Opracowanie metodyki zautomatyzowanego zasilania grafowej bazy ryzyka danymi z dowolnych zbiorów danych z wykorzystaniem uczenia maszynowego.

Szczegółowy opis przedmiotu zamówienia.

Spis treści

1. Cel badań 2
2. Podstawowe informacje o systemie TalRisk 2
3. Zakres zadań Wykonawcy 2
   3.1. Opracowanie metody doboru zbiorów uczących i testowych 3
   3.2. Opracowanie metody zasilania wierzchołków grafowej bazy danych i ich atrybutów 4
   3.3. Opracowanie metody zasilania krawędzi grafowej bazy danych i ich atrybutów 5
   3.4. Udoskonalenie metody zasilania wierzchołków i krawędzi grafowej bazy danych i ich atrybutów 5
4. Wymagania dodatkowe 6
5. Dane i zasoby udostępniane Wykonawcy przez Zamawiającego 6
   5.1. Opis modelu grafowej bazy ryzyka GBR 6
   5.2. Zbiory danych ZD_1 7
   5.3. Infrastruktura IT 7
6. Harmonogram realizacji 8

1. Cel badań

Celem projektu, na potrzeby którego będą realizowane przez Wykonawcę badania, jest stworzenie systemu zarządzania ryzykiem i ciągłością działania TalRisk poprzez eksplorację i analizę korelacji danych w czasie rzeczywistym przy użyciu bazy grafowej, z wykorzystaniem in-memory computing oraz uczenia maszynowego.

2. Podstawowe informacje o systemie TalRisk

TalRisk będzie systemem wspomagającym zarządzanie ryzykiem operacyjnym i ciągłością działania.

System TalRisk będzie systemem klasy Big Data, wykorzystującym techniki i technologie z obszaru nauki o danych (ang. data science), a w szczególności uczenia maszynowego.

System zapewni automatyzację najbardziej pracochłonnych i wrażliwych na czas procesów, których realizacja z akceptowalnym opóźnieniem bez wsparcia systemu informatycznego jest bardzo trudna, a w niektórych przypadkach wręcz praktycznie niemożliwa:

1. tworzenia i aktualizacji bazy ryzyka,
2. szacowania ryzyka,
3. wykrywania i prognozowania zdarzeń niepożądanych (incydentów).

Grafowa baza ryzyka (GBR) systemu TalRisk zawierać będzie m.in. dane o aktywach przedsiębiorstwa, realizowanych przez nie projektach i procesach oraz powiązanych z nimi zagrożeniach, podatnościach, zabezpieczeniach i incydentach. Aktywa w tym kontekście rozumiane są jako wszystkie pozostające w dyspozycji przedsiębiorstwa zasoby, niezbędne do realizacji jej celów biznesowych, w szczególności nieruchomości, urządzenia, jednostki organizacyjne, pracownicy i ich kompetencje.

Baza ryzyka GBR zarządzana będzie przez serwer OrientDB, z wykorzystaniem modelu grafowego.

3. Zakres zadań Wykonawcy

Zadaniem Wykonawcy jest opracowanie metodyki zautomatyzowanego zasilania grafowej bazy ryzyka GBR danymi z dowolnych zbiorów danych ZD z wykorzystaniem nienadzorowanego i częściowo nadzorowanego uczenia maszynowego.

Przez „dowolność” zbiorów danych ZD należy rozumieć brak ustandaryzowanej struktury ZD u poszczególnych klientów, u których system TalRisk będzie wdrażany.

Metodyka ma umożliwiać:

1. znaczące obniżenie pracochłonności inicjalnego zasilenia GBR w stosunku do pracochłonności podejścia opartego na ręcznym definiowaniu reguł transformacji danych,
2. stworzenie w ramach inicjalnego zasilenia modeli umożliwiających automatyczną aktualizację GBR na podstawie zmian w ZD z minimalnym opóźnieniem czasowym.

Ze względu na wymagane obniżenie pracochłonności, w tym pracochłonności etykietowania danych ze zbiorów uczących i testowych, Zamawiający oczekuje jak najszerszego zastosowania nienadzorowanego lub częściowo nadzorowanego uczenia maszynowego.

W ramach niniejszego zamówienia Wykonawca zobowiązany będzie do zaimplementowania narzędzi pomocniczych i zasilenia grafowej bazy ryzyka GBR_1 danymi z udostępnionych przez Zamawiającego zbiorów danych ZD_1, z wykorzystaniem opracowanej w ramach zamówienia metodyki i wykazania na tym przykładzie, że metodyka pozwala na osiągnięcie wymaganej skuteczności zasilenia grafowej bazy danych GBR.

Prace realizowane przez Wykonawcę muszą objąć opisane poniżej zadania szczegółowe.

3.1. Opracowanie metody doboru zbiorów uczących i testowych

W ramach zadania Wykonawca opracować ma metodykę i narzędzia doboru zbiorów uczących i testowych ZUT ze zbiorów danych ZD oraz za ich pomocą przygotować zbiory uczące ZUT_1 (bez etykiet) dla zbiorów danych ZD_1 (etykiety danych dla zbiorów uczących i testowych ZUT_1 uzupełnione zostaną przez Zamawiającego).

Przez zbiory uczące i testowe ZUT należy rozumieć zbioru uczące i testowe wykorzystywane w procesie budowy klasyfikatorów mapujących dane z ZD na obiekty z grafowej bazy ryzyka GBR.

Zbiory uczące i testowe ZUT z jednej strony będą musiały umożliwiać uzyskanie wymaganej skuteczności wszystkich klasyfikatorów, z drugiej strony ich wielkość powinna być jak najmniejsza, aby umożliwić uzyskanie zakładanego obniżenia pracochłonności całego procesu inicjalnego zasilenia grafowej bazy ryzyka GBR, w tym ręcznego etykietowania obiektów wybranych do zbiorów uczących i testowych ZUT.

Zamawiający wymaga, aby w ramach zamówienia Wykonawca zweryfikował możliwości zastosowania metod próbkowania na podstawie analizy skupień (cluster sampling). Analiza skupień (metoda nienadzorowana) miałaby być stosowana do wyznaczenia klastrów danych w zbiorach danych ZD. Jest wysoce prawdopodobne, że obiekty w ramach jednego klastra będą miały jednakowe etykiety, co może znacznie przyspieszyć proces etykietowania obiektów wytypowanych do zbiorów uczących i testowych ZUT.

Wykonawca może dodatkowo zastosować dowolne inne techniki optymalizacji próbkowania danych do zbiorów uczących i testowych ZUT.

Zamawiający wymaga, aby opracowane przez Wykonawcę metody wyboru obiektów do zbiorów uczących i testowych ZUT pozwalały na opracowanie klasyfikatorów o skuteczności nie niższej niż 70%, przy liczbie obiektów w zbiorach uczących nie większej niż 10% liczby obiektów w zbiorach danych ZD.

3.2. Opracowanie metody zasilania wierzchołków grafowej bazy danych i ich atrybutów

W ramach zadania Wykonawca opracować ma metodykę i narzędzia wspomagające zasilanie wierzchołków grafowej bazy danych i ich atrybutów w oparciu o częściowo nadzorowane uczenie maszynowe.

W ramach zadania należy uwzględnić tworzenie klasyfikatorów:

1. przypisujących obiekty ze zbiorów danych ZD do właściwych wierzchołków GBR,
2. mapujących atrybuty obiektów ze zbiorów danych ZD (i ich wartości) na atrybuty wierzchołków GBR (i ich wartości).

Wykonawca ma zasilić danymi ze zbiorów ZD_1 wierzchołki grafowej bazy ryzyka GBR_1 z wykorzystaniem opracowanej metodyki i narzędzi. Do stworzenia klasyfikatorów muszą zostać wykorzystane zbiory uczące i testowe ZUT_1.

Realizując prace Wykonawca musi uwzględnić fakt, że poszczególne atrybuty pojedynczego wierzchołka grafowej bazy ryzyka GBR mogą pochodzić z więcej niż jednego wewnętrznego źródła danych (np. dane dotyczące serwera mogą być pobierane z ewidencji środków trwałych (dane księgowe) oraz z systemu zarządzania infrastrukturą IT (parametry techniczne)). Opracowana metoda musi uwzględniać identyfikację danych cząstkowych i ich scalenie w ramach wierzchołka grafowej bazy ryzyka GBR. Również w ramach tego zadania istotne jest umożliwienie ręcznego poprawiania reguł generowanych przez algorytmy uczenia maszynowego:

- jawne reguły (np. drzewo decyzyjne) vs. czarne skrzynki (np. sieci neuronowe),
- pozyskiwanie reguł z czarnych skrzynek lub jawnych reguł w zależności od wyboru metody(na potrzeby ręcznego poprawiania reguł).

Kryteria odbioru:

- opracowana metoda etykietowania zasobów z wykorzystaniem uczenia maszynowego ze skutecznością minimum 80% (węzły w grafie),
- opracowana udoskonalona metoda etykietowania zasobów uwzględniająca sprzężenie zwrotne od użytkownika.

3.3. Opracowanie metody zasilania krawędzi grafowej bazy danych i ich atrybutów

W ramach zadania Wykonawca opracować ma metodykę i narzędzia wspomagające zasilanie krawędzi grafowej bazy danych i ich atrybutów w oparciu o częściowo nadzorowane uczenie maszynowe.

Problemem badawczym adresowanym w zadaniu jest określenie sposobu uwzględnienia powiązań między poszczególnymi wierzchołkami grafowej bazy ryzyka GBR. W ten sposób wierzchołki mogą być połączone w graf za pomocą krawędzi. Zadanie dotyczy zarówno krawędzi, które wynikają z opisów w zbiorach danych ZD, jak i krawędzi tworzonych na podstawie wykrywania nieoczywistych relacji.

Za pomocą opracowanej metodyki i narzędzi Wykonawca ma zasilić wierzchołki grafowej bazy ryzyka GBR_1 danymi ze zbiorów danych ZD_1. Również w ramach tego zadania istotne jest umożliwienie ręcznego poprawiania reguł generowanych przez algorytmy uczenia maszynowego:

- jawne reguły (np. drzewo decyzyjne) vs. czarne skrzynki (np. sieci neuronowe),
- określenie możliwości pozyskiwania reguł z czarnych skrzynek (na potrzeby ręcznego poprawiania reguł).

Kryteria odbioru:

- opracowana metoda łączenia danych (data blending),
- opracowana metoda etykietowania powiązań między zasobami (krawędzie w grafie) z skutecznością minimum 80%.

3.4. Udoskonalenie metody zasilania wierzchołków i krawędzi grafowej bazy danych i ich atrybutów

W ramach zadania Wykonawca może podjąć próbę poprawy skuteczności klasyfikatorów opracowanych w zadaniach opisanych w punktach 3.2 i 3.3.

Kryteria odbioru:

- Uzyskanie poprawy nie jest obligatoryjne.
- Zamawiający wymaga przekazania raportu z etapu.

4. Wymagania dodatkowe

1. Wykonawca będzie przygotowywał i przedstawiał do akceptacji przez Zamawiającego szczegółowe plany prac dla każdego trzymiesięcznego etapu realizacji prac, co najmniej 10 dni roboczych od dnia rozpoczęcia etapu, którego dotyczy plan.
2. Wykonawca będzie przedstawiał Zamawiającemu raport ze zrealizowanych w ramach kolejnych etapów prac najpóźniej 10 dni roboczych po zakończeniu etapu.
3. Zamawiający wymaga przeprowadzenia dla wybranych modeli klasyfikacyjnych (co najmniej trzech) dziesięciokrotnego sprawdzianu krzyżowego oraz sporządzenia dla nich krzywych uczenia (w zakresie zbioru danych ZD_1). Wykresy skuteczności dla poszczególnych foldów oraz krzywe uczenia powinny zostać zamieszczone w raporcie z prac.
4. Co najmniej 80% danych ze zbiorów danych ZD_1 musi zostać poprawnie przeniesionych do grafowej bazy ryzyka GBR_1. Ponieważ sprawdzenie będzie wymagało ręcznego przypisania etykiet do danych weryfikacyjnych, spełnienie powyższego warunku będzie badane na zbiorze weryfikacyjnym zawierającym 1% losowo wybranych rekordów ze zbiorów ZD_1, z pominięciem rekordów należących do zbiorów uczących i testowych ZUT_1 (aby w ocenie uwzględnić ewentualny efekt nadmiernego dopasowania zastosowanych klasyfikatorów).

5. Dane i zasoby udostępniane Wykonawcy przez Zamawiającego

5.1. Opis modelu grafowej bazy ryzyka GBR

Zamawiający udostępni Wykonawcy opis modelu grafowej bazy ryzyka GBR obejmujący:

1. klasy wierzchołków wraz z opisem ich atrybutów,
2. klasy krawędzi wraz z opisem ich atrybutów.

Każdy atrybut powyższych klas będzie opisany będzie przez następujące właściwości:

1. nazwę,
2. typ (zgodnie z typami atrybutów obsługiwanymi przez OrientDB),
3. czy wymagane jest przypisanie wartości do atrybutu,
4. czy dopuszczalna jest wartość NULL,
5. MIN i MAX oznaczające w zależności od typu atrybutu:
    a. String – minimalną i maksymalną długość łańcucha,
    b. Integer – minimalną i maksymalną wartość,

c. Date – najwcześniejszą i najpóźniejszą datę,

d. dla typów obsługujących wiele wartości atrybutu (list, set, map) – najmniejszą i największą liczbę wartości.

6. inne ewentualne ograniczenia wartości atrybutu definiowane za pomocą wyrażeń regularnych, w tym lista dopuszczalnych wartości (tylko dla atrybutów o stałej liście wartości).

Uwaga !!!

Nie wszystkie właściwości muszą być określone dla każdego atrybutu.

5.2. Zbiory danych ZD_1

Nieprzetworzone zbiory danych ZD_1 udostępnione zostaną na rozproszonym systemie plików Hadoop Distributed File System (HDFS) w postaci plików płaskich lub w bazach zarządzanych przez Apache HBase.

Sumaryczna liczba klas obiektów i atrybutów w zbiorach danych ZD_1 nie będzie większa od 1000.

5.3. Infrastruktura IT

Zamawiający udostępni Wykonawcy zdalne środowisko serwerowe ZS_1.

Wszystkie prace realizowane w ramach niniejszego zlecenia Wykonawca jest zobowiązany wykonywać na środowisku ZS_1. W szczególności Wykonawca nie może kopiować żadnych elementów udostępnionych przez Zamawiającego, w szczególności danych, poza środowisko ZS_1.

Na środowisku ZS_1 udostępnione zostaną następujące produkty:

1. Red Hat OpenShift,
2. Red Hat Enterprise Linux,
3. OrientDB,
4. Hortonworks Data Platform,
5. Hortonworks DataFlow,
6. Hazelcast,
7. Apache Flink,
8. Apache Spark,
9. Apache Spark MLlib,
10. scikit-learn,
11. TensorFlow,

12. Apache Solr,
13. Elasticsearch,
14. Python,
15. Django Web Framework,
16. Vue.js,
17. Tornado Web Server,
18. Nginx,
19. Node.js,
20. Varnish Cache.

Wykorzystanie przez Wykonawcę innych narzędzi i/lub technologii niż wymienione powyżej wymagać będzie każdorazowo zgody Zamawiającego.

6. Harmonogram realizacji

L.p.	Kamień milowy	Data
1	Udostępnienie zdalnego środowiska serwerowego ZS_1.	01.03.2019
2	Przekazanie opisu modelu grafowej bazy ryzyka GBR_1 i udostępnienie zbiorów danych ZD_1.	01.04.2019
3	Odbiór całości prac.	31.03.2020

Harmonogram zakłada, że:

1. Zamawiający uzupełni etykiety dla zbiorów uczących ZU_1 w ciągu 20 dni roboczych licząc od dnia następnego od dnia przekazania zbiorów uczących i testowych ZUT_1 przez Wykonawcę.
2. Zamawiający zobowiązany będzie do zgłoszenia ewentualnych uwag w ramach procedury odbioru (punkt 3 harmonogramu) w ciągu 13 dni roboczych. Wykonawca będzie zobowiązany do zgłoszenia gotowości do odbioru z odpowiednim wyprzedzeniem w stosunku do podanych w harmonogramie terminów zatwierdzenia i odbioru.