

Příloha č. 1 – Specifikace a rozsah předmětu plnění

1. Předmět plnění	2
2. Účel plnění	2
3. Návrh architektury řešení	3
3.1. <i>Stage</i>	3
3.2. <i>Datový sklad</i>	4
3.3. <i>Analytické a reportingové aplikace</i>	4
4. Fyzická architektura a technické parametry	5
4.1. <i>Přidružené aplikace</i>	5
5. Uživatelé datového skladu	5
6. Obecné vlastnosti komponent datového skladu	6
6.1 <i>Stage a řízení verzí dat</i>	6
6.2 <i>Evidence uživatelů a jejich rolí</i>	6
6.3. <i>Přístupová práva</i>	6
6.4. <i>Statistické výstupy</i>	6
7. Další vlastnosti	6
Příloha – upřesnění rozsahu dat	7

1. Předmět plnění

Objednatel požaduje po dodavateli implementaci a zprovoznění datového skladu Univerzity Karlovy, vytvoření portálu se statickými výstupy a on-premise periodickým ukládáním dat s možností generování předdefinovaných reportů.

2. Účel plnění

Implementace Business Intelligence řešení – datový sklad, reporting a prostředí pro jednoduchou analýzu dat, (dále jen „Datový sklad“) má sloužit zejména pro:

- Nahrání dat (jejich bližší specifikace je součástí Přílohy tohoto dokumentu – upřesnění rozsahu dat) do Datového skladu ze anonymizovaných dat ze SIS Univerzity Karlovy v rozsahu:
 - Studium
 - Mobility
 - Uchazeči
 - Poplatky a stipendium.
- Nahrání dat z jiných zdrojů – csv, případně jiný standardně zpracovatelný formát.
- Pročištění a úpravy dat k lepšímu chápání jejich významu.
- Zpracování statistických a jiných výkazů a přehledů z dat dodaných do Datového skladu.
- Tvorbu ad hoc reportů.

Datový sklad bude podporovat automatizovanou aktualizaci dat z jednotlivých zdrojů na denní bázi. Datový sklad bude provázán jak na interní databáze Univerzity Karlovy, tak v případě potřeby na externí veřejně přístupné databáze.

3. Návrh architektury řešení

Současná IT architektura zadavatele jakož i jeho personální kompetence jsou postavena na řešeních cloudových služeb Microsoft Azure. Zadavatel proto poptává řešení kompatibilní s tímto prostředím, kdy součástí dodávky není zajištění samotného provozu cloudových služeb Microsoft Azure. Zadavatel požaduje po dodavateli služeb datového skladu zajištění provozu, monitorování, zabezpečení, zálohování a další administrativní úkony spojené se správou serveru, databáze pro datový sklad a Business Intelligence aplikací.

Zadavatel předpokládá dodávku řešení složené ze dvou částí. První reprezentuje cloudovou část, kde se celé řešení napočítává v Azure cloudu včetně Power BI Reports. Druhá část je on-premise řešení, kde by se na periodické bázi nahrávaly snímky z výstupních dat z DWH cloudu na úložiště UK. A zde má fungovat i SQL Server Report Server, které umožní předdefinované (i další) statické výstupy s možností exportu. Druhá část představuje i referenční data použitá v daném období.

Celé cloudové řešení by bylo orchestrované ETL nástrojem Azure Data Factory.

Data budou na denní bázi extrahována do Stage vrstvy datového skladu, a to jedním z možných způsobů:

- Přímým přístupem do zdrojové databáze
- Přístupem do repliky zdrojové databáze
- Prostřednictvím extraktů dat ze zdrojových databází (formát csv, případně jiný standardně zpracovatelný formát)
- REST konektorem v případě dat dostupných pouze pomocí API.

Takto získaná data budou transformována do denormalizované vrstvy dimenzionálního datového skladu a optimalizována pro reporting a analýzu. V rámci transformačních procedur proběhne čištění dat, historizace a další potřebné procesy.

K vrstvě datového skladu bude možné připojit minimálně pomocí Power BI. Zadavatel požaduje možnost přístupu i dalších standardních analytických a reportingových aplikací, kterou má nebo bude mít UK k dispozici.

Pro on-premise část je určen dedikovaný server (infrastruktura) na UK, s on-premise SQL Serverem. Cílem je periodicky dle potřeby zadavatele (např. měsíc, kvartál) předávat vlastní data z cloudu do datových struktur on-premise. Primárním cílem je mít vlastní referenční data i možnost nad definovanou datovou strukturou nahlížet webovou aplikací SQL Server Reporting Services (součástí licenci SQL Server Standard), kde jsou dostupné statické html reporty (možnost zvolení vstupních parametrů). Přístupy se dají řídit standardně přes active directory.

3.1. Stage

V prostoru Stage budou uchovány snímky přenesených dat včetně dohodnuté historie. Snímky budou přesnou kopií dat ze zdrojových systémů včetně jmenných konvencí. Ve Stage budou vytvořeny také jakékoliv pomocné tabulky potřebné pro efektivní transformaci dat do reportingové vrstvy.

3.2. Datový sklad

V datovém skladu budou data ze Stage transformována do dimenzionální struktury, umožňující jednoduchou analýzu a reporting. Data v dimenzionální struktuře budou organizována ve formě:

- Dimenzí – tabulek s daty, podle kterých jsou analyzována čísla
- Faktů – tabulek s čísly, která jsou analyzována
- Můstků – tabulek obsahující složitější N:N vazby mezi dimenzemi a fakty, pokud takové vazby existují
- Agregačních tabulek obsahujících předpočítané hodnoty pro zrychlení reportingu a analýzy dat.

Datový sklad bude podporovat:

- Historizační procesy.
- Vytváření vlastních uživatelských reportů a analýz. Veškeré datové struktury budou pojmenovány a zdokumentovány v potřebné míře detailu.
- Všechny komponenty datového skladu budou maximálně optimalizovány pro dosažení optimální výkonnosti a snížení zátěže zdrojových systémů.
- V průběhu načítání dat budou vybraná data anonymizována s možností následného propojení s původními daty.
- Uživatelé budou k datovému skladu přistupovat primárně prostřednictvím BI aplikací, které obsahují vlastní integrované zabezpečení. Přímý přístup uživatelů bude probíhat přes zabezpečenou vrstvu databázových pohledů.
- Celý proces vývoje bude řízen metadatovým modelem, z něhož bude možné generovat kód na vytváření jednotlivých databázových objektů. Řešení bude umožňovat také automatické generování ETL pipeline a datasetů.

3.3. Analytické a reportingové aplikace

1. Portál se statickými reporty

Vydefinované statické reporty dle požadavků Zadavatele a jejich možná budoucí modifikace.

2. PowerBI

Pro klíčové uživatele Zadavatele (max. 5 osob). Možnost vytváření vlastních reportů a analýz nad daty z datového skladu včetně možnosti napojení externích dat pomocí Power BI.

4. Fyzická architektura a technické parametry

Na Univerzitě Karlově je provozováno MS Azure, cloudová platforma pro datový sklad. Platby za využívání Microsoft Azure nejsou součástí tohoto výběrového řízení. Z hlediska provozu si tyto služby zajišťuje Univerzita Karlova sama. MS Azure je pro autentizované uživatele přístupný výhradně zabezpečeným protokolem HTTPS.

Pro účely vývoje, testování a nasazování na produkční prostředí Univerzity Karlovy, bude vytvořeno vývojové prostředí, které bude mít podobnou strukturu jako prostředí produkční. Vývojové prostředí se může lišit v konfiguraci a výkonu. Toto prostředí bude primárně sloužit k testování nových funkcionalit a k prvotní kontrole transformovaných dat z databází Univerzity Karlovy. Po odsouhlasení provedených úprav zadavatelem nasadí dodavatel úpravy do provozního prostředí.

Z pohledu architektury se každé z těchto prostředí bude skládat z jednoho Azure tenantu.

4.1. Přidružené aplikace

K datovému skladu bude možné připojit jakoukoliv standardní analytickou aplikaci. Primárně budou uživatelé používat aplikaci Power BI, zajištění přístupu a licence si spravuje Univerzita Karlova sama.

Pro práci s datovým skladem mohou pověřené uživatelé využít například nástroj SQL Server Management studio. Zajištění přístupu a licence uživateli poskytne Univerzita Karlova sama, a proto není součástí dodávky.

Pro zajištění přístupu k datovému skladu prostřednictvím externí aplikace Power BI nebo SQL Server Management studia bude vyžadováno připojení přes VPN Zadavatele a v samotné aplikaci bude potřeba být přihlášen přes CAS nebo LDAP.

Pro potřebu on-premise řešení zadavatel má již potřebný HW i SW pro SQL server k dispozici.

5. Uživatelé datového skladu

Zadavatel požaduje přístup do portálu se statickými reporty a přístup pomocí Power BI (nebo jiných standardizovaných analytických nástrojů pro max. 5 uživatelů).

Z hlediska přístupových rolí Zadavatel požaduje tyto základní kategorie:

- pověřené osoby s oprávněním přistupovat k již vytvořeným reportům
- pověřené osoby s oprávněním (licencí reportingové aplikace) vytvářet reporty z dat v datovém skladu
- systémový administrátor – osoba provádějící nastavení klíčových parametrů a číselníků, přidělování uživatelských licencí, zajištění přístupů přes firewall ke zdrojovým databázím

Zadavatel dále požaduje možnost budoucí změny v podobě vyšší škálovatelnosti.

6. Obecné vlastnosti komponent datového skladu

6.1 Stage a řízení verzí dat

Datový sklad musí spravovat historická data a různé verze datových sad. Tato funkce umožňuje uživatelům sledovat vývoj dat v čase a analyzovat historické trendy. Stage bude kromě výše uvedených dat obsahovat historické verze dat. Extrakty ze zdrojových systémů budou ve Stage vrstvě ukládány formou snímků. Dohodnutý počet denních snímků bude po určitou dobu archivován a následně zálohován.

6.2 Evidence uživatelů a jejich rolí

Zadavatel požaduje k autentizaci uživatelů k přístupu do statického portálu využití jeho LDAP serveru. Dílčí oprávnění pro přístup pomocí Power BI nebo dalších analytických nástrojů bude řešeno na úrovni administrátora.

6.3. Přístupová práva

Přístupová práva k záznamům v databázi datového skladu se budou lišit podle přístupové role uživatele, budou záviset na fakultě/součásti a pracovišti uživatele, a na typu záznamu.

6.4. Statistické výstupy

Datový sklad umožní definovat a sestavovat statistické přehledy. Ty bude možné sestavovat zejména v členění podle časových hledisek, podle fakult/součástí nebo jiného organizačního členění UK či kritéria nahraného do datového skladu.

Sestavování těchto statistik bude přístupné v podobě reportů v analytické a reportingové aplikaci. Tyto reporty se mohou takto zpřístupnit vedení univerzity, vedení fakult/součástí nebo i anonymnímu uživateli.

7. Další vlastnosti

- Multidimenzionální pohled na data.
- Možnost uzamčení datového skladu pro uživatele, např. pro potřeby údržby;
- Možnost zakázat přihlášení vybraných uživatelů a naopak umožnit přihlášení vybraných uživatelů k celému datovému skladu nebo k jeho části;
- Zajištění plné auditovatelnosti. Celý transformační a čistící proces (ETL) bude plně auditován, hlavně na úrovni úspěšnosti dokončení ETL procedur, případných chybových hlášení, počtu vložených, změněných nebo smazaných záznamů nebo času zpracování. Provoz datového skladu bude dimenzován pro současný provoz minimálně 50 uživatelů.

Příloha – upřesnění rozsahu dat

Součástí popisu je zpracování dat studia a jejich nasazení do následujících oblastí datového skladu, kde hlavním zdrojem jsou data ze studijního systému. Přehledy je nutné brát jako orientační, může dojít k rozšíření či úpravě datového schématu. Modely datových hvězd viz Obrázky 1-4.

1) Studium

Základní datový model pro reporty ze studijní oblasti zachycuje Obr. 1. Ve středu pomyslné reportingové hvězdy je faktová tabulka studium s cizími klíči odkazujícími do dimenzí studovaného oboru, studijního programu, pracoviště garantujícího výuku, osoby studující dané studium a časové dimenze reprezentované tabulkou Kalendar. Pro účely snazšího reportingu počtů studií a absolventů jsou dále do faktové tabulky zařazeny atributy charakterizující dané studium jako je Pocet_Studujících, Pocet_absolventu, Pocet_Prerusujících apod. S posledním jmenovaným pak souvisí můstek Preruseni_Ve_Studiu, jenž umožňuje připojit detailní informaci o přerušení, zda se jedná o řádné přerušení, či přerušení z důvodu rodičovské dovolené, od kdy/do kdy bylo studium přerušeno apod. Za zmínku stojí i bridge (=můstek) v podobě tabulky Mobility_Po_Studiích, který odkazuje do faktové tabulky mobilit (viz níže). Konečně specifikem v levé části schématu je obecné namodelování víceoborů v rámci stejného programu a naopak, které je možno řešit i alternativními přístupy, pokud varianta vazeb M : N nenastane.

2) Uchazeči

Jedná se o základní datový model pro přijímací řízení, struktura je principiálně podobná hvězdě studia. Za zmínku stojí zejména můstek Prijimaci_Rizeni_Po_Studiu, které propojuje přihlášku uchazeče s následným studiem, pokud uchazeč skutečně nastoupil a zapsal se na studium, na které se hlásil.

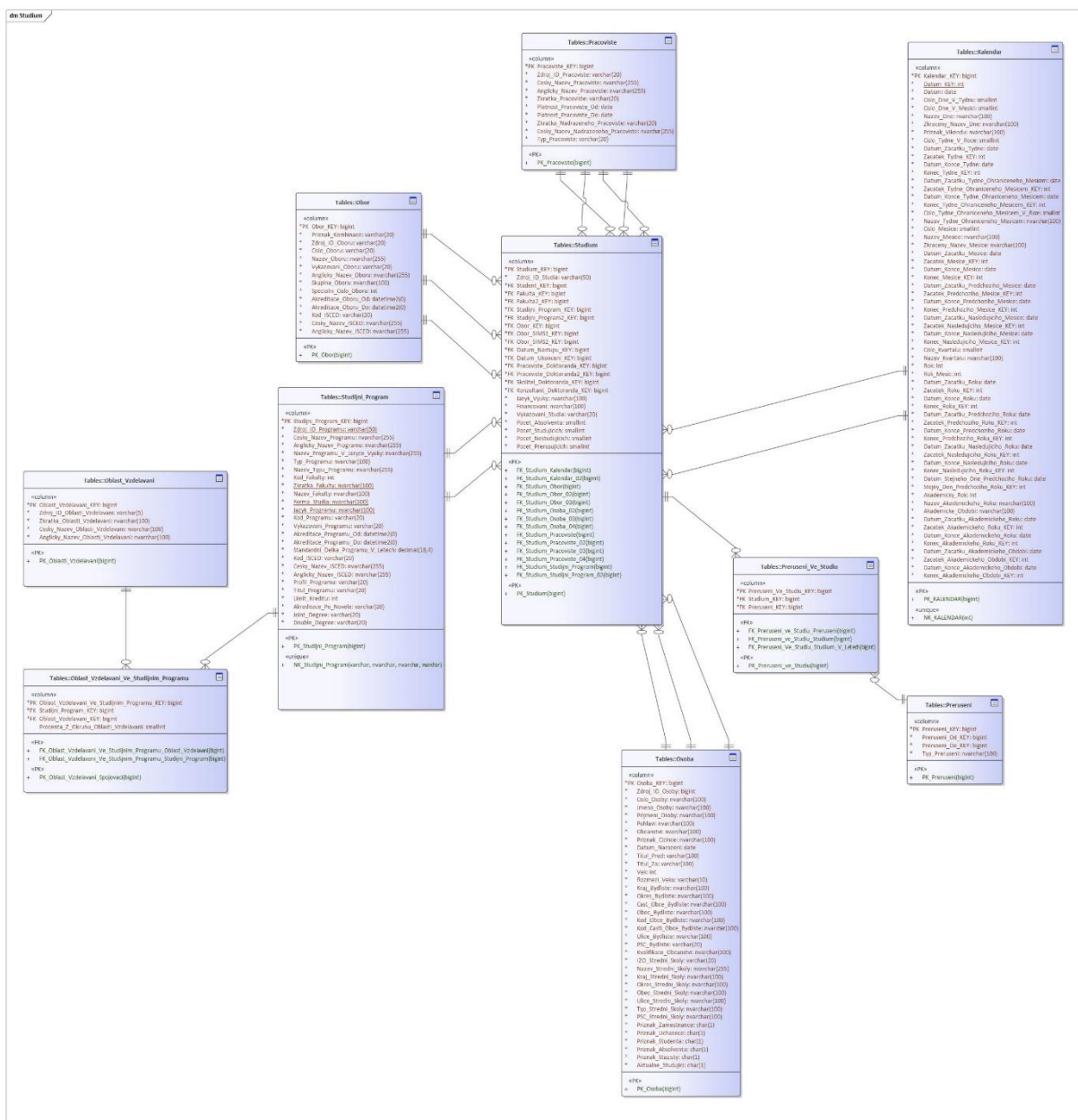
3) Mobilita

Základní datový model pro reporting studentských mobilit. Přes tabulku Mobility_Po_Studiích ji je možno napojit na hvězdu Studium. Ve faktové tabulce se pouze sleduje délka výjezdu a jinak odkazuje do dimenzí studovaného studijního programu, organizace přijímající/vysílající studenta, osoby studenta a konečně časové dimenze.

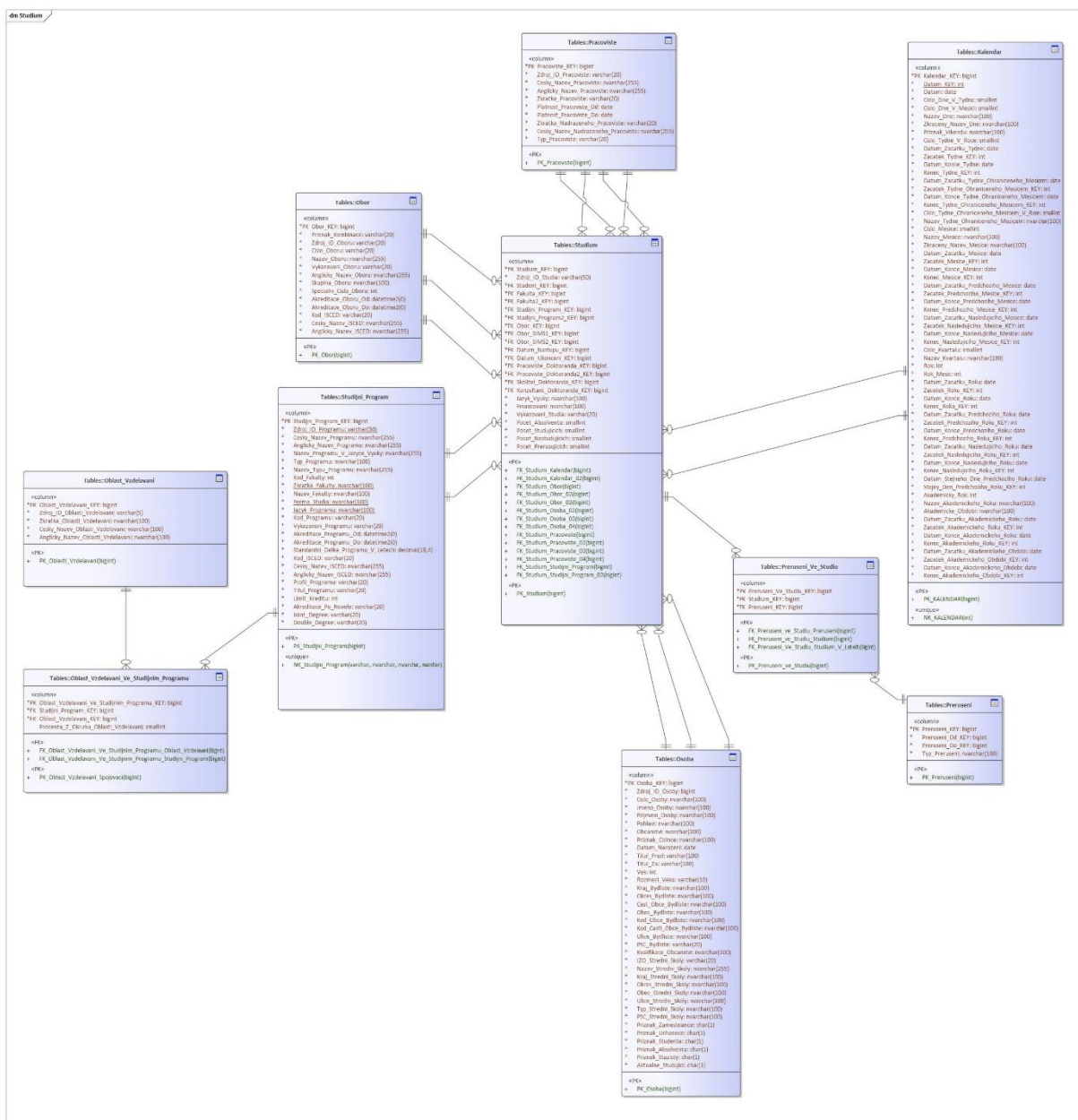
4) Poplatky a stipendia

Datová hvězda umožňující analyzovat poplatky přijaté od studentů a stipendia jim vyplacené. Je možné zkoumat tuto oblast přes různé typy poplatků a stipendií, jejich výše a vývoji v čase.

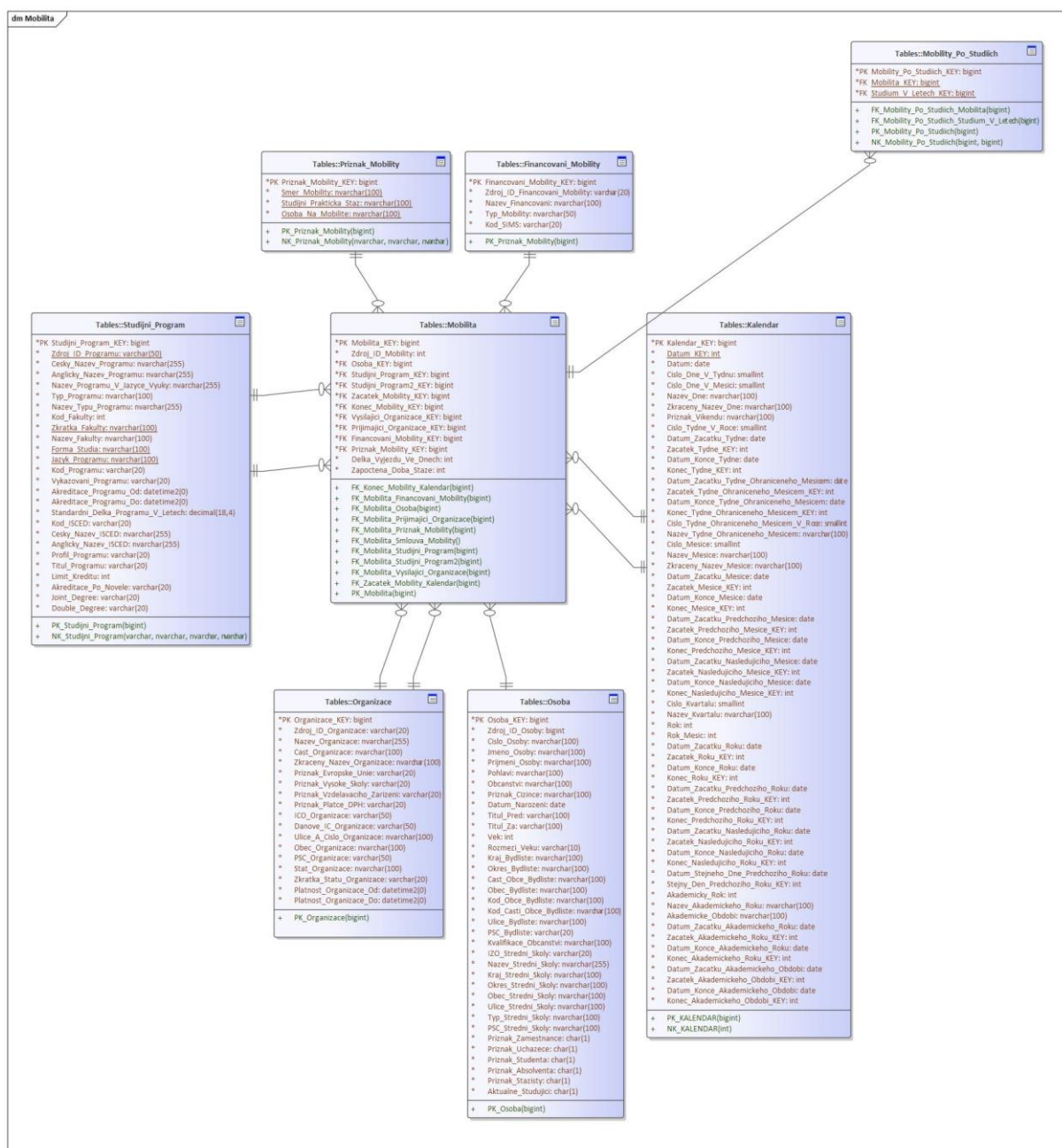
8



9



Obrázek č. 3 - Mobility (datová hvězda)



Obrázek č. 4 - Poplatky a stipendia (datová hvězda)

