

# **MSARCH - system archiwizacji przesyłek E-mail**

**Grzegorz Blinowski**  
**G.Blinowski@cc.com.pl**

**CC Otwarte Systemy Komputerowe Sp. z o.o.**

- Po co archiwizować e-mail?
- Problemy przy budowie archiwum e-mail
- Funkcje systemu MSARCH:
  - wyszukiwanie
  - dostęp do wiadomości
- Architektura MSARCH
  - przepływ danych
  - indeksy
- Wydajność

- Zastosowanie typowo backupowe - możliwość odzyskania przesyłek
- Przeznaczenie "dowodowe" - możliwość jednoznacznego stwierdzenia czy dana przesyłka została wysłana lub dostarczona:
  - przez/do kogo
  - kiedy
  - problemy: naruszenie tajemnicy przedsiębiorstwa, materiały niedozwolone, wirusy, ...
- Użytkownicy mogą przejrzeć archiwum własnych e-maili

- Kontrola poczty *wchodzącej* to obecnie standard
- W ciągu najbliższego roku ważniejsza stanie się kontrola poczty *wychodzącej*
  - e-mail stał się głównym medium wymiany informacji pomiędzy pracownikami wewnątrz firmy oraz w relacjach firma-firma
  - nietrudno o pomyłki
  - zdarza się też działanie w złej wierze
  - koszty odpowiedzialności prawnej są olbrzymie!

- Pomyłki / incydenty
  - Prezes Boeing-a usunięty z powodu molestowania przez e-mail
  - Microsoft: "knife the baby" (w odniesieniu do QuickTime)
- Badania wykazały, że wśród kadry zarządzającej średniego i wyższego szczebla:
  - 70% otrzymywał e-mail o treści pornograficznej
  - 24% otrzymywało poprzez e-mail informacje poufne lub tajne
  - 64% otrzymywało wiadomości o treści "niepoprawnej politycznie"
- W grupie "Fortune 500":
  - w 27% firm miały miejsce pozwy o molestowanie seksualne
  - w każdym z przypadków **e-mail** był jednym z dowodów
  - odszkodowania rzędu dziesiątek tysięcy USD
  - w przypadku naruszenia praw autorskich koszty dochodzą do mln USD

- Format wiadomości
  - zgodność z RFC 2822, itp. nie wystarcza
  - pewien odsetek przesyłek nie jest zgodny z RFC (zwłaszcza w zakresie kodowania załączników)
- Autoryzacja
  - nadawca / odbiorca ma prawo wglądu do wiadomości, ale dodatkowo:
    - pole "Bcc" nie jest przekazywane z definicji w nagłówku - odbiorca nie zawsze jest znany
    - należy obsługiwać aliasy grupowe
- Wydajność
  - rozmiar archiwum dla kilku tysięcy użytkowników w przeciągu kilku miesięcy wyniesie dziesiątki GB oraz setki tysięcy plików
  - efektywne indeksowanie archiwum tych rozmiarów jest trudne

- E-mail bez nagłówka
- CR lub LF jako koniec linii w nagłówku (powinno być CR LF)
- niepoprawne nagłówki, np.:
  - `Content-type :text/html`
  - `Content-type: text/html`
  - brak średników w nagłówku:  
`Content-type: text/html charset=iso8859-2` zamiast:  
`Content-type: text/html; charset=iso8859-2`
- Źle podane kodowanie:
  - iso-8859-2 wyrażane jako iso88592 lub iso8859-2
- Niedozwolone znaki przy kodowaniu quoted-printable i Base64
- Pusty charset:
  - `Content-type: text/plain; charset=""`
- Dwa różne Content-type w jednej wiadomości
- Spacja w ciągu boundary

# MSARCH v 2.2

- Scentralizowana archiwizacja przesyłek otrzymywanych i wysyłanych
- Indeksowane archiwum wiadomości i załączników
- Wyszukiwanie wg. kilkudziesięciu kryteriów:
  - w polach nagłówka wiadomości (nadawca, odbiorca, temat, itp.),
  - pełnotekstowe w zawartości przesyłki i załącznikach (txt, html, pdf)
- "User Authority" - użytkownicy mają dostęp do wiadomości wysłanych do/przez nich:
  - obsługa autoryzacji LDAP dla Active Directory, Sun Directory Server i innych systemów autoryzacji
  - "Send back" - użytkownik może ponownie otrzymać zamówiony zarchiwizowany mail do swojej skrzynki pocztowej
- Automatyczne usuwanie starych wiadomości z archiwum
- Interfejs WWW

- Interfejs:
  - użytkownika - WWW
  - wersja polska i angielska
- Administratora (konfiguracja) - GUI
- Dostęp do przesyłek:
  - nagłówek sformatowany
  - sformatowane ciało przesyłki
  - nagłówek w postaci źródłowej
  - załączniki jako lista

- Nadawca: from, sender
- Odbiorca: to, cc, SMTP "rcpt to" (bcc)
- Zakres czasowy od-do
- Temat i zawartość przesyłki:
  - fraza\*, fraza1 ~frazo2, fraza1 & fraza2, fraza1 | fraza2, fraza1 &! fraza2
  - podobne, zawiera (like, contains)
  - tak samo dla załączników
- Inne pola nagłówka
  - Message-ID, User-Agent, ...
- Załączniki (prócz treści):
  - nazwa: Content-Type, name; Content-Disposition filename
  - rozmiar: od - do

Msarch 2.2 - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Home Search Favorites Media

Address http://localhost:8800/archgui/query.aspx

### Pola podstawowe

Nadawca: @cc.com.pl

Odbiorca:

Temat:

Włączając pola:  from  sender

Włączając pola:  to  cc  SMTP "rcpt to" (bcc)

Metoda wyszukiwania CONTAINS:

Przykłady:

- email**  
pole zawiera 'email'
- market\***  
pole zawiera słowo zaczynające się od 'market'
- email ~ market**  
pole zawiera 'email' i 'market' na tej samej stronie
- email & market**  
pole zawiera 'email' i 'market'
- email | market**  
pole zawiera 'email' lub 'market'
- email &! market**  
pole zawiera 'email' ale nie zawiera 'market'

### Zakres czasowy

nie starsze niż: 2005 - 08 - 17 19 : 26 Wstaw Wyczyść

nie nowsze niż: 2005 - 08 - 24 19 : 26 Wstaw Wyczyść

### Przeszukiwanie treści

Treść (używaj składni CONTAINS)

### Pola dodatkowe

Message-ID:

Nazwa załącznika:

Rozmiar załącznika (w bajtach): - bajtów

Strefa:  wiadomości  załączniki

Klient poczty nadawcy:

Włączając pola:  name  filename

Odpowiedź do:

Składnia zapytania	Ile dokumentów	Ilość na stronie
<input type="radio"/> LIKE	<input checked="" type="radio"/> 1 000	<input checked="" type="radio"/> 10
<input checked="" type="radio"/> CONTAINS	<input type="radio"/> 10 000	<input type="radio"/> max

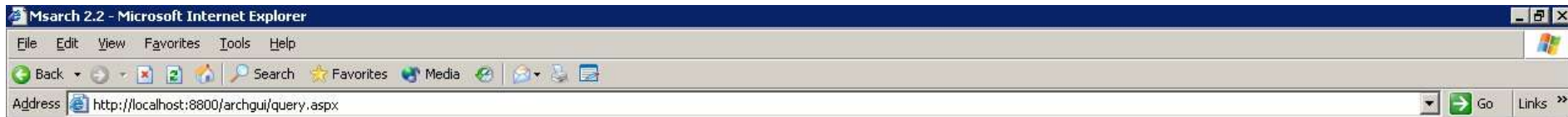
Uwaga:

- ~ jest wydajniejsze niż &
- Używając ~: im bliżej siebie są słowa, tym dopasowanie jest wyżej oceniane

[Pomoc szczegółowa](#)

Licencja dla: CC Open Computer Systems Ltd.  
Wygasa za dni: 36

Nowe szukanie    Sprawdź pola    **SZUKAJ**



## Msarch 2.2 - Wyniki przeszukiwania archiwum

[Zmodyfikuj zapytanie](#)

Znalezione wiadomości				Ilość	5
Link	Nadawca	Odbiorca	Temat	Czas	
<a href="#">zobacz</a>	"Tester CC" <test2@cc.com.pl>	<grzegorz.kaczor@cc.com.pl>	Fw: Msarch SMTP Intercept	2005-08-24 19:23:19	1
<a href="#">zobacz</a>	"Tester CC" <test2@cc.com.pl>	<grzegorz.kaczor@cc.com.pl>	Msarch SMTP Intercept	2005-08-24 19:23:19	
<a href="#">zobacz</a>	"Tester CC" <test2@cc.com.pl>	<grzegorz.kaczor@cc.com.pl>, <test2@cc.com.pl>	test proxy z outlook express	2005-08-24 18:51:59	
<a href="#">zobacz</a>	test2 <test2@cc.com.pl>	Grzegorz Kaczor <grzegorz.kaczor@cc.com.pl>	test prox	2005-08-24 18:51:59	
<a href="#">zobacz</a>	Tomasz Ramsza <tomasz.ramsza@cc.com.pl>	Grzegorz Kaczor <grzegorz.kaczor@cc.com.pl>	aaarrgghh	2005-08-22 17:35:44	1

Znalezione załączniki				Ilość	2
Link	Rozmiar	Czas	Tytuł		
<a href="#">zobacz (wiadomość)</a>	39211	2005-08-22 17:35:44	bez tytułu, rozszerzenie .pdf		1
<a href="#">zobacz (wiadomość)</a>	20544	2005-08-22 17:35:44	bez tytułu, rozszerzenie .pdf		

### Informacja na temat przetwarzania zapytania

```

Odpytano katalog, znalezione rezultaty: arch_m_7_2005_8_X4 , 5
Odpytano katalog, znalezione rezultaty: arch_m_7_2005_8_X3 , 0
Odpytano katalog, znalezione rezultaty: arch_a_7_2005_8_X4 , 2
Odpytano katalog, znalezione rezultaty: arch_a_7_2005_8_X3 , 0
  
```

Na tej stronie można:

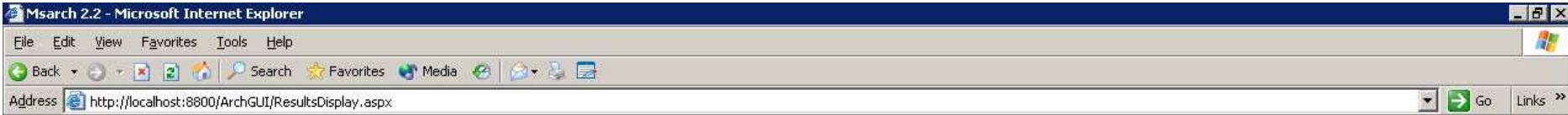
- zobaczyć widok szczegółowy wiadomości - kliknij na [zobacz](#) w kolumnie [Link](#)
- posortować wyniki - kliknij na nagłówek wybranej kolumny
- zobaczyć kolejne strony wyników - użyj numerowanych linków nad nagłówkami kolumn
- otworzyć załączniki - kliknij na [zobacz](#) w kolumnie [Link](#); **upewnij się, że Twój antywirus ma aktualną listę wirusów, zanim otworzysz załącznik**
- zobaczyć wiadomość, z którą przyszedł określony załącznik - kliknij na [wiadomość](#) w kolumnie [Link](#)
- zmodyfikować zapytanie bez ponownego wprowadzania wszystkich kryteriów - kliknij na [Zmodyfikuj zapytanie](#)

Jeśli wystąpią błędy przetwarzania zapytania, zobaczysz je w liście poniżej wyników.

Jeśli próba przejścia po linku zwróci **HTTP 404**, sprawdź rozszerzenie dokumentu. Może być konieczne zdefiniowanie dodatkowego typu MIME w konfiguracji IIS, żeby serwer zezwolił na jego pobranie. Jest także możliwe, że oprogramowanie antywirusowe usunęło plik.

Nie używaj *cofnij*w przeglądarce.

- Administrator - dostęp pełny
- Użytkownik - dostęp do wiadomości własnych (i grupy)
- Użytkownik - dostęp do załączników może być ograniczony:
  - Dlaczego? W przypadku gdy polityka archiwizacji zakłada archiwizowanie także wiadomości zawirusowanych dostęp do załączników jest silnie niewskazany



## Msarch 2.2 - Wyniki przeszukiwania archiwum

[Zmodyfikuj zapytanie](#) [Odzyskaj zaznaczone](#) [Odwróć zaznaczenie](#) [Wyloquij](#)

Znalezione wiadomości					Ilość	3
Link	Nadawca	Odbiorca	Temat	Czas	WYB	
<a href="#">zobacz</a>	"Tester CC" <test2@cc.com.pl>	<grzegorz.kaczor@cc.com.pl>, <test2@cc.com.pl>	test proxy z outlook express	2005-08-24 18:51:59	[X]	
<a href="#">zobacz</a>	test2 <test2@cc.com.pl>	Grzegorz Kaczor <grzegorz.kaczor@cc.com.pl>	test prox	2005-08-24 18:51:59	[ ]	
<a href="#">zobacz</a>	Tomasz Ramsza <tomasz.ramsza@cc.com.pl>	Grzegorz Kaczor <grzegorz.kaczor@cc.com.pl>	aaarrgghh	2005-08-22 17:35:44	[ ]	
						1

## Zalogowany: user3 \*

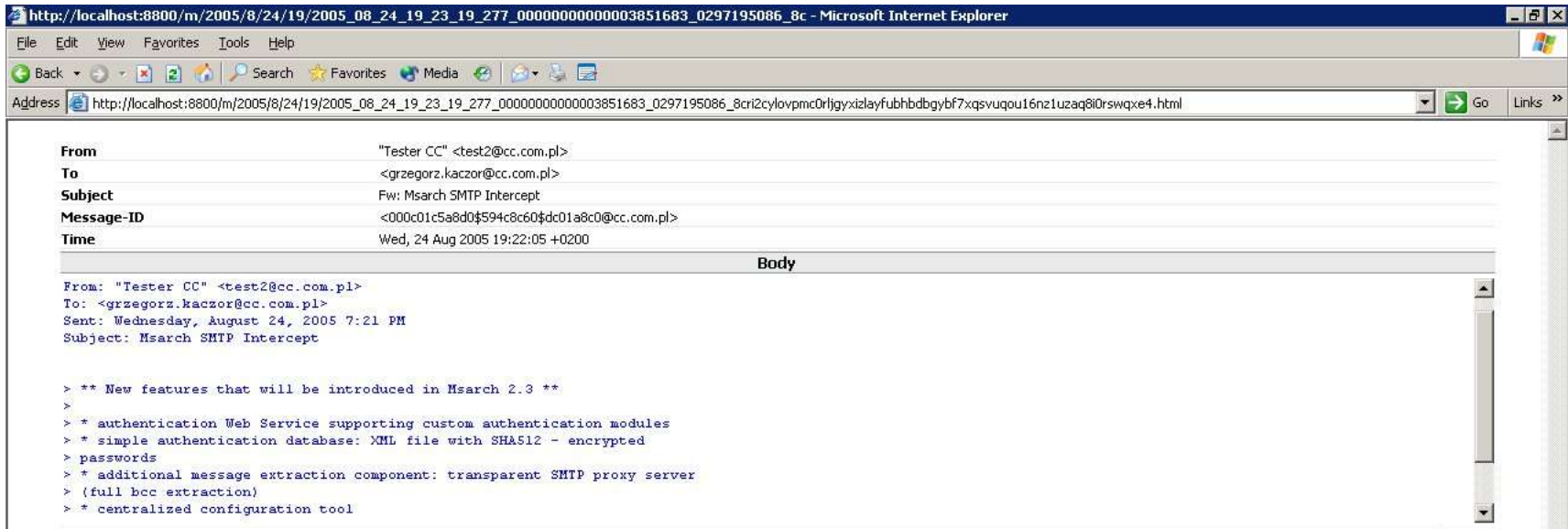
Na tej stronie można:

- zobaczyć widok szczegółowy wiadomości - kliknij na [zobacz](#) w kolumnie **Link**
- posortować wyniki - kliknij na nagłówek wybranej kolumny
- zobaczyć kolejne strony wyników - użyj numerowanych linków nad nagłówkami kolumn
- otworzyć załączniki - kliknij na [zobacz](#) w kolumnie **Link**; **upewnij się, że Twój antywirus ma aktualną listę wirusów, zanim otworzysz załącznik.**
- zobaczyć wiadomość, z którą przyszedł określony załącznik - kliknij na [wiadomość](#) w kolumnie **Link**
- zmodyfikować zapytanie bez ponownego wprowadzania wszystkich kryteriów - kliknij na [Zmodyfikuj zapytanie](#)

Jeśli wystąpią błędy przetwarzania zapytania, zobaczysz je w liście poniżej wyników.

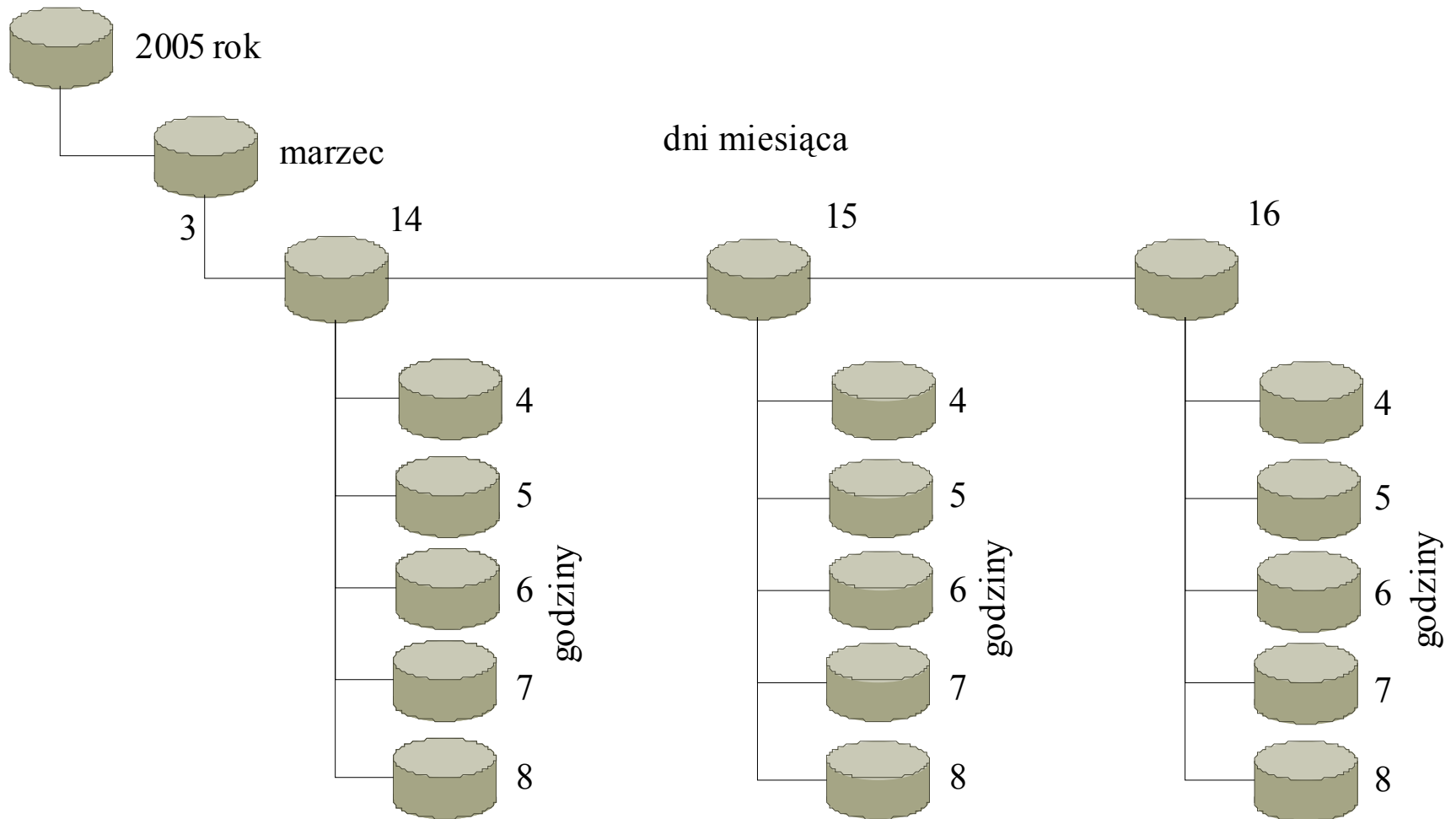
Jeśli próba przejścia po linku zwróci **HTTP 404**, sprawdź rozszerzenie dokumentu. Może być konieczne zdefiniowanie dodatkowego typu MIME w konfiguracji IIS, żeby serwer zezwolił na jego pobranie. Jest także możliwe, że oprogramowanie antywirusowe usunęło plik.

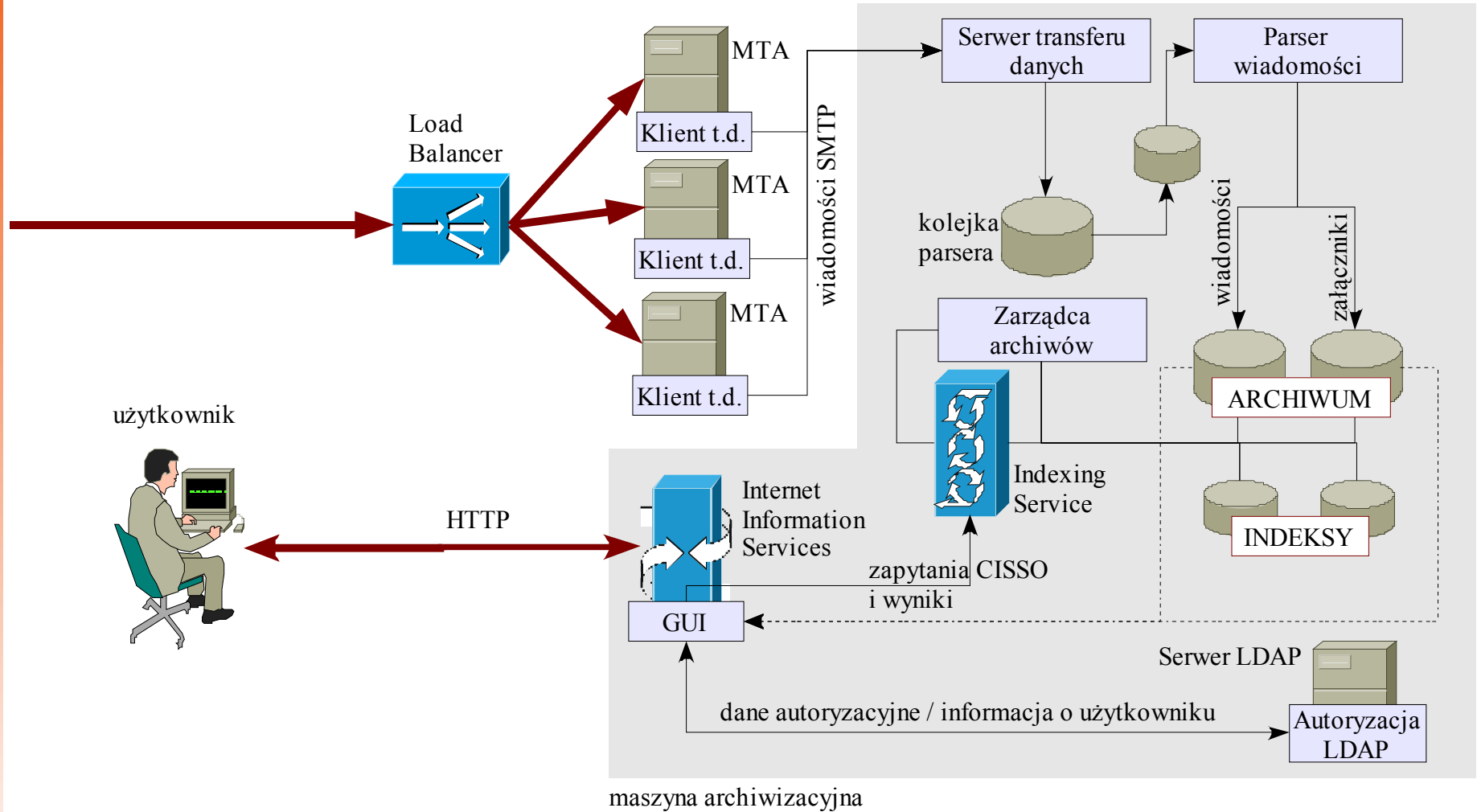
Nie używaj *cofnij* w przeglądarce.



- Architektura scentralizowana z wieloma źródłami danych
  - MSARCH agent - źródło danych dla centralnego archiwum
  - źródeł danych może być dowolna liczba - w typowych większych instalacjach dwie (osobno dla poczty wchodzącej i wychodzącej)
- Serwer archiwizacyjny: jeden
  - engine indeksujący: MS Indexing Service v3, dlaczego?
    - Bezproblemowe indeksowanie dokumentów CDA (Office: .doc, .xls, etc.)
    - bardzo wydajny: indeksier zintegrowany z systemem plików na poziomie jądra systemu operacyjnego
    - Indeksier zmodyfikowany w celu obsługi formatów przesyłek e-mail

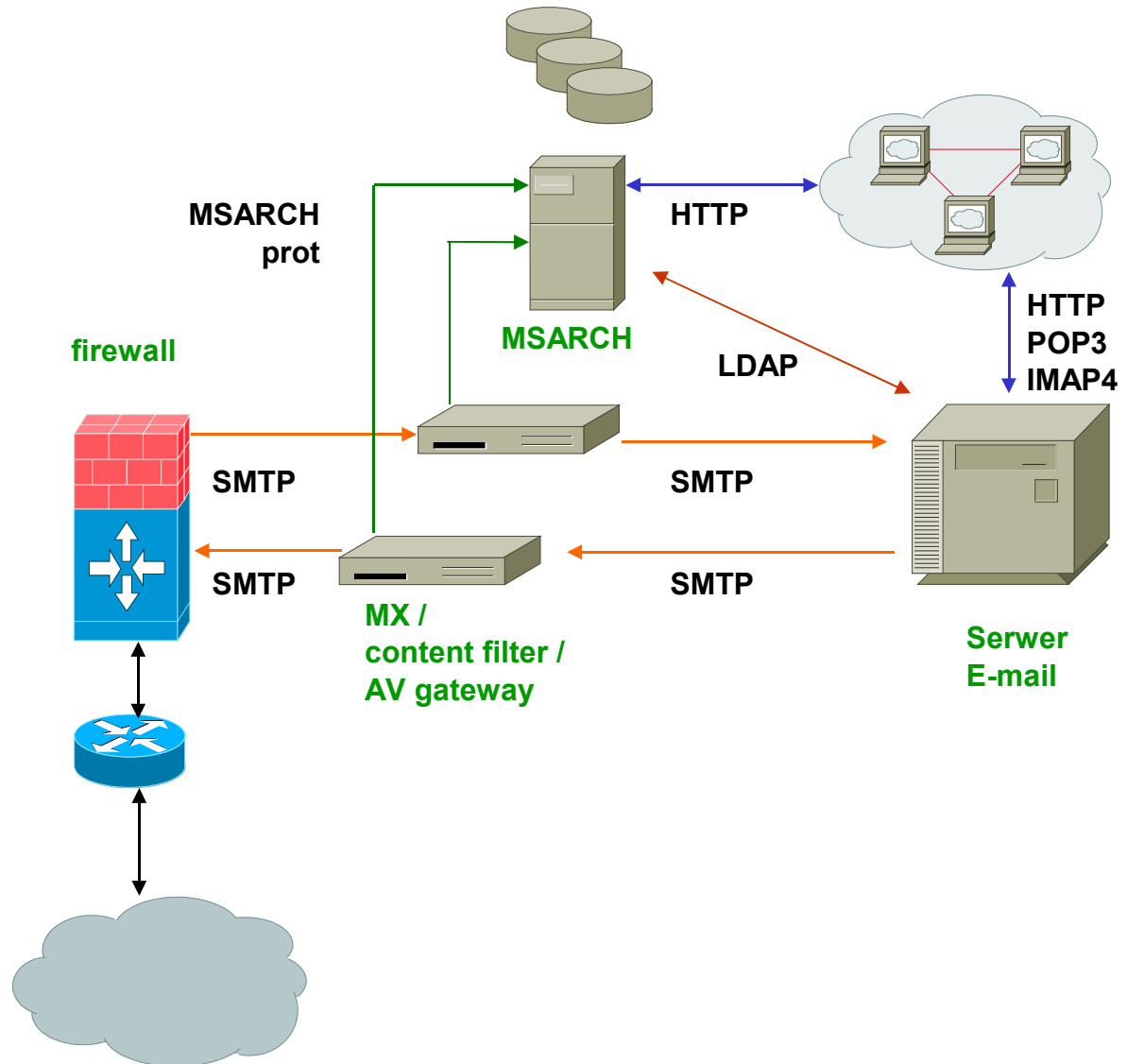
- Dwa główne nad-archiwa:
  - wiadomości
  - załączniki
- Archiwum wiadomości i załączników podzielone na odrębne wykazy - po jednym dla każdego kolejnego tygodnia
  - po upływie tygodnia dane przestają być dodawane do wykazu
  - po upływie kolejnego tygodnia wykaz przełączany jest w tryb RO
  - rotacja wykazów zapewnia też automatyczne usuwanie najstarszej części archiwum
  - struktura katalogów i plików odzwierciedla chronologię
- Podział na wykazy zapewnia:
  - zmniejszenie rozmiaru poszczególnych indeksów
  - przyspiesza wyszukiwanie i reindeksowanie
  - ułatwia usuwanie starych wiadomości





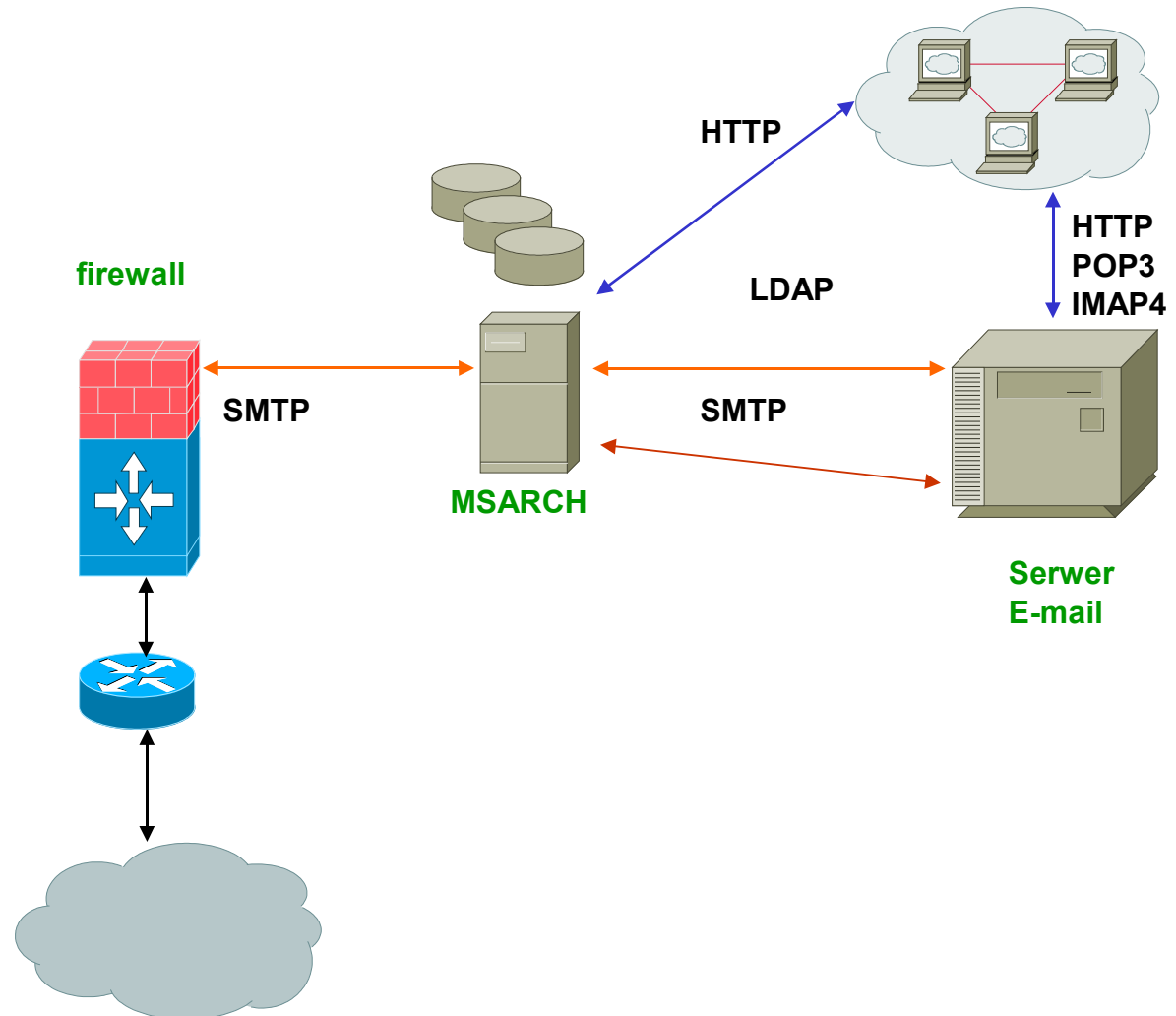
Przykładowa instalacja:

- Ruch **SMTP** rozdzielony na wchodzący i wychodzący
- Bramki SMTP pomiędzy firewall-em a serwerem pocztowym
- **Kopiowanie** wiadomości wykonane na bramkach
- **Autoryzacja** do serwera domeny (tutaj w uproszczeniu - do serwera e-mail)



Przykładowa instalacja:

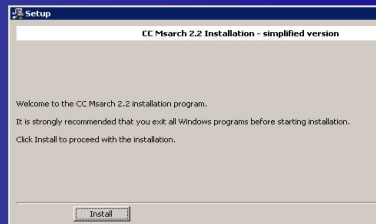
- Ruch **SMTP** przechodzi przez przezroczyste proxy wbudowane w MSARCH
- Na firewall musi funkcjonować bramka SMTP
- **Autoryzacja** do serwera domeny (tutaj w uproszczeniu - do serwera e-mail)
- Proxy zintegrowane z MSARCH może też być wdrożone na osobnej maszynie



- Instalator dostępny w wersji polskiej i angielskiej
- Instalator dokonuje niezbędnych modyfikacji MS IS oraz konfiguruje MS IIS.
- Istnieje wersja ewaluacyjna (klucz czasowy)

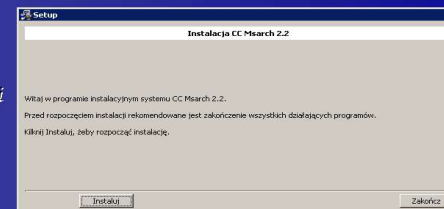
## *CC Msarch 2.2 - e-mail storage & search*

- \* central storage for multiple MTAs
- \* real-time indexing
- \* huge archives
- \* mail & attachment search
- \* rich criteria
- \* prefix search
- \* boolean query
- \* user access control
- \* mail send-back
- \* authentication
- \* flexible mail parser
- \* attachment filtering
- \* antivirus support
- \* download restrictions
- \* performance counters



## *CC Msarch 2.2 - archiwizacja i wyszukiwanie e-mail*

- \* centralne archiwum dla wielu MTA
- \* indeksowanie na bieżąco
- \* ogromne archiwa
- \* wyszukiwanie w wiadomościach i załącznikach
- \* bogate kryteria
- \* szukanie prefiksowe
- \* zapytania logiczne
- \* kontrola dostępu
- \* odsyłanie wiadomości
- \* autoryzacja
- \* elastyczny parser
- \* filtr załączników
- \* pomiar wydajności
- \* wsparcie dla antywirusa
- \* ograniczenia na pobieranie plików



**Msarch 2.X Visual Config**

**General settings**

Program path: C:\Program Files\CC Msarch 2.2

Repository location: C:\Program Files\CC Msarch 2.2\data

Index set location: C:\Program Files\CC Msarch 2.2\data\i

Swap partition: [v]

**Repository Manager**

Messages are kept in repository for days: 20

Indexing Service stop timeout [minutes]: 10

Indexing Service start timeout [minutes]: 10

Index rotation time [minutes past sunday 0:00]: 122

Old data deletion interval [days back]: 100

**Mail Parser Service**

Spool folder: C:\Program Files\CC Msarch 2.2\data\

Temporary folder: C:\Program Files\CC Msarch 2.2\data\

Max message size: 20000000

Number of parsing threads: 4

Initial inactivity timeout [seconds]: 10

Inactivity timeout increase [seconds]: 2

Max inactivity timeout [seconds]: 60

Performance dump interval [seconds]: 180

**Attachment extension filter**

Enable filter

Default action: \*

Save

Discard

Acceptable future message date [hours in future]: 1

Keep raw message data in repository

**Message Transfer Server**

Max message size [bytes]: 20000000

Output folder: C:\Program Files\CC Msarch 2.2\

Message name prefix: msgt-

Bind IP: 127.0.0.1 Bind port: 889

Message extension: .msg

1.0.w

**Allowed clients**

Client IP	Port	Shared secret	Spool	Max message size	Accept ext
▶ 127.0.0.1	900	whMINWyT3c0aV4I	C:\Program Files\CC	20000000	.msg, .eml
* [empty]					

**Search GUI**

IP address: 127.0.0.1 Port: 8800

WWW root path: C:\Program Files\CC Msarch 2.2\data\w

I18N: polski

Detailed errors

Custom error page  Local time fix

**Individual Access Mode**

Enable individual access (Forms Authentication)

Auth server IP: 192.168.1.11 Port: 41114

Transfer encryption key: [masked]

Enable Admin Mode (unrestricted search)

Enable message selection

Enable Send Back (SMTP message restore)

Mail relay IP: 192.168.10.6 Port: 26

Sender email: msarch@cc.com.pl

Enable non-admin attachment download

**ACTIONS**

**Import**

Load configuration

**Export**

Save configuration

Save config as...

Export clients' configuration

Export GUI configuration

C:\Program Files\CC Msarch 2.2\globconf.xml

**HELP**

General

Repository

Mail Parser

Search GUI

Individual Access

Send Back

Authentication

System

**Msarch 2.X Visual Config**

Copyright (C) 2005 CC Open Computer Systems Ltd.

Configuration loaded successfully

- Wielkość archiwum:
  - Teoretycznie rzędu kilku TB
  - największe wdrożone w praktyce: 0,5 TB
    - 450 GB danych
    - 60 GB indeksów
    - 3,5 mln wiadomości
    - 53 000 wiadomości, 9 GB danych dziennie
    - maszyna: 2 x Xeon 2,4 GHz, 1 GB RAM
- Czas dostępu do wiadomości w archiwum zawierającym kilka mln wiadomości 5-30 s

Na naszym stoisku można otrzymać CD  
testowe!

Lub pobrać z Internetu:

<http://www.cc.com.pl/pl/prods/msarch/home.php>

**Dziękuję za uwagę!**

**Pytania?**