# Příloha 1 Zadávací dokumentace

# **Služby v oblasti vývoje výpočetního systému pro robustní datovou analýzu založenou na matematické gnostice**

# Specifikace služeb

# Obsah

[Specifikace služeb 1](#_Toc64040879)

[Úvod 1](#_Toc64040880)

[Zadání smluvního výzkumu 1](#_Toc64040881)

[Gnostické metody analýzy dat 1](#_Toc64040882)

[Parametry gnostické metody 1](#_Toc64040883)

[Požadavky pro celý systém pokročilé informační analýzy 1](#_Toc64040884)

[Funkcionalita 1](#_Toc64040885)

[Správa registru formátů 1](#_Toc64040886)

[Import zdrojových dat 1](#_Toc64040887)

[Parametrizace výpočtu a jeho provedení 1](#_Toc64040888)

[Generování reportů 1](#_Toc64040889)

[Správa uživatelů 1](#_Toc64040890)

[Architektura a platforma celého systému, včetně datového skladu 1](#_Toc64040891)

[Klientská část 1](#_Toc64040892)

[Aplikační server 1](#_Toc64040893)

[Hostování aplikačního serveru 1](#_Toc64040894)

[Možnost nasadit službu in-house 1](#_Toc64040895)

[Předpoklady 1](#_Toc64040896)

# Úvod

Předmětem projektu je vyvinout robustní datově-analytický nástroj pro pokročilou informační analýzu založeném na teorii matematické gnostiky. Motivem je vytvoření nového způsobu analýzy hromadných dat pro environmentální, medicínské a jiné vědecko-výzkumné studie, pro které jsou stávající statistické nástroje nepostačující.

Předmětem smluvního výzkumu je vyvinout datové rozhraní mezi gnostickým jádrem a datovým skladem.

# Zadání smluvního výzkumu

Plnění smluvního výzkumu na datové rozhraní gnostického jádra se skládá z následujících dílčích cílů:

1. Vývoj výpočetních programů postavených na bázi matematické gnostiky
2. Optimalizace gnostického jádra s ohledem na požadavky vyvíjeného rozhraní
3. Tvorba datového rozhraní a GUI ve vývojovém prostředí podporující webové platformy
4. Implementace metod pro marginální a multivariantní datové analýzy
5. Realizace příkladových studií, prezentace výstupů pro environmentální a klinická data
6. Závěrečná zpráva

Detailní plnění těchto cílů je také popsáno v příloze 1b) Fáze a milníky této zadávací dokumentace.

# Gnostické metody analýzy dat

Metodika datové analýzy je založena na využití moderních metod matematické gnostiky, které zásadně rozšiřují možnosti standardních statistických metod a umožňují efektivní datovou analýzu i malých vzorků dat. Matematická gnostika vychází z teorie měření a na rozdíl od statistických metod je založena na vlastnostech každé individuální naměřené hodnoty. Vlastnosti datového souboru jsou získány agregací vlastností jednotlivých měření.

Matematická gnostika nepotřebuje apriorní předpoklad o tvaru distribuční funkce. Exploratorní analýza i vlastní datová analýza jsou sloučeny do jediného kroku, distribuční funkci chyb data prozradí sama. Gnostické estimátory vynikají přirozenou robustností, výsledky jsou necitlivé na odlehlá data. Matematická gnostika řeší i úlohy, které statistické metody řeší neúplně, nebo vůbec. V matematické gnostice jsou rozpracovány metody zpracování cenzorovaných dat, což je nesmírně důležité při ochraně životního prostředí (měření pod mezí detekce) a při klinických studiích (pacient přežil testovanou terapeutickou metodu tak dlouho, že např. kvůli přestěhování vypadl z evidence, ale je velká pravděpodobnost, že stále žije).

Matematická gnostika je založená na gnostické teorie neurčitých dat. Existuje několik faktů podporujících výběr této metodiky:

1. Tento přístup je založen na konzistentní axiomatické matematické teorii, která vznikla v průběhu doby z několika vědních oborů (teorie měření, neeuklidovská geometrie, klasická termodynamika, relativistická mechanika, algebře a tenzorové analýze).
2. Gnostická teorie datových vzorků, které se vztahuje na výsledky kompletní teorii jednotlivých nejistot datového balíku. Jako taková je stále použitelná na malé vzorky dat bez nutnosti použití ekvivalentu statistické centrální limitní věty.
3. Matematické modely neurčitosti, entropie, informace a pravděpodobnost nejsou zavedeny jako základní předpoklady. Jsou odvozeny od jediného věrohodného algebraického axiomu související s reálnými daty a podporuje teorii i praktické měření. Způsob měření nejistoty je aplikace Riemannovy geometrie, která je také přímým důsledkem odvozena z axiomu.
4. Optimální povaha gnostického zpracování dat spočívá v maximalizaci výsledné informace jako důsledek variačního teorému.
5. Nelineární metrika měření datových odchylek a neaditivní složení dat dává přístup použitelný dokonce i na velké datové chyby, pro které se centrální limitní teorém nevztahuje. Navíc gnostické znaky konvergují k těm klasickým statistických v případech velmi malé nejistoty. Gnostická teorie je tedy nejen v souladu se statistikou, ale i tak může být interpretována jako rozšíření statistiky v případech velké nejistoty.
6. Od doby prvních vědeckých publikací o přístupu ke statistice (1984) bylo věnováno úsilí vývoji gnostických algoritmů, jejich ověření a aplikace řadě oborů.
7. Všechny dlouhodobé zkušenosti s gnostickými přístupy byly použity ve vývoji gnostického softwarového balíčku platného v rámci open-source R-projektu.
8. Gnostické SW-balíčky obsahují funkce, které umožňují robustní a maximální informační řešení problémů mezí rozsahu normality jak jedno-a multi-dimenzionálních vzorků dat pouze na základě údajů, aniž by byly známy subjektivní apriorní předpoklady statistické povahy. Jedinečnost a objektivnost těchto řešení spolu s maximální dosaženou informací z nich dělá spolehlivé metody.
9. Aktuální rozhodování vyžaduje spolehlivé odhady rizika spolu s odhady pravděpodobnosti pozitivního výsledku z počátečního rozhodnutí. To vše můžeme robustně a objektivně dosáhnout postupy uvedenými v gnostické SW-balíčku.

# Parametry gnostické metody

Gnostická metoda splňuje tyto základní podmínky pro informační analýzu:

* Model použitý pro rozhodování musí objektivně vytáhnout všechny potřebné informace z dat, bez jakýchkoliv zvláštních a priori předpokladů o datovém modelu, protože datové modely se mohou měnit případ od případu.
* Maximální množství informací, které je třeba vytáhnout z údajů, musí být prokázány konzistentní teorií zacházení s daty.
* Meze oblasti pravděpodobných dat musí být odhadnuty z těchto údajů.
* Matematická definice rozsahu běžných údajů jako sub-intervalu pravděpodobného datového intervalu musí být k dispozici a meze intervalu budou odhadnuty z dostupných dat.
* Metoda odhadu hranice pravděpodobných a normální členů jednorozměrných datových vzorků musí být rovněž použitelná pro multidimenzionální vzorky.
* Metoda musí umožnit volbu mezi dvěma vzájemně protilehlými druhy robustnosti zacházení s daty:
* Robustnost s ohledem na odlehlá data (vnější robustnost),
* Robustnost, pokud jde o vnitřní poruchy / šum datového vzorku (vnitřní robustnost).
* Metoda musí být vhodná pro nehomogenní vzorky vytvořená jako směs dvou nebo více dílčích vzorků definována na různé rozmezí normálních dat.
* Metoda, která vyhoví těmto požadavkům, musí být k dispozici ve formě programů, fungujících na běžně dostupných výpočetních prostředcích.

Gnostické analytické metody umožňují analýzu dat podobně, jako statistické metody, dovolují tedy ze vzorku dat určit např. typ rozložení a jeho konkrétní parametry.

Na rozdíl od statistických metod, které z definice předpokládají velké množství vstupních dat, se gnostická metoda dá použít i tam, kde je vzorků pro statistické zpracování relativně málo, protože dokáže vytěžit informace i z dat typu "mimo měřitelný rozsah".

Gnostické metody jsou zatím relativně málo známé, přitom mají velice dobrý potenciál v řadě situací, kdy statistické metody selhávají.

# Požadavky pro celý systém pokročilé informační analýzy

Následně jsou uvedeny požadavky pro konečný produkt, včetně datového skladu, který není součástí tohoto výběrového řízení

## Funkcionalita

Základním předpokladem pro implementace funkčnosti je znalost zákonitosti a praktické realizace robustních algoritmů, typové výstupy a aplikace z rozsahu marginální i vícerozměrné analýzy.

## Správa registru formátů

Systém umožňuje správci dat spravovat sadu specifikací formátů, použitelných pro případné výpočty. Typickým formátem je tabulka, která má sadu polí s konkrétním významem a typy hodnot. Hodnoty mohou být v některých případech nejen číselné, ale také informace o tom, že naměřená hodnota je nad nebo pod mezí měřitelnosti, nebo že není hodnota dostupná vůbec.

### Import zdrojových dat

Systém umožňuje uživateli provést import dat upload svých dat do systému s tím, že přitom určí (poloautomaticky nebo explicitně), jaký konkrétní formát užívají.

Importovaná data mají svá metadata (tagy, čas importu, jméno, formát, kategorie atp.). Tato data budou uložena spolu s importovanými daty do datového skladu, dle uživatelských předvoleb.

### Parametrizace výpočtu a jeho provedení

Systém je navržen tak, umožní uživateli vytvořit zadání výpočtu, které sestává z odkazu na zdrojová data a ze sady dalších výpočetních parametrů.

Uložením zadání výpočtu se následně automaticky spouští samotný výpočet (alternativně lze vyžadovat explicitní spuštění výpočtu uživatelem). Parametry výpočtů pro jednotlivé typy mají být ukládány.

### Generování reportů

Systém umožní uživateli vybrat výsledek nějakého výpočtu, k němu přiřadit požadovaný typ reportu, dovolí zadat potřebné parametry pro report a následně report vygeneruje a uloží.

Systém uživateli umožní i následné prohlížení již hotových reportů.

### Správa uživatelů

Je základem customizace a ukládání předvoleb výpočtů. Postačuje jednoduchá správa.

##

## Architektura a platforma celého systému, včetně datového skladu

Systém je koncipován jako webová aplikace, aby se zajistila co nejširší dostupnost řešení na různých klientských platformách.

### Klientská část

Klienti aplikace budou užívat operační systém MS Windows, MacOS nebo Linux, přesné verze budou upřesněny během vývoje.

Předpokládá se dostupnost aktuální verze webového prohlížeče typu Firefox nebo Google Chrome.

### Aplikační server

Aplikační server může použít libovolný operační systém, protože bude své služby poskytovat formou webové aplikace.

Předpokládaná platforma je buď Linux nebo MS Windows, volba je na dodavateli.

### Hostování aplikačního serveru

Pokud bude použito hostování aplikačního serveru u nějakého poskytovatele internetových služeb, musí být server umístěn v jurisdikci České republiky. Příkladem mohou být např. Active24, Wedos atp.

### Možnost nasadit službu in-house

Aplikační server a související infrastrukturní služby (databáze, uložiště dat atp.) musí umožnit nasazení řešení in-house bez nutnosti užívat služeb třetích stran.

# Předpoklady

## Velikost importovaných dat

Maximální velikost importovaných dat je:

* počet záznamů: 500 tisíc
* počet polí: 10
* max. velikost vstupního souboru v nekomprimované podobě: 20 MB