

CONTRATTO DI RICERCA ENEA-CASPUR – CRESCO SPI.2

**ANALISI E REALIZZAZIONE DI TOOL INNOVATIVI A SUPPORTO DELLE FUNZIONALITÀ GRID**



**OR1 – CREAZIONE DI UN DATABASE STORICO DEI METADATI DI AFS**

# SOMMARIO

---

|                                                            |    |
|------------------------------------------------------------|----|
| 1. AMACA – Generalità .....                                | 3  |
| 1.1     Introduzione .....                                 | 3  |
| 1.2     – Analisi dei requisiti.....                       | 3  |
| 1.3 – Design di AMACA .....                                | 4  |
| 1.3.1 – Design del modulo di indicizzazione.....           | 4  |
| 1.3.2 – Design della web application.....                  | 6  |
| 1.3.2.1 – Security .....                                   | 7  |
| 1.3.2.2 – Interfaccia funzionale .....                     | 7  |
| 1.4 – Stato di avanzamento attuale e sviluppi futuri ..... | 8  |
| 1.5 – Dettagli sul database storico .....                  | 8  |
| 1.5.1 – Tabella Servers.....                               | 9  |
| 1.5.3 – Tabella Parts .....                                | 11 |
| 1.5.4 – Tabella VolFS.....                                 | 12 |
| 1.5.5– Tabella VolDB.....                                  | 14 |
| 1.5.6 – Tabella VolDBrep .....                             | 15 |
| 1.5.7 – Tabella Alarms .....                               | 16 |
| 2 - Riferimenti.....                                       | 18 |
| 2.1 – Riferimenti Web.....                                 | 18 |

# 1. AMACA – Generalità

---

## 1.1 Introduzione

Una delle attività che attualmente vengono svolte nell'ambito del progetto CRESCO [CRSC] è relativa alla costruzione di uno strumento di monitoraggio e tuning del filesystem su cui si appoggia l'architettura di calcolo distribuito di ENEA (ENEAGRID) [EGRID].

La realizzazione concreta di una griglia computazionale richiede la definizione di un insieme di strumenti standard capaci di realizzare in modo uniforme l'accesso alle risorse informatiche disponibili, sia in termini di sistemi di calcolo che di archiviazione di dati.

I componenti principali che permettono ad ENEA GRID di fornire all'utente questo unico sistema virtuale sono essenzialmente due: un gestore delle risorse (LSF, Load Sharing Facility) e il file system distribuito AFS (Andrew File System) [AFS] attualmente disponibile in ambito OpenSource.

Il progetto **AMACA** nasce in risposta alla necessità di effettuare uno *screening* di AFS, tramite la costruzione di un database storico che permettesse di tracciare informazioni sullo stato globale dei nodi del filesystem (oltre ad eventuali stati anomali), altrimenti individuabili poco agevolmente.

La struttura del documento è organizzata come segue: Nel paragrafo 1.2 è riportata l'analisi del pacchetto software, seguita dal design e dalla descrizione dello stato di avanzamento (e relativi sviluppi futuri) nei paragrafi 1.3 e 1.4.

Nella sezione 1.5 è dettagliata la struttura delle tabelle che compongono la base dati.

Il documento si conclude, al capitolo 2, con l'elenco dei riferimenti indicati nel corso del report.

## 1.2 – Analisi dei requisiti

Amministrare il monitoraggio del filesystem, nei termini accennati al precedente paragrafo introduttivo, ha richiesto la progettazione di un tool software che sapesse tracciare primariamente i seguenti stati:

*Volume-aware*

- Volumi Offline,
- Volumi Locked,

- Volumi Unreleased,
- Volumi Busy,
- Volumi Unattacheable,
- Volumi utente non correlati ad un *backup volume* corrispondente,
- Volumi montati in posizione non standard,
- Volumi non montati su nessun mount point.
- Volumi registrati solo nel Volume Database (VLDB) o solo sui file servers.
- Volumi modificati il giorno precedente a quello di indicizzazione
- Volumi con maggior numero di accessi
- Volumi con ID irregolare (ovvero non osservanti lo schema per cui, detto *i* il volumeID della “copia” R/W, valgano anche R/O ID = *i+1* e BackupID = *i+2*).

#### *User-aware*

Utenti titolari di un volume montato con nomi non standard.

#### *Partition-aware*

- Partizioni aventi meno del 20% di spazio residuo
- Partizioni *overbooked* (ovvero la cui dimensione è inferiore alla quota totale assegnata all'insieme dei volumi su di essa residenti).

#### *Difference-aware*

Mappa delle differenze tra server, partizioni e volumi, calcolate relativamente a differenti snapshots.

#### *Network-aware*

Rapporti resends/data (tracciabili, altrimenti, solo per mezzo di richieste RX) relativi a tutti i server della cella.

Nella prossima sezione 1.3 verranno affrontati gli aspetti progettuali inerenti al design dell'applicazione.

## **1.3 – Design di AMACA**

L'intero progetto si divide in due porzioni principali: da un lato troviamo il modulo di indicizzazione, che esegue lo *screening* del filesystem memorizzando i dettagli dello stato nelle tabelle di un database di backend.

Dall'altro lato, una web-application costituisce il nodo di analisi con cui un amministratore di sistema può esplorare i risultati della fase di *indexing* tramite una comoda interfaccia.

### **1.3.1 – Design del modulo di indicizzazione**

L'interrogazione di AFS relativamente agli stati elencati nel precedente paragrafo 1.2 prevede l'interazione con 3 dei servizi principali del filesystem:

- **Basic Overseer (BOS) Server**
- **Volume Server**
- **Volume Database**

Tale interazione è possibile mediante le *command-line utilities* fornite da AFS stesso.

La necessità di dover realizzare uno storico delle informazioni reperite, ha tuttavia reso necessario l'utilizzo di strumenti esterni.

In particolare, il tool fa un uso estensivo delle AFS Perl APIs scritte da Norbert Gruener [NGPA] (che fungono da wrapper per i comandi di AFS), che sono state scelte sia per ragioni di completezza e robustezza, sia per la loro “popolarità” nell'ambito della comunità online.

La struttura di AMACA è composta di una serie di 4 moduli, cooperanti nel modo descritto dalla seguente figura 1:

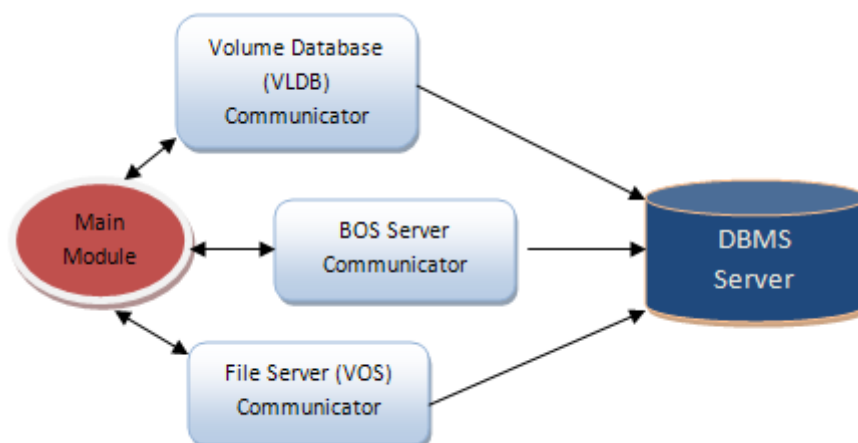


Figura 1 - Moduli di AMACA

Come si evince dalla precedente illustrazione, AMACA consta di:

Modulo principale (*Main Module*, che coordina lo startup dei moduli restanti e crea il canale di comunicazione con il Database server)

Modulo di comunicazione con il BOS Server (che recupera informazioni sullo stato dei servizi attivi su ogni nodo)

Modulo di comunicazione con il Volume Database (che interroga il VLDB sullo stato dei volumi in esso registrati, e utilizzato per reperire dati sui volumi in stato *locked*, sulle repliche e sulle release dei volumi)

Modulo di comunicazione con il Volume Server (Attivato solo nel caso in cui il nodo in analisi funga da fileserver, e invocato per ottenere i restanti rapporti).

Tutte le informazioni prodotte durante il processo di iterazione dei moduli (tranne quello principale) sui nodi AFS vengono trasferite nelle tabelle di un database di backend. La persistenza dei dati garantisce la possibilità di storicizzare gli eventi occorsi nell'intero filesystem, e permette -tramite elementari operazioni di *data mining*- di ottenere agevolmente informazioni (es: recupero del massimo *volume id*) che impiegherebbero uno sforzo più consistente, se ricercate con i tool “canonici”.

Nella sua versione finale, AMACA è stato ampliato per supportare anche

l'indicizzazione di celle esterne.

L'esecuzione concorrente e *non-invasive* di più utenti contemporanei

La differenziazione delle tabelle di indicizzazione *per-user*, in modo che i risultati dell'esecuzione di ogni istanza da parte di amministratori differenti coinvolgesse solamente delle strutture dati “personali”, e distinte in base alla *username*.

In particolare, sono stati creati due differenti database su cui AMACA è in grado di agire: il primo, di produzione, è acceduto soltanto da istanze comandate da un'apposita schedulazione cron. Il secondo, scratch, è aggiornabile invocando esplicitamente l'esecuzione dell'*indexer*, che, in base ai dati di login dell'utente, produrrà la creazione di una replica tabellare dedicata.

### 1.3.2 – Design della web application

Il modulo di analisi via web di AMACA è schematizzato secondo la seguente illustrazione 2:

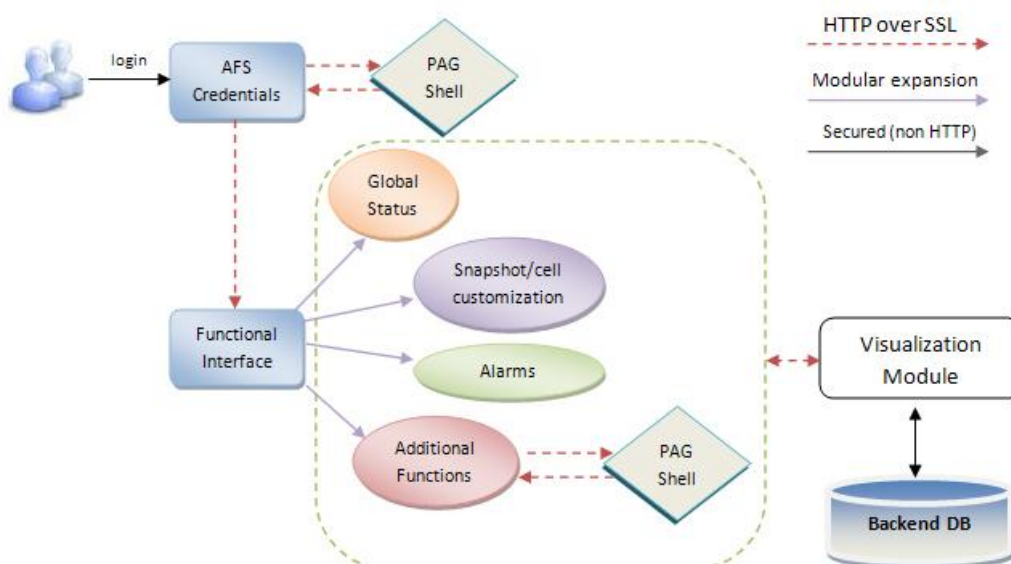


Figura 2 - Architettura della web application

la formalizzazione di progetto dei nodi componenti sarà chiarita nei successivi paragrafi 1.3.2.1 e 1.3.2.2.

### 1.3.2.1 – Security

L'accesso alle funzionalità dell'applicazione è subordinato all'inserimento, da parte dell'utente, delle corrette credenziali con cui egli è registrato su AFS.

Il processamento di tali credenziali d'accesso avviene confinando la transazione all'interno di una apposita **PAG shell**, che garantisce atomicità per la gestione concorrente di più *users* (tokens) che vogliano collegarsi al modulo web di amaca, e permette operazioni di logout regolari e pulite.

Alcune informazioni derivanti dal processo di autenticazione vengono memorizzate all'interno di un apposito cookie (che è utilizzato, come di regola, per mantenere lo stato tra le varie richieste HTTP che si susseguiranno durante l'utilizzo dell'applicazione).

Il progetto della security, a tal proposito, ha previsto l'esistenza di due layer che garantissero massima affidabilità nel controllo della protezione dei dati sensibili:

tunneling SSL, per incapsulare lo scambio dei dati tra client e server [OSSL]

crittografia dei dati *cookie-stored* mediante le funzionalità esposte dalla libreria *libmccrypt*[CRY].

### 1.3.2.2 – Interfaccia funzionale

L'interfaccia funzionale del modulo web di AMACA fornisce un quadro rapido ed esauritivo dello stato globale dell'intero filesystem.

In base a ciò che si evince dall'analisi dell'illustrazione 2, essa si scinde in 4 settori principali:

**Global Status** --> fornisce informazioni su *sync site*, stato di occupazione delle partizioni e dei volumi, Massimo *volumeID*)

**Customization** --> consente di restringere la visualizzazione dei risultati a snapshots e domini differenti da quelli di default

**Alarm Zone** --> visualizza situazioni di allarme che si stavano verificando all'atto della generazione dello snapshot selezionato. In particolare:

- Volumi Non attached
- Volumi rilasciati in modo non corretto
- Volumi Offline
- Volumi R/O mai rilasciati
- Volumi in {user.\*, p.\*} senza copia di backup
- Volumi non in {user.\*, p.\*} senza copia R/O
- Offline Servers
- IDs che identificano un volume R/W e che, invece, dovrebbero mappare una copia R/O o Backup di un altro volume R/W
- Server il cui indirizzo è non risolto da richieste DNS
- Volumi duplicati
- Volumi Busy
- Volumi R/O giacenti su partizioni/server differenti rispetto ai loro parent R/W.

**Additional functions** --> Permette lo svolgersi di operazioni di *online indexing* e di logout.

In particolare, AMACA permette ad un amministratore di sistema collegato tramite il suo modulo web di effettuare indicizzazioni online (ad esempio per verificare lo stato attuale di AFS dopo aver compiuto alcune operazioni amministrative). L'indicizzazione è avviabile a condizione che l'utente sia mappato in un'apposita membership (verificata per mezzo di una *call* al servizio **pts**). I risultati dell'indicizzazione sono memorizzati nella copia *scratch* (cfr. § 1.3.1) della base dati.

Il modulo di visualizzazione riceve le richieste dai singoli moduli componenti, le traduce in interrogazioni *pattern-based* verso le *DB-tables*, e invia i risultati al client web.

## 1.4 – Stato di avanzamento attuale e sviluppi futuri

Le informazioni descritte nel paragrafo precedente dettagliano con sufficiente accuratezza lo stato attuale di avanzamento di AMACA.

Al momento, le strategie di lavorazione risultano orientate sui seguenti fronti:

AMACA è attualmente in produzione sulla macchina 43p del C.R. ENEA di Frascati. Nel presente scenario, si svolgeranno:

- La prosecuzione della fase di testing del codice già prodotto, con strategie tese al miglioramento delle performances e all'ottimizzazione dei processi esistenti.
- L'individuazione di tecniche di progetto opportune per l'integrazione di funzionalità previste<sup>1</sup> ma non ancora implementate (per via delle onerose risorse di calcolo richieste, e dei particolari privilegi necessari alla loro esecuzione), e successiva realizzazione di un modulo aggiuntivo che le realizzi.

## 1.5 – Dettagli sul database storico

In questa sezione verrà esposta una descrizione dettagliata della base dati sottostante ad AMACA. Tutta la discussione è riferita alla sola porzione *production* del DB (costituendone le copie *scratch* delle pure repliche).

Sono presenti 7 tabelle, delle quali presenteremo, in appositi paragrafi, l'immagine della struttura e il *data dictionary*.

Esistono dei campi comuni, tra le tabelle, di cui forniamo subito un rendiconto:

---

1 Individuazione dei mount points di ciascun volume

**snap** -> è un progressivo che indica lo *snapshot* che si sta attualmente prendendo dello stato di AFS. E' incrementato ad ogni esecuzione dello script, e costituisce chiave primaria per tutte le tabelle.

**tstamp** -> E' un campo che rappresenta l'istante in cui la riga corrente viene scritta sulla tabella considerata.

**tday** -> Rappresenta il timestamp di esecuzione dello script.

Il tipo di dato è omogeneo su tutto il database, e verrà ogni volta riportato assieme alla struttura delle singole tabelle in analisi.

### 1.5.1 – Tabella Servers

E' la tabella elencante tutti i server di una cella (compresi quelli momentaneamente offline). Sono presenti due indici, creati sulle chiavi primarie **name** e **snap**.

#### DATA DICTIONARY

|        |                                   |
|--------|-----------------------------------|
| NAME   | Nome del server                   |
| STATUS | Stato del server (online/offline) |

I campi {**fserv**, **dbserv**, **buserv**} indicano se, sul server considerato, siano in esecuzione servizi di file serving, db serving o bakup serving.

I campi {**vlsync**, **ptsync**, **busync**, **kasync**}, servono a derivare il *sync site*.

Il campo **sendretrato** indica il rapporto resends/data per il server in analisi.

| Field     | Type          | Null | Key | Default | Extra |
|-----------|---------------|------|-----|---------|-------|
| name      | varchar(255)  |      | PRI |         |       |
| cell      | varchar(50)   |      | PRI |         |       |
| localcell | varchar(50)   |      |     |         |       |
| status    | varchar(255)  |      |     |         |       |
| fserv     | enum('y','n') | YES  |     | NULL    |       |
| dbserv    | enum('y','n') | YES  |     | NULL    |       |
| buserv    | enum('y','n') | YES  |     | NULL    |       |
| vlsync    | enum('y','n') | YES  |     | NULL    |       |
| ptsync    | enum('y','n') | YES  |     | NULL    |       |



## DATA DICTIONARY

|            |                                             |
|------------|---------------------------------------------|
| PROCNAME   | Nome del processo                           |
| SERVERNAME | Nome del server che ospita il processo      |
| STATUS     | Stato del processo                          |
| AUXSTATUS  | Stato ausiliario del processo               |
| COMMAND    | Comando sottomesso per eseguire il processo |
| LASTSTART  | Ultima data di esecuzione del processo      |
| LASTEXIT   | Ultima data di terminazione del processo    |

### 1.5.3 – Tabella Parts

Questa tabella elenca tutte le partizioni presenti sui vari fileserver.

Gli indici sono costruiti sulle chiavi **name** e **server**.

| Field     | Type                | Null | Key | Default             | Extra |
|-----------|---------------------|------|-----|---------------------|-------|
| name      | varchar(255)        |      | PRI |                     |       |
| cell      | varchar(50)         |      | PRI |                     |       |
| server    | varchar(255)        |      | PRI |                     |       |
| totVols   | bigint(20) unsigned | YES  |     | NULL                |       |
| busyVols  | bigint(20) unsigned | YES  |     | NULL                |       |
| notokVols | bigint(20) unsigned | YES  |     | NULL                |       |
| freeSpace | bigint(20) unsigned |      |     | 0                   |       |
| maxSpace  | bigint(20) unsigned |      |     | 0                   |       |
| tstamp    | timestamp           | YES  |     | CURRENT_TIMESTAMP   |       |
| snap      | bigint(20) unsigned |      | PRI | 0                   |       |
| tday      | datetime            |      |     | 0000-00-00 00:00:00 |       |

## DATA DICTIONARY

|           |                                          |
|-----------|------------------------------------------|
| NAME      | Nome della partizione                    |
| SERVER    | Nome del server che ospita la partizione |
| TOTVOLS   | Numero totale di volumi                  |
| BUSYVOLS  | Numero di volumi 'busy'                  |
| NOTOKVOLS | Numero di volumi corrotti                |
| FREESPACE | Spazio libero nella partizione           |
| MAXSPACE  | Massimo spazio disponibile               |

### 1.5.4 – Tabella VolFS

Questa tabella elenca tutti i volumi presenti sui vari fileserver.

Le chiavi primarie sono realizzate sui campi **valid**, **internalID**, **server** e **partition**

| Field      | Type                | Null | Key | Default | Extra |
|------------|---------------------|------|-----|---------|-------|
| validID    | bigint(20) unsigned |      | PRI | 0       |       |
| internalID | int(11)             |      | PRI | 0       |       |
| cloneID    | bigint(20) unsigned | YES  |     | NULL    |       |
| backupID   | bigint(20) unsigned | YES  |     | NULL    |       |
| parentID   | bigint(20) unsigned | YES  |     | NULL    |       |
| creatDate  | date                | YES  |     | NULL    |       |
| updateDate | date                | YES  |     | NULL    |       |
| name       | varchar(255)        |      |     |         |       |
| cell       | varchar(50)         |      | PRI |         |       |
| server     | varchar(255)        |      | PRI |         |       |
| partition  | varchar(255)        |      | PRI |         |       |
| inUse      | varchar(255)        | YES  |     | NULL    |       |
| type       | varchar(255)        | YES  |     | NULL    |       |
| status     | bigint(20) unsigned |      |     | 0       |       |

```

| size          | bigint(20) unsigned | YES | | 0 | | |
| fileCount     | bigint(20) unsigned | YES | | 0 | | |
| dayUse        | bigint(20) unsigned | YES | | 0 | | |
| maxQuota      | bigint(20) unsigned | YES | | 0 | | |
| attached      | char(1)              | YES | | n | | |
| busy          | char(1)              | YES | | n | | |
| tstamp        | timestamp             | YES | | CURRENT_TIMESTAMP | | |
| snap          | bigint(20) unsigned |     | PRI | 0 | | |
| tday          | datetime              |     | | 0000-00-00 00:00:00 | | |
+-----+-----+-----+-----+-----+-----+

```

## DATA DICTIONARY (VoIFS)

|            |                                                                            |
|------------|----------------------------------------------------------------------------|
| VOLID      | ID del volume                                                              |
| INTERNALID | Contatore incrementale utilizzato per prevenire situazioni di duplicazione |
| CLONEID    | ID del clone                                                               |
| BACKUPID   | ID del volume di backup                                                    |
| PARENTID   | ID del volume R/W nel caso questo sia una replica                          |
| CREATDATE  | data di creazione del volume                                               |
| UPDATEDATE | ultima data di aggiornamento                                               |
| NAME       | nome del volume                                                            |
| SERVER     | server che ospita il volume                                                |
| PARTITION  | partizione che contiene il volume                                          |
| INUSE      | volume online/offline                                                      |
| TYPE       | tipo di volume                                                             |
| STATUS     | stato del volume                                                           |
| SIZE       | dimensione (in Kbytes) del volume                                          |

## DATA DICTIONARY (VoIFS)

|           |                                     |
|-----------|-------------------------------------|
| FILECOUNT | numero di files ospitati nel volume |
| DAYUSE    | utilizzo giornaliero                |
| MAXQUOTA  | quota massima                       |

### 1.5.5- Tabella VolDB

Questa tabella elenca tutti i volumi registrati sul Volume Database che risultino in modalità Read-Write.

Le chiavi primarie sono realizzate sui campi **valid**, **internalID**, **server**, **partition** e **volstatus**.

| Field       | Type                | Null | Key | Default             | Extra |
|-------------|---------------------|------|-----|---------------------|-------|
| volname     | varchar(255)        | YES  |     | NULL                |       |
| internalID  | int(11)             |      | PRI | 0                   |       |
| volID       | bigint(20) unsigned |      | PRI | 0                   |       |
| backupID    | bigint(20) unsigned | YES  |     | NULL                |       |
| cell        | varchar(50)         |      | PRI |                     |       |
| server      | varchar(255)        |      | PRI |                     |       |
| partition   | varchar(255)        |      |     |                     |       |
| volstatus   | varchar(255)        |      | PRI |                     |       |
| vRelease    | varchar(255)        | YES  |     | NULL                |       |
| mpoint      | varchar(255)        | YES  |     | NULL                |       |
| nReplica    | bigint(20) unsigned | YES  |     | NULL                |       |
| flags       | bigint(20)          | YES  |     | NULL                |       |
| serverFlags | bigint(20)          | YES  |     | NULL                |       |
| type        | enum('RW','BK')     | YES  |     | NULL                |       |
| tstamp      | timestamp           | YES  |     | CURRENT_TIMESTAMP   |       |
| snap        | bigint(20) unsigned |      | PRI | 0                   |       |
| tday        | datetime            | YES  |     | 0000-00-00 00:00:00 |       |

+-----+-----+-----+-----+-----+

## DATA DICTIONARY (VolDB)

|             |                                                                            |
|-------------|----------------------------------------------------------------------------|
| VOLNAME     | Nome del volume                                                            |
| INTERNALID  | Contatore incrementale utilizzato per prevenire situazioni di duplicazione |
| VOLID       | ID del volume                                                              |
| BACKUPID    | ID del volume di backup                                                    |
| SERVER      | server che ospita il volume                                                |
| PARTITION   | partizione che contiene il volume                                          |
| VOLSTATUS   | stato del volume (locked/unlocked)                                         |
| RELEASE     | stato release del volume                                                   |
| MPOINT      | mount point in AFS                                                         |
| NREPLICA    | numero di volumi replica                                                   |
| FLAGS       | flags                                                                      |
| SERVERFLAGS | flags del server a cui appartiene il volume                                |

## 1.5.6 – Tabella VolDBrep

Questa tabella elenca tutti i volumi replica registrati nel VLDB.

le chiavi sono sui campi **Volid**, **internalID**, **server** e **partition**.

+-----+-----+-----+-----+-----+

| Field                           | Type                | Null | Key | Default | Extra |
|---------------------------------|---------------------|------|-----|---------|-------|
| +-----+-----+-----+-----+-----+ |                     |      |     |         |       |
| volid                           | bigint(20) unsigned |      | PRI | 0       |       |
| internalID                      | int(11)             |      | PRI | 0       |       |
| parentID                        | bigint(20) unsigned | YES  | MUL | NULL    |       |



|                                       |                      |     |     |      |  |  |
|---------------------------------------|----------------------|-----|-----|------|--|--|
| valid                                 | int(11)              |     | PRI | 0    |  |  |
| volname                               | varchar(255)         | YES |     | NULL |  |  |
| voltype                               | enum('RW','RO','BK') | YES |     | NULL |  |  |
| cell                                  | varchar(50)          |     | PRI |      |  |  |
| server                                | varchar(255)         |     | PRI |      |  |  |
| partition                             | varchar(10)          |     | PRI |      |  |  |
| rovolname                             | varchar(255)         |     | PRI |      |  |  |
| bkvolname                             | varchar(255)         |     | PRI |      |  |  |
| vRelease                              | varchar(15)          | YES |     | NULL |  |  |
| alarmkind                             | int(11)              |     | PRI | 0    |  |  |
| snap                                  | int(11)              |     | PRI | 0    |  |  |
| +-----+-----+-----+-----+-----+-----+ |                      |     |     |      |  |  |

#### DATA DICTIONARY (Alarms)

|           |                                   |
|-----------|-----------------------------------|
| VOLID     | ID del volume                     |
| VOLNAME   | Nome del volume                   |
| VOLTYPE   | Tipo di volume                    |
| SERVER    | server name                       |
| PARTITION | partizione che contiene il volume |
| ROVOLNAME | Nome del volume R/O               |
| BKVOLNAME | Nome del volume di backup         |
| VRELEASE  | Stato di rilascio del volume      |
| ALARMKIND | Codice allarme                    |

## 2 - Riferimenti

---

### 2.1 – Riferimenti Web

[CRSC] – <http://www.afs.enea.it/project/cresco/>

[EGRID] - <http://www.afs.enea.it/project/eneagrid/>

[AFS] – <http://www.openafs.org>

[NGPA] – <http://search.cpan.org/~nog/>

[OSSL] – <http://www.openssl.org>

[CRY] – <http://mccrypt.hellug.hr/lib/>