Standard datawarehouse
Il disegno fisico degli ambienti

INDICE

Introduzione

Il presente documento fa parte della collezione di documenti "data warehouse", che raccoglie le indicazioni specifiche delle varie componenti sotto un unico indirizzo, separando al contempo le varie tematiche, al fine di proporre una più agevole lettura e meglio indirizzare le necessità di aggiornamento dei singoli documenti.

I documenti che compongono la collezione sono:

Architettura Data warehouse
Descrive i concetti base del Data Warehouse e sottolinea le differenze di obiettivi e caratteristiche rispetto ai sistemi di tipo gestionale. E' è orientato soprattutto alla definizione di standard tecnologici, pur delineando brevemente anche gli aspetti di approccio metodologico legati allo sviluppo di progetti di Data Warehouse.
Per le tre componenti acquisizione di dati, gestione e distribuzione sono descritti i requisiti e le funzionalità richieste, e vengono inoltre indicati anche i prodotti di mercato che meglio rispondono alle esigenze e si integrano con il sistema informativo del Ministero.

Indicazioni per lo sviluppo e realizzazione di progetti di Data warehouse
Il documento, rivolto in modo particolare ai responsabili dei progetti di Data Mart, intende rappresentare un ausilio concreto per coloro che partecipano alle singole iniziative progettuali del Ministero e non una linea guida teorica. Lo scopo è quello di affrontare lo sviluppo e la realizzazione delle inziative di natura informativa in corso presso il Ministero in modo integrato e uniforme, in coerenza con le scelte architetturali.

Il disegno fisico degli ambienti
E' il presente documento.

1.1 *Destinatari del documento*

Il documento è rivolto a tutti coloro che, operando nell'ambito del Ministero del Tesoro, del Bilancio e della Programmazione Economica, sono interessati alla realizzazione di progetti di Data Warehouse.

1.2 *Scopo del documento*

Lo scopo del presente documento è quello di fornire un'indicazione sulla configurazione e dislocazione degli ambienti di data warehouse e data mart.

1.3 *Fonti*

Le scelte indicate nel presente documento sono frutto degli studi effettuati da Consip. L'infrastruttura è stata congruita da Aipa con parerie n. 5 e 6 del 27 gennaio 2000.

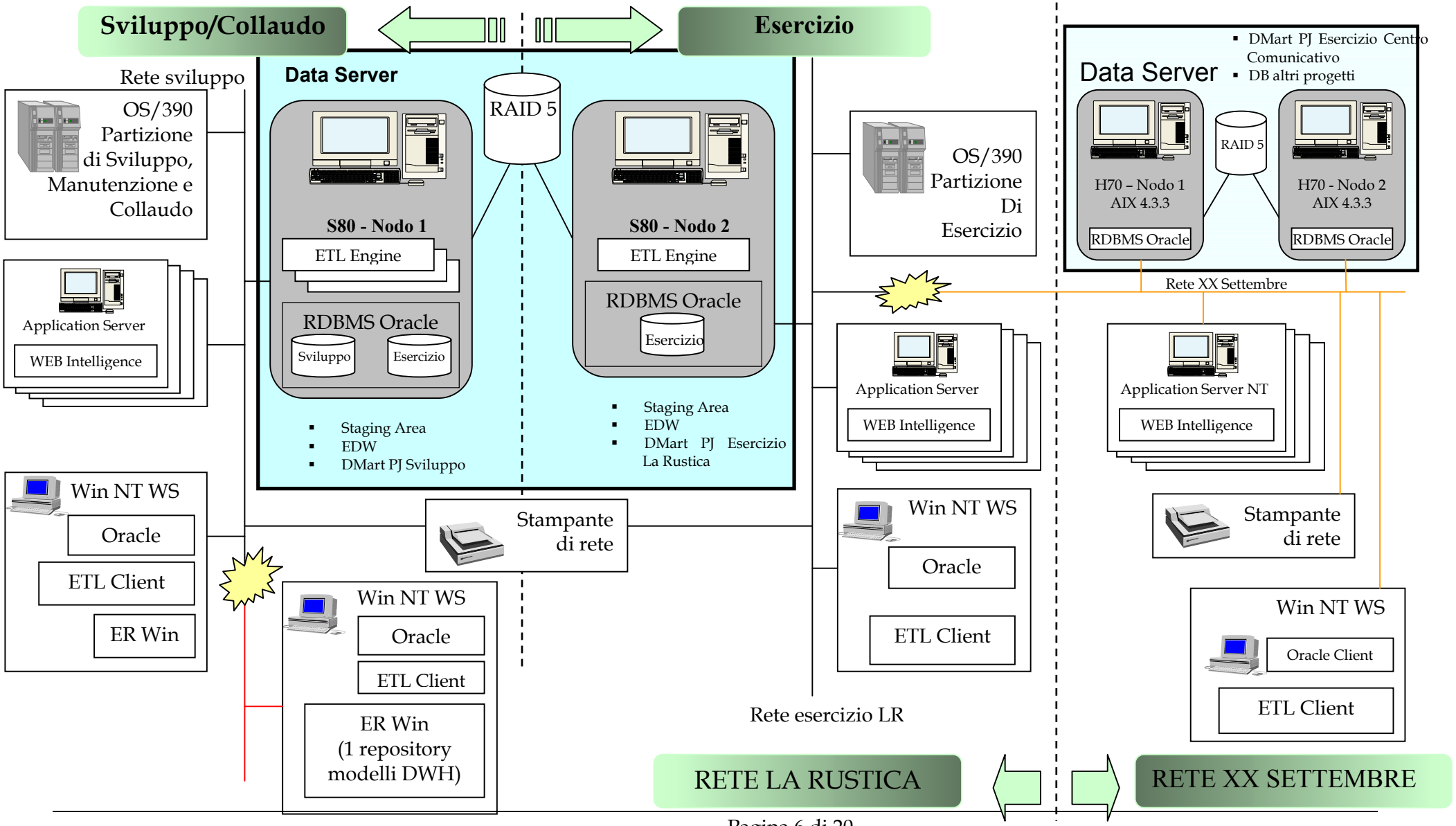
Disegno del sistema

Nel disegno in allegato è rappresentata l'infrastruttura dell'intero sistema. Sono previsti due ambienti, uno presso il CED di La Rustica e uno presso la sede del Ministero di Via XX settembre.

L'ambiente di La Rustica è dedicato ad ospitare le aree tecniche (cosiddette aree di staging) e l'Enterprise Data Warehouse, non accessibili direttamente agli utenti finali.
Nella scelta dell'ubicazione si è tenuto conto che l'alimentazione delle aree di staging e dell'Enterprise Data Warehouse avviene principalmente a partire dai sistemi gestionali ospitati su piattaforma mainframe; pertanto le operazioni di alimentazione del sistema beneficeranno di un collegamento via canale tra i due sistemi.

Tale ambiente rappresenta inoltre, il punto di ingresso e di uscita delle informazioni verso e dal Ministero, consentendo la razionalizzazione dei flussi esistenti e garantendo univocità e correttezza dell'interpretazione del significato semantico dei dati.
L'ambiente di La Rustica ospiterà poi, in modo dedicato, singoli progetti di Data Mart, le cui dimensioni e requisiti di sicurezza logica e fisica richiedono un adeguato presidio gestionale e sistemistico.

L'ambiente di via XX settembre sarà invece dedicato all'ambiente di esercizio dei singoli progetti di Data Mart. In questo caso, l'ubicazione del sistema intende privilegiare le prestazioni di accesso da parte degli utenti che risiedono nella stessa sede.


Sviluppo/Collaudo
Esercizio
Rete sviluppo
OS/390 Partizione di Sviluppo, Manutenzione e Collaudo
Application Server
WEB Intelligence
Win NT WS
Oracle
ETL Client
ER Win
Data Server
RAID 5
S80 - Nodo 1
ETL Engine
RDBMS Oracle
Sviluppo
Esercizio
S80 - Nodo 2
ETL Engine
RDBMS Oracle
Esercizio
Staging Area
EDW
DMart PJ Sviluppo
Staging Area
EDW
DMart PJ Esercizio La Rustica
Stampante di rete
Win NT WS
Oracle
ETL Client
ER Win (1 repository modelli DWH)
OS/390 Partizione Di Esercizio
Application Server
WEB Intelligence
Win NT WS
Oracle
ETL Client
Rete esercizio LR
Data Server
DMart PJ Esercizio Centro Comunicativo
DB altri progetti
H70 – Nodo 1 AIX 4.3.3
RAID 5
H70 - Nodo 2 AIX 4.3.3
RDBMS Oracle
RDBMS Oracle
Rete XX Settembre
Application Server NT
WEB Intelligence
Stampante di rete
Win NT WS
Oracle Client
ETL Client
RETE LA RUSTICA
RETE XX SETTEMBRE

Sono stati previsti due Data Server, dislocati uno presso il CED di La Rustica ed uno presso il Centro Comunicativo di via XX Settembre.

I due Data Server sono preposti a svolgere funzioni leggermente diverse ma entrambi devono garantire la costante affidabilità e disponibilità del sistema. Per questo motivo è stata scelta una architettura di tipo Cluster SMP a due nodi.

Il sistema dislocato alla Rustica rappresenterà il Data Server per l'Enterprise Data Warehouse e per i Data Mart dei progetti che abbiano un significativo volume di dati ed un utenza ampia.

L'ambiente ospiterà, inoltre, le aree di staging comuni, cioè aree tecniche, non accedibili direttamente dall'utente finale. Su queste aree verranno consolidati gli output dei processi di estrazione dai vari sistemi sorgente e centralizzate le operazioni di cleaning e trasformazione preliminari al popolamento dei Data Mart .

I due nodi Unix ospiteranno rispettivamente gli ambienti di sviluppo e di esercizio. In caso di malfunzionamento del nodo di esercizio il software di clustering sarà configurato in modo da trasferire automaticamente l'ambiente di esercizio sul nodo di sviluppo che ne prende il posto.

Le macchine impiegate sono di tipo IBM S80 con sistema operativo AIX 4.3.3 con 8 processori e 8GB di RAM ciascuna. La memoria di massa è costituita da un sistema di 30 dischi esterni configurati in architettura RAID5 e collegati ad entrambi i nodi.

Presso il Centro Comunicativo risiederà un altro ambiente applicativo volto all'elaborazione di Datamart di piccole dimensioni e con tipologia di utenza ridotta e, comunque, locale. Le caratteristiche architetturali di tale ambiente sono simili al precedente; i due nodi Unix, di fascia inferiore rispetto ai nodi richiesti per il CED di "La Rustica", svolgeranno funzioni di database server.

Dal punto di vista applicativo i due nodi sono indipendenti, servendo applicazioni diverse e montando dischi diversi. In caso di malfunzionamento di uno dei due nodi il software di clustering sarà configurato per spostare gli ambienti sull'altro e vi monta i relativi dischi. Ciascuno dei due nodi sarà dunque in grado di sostenere da solo tutto il carico di lavoro.

Le macchine impiegate sono di tipo IBM H70 con sistema operativo AIX 4.3.3 con 2 processori e 2GB di RAM ciascuna. La memoria di massa è costituita da un sistema di 15 dischi esterni configurati in architettura RAID5 e collegati ad entrambi i nodi.

Per entrambi i sistemi l'ambiente Datamart sarà implementato su una piattaforma hardware eterogenea secondo un'architettura a 3 livelli: il Database Server risiederà sui nodi Unix mentre il Web/Application Server sarà su piattaforma NT.

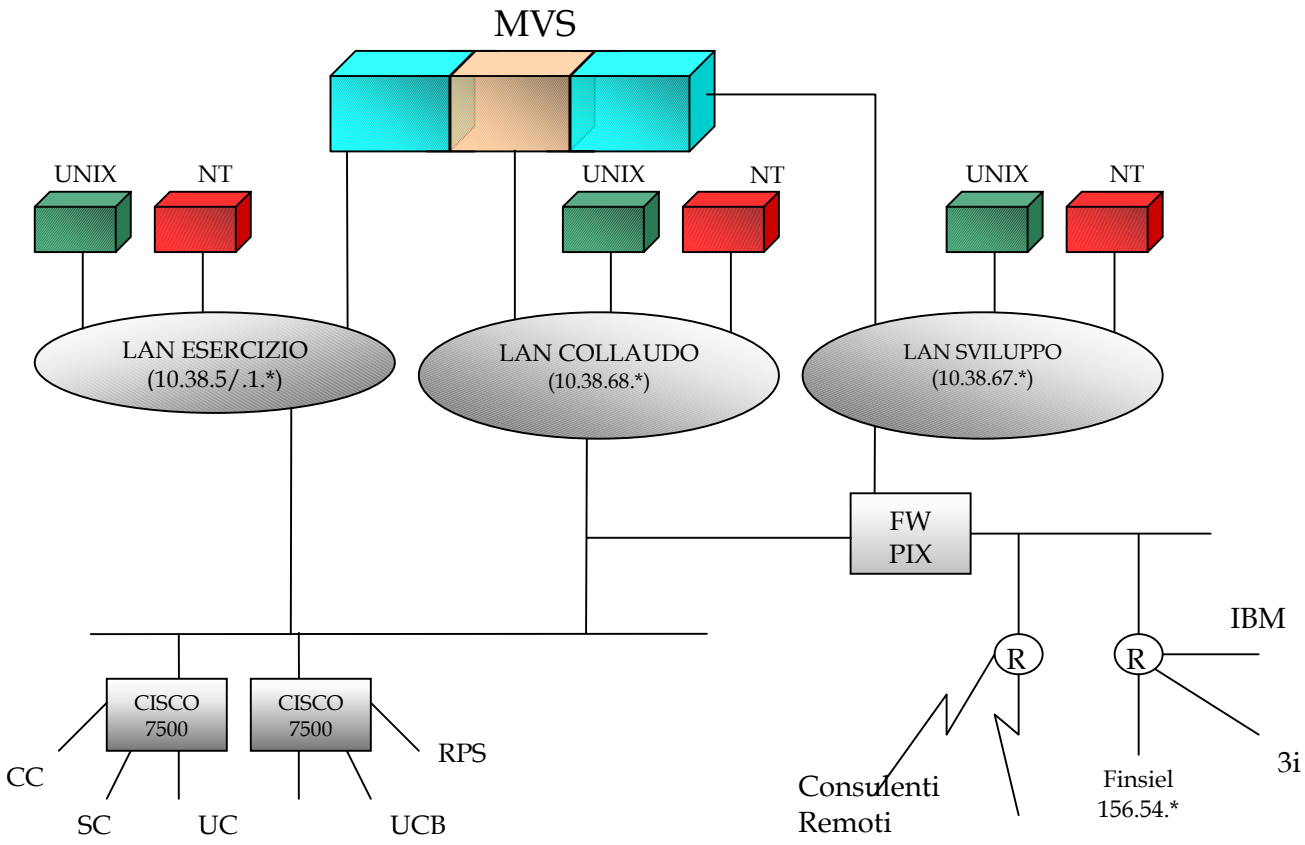
Per tale ambiente di sviluppo sono disponibili quattro macchine a La Rustica, quattro per l'ambiente di esercizio alla Rustica e quattro per l'ambiente di esercizio del Centro Comunicativo. Le macchine impiegate sono di tipo COMPAQ PROLIANT 5500.

Il database installato è Oracle 8. In più sui nodi Unix alla Rustica sono installati due motori ETL (PowerCenter).
I due ambienti di esercizio hanno a disposizione ciascuno una stampante di rete mod. EPSON 8200.
Le postazioni client, sia di La Rustica che di XX Settembre, sono PC MEGABYTE mod. REFERENCE con annessa stampante di tipo BROTHER HL 1250. Tre postazioni del centro comunicativo sono dotate di capacità grafiche.
Sui client sono installati Oracle client ed ETL client.

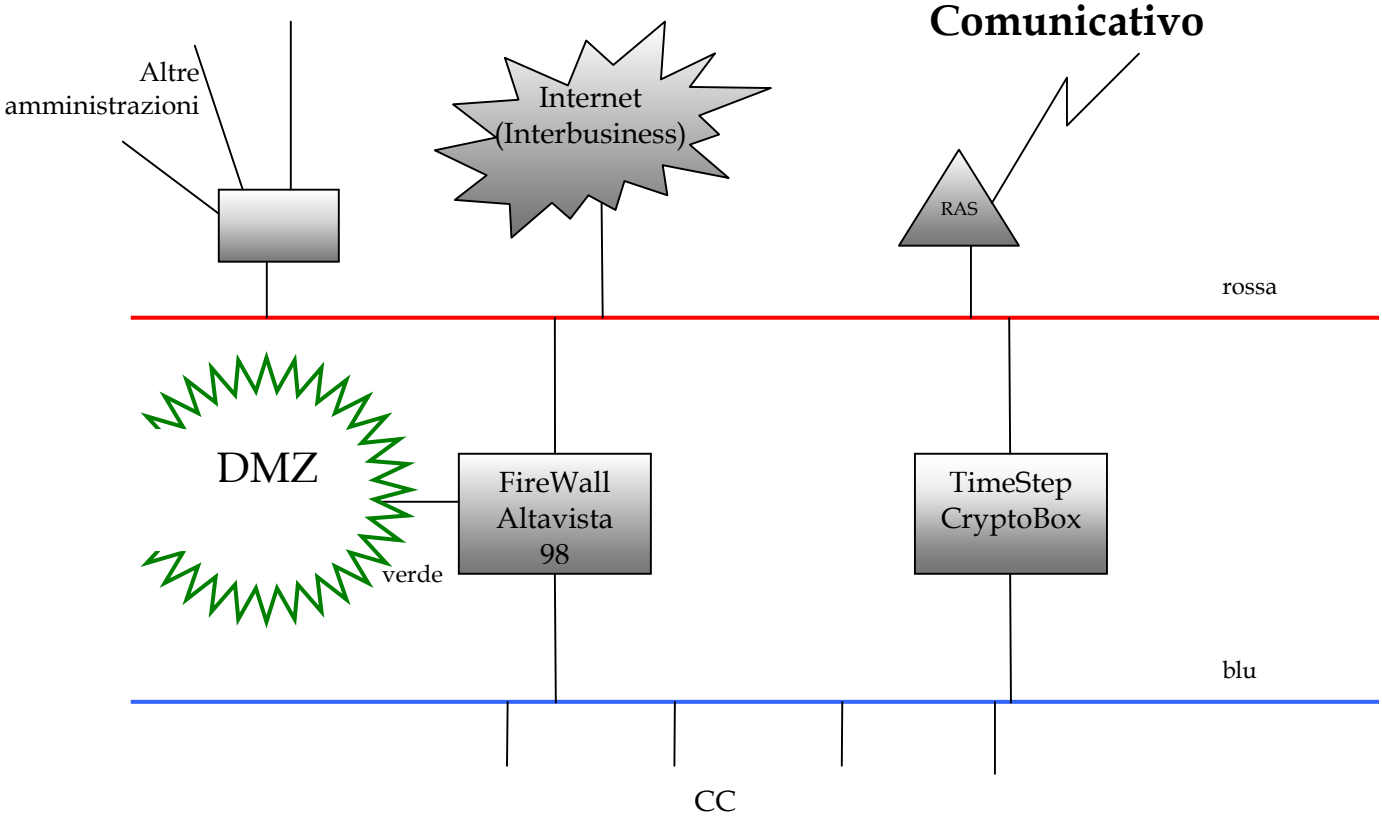
Sottosistema di rete

Di seguito sono riportate due figure schematiche che sintetizzano graficamente la situazione:

Schema logico rete La Rustica



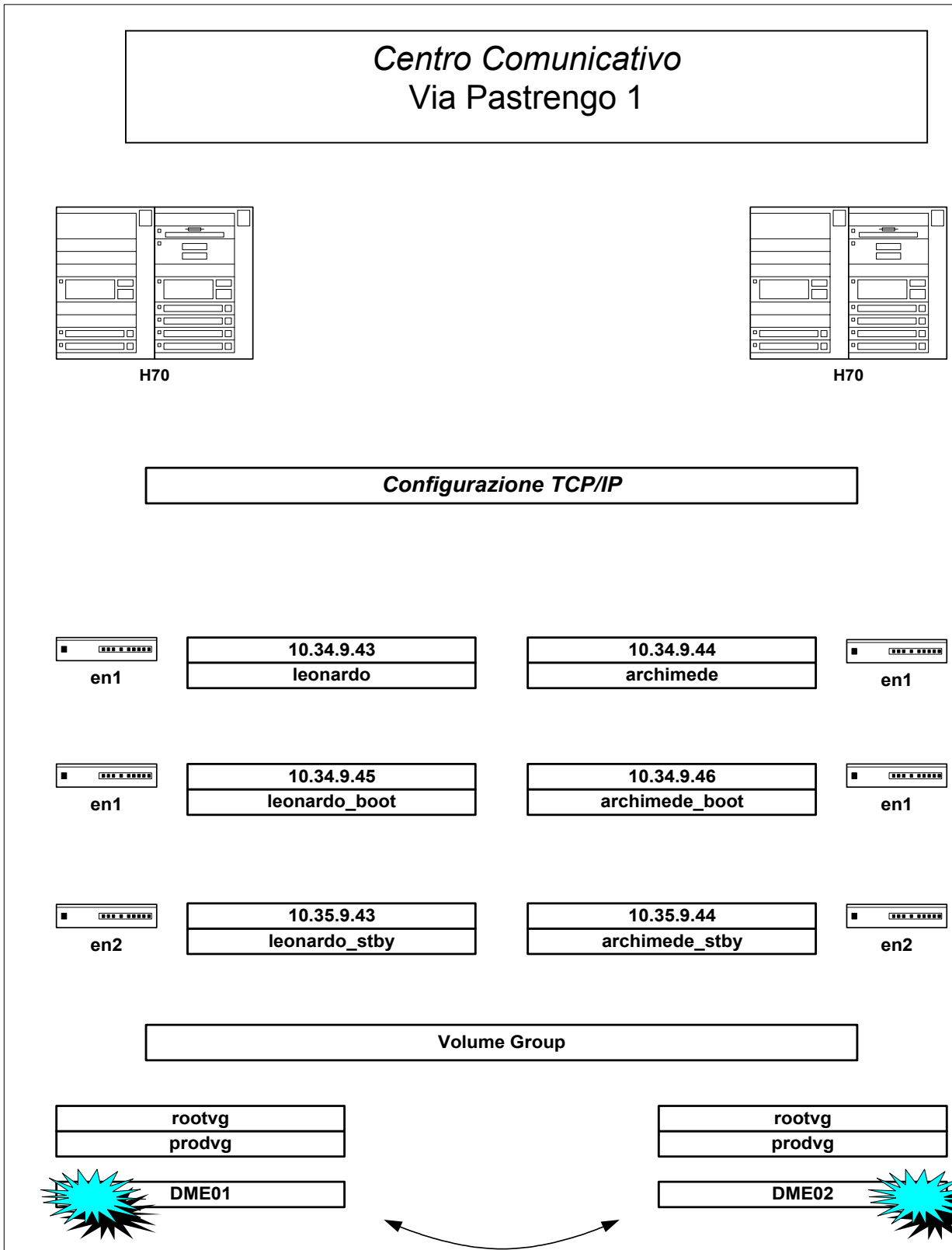
Schema logico rete Centro Comunicativo



L'installazione dei nodi Unix presso il CED de La Rustica avverrà sulla rete lan di esercizio e su quella di riferimento (collaudo e sviluppo), essendo la rete lan di sviluppo completamente dedicata all'attività dei fornitori.

Attualmente le due macchine sono collegate in rete con i seguenti indirizzi IP:
nodo zeus 10.38.5.241 e 10.38.67.3
nodo era 10.38.5.201

Presso il Centro Comunicativo l'istallazione dei nodi Unix è stata effettuata sul tratto di lan che attualmente garantisce i migliori requisiti sotto il profilo della sicurezza, ovvero quella convenzionalmente indicata come 'blu'. La configurazione IP delle due macchine, denominate Leonardo e Archimede, è rappresentata nella seguente figura:


Centro Comunicativo
Via Pastrengo 1
H70
H70
Configurazione TCP/IP
en1
10.34.9.43
leonardo
10.34.9.44
archimede
en1
en1
10.34.9.45
leonardo_boot
10.34.9.46
archimede_boot
en1
en2
10.35.9.43
leonardo_stby
10.35.9.44
archimede_stby
en2
Volume Group
rootvg
prodvg
DME01
rootvg
prodvg
DME02

Area sistema UNIX

Il Sistema operativo scelto è AIX 4.3.3. Sia al Centro Comunicativo che presso il CED de La Rustica i nodi Unix saranno installati in configurazione Cluster per mezzo del software HACMP.
Lo schema riassuntivo per il partizionamento delle macchine è il seguente:
Centro Comunicativo

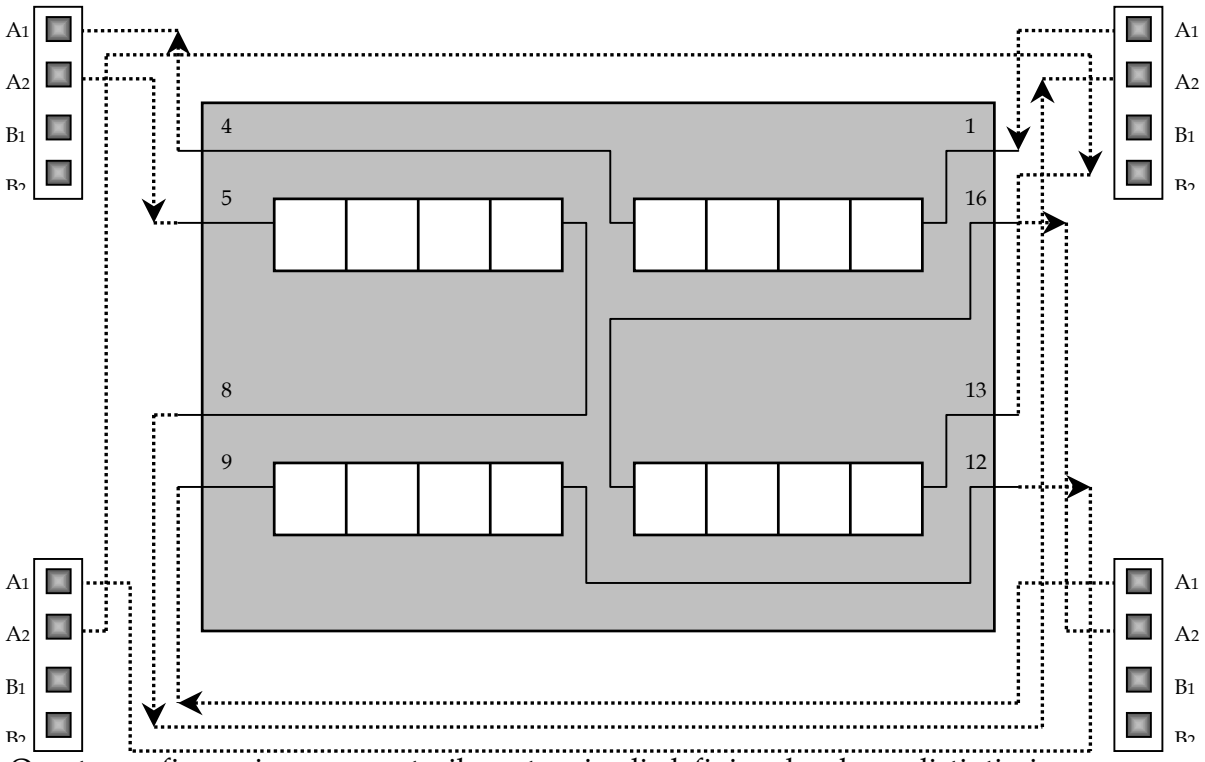
Complessivamente sono disponibili:
2 dischi interni da 9,1 GB per ciascun nodo
15 dischi esterni da 18,2 GB montati in un rack IBM 7133-D40 e condivisi dai due nodi.

Dischi interni:
Ciascuna macchina è configurata nella maniera seguente:

Rootvg (Swap)	9,1 GB	Prodvg (Swap)	9,1 GB

Almeno in prima istanza i dischi interni non verranno mirrorati.

Dischi esterni:
E' stata adottata ls seguente configurazione SSA.



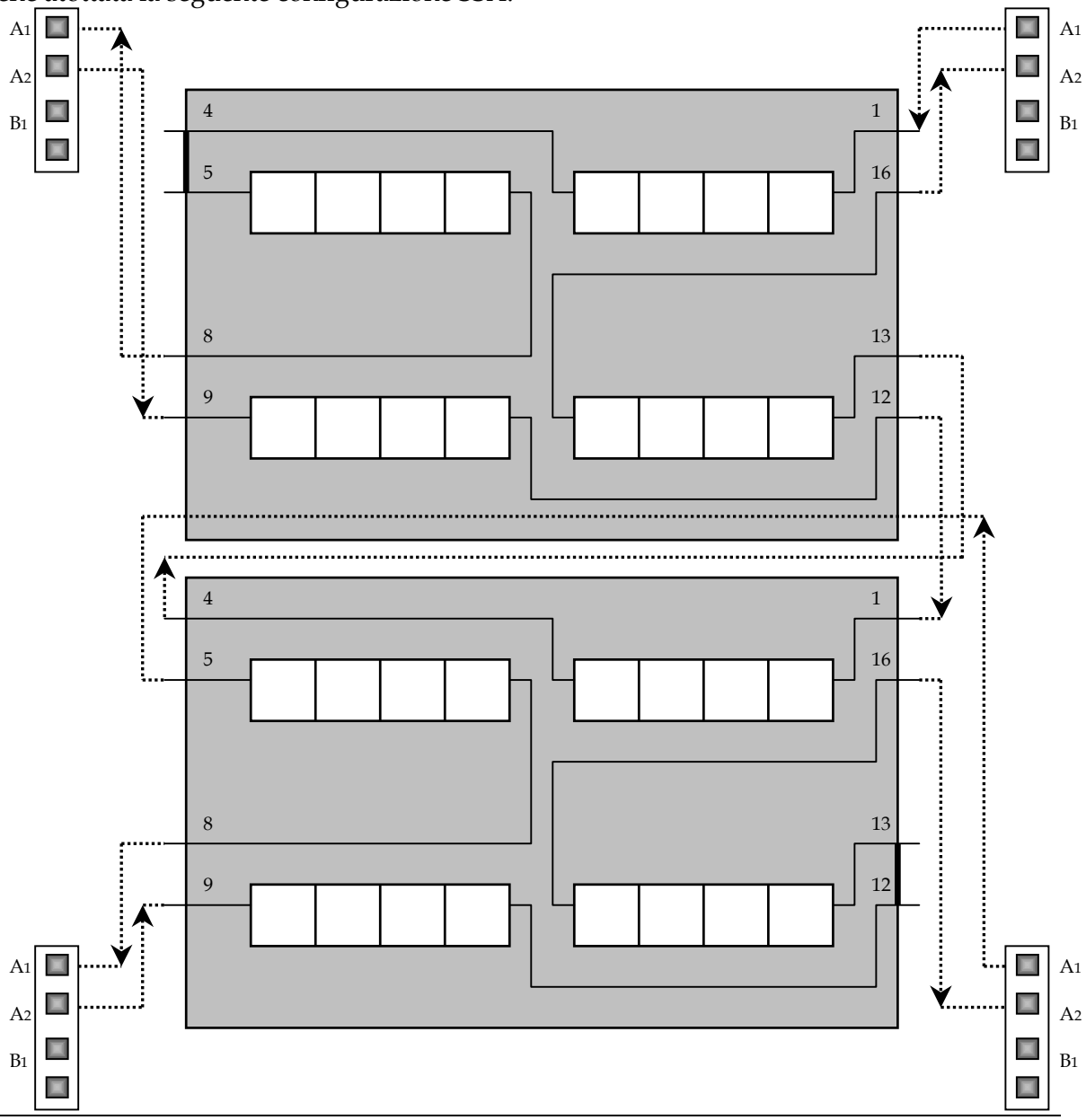
Questa configurazione presenta il vantaggio di definire due loop distinti ciascuno su un controller di una scheda diversa rendendo così possibile la creazione di due RAID5 indipendenti da 7 dischi più uno ciascuno di spare.
Sui due RAID5 sono realizzati altrettanti Volume Group:

RAIDvg1 (7 dischi)	109,2 GB	RAIDvg2 (7 dischi)	109,2 GB

La Rustica
Complessivamente sono disponibili:
2 dischi interni da 9,1 GB per ciascun nodo
30 dischi esterni da 18,2 GB montati in un rack IBM 7133-D40 e condivisi dai due nodi.
Dischi interni:
Ciascuna macchina è configurata nella maniera seguente:

Rootvg (Swap)	9,1 GB	Prodvg (Swap)	9,1 GB

Anche qui, inizialmente i dischi interni non verranno mirrorati.
Dischi esterni:
Viene atottata la seguente configurazione SSA.



Questa configurazione permette la realizzazione di due loop, uno comprendente 20 dischi ed uno 9. Il loop più grande sarà quello normalmente montato dal nodo di esercizio, l'altro quello montato dal nodo di sviluppo.

I 20 dischi del primo loop sono configurati nel modo di seguito descritto (un disco viene lasciato di spare):

RAIDvg1 (6 dischi)	91 GB	RAIDvg2 (5 dischi)	72,8 GB	RAIDvg3 (8 dischi)	127,4 GB

I 9 dischi del secondo loop sono utilizzati nel modo seguente (anche qui un disco di spare):

RAIDvg1 (8 dischi)	127,4 GB

Area Database

Come già indicato in precedenza si è scelto di utilizzare l'RDBMS Oracle, nella versione 8i (8.1.6). Il motore Oracle è installato sia a La Rustica che al Centro Comunicativo su entrambi i nodi dei cluster Unix. Il prodotto è installato sui dischi interni dei server, riservando lo spazio sui RAID5 ai soli datafiles.

Al Centro Comunicativo sono state implementate due istanze Oracle, una per nodo, che saranno dunque a completa disposizione delle applicazioni di Data Mart. Ogni singolo progetto avrà a disposizione un suo schema, dislocato su di un tablespace dedicato.

Le istanze conterranno anche uno schema dedicato al repository BO e WebI su di un tablespace dedicato, il cui dimensionamento è attualmente in fase di definizione.

Per quanto riguarda il CED de La Rustica, sull'ambiente di Sviluppo sono state inizialmente create tre istanze Oracle:

- una dedicata ai vari repository per i Power Center server. Ciascun progetto avrà a disposizione per il suo repository uno schema dedicato su di un apposito tablespace con circa 200Mb di spazio disco. Inizialmente si dovranno prevedere almeno cinque di questi ambienti più uno per il repository globale contenente le informazioni a carattere comune.
- una dedicata ai Data Mart di sviluppo. Anche in questo caso ciascun progetto di Data Mart avrà a disposizione un suo schema dedicato su di un apposito tablespace, con uno spazio disco che sarà assegnato caso per caso in base ai requisiti applicativi.
- una dedicata all'ambiente Data Warehouse ed alle aree di staging di sviluppo. Questa istanza sarà suddivisa su almeno due schema, una per le arre di staging, l'altra per l'Enterprise Data Warehouse vero e proprio.

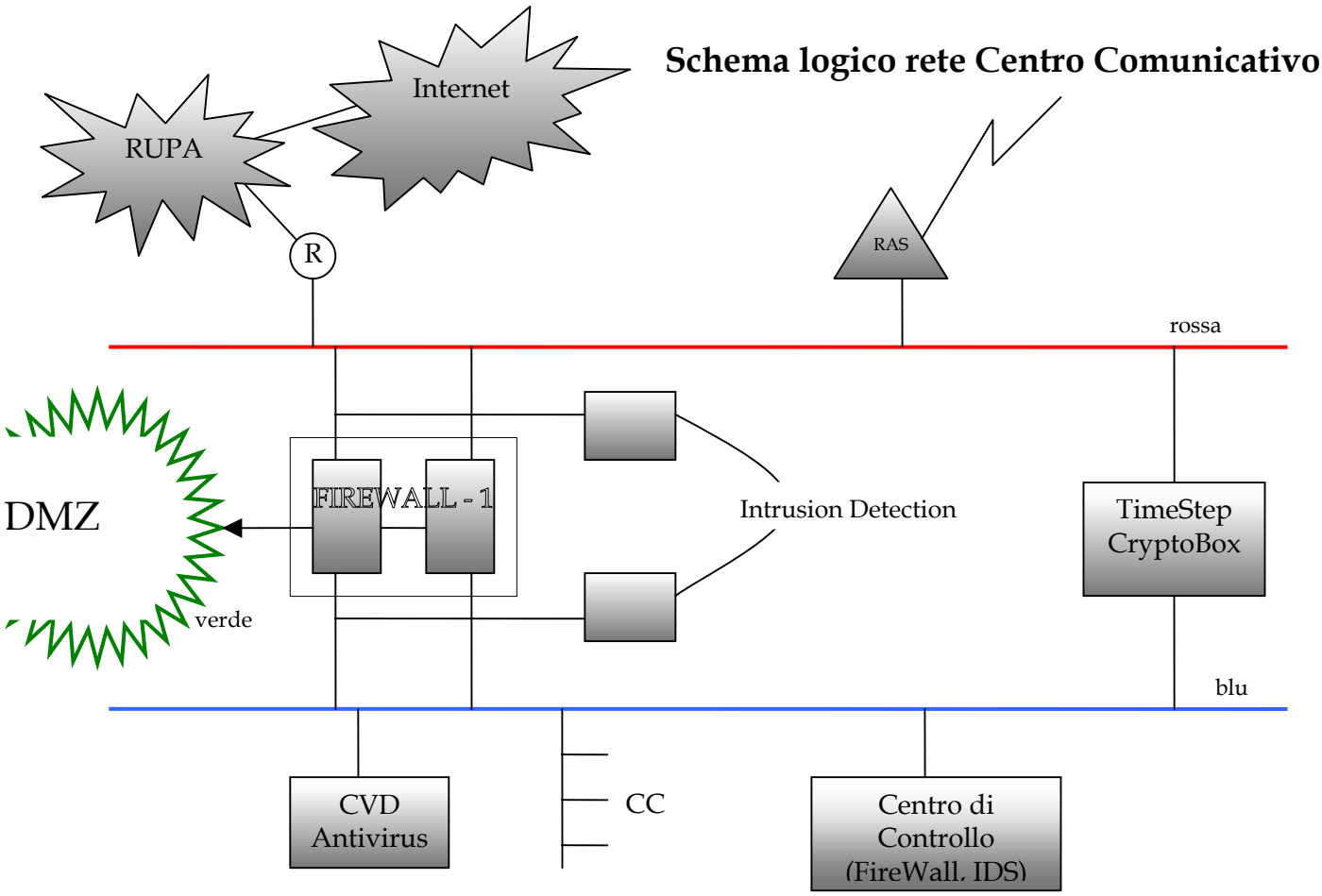
Per gli ambienti di collaudo non saranno create istanze di database dedicate. Tutte le operazioni relative saranno effettuate in apposite aree delle istanze di sviluppo.

La configurazione degli ambienti di esercizio ricalca quella degli ambienti di sviluppo, con la differenza che l'istanza dedicata ai repostitory conterrà tutti i metadati in un solo schema

Area Rete

Entro la fine dell'anno la configurazione dei sistemi passerà da quella descritta nelle pagine precedenti a quella riportata nei prossimi paragrafi.

1.4 *8.1. Il Centro Comunicativo*

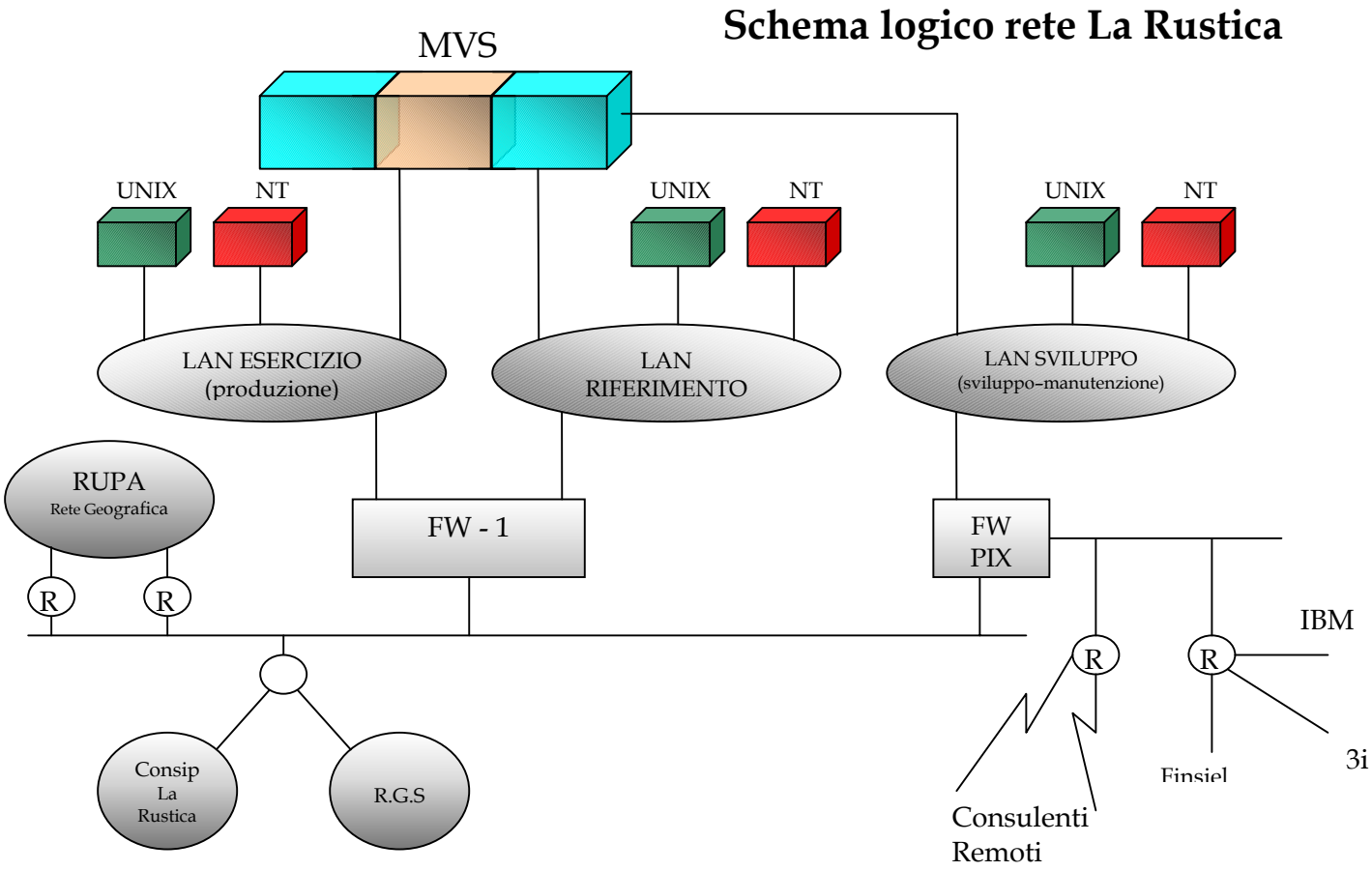


In particolare sono individuate tre zone logiche, differenziate per caratteristiche di sicurezza crescenti:

- Rossa, separata dalla Rupa per semplice interposizione di una router;
- Verde, a valle dei firewall (zona demilitarizzata, DMZ);
- Blu, a valle del sistema di intrusion detection e del timestep cryptobox.

Quest'ultima è la zona più sicura, ed è in questo tratto della lan del CC che saranno disposte tutte le macchine per il sistema di Data Warehouse.

8.2. La Rustica

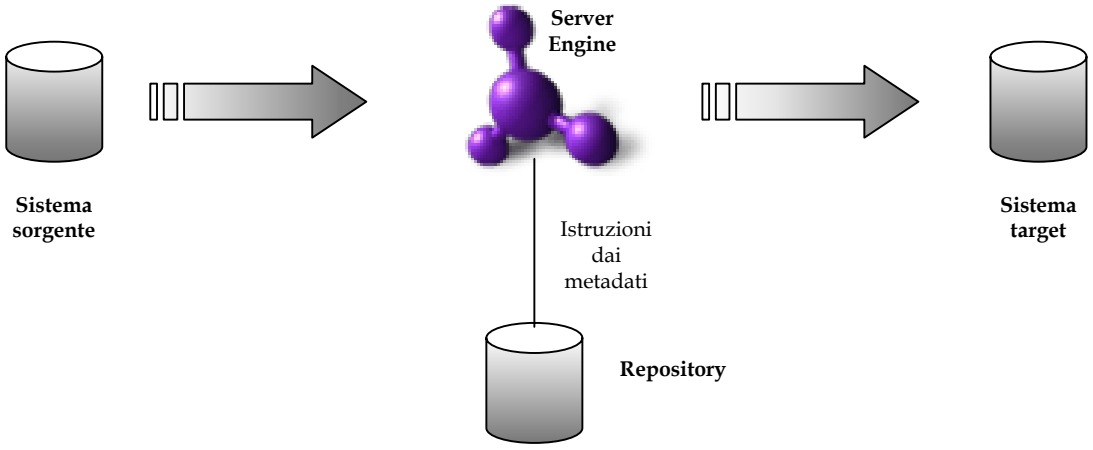


Nella nuova configurazione queste lan non saranno più direttamente esposte alla RUPA, ma protette dall'interposizione del Firewall-1.

Area ETL

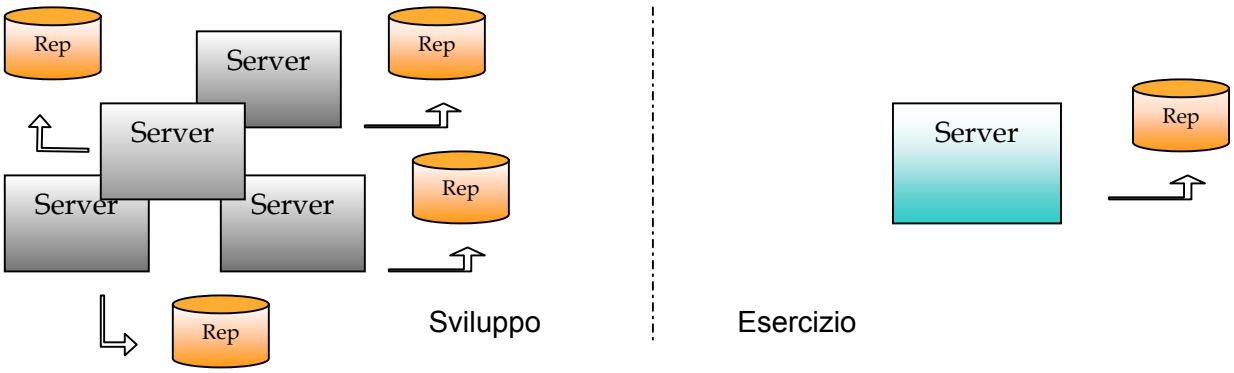
Il prodotto di ETL utilizzato da Consip per i progetti di Datawarehouse e Data Mart è Informatica Power Center.

Il prodotto permette di estrarre i dati da un sistema sorgente (che può essere un DBMS piuttosto che un file sequenziale), trasformarli secondo determinate regole (metadati) e caricarli nel sistema target.



Il server Power Mart verrà installato sui nodi Unix a La Rustica.

E' stato deciso di implementare tanti repository quanto saranno i progetti Unix di sviluppo. Ciascun gruppo avrà a disposizione un suo utente Unix e potrà lanciarsi il suo processo Power Center server che afferirà ad un repository distinto.
In esercizio il repository sarà centralizzato. Il singolo Server Power Center è comunque in grado di generare più processi figlio per far fronte a eventuali picchi di carico di lavoro.



Area Windows 2000

I server Win 2000 saranno completamente dedicati a Web Intelligence, il cui funzionamento è descritto nel paragrafo successivo. Non esistendo al momento alcuna particolare esigenza si è scelto di configurare i tre dischi di ciascun server in modalità RAID5 con una sola partizione da 36 GB.
I server sono attualmente installati con i seguenti indirizzi ip:

La Rustica:	10.38.5.13	10.38.1.102
	10.38.67.51	10.38.67.53
	10.38.67.52	10.38.1.104
	10.38.5.17	

Centro Comunicativo:	10.34.9.47
	10.34.9.48
	10.34.9.49
	10.34.9.51
	10.34.80.5

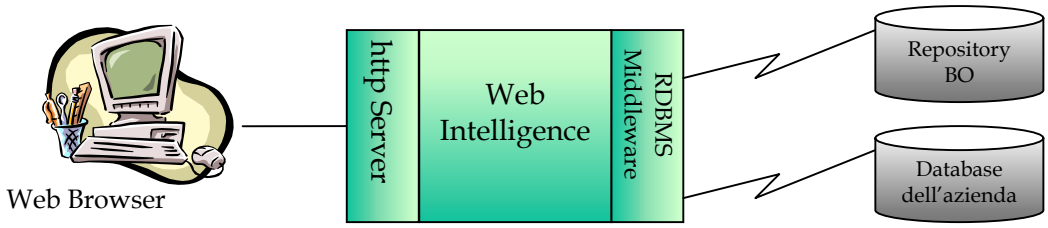
Area WebI

Il principale strumento di indagine OLAP che verrà impiegato per il sistema di Datawarehouse è Business Object, sia in versione Client/Server che Web Intelligence.

I realtà di Business Objects vengono sfruttate solamente alcune caratteristiche di disegno dei report che ancora non sono state integrate in Web Intelligence. La pubblicazione dei report risultanti verrà comunque affidata a WebI.

Web Intelligence è una soluzione di decision support system (DSS) a tre livelli espressamente studiata per effettuare analisi OLAP sul World Wide Web. L'applicazione permette ad utenti non tecnici di effettuare query ad hoc, rporting ed analisi di informazioni archiviate su data warehouse e data mart aziendali.

Lo schema di funzionamento di WebI è semplice:



Il browser http dell'utente effettua le interrogazioni al sistema tramite un server http che può essere sia IIS che Netscape Enterprise Server, il quale poi si farà carico di inoltrare la richiesta al motore Web Intelligence. Quest'ultimo esegue le transazioni che gli sono richieste (per esempio la definizione di un report, piuttosto che la sua esecuzione), eventualmente interagendo con il repository di BO o con il database aziendale, dopodichè restituisce il risultato in formato http.al web server, che a sua volta lo inoltra al thim client

Una delle caratteristiche più interessanti supportate da Web Intelligence è la possibilità di creare cluster per suddividere il carico di transito e ottenere funzioni di failover. I moduli WebI possono infatti essere su diverse macchine e il sistema riconosce automaticamente tutte le macchine della soluzione.

Una di queste macchine viene configurata per fare da gestore del cluster, mentre le restanti fungono da nodi.

Il cluster può essere implementato anche utilizzando macchine dotate di sistemi operativi eterogenei (quindi possono essere contemporaneamente utilizzati per lo stesso cluster sia server Unix che NT).

In considerazione dell'elevato numero di utenti che il sistema sarà chiamato a gestire, per il data warehouse del MTBPE si è scelto di implementare inizialmente un cluster di server NT. Grazie alle elevate caratteristiche di scalabilità di WebI sarà possibile far fronte ad eventuali semplicemente aggiungendo ulteriori nodi.