

UNIVERSITÀ
DELLA CALABRIA

Dipartimento di Matematica e Informatica
Corso di Laurea Magistrale in Informatica

Progettazione di un Data Warehouse per analisi OLAP relativamente agli attacchi terroristici dal 1970 al 2016

Sintesi:

Il sistema progettato, raccolte le informazioni sugli attacchi terroristici dal 1970 al 2016, permette di interrogare un data mart appositamente sviluppato, in grado di supportare interrogazioni di tipo OLAP. L'obiettivo principale è riuscire a mostrare ed evidenziare i vari trend terroristici relativamente ad esempio al numero di attacchi, al numero di feriti ma alla tipologia di arma più usata o alla nazione più soggetta ad attacchi.

Docente

Prof. G. Terracina

Studente

Antonio Fortino
193451

Obiettivo:

L'obiettivo principale è quello di costruire un DataWarehouse completo seguendo quelle che sono le fasi progettuali. L'approccio utilizzato è "source oriented" il che significa la fase di selezione, analisi e trasformazione della sorgente ha avuto un ruolo molto importante in quanto è stato necessario comprenderne al meglio la qualità e il dominio. In particolare tutte la progettazione è stata definita puntando a quello che è il fatto di interesse scelto, ovvero il verificarsi proprio di un attacco terroristico. Il risultato finale è stato un data mart capace di supportare analisi di tipo OLAP, le quali sono state realizzate attraverso l'utilizzo del software Tableau e individuano eventuali trend terroristici.

Descrizione della sorgente operazionale:

La sorgente è stata scaricata in formato Excel, e poi convertita in formato CSV, dal sito <http://www.start.umd.edu/gtd/>, e contiene le informazioni relative agli attacchi terroristici verificatisi dal 1970 al 2016, per un totale di circa 170.000 eventi registrati.

Giacché il numero di attributi presenti nel file sorgente sono più di 120, di seguito elencherò solo quelli selezionati in quanto ritenuti più rilevanti, successivamente inseriti nello schema del livello riconciliato.

| Incident | | |
|-------------|---------------------|--|
| Nome | Tipo | Descrizione |
| Eventid | Numerico | Id dell'evento |
| Iyear | Numerico | L'anno in cui si è verificato l'evento |
| Imonth | Numerico | Il mese in cui si è verificato l'evento |
| Iday | Numerico | Il giorno in cui si è verificato l'evento |
| Approxdate | Testo | Data approssimativa, nel caso in cui non è presente la data esatta |
| Country_txt | Testo | Stato in cui si è verificato l'evento |
| Region_txt | Testo, 12 valori | Continente in cui si è verificato l'evento (es. Eastern Europe) |

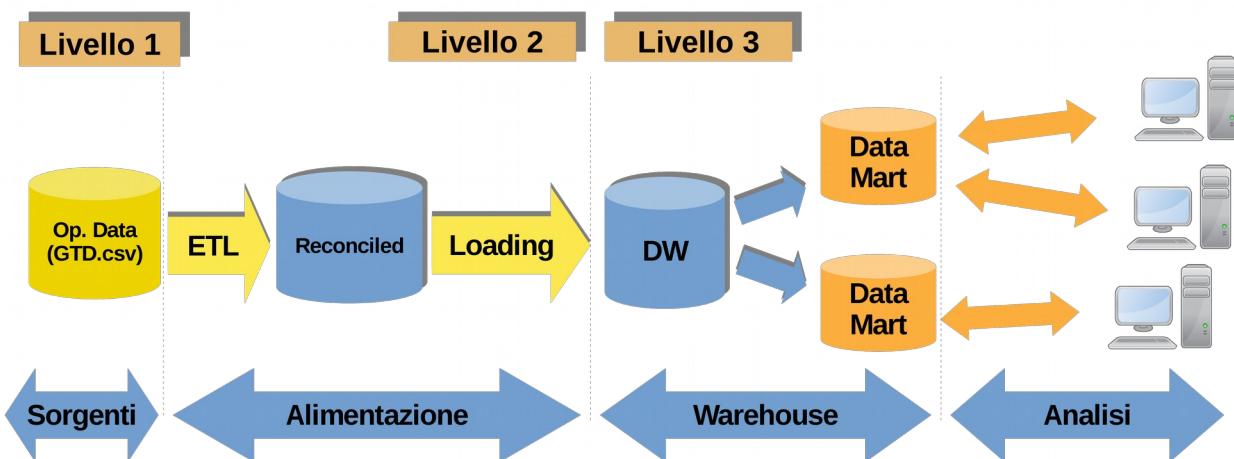
| Incident | | |
|------------------|--------------------|--|
| Nome | Tipo | Descrizione |
| Provstate | Testo | Provincia in cui si è verificato l'evento |
| City | Testo | La città |
| Latitude | Numerico | |
| Longitude | Numerico | |
| Crit1 | Booleano | Criterio di inclusione dell'evento : “Politico, economico, religioso o sociale” |
| Crit2 | Booleano | Criterio di inclusione dell'evento : “Intento di coercizione, intimidamento ” |
| Crit3 | Booleano | Criterio di inclusione dell'evento : “Violazione delle leggi umanitarie internazionali” |
| Doubtterr | Booleano | Indica se l'evento è stato inserito ma non è sicuro che rispetti i criteri di inclusione |
| Success | Booleano | Indica se l'attacco è stato completato |
| Suicide | Booleano | Indica se l'attacco è di tipo suicida |
| attacktype1_txt | Testo, 9 valori | Tipologia di attacco (es. Assassinio) |
| targtype1_txt | Testo | Tipologia della vittima |
| targsubtype1_txt | Testo | Tipologia della vittima |
| Corp1 | Testo | Nome della vittima/entità |
| Target1 | Testo | Indica la persona specifica o l'edificio esatto |
| natlty1_txt | Testo | Nazionalità |
| Gname | Testo | Gruppo di appartenenza degli attentatori |

| Incident | | |
|--------------------|----------------------|---|
| Nome | Tipo | Descrizione |
| Motive | Testo | |
| Nperps | Numerico | Numero degli attentatori |
| Nperpscav | Numerico | Numero degli attentatori catturati |
| Claimed | Booleano | Indica se l'attacco è stato rivendicato |
| claimmode_txt | Testo | Indica il tipo di rivendicazione |
| weaptype1_txt | Testo | Indica il tipo di arma usata |
| weapsubtype1_txt | Testo | Indica il tipo di arma usata |
| Nkill | Numerico | Numero di morti fra le vittime |
| Nkillter | Numerico | Numero di attentatori morti |
| Nwound | Numerico | Numero di feriti fra le vittime |
| Nwoundte | Numerico | Numero di attentatori feriti |
| propertyextent_txt | Testo | Indica l'eventuale range in dollari di danni agli oggetti causati dall'attacco |
| Ishostkid | Booleano | Indica se gli ostaggi sono stati rapiti |
| nhostkid | Booleano | Indica il numero di ostaggi rapiti |
| hostkidoutcome_txt | Testo | Indica come o se gli ostaggi sono stati salvati |
| INT_LOG | Enumeratore {0,1,-9} | Indica se è un attacco internazionale, comparando il luogo dell'attacco con il luogo di appartenenza del gruppo |
| INT_IDEO | Enumeratore {0,1,-9} | Indica se è un attacco internazionale, comparando l'ideologia dell'attacco con quella del gruppo |
| INT_MISC | Enumeratore {0,1,-9} | Indica se è un attacco internazionale, comparando il luogo dell'attacco con la nazionalità del gruppo |
| INT_ANY | Enumeratore {0,1,-9} | Indica se è un attacco internazionale, relativamente ai parametri precedenti |

Architettura del Data Warehouse:

L'architettura che si è deciso di adottare per questo sistema è conosciuta come Architettura a tre livelli , che differisce dalle altre per l'aggiunta di un livello intermedio , detto di dati riconciliati. Anche se il numero di sorgenti operazionali nel nostro caso si ferma ad uno e quindi non è stato necessario integrare più sorgenti, si è deciso di adottare questa specifica architettura in quanto il nuovo livello introdotto permette di materializzare i dati operazionali, separando la procedura di alimentazione del Data Warehouse e del data mart dal delicato processo di ripulitura e trasformazione dei dati sorgente. Questo livello perciò è servito soprattutto per definire le dipendenze funzionali nei dati.

Nella seguente figura è mostrata in sintesi l'architettura adottata:

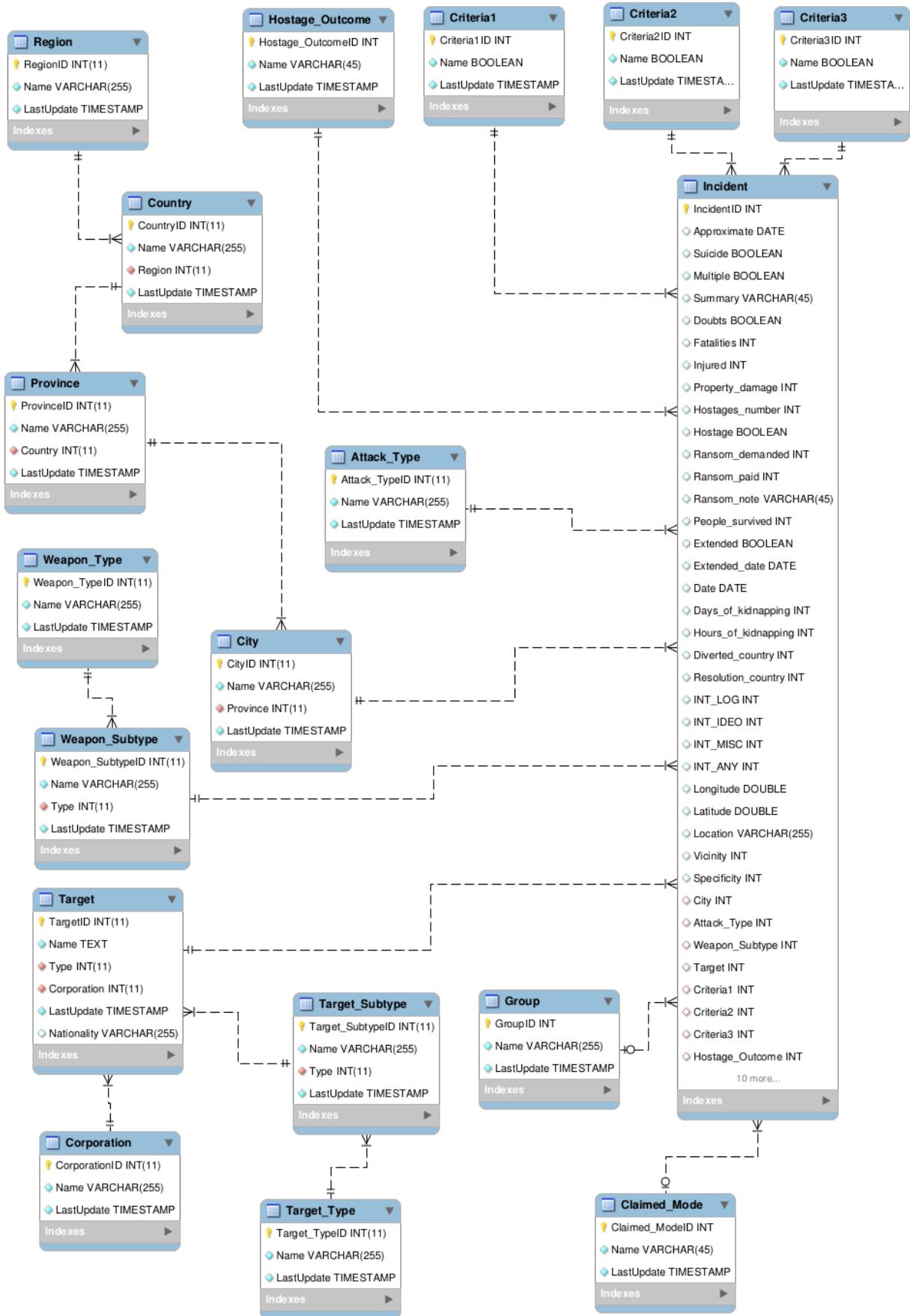


Come si può notare i dati sorgente grezzi vengono innanzitutto estratti staticamente e poi trasformati opportunamente così da ottenere una migliore qualità degli stessi. Segue la fase di caricamento dei dati nel DW che avviene attraverso la modalità di refresh, in quanto questa modalità viene normalmente abbinata con l'estrazione statica ed inoltre perché si assume che tutte le eventuali e successive modifiche/inserimenti verranno effettuate direttamente sulla base di dati del livello riconciliato.

Progettazione del database riconciliato:

Lo schema del database dei dati riconciliati è stato definito e sviluppato a valle di due attività il cui scopo è quello di acquisire un'approfondita conoscenza della sorgente di dati. La prima attività consiste in un'analisi dettagliata del dominio applicativo conosciuta come cognizione, la seconda invece è conosciuta come normalizzazione in quanto, questa attività, ha permesso di ricostruire e arricchire tutte le dipendenze funzionali tra i dati.

Si è ottenuto così il seguente schema entità relazioni:



Progettazione concettuale:

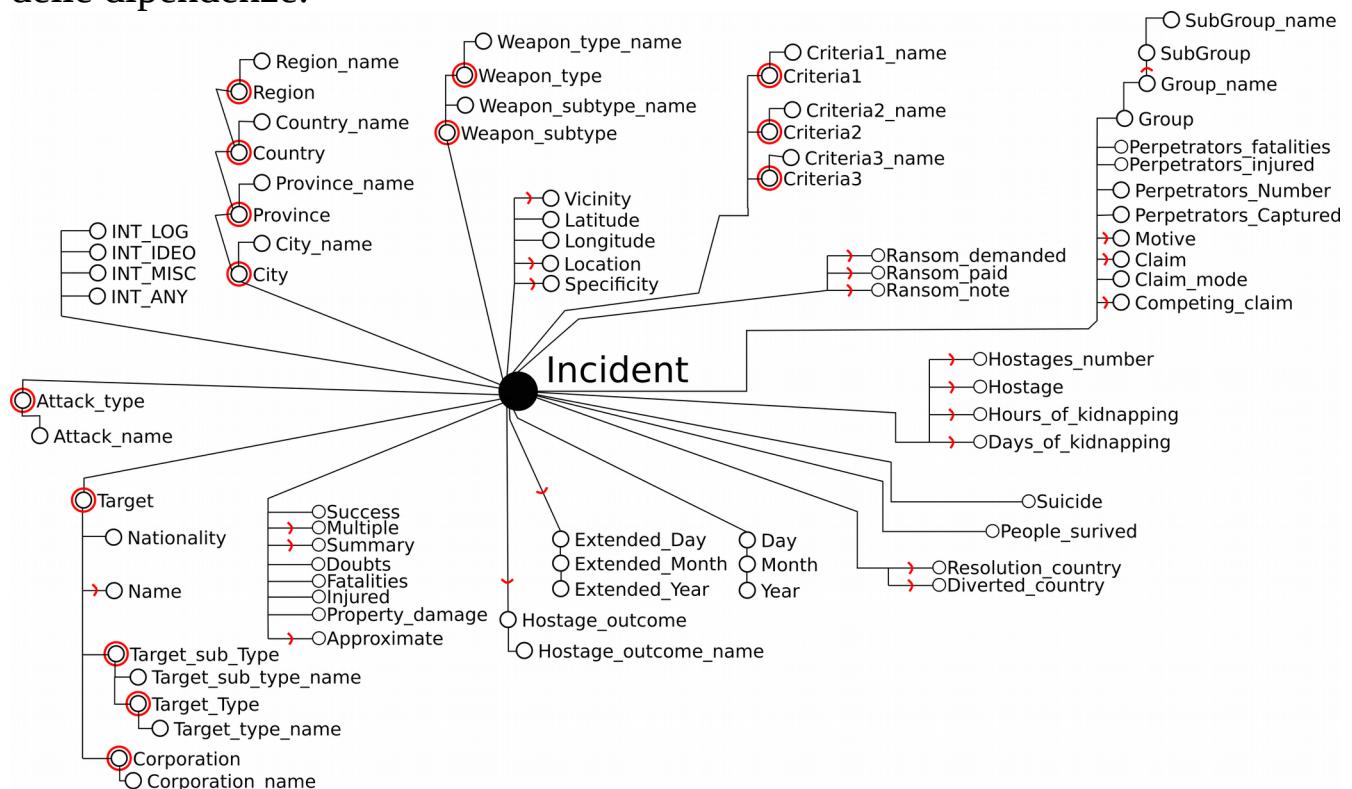
La tecnica per la progettazione concettuale di un data mart a partire dalle sorgenti operazionali consiste nel seguire determinati passaggi.

Per cominciare bisogna definire il fatto di interesse, e cioè il concetto di interesse primario sul quale verranno effettuate tutte le future analisi, per questo progetto si è scelto di concentrarsi direttamente sul verificarsi del attacco terroristico, rinominato per semplicità un Incidente.

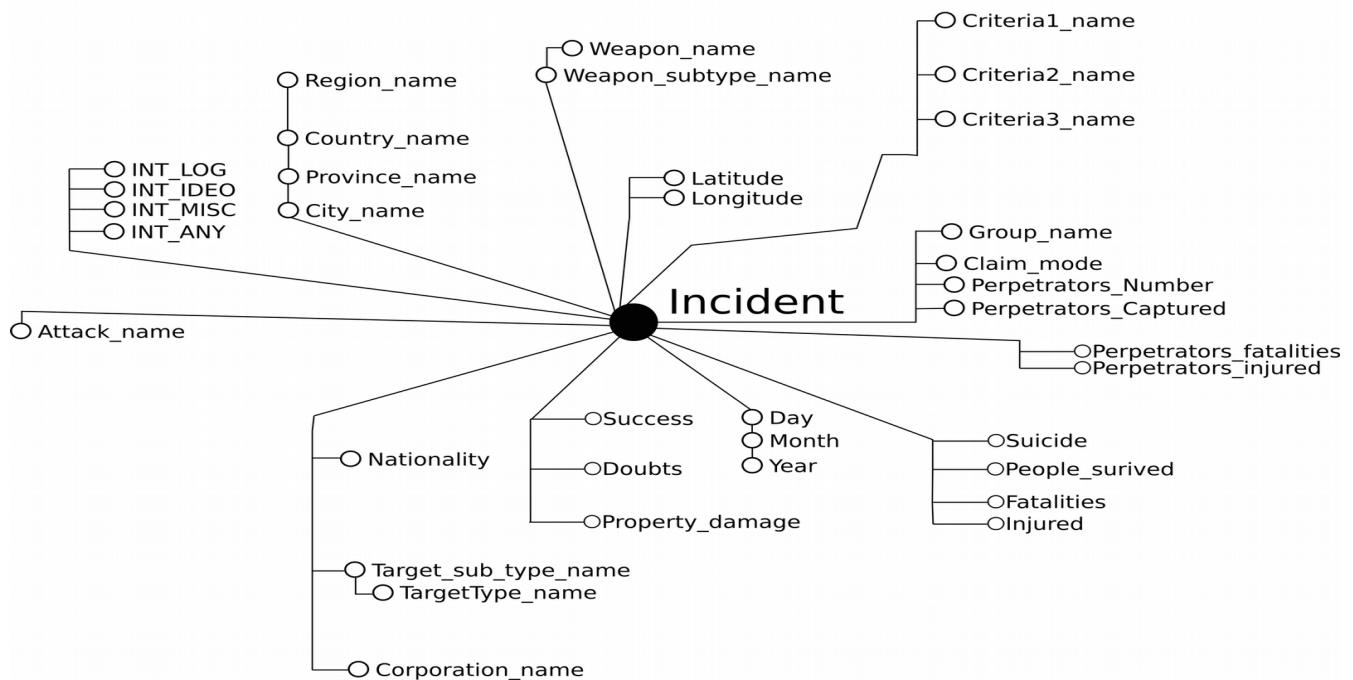
Albero degli attributi:

Continuiamo costruendo quello che è l'albero degli attributi, attraverso il quale si delimita l'area di interesse dello schema di fatto eliminando dapprima gli attributi irrilevanti ai fini dell'analisi, eventualmente modificare alcune dipendenze tra attributi ed infine definire misure e dimensioni.

Di seguito sono mostrate le varie operazione di pruning, grafting e modifica delle dipendenze:

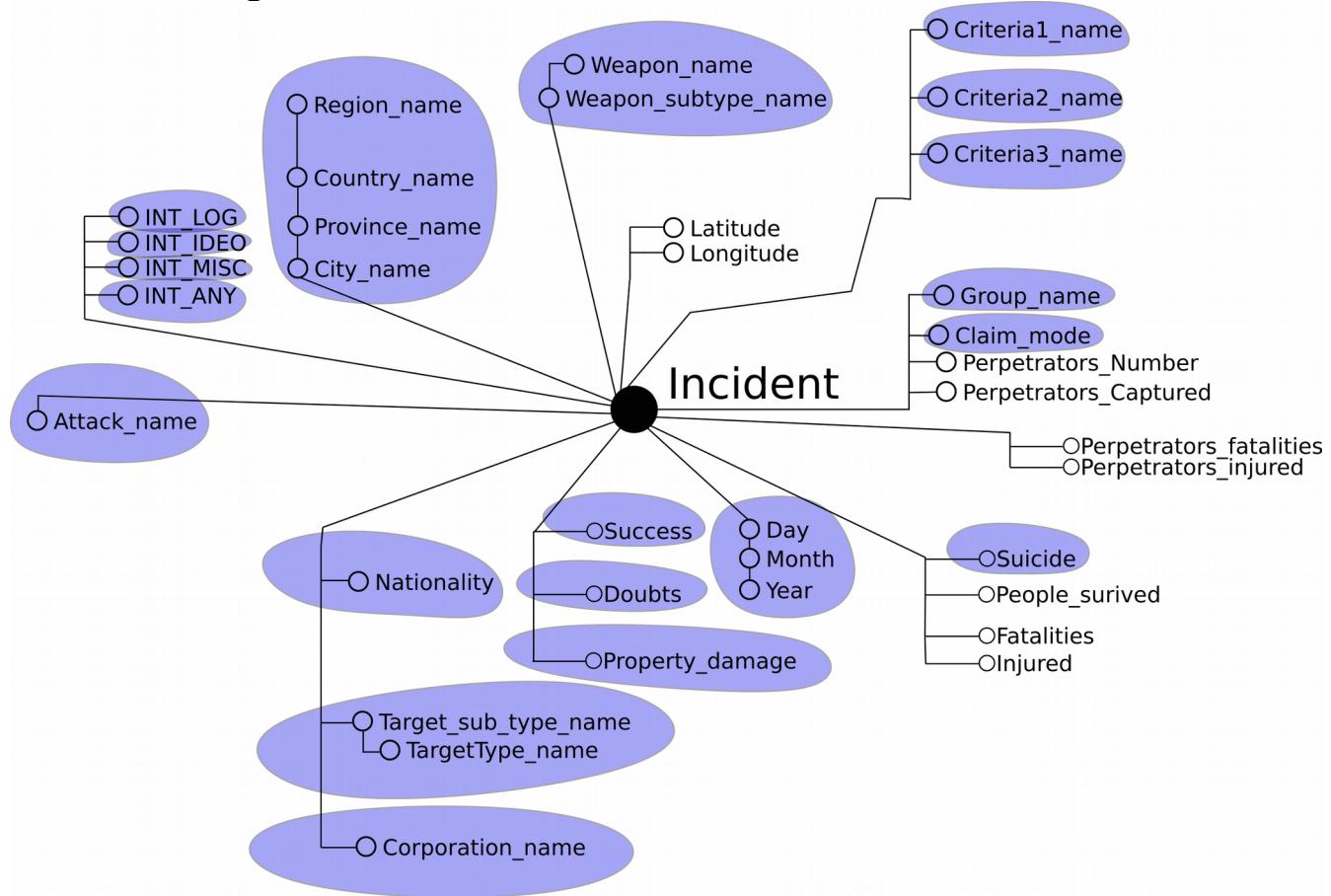


L'albero degli attributi conclusivo è il seguente:



Identificazione delle dimensioni:

Per determinare in che modo gli eventi potranno essere aggregati bisogna definire quelle che sono le dimensioni associate al nostro fatto di interesse, mostrate di seguito:



In questo progetto lo schema risultante è di tipo transazionale: dato che tutti gli attributi che identificano l'entità del fatto compaiono come dimensioni.

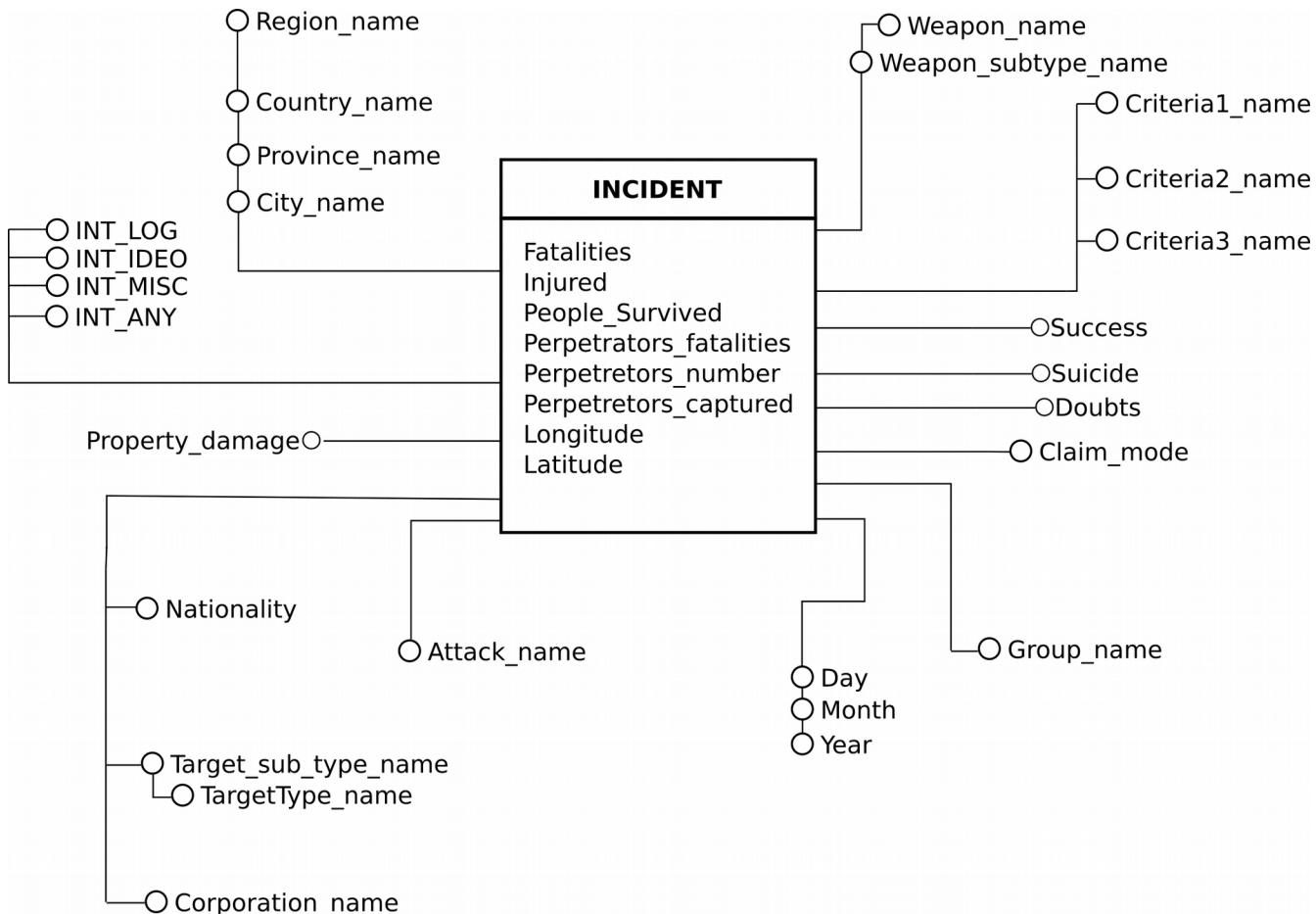
Identificazione delle misure:

Di seguito sono invece riportate le misure scelte e le varie funzioni di aggregazione, scelte sempre tra gli attributi direttamente collegati al fatto:

- SUM(Fatalities)
- SUM(Injured)
- AVG(Longitude)
- AVG(Latitude)
- SUM(People survived)
- SUM(Perpetrators fatalities)
- SUM(Perpetrators number)
- SUM(Perpetrator captured)

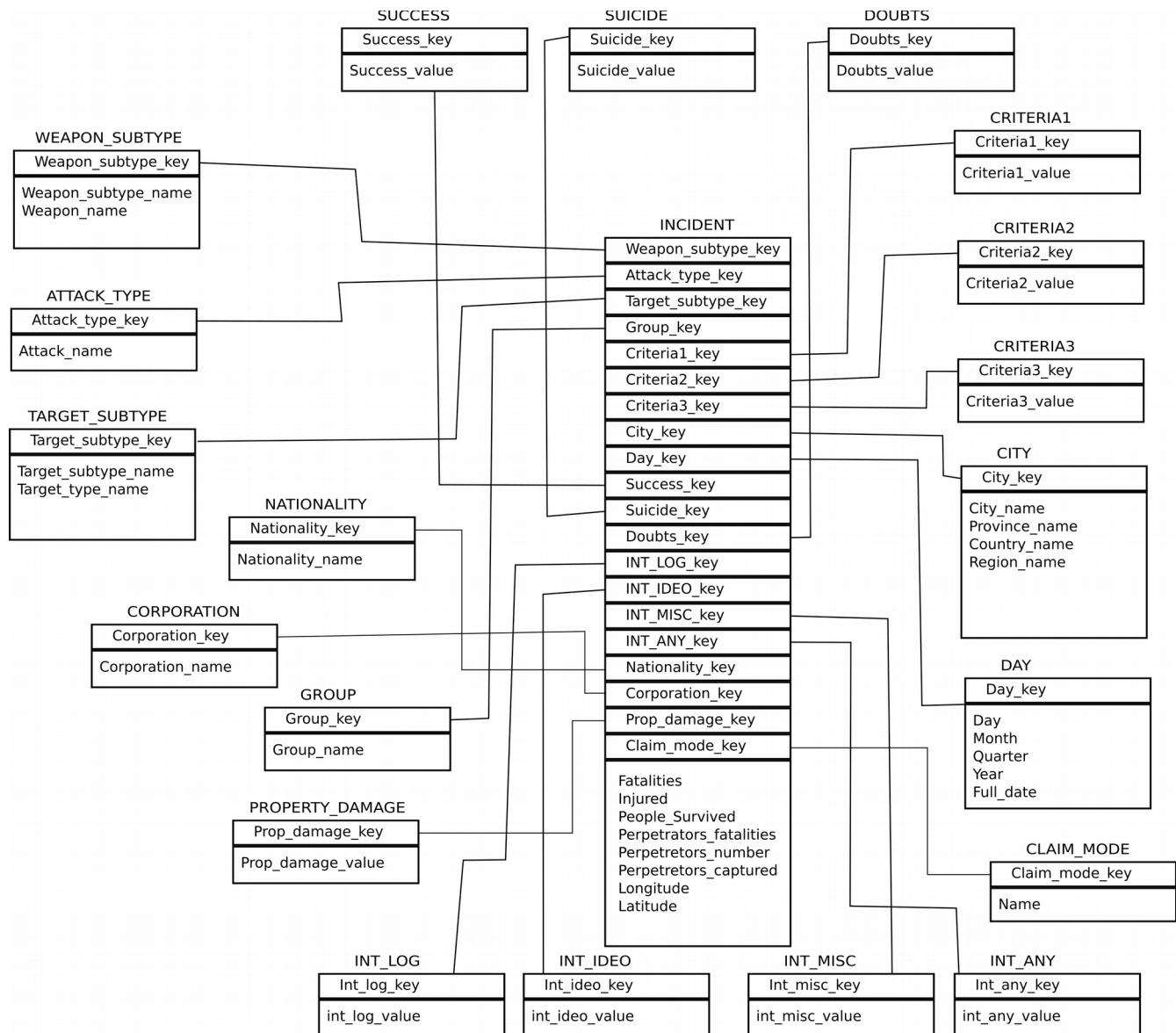
Costruzione del Fact Schema:

Procediamo adesso alla costruzione dello schema del fatto (fact schema), traducendo direttamente l'albero degli attributi includendo dimensioni e misure, come mostrato di seguito:



Progettazione dello Star Schema:

Di seguito è mostrato lo star schema che non è altro che la traduzione in termini di schema entità/relazioni dello schema di fatto: modellando inizialmente la fact table, la quale contiene tutte le misure e gli attributi descrittivi direttamente collegati al fatto, e per ogni gerarchia viene creata una dimension table che ne contiene tutti gli attributi.



Progettazione dell'Alimentazione:

In questa fase della progettazione vengono definite le procedure necessarie a caricare all'interno del data mart i dati provenienti dalle sorgenti operazionali, avendo però scelto un'architettura a tre livelli questo processo si suddivide ulteriormente in due fasi: la prima definisce le operazioni ETL dalle sorgenti operazionali al livello riconciliato, che relativamente al caricamento si è deciso di adottare la tecnica del caricamento statico, insieme alle procedure di pulizia dei dati le quali, in questo progetto includono anche tecniche di verifica e correzione dei valori basate sui dizionari (vedi Città, Province e Stati).

La seconda fase invece definisce le procedure di caricamento e aggiornamento dei dati dal livello riconciliato al livello del data mart.

Dato che il tempo è un fattore fondamentale nei sistemi di data warehousing, si è deciso di adottare diverse tecniche per gestire la dinamicità delle gerarchie presenti nello schema del fatto:

Per il caricamento dei dati si è scelto di utilizzare la tecnica basata sulle marche temporali, in modo tale da riuscire a contrassegnare i campi modificati/inseriti(ETL update e insert) rispetto all'ultima esecuzione del processo di estrazione, la quale avviene con una frequenza di aggiornamento prestabilità.

Per l'aggiornamento delle tabelle delle dimensioni invece di è deciso di adottare diverse tecniche in base alla tipologia di gerarchia dinamica utilizzata: le gerarchie di tipo 1 non prevedono di preservare la storicitizzazione, quindi l'aggiornamento avviene semplicemente sovrascrivendo la tupla modificata.

Le gerarchie di tipo 3 invece prevede di preservare la storicitizzazione, ed infatti per ogni dimension table, relativa ad una gerarchia di questo tipo, sono stati introdotti due ulteriori marche temporali che definiscono la validità temporale della tupla, e un attributo che ne identifica la versione.