Aspekty wydajności systemu dLibra

1. Wprowadzenie

Wydajność jest jednym z kluczowych czynników branych pod uwagę przy tworzeniu kolejnych wersji oprogramowania dLibra. Architektura systemu została zaprojektowana tak, aby umożliwiać jego skalowalność poprzez rozkładanie obciążenia na wiele komputerów. Zostało to opisane w dokumencie Architektura oraz możliwości skalowania systemu dLibra. Kluczowymi czynnikami wpływającymi na obciążenie generowane przez system dLibra są:

- liczba nowych publikacji (np. dziennie),
- liczba czytelników,
- liczba publikacji w systemie.

Pierwszy z tych parametrów związany jest z koniecznością zaindeksowania każdej nowej publikacji w bibliotece cyfrowej. Opisany on został w następnym punkcie. Pozostałe dwa parametry związane są z potrzebą prezentacji czytelnikowi danych gromadzonych w systemie. Gdy są to jednorazowo duże ilości danych (np. indeks wszystkich tytułów w bibliotece cyfrowej), system jest poważnie obciążany. Związane z tym kwestie opisane są w punkcie 3.

2. Indeksowanie zasobów

Za indeksowanie zasobów odpowiada w systemie dLibra usługa Search Server. W wersji 2.5 usługa ta indeksuje zasoby lokalna i zdalne, w wersji 3.0 indeksowanie zasobów zdalnych zostało przeniesione do usługi Distributed Search Server.

2.1. Zasoby lokalne

Każda opublikowana publikacja jest indeksowana. Osobno indeksowane są jej metadane (w schemacie Dublin Core i w schemacie charakterystycznym dla danej biblioteki cyfrowej), a osobno treść. Największe obciążenie generuje tutaj:

- analiza plików publikacji w poszukiwaniu tekstu i ekstrakcja tego tekstu (jeżeli istnieje),
- włączanie metadanych/treści do indeksów wyszukiwawczych biblioteki cyfrowej.

Obciążenie pierwszej z tych operacji zależy od wielkości plików publikacji, ich liczby oraz formatu. W przypadku drugiej operacji dodatkowo istotna jest wielkość indeksu, do którego dane są włączane, gdyż indeks ten jest przebudowywany. Od wersji 2.0 dLibry nowe publikacje do indeksu treści wprowadzane są raz na dobę - okazało się, że przy dużych indeksach (np. gdy większość publikacji posiada OCR) przebudowa indeksu dla każdej publikacji osobno łatwo może prowadzić do przeciążenia systemu. W wersji 3.0 możliwe jest również zmniejszenie częstotliwości przebudowy wszystkich pozostałych indeksów.

2.2. Zasoby zdalne

Mechanizm wyszukiwania rozproszonego polega na tym, iż każda z bibliotek cyfrowych pobiera metadane na temat publikacji znajdujących się w innych bibliotekach i indeksuje je. Indeksowanie metadanych pojedynczej publikacji nie jest procesem czasochłonnym, jednak w skali całej platformy bibliotek cyfrowych takich nowych publikacji może być nawet kilkaset dziennie. Z tego powodu czasami również indeksowanie zasobów rozproszonych mocniej obciąża system. Indeksowanie to standardowo rozpoczyna się o 4 nad ranem.

Do wersji 2.5 włącznie indeksowanie zasobów rozproszonych było obsługiwane w tej samej kolekcje zadań, co indeksowanie zasobów lokalnych. W związku z tym mogła się zdarzyć sytuacja, iż duża ilość publikacji dodanych w zdalnych bibliotekach powodowała opóźnienie indeksowania zasobów lokalnych. Od wersji 3.0 operacje te zostały rozdzielone i wzajemne blokowanie nie powinno już mieć miejsca.

3. Aplikacja czytelnika

Aplikacja Czytelnika jest punktem dostępu do systemu pozwalającym użytkownikom przeglądać, przeszukiwać i pobierać zasoby gromadzone w bibliotece cyfrowej. Duża liczba równoczesnych czytelników jest oczywistą przyczyną obciążenia systemu. Dodatkowy problem sprawiają żądania użytkowników, w odpowiedzi na które system musi wygenerować dużą ilość danych. Poniżej opisano dwie najważniejsze tego typu sytuacje.

3.1. Indeksy

Jednym z najbardziej kłopotliwych dla Aplikacji Czytelnika żądań są:

- indeks wartości atrybutów dla danej kolekcji (np. autorów, tytułów, słów kluczowych),
- indeks publikacji dla danej kolekcji.

Dodatkowo poważnym obciążeniem jest też wyświetlanie użytkownikom automatycznych podpowiedzi (rozwijanych list) w formularzach wyszukiwania - ono również bazuje na indeksach wartości poszczególnych atrybutów z dodatkową informacją o częstości występowania poszczególnych wartości. Informacje niezbędne do obsługi wspomnianych żądań użytkowników w przypadku małej biblioteki cyfrowej (kilkadziesiąt publikacji) mogą być generowane zawsze bezpośrednio z bazy danych. Jednak w przypadku większej liczby publikacji konieczne jest wcześniejsze utworzenie odpowiednich list w pamięci i okresowe (standardowo raz na dobę) odświeżanie ich zawartości. List te generowane są przy starcie Aplikacji Czytelnika. Powoduje to rosnący wraz z liczbą publikacji czas startu serwera Tomcat. Żeby skrócić ten czas startu w wersji 3.0 zdecydowaliśmy się na zwiększenie ilości informacji, które są przechowywane w bazie danych. Dodane zostały wprost informacje o dziedziczeniu tzn. dla przykładu:

- w wersji 2.5 w bazie danych przechowywana była tylko informacja o bezpośrednim przypisaniu publikacji do jakiejś kolekcji,
- dodatkowo można było wyznaczyć do jakich jeszcze kolekcji publikacja należy ze względu na strukturę kolekcji (nadkolekcje) i dziedziczenie przynależności z publikacji grupowych.

W wersji 3.0 w bazie danych dodatkowo przechowywana jest informacja o tym, do jakich kolekcji publikacja należy nie wprost, lecz właśnie przez dziedziczenie. Podobnie jest z wartościami atrybutów. Efektem jest z jednej strony wzrost rozmiaru bazy danych, a z drugiej przyspieszenie działania systemu. Dla przykładu - czas startu Aplikacji Czytelnika na WBC po migracji do wersji 3.0 spadł z około 1h do około 20 minut.

3.2. Struktura publikacji grupowych

Ostatnio zauważonym problemem była kwestia wyświetlania struktury bardzo "płaskich" publikacji grupowych - takich jak np.: Akta Braci Czeskichw WBC. Jednorazowo trzeba było pobrać z bazy kilkaset tytułów i zaprezentować je czytelnikowi. W ramach poprawki do wersji 2.5 wprowadziliśmy przy obsłudze tej strony stosowną pamięć cache dla struktury publikacji. Po tej poprawce obciążenie związane z takimi publikacjami bardzo wyraźnie spadło.