

UNIVERSITÀ DEGLI STUDI DI BARI

DIPARTIMENTO DI INFORMATICA

Caso di studio del corso di Integrazione e test di sistemi software

Business Intelligence: dai dati alla conoscenza

Gruppo: 4Designers

Federica Lisco, Davide Di Pierro, Domenico Gigante, Graziano Castellano

# Abstract

Al giorno d’oggi le aziende di ogni calibro sono sempre alla ricerca di strumenti, strategie e soluzioni per migliorare la propria capacità competitiva nei confronti della concorrenza. Le sorti del business sono influenzate fortemente, oltre che dalle idee, dalla risolutezza e dalla brillantezza di chi compie le scelte. I mezzi aiutano ad assumere le corrette valutazioni e le corrette decisioni.

La disciplina che si occupa della creazione di conoscenza è la Business Intelligence. Questa consente alle organizzazioni di diventare competitive prendere decisioni in maniera consapevole attraverso l’analisi delle informazioni.

Quindi “*After the information has been extracted from the data the information is yet to be interpreted the process used to interpret and derive value from information is often called as information value chain. The first step in the value chain is the extraction of data from different sources; applying different logics and business contexts to this data creates information; information is then consumed by BI users; Based on these information different decisions are made and executed; thus increasing the business value.*” [1]

Nel nostro caso di studio si è deciso di fare BI sugli incidenti stradali avvenuti in Lombardia tra gli anni 2000 e 2011.

I risultati ottenuti sono i seguenti:

//VALUTARE

“*The collected data is used to form the Data Warehouse of the enterprise and will be analyzed by the BI applications. The analysis from the BI applications is used to understand business behavior and make strategic decisions.*”[2]

# Indice

1. Introduzione
2. Letteratura a supporto
3. Procedura ETL
4. Strumenti utilizzati
5. Esperienza
6. Conclusione
7. Bibliografia

# Sommario

[Abstract 1](#_Toc532720442)

[Indice 2](#_Toc532720443)

[Sommario 2](#_Toc532720444)

[Introduzione 2](#_Toc532720445)

[Letteratura a Supporto 5](#_Toc532720446)

[Strumenti utilizzati e ETL // Cosa? 5](#_Toc532720447)

[Caso di studio 5](#_Toc532720448)

[Procedura ETL 9](#_Toc532720449)

[Conclusioni 12](#_Toc532720450)

[Bibliografia 12](#_Toc532720451)

# Introduzione

Compiere decisioni può risultare complicato specialmente quando c’è una considerevole mole di dati alla base. Il principale problema da affrontare diventa, quindi, quello di estrapolare le informazioni a partire da questi dati. Sebbene dato e informazione siano due termini spesso utilizzati come sinonimi, hanno un significato differente, così come sottoscrive Drucker affermando che *“Information is data endowed with relevance and purpose”*[3].

Questa differenza viene evidenziata anche da altri tra cui Hicks che definisce il dato come: “*A representation of facts, concepts or instructions in a formalised manner suitable for communication, interpretation, or processing by humans or by automatic means*.”[4], o ancora è definito come *“Raw facts that can be shaped and formed to create information.”.*[5]

Mentre l’informazione è *“data that has been processed so that it is meaningful to a decision maker to use in a particular decision. ”*[4], oppure *“Data that have been shaped or formed by humans into a meaningful and useful form.”*[5]

Tuttavia, dopo aver estrapolato le informazioni dai dati, il problema consiste nel non riuscire ad usare questa ingente mole di informazioni per ottenere conoscenza e valutare le prestazioni di un determinato processo aziendale. Definiamo i processi aziendali come *“quite simply, the processes a business must execute in order for the business to operate.”*[6]

Per conoscenza, invece, si intende *“a fluid mix of framed experience, values,contextual information, and expert insight that provides a framework for evaluating and incorporating new experiences and information.”*[7]

Il processo che trasforma i dati e le informazioni in conoscenza è detto Business Intelligence ed è definito come: *“the intelligence as information valued for its currency and relevance. It is expert information, knowledge and technologies efficient in the management of organizational and individual business. Therefore, in this sense, business intelligence is a broad category of applications and technologies for gathering, providing access to, and analyzing data for the purpose of helping enterprise users make better business decisions.”* [8]

Un’altra definizione di Business Intelligence (BI) è “*the techniques or practices which utilize different technologies to create different methods or applications which analyze the business data available with the organization to help the enterprise to take decisions based on the predictions made by the data.*”[9]

(Integrare con la frase precedente).È possibile ottenere la suddetta conoscenza attraverso l’uso di un complesso strumento chiamato Business Intelligence System (BI System), cioè *“an integrated set of tools, technologies and programmed products that are used to collect, integrate, analyse and make data available”*[10]

È importante sottolineare che la BI non comprende soltanto le tecnologie di elaborazione e analisi dei dati ma anche molte pratiche e metodi di business che possono essere applicati a vari settori (e-governance, assistenza sanitaria, sicurezza, …). [1], [9]

Il BI System elabora informazioni a partire da una sorgente, chiamata Data Warehouse definita come: “*an integrated and time-varying collection of data primarily used in strategic decision making by means of online analytical processing (OLAP) techniques. It is essentially a database that stores integrated, often historical, and aggregated information extracted from multiple, heterogeneous, autonomous, and distributed information sources.”*[11]

Costruire un grande Data Warehouse spesso porta a un maggiore interesse nell'analisi e nell'utilizzo dell'accumulo di dati storici. Una soluzione sarebbe analizzare i dati in un Data Warehouse mediante analisi analitica on-line con strumenti di elaborazione (OLAP). Per OLAP (On-Line Analysis Processing) si intende “*software for manipulating multidimensional data from a variety of sources that has been stored in a data warehouse. The software can create various views and representations of the data. OLAP software provides fast, consistent, interactive access to shared, multidimensional data. These systems are used to discover trends, analyze critical factors and perform statistical analysis*.[12]

L’OLAP nasce però grazie al mondo OLTP (On-Line Transaction Processing) che è definito come: “*a class of systems that supports or facilitates high transaction-oriented applications. OLTP’s primary system features are immediate client feedback and high individual transaction volume.”[13]*

I sistemi informativi si possono tassonomizzare nelle seguenti quattro tipologie utili al fine di costituire quel valore aggiunto indispensabile per un’azienda che opera nella società dell’informazione:

* TPS (Transaction Processing System). Sistema informativo a livello operativo di elaborazione delle transazioni.
* MIS (Management Information System). Sistema informativo di gestione delle informazioni per i middle manager.
* DSS (Decision Support System). Sistema informativo di supporto alle decisioni.
* ESS (Executive Support System). Sistema informativo a livello strategico e di supporto direzionale.

I primi due fanno riferimento al mondo OLTP, mentre i restanti al mondo OLAP.

Per TPS si intende “*an information processing system for business transactions involving the collection, modification and retrieval of all transaction data. Characteristics of a TPS include performance, reliability and consistency.*” [14]

Per MIS si intende “*a computerized database of financial information organized and programmed in such a way that it produces regular reports on operations for every level of management in a company.*” [15]

Per DSS si intende “*a computer program application that analyzes business data and presents it so that users can make business decisions more easily. It is an informational application that collects the*[*data*](https://searchdatamanagement.techtarget.com/definition/data)*in the course of normal business operation.*” [16]

Un’ altra definizione concerne il DSS è: “*Decision support systems (DSS) are computer technology solutions that can be used to support complex decision making and problem solving*.”[17].

Per ESS si intende “*a software that allows users to transform enterprise data into quickly accessible and executive-level reports, such as those used by billing, accounting and staffing departments. It enhances decision making for executives.*” [18]

Riassumendo, possiamo affermare che lo scopo principale della tecnologia dei Data Warehouse è proprio quello di riorganizzare e sintetizzare le informazioni immagazzinate dai sistemi operazionali, permettendo di condurre analisi immediate sull’andamento di determinati processi.

Definiamo Data warehousing come *“the process of constructing and using data warehouses”*[19]*.* Completata la fase di datawarehousing, si può passare all’analisi dei dati contenuti nel DW. Questa fase è detta OLAP: *“OLAP (On-Line Analytical Processing) refers to a set of data analysis techniques developed for analyzing data in data warehouses since 1990s.”*[20]

I componenti di un Data warehouse sono:

* Sistemi Sorgenti operazionali – Data Source
* Area di staging dei dati - Sistema di alimentazione;
* Area di presentazione dei dati – Data Warehouse Server
* Strumenti di accesso ai dati

La componente di Data Source è costituita dall’insieme delle sorgenti di dati che un’azienda dispone. Questi possono essere correlati o meno ed è per questo necessaria una fase di integrazione iniziale.

L’area di staging è un’area di memorizzazione ma anche un insieme di processi ETL (Extraction Transform and Loading).

Per procedura ETL si intende: “*a process in data warehousing responsible for pulling data out of the source systems and placing it into a data warehouse. ETL involves the following tasks:*

* *Extracting the data;*
* *Transforming the data;*
* *Loading the data;”* [21]

L’area di presentazione dei dati descrive la modalità con cui sono risolti i problemi legati ai database relazionali. L’area di presentazione si può basare su sistema relazione (ROLAP) o multidimensionale (MOLAP).

Per ROLAP si intende “*a form of online analytical processing that performs dynamic multidimensional analysis of data stored in a*[*relational database*](https://searchdatamanagement.techtarget.com/definition/relational-database)*.*” [22] Per MOLAP si intende “*an online analytical processing that indexes directly into a*[*multidimensional database*](https://searchoracle.techtarget.com/definition/multidimensional-database).” [23]

# Letteratura a Supporto (cosa?)

# Caso di studio

Per il nostro caso di studio si è preso in considerazione il database degli incidenti stradali accaduti in Lombardia tra il duemila e il duemilaundici presi dal seguente Open Data: https://www.dati.gov.it/dataset/incidenti-stradali-lombardia-0.

Il database è espresso in formato csv e conteneva in un’unica tabella i seguenti campi:

*anno\_incidente, prov, cod\_com, comune, tot\_inc, di\_cui\_inc\_mortali, tot\_feriti, tot\_morti, veic\_coinvolti\_(identificato\_conducente), di\_cui\_altri\_veic\_coinvolti, conduc\_totali, di\_cui\_conduc\_femmine, passeggeri\_infortunati, di\_cui\_passeggeri\_infortunati\_femmine, pedoni\_infortunati, di\_cui\_pedoni\_infortunati\_femmine, conduc\_feriti, conduc\_morti\_entro24h, conduc\_ignoti, conduc\_morti\_entro30gg, conduc\_età\_0-14\_inc, conduc\_età\_15-19\_inc, conduc\_età\_20-64\_inc, conduc\_età\_65\_e\_oltre\_inc, conduc\_età\_0-14\_feriti, conduc\_età\_15-19\_feriti, conduc\_età\_20-64\_feriti, conduc\_età\_65\_e\_oltre\_feriti, conduc\_età\_0-14\_morti, conduc\_età\_15-19\_morti, conduc\_età\_20-64\_morti, conduc\_età\_65\_e\_oltre\_morti, passeggeri\_feriti, passeggeri\_morti, pedoni\_feriti, pedoni\_morti, altri\_veic\_feriti, altri\_veic\_morti, intersezione\_inc, intersezione\_feriti, non\_intersezione\_morti\_entro24h, non\_intersezione\_inc, non\_intersezione\_feriti, non\_intersezione\_morti\_\_entro30gg, strade\_urbane\_inc, strade\_urbane\_feriti, strade\_urbane\_morti, strade\_extraurb\_(no\_autostr)\_inc, strade\_extraurb\_(no\_autostr)\_feriti, strade\_extraurb\_(no\_autostr)\_morti, autostr\_inc, autostr\_feriti, autostr\_morti, inc\_tra\_veic\_in\_marcia, inc\_tra\_veic\_in\_marcia\_feriti, inc\_tra\_veic\_in\_marcia\_morti, inc\_tra\_veic\_e\_pedone, inc\_tra\_veic\_e\_pedone\_feriti, inc\_tra\_veic\_e\_pedone\_morti, inc\_tra\_veic\_isolati, inc\_tra\_veic\_isolati\_feriti, inc\_tra\_veic\_isolati\_morti, weekend\_inc, weekend\_feriti, weekend\_morti, feriali\_inc, feriali\_feriti, feriali\_morti, notte\_inc, notte\_feriti, notte\_morti, giorno\_inc, giorno\_feriti, giorno\_morti, ore\_di\_punta\_07-09\_inc, ore\_di\_punta\_07-09\_feriti, ore\_di\_punta\_07-09\_morti, ore\_di\_punta\_17-19\_inc, ore\_di\_punta\_17-19\_feriti, ore\_di\_punta\_17-19\_morti, sereno\_inc, sereno\_feriti, sereno\_morti, nebbia\_inc, nebbia\_feriti, nebbia\_morti, pioggia-neve\_inc, pioggia-neve\_feriti, pioggia-neve\_morti, veic\_coinvolti\_autovet\_(priv\_e\_pub), autovet\_(priv\_e\_pub)\_feriti, autovet\_(priv\_e\_pub)\_morti, veic\_coinvolti\_autocar\_e\_simili, autocar\_e\_simili\_feriti, autocar\_e\_simili\_morti, veic\_coinvolti\_motocicli, motocicli\_feriti, motocicli\_morti, veic\_coinvolti\_velocipedi, velocipedi\_feriti, velocipedi\_morti*

A questa tabella si è poi deciso di eliminare i seguenti campi:

*cod\_comune, conduc\_morti\_entro30gg, altri\_veic\_feriti, altri\_veic\_morti, intersezione\_inc, intersezione\_feriti, non\_intersezione\_morti\_entro24h, non\_intersezione\_inc, non\_intersezione\_feriti,*

*non\_intersezione\_morti\_\_entro30gg, inc\_tra\_veic\_in\_marcia\_feriti, inc\_tra\_veic\_in\_marcia\_morti,*

*inc\_tra\_veic\_e\_pedone\_feriti, inc\_tra\_veic\_e\_pedone\_morti, inc\_tra\_veic\_isolati\_feriti, inc\_tra\_veic\_isolati\_morti, weekend\_feriti, weekend\_morti, feriali\_feriti, feriali\_morti, notte\_feriti, notte\_morti, giorno\_feriti, giorno\_morti, ore\_di\_punta\_07-09\_feriti, ore\_di\_punta\_07-09\_morti,*

*ore\_di\_punta\_17-19\_feriti, ore\_di\_punta\_17-19\_morti, sereno\_feriti, sereno\_morti, nebbia\_feriti,*

*nebbia\_morti, pioggia-neve\_feriti, pioggia-neve\_morti, autovet\_(priv\_e\_pub)\_feriti,*

*autovet\_(priv\_e\_pub)\_morti,autocar\_e\_simili\_feriti,autocar\_e\_simili\_morti,motocicli\_feriti,*

*motocicli\_morti,velocipedi\_feriti,velocipedi\_morti passeggeri\_infortunati,*

*di\_cui\_passeggeri\_infortunati\_femmine, pedoni\_infortunati,di\_cui\_pedoni\_infortunati\_femmine.*

Per motivi di prestazione, è stato poi deciso di inserire i seguenti campi che saranno utili nelle interrogazioni:

*di\_cui\_inc\_non\_mortali, di\_cui\_conduc\_maschi, veicoli\_coinvolti\_altri.*

Seguendo il modello relazione, bisogna stabilire per ogni campo se fa parte della tabella dei fatti o della tabella delle dimensioni. Si approfitta per descrivere brevemente ogni campo:

* *anno\_incidente Descrizione: anno in esame Tabella: fatti;*
* *prov Descrizione: provincia in esame Tabella: fatti;*
* *comune Descrizione: comune in esame Tabella: dimensioni;*
* *tot\_inc Descrizione: incidenti nell’anno Tabella: fatti;*
* *di\_cui\_inc\_mortali Descrizione: incidenti mortali Tabella: fatti;*
* *tot\_feriti Descrizione: Numero feriti Tabella: fatti;*
* *tot\_morti Descrizione: Numero morti Tabella: fatti;*
* *veic\_coinvolti\_(identificato\_conducente) Descrizione: Veicoli incidentati con conducente Tabella: fatti;*
* *di\_cui\_altri\_veic\_coinvolti Descrizione: Veicoli incidentati senza conducente Tabella: fatti;*
* *conduc\_totali Descrizione: Conducenti totali Tabella: fatti;*
* *di\_cui\_conduc\_femmine Descrizione: Conducenti donne Tabella: fatti;*
* *conduc\_feriti Descrizione: Conducenti feriti Tabella: fatti;*
* *conduc\_morti\_entro24h Descrizione: Conducenti morti entro la giornata Tabella: fatti;*
* *conduc\_ignoti Descrizione:Conducenti non riconosciuti Tabella: fatti;*
* *conduc\_età\_0-14\_inc Descrizione: Conducenti da zero a quattoridici anni Tabella: fatti;*
* *conduc\_età\_15-19\_inc Descrizione: Conducenti da quindici a diciannove anni Tabella: fatti;*
* *conduc\_età\_20-64\_inc Descrizione: Conducenti da venti a sessantaquattro Tabella: fatti;*
* *conduc\_età\_65\_e\_oltre\_inc Descrizione: Conducenti con almeno sessantacinque anni Tabella: fatti;*
* *conduc\_età\_0-14\_feriti Descrizione: Conducenti feriti da zero a quattoridici Tabella: fatti;*
* *conduc\_età\_15-19\_feriti Descrizione: Conducenti feriti da quindici a diciannove anni Tabella: fatti;*
* *conduc\_età\_20-64\_feriti Descrizione: Conducenti feriti da venti a sassantaquattro anni Tabella: fatti;*
* *conduc\_età\_65\_e\_oltre\_feriti Descrizione: Conducenti feriti con almeno sessantacinque anni Tabella: fatti;*
* *conduc\_età\_0-14\_morti Descrizione: Conducenti morti da zero a quattoridici Tabella: fatti;*
* *conduc\_età\_15-19\_morti Descrizione: Conducenti morti da quindici a diciannove anni Tabella: fatti;*
* *conduc\_età\_20-64\_morti Descrizione: Conducenti morti da venti a sassantaquattro anni Tabella: fatti;*
* *conduc\_età\_65\_e\_oltre\_morti Descrizione: Conducenti morti con almeno sessantacinque anni Tabella: fatti;*
* *passeggeri\_feriti Descrizione: Numero passeggeri feriti Tabella: fatti;*
* *passeggeri\_morti Descrizione: Numero passeggeri morti Tabella: fatti;*
* *pedoni\_feriti Descrizione: Numero pedoni feriti Tabella: fatti;*
* *pedoni\_morti Descrizione: Numero pedoni morti Tabella: fatti;*
* *strade\_urbane\_inc Descrizione: Incidenti su strade urbane Tabella: fatti;*
* *strade\_urbane\_feriti Descrizione: Feriti su strade urbane Tabella: fatti;*
* *strade\_urbane\_morti Descrizione: Morti su strade urbane Tabella: fatti;*
* *strade\_extraurb\_(no\_autostr)\_inc Descrizione: Incidenti su strade extraurbane Tabella: fatti;*
* *strade\_extraurb\_(no\_autostr)\_feriti Descrizione: Feriti su strade extraurbane Tabella: fatti;*
* *strade\_extraurb\_(no\_autostr)\_morti Descrizione: Morti su strade extraurbane Tabella: fatti;*
* *autostr\_inc Descrizione: Incidenti su autostrade Tabella: fatti;*
* *autostr\_feriti Descrizione: Feriti su autostrade Tabella: fatti;*
* *autostr\_morti Descrizione: Morti su autostrade Tabella: fatti;*
* *inc\_tra\_veic\_in\_marcia Descrizione: Incidenti tra veicoli in marcia Tabella: fatti;*
* *inc\_tra\_veic\_e\_pedone Descrizione: Incidenti tra veicolo e pedone Tabella: fatti;*
* *inc\_tra\_veic\_isolati Descrizione: Incidenti tra veicoli isolati Tabella: fatti;*
* *weekend\_inc Descrizione: Incidenti nei weekend Tabella: fatti;*
* *feriali\_inc Descrizione: Incidenti nei giorni feriali Tabella: fatti;*
* *notte\_inc Descrizione: Incidenti di notte Tabella: fatti;*
* *giorno\_inc Descrizione: Incidenti di giorno Tabella: fatti;*
* *ore\_di\_punta\_07-09\_inc Descrizione: Incidenti tra le 7.00 e le 9.00 Tabella: fatti;*
* *ore\_di\_punta\_17-19\_inc Descrizione: Incidenti tra le 17.00 e le 19.00 Tabella: fatti;*
* *sereno\_inc Descrizione: Incidenti con clima sereno Tabella: fatti;*
* *nebbia\_inc Descrizione: Incidenti con nebbia Tabella: fatti;*
* *pioggia-neve\_inc Descrizione: Incidenti con pioggia o neve Tabella: fatti;*
* *veic\_coinvolti\_autovet\_(priv\_e\_pub) Descrizione: Numero autovetture coinvolte Tabella: fatti;*
* *veic\_coinvolti\_autocar\_e\_simili Descrizione: Numero autocarri coinvolti Tabella: fatti;*
* *veic\_coinvolti\_motocicli Descrizione: Numero motocicli coinvolti Tabella: fatti;*
* *veic\_coinvolti\_velocipedi Descrizione: Numero velocipiedi coinvolti Tabella: fatti;*
* *di\_cui\_inc\_non\_mortali Descrizione: Numero incidenti non mortali Tabella: fatti;,*
* *di\_cui\_conduc\_maschi Descrizione: Numero conducenti uomini Tabella: fatti;,*
* *veicoli\_coinvolti\_altri Descrizione: Numero veicoli diversi dai precedenti coinvolti Tabella: fatti;.*

Per l’analisi dei dati si è deciso di utilizzare un tool chiamato Jaspersoft iReport che, però, ha comportato la necessità di rinominare alcuni campi della tabella troppo lunghi.

Per una spiegazione dettagliata sul suddetto software, si rimanda a pagina 12.

Si riportando di seguito le modifiche:

ANNO\_INCIDENTE 🡪 ANNO;

VEIC\_COINVOLTI\_(IDENTIFICATO\_CONDUCENTE) 🡪 TOT\_VEIC\_COINVOLTI;

DI\_CUI\_ALTRI\_VEIC\_COINVOLTI 🡪 TOT\_VEIC\_CONDUC\_IGNOTO;

CONDUC\_TOTALI 🡪 TOT\_CONDUC;

DI\_CUI\_CONDUC\_FEMMINE 🡪 CONDUC\_FEMMINE;

CONDUC\_ETÀ\_65\_E\_OLTRE\_INC 🡪 CONDUC\_ETA\_65+\_INC;

CONDUC\_ETÀ\_65\_E\_OLTRE\_FERITI 🡪 CONDUC\_ETA\_65+\_FERITI;

CONDUC\_ETÀ\_65\_E\_OLTRE\_MORTI 🡪 CONDUC\_ETA\_65+\_MORTI;

*DI\_CUI\_INC\_NON\_MORTALI 🡪 TOT\_INC\_NONMORTALI;*

*DI\_CUI\_CONDUC\_MASCHI 🡪 CONDUC\_MASCHI;*

# Procedura ETL

La procedura ETL che si vuole realizzare dev’essere in grado di prendere in input un file sorgente (incidenti.csv) in formato csv e di scrivere un nuovo file (new\_incidenti.csv) che contenga i vecchi record più quelli nuovi. Deve anche essere prodotto un report in formato HTML (risultato[anno]\_[mese]\_[giorno].html) per avere i risultati sommativi, leggibili tramite browser, della procedura dove per [anno], [mese] e [giorno] si intende rispettivamente anno, mese e giorno della data in cui viene lanciata la procedura.

Il file sorgente deve contenere, per un anno, il numero di incidenti per ogni comune della Lombardia. È quindi plausibile che questa procedura debba essere lanciata esattamente una volta all’anno.

Per quanto concerne la correttezza, la procedura dovrà essere stringente su alcuni controlli e più elastica su altri in quanto non si conoscono informazioni su tutti i vincoli che legano i vari campi tra loro.

La procedura dovrà innanzitutto verificare che il numero di colonne che si vogliono inserire sia esattamente cinquantasei e che le colonne abbiano questi titoli, rispettivamente:

anno, prov, comune, tot\_inc, tot\_inc\_mortali, tot\_feriti, tot\_morti, tot\_veic\_coinvolti, tot\_veic\_conduc\_ignoto, tot\_conduc, conduc\_femmine, conduc\_feriti, conduc\_morti\_entro24h, conduc\_ignoti, conduc\_eta\_0-14\_inc, conduc\_eta\_15-19\_inc, conduc\_eta\_20-64\_inc, conduc\_eta\_65+\_inc, conduc\_eta\_0-14\_feriti, conduc\_eta\_15-19\_feriti, conduc\_età\_20-64\_feriti, conduc\_eta\_65+\_feriti, conduc\_eta\_0-14\_morti, conduc\_eta\_15-19\_morti, conduc\_eta\_20-64\_morti, conduc\_eta\_65+\_morti, passeggeri\_feriti, passeggeri\_morti, pedoni\_feriti, pedoni\_morti, strade\_urbane\_inc, strade\_urbane\_feriti, strade\_urbane\_morti, strade\_extraurb\_inc, strade\_extraurb\_feriti, strade\_extraurb\_morti, autostr\_inc, autostr\_feriti, autostr\_morti, inc\_tra\_veic\_in\_marcia, inc\_tra\_veic\_e\_pedone, inc\_tra\_veic\_isolati, weekend\_inc, feriali\_inc, notte\_inc, giorno\_inc, ore\_07-09\_inc, ore\_17-19\_inc, sereno\_inc, nebbia\_inc, pioggia-neve\_inc, veic\_coinvolti\_autovet\_(priv\_e\_pub), veic\_coinvolti\_autocar\_e\_simili, veic\_coinvolti\_motocicli, veic\_coinvolti\_velocipedi, tot\_inc\_nonmortali, conduc\_maschi, veic\_coinvolti\_altro.

L’ordine è fondamentale. Se la prima riga del file sorgente ha questi campi ma qualcuno non è nella sua posizione stabilita, la procedura si arresterà facendo evidenziare il problema nel report in formato HTML.

Successivamente, viene eseguito il controllo sul dominio dei dati. Tra i diversi campi ce ne sono alcuni numerici (interi) e altri testuali.

I campi provincia e comune sono testuali mentre gli altri sono numerici.

Per ogni campo numerico viene controllato che il numero sia positivo (maggiore o uguale a 0). In particolare, l’anno deve essere un valore compreso tra 0 e l’anno corrente. In questo modo si garantisce che la procedura ETL possa funzionare correttamente nel tempo.

Dopo aver eseguito i vari controlli di dominio, si cerca invece se i campi provincia e comune sono consistenti.

I comuni devono essere necessariamente della Lombardia e la provincia deve essere corretta a seconda del comune.

La corrispondenza comune-provincia viene verificata attraverso un file che ha messo a disposizione la regione Lombardia.

Se la coppia comune-provincia non trova corrispondenza nel file messo a disposizione dalla regione, il dato viene considerato errato e il record non sarà inserito nel file dei risultati.

Vengono poi eseguiti dei controlli sulla consistenza dei dati.

Se questi controlli generano errori, però, il record verrà comunque considerato corretto e, se non duplicato, inserito nel file dei risultati. Nel report però verrà segnalato che eventualmente su una riga c’è un dato anomalo (con riga e descrizione dell’anomalia).

I controlli di consistenza per record sono i seguenti:

* *tot\_inc >= strade\_urbane\_inc + strade\_extraurbane\_inc+autostrade\_inc;*
* *tot\_inc >= inc\_tra\_veic\_in\_marcia + inc\_tra\_veic\_pedone + inc\_tra\_veic\_isolati;*
* *tot\_inc = weekend\_inc + feriali\_inc;*
* *tot\_inc >= giorno\_inc+ notte\_inc;*
* *tot\_inc >= ore\_di\_punta\_inc\_7-9 + ore\_di\_punta\_inc\_17-19;*
* *tot\_inc >= sereno\_inc + nebbia\_inc + pioggia\_neve\_inc;*
* *tot\_feriti >= conduc\_feriti + passeggeri\_feriti + pedoni\_feriti;*
* *tot\_morti >= conduc\_eta\_0-14\_mor + conduc\_eta\_15-19\_mor + conduc\_eta\_20-64\_mor + conduc\_eta\_65+ + passeggeri\_morti + pedoni\_morti;*
* *veicoli\_coinvolti\_(ident\_conduc) >= veic\_coinvolti\_autovet\_(priv\_e\_pub) +*

*veic\_coinvolti\_autocar\_e\_simili + veic\_coinvolti\_motocicli + veic\_coinvolti\_velocipedi;*

* *conduc\_tot >= conduc\_eta\_0-14\_inc + conduc\_eta\_15-19\_inc + conduc\_eta\_20-64\_inc + conduc\_eta\_65+;*
* *conduc\_feriti >= conduc\_eta\_0-14\_feriti + conduc\_eta\_15-19\_feriti + conduc\_eta\_20-64\_feriti + conduc\_eta\_65+\_feriti;*

Verificata la correttezza, si determina se il record è o meno un duplicato.

Ogni record deve essere necessariamente distinto dalla coppia anno-comune che fungono quindi da chiave composta.

La scelta viene evinta dallo specifico caso in cui per ogni anno vengono inseriti informazioni riguardo un singolo comune.

Si arriva quindi a determinare se un record è duplicato analizzando quei due campi.

In memoria viene salvata una struttura a dizionario (chiave-valore) in cui per ogni anno c’è un insieme di comuni già inseriti. Se la specifica coppia che si intende inserire è già presente, il record sarà segnalato come duplicato e non verrà inserito.

I record del file sorgente sono indipendenti tra loro: se uno di loro è incorretto o duplicato, gli altri possono tranquillamente essere analizzati ed eventualmente inseriti.

Al termine dell’inserimento delle tuple, viene calcolato il numero di record mancanti. Se la procedura prevede che per ogni anno vengano inseriti tutti i comuni della Lombardia ci si aspetta di trovare milleecinquecentosedici comuni per anno.

La procedura memorizza tutte le possibili combinazioni chiave-valore avente come chiave l’anno e come valore l’insieme dei possibili comuni.

Per ogni coppia chiave-valore presente, si elimina dalla lista di tutte le possibili combinazioni quest’ultima e, alla fine, si calcola il numero di valori ancora presenti.

**Analisi delle prestazioni**

Si è verificato sperimentalmente che la procedura ETL impiega complessivamente 30-40 secondi ricevendo in input poco più di 150000 record.

Si può quindi stimare che, se in un anno debbano essere caricati 1516 record (numero di comuni), la procedura mediamente non dovrebbe terminare dopo 4 secondi.

T15000 = 40 s.

T1516 = T15000/15000 \* 1516 ≈ 4 s.

Possiamo considerare questo ritardo ampiamente accettabile visto che la procedura, in media, dovrebbe essere lanciata una volta all’anno.

# Strumenti utilizzati

Gli strumenti utilizzati sono stati:

* NetBeans IDE 8.2 per la realizzazione del codice Java che alla base della procedura ETL;
* Jaspersoft iReport – NetBeans Plugin;
* RapidMiner per la realizzazione di grafici a partire dai risultati;

## Jaspersoft iReport

Inizialmente, per la realizzazione di un esempio di report finale, si era deciso di utilizzare il software di Business Intelligence (BI) “TIBCO Jaspersoft"[[1]](#footnote-1). Tuttavia, tale software non è open source e, vista la natura accademica del progetto, si è deciso di optare per la componente open source del suddetto software, chiamata “Jaspersoft iReport – NetBeans Plugin”[[2]](#footnote-2), il quale offre un set di funzionalità ridotto ed una interfaccia grafica un po’ meno usabile.

Attraverso Jaspersoft iReport è stato possibile creare un report eseguendo le seguenti operazioni:

1. Selezionare il layout, a scelta tra un set di layout predefiniti
2. Selezionare la sorgente dati (nel nostro caso il file .csv generato dalla procedura ETL)
3. Selezionare i campi di interesse e, eventualmente, rinominarli secondo le proprie necessità
4. Impostare la sorgente dati anche nel grafico

Al termine di queste fasi, si verrà trasferiti in un editor in cui è possibile personalizzare ulteriormente il layout del report.

Per vedere un’anteprima del report che verrà generato, è possibile cliccare sulla voce “Run”: così facendo, verrà mostrato il report con i dati che il software è riuscito a recuperare dalla sorgente dati indicati.

Infine, è possibile esportare tale report in molti formati standard (tra cui .PDF).

**RapidMiner**

Come strumento rapido per la produzione di grafici e tabelle si è deciso di utilizzare RapidMiner che necessita solo di un file in formato CSV (in alternativa PDF, XLS, …) per creare diversi tipi di diagramma (a torta, istogrammi, ecc.) utilizzando i dati numerici che ha a disposizione.

È possibile ovviamente utilizzare funzioni complesse su dati numerici quali somma, media, deviazione standard ecc.

All’interno dell’applicativo basterà creare un oggetto di tipo Data Resource, caricare il file sorgente e collegare la risorsa creata precedentemente a quest’ultimo.

L’unico difetto è che non è possibile esportare i grafici in formati standard.

# Conclusioni

# Bibliografia

[1] P. Wazurkar, R. S. Bhadoria, and D. Bajpai, “Predictive analytics in data science for business intelligence solutions,” in *2017 7th International Conference on Communication Systems and Network Technologies (CSNT)*, 2017, pp. 367–370.

[2] A. D. Patil and N. D. Gangadhar, “OLaaS: OLAP as a Service,” in *Proceedings - 2016 IEEE International Conference on Cloud Computing in Emerging Markets, CCEM 2016*, 2017, pp. 119–124.

[3] P. F. Drucker, “The Coming of the New Organization Harvard Business Review,” 1987.

[4] J. O. Hicks, *Management Information Systems: A User Perspective*. West Publishing Company, 1993.

[5] K. C. Laudon and J. P. Laudon, “Business information systems: a problem solving approach.” Dryden Press, Chicago, p. xxiii, 631 p., 1991.

[6] A. Elkin and S. Opitz, “Method and system for top-down business process definition and execution.” Google Patents, 2007.

[7] T. H. Davenport, P. D. M. I. S. M. T. H. Davenport, L. Prusak, and I. York University - NetLibrary, *Working Knowledge: How Organizations Manage what They Know*. Harvard Business School Press, 1998.

[8] J. Ranjan, “Business intelligence: Concepts, components, techniques and benefits,” *J. Theor. Appl. Inf. Technol.*, vol. 9, no. 1, pp. 60–70, 2009.

[9] G. Shmueli and O. R. Koppius, “PREDICTIVE ANALYTICS IN INFORMATION SYSTEMS RESEARCH 1.”

[10] W. Yeoh and A. Koronios, “Critical success factors for business intelligence systems,” *J. Comput. Inf. Syst.*, vol. 50, no. 3, pp. 23–32, 2010.

[11] B. Hüsemann, J. Lechtenbörger, and G. Vossen, *Conceptual data warehouse design*. Universität Münster. Angewandte Mathematik und Informatik, 2000.

[12] “On-line Analytical Processing (OLAP).” [Online]. Available: http://dssresources.com/glossary/142.php. [Accessed: 01-Dec-2018].

[13] “What is Online Transaction Processing (OLTP)? - Definition from Techopedia.” [Online]. Available: https://www.techopedia.com/definition/24436/online-transaction-processing-oltp. [Accessed: 09-Dec-2018].

[14] “What is a Transaction Process System (TPS)? - Definition from Techopedia.” [Online]. Available: https://www.techopedia.com/definition/707/transaction-process-system-tps. [Accessed: 09-Dec-2018].

[15] “Management Information Systems (MIS) - Encyclopedia - Business Terms | Inc.com.” [Online]. Available: https://www.inc.com/encyclopedia/management-information-systems-mis.html. [Accessed: 09-Dec-2018].

[16] “What is decision support system (DSS)? - Definition from WhatIs.com.” [Online]. Available: https://searchcio.techtarget.com/definition/decision-support-system. [Accessed: 09-Dec-2018].

[17] J. P. Shim, M. Warkentin, J. F. Courtney, D. J. Power, R. Sharda, and C. Carlsson, “Past, present, and future of decision support technology $.”

[18] “What is an Executive Support System (ESS)? - Definition from Techopedia.” [Online]. Available: https://www.techopedia.com/definition/543/executive-support-system-ess. [Accessed: 09-Dec-2018].

[19] W. H. Inmon, “What is a data warehouse,” *Prism Tech Top.*, vol. 1, no. 1, 1995.

[20] J. Han, “OLAP mining: An integration of OLAP with data mining,” in *Proceedings of the 7th IFIP*, 1997, vol. 2, pp. 1–9.

[21] “ETL.” [Online]. Available: https://www.datawarehouse4u.info/ETL-process.html. [Accessed: 09-Dec-2018].

[22] “What is relational online analytical processing (ROLAP)? - Definition from WhatIs.com.” [Online]. Available: https://searchoracle.techtarget.com/definition/relational-online-analytical-processing. [Accessed: 09-Dec-2018].

[23] “What is MOLAP (multidimensional online analytical processing)? - Definition from WhatIs.com.” [Online]. Available: https://searchsqlserver.techtarget.com/definition/MOLAP. [Accessed: 09-Dec-2018].

1. https://www.jaspersoft.com/it [↑](#footnote-ref-1)
2. http://plugins.netbeans.org/plugin/4425/ireport [↑](#footnote-ref-2)