Výzkumný ústav bezpečnosti práce, v. v. i.
JERUZALÉMSKÁ 1283/9
110 00 PRAHA 1 – NOVÉ MĚSTO
ČESKÁ REPUBLIKA

Příloha č. 2 – Popis zakázky

Veřejná zakázka malého rozsahu

Datový repozitář

Předmětem plnění je dodání a implementace datového repositáře, který rozšíří stávající softwarové vybavení Výzkumného ústavu bezpečnosti a práce.

Technická specifikace a požadavky na funkcionality

Datový repozitář musí:

- obsahovat centralizované úložiště (repozitář) pro správu datových a analytických zdrojů (Modeler streamů) za účelem automatizace dataminingových procesů
- být postaven na MS SQL Server (Microsoft).
- zajistit, že datové zdroje a objekty jsou skladovány v definovatelné struktuře, vybavené podrobnými popisy, textovými poznámkami a klíčovými slovy pro pokročilé vyhledávání.
- zachovávat všechny verze analytických zdrojů a objektů.
- být přístupný pro neomezený počet uživatelů pro efektivní sdílení mezi definovanými uživateli.
- obsahovat nástroj pro automatické zpracování nových datových zdrojů z externích úložišť na základě dávkového zpracování nebo zpracování v reálném čase.
- obsahovat nástroj pro automatizaci a plánování spouštění dataminingových analytických procesů ve formátu IBM SPSS Modeler.
- mít zajištěnou správu uživatelů co do přihlašování a do oprávnění. Nezbytné stavy jsou: jen pro čtení, úpravy.
- být vybaven možností definovatelné struktury s různými typy oprávnění pro viditelnost konkrétních souborů s konkrétním oprávněním.
- umožňovat nahrávat datové soubory v následujících formátech: všechny **typy databází** (Microsfoft SQL Server, Oracle a IBM Netezza**), tabulek, datových souborů** (jako jsou soubory IBM SPSS Statistics, SAS a Excel**), textové soubory, zdroje z Web 2.0** (například RSS), **PS Quaestio PRO**, **systémy s IBM Classic Federation server a zDB2 pro z/OS**
- obsahovat nadstavbu v podobě reportovacího nástroje, pomocí kterého budou uživatelé schopni zobrazit základní datové zdroje.
- mít zabezpečený přístup k datům, ochranu dat i výstupů, transparentnost a auditovatelnost změn.
- musí umožňovat integrace s jinými portály.
- musí obsahovat pokročilé analytické metody pro další zpracování datových zdrojů s graficky orientovaným uživatelským prostředím, bez nutnosti znalosti programovacích jazyků.
    - Požadované funkcionality:
        - široká škála interaktivních grafů
        - pavučinový graf pro analýzu vztahů v datech
        - interaktivní výběr dat z grafu pro vizualizaci nebo modelování
        - přístup k procedurám a grafům z programu IBM SPSS Statistics

- přístup k datům z datového repozitáře, IBM DB2®, Oracle®, Microsoft SQLServer™, IBM Informix®, IBM Netezza, mySQL (Oracle), Hadoop Distributed File System, datovým zdrojům Teradata, stejně tak jako k databázím zDB2 a IBM Classic Federation Server Support
- import textových souborů pevné délky nebo s oddělovači, import datových souborů IBM SPSS Statistics, SAS, IBM SPSS Quaestio PRO nebo XML
- paleta nástrojů pro čištění dat od odebrání či nahrazení chybných údajů až po automatické vkládání chybějících hodnot a zmírnění vlivu odlehlých pozorování a extrémních hodnot
- automatické ověření kvality dat a jejich příprava k modelování
- výběr proměnných, přejmenování, odvození nových proměnných, kategorizace, nahrazení hodnot a změna pořadí proměnných
- výběr případů, náhodné výběry, spojení dat a textových řetězců, třídění, agregace a vážení
- restrukturalizace dat, rozdělení na tréninkovou a testovací množinu a transpozice
- funkce pro práci s textovými řetězci: tvorba řetězců, nahrazování znaků, vyhledávání, ořezávání a odebírání mezer
- RFM skórování, agregace transakčních dat pro kompletní RFM analýzu
- export dat do databází, IBM Cognos Business Intelligence, IBM SPSS Statistics, textových dokumentů, Excel, SAS, XML.
- pokročilé data miningové algoritmy
- automatická klasifikace a seskupování
- interaktivní prohlížeč modelů a přehledné statistické výstupy
- vizualizace analytických výsledků na geografických mapách
- grafické zobrazení relativní důležitosti prediktorů pro závislou proměnnou
- kombinace několika modelů (metamodelování), nebo analýza jednoho modelu pomocí druhého
- Component-Level Extension Framework (CLEF) pro tvorbu vlastních aplikací
- přístup k nástrojům jazyka R
- možnost práce v jazyku Python
- propojení s IBM SPSS Statistics
- simulování dat metodou Monte Carlo
- C&RT, C5.0, CHAID & QUEST – rozhodovací a klasifikační stromy s možností interaktivního růstu
- Decision List – interaktivní algoritmus pro vytváření pravidel
- Kohonenovy sítě, metody K-Means a Two Step, diskriminační analýza a metoda podpůrných vektorů (SVM) – seskupovací a segmentační algoritmy

- Faktorová analýza, analýza hlavních komponent – algoritmy pro redukci dimenzionality
- Lineární regrese, zobecněná lineární regrese (GLM) a zobecněné lineární smíšené modely (GLMM) – odhady parametrů v lineárních modelech
- Logistická regrese - modelování kategorizovaných proměnných
- SLRM – bayesovský model s postupným učením
- Analýza časových řad – automatické generování a odhady parametrů časových řad
- Neuronové sítě – vícevrstvá síť se zpětnou propagací, síť s radiální bazickou funkcí
- Podpůrné vektory (SVM) – pokročilý algoritmus vhodný pro rozsáhlé datové soubory
- Bayesovské sítě – modely založené na podmíněné pravděpodobnosti
- Coxova regrese – odhad času do konkrétní události
- Detekce anomálií – nalezne neobvyklé záznamy pomocí seskupovacích algoritmů
- KNN – klasifikace metodou nejbližších sousedů
- Apriori – oblíbený asociační algoritmus s pokročilými funkcemi pro vyhodnocení výsledků
- CARMA – asociační algoritmus s možností vícenásobných důsledků
- Sequence – nalezení asociací v záznamech uspořádaných podle času

Dodavatel v rámci implementace využijte stávající IT a databázovou infrastrukturu VÚBP.

Dodavatel bude zodpovědný za analýzu datových procesů a postupů, specifikaci a popis zdrojů, zajištění dostupnosti a nastavení oprávnění jednotlivých uživatelů/skupin analytického repositáře a přípravy procesů pro jejich automatizaci. Automatizovaným procesem se myslí časové nebo jiné (SOAP) spuštění procesu/ů, které načtou data, zpracují, validují, vyčistí, připraví do požadované struktury a zapíše do příslušných umístění – databáze/soubory.
Protože aktuálně není znám ani přibližný rozsah a obsah uvažovaných datových zdrojů (bude upřesněno v průběhu analýzy) je v tuto chvíli pro zavedení datového repozitáře počítáno se dvěmi datovými zdroji a maximálně 3 reporty pro každý zdroj. Reportem je myšleno max 10 stránkový webový report s využitím technologie v prostředí VÚBP *(PS PORTAL)* nebo 5 stránkový tabulkový Excelovský/PowerQuery report. Obsah reportů bude stanoven v závislosti na datovém zdroji a bude specikován pozdější analýzou, ale předpokládá se základní popisná statistika (četnosti, průměry) a pro vybrané datové zdroje i sofistikovanější metody jako detekce anomálií a extrémních hodnot, vztahy v datech (korelace, regrese), kategorizace dat a souhrny za nově vzniklé kategorie apod.

Pro automatizaci je klíčové a nezbytné verzování analytických aktiv v repozitáři, možnost vytvoření sady automatizovaných úloh pomocí těchto aktiv a vedení auditního logovacího záznamu o jejich (ne)úspěšném spuštění a kompletním zpracování. Musí být dohledatelné, kdy, co a s jakým výsledkem bylo zpracováno. Úlohy musí být automatizovány časově, ale musí

umožňovat i ad hoc spuštění, případně spuštění jiným procesem nebo systémovou událostí (SOAP, změna v datech). S ohledem na ostatní použité technologie v prostředí VÚBP (*PS Quaestio PRO/PS Imago PRO/ PS Clementine PRO*) by měly být automatizované procesy založeny na dataminingové technologii IBM SPSS Modeler a metodologii CRISP-DM.

Činnosti spojené s implementací Datového repozitáře

1. Identifikace uživatelů/skupin/organizací, které budou datově přispívat do repozitáře
    a. Formát dat (csv, txt, dtb – napojení)
    b. Frekvence importu dat (pravidelně, adhoc)
    c. Metadata (význam kategorií, správa číselníků)
2. Softwarové zajištění ETL procesů a databázového úložiště
    a. Příprava ETL a automatizace
    b. Příprava a správa dtb úložiště (databáze, tabulky, práva přístupu, zálohování)
    c. Nastavení prostředí pro adhoc import dat
3. Identifikace uživatelů/skupin/organizace, které budou číst obsah repozitáře
    a. Nastavení přístupů do repozitáře ke konkrétnímu obsahu
    b. Zajištění přístupu z vnitřní/vnější sítě
    c. Příprava dotazů, které bude moci uživatel editovat podle svých potřeb