Università degli Studi di Camerino

Scuola di Scienze e Tecnologie

Corso di Laurea in Informatica

Big Data

Tecniche e tool di analisi

Elaborato Finale

Laureando

Emanuele Gentiletti

Matricola: **090150**

Relatore

Prof. Diletta Romana Cacciagrano

Anno Accademico 2016/2017

Indice

Introduzione

Definizione

Per Big Data si intendono collezioni di dati non gestibili da tecnologie "tradizionali". La ragione per cui queste collezioni non sono gestibili è rilevabile in tre fattori:

- Il **volume** della collezione;
- La **varietà**, intesa come la varietà di *fonti* e di *possibili strutturazioni* dell'informazione;
- La **velocità** dell'informazione, intesa come la velocità di produzione di nuova informazione.

Questo modello per descrivere i Big Data viene chiamato **modello delle 3V**. Ognuno dei punti di questo modello deriva da esigenze relativamente recenti (a volte per la tipologia, a volte per la scala di necessità), in particolare:

- Il volume delle collezioni dei dati è aumentato esponenzialmente in tempi recenti, con l'avvento dei Social Media, dell'IOT, e degli smartphone ~attrezzati con molte tipologie di sensori diversi. Generalizzando, i fattori che hanno portato a un grande incremento del volume dei data set sono un aumento della generazione automatica di dati da parte di dispositivi (sensori a basso costo e smartphone), in opposizione all'inserimento manuale dei dati da parte di operatori, e di un grande incremento dei contenuti prodotti dagli utenti rispetto al passato.
- La varietà delle collezioni di dati è aumentata, perché ci sono più fonti rispetto che in passato da cui è desiderabile attingere dati, e molte fonti forniscono dati che non sono strutturati uniformemente rispetto alle altre. Le fonti possono differire in struttura, o possono essere non strutturate affatto, come nel caso dei documenti JSON o del linguaggio naturale. Una struttura uniforme è una condizione necessaria per l'elaborazione corretta dei dati, e a volte può non essere triviale giungere a questa condizione. Ci sono molti casi in cui le fonti di dati possono avere informazioni non corrette che richiedono di essere filtrate, o in cui è necessario applicare strategie difensive nei confronti dei dati ricevuti, per la possibilità che questi siano mal filtrati o provengano da una fonte non sicure.
- Si possono fare le stesse considerazioni fatte per il volume dei dati per quanto riguarda la velocità. I flussi di dati vengono generati dai dispositivi e dagli utenti, che li producono a velocità molto maggiori rispetto a degli operatori.

La definizione di Big Data che ho fornito parla di collezioni di dati non gestibili da tecnologie tradizionali. Definite le caratteristiche di queste collezioni, le domande consequenziali a questa definizione sono, *quali sono le tecnologie tradizionali*, e *perché non sono adeguate?*

Data Warehousing ed ETL

Le tecnologie tradizionali, nell'ambito della collezione e della trasformazione dei dati, sono Data Warehousing ed ETL.

Il Data Warehousing è un'architettura