Una miniera di dati sul comportamento degli utenti del Web

Organizzare le informazioni sull'utilizzo della rete in PostgreSQL utilizzando ht://Miner, un sistema open-source di data mining e data warehousing

comune di
PRATO

Gabriele Bartolini
Comune di Prato – Sistema Informativo
Servizi di E-government e Open-Source

g.bartolini@comune.prato.it
www.htminer.it

Festa della Creatività 2007
I Diritti della Rete
Fortezza da Basso, Firenze, 26 ottobre 2007

ht://miner

Premessa e obiettivi della presentazione

- E' **impossibile concentrare** in un talk di 10 minuti:
  - Knowledge Discovery from Data (KDD)
  - Sistemi di supporto alle decisioni (DSS)
  - Data mining e web usage mining
  - Data warehouse e data webhouse
  - Un sistema open-source con:
    - oltre 300 classi C++ e 75 mila linee di codice
    - un'architettura a moduli abbastanza complessa
    - un database con centinaia di tabelle e indici
- Pertanto gli **obiettivi** sono quelli di rendere una idea circa:
  - il campo di applicazione
  - la soluzione open-source proposta (ht://Miner)
  - le possibilità in termini di scoperta di informazioni sia automatica che semiautomatica offerte dal sistema
  - le potenzialità ancora inesplorate
- **Presentazione scaricabile da Internet**
- **Disponibilità del sottoscritto a fornire maggiori informazioni**

Sommario

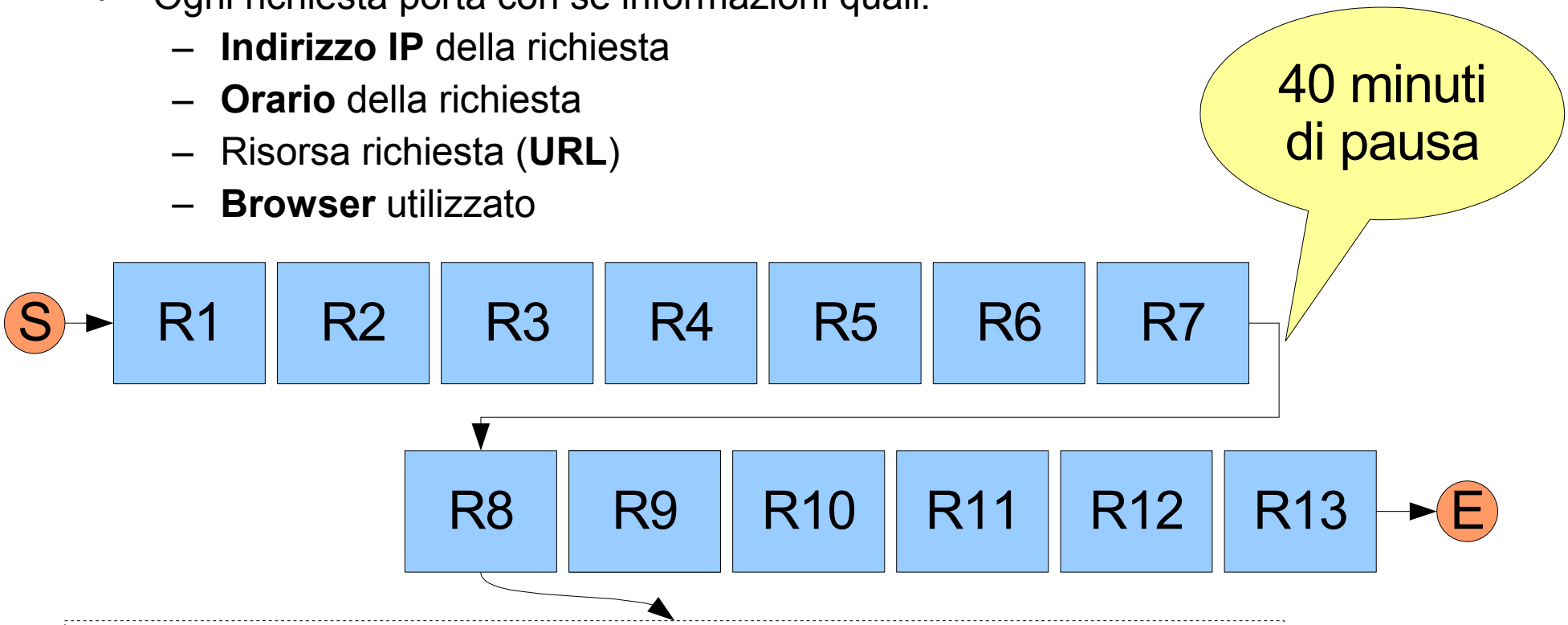
PARTE I - Introduzione

Si parla di “Web Usage Mining”

- E' parte integrante delle seguenti discipline di *Information Technology (IT)*:
  - *Knowledge Discovery from Data (KDD)*
  - *Data mining*, in particolare *Web mining* (data mining applicato al web)
- Dall'inglese:
  - *Web usage*: utilizzo del web (riferito agli utenti navigatori di Internet)
  - *Mining*: attività di estrazione di conoscenza *nascosta* da dati
- **Processo di scoperta e analisi di modelli** (*pattern*) **che concentra l'attenzione sui dati relativi agli accessi effettuati dagli utenti** (*Web usage data*)
- Rientrano in questa categoria i processi di analisi degli accessi raccolti a livello di:
  - **HTTP server (server web)**
  - HTTP proxy server
  - ISP
  - HTTP client
- L'argomento di questa presentazione verterà su accessi a server HTTP ma ...
- ... può essere esteso anche agli altri casi (con delle differenze)

La navigazione su Internet ***for dummies***

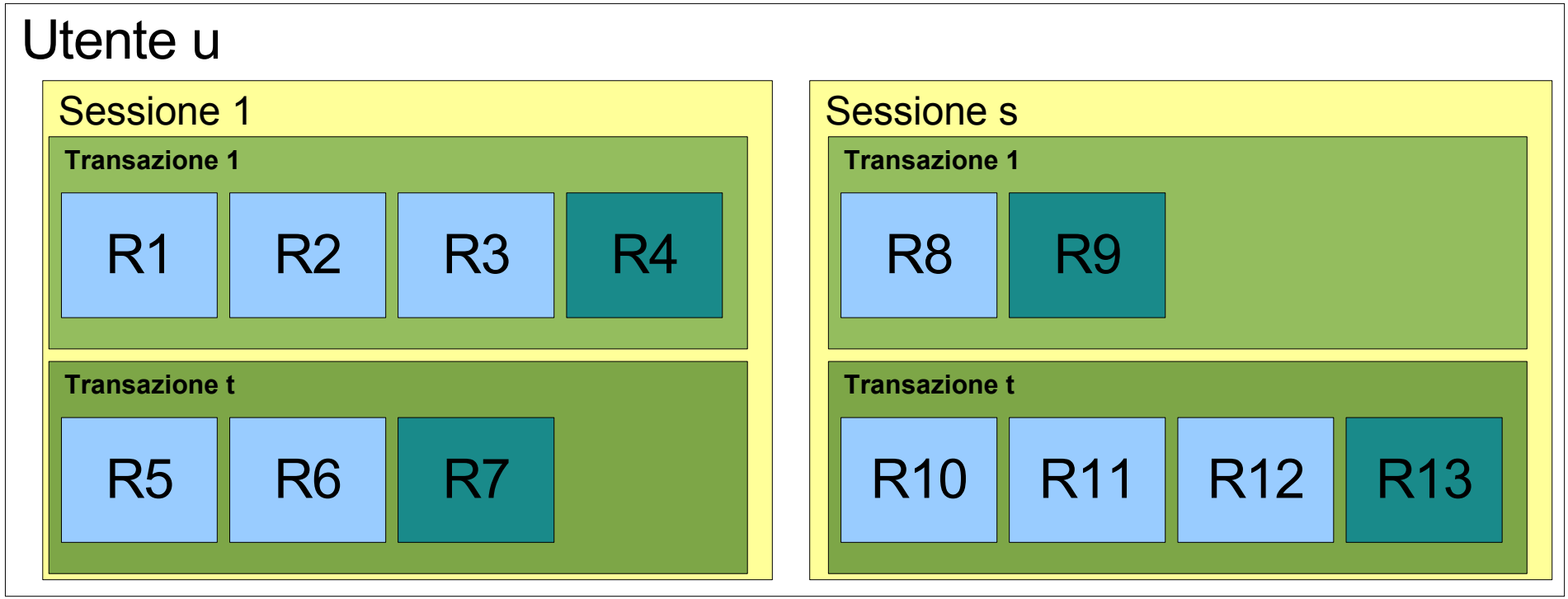
- L'utente tramite il browser accede a una risorsa sul server web X
- Il server web X memorizza la richiesta in un **file di testo** chiamato *access log*
- Il file *access log* è un elenco sequenziale delle richieste HTTP fatte al server
- Ogni richiesta porta con sé informazioni quali:
  - **Indirizzo IP** della richiesta
  - **Orario** della richiesta
  - Risorsa richiesta (**URL**)
  - **Browser** utilizzato



IP: 999.999.999.999
Orario: Mer 17 Ott 2007 ore 9.40.35 CEST
URL: http://www.comune.prato.it/concorsi/htm/z4.htm
Browser: Mozilla/5.0 (X11; U; Linux i686; it; rv:1.8.1) Gecko/20060601 Firefox/2.0 (Ubuntu-edgy)

Utenti, sessioni, transazioni e richieste

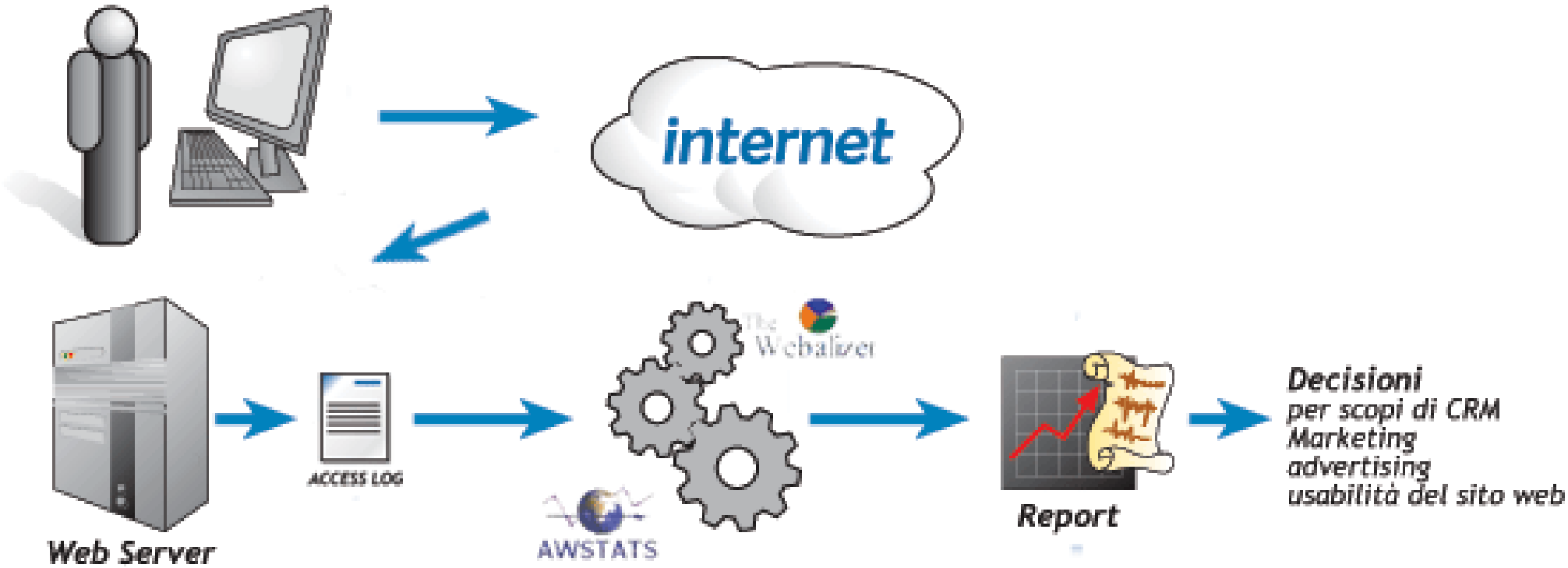
- Logicamente, è possibile raggruppare le richieste di uno **stesso utente** in:
  - **Sessioni** o visite (gruppo di richieste con tempo inattività < 30 minuti)
  - **Transazioni** (gruppi logici di richieste)
  - **Richieste**



orario(R8) – orario(R7) > 30 minuti

L'analisi *tradizionale* degli *access log (1/2)*

Non è possibile estrarre informazioni utili senza un'azione di *processing*



L'analisi *tradizionale* degli *access log (2/2)*

- E' norma utilizzare software di analisi dei log (*AWStats, Analog, Webalizer, ecc.*)
- Questo tipo di software, seppure funzionale e configurabile, ha dei limiti:
  - Produce report **statici** (istantanee)
  - Richiede una rielaborazione (parziale o totale) per modifiche ai report
  - Si integra difficilmente con fonti di dati ausiliarie
  - **Manca** di un **DB relazionale** sul quale effettuare interrogazioni puntuali
  - **Manca** di un **DB relazionale pluriennale** modellato sulle esigenze di business

```
SELECT u.url, r.tempo
    FROM richieste r, url u
    WHERE id_sessione = 14483884 AND u.id_url = r.id_url;
```

url	tempo
http://www.comune.prato.it/concorsi/	6
http://www.comune.prato.it/concorsi/htm/z4.htm	10
http://www.comune.prato.it/concorsi/	2
http://www.comune.prato.it/concorsi/htm/z301.htm	15

PARTE II - ht://Miner

Nasce un nuovo prodotto open-source: ht://Miner

- Esperienza pluriennale nel campo delle statistiche degli accessi ai server web
- Esperienza in campo accademico da parte del sottoscritto:
  - Semestre di ricerca su data mining c/o Monash University, Melbourne, Australia
  - Laurea in Statistica (Università degli Studi di Firenze) con particolare interesse al data mining (web usage mining)
  - Progettazione di un prototipo sperimentale di web usage mining
- Esperienza di programmazione open-source in C++
- Esperienza di modellazione di basi di dati relazionali
- Ricerca nel settore del data warehousing
- Conoscenza di PostgreSQL e di sistemi operativi open-source (GNU/Linux)
- Contesto: Rete Civica Po-Net

Tutti questi fattori portano alla nascita nel 2003 del progetto sperimentale ht://Miner da parte del Comune di Prato all'interno della rete civica di Prato

Il progetto è distribuito secondo la licenza GNU GPL 2.0
ed è scaricabile liberamente dal sito di SourceForge.net

Prato e la sua rete civica: Po-Net

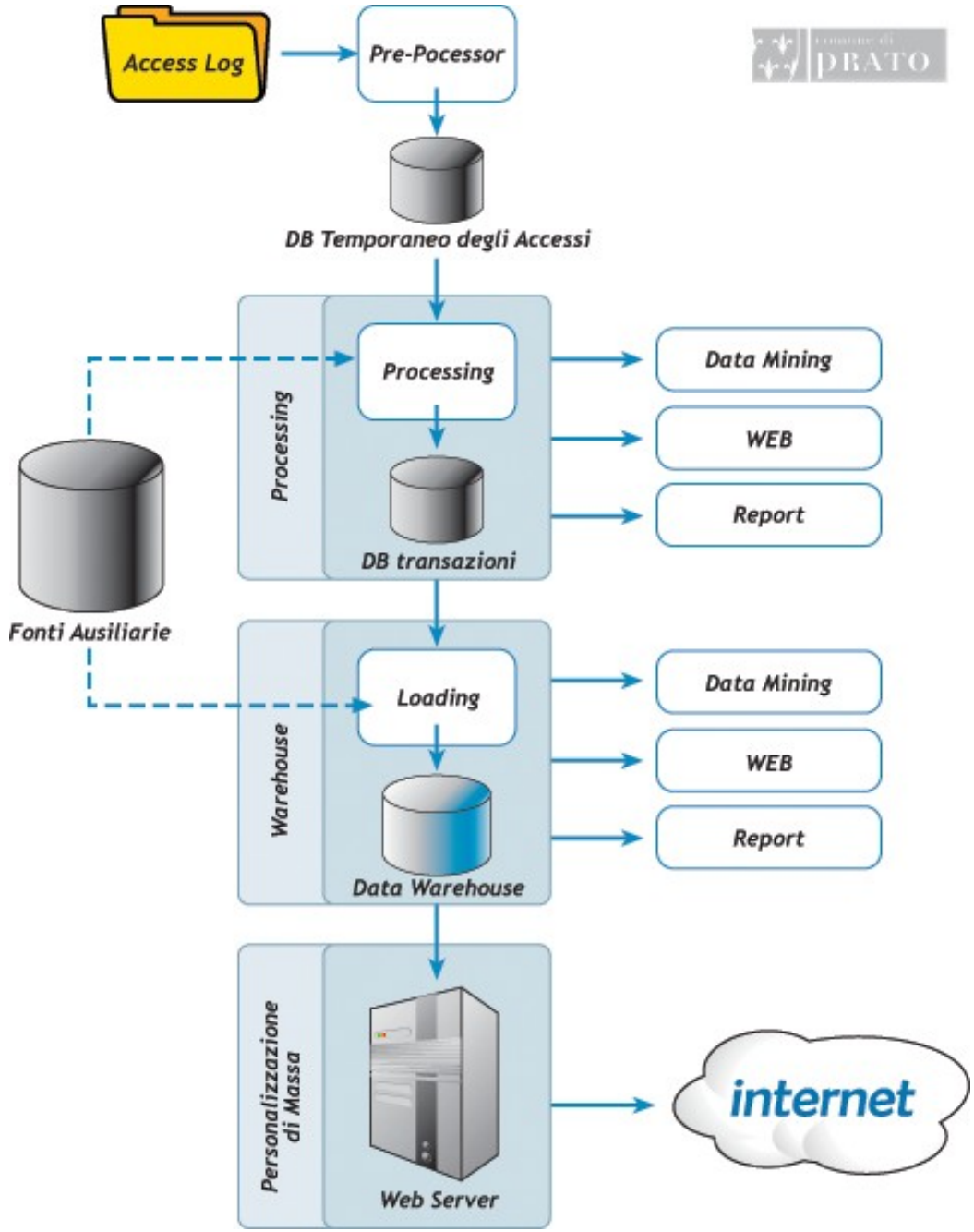
- Popolazione residente nella provincia di Prato al 31/12/2005: 242.497
- Progetto Po-Net nasce nel 1995 (fra i primi in Italia) e vede coinvolti:
  - Comuni della Provincia (7)
  - Provincia
  - Prefettura
  - Camera di Commercio
  - ASL
  - Aziende a partecipazione pubblica
  - Biblioteche
  - Musei
  - Istituzioni culturali
  - Scuole
  - Associazioni
- Fa parte di RTRT (Rete Telematica Regionale Toscana)
- E' composto da 140 gruppi di lavoro e redazioni
- Il coordinatore è l'ufficio Rete Civica del Comune di Prato
- **40.913.165 accessi ai siti web di Po-Net dal 1 gennaio al 21 ottobre 2007**
- **ht://Miner è utilizzato ufficialmente dal 1 gennaio 2007**

Alcuni fra gli obiettivi di ht://Miner

- Memorizzazione automatica delle informazioni in un RDBMS
- Creazione di un data warehouse per il **supporto alle decisioni** (DSS) sul comportamento degli utenti di uno o più siti web
- Predisposizione al data mining:
  - Regole associative (*market basket analysis*)
  - Path analysis (simile al concetto di *conversion*)
  - Clustering
- Individuazione automatica dei visitatori unici, delle sessioni, delle transazioni e del tempo speso
- Rilevazione supervisionata e automatica degli spider
- Supporto per la localizzazione degli indirizzi IP tramite GeoIP
- Classificazione delle URL in strutture gerarchiche organizzate a categorie
- Creazione di un framework di astrazione in PHP (libreria middleware) per l'interrogazione online
- Integrazione con fonti di dati ausiliarie
- Produzione di report con le più diffuse metriche di analisi degli accessi (visite, richieste di pagine, browser e sistemi operativi utilizzati, provenienza, ecc.)
- Rispetto e garanzia della privacy (utenti anonimi)
- ...

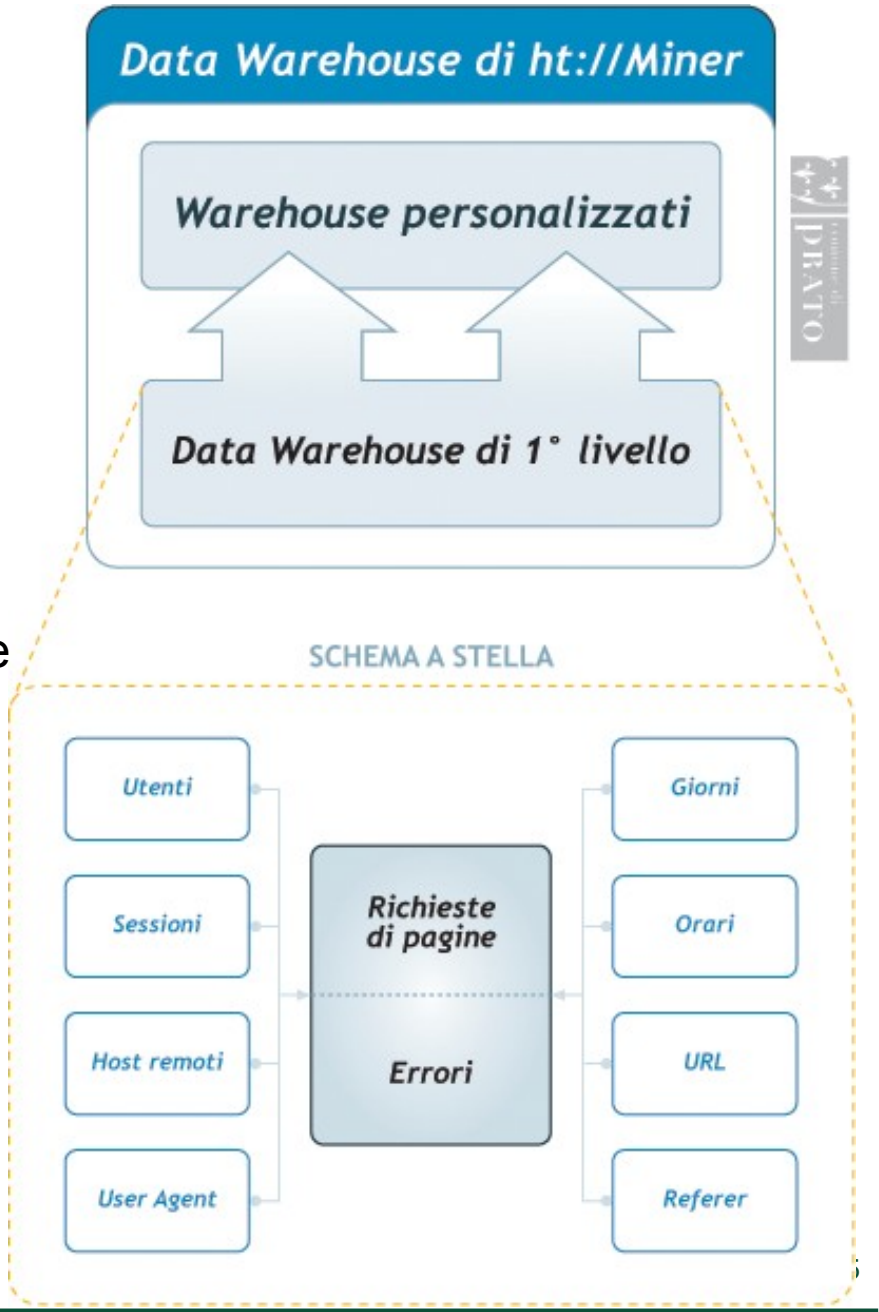
Architettura di ht://Miner

- Architettura a stack su 5 livelli:
  - Pre-processing
  - Processing
  - **Data warehousing**
  - Analisi:
    - **Data mining**
    - Report
    - Interrogazione via Web
  - *Personalizzazione di massa*



Data Warehouse

- Data warehouse di primo livello:
  - Schema a stella *(star schema)*:
    - Fatti
    - Dimensioni
  - Memorizza i *subject*:
    - Richieste di pagine
    - Richieste di errori
- Data warehouse (mart) di secondo livello modellati sulla base delle esigenze specifiche a partire dal data warehouse di primo livello



Un esempio di data warehouse personalizzato (data mart)

- **Stato di sviluppo:** stabile
- **Punto di partenza:** data warehouse di primo livello
- **Obiettivi:** creare uno o più data mart (sottoinsiemi) con accessi giornalieri e mensili organizzati logicamente in una o più gerarchie di categorie
- **Task:**
  - Classificare le URL in gerarchie di categorie (struttura ad albero) tramite espressioni regolari
  - Mantenere aggiornata la gerarchia nel tempo
  - Permettere la creazione e la navigazione automatica di data mart con diverso grado di aggregazione (per motivi di efficienza)
- **Caso d'uso:**
  - http://statistiche.po-net.prato.it/ (dal 1 gennaio 2007)

Data mining: una sperimentazione di *market basket analysis*

- **Stato di sviluppo:** alpha
- **Punto di partenza:** data warehouse di secondo livello
- **Obiettivi:** memorizzare e recuperare regole associative semplici del tipo:

SE "Servizi comunali del Comune di Prato"
ALLORA "Statuto e regolamenti del Comune di Prato"
con confidenza pari a 11,57%

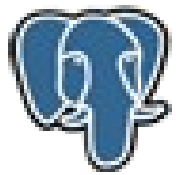
con query del tipo:

```
SELECT categoria1, categoria2, confidenza FROM arules WHERE confidenza >
   0 AND supporto > 0.01 ORDER BY confidenza desc;
```

```
  categoria1      |  categoria2     | confidenza
------------------+-----------------+------------
 Ist. F. Datini   | Bibl. Datini    |     0.8825
 Canale giovani   | E-move          |     0.8777
 Catalogo Prov.   | Biblioteche     |     0.8750
...
```

PARTE III - Conclusioni

ht://Miner e PostgreSQL



- **PostgreSQL** è il più avanzato sistema per la gestione di database open-source
- RDBMS di classe *enterprise*
- Multipiattaforma
- Supporta stabilmente:
  - Integrità referenziale
  - **Transazioni**
  - Viste
  - Stored Procedure
  - Sub-query
- Mette a disposizione gli **schemi**
- Mette a disposizioni i **tablespace** (fondamentali per il data warehousing)
- Ha una gestione degli indici molto efficiente anche con database di grandi dimensioni
- 17.10.07: sul server di Po-Net, ht://Miner ha un database su PostgreSQL con:
  - 189 tabelle e 725 indici
  - 229.436.843 record
  - 125 GB di spazio fisico occupato (su due tablespace)

Riferimenti e notizie

Riferimenti Utili:

- Sito principale di presentazione del progetto
  - http://www.htminer.it

- Sito di sviluppo su Sourceforge.net
  - http://www.sourceforge.net/projects/htminer

Novità in arrivo per la comunità italiana di PostgreSQL:

- **Italian PostgreSQL Users Group (IT-PUG):**
  - http://www.itpug.org/
- **PostgreSQL 8.3:**
  - http://www.postgresql.org/

Domande?

GRAZIE

Licenza della presentazione: Creative Commons 3.0 BY NC SA