SRC Computers d.o.o., Tržaška 116, 1001 Ljubljana, Slovenija
tel.: 061/123 32 32
fax: 061/123 41 73
e-mail: srccomputers@src-comp.si
http:/www.src-comp.si

ARHIVIRANJE (SKENIRANIH) DOKUMENTOV

Bine Žerko, SRC Computers d.o.o.
(Email: bzerkoŽsrc.si ali bine.zerkoŽsrc-comp.si)

Povzetek:

Ob različnih priložnostih vedno znova ugotavljam, da se večina sogovornikov nekako ne zaveda vloge in pomena, ki ga ima hierarhično arhiviranje v sistemu računalniškega upodabljanja (in arhiviranja) dokumentov. V prispevku si bomo ogledali nekaj najpomembnejših lastnosti, ki jih mora zagotavljati učinkovit sistem za arhiviranje (skeniranih) dokumentov. Na kratko pa smo opisali tudi značilnosti programskih produktov Lotus Notes:Document Imaging.

HIERARHIČNO ARHIVIRANJE DOKUMENTOV

oziroma HSM (Hierarchical Storage Management) je obvezni podsistem vsakega sistema za računalniško upodabljanje in arhiviranje dokumentov (angl. imaging & archiving), zato takoj poudarimo, da tako, kot:

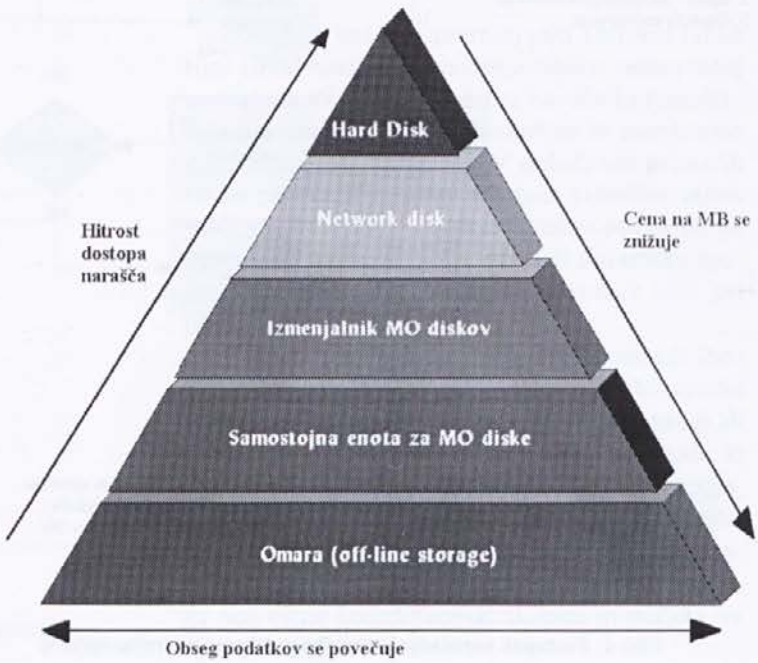
- ni sistema za upravljanje z dokumenti brez imaging-a, tudi
- ni imaging-a brez sistema HSM.

Za implementacijo HSM na nivoju datotečnih sistemov, lahko izbiramo med večjim številom različnih produktov (obstaja jih več kot 30), vendar je pri tem treba natančno poznati predvsem namen, oziroma funkcionalnost posameznih sistemov, da bi se lahko optimalno odločili. Pri tem moramo prvenstveno upoštevati način delovanja aplikacij in pa infrastrukturo (strojna oprema, operacijski sistem(i)), ki jo uporabljamo.

HSM v bistvu simulira klasične metode arhiviranja, pri katerih imajo uporabniki na svojih mizah najbolj aktualne dokumente, medtem, ko se starejši nahajajo v omarah, ali v kleteh. Razlika med ročnim in avtomatičnim sistemom je v tem, da uporabniku ni treba skrbeti tako za iskanje in dostavo dokumentov (iz arhiva), kot tudi ne za pospravljanje (nazaj v arhiv). Na ta način preprečimo sicer pogosto izgubljanje dokumentov in njihovo nepotrebno fotokopiranje za primer, ko isti dokument hkrati potrebuje več uporabnikov. Z uporabo različnih medijev (dražjih in cenejših pomnilniških enot) tako zgradimo neke vrste (elektronski) arhiv, v katerem dokumenti "potujejo" iz medija na medij - odvisno od pogostosti uporabe.

Za ponazoritev sistema HSM si predstavljajmo piramido (glej sliko), ki predstavlja različne medije za arhiviranje. Piramida ponazarja posamezne nivoje arhiviranja, in sicer:

- trdi disk v strežniku HSM (ena ali več particij)
- trdi disk v datotečnem strežniku (eden ali več volumnov)
- izmenjalnik (angl. jukebox) magnetno-optičnih (MO) diskov (število volumnov je odvisno od števila diskov)
- samostojna, zunanja enota za magnetno-optične diske
- tračna enota in navadna, oziroma ognjevarna omara (off-line arhiv), kamor shranjujemo magnetno-optične diske.



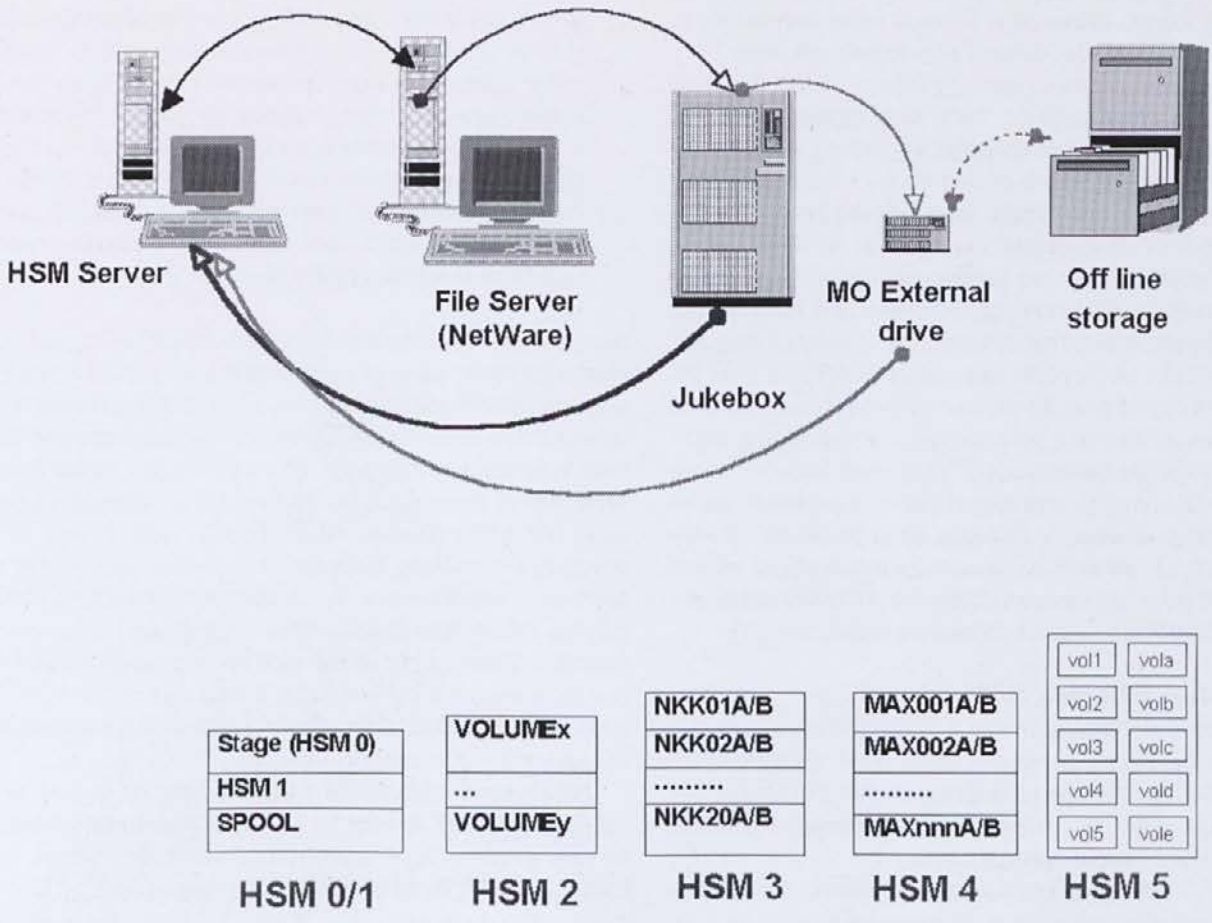
Vrh piramide predstavlja (lokalni) disk strežnika HSM, na katerem se nahajajo najbolj aktualni podatki; tj. najnovejši in pa tudi tisti, ki jih uporabniki najpogosteje uporabljajo. Zaradi narave delovanja HSM so odzivni časi najkrajši. V tem delu piramide je tudi najmanj prostora (velikost diska); gre pa za (naj)dražji pomnilni medij.

Za naslednji nivo (lahko) uporabimo enega ali več diskov (oz. volumnov) v datotečnem, oziroma mrežnem strežniku in odzivni časi pri dostopanju do teh podatkov so le malenkost slabši, ker mora HSM "poiskati" ustrezno datoteko in jo potem posredovati uporabniku.

Tretji nivo v našem primeru predstavlja izmenjalnik magnetno-optičnih diskov. Dostop do podatkov je še vedno avtomatičen, tj. brez posegov operaterja, vendar se dostopni časi lahko že opazno podaljšajo; še posebej v primeru, ko več uporabnikov zahteva datoteke, ki se nahajajo na različnih diskih.

Za četrti nivo smo uporabili samostojno enoto za magnetno-optične diske, s pomočjo katere postopoma nastaja "off-line" arhiv. Dostop do teh podatkov že ni več avtomatičen, ker mora operater ročno zamenjevati posamezne diske, ki se nahajajo na petem nivoju, oziroma v omari. Kot medij arhiviranja na petem nivoju bi sicer lahko uporabili tudi tračne enote z ustreznim izmenjevalnikom (v tem primeru predstavlja omara šesti nivo), vendar po mojem to skoraj ni smiselno, saj z magnetno optičnimi diski dosežemo bistveno večjo fleksibilnost kot s trakovi.

Seveda pa lahko sistem HSM oblikujemo na različne načine in s poljubnim številom nivojev - odvisno predvsem od tega, za kolikšen obseg podatkov moramo zagotoviti minimalne odzivne čase in dostopnost brez operaterjevih posegov (zamenjava magnetno-optičnih diskov). Pri konfiguriranju sistema HSM si pomagamo z analizo



Slika: **Prikaz petnivojskega sistema HSM (HSM15.BMP)**

trenutnega obsega in napovedjo predvidenega števila podatkov in izračunamo vsaj približne potrebne kapacitete pomnilnih medijev na posameznem nivoju.

Na naslednji sliki je okvirno prikazan postopek delovanja sistema HSM, oziroma proces migracije datotek med različnimi nivoji HSM.

Migracija

je proces "selitve" posameznih dokumentov po hierarhični lestvici (iz medija na medij) navzdol ali navzgor na osnovi predhodno definiranih pravil, oziroma postopkov. Na sliki je prikazan večnivojski HSM, ki smo ga obravnavali pri razlagi piramide.

Vsak objekt, ki je shranjen v HSM, je vezan na profil, s katerim je določen način, kako se bo objekt selil po posameznih nivojih, oziroma po medijih, ki sestavljajo sistem za hierarhično arhiviranje. Posamezne skupine objektov so lahko vezane na skupen profil, kar pomeni, da predstavljajo v procesu migracij množico objektov, za katero veljajo ista, oziroma skupna pravila. Vsak profil ima lahko svoj nabor medijev (diskov), lahko pa si jih med seboj tudi delijo, tj. več profilov uporablja (deloma ali v celoti) iste diske.

V procesu migracij ima posebno vlogo delovno področje, ki predstavlja t.i. nulti nivo sistema HSM. Na tem področju (disku) se nahajajo vsi najbolj aktualni objekti, to so tisti, ki so bili poskenirani v obdobju po zadnji migraciji. Na tem nivoju se prav tako (začasno) nahajajo tudi vsi tisti objekti, ki so sicer že shranjeni na nižjih nivojih HSM in do katerih so uporabniki kasneje ponovno dostopali. S tem je zagotovljen zelo hiter dostop do najbolj aktualnih dokumentov, saj "povratna migracija" ne poteka po nivojih (glej smer puščic na sliki 2), temveč se objekt prepiše na omenjeni nulti nivo HSM. Ko se ta nivo postopoma spet zapolni, se (načeloma avtomatično) izvede postopek čiščenja, med katerim se vsi novonastali objekti preselijo na prvi nivo, ostali pa se enostavno pobrišejo. V primeru, da je prišlo do kakšnih sprememb na objektih, ki so sicer že shranjeni na nižjem nivoju, se tako spremenjeni objekti obravnavajo kot novonastali (in brišejo se stari objekti na nižjih nivojih).

Kategorije sistemov HSM

Posameznih nivojev arhiviranja (medijev) pa ne smemo enačiti, oziroma poistovetiti z nivoji HSM, v katere lahko "popredalčkamo" tovrstno programsko opremo. HSM razvrščamo glede na lastnosti in značilnosti programske opreme v pet skupin, oziroma nivojev:

1. Nivo 1 omogoča enostavno, avtomatično migracijo podatkov (datotek) in transparenten dostop do njih. Datoteke migrirajo na naslednji nivo, vendar uporabniki "ne vidijo" nobenih sprememb. Datoteka se navidezno še vedno nahaja npr na disku G:, čeprav je fizično že na naslednjem nivoju. Ko uporabnik "odpre" datoteko, posredovanje opravi HSM.
2. Nivo 2 podpira vsaj dva ali tri nivoje (piramida) in migracija poteka na osnovi predhodno definiranih postopkov in pravil; odvisno od (pre)zasedenosti medija na posameznih nivojih.
3. Nivo 3 podpira tri ali štiri nivoje (piramida); kriterije migracij lahko dinamično spreminjamo in prilagajamo, oziroma dodajamo medije na posameznih nivojih. Sistemi HSM nivoja 3 že podpirajo optične in tračne enote.
4. Nivo 4 omogoča definiranje pravil in postopkov za migracijo na osnovi klasifikacije (pomembnosti) datotek. Podprti so praktično vsi nivoji arhiviranja (piramida) in uporabljamo lahko vse enote v celotnem omrežju.
5. Nivo 5 predstavlja najvišjo stopnjo sistemov HSM in se od predhodnega nivoja razlikuje po tem, da ne gre več za t.i. datotečni, temveč za objektni sistem. In takšni sistemi so vsaj po mojem mnenju tudi najbolj primerni za arhiviranje skeniranih dokumentov. Vzrok najdemo v objektnem načinu obravnavanja skeniranih dokumentov (torej ne gre zgolj za datotečni sistem), ki niso sestavni del baze podatkov, temveč se v bazo shranjujejo samo referenčni podatki o tem, kje se posamezen objekt (trenutno) nahaja. Objekt sam pa pomeni samostojno entiteto v procesu migracij; seveda pod kontrolo HSM-a. Sodoben sistem HSM omogoča tudi replikacijo podatkov med različnim strežniki HSM v okolju WAN.

Izbrati ustrezno programsko opremo za HSM ni najbolj preprosta stvar, saj smo v večini primerov vezani na podsistem, ki je vkomponiran v celoten sistem za računalniško upodabljanje in arhiviranje dokumentov. Po mojem mnenju moramo pri izbiri celotnega sistema posvetiti največ pozornosti ravno tovrstni programski opremi in natančno preučiti vse možnosti, ki jih (n)ima. Pri odločanju o nakupu sistema za celovito upravljanje z dokumenti moramo imeti že na samem začetku pred očmi celotno sliko o tem, kako bo stvar izgledala po nekaj letih uporabe. V bistvu gre za povsem konceptualna vprašanja o tem, kje in na kakšen način bomo zajemali in hranili (skenirane in ostale) dokumente, kakšen bo format zapisa, oziroma način zgoščevanja, ipd.

Neprimerno izbran HSM v začetku sicer ne bo povzročal nobenih težav, vendar bo kasnejše reševanje problemov prav gotovo zelo draga pustolovščina; še posebej, če boste ugotovili, da bo potrebno programsko opremo (skoraj) v celoti zamenjati. Res je tudi to, da je programska oprema vedno cenejša, vendar k ceni licenc prištejte še stroške z izobraževanjem in uvajanjem uporabnikov in seveda s konverzijo podatkov iz starega v nov sistem.

Še par besed o strojni opremi. Strežnik HSM naj bo resnično strežnik in ne navadna delovna postaja, z nekoliko večjimi diski in večjim pomnilnikom. Zelo priporočljivo je, da za upravljanje HSM dejansko uporabimo namenski računalnik, s stabilnim in zanesljivim operacijskim sistemom (avtor je zagovornik OS/2). Prav tako poskrbimo za ustrezno varovanje samega strežnika (sistem za neprekinjeno napajanje) in podatkov (backup).

O izmenjalnikih magnetno-optičnih diskov

Uporaba izmenjalnikov za magnetno-optične diske je praktično nujna v vseh sistemih, kjer imajo opravka z več tisoč dokumenti, ker lahko le tako zagotovimo dovolj uporaben arhiv. Ponovno pa naj poudarim, da moramo vsaj približno izračunati pričakovan obseg podatkov, da lahko pripravimo ustrezne kapacitete tudi na predhodnih nivojih HSM.

Vzemimo za primer, da imamo letno opraviti s 100.000 dokumenti, oziroma stranmi (A4 format). V najboljšem primeru (skeniramo pri 200 dpi, zgoščevanje CCITT/ITU G4) to predstavlja približno 3 GB (30 KB/stran). Če želimo zagotoviti popolnoma transparenten dostop do vseh teh podatkov tudi po petih letih, potrebujemo izmenjalnik s kapaciteto 15 GB. V primeru, da imamo na prvem in drugem nivoju HSM na voljo vsaj 3 GB, to pomeni, da smo zagotovili minimalne dostopne čase za 100.000 najbolj aktualnih dokumentov, ki se bodo vedno nahajali na prvem in drugem nivoju. Hkrati pa to pomeni, da lahko z nakupom izmenjalnika počakamo vsaj leto dni; cene bodo padle, kapacitete in ostale lastnosti pa se bodo ponovno izboljšale, saj dobimo danes izmenjalnik kapacitete 100 GB praktično za enak denar, kot smo odšteli v začetku 90-ih za 10 GB.

Kaj se torej skriva v izmenjalnikih magnetno-optičnih diskov?

V navzven dokaj dolgočasnem ohišju, se nahajajo naslednji elementi izmenjalnika:

- odlagališče, shramba za magnetno-optične diske
- najmanj ena diskovna enota
- mehanski in elektronski del za pomik robotske roke in
- matična plošča za nadzor delovanja izmenjalnika.

Največji vpliv na odzivne čase pri posredovanju želenih podatkov uporabnikom imajo poleg hitrosti robotske roke predvsem lastnosti magnetno-diskovne enote in diskov. Pri tem mislim na dostopne čase in hitrost prenosa podatkov na posameznih diskih. Robotske roke v sodobnih izmenjalnikih lahko zamenjajo magnetno-optični disk v nekaj sekundah, k temu pa moramo prišteti še čas, ki ga potrebuje enota za pozicioniranje bralno-pisalne glave. Trenutno znaša rekord 0,8 sekunde (iz odlagališča v enoto ali obratno), kar je verjetno že skoraj teoretični minimum; realno pa to pomeni, da boste do želenega podatka v povprečju prišli v približno dveh sekundah.

V letu 1996 je postala standardna (ISO) kapaciteta 5,25 palčnih magnetno-optičnih diskov 2,6 GB in kapacitete se bodo po napovedih podvojile že do konca tega leta. Dostopni časi (average access time) znašajo od 19-35 ms (odvisno od proizvajalca in modela), večina magnetno-optičnih enot pa se lahko pohvali s hitrostjo prenosa 5 MB/sec; čeprav je dejanska hitrost nižja (posebej pri zapisovanju podatkov na disk). Pri izbiri opreme pa ne pozabimo na pomen standardov (tudi) na tem področju, saj je osnovna zahteva za vsako (novo) magnetno-optično enoto ali magnetno-optični disk predvsem kompatibilnost z ostalimi proizvajalci in seveda združljivost s predhodnimi modeli.

LOTUS NOTES:DOCUMENT IMAGING

LN:DI Professional (Lindy Pro)
je osnovni programski modul, oziroma produkt, ki omogoča implementacijo računalniškega upodabljanja in arhiviranja dokumentov s pomočjo aplikacij Lotus Notes. V bistvu gre za povsem samostojen program, ki ga uporabljamo za nadzor delovanja skenerja in s pomočjo OLE (Object Linking and Embedding) posredujemo referenčne podatke o skeniranih dokumentih v datoteko Notes, oziroma v sistem HSM.

Poglavitna prednost, ki jo ima Lindy pred konkurenco, se izraža v izredni arhitekturi celotnega sistema Notes, v kateri predstavlja zajem (skeniranje) dokumentov samo enega od procesov, oziroma funkcij v sistemu za upravljanje z dokumenti.

LN:DI Mass Storage System (Lindy MSS)
je najpomembnejši programski modul iz družine Lindy, saj je namenjen vzpostavitvi sistema za hierarhično arhiviranje skeniranih dokumentov. Njegova poglavitna lastnost je, da obravnava računalniško upodobljene dokumente kot objekte in ne kot (samostojne) datoteke, zato MSS po specifikaciji HSM sodi v kategorijo petega nivoja. Ponovimo, da imamo pri tem opravka z objekti, ki so "povezani" s posameznimi (Notes) dokumenti kot njegov sestavni del, vendar se fizično nahajajo izven (Notes) baze podatkov; v bazi sami so shranjeni samo referenčni podatki o posameznem (OLE) objektu.

Osnovni namen tega produkta je torej zagotovitev in upravljanje dovolj fleksibilnega sistema za arhiviranje dokumentov; tj. takšnega sistema, ki lahko dejansko obravnava velike količine (skeniranih) dokumentov.

Drugi produkti iz družine Lindy
V družino izdelkov Lindy spadajo poleg omenjenih dveh programov še:

- **Lindy Image Viewer (LIV)**, ki omogoča pregledovanje skeniranih dokumentov.
- **Image Processing Server (IPS)** je namenjen nadzoru, administraciji in povezovanju okolja Lindy z aplikacijami ostalih proizvajalcev.
- **Lotus Fax Server (LFS)** je namenjen prejemanju in pošiljanju faxov.
- **Print-to-Fax Driver** omogoča, da lahko iz katerekoli aplikacije Windows pošiljamo datoteke na LFS.
- **Workgroup OCR Option 2.6** omogoča optično razpoznavanje znakov skeniranih dokumentov. Podobno kot IPS in LFS, je tudi to samostojni (aplikacijski) strežnik v okolju Windows in ima status odjemalca Notes.
- **Lotus Notes to IBM ImagePlus Connection 1.0.** ImagePlus je IBM-ov produkt, oziroma družina produktov v katero sodi tudi VisualInfo. Bistveno (in zanimivo) pa je, da se ohranja arhitektura družine Lindy z MSS na čelu, kar samo potrjuje nivo kakovosti in uporabnosti okolja Lotus Notes. Sestavni del tega produkta je tudi Lindy IPS, ki dejansko opravlja povezovalno funkcijo med obema platformama.

Zaključimo

z vabilom, da si lahko celovit sistem za računalniško upodabljanje in arhiviranje dokumentov, delujoč v Lotus Notes okolju, ogledate v skupini podjetij SRC. Ob tem pa povejmo še to, da smo za dosežke na imaging področju na letošnji konferenci Lotusphere (Nica, Francija), prejeli tudi nagrado Lotus European Beacon Award 1997. Priznanje ima še večjo težo ob upoštevanju dejstva, da na ameriškem delu prireditve za to področje ni bilo podeljeno priznanje. Podjetje IBM/Lotus vsako leto podeli nagrade podjetjem, ki s svojim delom najvidneje uveljavljajo in razširjajo uporabnost programskega izdelka Lotus Notes. V osemnajstih kategorijah je bilo več kot 60 finalistov.

Podjetje SRC INFO d.o.o. je dobilo nagrado za implementacijo projekta Lotus Notes Document Imaging v državni upravi Republike Slovenije.