

Specyfikacja przedmiotu zamówienia

Przedmiotem zamówienia jest wykonanie oraz przeniesienie praw autorskich do 6 modułów programistycznych systemu teleinformatycznego stanowiącego szkielet infrastruktury badawczej, która umożliwi efektywne prowadzenie prac badawczo rozwojowych z zakresu zastosowania uczenia maszynowego i głębokiego uczenia (sieci neuronowe) do ekstrakcji wiedzy z dużych zbiorów danych, wyznaczania Potencjału Zasobu Wiedzy oraz tworzenia profili semantycznych dla zasobów wiedzy.

W zakres zamówienia wchodzi zaprojektowanie, wytworzenie, rozruch, wdrożenie, świadczenie asysty technicznej, przeniesienie praw autorskich oraz uruchomienie specjalistycznego oprogramowania informatycznego stanowiącego szkielet infrastruktury badawczej zgodnie z poniższą specyfikacją:

1. Moduły

1.1. Specyfikacja ogólna modułów

1.1.1. Moduł nr 1. „Importer danych” - Zadaniem modułu jest zasilanie hurtowni analitycznej o nowe dane pozyskane z zewnętrznych źródeł (DarkData). Zakładamy, że moduł ten będzie w ramach prac badawczych zasilany o nowe „wtyczki” pozyskujące wartościowe dane, wytypowane w oparciu o wyniki analiz przeprowadzone na gotowym systemie.

1.1.1.1. Importer danych ciągłych (DataImporter1) - Zadaniem komponentu importu danych ciągłych jest zasilanie systemu o dane pozyskane ze źródeł „ciągłych”, czyli dostarczających dane w sposób stały i przyrostowy (dane typu DarkData). Będą to takie źródła danych, które są stale zasilane i które warto pozyskać. Zakładamy, że dla każdego ze źródeł danych będzie możliwe przygotowanie oddzielnej „wtyczki”, pozyskującej przyrostowo dane w określonych przedziałach czasu, z określonego źródła. Zakładamy, że komponent w ramach prac badawczych będzie rozbudowywany o nowe „wtyczki” w razie potrzeb i pozyskiwania nowych źródeł danych.

1.1.1.2. Importer wsadowy (DataImporter2) - Zadaniem komponentu importu wsadowego, jest zasilanie systemu o dane pozyskane ze źródeł danych serwujących dane nie w sposób ciągły (dane typu DarkData). Często będą to dane „jednorazowe”. Zakładamy, że dla każdego źródła danych, będzie możliwe przygotowanie oddzielnej „wtyczki”, która będzie potrafiła przyjąć dane do systemu. Dodatkowo, komponent będzie realizował także możliwość wprowadzania danych w trybie „ręcznym”. Dane będą importowane do systemu na żądanie. Zakładamy, że komponent w ramach prac badawczych będzie rozbudowywany o nowe „wtyczki” w razie potrzeb i pozyskiwania nowych, wartościowych źródeł danych.

- 1.1.1.3.** Importer danych – konwerter danych (DataConverter) - Zadaniem komponentu konwertera danych jest konwersja danych z formatu źródłowego (DarkData) danego źródła danych, na opracowany wcześniej, zunifikowany format danych. Przygotowane przez kolejne „wtyczki” komponentu dane, będą gotowe do oczyszczenia i dalszego przetwarzania przez kolejne moduły systemu. Ze względu na efektywność, komponent ten może być zintegrowany z konkretnymi wtyczkami dla poszczególnych źródeł danych. Zakładamy, że komponent w ramach prac badawczych będzie rozbudowywany o nowe „wtyczki” w razie potrzeb, ewentualnych zmian w zunifikowanym modelu i pozyskiwania nowych źródeł danych.
- 1.1.1.4.** Importer danych – czyszczenie (DataCleaner) - Zadaniem komponentu jest wstępne oczyszczanie zunifikowanych danych. Zakładamy, że dane wychodzące z komponentów importera i konwertera będą wymagały wstępnego oczyszczenia i konwersji bezstratnej (czyli takiej, gdzie konwersja jest jednoznaczna). Komponent będzie uruchamiał na zunifikowanych danych kolejne „wtyczki”, których zadaniem będzie oczyszczanie danych.
- 1.1.1.5.** Importer danych – filtrowanie (DataFilter) - Zadaniem komponentu jest odpowiednie oznaczenie danych zakwalifikowanych jako „błąd oczywisty” tak, aby dalsze moduły systemu pominęły je.
- 1.1.2. Moduł nr 2 „Ekstrakcja danych”** - Zadania modułu Ekstrakcji danych to transformacja, czyszczenie i ekstrakcja danych ze zunifikowanego formatu źródłowego na dane w postaci obiektów i metadanych obiektów, które następnie zasilą główne repozytorium danych systemu, a więc będą stanowiły dane źródłowe dla systemów prezentacyjnych i laboratoryjnych (analitycznych). Wszystkie dane powstałe w wyniku działania modułu docelowo trafią do głównego repozytorium danych systemu będącego głównym repozytorium danych do dalszych prac analityczno-badawczych. Każdy kwant informacji powinien mieć oznaczenie importu i źródła. Działanie poszczególnych komponentów będzie polegało na ekstrakcji odpowiednich danych, a następnie ich dokładnym czyszczeniu i opisanu wyznaczonym wskaźnikiem poziomu ufności. Całość będzie opierała się na systemie „wtyczek”, czyli mniejszych komponentów, które można przypinać, modyfikować, dodawać i usuwać z poszczególnych komponentów, a których zadaniem będzie wyczyszczenie danych.
- 1.1.2.1.** Ekstraktor obiektów (ObjectExtractor) - Głównym zadaniem komponentu jest ekstrakcja obiektów ze zunifikowanego źródła danych, ich analiza i czyszczenie w oparciu o kolejne, testowane „wtyczki”.
- 1.1.2.2.** Ekstraktor metadanych (MetadataExtractor) - Głównym zadaniem komponentu jest ekstrakcja metadanych dla obiektów i powiązań z zunifikowanego źródła danych w oparciu o kolejne, testowane „wtyczki”.

1.1.2.3. Ekstraktor powiązań (AsociacionExtractor) - Głównym zadaniem komponentu jest ekstrakcja powiązań pomiędzy obiektami z zunifikowanego źródła danych w oparciu o kolejne, testowane „wtyczki”. Główne zadania to ekstrakcja powiązań, ich oczyszczenie, mapowanie ich cech na słowniki oraz wyznaczanie dla nich poziomów ufności.

1.1.3. Moduł nr 3 „DataIndexer” - Zadania modułu inteligentnego indeksera, to identyfikacja, nazwanie i numerowanie danych jednakowych. Do głównych zadań modułu inteligentnego indeksera należy właśnie wykrywanie potencjalnych duplikatów (zarówno obiektów, jak i ich powiązań) i ich odpowiednie opisanie i oznaczenie. W ten sposób zbudowana zostanie baza unikalnych obiektów i ich powiązań.

1.1.3.1. Baza parametrów deduplikacji (DeduplicationParameterHandler) - Komponent będzie odpowiedzialny za możliwość tworzenia własnych grup ważonych parametrów pozwalających na określenie duplikatu. Dana grupa parametrów będzie przypisana dla danej klasy obiektu, jedna klasa obiektu może posiadać więcej niż jedną grupę parametrów – jest to zależne od liczby wykorzystywanych algorytmów. Zakładamy, że parametry powstaną w wyniku pracy analitycznej. Komponent będzie stanowił bazę wiedzy oraz podstawę pracy algorytmów deduplikacji danych. Odpowiednia definicja parametrów umożliwi wykorzystanie zaawansowanych algorytmów deduplikacji danych, takich jak sieci neuronowe, drzewa decyzyjne, itp.

1.1.3.2. Konwerter parametrów (DeduplicationParameterConverter) - Komponent będzie odpowiedzialny za konwersję danych pochodzących z głównej bazy danych zgodnie z przypisanymi algorytmami. Oprócz konwersji zgodnie z parametrami komponent jest również odpowiedzialny za normalizację danych wejściowych do algorytmów (np. normalizacja za pomocą standardowego odchylenia średniokwadratowego).

1.1.3.3. Deduplikator danych (DeduplicationAlgorithms) - Komponent będzie odpowiedzialny za przetwarzanie danych w celu wykrycia duplikatów. Będzie on bazował na przygotowanych testowanych „wtyczkach”, między innymi na algorytmach prostych opartych na porównaniu pól, ewentualnie z uprzednim oczyszczeniem, po algorytmy uczenia maszynowego (sieci neuronowe).

1.1.3.4. Numerator danych (GICIDHandler) - Komponent ten będzie zarządzał numeracją obiektów. Zakładamy, że każdy nowy, zdeduplikowany obiekt zostanie oznaczony nowym, unikalnym identyfikatorem (w celu jego jednoznacznej identyfikacji).

1.1.3.5. Rozstrzyganie sporów (DataStewardComponent) - Ze względu na wysoką staranność o jakość wyników przewidywana jest w wyjątkowych

przypadkach pomoc człowieka rozumianego jako DataStewarda, który ostatecznie rozstrzygnie o wyniku.

1.1.4. Moduł nr 4 „DataProfiler” - Zadaniem modułu Inteligentnego Profilera jest przygotowanie bazy profili obiektów (najlepszych wersji). Przyjmujemy, że istnieje możliwość aktualizacji najlepszych wersji obiektów wraz z rosnącą liczbą ich wersji pochodzących z różnych źródeł zewnętrznych. Moduł Inteligentnego Profilera zostanie zrealizowany przez kilka głównych komponentów, opartych o testowane w procesie badawczym „wtyczki” wyłuskujące wartościowe dane i wyznaczające wartości parametrów.

1.1.4.1. Baza parametrów profili (ProfileParameters) - Komponent ten będzie pozwalał na budowanie charakterystycznych grup parametrów ilościowych oraz jakościowych dla danej klasy obiektów. Parametry w głównej mierze będą wydobywane dzięki pracy analitycznej oraz z biegiem czasu aktualizowane. Celem jest nie tylko wyłonienie najlepszej wersji danych opisujących obiekt, ale również określenie jej jakości na podstawie źródła pochodzenia oraz metadanych i ich poziomu ufności.

1.1.4.2. Konwerter parametrów (ProfileParameterConverter) - Głównym zadaniem będzie konwertowanie obiektu zgodnie z przypisanymi do jego typu parametrami. Będzie stanowił więc podstawę do pracy „wtyczek” algorytmów Profilera.

1.1.4.3. Profiler danych (ProfilerAlgorithms) - Komponent będzie odpowiedzialny za odnajdowanie, budowanie i ocenę gotowych profili obiektów. Gotowe profile, jak i wszystkie wytworzone przez komponent dane, będą składowane w głównej bazie profili (MasterDB).

1.1.5. Moduł nr 5 „Portal prezentacyjny” - Moduł ten będzie odpowiedzialny za prezentację danych pochodzących ze wszystkich podsystemów laboratorium. Moduł zostanie zrealizowany jako portal internetowy, serwujący i pozwalający na analizę zgromadzonych danych. Dzięki niemu, możliwe będzie zarządzanie procesem badawczym i porównywanie wyników działań poszczególnych algorytmów. Komponent będzie prezentował informacje o wyekstrahowanych przez system danych o obiektach, ich powiązaniach i profilach. Portal będzie opierał się o dane zgromadzone w bazie MasterDB, a więc będzie prezentował informacje zakwalifikowane jako najlepsze i najpewniejsze. Będzie umożliwiał ich przeszukiwanie, prezentowanie i porównywanie zgromadzonych informacji, wraz z historią ich „utworzenia”.

1.1.5.1. Baza obiektów, powiązań i profili - Komponent będzie prezentował informacje o wyekstrahowanych przez system danych o obiektach, ich powiązaniach i profilach. Portal będzie opierał się o dane zgromadzone w bazie MasterDB, a więc będzie prezentował informacje zakwalifikowane jako najlepsze i najpewniejsze. Będzie umożliwiał ich przeszukiwanie,

prezentowanie i porównywanie zgromadzonych informacji, wraz z historią ich „utworzenia”.

1.1.5.2. Baza analityczna - Komponent będzie pozwalał na analizę danych zgromadzonych w systemie.

1.1.5.3. Użytkownicy i portal typu SocialMedia - Komponent będzie umożliwił tworzenie i zarządzanie kontami użytkowników w module prezentacyjnym. Użytkownicy będą mogli zgłaszać informacje o nieprawidłowych danych, bądź propozycje ulepszeń i modyfikacji kolejnych iteracji badań. Zgłoszenia takie trafią do operatora, który będzie mógł (bądź nie) wprowadzić stosowne poprawki. Każdy użytkownik będzie przypisany do odpowiednich ról, np. aktywni użytkownicy będą mogli dostawać od systemu dane do potwierdzenia.

1.1.6. Moduł nr 6 „DataAnalyser” - Moduł będzie stanowił główną bazę analityczną, pozwalającą na dogłębną analizę pozyskanych, zgromadzonych i przetworzonych przez system danych. Zakładamy, że wszystkie dane zgromadzone w systemie, w tym dane zgromadzone lub wytworzone w kolejnych modułach trafią, bądź zostaną skonsolidowane w dużej bazie badawczej pozwalającej na skuteczną analizę dużych baz danych. Wykorzystane mogą tu zostać takie narzędzia jak Apache Hadoop, Apache Spark, MS HDInsight i tym podobne. Umożliwi to dowolne i dogłębne prowadzenie analiz i weryfikacji jakości przygotowanych danych. Wynikiem prac analitycznych będą także nowe i poprawione „wtyczki” algorytmów do przebadania, między innymi wtyczki czyszczące dane, algorytmy deduplikacyjne, profilujące itp. Dane zawarte w bazie będą mogły posłużyć np. do uczenia sieci neuronowych czy modeli uczenia maszynowego.

1.2. W ramach powyższych modułów powinny zostać zrealizowane również podmoduły wspólne

1.2.1. „CAS” – Moduł ten będzie odpowiedzialny za zarządzanie kontami oraz uprawnieniami użytkowników,

1.2.2. „System powiadomień” – Moduł ten będzie odpowiedzialny za informowanie użytkowników o podejmowanych działaniach w systemie oraz ich wynikach, w szczególności w przypadku niepowodzenia,

1.2.3. „Zarządzanie procesem” – Moduł ten będzie odpowiedzialny za udostępnienie funkcjonalności umożliwiających konfigurację procesów realizowanych w Centrum Badawczo – Rozwojowym na przestrzeni modułów: Importera Danych, Ekstrakcji Danych, DataIndexera, DataProfiler, Portalu Prezentacyjnego oraz Analizy Danych.

1.2.4. Budowa interfejsów wymiany danych między modułami.

1.2.5. Budowa i obsługa procesów utajniania danych.

1.3. Algorytmy – w ramach przedmiotu zamówienia należy przygotować zestaw podstawowych algorytmów umożliwiających przetestowanie poprawności działania laboratorium i podsystemów wchodzących w jego skład.

1.3.1. Algorytmy ekstrakcji obiektów poprzez dekompozycję przypisów bibliograficznych będących ciągami znaków.

1.3.2. Algorytmy asocjacji obiektów.

Wymagania dodatkowe w ramach przedmiotu zamówienia:

Przedmiot zamówienia obejmuje następujące elementy:

2. Przeprowadzenie analizy wymagań (weryfikacja i uszczegółowienie wymagań funkcjonalnych i niefunkcjonalnych oraz opracowanie koncepcji modułów).

3. Stworzenie projektu technicznego:

3.1. Rozwiązania infrastruktury teleinformatycznej (sprzętowej, sieciowej i programowej) modułów dla wymaganych środowisk pracy - uwzględniający dedykowaną przez Zamawiającego infrastrukturę na potrzeby tych modułów.

3.2. Referencyjnej bazy danych (zawierający m.in. opis modułów funkcjonalnych, model danych, opis procesów przetwarzania danych, opis interfejsów, opis bezpieczeństwa systemu itp.).

3.3. Środowiska informatycznego umożliwiającego uruchamianie, zarządzanie i konfigurację algorytmów w wydzielonych obszarach, będących docelowym zadaniem badań w laboratorium.

4. Budowa, rozruch oraz wdrożenie modułów, implementacja algorytmów

4.1. Budowa i konfiguracja infrastruktury teleinformatycznej modułów.

4.1.1. Wdrożenie infrastruktury teleinformatycznej (przeznaczonej w tym celu przez Zamawiającego).

4.1.2. Instalacja, konfiguracja i wdrożenie oprogramowania.

4.1.3. Konfiguracja środowisk.

5. Dodatkowe założenia dotyczące realizacji przedmiotu zamówienia

5.1. Infrastruktura teleinformatyczna i moduły powinny:

5.1.1. Zostać zoptymalizowane do przetwarzania dużych zbiorów danych.

5.1.2. Umożliwiać zrównoleglanie wykonywania algorytmów.

5.1.3. Umożliwiać sterowanie przebiegiem całościowego procesu przetwarzania danych za pomocą panelu zarządczego z możliwością wyboru źródeł danych oraz zakresu danych.

5.1.4. Umożliwiać łatwą integrację z systemami zewnętrznymi.

5.1.5. Umożliwić prowadzenie jednocześnie kilku niezależnych badań na wydzielonych obszarach badawczych.

5.2. Realizacja podstawowych algorytmów dekompozycji, ekstrakcji, asocjacji powinna odbywać się w jak największym stopniu poprzez mechanizmy uczenia maszynowego, głębokiego uczenia z wykorzystaniem sieci neuronowych.

5.3. Realizacja podstawowych algorytmów deduplikacji oraz indeksacji danych powinna odbywać się w jak największym stopniu poprzez mechanizmy uczenia maszynowego, głębokiego uczenia z wykorzystaniem sieci neuronowych.

5.4. Realizacja podstawowych algorytmów profilowania danych powinna odbyć się w jak największym stopniu poprzez mechanizmy uczenia maszynowego, głębokiego uczenia z wykorzystaniem sieci neuronowych w oparciu o wskazane narzędzia analityczne.

6. Testowanie i odbiór modułów:

6.1. Import i ekstrakcja obiektów biznesowych ze wskazanego źródła danych, w tym czyszczenie, konwersja, filtrowanie oraz walidacja danych, w szczególności skompletowanie danych danego obiektu z różnych zdarzeń, które go dotyczą. Liczba rekordów opisujących zdarzenia - ok. 6.000.000, liczba ekstrahowanych typów obiektów - minimum 3, liczba obiektów, które powinny zostać wyekstrahowane - około 1.000.000 obiektów referencyjnych.

6.1.1. Konfiguracja procesu importu (w trybie ciągłym i wsadowym) i ekstrakcji.

6.1.2. Import zadanego źródła danych o zdarzeniach.

6.1.3. Czyszczenie, konwersja, filtrowanie i walidacja danych wg skonfigurowanego wcześniej procesu.

6.1.4. Przeprowadzenie procesu ekstrakcji obiektów wraz z ich metadanymi i powiązaniem z zaimportowanym źródłem wg skonfigurowanego wcześniej procesu.

6.1.5. Testy ilościowe i jakościowe wyekstrahowanych obiektów wraz z ich metadanymi i powiązaniem wg zadanych kryteriów. Ocena poprawności procesu na podstawie bazy wzorcowej.

6.1.6. Testy odporności na awarie (utrata lub desynchronizacja przetwarzanych danych).

6.2. Import, ekstrakcja oraz indeksowanie obiektów biznesowych ze wskazanego źródła danych, w tym czyszczenie, konwersja, filtrowanie, walidacja, deduplikacja i numerowanie danych, pochodzących z kilku niezależnych źródeł danych w celu sporządzenia referencyjnej bazy danych. Liczba niezależnych źródeł danych - minimum 3, łączna liczba rekordów we wszystkich źródłach danych - minimum 10.000.000 rekordów, liczba typów obiektów - minimum 3, liczba obiektów, które powinny zostać wyekstrahowane - około 5.000.000 obiektów, liczba obiektów referencyjnych - minimum 2.000.000.

- 6.2.1.** Konfiguracja procesu importu (w trybie ciągłym i wsadowym), ekstrakcji, indeksowania, deduplikacji i profilowania danych.
- 6.2.2.** Import zadanych źródeł danych.
- 6.2.3.** Czyszczenie, konwersja, filtrowanie i walidacja danych wraz z określeniem poziomu ufności dla każdego źródła wg skonfigurowanego wcześniej procesu.
- 6.2.4.** Przeprowadzenie procesu ekstrakcji obiektów wraz z ich metadanymi i powiązaniem z zaimportowanymi źródłami wg skonfigurowanego wcześniej procesu.
- 6.2.5.** Przeprowadzenie procesu indeksowania obiektów, w tym deduplikacja, nazwanie i indeksacja danych oraz wskazanie rekordów do obsługi ręcznej (DataSteward) wg skonfigurowanych wcześniej parametrów procesu.
- 6.2.6.** Przeprowadzenie procesu tworzenia referencyjnej bazy danych obiektów wg skonfigurowanych wcześniej parametrów procesu.
- 6.2.7.** Testy ilościowe i jakościowe obiektów referencyjnych wraz z ich metadanymi i powiązaniem wg zadanych kryteriów. Ocena poprawności procesu na podstawie bazy wzorcowej.
- 6.2.8.** Testy odporności na awarie (utrata lub desynchronizacja przetwarzanych danych).
- 6.3.** Testy portalu prezentacyjnego w zakresie prezentacji wyników poszczególnych etapów importu, ekstrakcji, indeksowania i profilowania obiektów pochodzących z testowych zbiorów danych. Liczba niezależnych źródeł danych - minimum 3, łączna liczba rekordów we wszystkich źródłach danych - minimum 10.000.000 rekordów, liczba typów obiektów - minimum 3, liczba obiektów, które powinny zostać wyekstrahowane - około 5.000.000 obiektów, liczba obiektów referencyjnych - minimum 2.000.000, liczba użytkowników jednoczesnych - minimum 100.
 - 6.3.1.** Przeprowadzenie procesu tworzenia oraz zarządzania kontami użytkowników na zadanym źródle danych.
 - 6.3.2.** Prezentacja bazy obiektów, profili i powiązań wraz z historią ich pozyskania w celu analizy przeprowadzanych testowych procesów.
 - 6.3.3.** Graficzna wizualizacja danych umożliwiająca analizę procesowanych obiektów (np. grafy, wykresy) w tym testy integracji modułu portalu prezentacyjnego z platformą Apache Spark.
 - 6.3.4.** Testy integracji modułu portalu prezentacyjnego z modułem inteligentnego indeksera w celu analizy funkcjonalności DataStewarda.
 - 6.3.5.** Testy integracji modułu portalu prezentacyjnego z modułem analizy danych.
 - 6.3.6.** Ocena interfejsów użytkownika w pod kątem intuicyjności, przejrzystości oraz czasu odpowiedzi systemu na przesyłane przez użytkowników żądania (w tym weryfikacja średniego, minimalnego, maksymalnego czasu odpowiedzi).
 - 6.3.7.** Testy ilościowe i jakościowe zebranych informacji o zachowaniu użytkowników.

6.3.8. Testy odporności na awarie (utrata lub desynchronizacja przetwarzanych danych).

6.4. Testy modułu analizy danych. Liczba niezależnych źródeł danych – minimum 3, łączna liczba rekordów we wszystkich źródłach danych – minimum 30.000.000, liczba typów obiektów – minimum 3.

6.4.1. Testy integracji modułu z narzędziami analitycznymi (Apache Spark/Apache Hadoop/ MS HDInsight).

6.4.2. Testy dostarczonych bibliotek zawierających algorytmy uczenia maszynowego na zadanych zbiorach testowych oraz wzorcowych, w tym testy funkcjonalności opisanych w szczegółowej specyfikacji projektu.

6.4.3. Ocena mechanizmów przetwarzania dużych zbiorów danych pod kątem optymalizacji czasu działania (wyniki poniżej określonego progu czasu działania algorytmów).

6.4.4. Testy odporności na awarie (utrata lub desynchronizacja przetwarzanych danych).

6.4.5. Wymagany zestaw testów jednostkowych pokrywających algorytmy oraz udostępniane funkcjonalności.

7. Technologie

Moduły programistyczne powinny zostać wykonane w technologii Java SE 9, Java EE 8 lub równoważnej, która jest podstawową technologią systemów informatycznych Zamawiającego. Dodatkowe preferowane technologie to:

- Relacyjne i nierelacyjne bazy danych
- Scala, Python, R
- Spring 5 (wraz z podprojektami takimi jak Spring MVC, Spring Web-Flow, Spring Data, Spring Security, itp..), Hibernate, JPA2
- Apache Spark, deeplearning4j, WEKA, H2O.io, jblas
- Apache Hadoop (wraz z elementami ekosystemu aplikacji)
- Automatyzacja budowy oprogramowania: Maven, Gradle
- System kontroli wersji GIT
- CSS, JavaScript, HTML5, Angular4
- JSF 2.x
- JUnit
- Solr, Apache Lucene