Technická specifikace

Datová analytika a čistota dat

Obsah dokumentu

1. Úvod

1.1. Cíle projektu

Cílem tohoto projektu je navrhnout a implementovat analytické řešení a postupy pro „analýzu a zpracování kmenových dat“, včetně definice jednotlivých technických kroků pro zapojení a integraci řešení do IS PGRLF.

1.2. Seznam zkratek a pojmů

V dokumentu se vyskytují následující zkratky a termíny:

Zkratka/Pojem	Vysvětlení
CRM	Z anglického jazyka: „customer relationship management“ – systém pro řízení vztahů se zákazníky.
DAIS	Z anglického jazyka: „Data Analysis and Integration System“ – vlastní ETL řešení dodavatele.
DOD	Z anglického jazyka: „Definition of Done“ – seznam výstupů podmiňujících ukončení aktivity.
DOR	Z anglického jazyka: „Definition of Ready“ – seznam požadavků podmiňujících zahájení aktivity.
ETL	Z anglického jazyka: „extract, transform, load“ – popis procesu, který data načítá, transformuje je a následně je zpřístupňuje navazujícím systémům nebo aplikacím.
PGRLF	Podpůrný a garanční rolnický a lesnický fond, a. s.
MDM	Z anglického jazyka: „master data management“ – systém klasifikace dat, proces práce s nimi, datový model a principy kontinuálního řízení a kontroly datové kvality.
PoC	Z anglického jazyka: „proof of concept“ – pilotní fáze projektu, která ověřuje použitelnost postupu vzhledem k požadovaným výsledkům.
RO	Z anglického jazyka: „read only“ – přístup do databáze, který umožňuje pouze čtení/export dat a nikoliv jejich úpravu nebo mazání.
SW	Z anglického jazyka „Software“ – Aplikace
DB	Z anglického jazyka „Database“ – Databáze využívaná pro skladování informací
IS	Z anglického jazyka „Information System“ – Informační systémy zahrnují více aplikací, kterou jsou ve společnosti běžně využívány v rámci agend ke splnění své činnosti.

Tabulka 1: Seznam zkratek a pojmů

1.3. Současný stav a aktivity

Zadavatel, Podpůrný a garanční rolnický a lesnický fond, a. s. (dále jen PGRLF), v rámci obchodních vztahů s klienty udržuje databázi údajů o cca 30.000 subjektech (právnických osobách), včetně cca dvojnásobného počtu údajů o kontaktních osobách jednotlivých subjektů. Kvalita kmenových dat je identifikována jako důležitá a nezbytná pro další rozvoj informačních systémů společnosti a zároveň je klíčová pro digitalizaci procesů ve společnosti.

V rámci informačních systémů PGRLF došlo historicky k nekonzistentnosti kmenových dat jednotlivých systémů, jejich oddělené správě a evidenci. Data v konsolidované podobě vykazují významné množství nekonzistencí v podobě duplicit, neúplných údajů, překlepů, formálních, faktických či datových chyb.

Cílem plnění je provedení celkové revize a optimalizace kmenových dat, jejich sloučení do jediného datového zdroje (DB, nebo datového skladu), integrace existujících systémů na tento zdroj a nastavení kontrolních procesů, které zamezí znehodnocování nebo degradaci kmenových dat v budoucnu – v odborné terminologii shrnutelné jako zavedení Master Data Managementu (MDM). Jde náročný a dlouhodobý projekt, jehož přínosy jsou evidentní.

2. Fáze projektu

Zadavatel požaduje realizaci plnění ve dvou fázích:

Fáze 1:

a) Návrh a implementace ETL řešení včetně připojení požadovaných služeb
b) Integrace ETL řešení na jednotlivé systémy zadavatele
c) Analýza a zpracování kmenových dat
   - Datové profilování,
   - Analýza kmenových dat
   - Zpracování kmenových dat a generování výstupů

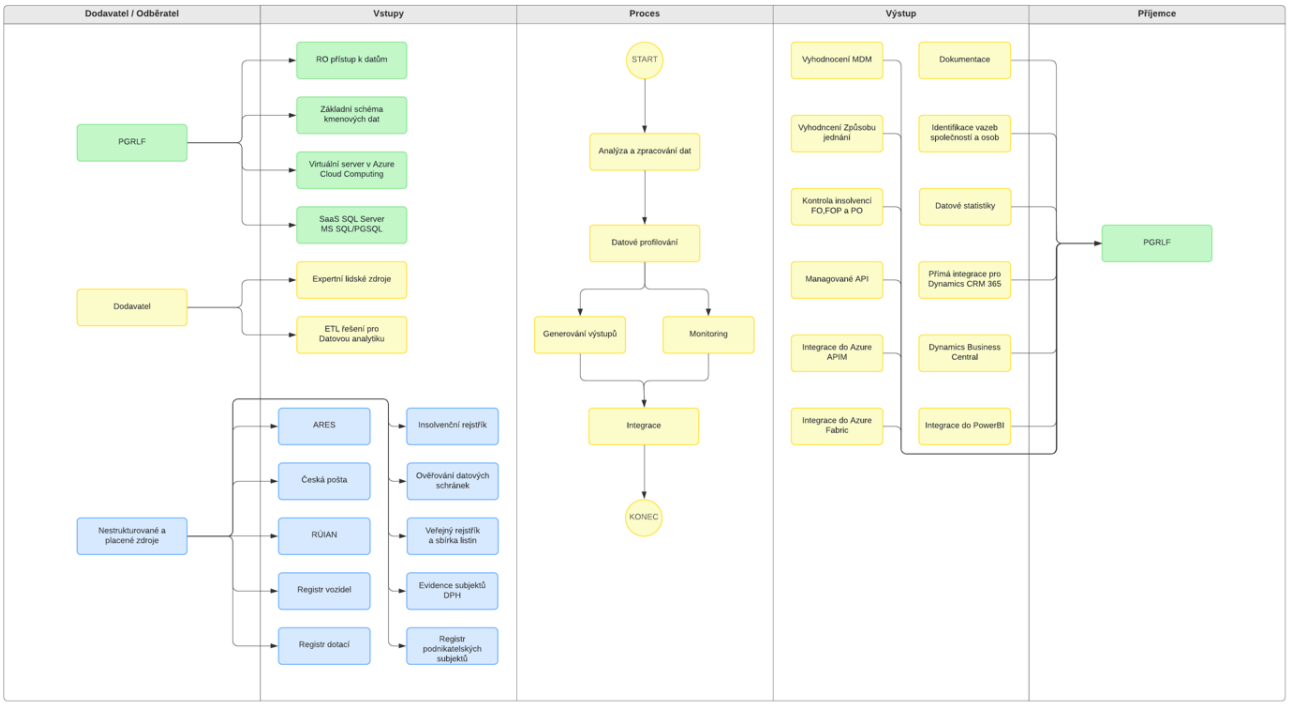
Zadavatel požaduje, aby činnosti dle odst. a) a b) Fáze 1 byly realizovány a výstupy předány **do 1 měsíce** od podpisu smlouvy. Činnosti dle odst. c) Fáze 1 musí být realizovány tak, aby návrh Master Data Management (MDM) řešení s novým datovým modelem a příslušnými kontrolními procesy předal dodavatel k akceptaci **do 2 měsíců** od podpisu smlouvy a došlo k zahájení vlastní analýzy a zpracování kmenových dat. Dále budou již činnosti probíhat průběžně.

Fáze 2:

a) Zajištění datových kontrol a statistik nad kmenovými daty
b) Monitoring změn subjektů v příslušných evidencích a registrech
c) Rozvoj API a integrací dle potřeb zadavatele
d) Provoz a rozšiřování analytických procesů a činností

Činnosti specifikované ve Fázi 2 budou realizovány průběžně dle aktuálních potřeb zadavatele a to od termínu akceptace MDM řešení ve Fázi 1.

Rozdělení odpovědností v rámci realizace plnění znázorňuje následující obrázek:



Obrázek 1: Schéma projektové fáze Datová analytika a čistota dat

Zadavatel předpokládá následující objem plnění, resp. náročnost na zdroje dodavatele:

- Fáze 1: až 280 MD za dobu trvání smlouvy (4 roky) při odhadované náročnosti 4-6 MD/měs.
- Fáze 2: až 80 MD za dobu trvání smlouvy (4 roky) při odhadované náročnosti 1-2 MD/měs.

Z výše uvedených požadavků na plnění musí dodavatel předpokládat, že činnosti Fáze 1 a Fáze 2 budou probíhat v určitou dobu paralelně a tomuto faktu přizpůsobit alokaci zdrojů na své straně.

Z důvodu značných nároků na termíny a zajištění potřebné úrovně bezpečnosti dat požaduje zadavatel realizovat plnění dle odst. a) a b) Fáze 1 v sídle zadavatele.

3. Požadavky na architekturu řešení

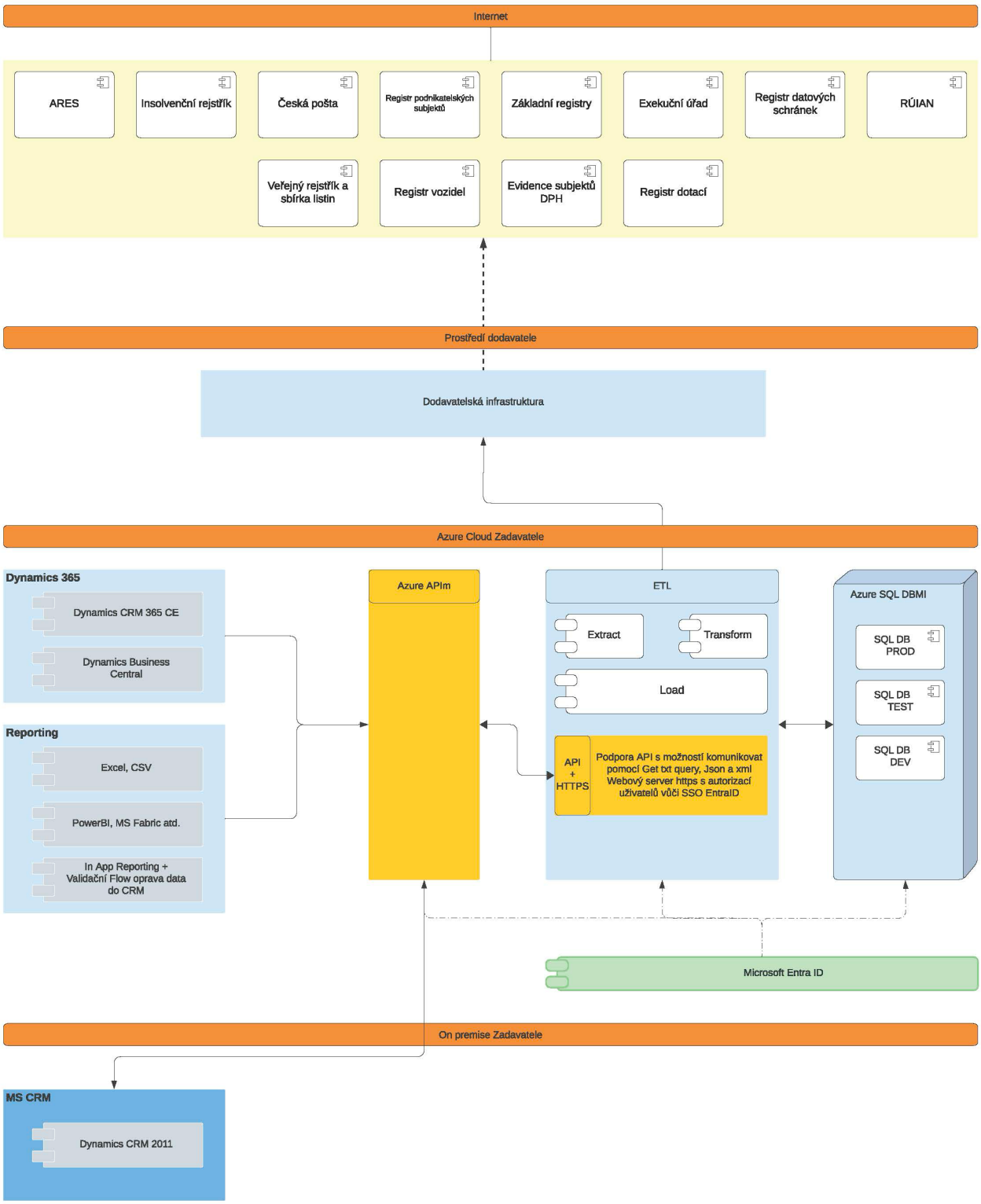
V rámci Fáze 1 bude implementováno řešení pro MDM využitelné jako transportní můstek mezi veřejně dostupnými zdroji a interními systémy PGRLF.

Řešení musí umožňovat zpracovat data ze zdrojových systémů a transformovat je do požadovaného formátu a následně integrovat do potřebných Interních informačních systémů. ETL bude navrženo a integrováno tak, aby bylo rychlé, spolehlivé a snadno konfigurovatelné s maximální využitím standardních aplikací a nástrojů v prostředí Azure Cloud. Musí být schopno provádět pravidelné kontroly kvality dat, identifikovat a opravovat chyby, generovat reporty a pomocí Azure APIM komunikovat s integrovanými aplikacemi.

Základní požadavky na architekturu jsou následující:

- z IS zadavatele nebude vedená přímá komunikace ani dotazy na připojené veřejné registry, dodavatel je povinen veškerá data z těchto registrů kompletovat ve svém prostředí,
- externí komunikace z vytvořeného ETL prostředí bude povolena pouze do prostředí dodavatele
- ETL prostředí nesmí do prostředí dodavatele zasílat žádná data o klientech z IS zadavatele, veškerá práce nad daty zadavatele musí probíhat výhradně ve vytvořeném ETL prostředí,
- ETL prostředí nebude mít oprávnění ke změnám dat v IS zadavatele, pouze bude na základě výsledků analytických algoritmů „předkládat“ návrhy na změny a opravu dat.
- ETL řešení bude plně vytvořeno v prostředí Azure Cloudu na platformě Microsoft produktů, které zadavatel využívá a provozuje.

Zadavatel požaduje respektování následující architektury:



Obrázek 2: Požadovaná architektura řešení

Zadavatel provozuje veškeré systémy ve 3-úrovňovém landscape (DEV, TEST, PROD). Stejným způsobem musí být implementováno i řešení ETL.

V rámci procesu vyhodnocení MDM řešení budou dostupné zejména následující analytické výstupy:

- Jednotný přehled všech registrů, jejichž data jsou k dispozici pro daný subjekt
- Přehled chybných či jinak nekvalitních dat pro subjekt i celkově nad databází

- Vyhodnocení způsobu jednání FOP a PO, použitelné pro další automatizované zpracovaní
- Identifikace vazeb společností a osob, včetně historie
- Kontrola a monitoring insolvencí FO, FOP a PO
- Datové statistiky (pro subjekt i celkovou databázi)

Aktuálnost dat bude záviset na jednotlivých připojených registrech. Data však nesmí být starší **více jak 5 dní**.

Z hlediska integrací musí být využitelná služba Azure APIm zadavatele, ke které bude řešení přistupovat prostřednictvím „managed API". Dodavatel zajistí integraci minimálně na následující systémy a aplikace zadavatele:

- Integrace do Dynamics CRM365 Customer Engagement
- Integrace do Dynamics Business Central
- Integrace do Dynamics CRM 2011 (on-premise, přístupné přes Azure Application Gatteway)
- Integrace do Azure APIM SaaS
- Integrace do Azure Fabric
- Integrace do PowerBI

K řešení MDM dodavatel předloží:

- Přehled procesů a analytických algoritmů MDM
- Technickou dokumentaci řešení a schéma architektury
- Dokumentaci datového modelu a datových vazeb

4. Požadovaný rámec provedení

4.1.1. Analýza a zpracování kmenových dat

Analýza a zpracování kmenových dat bude vedená s cílem provést revizi a optimalizaci kmenových dat ve všech systémech a databázích zadavatele. Z databází je nutné data konsolidovat a očistit, poskytnout statistiku chybovosti a dokumentaci řešení a v neposlední řadě navrhnout Master Data Management (MDM) s novým datovým modelem. Vybraná kmenová data budou profilována, konsolidována, analyzována a čištěna a následně průběžným vyhodnocováním a nastavením monitoringu budou sledována a kontrolována. Nově přidaná data (subjekty) budou přidána, kontrolována a profilována a následně zařazena do průběžného monitoringu a vyhodnocování. Dodavatel bude integrovat další analytické nástroje a potřebné zdroje průběžně po celou dobu trvání smlouvy dle požadavků zadavatele.

Provedení analýzy kmenových dat a jejich následné zpracování předpokládá, že se v rámci zpracování provede očištění, validace, ověření formátu a deduplikace dat, za kterou bude následovat kontrola úplnosti a přesnosti skrze referenční srovnání s veřejně dostupnými rejstříky:

- ARES
- Insolvenční rejstřík
- Česká pošta
- Registr datových schránek
- RÚIAN
- Veřejný rejstřík a sbírka listin
- Registr vozidel
- Evidence subjektů DPH
- Registr dotací
- Registr podnikatelských subjektů

- Základní registry (nepovinné)
- Exekuční úřad (nepovinné)

Pokud jsou identifikovány nedostatky či anomálie, provede se validace v těch případech, ve kterých to je technicky a reálně možné. S ohledem na velmi velké množství záznamů ve zdrojových datech je nutné provádět validaci chyb a odchylek automatizovaně.

Zprovoznění a integrace – DOR
- Kmenová data budou dostupná v rámci jedné centralizované databáze ETL řešení
- Přístup k internetu z ETL na server dodavatele obsahující zdrojová veřejná data.
- Veřejné rejstříky poskytují data v očekávaném formátu a bez omezení.

Zprovoznění a integrace – DOD
- Kmenová data budou připravena k profilování, kategorizaci, analýze a zpracováním očištěna/validována a doplněna
- Dokumentace řešení je zpracována

Analýza a zpracování dat – DOR
- Definice profilovaných činností
- Určení rozsahu profilovaných dat – technická definice
- Definice kategorií a struktury dat

Analýza a zpracování dat – DOD
- Výstup v datových strukturách zadavatele
- Integrace do požadovaných systémů
- Příprava do druhé fáze projektu

4.1.2. Datové profilování

Při datovém profilování (data profiling) se kategorizují data v závislosti na obsahu pro získání nezbytných podkladů pro navazující hodnocení a statistiky. Předpokládá se profilování sloupců, profilování mezi sloupci, profilování napříč tabulkami, zkoumání datové struktury, zkoumání integrity klíčů, vztahy mezi daty a datovými sadami, nebo profilování distribuce vzorů a četností. Data budou v základu kategorizována dle identifikovatelnosti a struktury do následujících 4 kategorií:

- Data identifikovatelná – taková, která lze jednoznačně určit na základě klíčových atributů. Jeden z atributů může být například IČ anebo DIČ.
- Data bez klíče – taková, která nemají žádný unikátní klíč, který by je odlišil od ostatních dat. Například data jen s emailovou adresou, PSČ anebo společnost se špatným názvem.
- Data dohledatelná – taková, která je možné identifikovat pomocí dalších dostupných zdrojů, například pomocí datové integrace, externích databází, nebo manuální validace.
- Data neidentifikovatelná – taková, která se nepodaří identifikovat, je možné dále třídit podle struktury, tedy podle toho, kolik sloupců a řádků obsahují data:
    - Data s úzkou strukturou jsou taková, která mají malý počet dat a vysoký počet řádků. Tato data mohou být vhodná pro statistickou analýzu, ale ne pro podrobné reporty.
    - Data s širokou strukturou jsou taková, která mají velký počet dat a malý počet řádků. Tato data mohou být vhodná pro podrobné reporty.

- Data s možnou identifikací jsou taková, která nemají konzistentní informace, nebo obsahují opravitelné chyby, duplikáty, nebo neúplné informace. Tato data mohou být vhodná pro jakoukoliv analýzu, ale je potřeba je buď opravit anebo vyřadit.

Datové profilování – DOR

- ETL řešení je dostupné a funkční
- Dodavatel má R-O přístup k databázím všech systémů a aplikací zadavatele
- Dodavatel dostal základní schéma a popis kmenových dat

Datové profilování – DOD

- Kmenová data v určených databázích byla profilována a kategorizována
- Existuje dokumentace postupu řešení
- Jsou vstupy pro datové statistiky, integrace a monitoring

4.1.3. Generování výstupů

Generováním výstupů je myšleno zpracování vstupů z předchozích kroků do srozumitelných výstupů pro zadavatele. Na základě zpracovaných datových statistik a zkušeností dojde k návrhu nejvhodnější strategie MDM včetně nového datového modelu. Finalizuje a formalizuje se dokumentace postupu a řešení. Očištěná kmenová data je možno zobrazit několika způsoby:

- Ve formátu pro Excel, který umožňuje snadnou manipulaci s daty, filtrování, třídění a další funkce
- Ve formátu csv, který je jednoduchý, univerzální a vhodný pro velké objemy dat
- K datům je možno přistoupit přes interface nástroje Power BI
- Data lze zobrazit přímo v ETL řešení formou komparativního zobrazení, kdy bude jasně viditelný stav před navrhovanou opravou a po opravě. Chyby budou barevně zvýrazněny.

Generování výstupů – DOR

- Všechny předchozí kroky byly úspěšně ukončeny

Generování výstupů – DOD

- MDM vč. datového modelu
- Dokumentace postupu a řešení
- Datové statistiky
- Očištěná kmenová data

4.1.4. Integrace

Integrace mezi ETL a IS zadavatele bude vždy realizována rozhraním Azure APIm se zajištěním plynulého a bezpečného přenosu dat. API rozhraní dodaného řešení bude splňovat běžné standardy bezpečené komunikace se SSL min. v1.3. Rozhraní musí být schopné komunikovat pomocí formátů JSON, XML a TXT GET Query.

5. Požadavky na projektovou metodika, realizaci a nástroje

Zadavatel požaduje agilní metodu vývoje s využitím prostředí Azure DevOps zadavatele. Veškeré zdrojové kódy zůstávají vlastnictvím zadavatele. Pro použité prostředí je dodavatel povinen zpracovat konfigurační parametry a postup pro deployment a provisioning ve formě IaC (Infrastructure as a Code) s využitím podporovaných platforem pro Azure Cloud (Bicep, Azure Resource Manager, atd.).

Zadavatel předpokládá, že analytické know-how dodavatele, transformované do podoby funkcionality ETL, bude podléhat udělení licenčních oprávnění dodavatele. Tato licenční oprávnění musí být součástí nabídkové ceny.

Primární metodou komunikace mezi zadavatelem a objednatelem bude email a ServiceDesk dodavatele. Zadavatel bude pravidelně informován o stavu plnění písemnou formou v intervalu alespoň 1x měsíčně.