stato e un anno, in che misura ogni forma tumorale ha inciso sul totale dei decessi stimati? E per un dato sesso?
- Dopo aver raggruppato le forme tumorali in base al sesso in cui si manifestano, in che modo nel corso degli anni è variato il numero dei decessi stimati? In che misura il numero dei decessi stimati per ogni categoria ha subito una variazione percentuale nel decennio 2000-2010?
- Dato uno stato e dato un sesso, a quanto ammonta la percentuale dei decessi stimati rispetto alla popolazione stimata in un dato anno? A quanto ammonta la popolazione viva stimata? E la popolazione deceduta stimata?

Di seguito le dashboard realizzate per l'esecuzione delle analisi OLAP.

- Dashboard #1

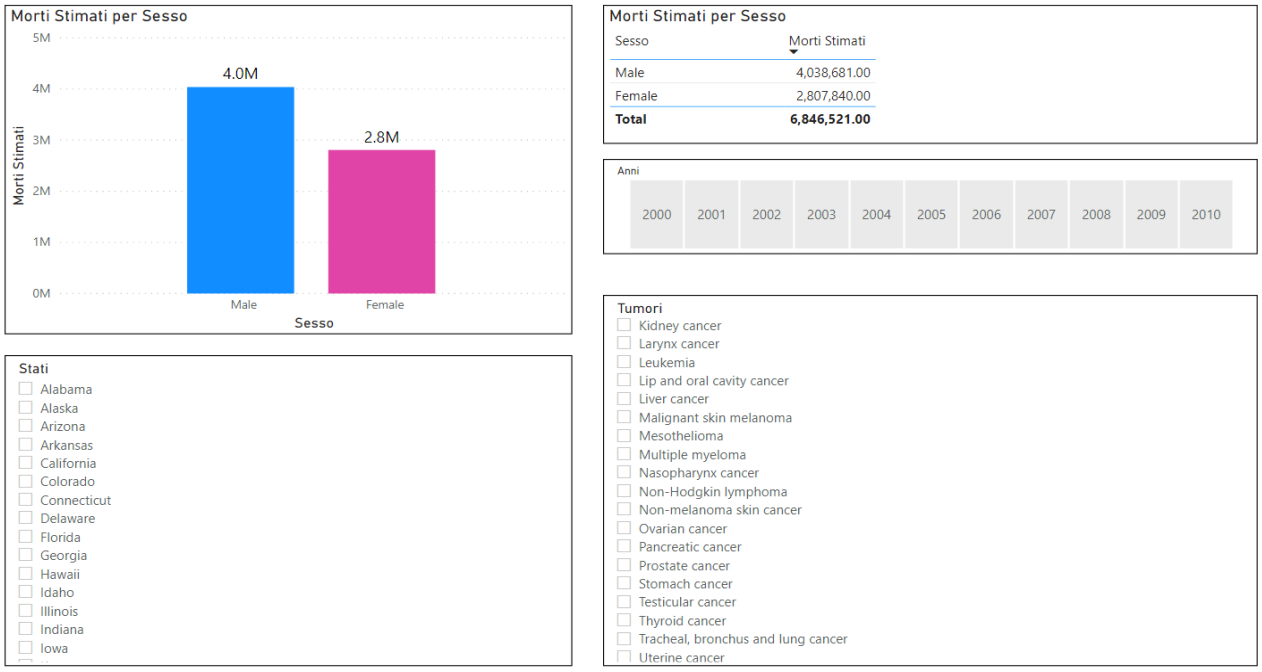

Figura 5.1 Morti stimati per sesso

Questa dashboard ci permette di confrontare i decessi stimati divisi per sesso. Inoltre, è possibile filtrare i risultati selezionando uno o più stati, uno o più forme tumorali ed uno o più anni.

Come si può vedere abbiamo modo di confrontare l'ammontare dei decessi stimati divisi per sesso nello stato dell'Alabama nell'anno 2000 per il tumore al pancreas.

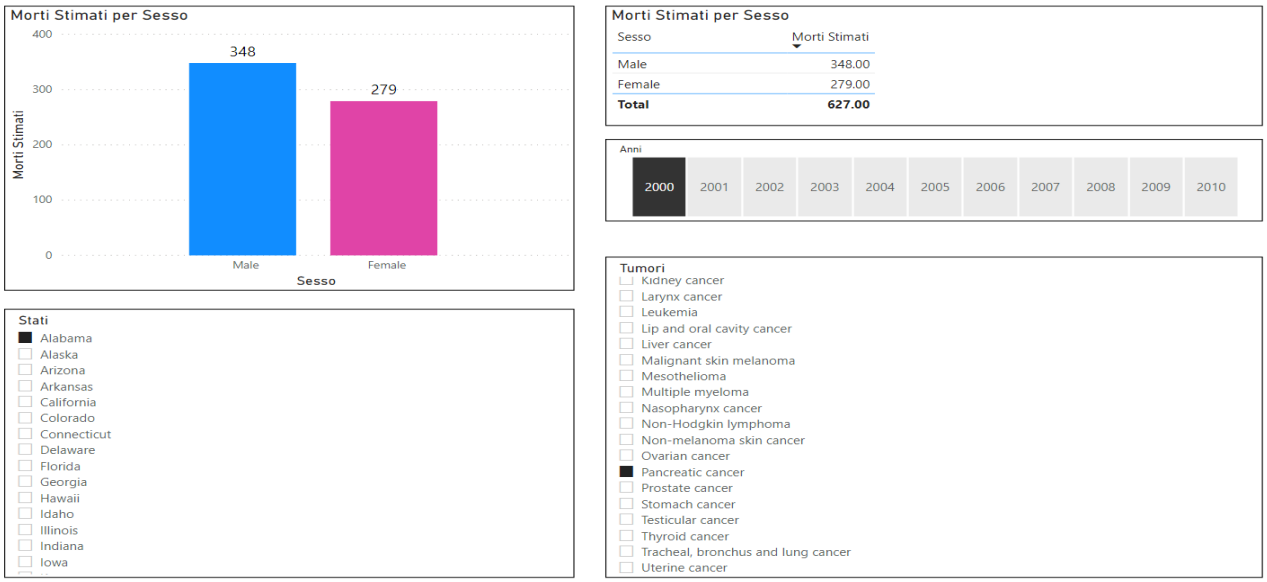

Figura 5.2 Morti stimati per sesso nello stato dell'Alabama nell'anno 2000 per il tumore al pancreas

- Dashboard #2

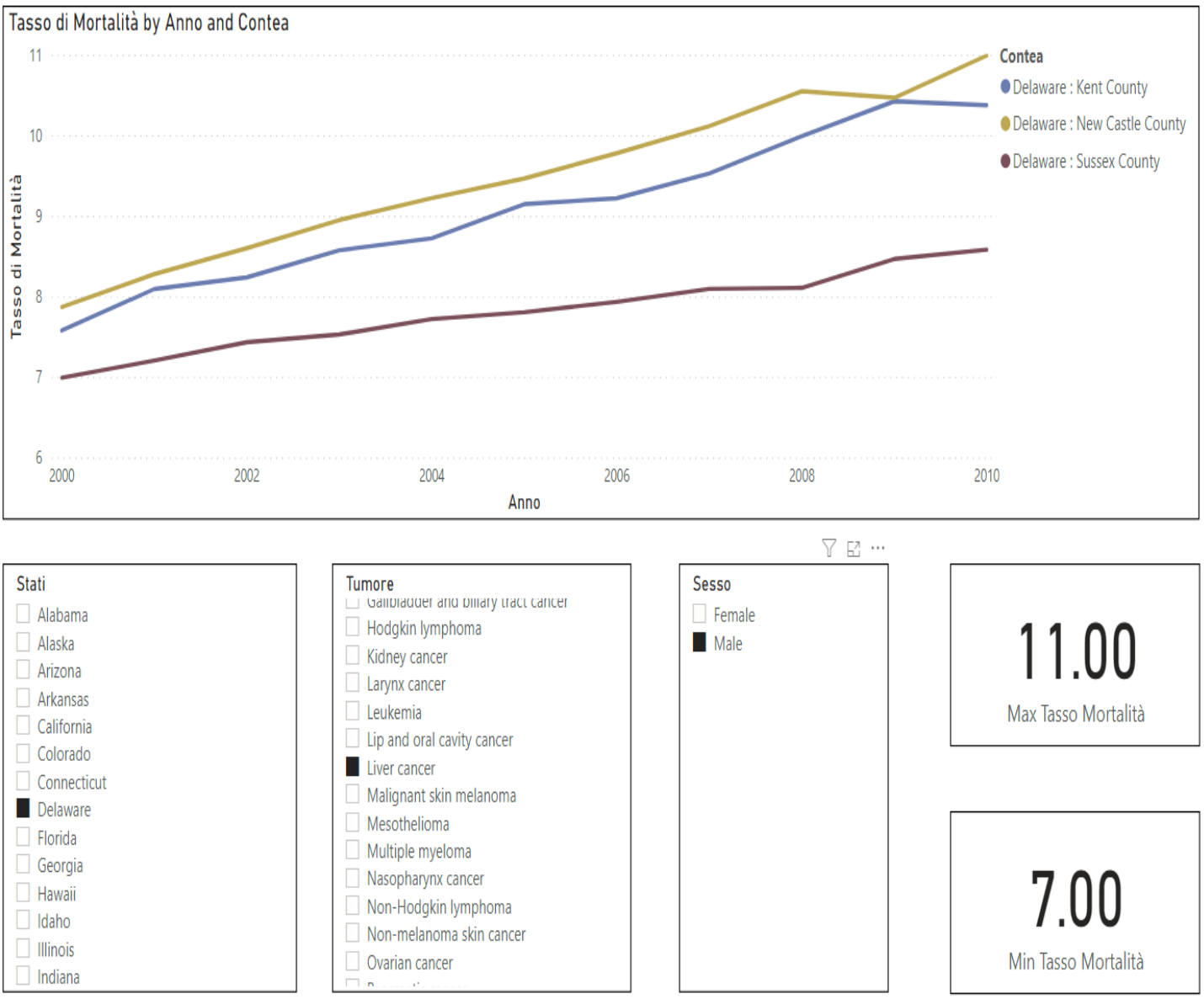

Figura 5.3 Andamento tassi di mortalità del tumore al fegato nelle contee del Delaware per il sesso maschile

La dashboard ci permette di selezionare uno stato, una forma tumorale e un sesso, mostrandoci come i tassi di mortalità associati al tumore selezionato hanno subito variazione nel corso degli anni nelle singole contee dello stato. Inoltre, è possibile mostrare il tasso di mortalità più alto e più basso registrato nello stato selezionato. Nell'esempio sopra mostriamo come il tasso di mortalità del tumore al fegato è variato negli uomini nello stato del Delaware.

- Dashboard #3

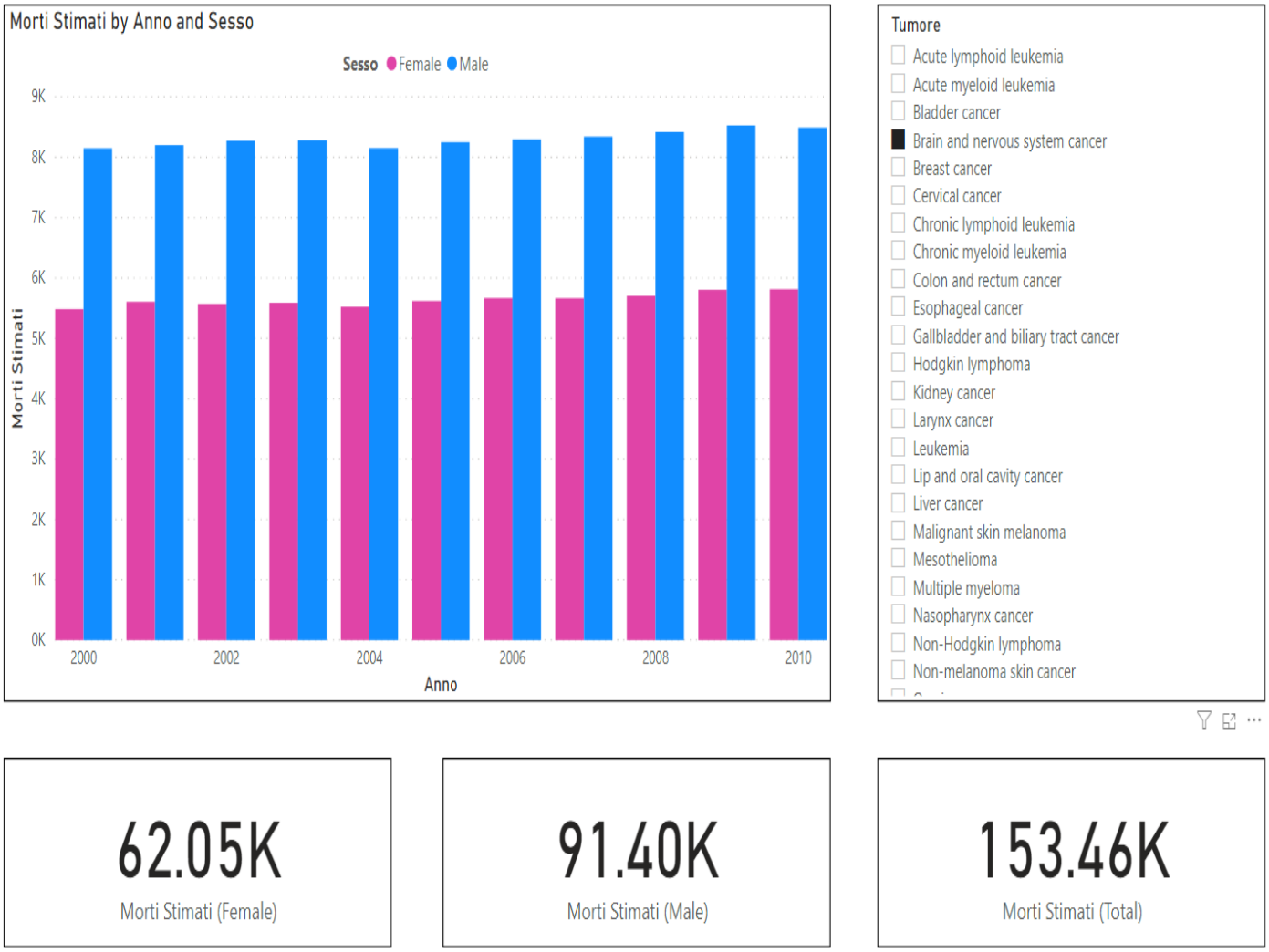

Figura 5.4 Morti stimati nel corso degli anni divisi per sesso per il tumore al cervello e sistema nervoso

La dashboard ci permette di selezionare una specifica forma tumorale e ci mostra come nel corso degli anni è variato il numero di decessi stimati divisi per sesso per la forma tumorale selezionata. Inoltre, ci mostra il numero totale di decessi stimati e il numero totale di decessi divisi per sesso per quella data forma tumorale.

Ad esempio, nella dashboard mostriamo l'ammontare dei decessi causati dal tumore al cervello e al sistema nervoso divisi per sesso nell'arco del decennio 2000-2010. Infine, mostriamo il totale dei decessi per lo stesso tumore divisi per sesso e il totale globale.

- Dashboard #4

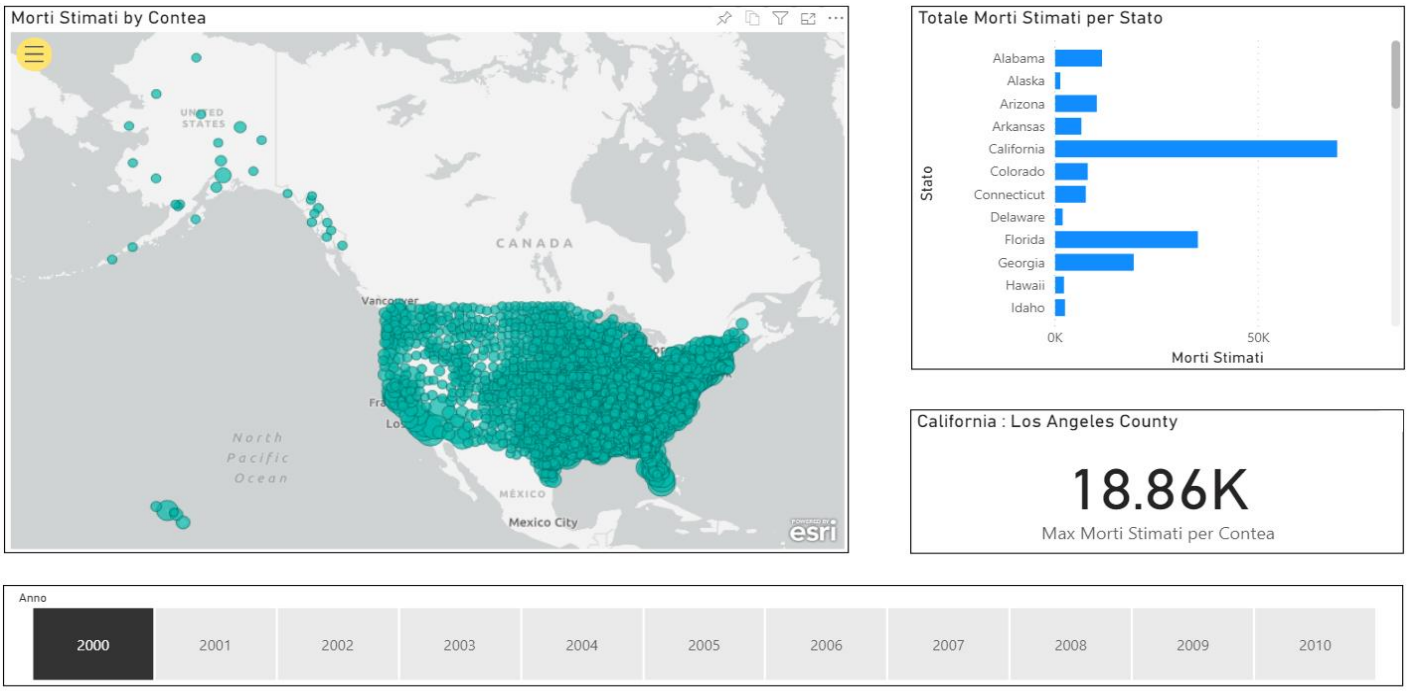

Figura 5.5 Morti stimati per contea nell'anno 2000

La dashboard ci mostra per un dato anno il numero totale dei decessi stimati nelle singole contee. Ci viene mostrano il numero totale di decessi avvenuti in ogni stato ed infine la contea che ha registrato il più alto numero di decessi.

Inoltre, selezionato un anno e selezionato uno stato è possibile visualizzare la contea che ha registrato il più alto numero di decessi stimati.

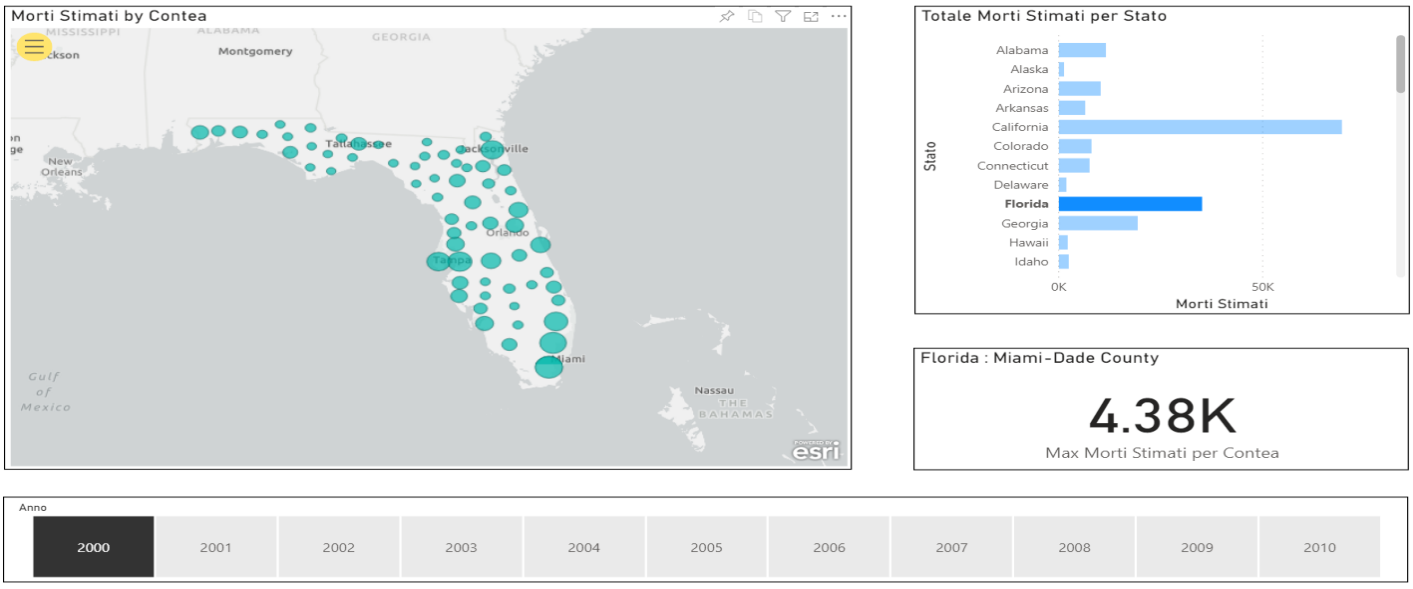

Figura 5.6 Morti stimati nelle contee della Florida

- Dashboard #5

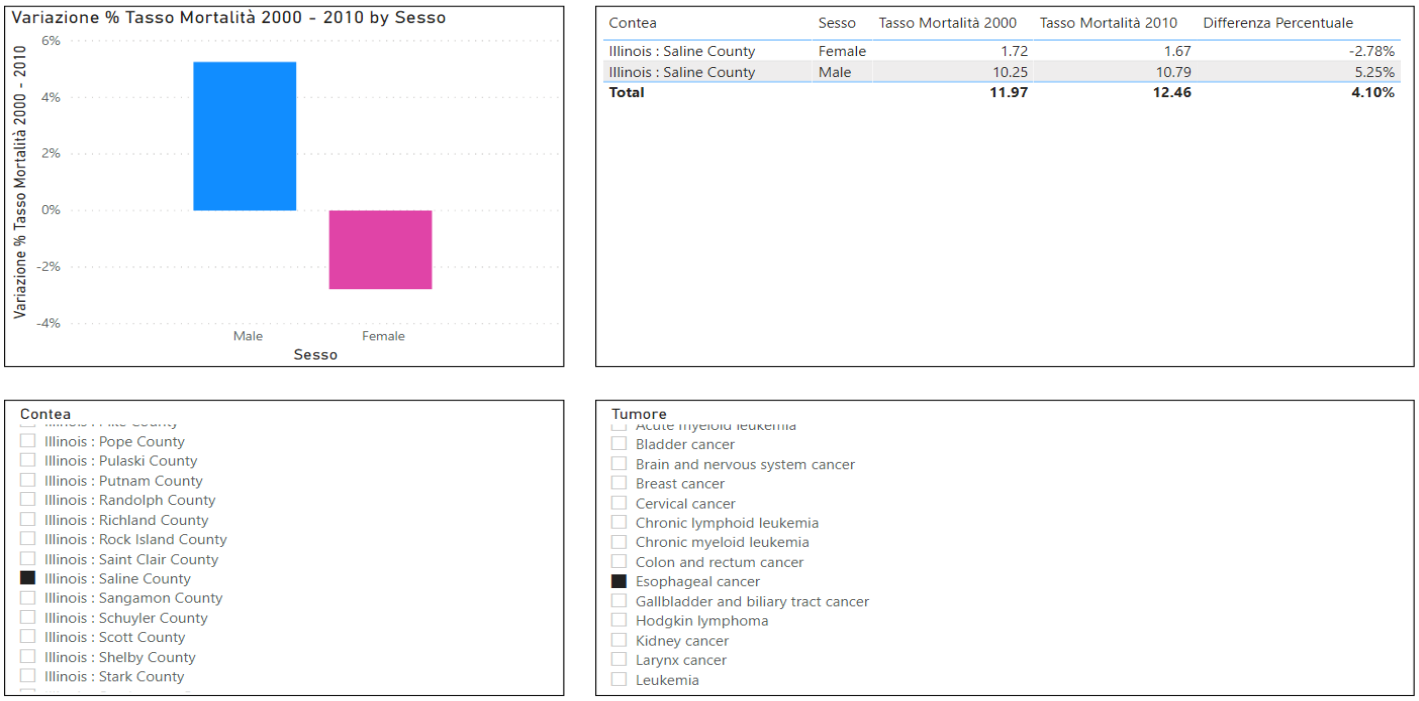

Figura 5.7 Variazione % tassi di mortalità del tumore all'esofago divisi per sesso nella contea Saline County dello stato dell'Illinois

La dashboard ci mostra in che misura i tassi di mortalità di un certo tumore in una certa contea hanno subito una variazione percentuale nell'arco di tempo 2000-2010 divisi per sesso.

In particolare, nell'esempio riportato ci viene mostrato in che misura i tassi di mortalità del tumore all'esofago sono variati nella contea di Saline nello stato dell'Illinois divisi per sesso. Possiamo quindi notare come per gli uomini dall'anno 2000 all'anno 2010 ci sia stato un incremento del tasso di mortalità del 5.25%. Per le donne possiamo notare una riduzione del tasso di mortalità del 2.78% rispetto all'anno 2000.

- Dashboard #6

La dashboard seguente mostra l'incidenza di ciascun tumore sul totale dei decessi stimati divisi per sesso e globalmente. In particolare, è possibile filtrare i risultati in base allo stato e all'anno di interesse.

Ad esempio, la dashboard riportato mostra la percentuale di incidenza di ogni tumore sui decessi registrati nello stato dell'Alabama nell'anno 2000. È

possibile notare come nel caso degli uomini, delle donne ed in generale il tumore che registra il maggior numero di morti è quello dell'apparato respiratorio. Segue per gli uomini il tumore alla prostata e per le donne il tumore al seno.

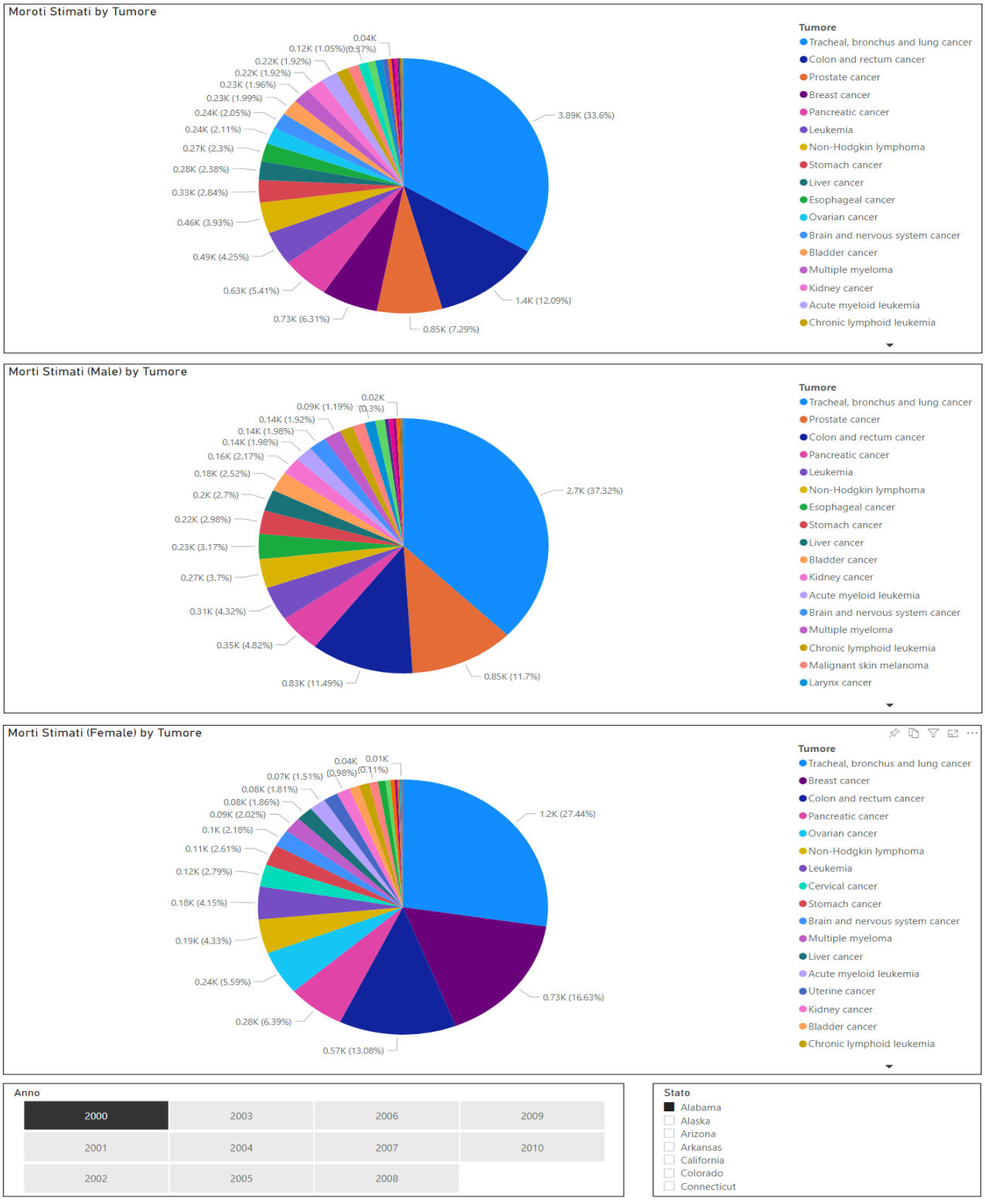

Figura 5.8 Morti stimati in termine percentuale per tumore nell'anno 2000 nello stato dell'Alabama divisi per sesso

- Dashboard #7

Abbiamo raggruppato le forme tumorali in tre categorie: tumori femminili, tumori maschili e tumori comuni ad entrambi. La dashboard pertanto ci mostra come sia variato nel corso degli anni il numero di decessi per ogni categoria. Inoltre, ci mostra la variazione percentuale del numero dei decessi legati ad ogni categoria nell'arco di tempo compreso tra il 2000 e il 2010. Notiamo come i tumori femminili abbiamo avuto un incremento in termine di decessi di un valore inferiore al 5%; per i tumori comuni e per i tumori maschili notiamo una riduzione di circa il 5% per il primo e oltre il 10% per il secondo.

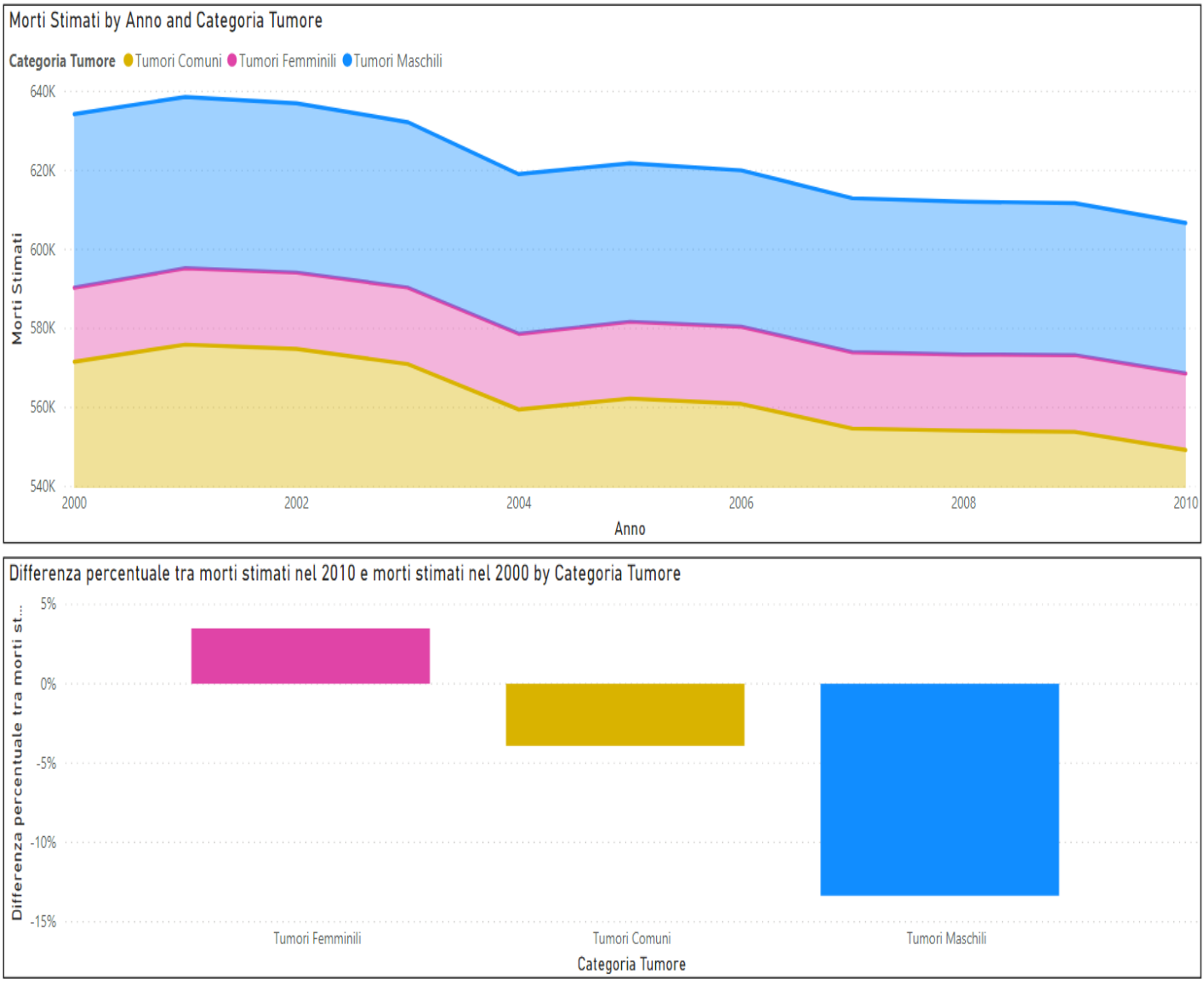

Figura 5.9 Morti stimate per anno e categoria di tumore; Variazione percentuale dei tassi di mortalità per ogni categoria

Infine, abbiamo la possibilità di selezionare una particolare categoria per visualizzare l'andamento dei decessi stimati nel corso degli anni.

Ad esempio, selezionata la categoria "Tumori Femminili" possiamo notare come il minor numero di decessi si sia manifestato nell'anno 2000 con un numero di morti inferiore a 19000, mentre il maggior numero di decessi è stato registrato nell'anno 2006 con un numero di morti superiore a 19500.

In generale, il numero di decessi è compreso tra 18500 circa e non oltre 19600.

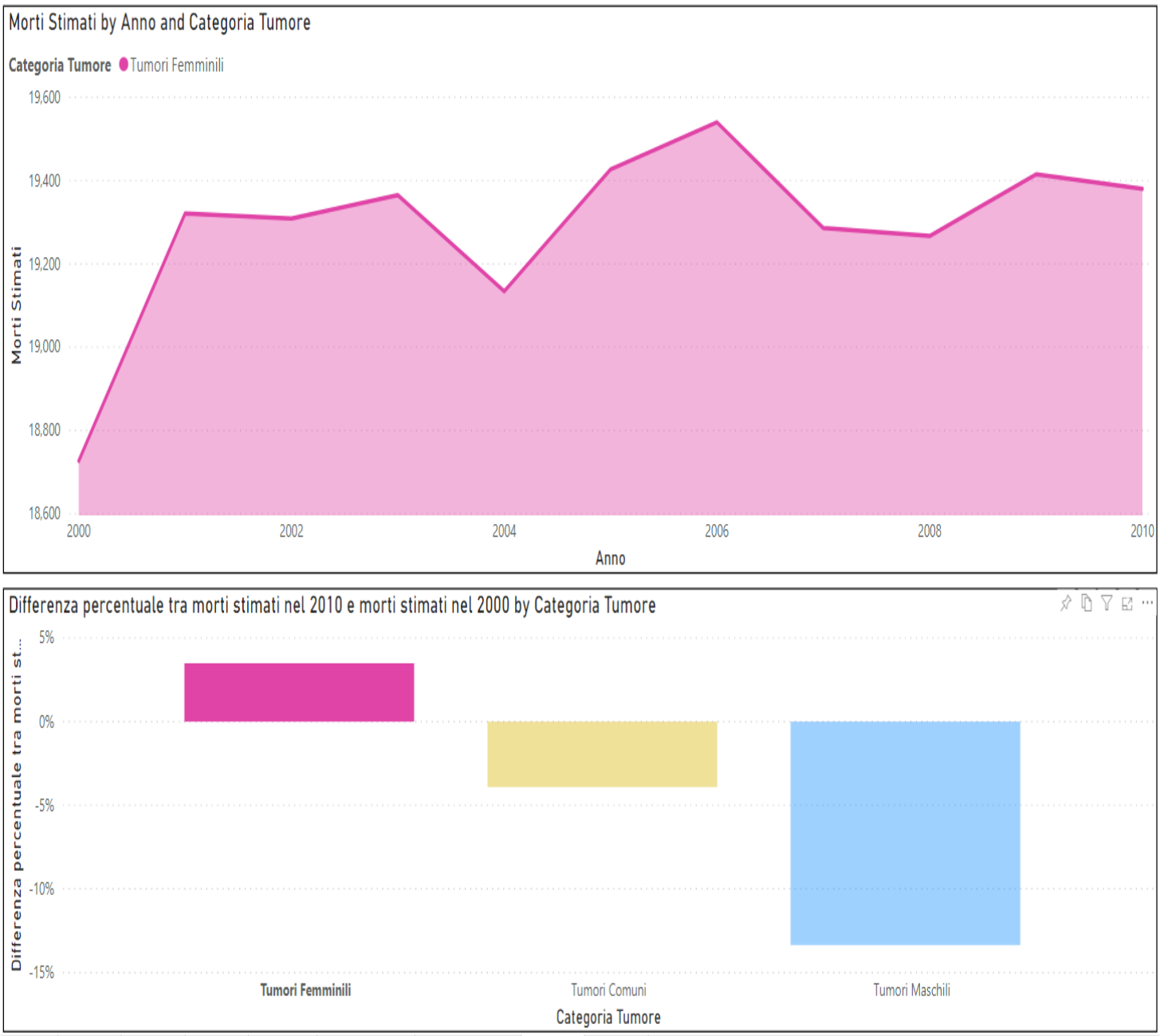

Figura 5.10 Andamento decessi registrati nel corso degli anni per i tumori femminili

- Dashboard #8

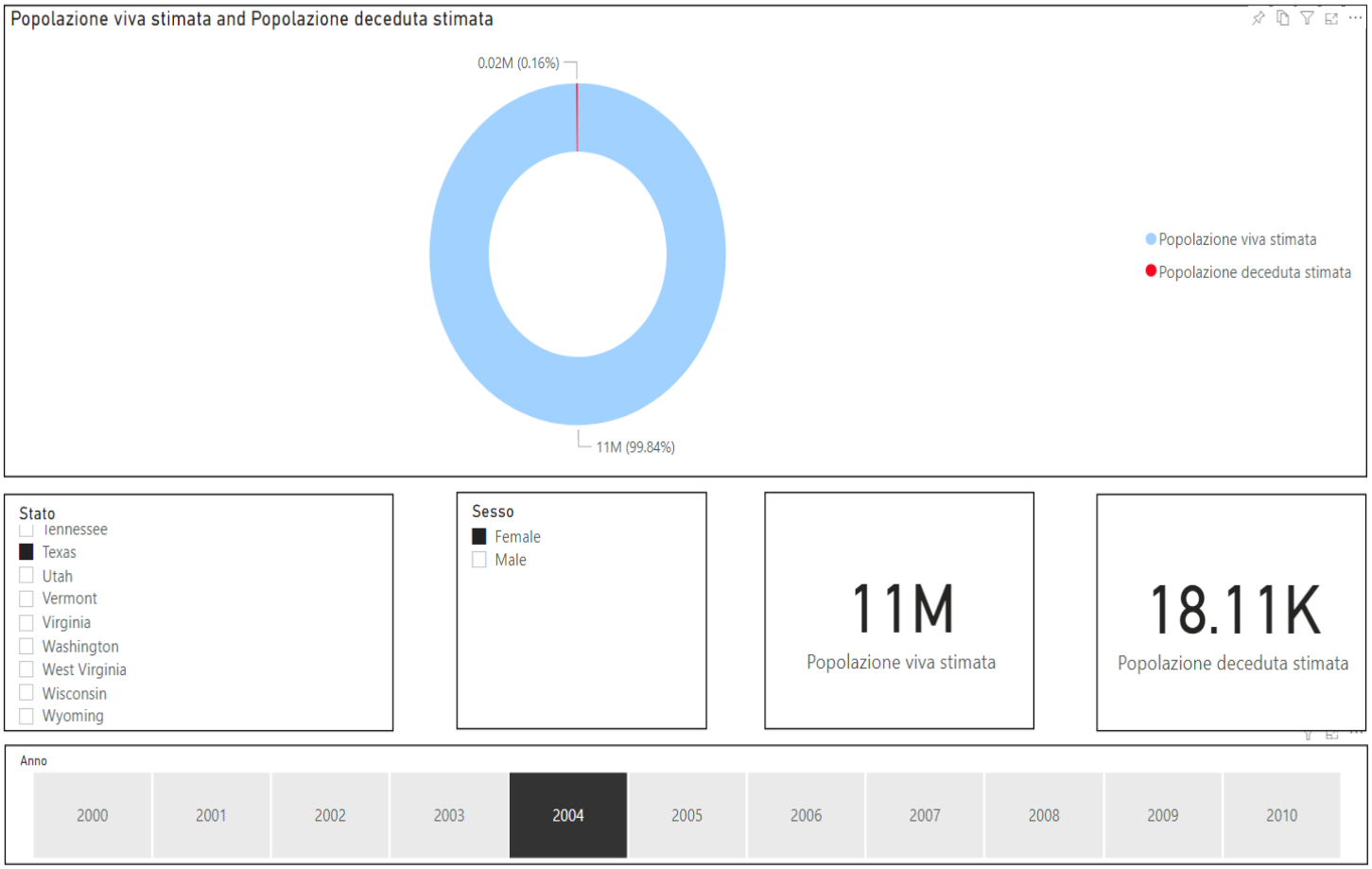

Figura 5.11 Popolazione femminile viva e deceduta stimata nel Texas nell'anno 2004

La dashboard ci mostra in termini percentuali il numero di decessi registrati ed il numero di popolazione viva stimata. In particolare, possiamo filtrare i risultati in base allo stato di interesse, al sesso di interesse e all'anno. Inoltre, ci mostrerà il totale della popolazione viva stimata in quello stato per quel sesso in quell'anno e il totale di decessi registrati in base allo stesso criterio. È possibile filtrare i risultati combinando vari criteri su ciascuna categoria.

Ad esempio, la dashboard ci mostra come nello stato del Texas nell'anno 2004 per il sesso maschile si è registrato un numero di decessi pari a 18000 circa mentre la popolazione viva stimati è pari ad 11 milioni circa.

In termini percentuali notiamo come sulla popolazione femminile totale presente nel Texas nel 2004 solo 0,16% è deceduto a causa di una forma tumorale.

6 Conclusioni

In questo articolo, è stata fornita una panoramica generale sul lavoro svolto dall' IHME da cui sono stati reperiti i dati oggetti di analisi. Dopo una breve introduzione su cosa sia la business intelligence, sono state fornite le definizioni di data warehouse e di OLAP, evidenziando le principali differenze rispettivamente con il database relazionale e l'OLTP.

È stata effettuata inoltre una ricerca nella letteratura mirata ad evidenziare i seguenti aspetti: architettura di un data warehouse, rappresentazione multidimensionale dei dati, realizzazione di un data warehouse e metodologia generale per la progettazione di un data warehouse.

Successivamente, è stato progettato un data warehouse che permette di studiare accuratamente come si sono evolute nel corso degli anni le diverse forme tumorali, utilizzando come tool Power BI.

Per quanto riguarda gli sviluppi futuri del data warehouse si potrebbe pensare di estendere l'analisi su intervalli di tempo via via più grandi, ampliando l'insieme di malattie oggetto di analisi ed eventualmente estendo la ricerca su stati o continenti differenti.

7 Riferimenti

Golfarelli, Matteo; Rizzi, S. (2006). *Data Warehouse Teoria e Pratica della Progettazione*.

Institute for Health Metrics and Evaluation. (n.d.). Https://En.Wikipedia.Org/Wiki/Main_Page. Retrieved August 3, 2020, from https://en.wikipedia.org/wiki/Institute_for_Health_Metrics_and_Evaluation

MONTEBELLI, M. R. (2016). *Usa, il cancro diventa la prima causa di morte in 22 stati*. Https://Www.Repubblica.It/. https://www.repubblica.it/oncologia/news/2016/09/27/news/usa_tumore_diventa_prima_causa_di_morte_in_22_stati-148590223/

Paolo Atzeni, Stefano Ceri, Stefano Paraboschi, & Riccardo Torlone. (2002). *Basi di dati. Modelli e linguaggi di interrogazione*.