MATRIX

Implementazione dei sistemi di Analytics con Netezza TF6

Data	**05/06/2012**	**Versione 1.1**

TELECOM ITALIA

Chi è Matrix - Audience e Tecnologia per la Comunicazione Digitale

➔**14,8 Milioni utenti unici/mese** su Virgilio (Fonte:Audiweb View EIA marzo 2011)

➔**7,6 Milioni clienti TIM** per direct marketing su Mobile

➔**2.5 Milioni** utenti web su TIM mobile portal e Virgilio Mobile

➔**8 Milioni chiamanti/anno** al voice directory assistance

Virgilio è il primo portale italiano: 14,8M Utenti/mese, 55M PV/day

E’ un portale orizzontale con servizi e contenuti altamente qualificati

- 70K Blog attivi (+ 30K Siti Personali)
- 5 Mio Mailbox attive
- 2° motore di ricerca Italiano
- Notizie, Sport, Viaggi, Auto, Lifestyle, Green, ecc
- Social News, Social Knowledge (Q&A)
- Canale Musica : Testi & Karaoke
- Video sharing platform
- Portali Locali: 8103 siti, uno per ogni comune italiano, con informazione e servizi locali

Architettura BI 2008 (Oracle)

- Applicazioni di Business Intelligence distribuita su vari sistemi
- Volumi di dati e tempi di elaborazione non soddisfacenti
- Analisi disgiunte in ambito Virgilio

Home Page Virgilio
6mio pv/giorno
Tempi di caricamento e calcolo delle visite:12 ore

Architettura BI 2009 – 2011 (Mustang 10050)

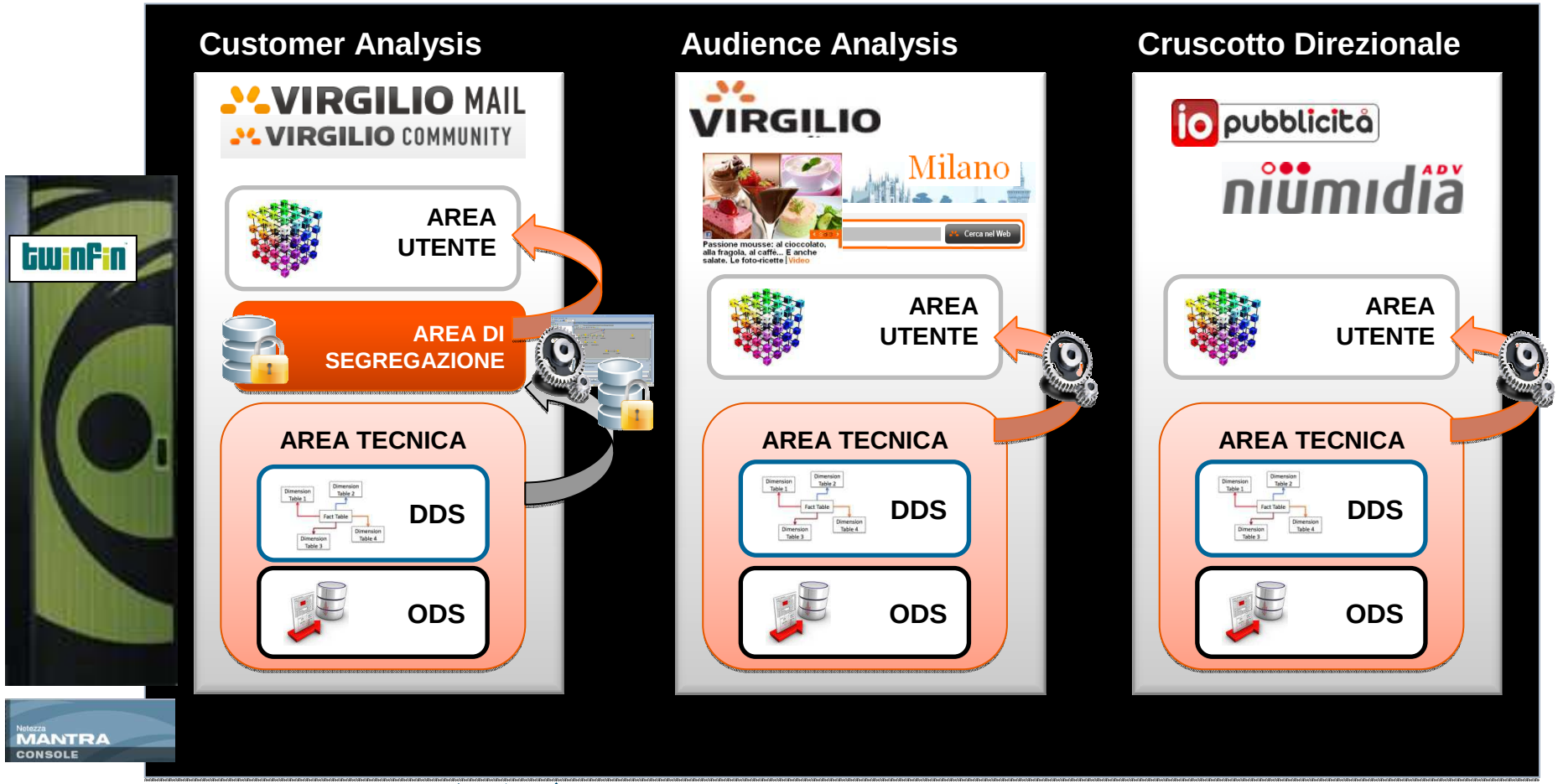
- Progetto di revisione e consolidamento dei sistemi di Business Intelligence con il coinvolgimenti del partner ICARE
- Web Analytics/Path Analysis
- Necessità di analisi cross sui diversi ambiti
- Incremento dei volumi
- Incremento delle prestazioni

Home Page Virgilio + tutti i canali
22mio pv/giorno
Tempi di caricamento e calcolo delle visite: 6 ore
Click Virgilio
Tempi di caricamento ed elaborazione: 8 ore

Nuova architettura BI 2011 (Twinfin 6)

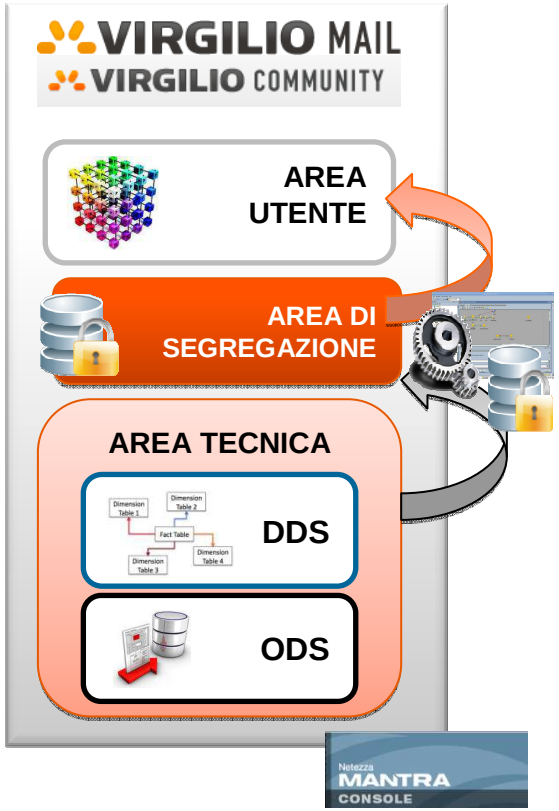
- Revisione architettura logica della BI
- Adeguamento alle linee guida di Telecom Italia sui sistemi di DWH
- Evoluzione dell'analisi editoriale e commerciale dell'audience (Clustering/Profilazione)
- Evoluzione della reportistica direzionale
- Utilizzo delle funzioni analitiche native di Netezza
- Più capacità di calcolo per le analisi degli utenti finali

Nuova architettura 2011 (Twinfin 6)



Piattaforma Mail	Contenuti Virgilio	Ordini
Piattaforme Community	Ricerca Virgilio	Web analytics	Billing	Gestionale	SDP
	Erogato ADV	Erogato ADV

• Dai dati personali aggregati non deve essere possibile risalire immediatamente ai dati di dettaglio da cui derivano (Architettura dei sistemi DWH/Profilazione a layer) secondo il principio del "need to know".
• I dati personali identificativi trattati dai sistemi di profilazione devono essere mascherati in area utente



• Nessuna utenza nominale accede alle tre aree, ma solo l'utenza applicativa dell'area di segregazione
• I dati vengono caricati in una tabella temporanea per il mascheramento on-the fly utilizzando 3 UDF, una per le stringhe, una per le mail ed una per gli interi
• Le UDF implementano l'algoritmo di "De Vigenere" scritto in C++ e compilato su Netezza
• La chiave non è facilmente reperebile e non passa in chiaro nei log
• Non viene gestito alcun incrementale ma ogni giorno viene eseguito un full
• L'accesso e le query sono monitorati e loggati con Mantra

AREA TECNICA

SPAZIO OCCUPATO	0,9 TB
RECORDS TOTALI	24 MRD
RECORDS GIORNALIERI	110 MIO
TEMPO DI ELABORAZIONE	2h
N.TABELLE	<100

RIBALTAMENTO IN AREA UTENTE
250 GB
TEMPO DI RIBALTAMENTO 30 min

MASCHERAMENTO IN AREA UTENTE
100GB
TEMPO DI MASCHERAMENTO 1h

ANALISI

•Clustering/Profilazione interessi/sociodemo

• Analisi dell'audience/Bacini ADV

• Ottimizzazione contenuti

• Clickstream GEO

STRUMENTI

• Oracle (Reporting e anailsi libera)

• SAS

• R-REvolution

Clustering

• Preparazione delle tabelle base . Si movimentano circa 900.000.000 di record in circa 1 ora.

• Profilazione per Interessi : per ogni utente vengono individuate 6 categorie di interesse. La profilazione avviene quotidianamente con una durata di circa ½ ora

• Profilazione Socio-Demografica : ad ogni utente viene attribuito sesso, fascia età livello istruzione . La profilazione avviene quotidianamente con una durata di circa 2 h

Click Home Page Virgilio

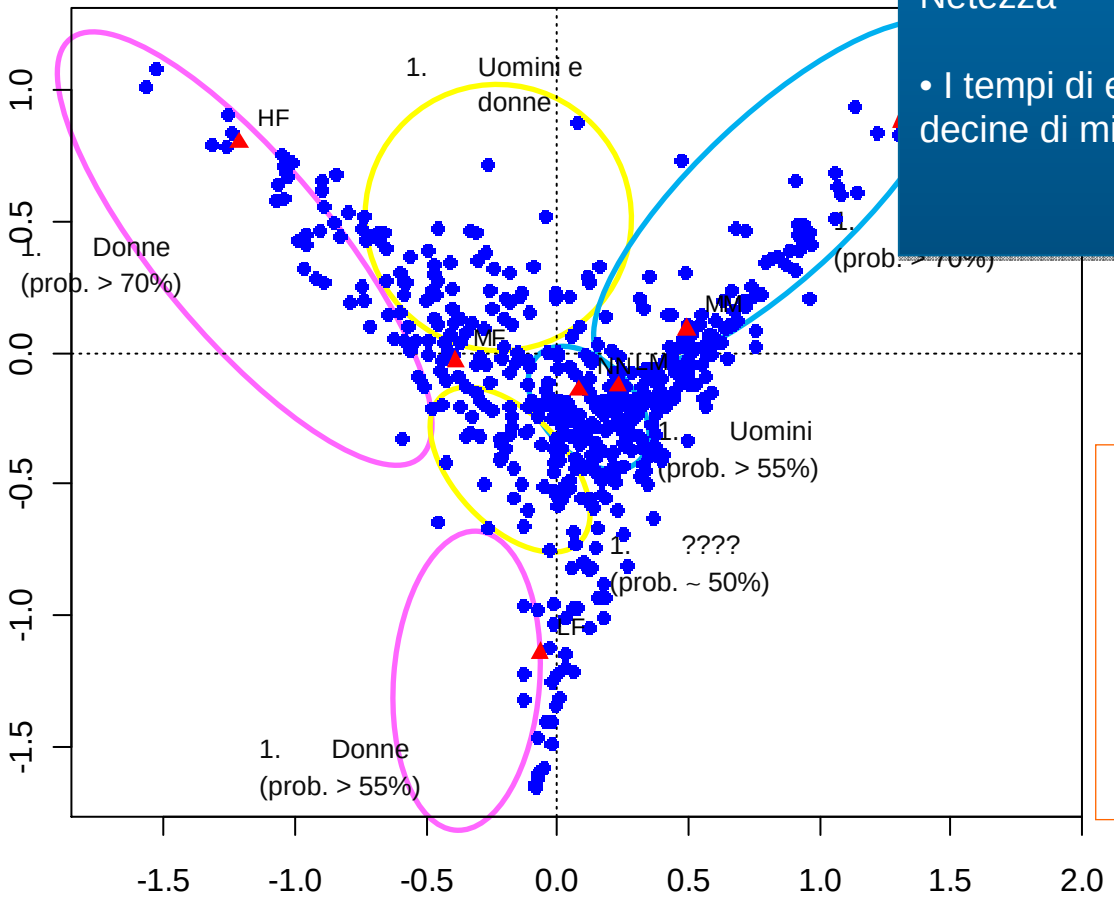
Tempi di caricamento ed elaborazione 50 minuti

TOTALI	
SPAZIO OCCUPATO	2 TB
RECORDS TOTALI	75 MRD
BYTES GIORNALIERI	10 GB
TEMPO DI ELABORAZIONE	5h
N.TABELLE	230

SOCIODEMO CON R

•Questo prototipo di clustering sociodemo si basa sull'analisi delle corrisponenze implementata con R utilizzando le librerie di funzioni analitiche native di Netezza

• I tempi di elaborazione sono nell'ordine delle decine di minuti



7 Profili:
- •HM: altamente Maschili
- •M: Maschili
- •N: Neutri
- •F: femminili
- •HF: altamente Femminili
- •LF: leggermente Femminili
- •LM: leggermente Maschili

ANALISI
- Performance/Venduto
- Programmazione commerciale
- Giacenze
- Profit & Loss

STRUMENTI
- Oracle (Reporting e analisi libera)

TOTALI	
SPAZIO OCCUPATO	190 GB
RECORDS TOTALI	3,8 MRD
RECORDS GIORNALIERI	38 MIO
TEMPO DI ELABORAZIONE	1h
N.TABELLE	180