V2

Vytvořený nástroj pro sjednocení dat na úrovni druhé vrstvy datového skladu, včetně metodiky pro její dlouhodobé udržování činnosti: Vytvoření druhé vrstvy DW, vybrané typy ukazatelů (jednoduché, složené) pro hodnocení studijních programů (data o uchazečích, studentech, absolventech), analýzy objektů a jejich vlastností ve studijních informačních systémech.

O výstupu:

Jedná se o nástroj, který vychází ze zkušenosti s budováním datových skladů na několika VŠ a shrnuje možnosti sjednocení dat na úrovni druhé vrstvy datového skladu, analyzuje používané typy ukazatelů a zhodnocuje jejich využitelnost. Zkušenost s různými datovými zdroji, informačními systémy i přístupy k budování datového skladu je transformována ve sdílenou sadu případových studií, utvářejících jednotný a z praxe vycházející metodický materiál.

Cíle výstu:

Podpořit vytváření objektivní datové podpory pro získávání dat o hodnotě ukazatelů relevantních pro hodnocení kvality vzdělávací činnosti s ohledem na potřeby a požadavky různých zainteresovaných stran v procesu hodnocení, s důrazem na neúspěšnost studia.

Aktivity vázané k výstupu:

- Vytváření druhé vrstvy dat v datovém skladu – vybrané různé typy ukazatelů (jednoduché, složené) pro hodnocení studijních programů (data o uchazečích, studentech a absolventech), analýzy objektů a jejich vlastností ve studijních informačních systémech.
- Sdílení informací a postupů mezi zúčastněnými školami, reflexe zkušeností, sumarizace metodiky.

Obsah

Případová studie 1: Datový sklad na Západočeské univerzitě v Plzni

Pro hodnocení ukazatelů kvality studijních programů je využíván datový sklad (= Datawarehouse, dále jen DW), který centralizuje data s provozních systémů Západočeské univerzity v Plzni (dále jen ZČU).

Těmito zdroji jsou:

- IS STAG – je primárním provozním systémem pro hodnocení, tedy informační systém určený pro administraci studijní agendy ZČU. Pokrývá funkce od přijímacího řízení až po vydání diplomu. Eviduje studenty prezenční i kombinované formy studia, studenty celoživotního vzdělávání i účastníky univerzity třetího věku.
- EIS MAGION – je ekonomický informační systém, který zpracovává ekonomickou evidenci včetně evidence majetku a personální i mzdovou agendu. Je tedy zdrojem informací o personálním zabezpečení ZČU.
- INIS – je integrovaný informační systém ZČU, který slučuje data z oblasti internacionalizace a externího pedagogického působení. Je to systém umožňující zaměstnancům ZČU získávat informace ze všech výše zmíněných oblastí podle přidělených přístupových práv, doplňovat a aktualizovat evidenci svých aktivity zejména v oblasti pedagogického a vědeckovýzkumného působení.
- OBD + GaP – osobní bibliografická databáze a Granty a projekty jsou aplikace, ve kterých se evidují a spravují data o publikační a grantové činnosti, což jsou nástroje pro správu dat výsledků vědecké a výzkumné činnosti dle specifikace IS VaVaI.
- Externí zdroje – pro kompletní zhodnocení kvality studijních programů byly připojeny další externí zdroje poskytující relevantní informace ve vztahu ke studijním programům (data ze SIMS, MPSV, data získaná metodou focus groups apod.).

V DW jsou dle vhodných vzájemných vazeb vytvářeny datové kostky, které poskytují širší pohled na celou sféru souvislostí mezi získávanými daty a umožňují tak vytvářet komplexní analýzy a vyhledávat všemožné korelace napříč všemi informačními systémy.

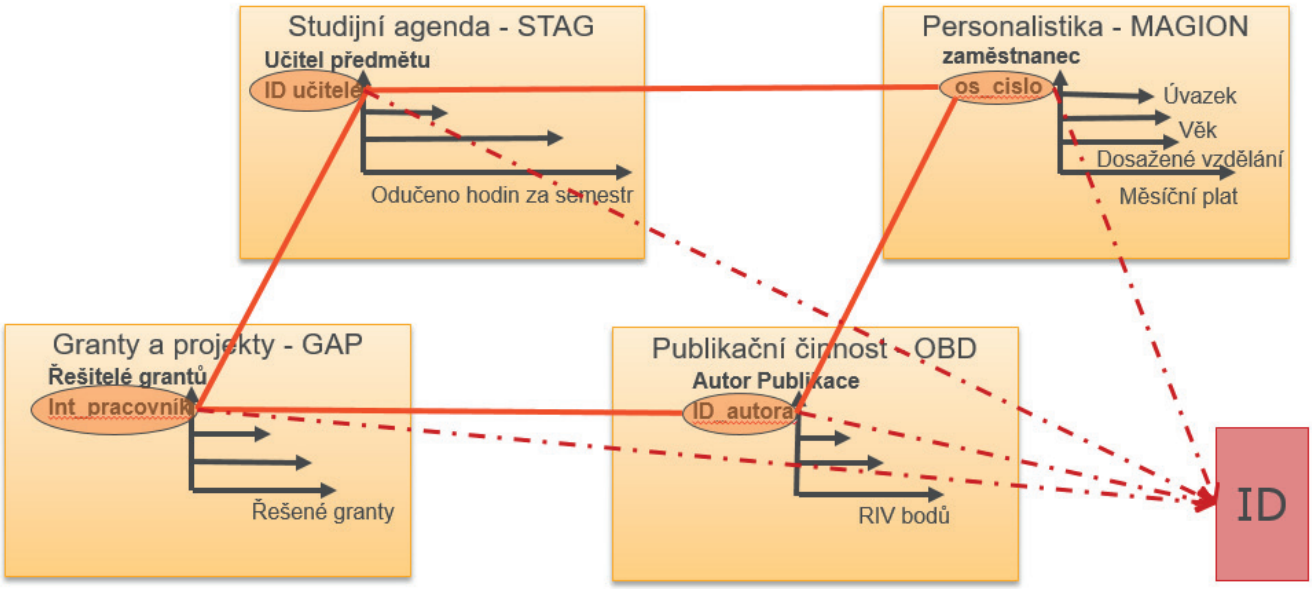
Po konzultacích a sdílení informací s dalšími vysokými školami, především s UPa, prošel DW velkou restrukturalizací nejen z hlediska datového, ale i z hlediska funkčního.

DW je nyní rozdělen na dvě databáze – vývojovou a produkční.

Produkční databáze (DW_PROD) je členěná na jednotlivá schémata dle relevance s definovaným přístupem k datům. Je rozdělena na tři části. První jsou datové kostky obsahující zpracovaná data a základní tabulky včetně jejich historizace (= časové řezy). Tyto dvě oblasti jsou určeny především pro administrátory a IT analytiky. Poslední část je část pohledová. Ta obsahuje předem připravené výstupy pro definované koncové uživatele. Součásti (fakulty a ústavy) mají přístupy implicitně, ostatní uživatelé jsou postupně doplňováni dle pokynů prorektorky pro studium. K tomuto je vytvářena směrnice. K přidávání uživatelů bude vedena dokumentace uložena v DW, zatím nelze automatizovat.

Všechna metadata jsou uložena ve **vývojové databázi** (DW_DEV) a přístupy k ní jsou omezeny pouze na administrátory DW.

Dalším významným krokem, který vzešel ze spolupráce s ostatními univerzitami je úprava vlastních datových kostek **dle relevance** (logických rámců) nikoliv vazby přímo na výstupní ukazatele. Což znamená, že data nejsou navázána na výstupy, ale kostky jsou tvořeny tak, aby byla jejich provazba schopná komplexně pokrýt veškeré uživatelské dotazy. Tyto kostky obsahují co nejvíce atributů vztahujících se k vybrané oblasti. Například datová kostka vyučujícího obsahuje data ze studijní agendy, ale je v ní uloženo i osobní číslo z personalistiky pro snadnější propojení. Toto bude v budoucnu ošetřeno jednotným ID (vnitřním číselníkem osob přímo pro DW), kde jedna osoba bude mít jedinečné číslo.



Příklad propojení dle osobních údajů

Konkrétně k postupu: Data jsou pomocí nástroje Oracle Fussion Middleware – Data Integrator sloučena dle souvislostí skrz různé databázové zdroje a rozčleněna na oblasti se stejným základem pro zjednodušení dalších analytických operací.

Oblasti analýz

Tři základní části produkční databáze jsou dále členěné dle oblastí:

Část studenti

- informace o osobě (adresa, pohlaví, původ,…),
- informace o studiu (předměty, obory, programy, plány,…),
- informace o mobilitách (příjezdy, výjezdy)

Část zaměstnanci

- osoby, vzdělání, mobility, úvazky, pracoviště, rozvrhové akce

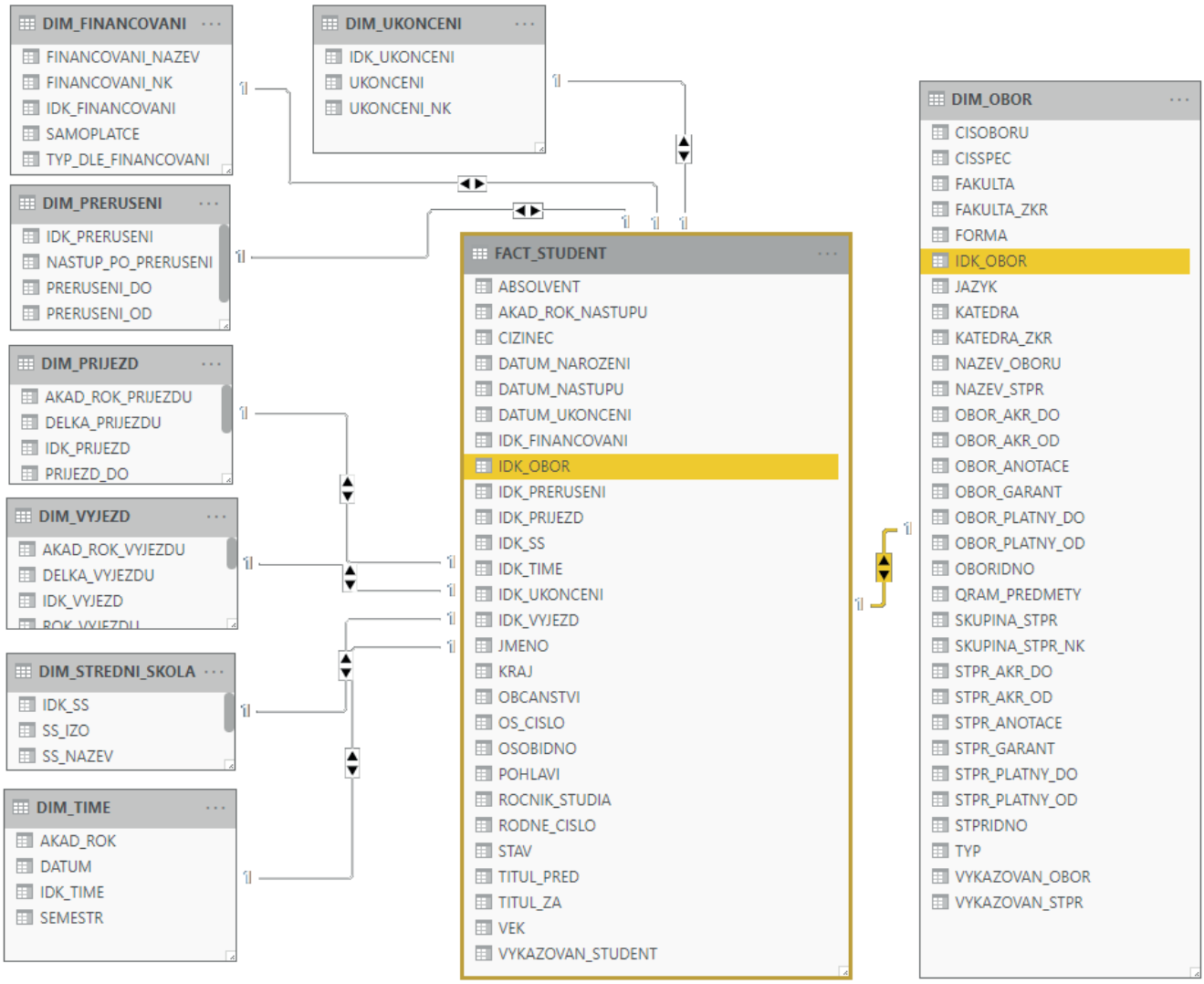
Část publikací a grantů

- informace o publikačních činnostech, zapojení do projektů, řešitelé, spolupráce na výzkumu apod.

Datové kostky

Speciálně pro hodnocení kvality studijních programů byly vytvořeny, resp. uzpůsobeny, tyto datové kostky – abecedně:

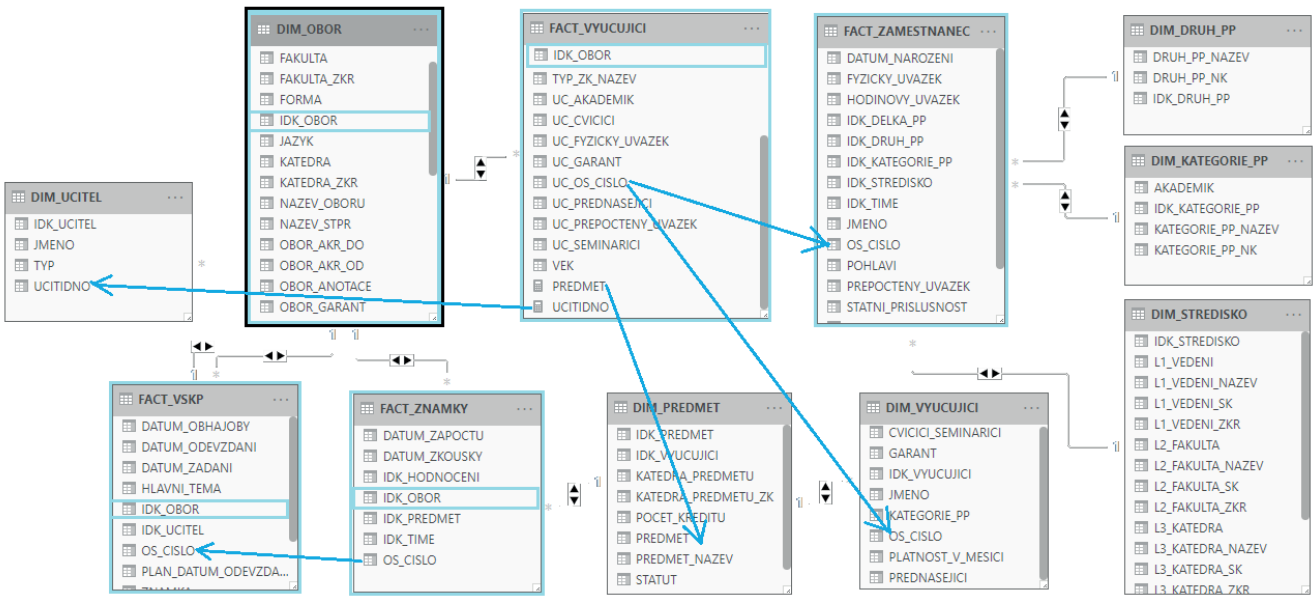
- Granty – zachycují řešitelskou činnost a zatížení učitele
- Publikace – publikační činnost učitele
- Studenti – nejobsáhlejší kostka evidující informace o studentovi (osoby, studium, mobility,…)
- Uchazeči – obsahuje veškeré informace o uchazeči
- VŠKP – zachycuje informace o závěrečných pracích a obhajobách studenta
- Vyučující – datová kostka zaměřená na informace o zaměstnancích z hlediska výuky (vyučované rozvrhové akce, garantování předmětu)
- Zaměstnanci – datová kostka zaměřená na informace o zaměstnancích z hlediska personálního (pracovní zařazení, úvazky,…)
- Známky – eviduje data o zapsaných předmětech a výsledcích učení studenta
- Všechny datové kostky mají identické dimenze času pro sledování vývoje studia v různých akademických rocích po semestru (nejmenší jednotka členění je jeden kalendářní měsíc). Jednotná je dimenze Oboru, její složení je znázorněno na ERA modelu datové kostky Studenta, kdy studijní program má zkratku STPR



ERA model datové kostky studentů

Datové kostky jsou uzpůsobeny tak, aby se dali jednoduše **propojit mezi sebou dle určené závislosti**. Na obrázku číslo 3 jsou znázorněny možnosti propojení několika datových kostek týkající se výuky. Kostky lze propojovat různými způsoby dle ukazatele, který sledujeme. Například pro zjištění zatížení akademického pracovníka v pozici lektor potřebujeme jeho personálie a zároveň napojení na obor a vyučované či garantované předměty. Již z tohoto náhledu je jasné, že je nejprve třeba na manažerské úrovni vytvořit několik zásadních rozhodnutí. Například jaký obor je pro nás primární – obory vázané na předměty a jejich studijní plány, na studenty nebo na vyučující?

Spojení datových kostek dle vybraných ukazatelů je prováděno, je-li třeba, na databázové úrovni vytváření view nebo materilized view nebo přímo v bussines inteligence nástroji power BI.



Propojení IS STAG a MAGION pro zjištění zajištění výuky

Typy používaných ukazatelů

Ve spolupráci s managementem univerzity při tvoření celkové koncepce výstupních sestav byly vytvořeny skupiny ukazatelů, jež lze rozdělit na několik typů (jejichž klasifikace byla popsána v V1). Vzhledem k tomu, že ukazatele mají být využívány pro analytické a rozhodovací procesy jako podklad pro zvyšování kvality studijních programů, je důležité sledovat především jejich dynamické chování. Většina ukazatelů tedy odpovídá typu C. Vhodnost použití poměrového nebo početního výstupu závisí vždy na daném ukazateli. Kde je to možné, je standardně vyžadován a využíván poměrový ukazatel (indikátor). Jak již ovšem bylo zmíněno, někdy je využití poměrového ukazatele zcela nevhodné – jeho vypovídající hodnota je téměř nulová. Ideálně pro strategická rozhodnutí využíváme pro přehled oba typy, jak jednoduchý počet, tak poměr.

Tyto dříve definované typy spadají do oblasti jednoduchých ukazatelů. Při definování ukazatelů se ovšem ukázala potřeba sledovat jevy ještě komplexněji – neboli vytvořit ukazatele složené. Složený ukazatel, jak již název napovídá je složen z několika jednoduchých ukazatelů, jejichž složením zvýšíme jejich vypovídající schopnost. Složený ukazatel již zpravidla většinou nebývá zahrnut ve výstupních sestavách interních univerzitních systémů a je potřeba jej vytvořit dotazováním, kde jinde, než na úrovni datového skladu.

Přibližme si jejich definici na příkladu datové kostky. Datová kostka se skládá z faktové tabulky, která vyjadřuje výsledek dotazu. Na ni jsou připojeny různé dimenze, kterými definujeme výběr (řekněme, že nastavujeme filtr) :

Jednoduché ukazatele (indikátory):

- pro vyjádření ukazatele využíváme vždy maximálně jeden výběr z dané oblasti (dimenze)
- výhody: lze je snadněji získat ze základních sestav interních univerzitních systémů
- příklad: přehled studentů na oboru, časová řada poměru přijatých a zapsaných uchazečů o studium na ZČU dle fakult

Year	POČET
2012	**3**
Kanada	1
Španělsko	2
2013	**12**
Belgické království	1
Chorvatská republika	1
Portugalská republika	1
Řecká republika	1
Rumunsko	1
Slovenská republika	1
Španělsko	3
Spojené státy americké	3
2014	**40**
Belgické království	3
Chorvatská republika	1
Čínská republika (Tchaj-wan)	1
Estonská republika	1
Francouzská republika	1
Lotyšská republika	2

Příklad jednoduchého ukazatele – počet studentů, kteří vyjeli během svého studia do zahraničí

Složené ukazatele (indikátory):

- pro vyjádření ukazatele slučuje několik jednoduchých ukazatelů, resp. využíváme výběru více položek z jedné oblasti)
- výhodou je detailnější pohled na sledovaný jev
- příklad: Přehled (časová řada) zapsaných a absolvovaných předmětů podle jazyka, z toho zapsaných jako povinných a povinně volitelných; poměr (časová řada) výjezdů studentů do zahraničí vůči všem studentům, z toho výjezdy delší než 3 kalendářní měsíce (lze připojit další ukazatel a rozdělit podle typu studia)

Year	POČET	POČET_NAD_3M	POČET Bc
FDU	1		
FEK	1		1
FPR	2	1	
FST	1		
Spojené království Velké Británie a Severního Irska	**2**		**2**
FPE	2		2
Spolková republika Německo	**12**	**3**	**10**
FEK	3	1	3
FEL	1		1
FF	3		3
FPE	3		2
FPR	2	2	
2015	**125**	**20**	**67**
Belgické království	**1**		
FF	1		
Bulharská republika	**3**		**2**
FAV	1		
FDU	2		2
Česká republika	**1**		**1**
FDU	1		1
Čínská republika (Tchaj-wan)	**2**		**1**
FEK	1		

Příklad složeného ukazatele ze čtyř jednoduchých – počet studentů, kteří vyjeli během svého studia do zahraničí a z nich počet těch, kteří měli výjezd delší než 3 měsíce, z toho počet studentů bakalářského studia, rozděleno dle fakult.

Řešení problémů

Při tvorbě ukazatelů bylo řešeno mnoho problémů, které se týkali nejen jejich samotné definice, ale i datových podkladů či databázových nejasností. Příklady jsou uvedeny v následující tabulce:

PROBLEMATIKA NA ÚROVNI DEFINICE UKAZATELŮ	
PROBLÉM	POPIS PROBLÉMU
ukazatel jazyk předmětu	základní tabulka varianty předmětu obsahuje atribut jazyk, ve kterém je možné předmět vyučován tabulka předmětů vázaných na studenta obsahuje atribut jazyk, který ovšem nemusí být definován u varianty předmětu
ukazatel splnění přijímacího řízení	atribut rozhodnutí komise/resp.děkana/resp.rektora…kód 20=přijat atribut uchazeč vyhověl atribut vyhověl bez ...tyto atributy nemají žádnou souvislost...
definice časové osy	obecně problematika přiřazení k časové osy např. Vysokoškolské práce obsahují datum zadání, plánovaného odevzdání, odevzdání, datum obhajoby – které datum přiřadit k časové ose?
propojení vzájemně nezávislých atributů	např. propojení učitele a oboru – mnoho variant: podle kmenového pracoviště zaměstnance,po dle vyučovaných předmětů a studijních plánů, podle oboru studentů, kteří navštěvují předmět učitele, podle rovrhových akcí…
ukazatel zatížení učitele	př. učitel je garantem předmětu, dle základních informací je přednášejícím i cvičícím předmětu, ale fyzicky nemá žádnou rovrhovanou akci… je to zatížení? Kolik hodin stráví garantováním předmětu?
ukazatel studenti recyklanti	chceme-li sledovat recyklanty dle oboru, kteří jsou to studenti? Ti, kteří zrecyklují k nám na obor; ti, co odcházejí z oboru a recyklují jinam; jen ti recyklující z oboru na stejný; recyklující na jiný obor ale na stejnou fakultu/školu…

PROBLEMATIKA NA ÚROVNI DAT	
PROBLÉM	POPIS PROBLÉMU
datová irelevance	př. praxe – jasné databázové položky, nevyplněné, či vyplněné nahodile, nesouvisle
kombinace datových položek na různých úrovních	př. ukazatel rozložení studentů na oboru – je třeba sledovat i studenty na kombinacích, které jsou na nižší resp. na nezávislé úrovni, ale je potřeba je přiřadit na stejnou úroveň k oboru = úprava základních ientifikátorů
více vstupních databází s informacemi	více zdrojů pro jedny informace - př. Výjezdy studentů - uloženy ve STAGu, v INISU, některé duplicitně, některé nikde, některé jednou

PROBLEMATIKA NA ÚROVNI DATABÁZE	
PROBLÉM	POPIS PROBLÉMU
nestejné cizí klíče pro stejné datové položky	př. uchazeči vs studenti dimenze oborů nesoucí stejné atributy jsou plněny zvlášť a obsahují položky omezené jen na danou oblast… nutno sjednotit
nejednotné IDM	každá databáze obsahuje jednoznačné klíče na identifikaci osoby, ale vzájemně nejsou propojené jednoznačným ID nejednoznačné vazby - př. u pseudorodných čísel - RČ nesedí; tituly osob v různých zdrojích nestejné… ruční opravy a vyhledávání - DWH zavádí jednotné IDM

Případová studie 2: Univerzita Pardubice

Klíčovým faktorem pro hodnocení ukazatelů kvality studijních programů je jednotný datový zdroj dostupný pro všechny zainteresované strany. Tímto zdrojem v prostředí UPa bude do budoucna datový sklad (data *warehouse, DWH*), který je součástí širšího konceptu označovaného zkratkou BI (*business intelligence*).

Kromě datového skladu (centrálního úložiště integrovaných dat) zahrnuje BI i další technologické a procesní prvky. Z technologických prvků lze jmenovat zejména problematiku načítání a transformace dat z různorodých primárních datových zdrojů, stahování dat z externích zdrojů, monitorování funkčnosti celého řetězce, auditování a efektivního zpřístupňování dat a vytváření reportů. Mezi organizační prvky BI pak patří její ukotvení do interního fungování organizace v souladu s platnou legislativou, stanovení odpovědnosti za provoz, procesy související se zpřístupňováním dat apod.

Základní popis technologického řešení

IT prostředí Univerzity Pardubice (UPa) je založeno na hydridním využívání on-premise technologií (tzn. lokálně instalovaných a spravovaných databází a dalšího softwarového i hardwarového vybavení) a cloudových technologií, s převážným zastoupením technologií od společnosti Microsoft (O365, Azure). Tuto hybridní podstatu kopíruje i realizace BI, kdy primární datové zdroje se nachází v on-premise prostředí, zatímco co většina technologií implementující BI je umístěna v cloudu MS Azure. Zároveň probíhá orchestrace některých funkcionalit mezi cloudovým a on premise prostředí, kdy např. některé externí zdroje mají provozovatelem definovanou přístupovou politiku umožňující dotazování pouze z IP adres univerzity.

Základní schéma technického řešení:

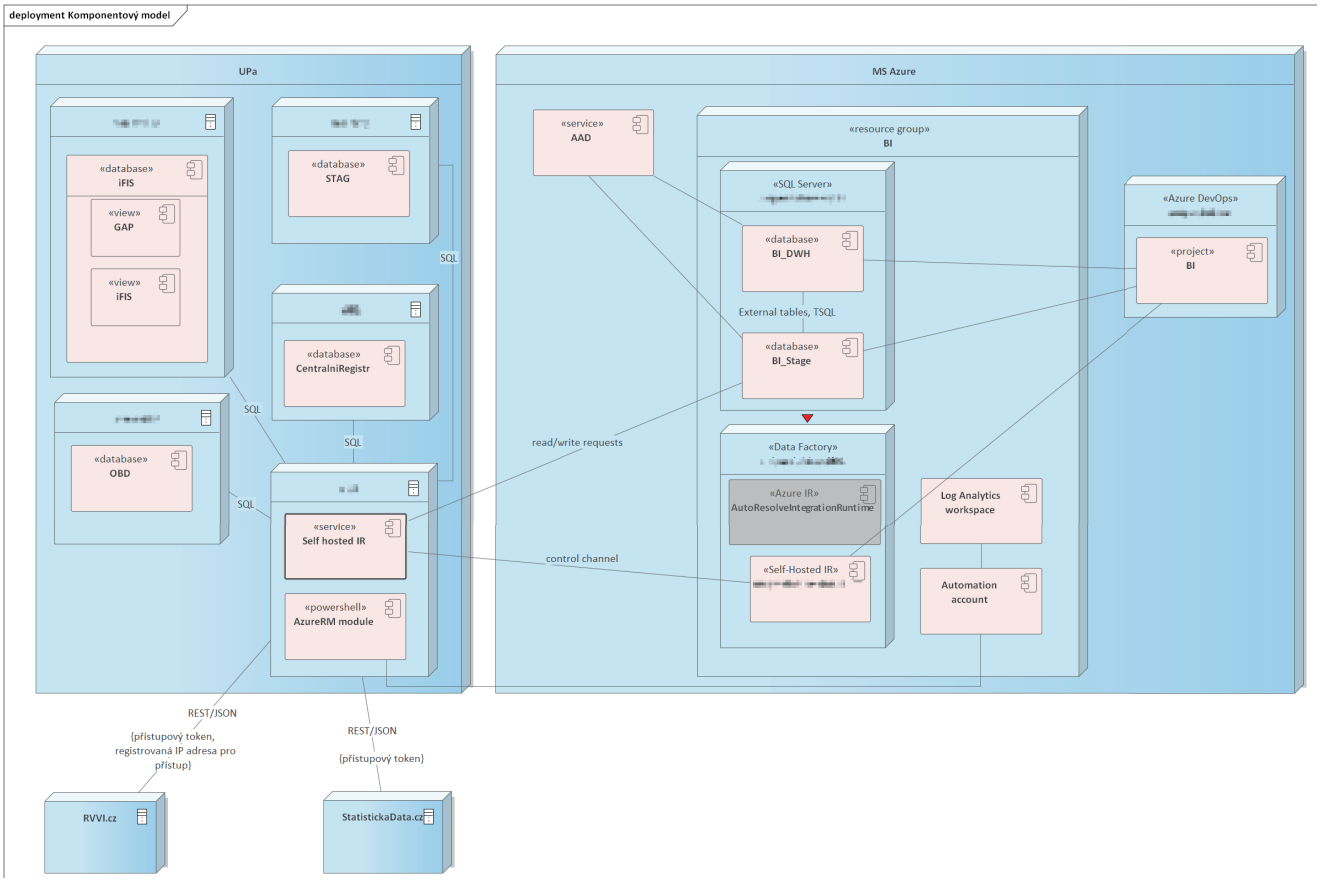


Schéma implementace BI a jeho rozdělení mezi on-premise a cloudové prostředí

V lokálním prostředí je instalována naprostá většina primárních datových zdrojů. Jedná se o databáze těchto systémů:

- IS STAG – informační systém pro správu studijní agendy. S ohledem na *raison d'être* Univerzity Pardubice se jedná o primární provozní systém a klíčový datový zdroj pro hodnocení a řízení kvality.
- iFIS – ekonomický informační systém. Představuje důležitý datový zdroj především pro ekonomické hodnocení (rozpočty, čerpání finančních prostředků, …). S ohledem na implementaci dalších funkcionalit v rámci tohoto systému (např. spisová služba, registr smluv) lze v budoucnu očekávat další rozšiřování získávaných dat.
- CRO – centrální registr osob. Univerzitní implementace *identity managementu*, řízení identit, zajišťující jednoznačnou identifikaci osob (zaměstnanců, studentů) a eliminující případné duplicity plynoucí z historického vývoje vztahu osoby k univerzitě (např. „bývalí studenti, kteří se stali zaměstnanci") apod.
- OBD + GAP – osobní bibliografická databáze a evidence grantů a projektů. Jedná se o systémy evidující publikační aktivitu zaměstnanců a studentů univerzity, jakož i vztah této aktivity ke grantům, ze kterých je tato činnost financována.
- Active Directory – implementace adresářových služeb od společnosti Microsoft. Udržuje evidenci uživatelských účtů, jejich zařazení do skupin, zajišťuje ověřování identity atd. Hraje podpůrnou, ale důležitou roli, pro správnou činnost BI, zejména s ohledem na řízení přístupových oprávnění.

Data z výše uvedených zdrojů v on-premise prostředí jsou prostřednictvím tzv. *self-hosted integration runtime* [1] načítána a přenášena do cloudového prostředí. Data jsou načítána tak, jak jsou, pro jejich další zpracování se využívá výpočetní síly cloudových prostředků. Procesně se tak jedná o modifikovaný mechanismus klasického ETL (extract-transform-load), tzv. ELT (extract-load-transform), v čemž se řešení UPa odlišuje od většiny ostatních VVŠ zapojených do projektu.

Na straně cloudu je celý proces řízen službou data factory, která fakticky řídí lokálně instalovaný *self-hosted integration runtime*. Výsledkem její činnosti je periodický přenos nemodifikovaných dat z on-premise prostředí do první ze dvou využívaných databází (tzv. *STAGE* databáze). Dále zajišťuje transformaci již přenesených dat do tvaru vhodného pro zpracování statistickými nástroji. Transformovaná data následně ukládá do druhé databáze, která fakticky představuje vlastní datových sklad (tzv. *DWH* databáze).

Tento základní proces doplňují další služby poskytované prostředím MS Azure, jako jsou:

- *Azure Active Directory* (AAD) obsahující synchronizované účty z on-premise *Active Directory* a zajišťující autentizaci a autorizaci osob v cloudovém prostředí.
- *Automation Account* zajišťující automatizaci rutinních činností, jako je např. stahování logů, jejich propagaci do *Log Analytics* (analyzátor logů, další využívaná služba v rámci MS Azure), orchestraci volání webových služeb externích datových poskytovatelů (např. RVVI) apod.
- Datový model a definice transformací dat jsou ukládány ve verzovacím systému GIT v rámci služby Azure *DevOps*. To umožňuje řízení celého vývojového cyklu a udržitelnost řešení v čase (možnost snadného návratu k předchozím verzím v případě problémů, sledování změn v čase, apod.).

Popis databází

Datové úložiště BI je realizováno pomocí dvou databází:

- BI_STAGE
- BI_DWH

BI_STAGE obsahuje data z primárních systémů, přičemž pro každý primární systém je definováno samostatné schéma. Aktuálně existují schémata pro interní systémy AAD, CRO, GAP, iFIS, OBD a STAG a externí systém RIV.

1 Softwarový klient nainstalovaný na univerzitním serveru spárovaný s cloudovým prostředím univerzitního tenantu

Jsou definována i speciální schémata LOG, MANUAL a METADATA.

Schéma LOG obsahuje informace o průběhu přenosu a transformace dat, v budoucnu se předpokládá i jeho statistické využití pro monitoring průběhu plnění datového skladu.

Schéma METADATA slouží pro správu dodatečných informací k modelu datové struktury (udržovanému v CASE nástroji *Enterprise Architect*). Využití modelu spolu s metadaty umožňuje dosažení značného stupně automatizace při rozšiřování struktury datového skladu, což snižuje nároky na manuální úpravy celého systémy.

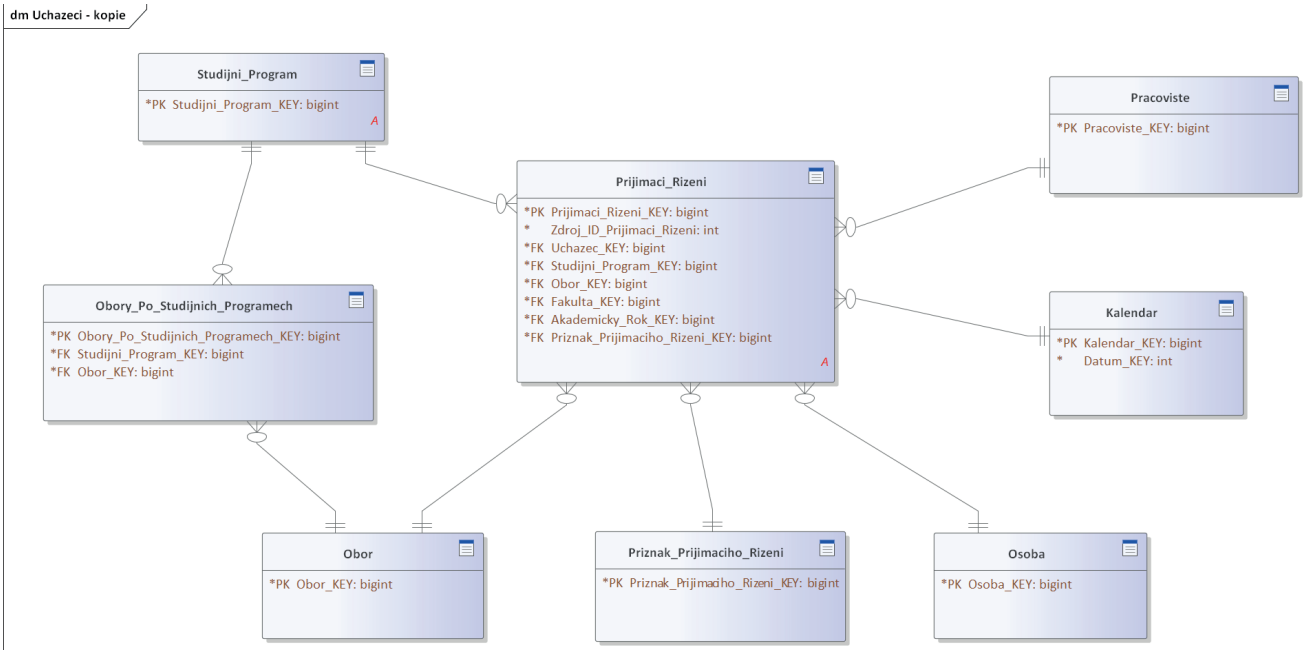
Schéma MANUL slouží pro správu ručních vstupů, např. v případě neexistence oficiálních datových zdrojů různých převodních číselníků apod. O problematice ručních vstupů blíže hovoří kapitola Ruční vstupy.

BI_DWH má formou *external tables* zpřístupněná data z BI_STAGE. Z těchto dat prostředním T-SQL procedur plní vlastní tabulky, které jsou primárně rozdělené do dvou schémat: *DWH_daily* obsahující denní přírůstky a *DWH_hist* obsahující historicky veškerá data.

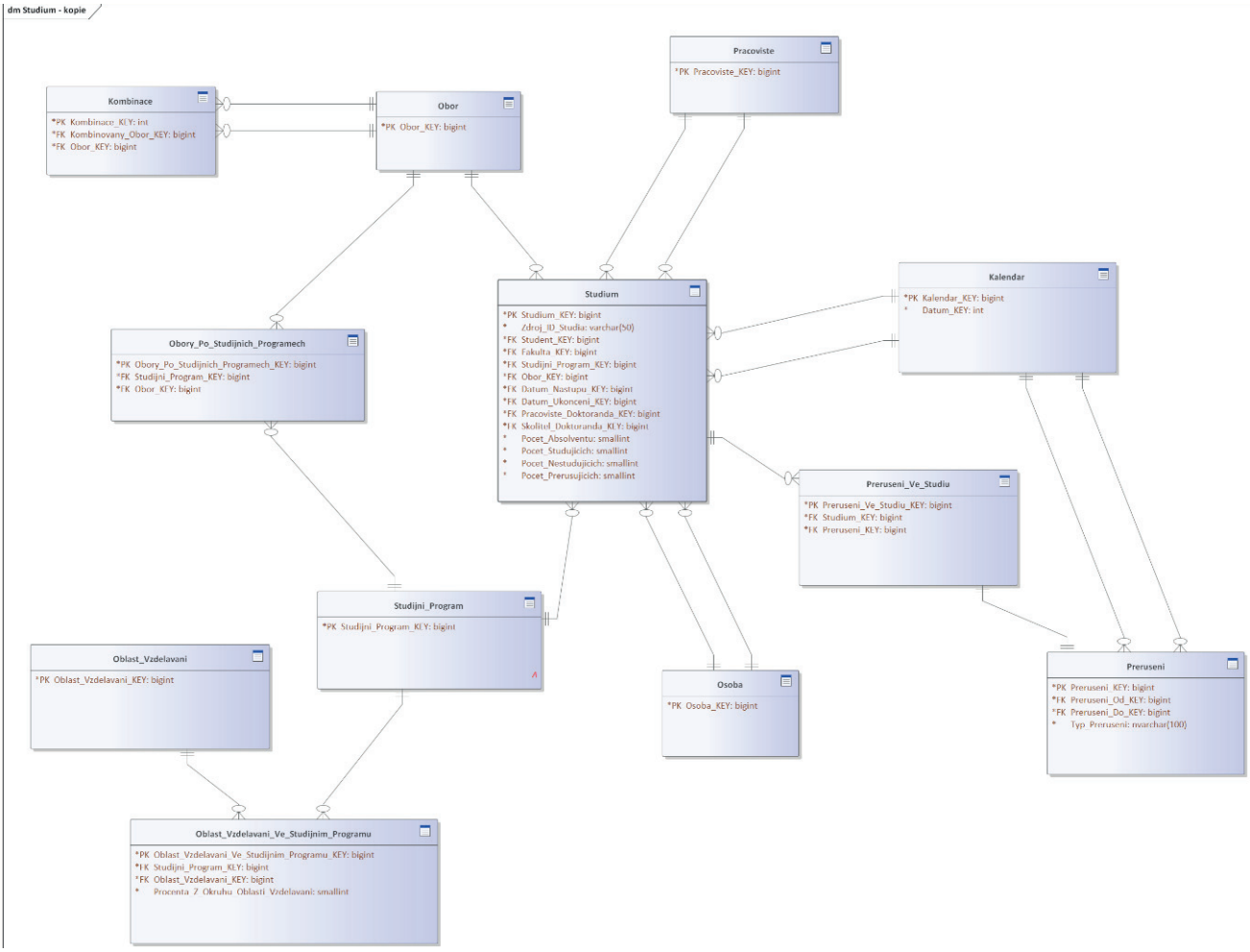
Aktuálně jsou zpracované datové kostky pro:

- Uchazeče a přijímací řízení
- Studium (dvě datové kostky)
- Studentské mobility
- Výsledovka (ekonomická sestava)

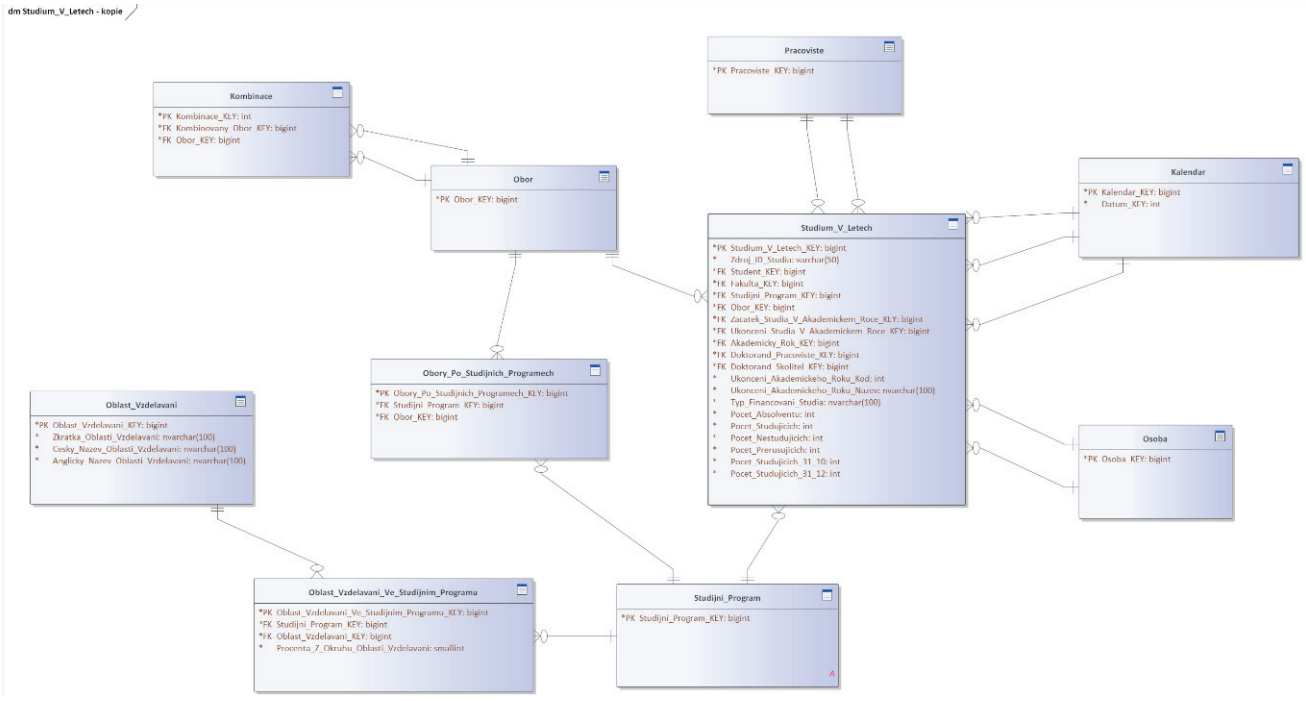
Ukázky modelů jsou uvedeny na následujících třech vyobrazeních. S ohledem na rozsáhlost datového modelu jsou uvedeny jen klíčové atributy a relace mezi jednotlivými tabulky, rozsah evidovaných atributů je však výrazně vyšší.



Model datové struktury pro analýzu přijímacích řízení (zobrazeny jen klíčové atributy)



Model datové struktury pro analýzu studia (zobrazeny jen klíčové atributy)



Model datové struktury pro analýzu studia v jednotlivých letech (zobrazeny jen klíčové atributy)

Specifickou dimenzi představuje tabulka Kalendar. Jejím účelem je maximálně zjednodušit analytickou činnost koncových uživatelů a odstínit je od problémů typu „patřil tento den do letního zkouškového období?“. Pro každý den tak systém generuje řadu atributů popisujících datum z nejrůznějších možných hledisek.

Drobný výsek informací ke konkrétnímu datu:

Results Messages

	Datum	Cislo_Dne_V_Tydnu	Cislo_Dne_V_Mesici	Nazev_Dne	Zkraceny_Nazev_Dne	Priznak_Vikendu	Cislo_Tydne_V_Roce	Datum_Zacatku_Tydne	Datum_Konce_Tydne	Cislo_Mesice	Nazev_Mesice	Zkraceny_Nazev_Mesice	Datum_Zacatku_Mesice	Datum_Konce_Mesice	Datum_Zacatku_Pr
1	2019-01-01	2	1	Úterý	Út	Ne	1	2018-12-31	2019-01-06	1	Leden	Led	2019-01-01	2019-01-31	2018-12-01

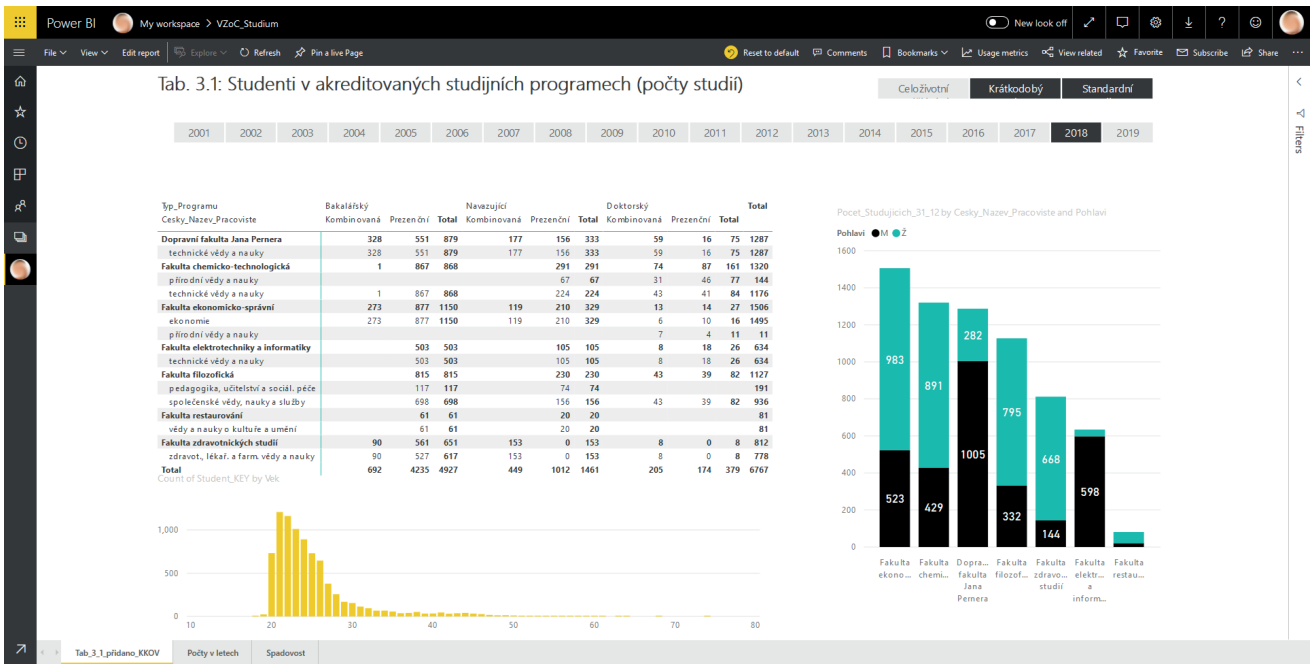
Ilustrace dat v dimenzi Kalendar

Nad daty ve schématu DWH_hist pak existují automaticky generované pohledy, které skrývají některé technické sloupce, které z uživatelského hlediska nepřináší žádný benefit. Implementace pohledů zároveň do budoucna zjednodušují údržbu celého systému – např. při refaktoringu datové struktury mohou být tyto pohledy zachovány bez nutnosti předělávat všechny navazující sestavy, apod.

Sestavy

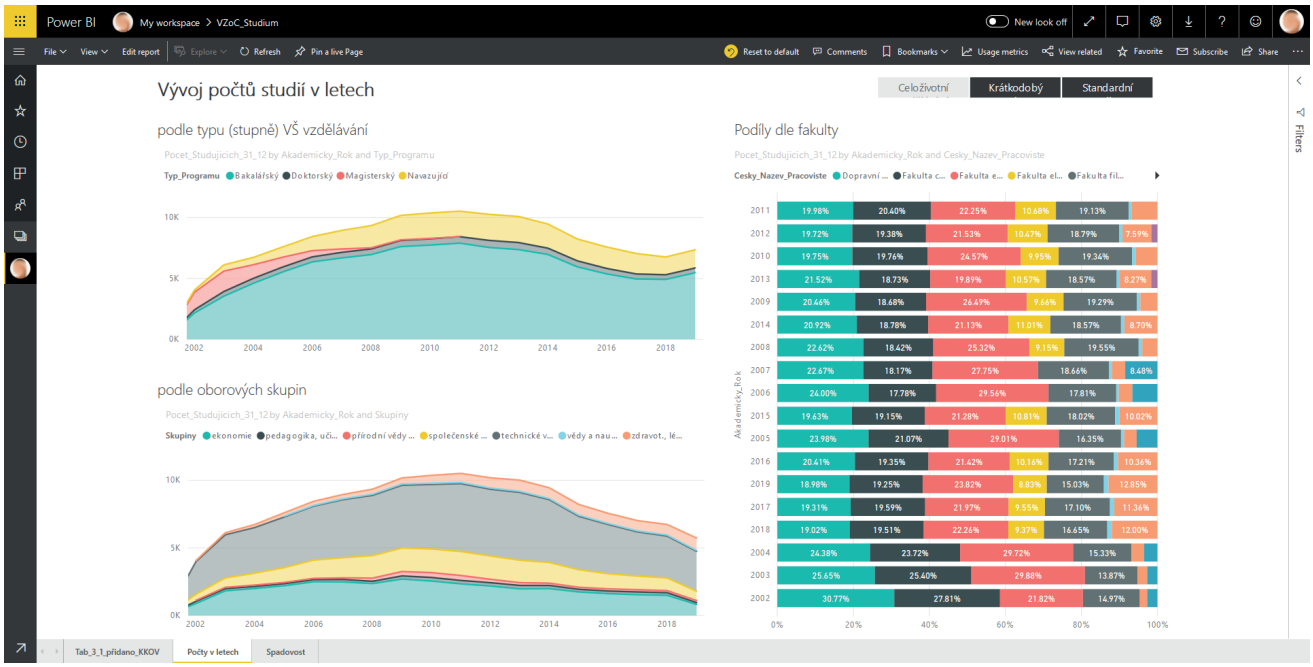
Funkčnost a životaschopnost celého řetězce tvorby datového skladu je možné ověřit pouze prostřednictvím sestav a reportů. Existence dat je základním předpokladem pro jejich tvorbu. Transformace dat z podoby uložené v primárních systémech do podoby vhodné pro tvorbu sestav a reportů představuje náročnou činnost (zejména na úrovni modelování) a tedy bylo nutné prioritizovat, která data se „přenesou“ z vrstvy STAGE do vrstvy DWH dříve a které později. Aby bylo možné technologické řešení ověřit, byla do vrstvy DWH přednostně zpracovávána data, ze kterých vychází výroční zprávy každoročně univerzitou zpracovávané; nicméně ve vrstvě STAGE je integrována širší množina dat.

Z prvotního zaměření na oblast výročních zpráv tak vyplývají i sledované ukazatele, které jsou fakticky určené požadovanou strukturou výroční zprávy. Výhodou těchto ukazatelů je, že je možné snadno komparovat data získaná z datového skladu s již existujícími daty uvedenými v dřívějších výročních zprávách.



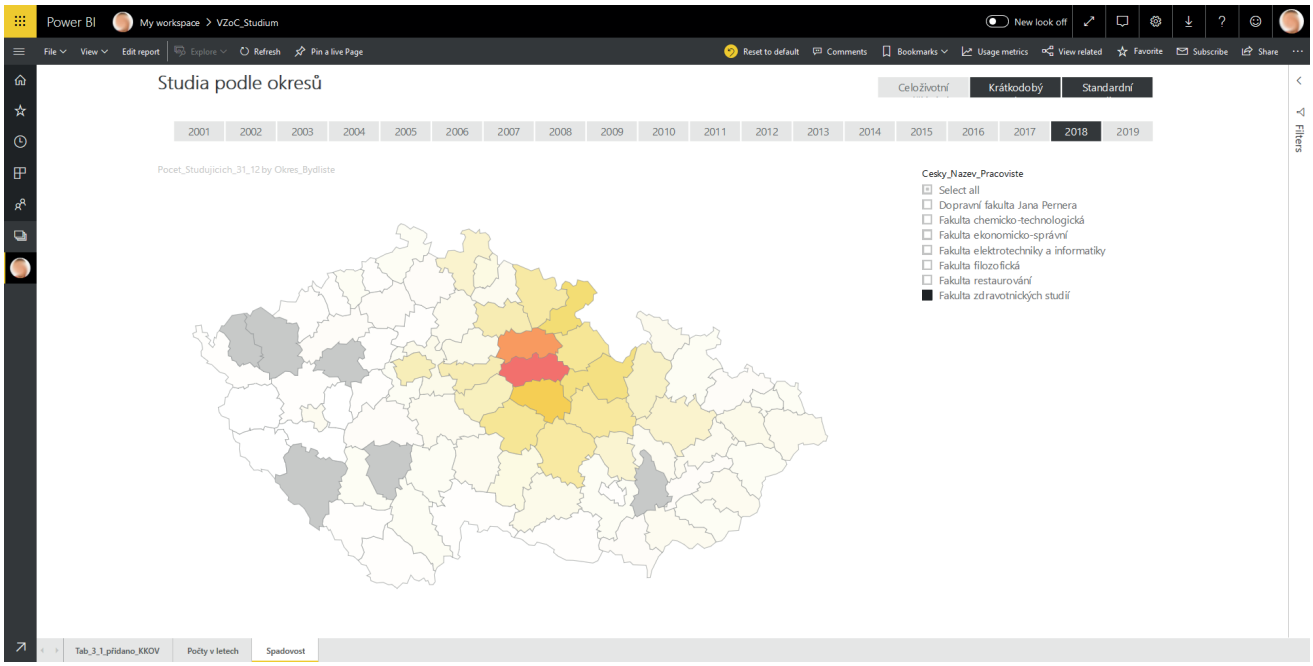
Ilustrace vyhodnocení ukazatele počtu studentů v akreditovaných studijních programech

Kromě samotné hodnoty ukazatele je pak možné v reálném čase analyzovat i jeho další aspekty, jako např. jeho vývoj v čase.



Vývoj ukazatele v času

Díky moderním způsobům vizualizace je snadné rovněž vizualizovat složitější atributy ukazatele, například geografické rozložení.



Spádovost studentů

Řízení přístupů

Přestože účel datového skladu spočívá primárně ve statistických úlohách, nelze opomenout fakt, že data, ze kterých jsou výsledky získávány, mohou mít charakter osobních údajů, a i samotné anonymní statistické výstupy mohou být z hlediska organizace někdy vnímány jako citlivé. Podstatnou součástí celého řešení tak musí být i zabezpečení respektující legislativní požadavky (např. ZKB, GDPR, aj.) a zvyklosti interního prostředí.

Zabezpečení administrátorského přístupu spočívá především v minimalizaci počtu osob, které mohou k datovému skladu přistupovat prostřednictvím vysoce privilegovaných účtů. Účty, prostřednictvím kterých technickou administraci zajišťují, jsou odděleny od standardně používaných účtů těchto osob a zabezpečeny multifaktorovou autentizací. Tento požadavek je o to intenzivnější v situaci, kdy je řešení realizováno v cloudovém prostředí.

Zabezpečení výstupů z BI je možné realizovat na dvou úrovních:

- Na úrovni výsledných sestav
- Na úrovni dat, ze kterých sestavy vznikají

Zabezpečení na úrovni výsledných sestav je jednodušší na realizaci. Vychází z faktu, že navzdory uživatelské přívětivosti moderních analytických nástrojů je jejich obsluha pro většinu běžných uživatelů stále komplikovaná. K tomu je zapotřebí připočíst fakt, že je zapotřebí znát základní relace mezi jednotlivými daty a zvládat jejich interpretaci. V neposlední řadě se složení koncových uživatelů často mění, např. v případě periodicky volených funkcí apod.

Lze proto oprávněně předpokládat, že tito uživatelů budou pouze konzumenty sestav a reportů, které jim na míru – a to včetně zohlednění oprávnění – připraví někdo jiný (např. operátor BI, oddělení datové analýzy, …). Koncovému uživateli pak bude poskytnuta pouze sestava pracující nad daty, na která má dotyčný oprávnění. Pro jiného obdobného uživatele (např. osobě na stejné pozici z jiné fakulty) pak bude vytvořena stejná sestava, jen s jinými daty.

Nevýhoda tohoto přístupu spočívá v potřebě tvořit více kopií téže sestavy (vždy na jinou množinou dat) a z toho vyplývající náročnost údržby a rozvoj těchto sestav – každou změnu je nutné zanést do více instancí téže sestavy. Rovněž existuje riziko, že k sestavě přistoupí i jiná, než oprávněná osoba, nebo že osoba disponující určitou sestav v čase pozbyde svého oprávnění. V neposlední řadě jakékoli změny v dodaných sestavách musí provádět operátor celého systému, který se tak snadno může stát úzkým hrdlem celého systému.

Zabezpečení na úrovni dat řeší některé výše uvedené problémy. Datový zdroj v tomto případě v sobě zahrnuje potřebnou definici oprávnění, díky které uživatelům poskytuje pouze ta data, která jim náleží. Přestože stále platí, že většině uživatelů budou zřejmě poskytovány na míru připravené sestavy, lze v případě zabezpečení na úrovni dat rovněž poskytnout vybraným pokročilým uživatelům přístup přímo k datovému zdroji. Tito uživatelé si tak mohou podle svých potřeb vytvářet vlastní sestavy a využívat k tomu libovolné nástroje podle svého uvážení.

Jedná se o náročnější řešení na technickou realizaci i na organizační zajištění. UPa má výhodu v tom, že v minulosti implementovala systém *identity managementu*, který zajišťuje, že každá osoba má jasně určenou identitu. Zásadní pro zabezpečení na úrovni dat je na něj navazující role *management*, který **automatizovaně** zařazuje osobu na základě jejího organizačního zařazení do skupin (*security groups*) a zároveň zajišťuje, že je z daných skupin zase vyřazena ve chvíli, kdy se její organizační zařazení změní. Díky tomu je možné udržitelným způsobem definovat přístupová oprávnění.

Princip zabezpečení spočívá v tom, že přístupové údaje nejsou vázány na konkrétního uživatele, ale na skupiny, do kterých je uživatel zařazen. Uživatelé se k datovému přihlašují svým doménovým účtem, stejným, jakým se přihlašují např. ke svým počítačům[2]. Datový zdroj je schopen ověřit, zda je konkrétní

2 Díky tomu, že je celé řešení postaveno nad technologiemi od firmy Microsoft, je jejich přihlášení konzistentní napříč všemi komponentami řešení.

uživatel zařazen alespoň do jedné ze skupin, na která jsou vázána oprávnění k datům. Pokud ano, zpřístupní uživateli data odpovídající všem skupinám, do kterých je zařazen. Pokud ne, zamítne uživateli přístup.

Problematický bod řešení spočívá v tom, že databázový engine typicky zná identitu přihlášené osoby, ale nemá k dispozici všechny skupiny, do kterých je zařazena. To je způsobeno tím, že autentizaci osoby neprovádí databázový engine jako takový, ale tato autentizace je zprostředkována *Active directory*. Bylo proto zapotřebí doplnit datový sklad o informace o uživatelích a jejich zařazení do definovaných skupin a implementovat logiku, která bude tyto informace aktualizovat.

V okamžiku, kdy je známá informace o identitě osoby a jejím zařazením do skupin, je možné definovat:

- Na které objekty z datového skladu může přistupovat (např. tabulky obsahující ekonomická data)
- Na která data z těchto objektů může přistupovat (např. ekonomická data pouze jedné fakulty)
- V jaké míře detailu jí budou data zpřístupněna (např. studijní data bez totožnosti studentů)

V prostředí SQL Serveru, na kterém je datový sklad UPa realizován, jsou jako klíčové technologie určené pro řízení přístupu použity tyto:

- Row-Level Security (RLS) – definice filtračních podmínek, které jsou vždy doplňovány k dotazům na zdrojová data, bez možnosti uživatele je změnit. Podmínky jsou definovány na úrovni databáze a jsou tak uplatňovány bez ohledu na to, z jakého nástroje uživatel data poptává.
- Dynamic Data Masking (DDM) – skrývání vybraných dat neprivilegovaným uživatelům. Uplatní se především v oblasti ochrany osobních údajů, kdy některé údaje přímo identifikují konkrétní osobu. Nejedná se nutně o údaje typu jméno a příjmení nebo e-mailová adresa, ale daleko zásadněji i o různé identifikátory typu identifikátor osoby z *IdM*. Tyto údaje uživatel BI typicky potřebuje mít k zpřístupněné pro to, aby mohl analyzovat data z různých oblastí, ale nepotřebuje znát jejich konkrétní hodnotu. Např. při analýze uchazečů a absolventů může potřebovat informace vztahující se ke stejné fyzické osobě, ale již nepotřebuje znát, jak se tato osoba jmenuje. Technologie DDM tak „nahradí" skutečnou hodnotu atributu jinou nesouvisející hodnotou.

Schematická ilustrace technologií RLS a DDM:

Ilustrace funkce zabezpečení pomocí Row Level Security a Dynamic Data Masking[3]

Technologická připravenost řešení musí být následována i organizační připraveností. Ze strany UPa bude nutné definovat pravidla, na základě kterých budou přístupy jednotlivým skupinám uživatelů přidělovány (automaticky či manuálně) a rovněž definovat, kterým uživatelům mají být zpřístupněna jaká data. Za tímto účelem by měla vzniknout vnitřní norma, která bude tyto definice obsahovat.

3 Ilustrace přebrány z https://docs.microsoft.com/en-us/sql/relational-databases/security/row-level-security?view=sql-server-2017 a https://docs.microsoft.com/en-us/sql/relational-databases/security/dynamic-data-masking?view=sql-server-2017

Ruční vstupy

Již vstupní analýzy ukázaly, že ne všechna data, která jsou pro činnost datového skladu potřebná, jsou v primárních systémech k dispozici.

Jedná se typově o situace, kdy:

- Jsou zapotřebí číselníky, které nejsou k dispozici v žádném z primárních systémů a leckdy nejsou vůbec dostupné ve strukturované podobě.
- Je zapotřebí převodník mezi různým kódováním určitých objektů, kdy primární systém pracuje s kódováním podle vlastních potřeb, ale vykazování je nutné provádět v jiném kódování. Častým doprovodným jevem je, že tento převodník existuje pouze někde na webu ve formě PDF či jiného, strojově obtížně zpracovatelného formátu.
- Je zapotřebí doplňovat k datům různý dodatečná data nebo korekční faktory, pro které primární systém opět nemá k dispozici žádnou datovou strukturu, nebo pracovník, který s výstupy pracuje, nemá oprávnění přistupovat k primárnímu systému.

Byly analyzovány možnosti, jakým uživatelům umožnit podobné informace zadávat. Při analýze byla brána v potaz především tato kritéria:

- Uživatelům by neměla být navyšována oprávnění nad rámec jejich pracovního zařazování.
- Uživatelé by měli mít možnost doplňovat potřebná data autonomně, bez potřeby součinnosti ze strany administrátora BI.
- Uživatelé by měli být schopni zadávat pouze validní data ve správných formátech apod.
- Nemělo by být zapotřebí aktivovat interní vývojový tým pro vytváření jednorázových a jednoúčelových aplikací.

Uživatelsky nejpříznivější možnost by i s ohledem na předpokládanou skupinu uživatelů byl tabulkový procesor typu Excel, který ale uživatelům poskytuje příliš velkou volnost z hlediska zadávaných dat, změn struktury apod., a navíc by jeho import musel provádět administrátor BI. To by vedlo jednak prodlevám doplňování dat a dále dříve či později k zahlcení administrátora BI požadavky na import.

Klasický vývoj aplikací by byl patrně ekonomicky nevýhodný, především s ohledem na omezené vývojové kapacity, nároky nejen na samotný vývoj ale i na infrastrukturu a s ohledem na to, že často jediným účelem těchto aplikací by bylo doplňování několik málo údajů.

Jako vhodná technologie se nakonec ukázala technologie *PowerApps*, kterou má většina VVŠ k dispozici v rámci O365. Tato technologie umožňuje snadnou implementaci „webových" či „mobilních" aplikací, bez znalosti standardního programování – většinou stačí schopnosti na úrovni pokročilejší práce se vzorci v programu Excel. Zároveň však umožňuje bez dalšího implementovat např. zabezpečení vůči AD apod.

Pomocí aplikací vytvořených v technologiích *PowerApps* budou uživatelé schopni zadávat potřebné informaci přímo do STAGE databáze (za tím účelem bude na úrovni DB vytvořen samostatný technický uživatel s přístupem jen a pouze k objektům, u kterých se předpokládá ruční vstup – viz dříve popsané schéma MANUAL). Je to zároveň asi jediná výjimka, kdy někdo jiný, než administrátoři, bude jakkoli přistupovat ke STAGE databázi.

Řešení problémů

Při budování datového skladu hraje velkou roli celkový kontext a stav IT prostředí na jednotlivých VVŠ. Některé problémy, se kterými se jiné VVŠ potýkají, tak v prostředí UPa neexistují nebo jsou minimalizovány. To je případ např. nejednotnosti identifikace jednotlivých osob napříč různými primárními systémy, kterou eliminuje existence Centrálního registru osob, který všechny dílčí identifikátory mapuje na jeden identifikátor.

Jiné problémy jsou naopak společné, typicky se jedná o definici ukazatelů, kdy není jasný výklad toho, „kdo je recyklant", „kdo je nový student" apod. Za tímto účelem byly zahájeny práce na tzv. **byznys slovníku**, který by měl sloužit jako zdroj informací o výkladu jednotlivých datových položek z datového skladu. Jeho vznikem by mělo být dosaženo jednotnosti výkladu položek a zároveň udržení kontinuity tohoto výkladu v průběhu personálních výměn (nejen) na volených pozicích. Předpokládáme, že i byznys slovník bude vytvořen svépomocně pomocí technologie *PowerApps* a že data budou uložena v jedné z databází datového skladu. Hlavní objem práce na tvorbě byznys slovníku předpokládáme v průběhu příštího roku.

Obecným problém, se kterým se potýkáme při práci na BI, je značná „fluidnost" celého vysokoškolského prostředí, kdy vysokoškolský zákon poskytuje vysokým školám značnou volnost v nastavení jejich procesů. Informační systémy tuto volnost musí podporovat nejen napříč jednotlivými vysokými školami, ale často i napříč jednotlivými fakultami téže školy. Proto je často náročné spolehlivě nastavit proces extrakce dat, protože i stejné situace mohou být často realizovány různými způsoby, přičemž množina těchto způsobu nemusí být nutně konečná.

S tím souvisí i problémy při komparaci výstupů z BI a výstupů z dříve vykazovaných VZ, kdy se často obtížně dohledává příčina rozdílů. Často může být zapříčiněna jinou metodikou výkladu pojmů, personálními změnami s různou mírou předání know-how, apod. Podobným problémům by do budoucna mělo aktuálně budované BI v kombinaci s byznys slovníkem zamezit, nebo jejich dopad alespoň výrazně eliminovat.

Případová studie 3: Vysoká škola Báňská – Technická univerzita Ostrava

Datový sklad (DWH) na VŠB-TUO má za cíl integrovat a historizovat data z několika interních informačních systémů a externích zdrojů. V první fázi studijní agendu (IS EDISON), *granty a projekty* (IS OBD, GAP) a *bibliometrická data* publikací vytvářených na univerzitě (IS OBD, Scopus, WOS).

Staging area datového skladu, kde se zrcadlí data ze vstupních systémů, i *Presentation area*, kde se ukládají již transformovaná a integrovaná data do jednotlivých *Data martů*, běží na samostatných virtuálních linuxových serverech jako PostgreSQL databáze.

ETL procesy jsou realizovány buď jako úlohy v nástroji Pentaho Data Integration (od tohoto řešení ustupujeme) nebo jako scripty v jazyce Python. K těmto účelům máme dedikovaný další samostatný virtuální server (ETL Server). Databáze i ETL procesy jsou zatím pouze v produkční verzi, do budoucna plánujeme nasadit i testovací.

Extrakce dat ze zdrojových systémů

Ze zdrojových databází extrahujeme data do Landing vrstvy Staging area Python skriptem s využitím knihoven SQLAlchemy a Pandas. Skript je konfigurován externím souborem s výčtem metadat.

Z externích biblografických databází (Scopus, Scimago, WOS) provádíme jednorázově export do CSV souboru, v plánu je automatizovaný periodický přístup přes API.

Transformace a 'Load'

Transformace dat a Load do Data martů probíhají mezi Staging area a Presentation area prostřednictvím Python skriptů. Ladění a prototypování skriptů provádíme v prostředí Jupyter Notebooku, která běží jako webová služba na ETL serveru.

Reporty

Uživatelé přistupují k datům a vizualizacím přes Power BI reporty, které jsou publikovány na web, zatím bez řízeného přístupu. Již funkční nebo v testovací verzi provozujeme report s vývojem počtu přihlášek v čase, analýzu počtu přihlášek dle místa bydliště nebo místa SŠ a report analyzující publikační produktivitu fakult a útvarů.

Prostor pro zlepšení

Při přípravě dat a tvorbě bibliometrických analýz jsme narazili na nedostatky zdrojových systémů (konkrétně OBD). Systém v současné verzi neumožňuje zadávat u autorů více afiliací k různým pracovištím a procentuální podíl mezi nimi. To při rozborech po organizačních jednotkách komplikuje spravedlivé dělení kreditu (a následně finančních prostředků) za publikace a může být zdrojem napětí mezi pracovišti.

Další nedostatek spočívá v přiřazování oborové kategorie k publikacím. Ta by měla být určena dle oborové kategorie časopisu, ve které je publikována. Z analýzy dat ale vyplývá, že je přiřazována intuitivně (z číselníku FORD). Celou situaci komplikuje fakt, že publikace mohou být indexovány v databázích Web of Science nebo Scopus, které mají odlišné kategorie, které jsou navíc odlišné od kategorií FORD, se kterými pracuje OBD.

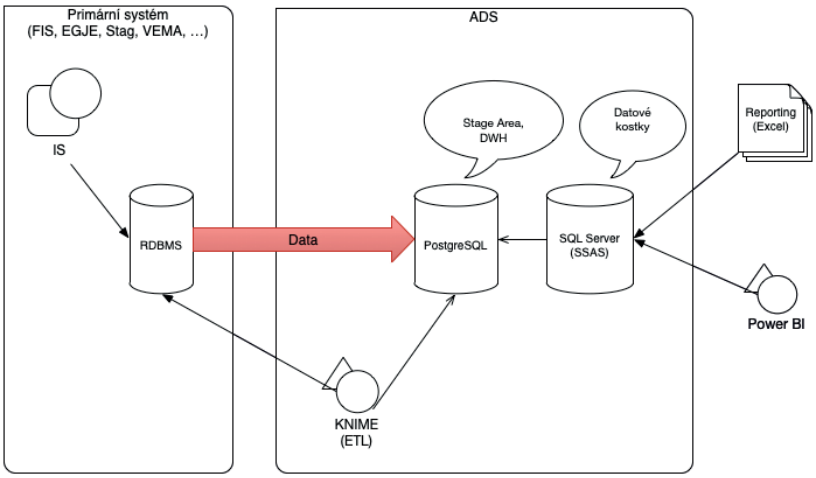
Dále se potýkáme s chybějícími nebo nekvalitními daty u některých atributů záznamů, neboť jsou v systému nepovinné anebo nejsou nabízeny z číselníku.

Uvedené nedostatky je nutno řešit kombinací administrativních opatření (úprava Směrnice) a úpravy IS OBD.

Případová studie 4: České vysoké učení technické v Praze

Technologický stack založený na:

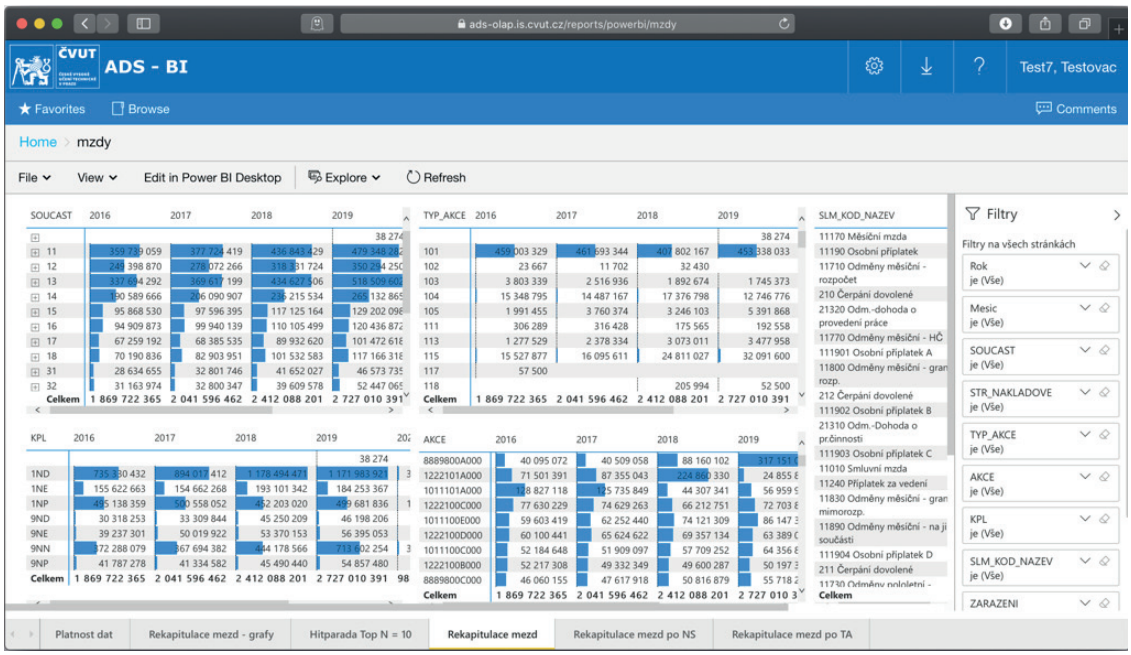
- KNIME (Konstanz Information Miner) – ETL – integrační vrstva
- ulozeni dat DWH (+ stage area) = PostgreSQL
- OLAP - MS SQL Analytic Service - Analytické služby
- prezentacni vrstva - MS Report Server BI (Power BI Desktop, MS Excel)

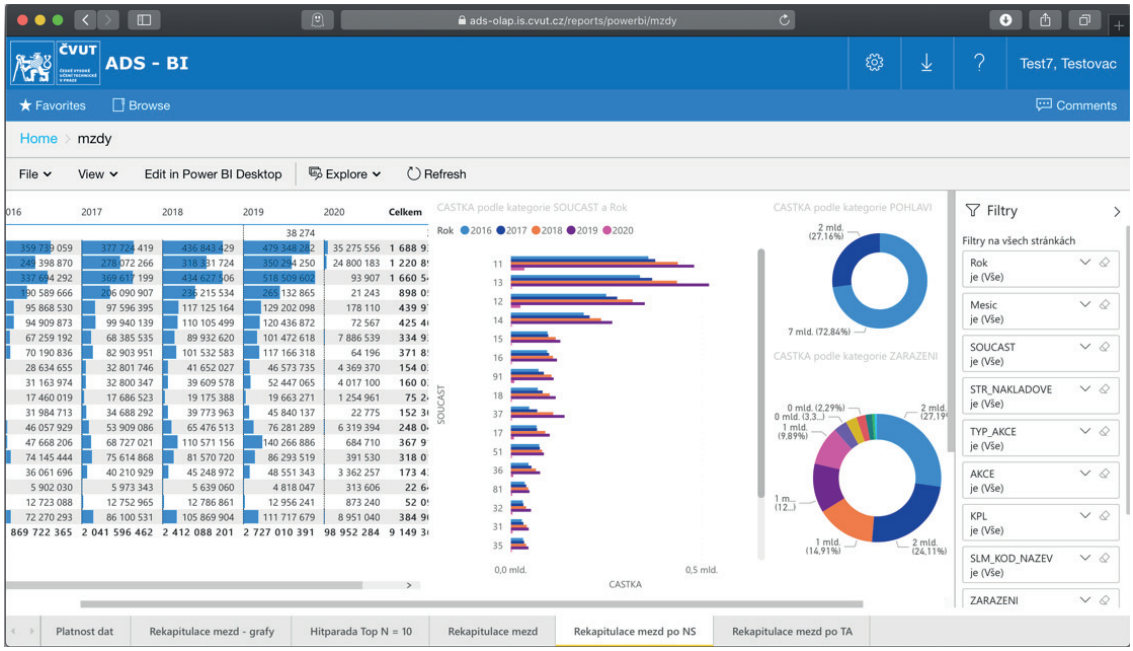


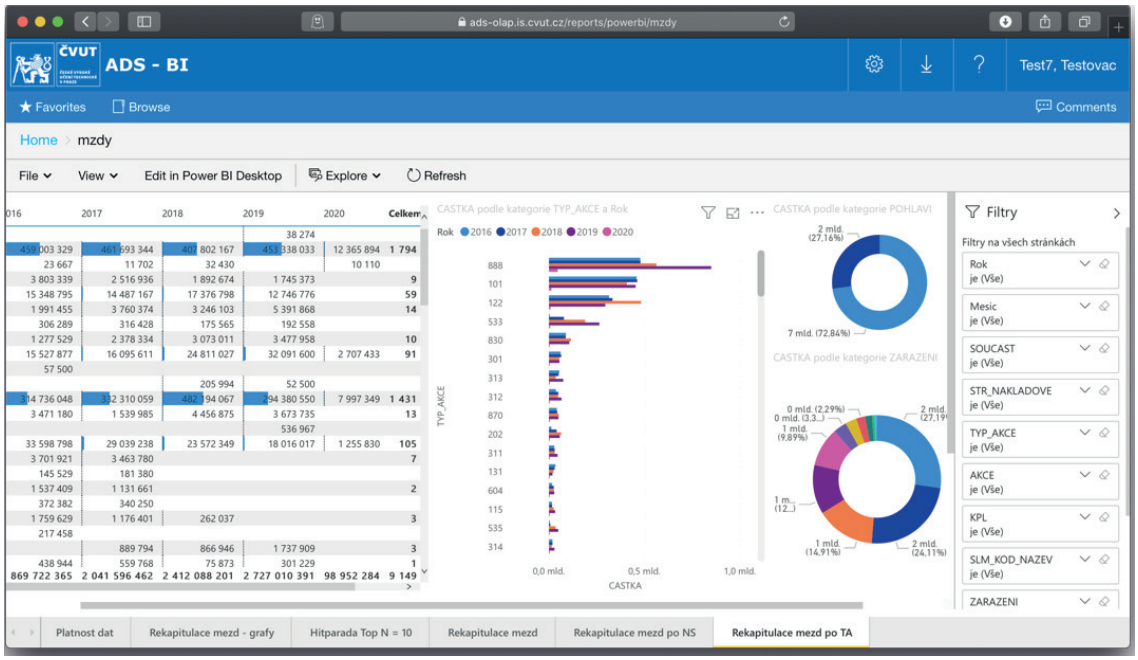
Kroky v realizaci:

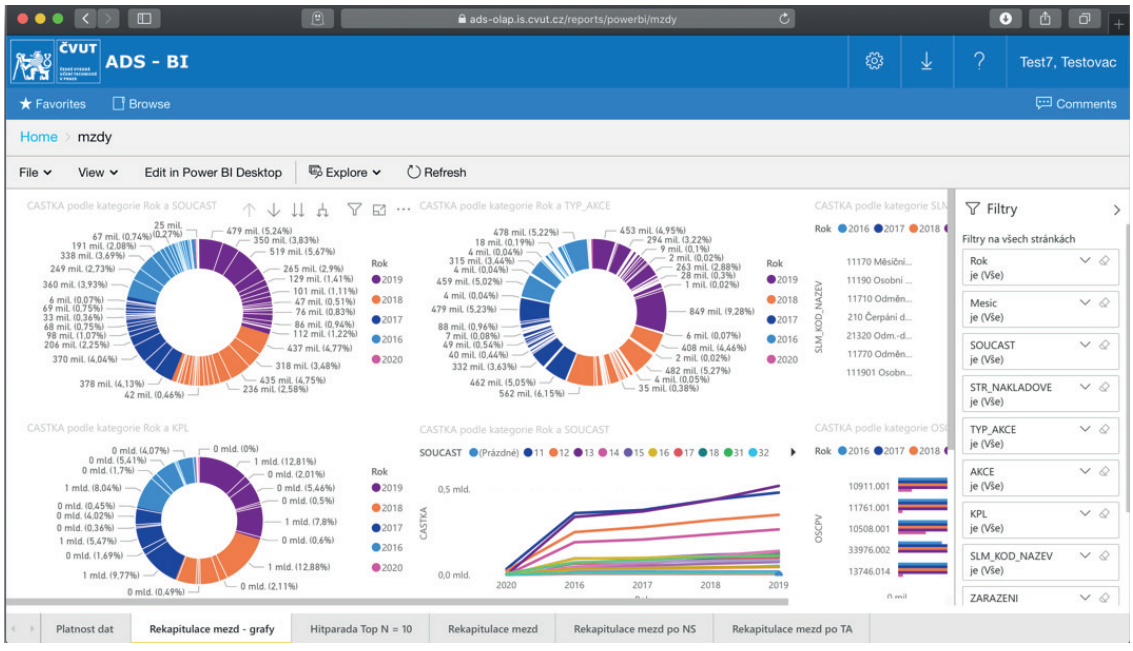
- naplnění EIS kostky
- nastavení row level security (RLS)
- vytvoření základních analytických reportů
- vytvoření základních analytických nástěnek
- výstupy v rozsahu výroční zprávy o činnosti a hospodaření
- další KPI dle ...

Ukázka výstupů:




ads-olap.is.cvut.cz/reports/powerbi/mzdy
ČVUT
ADS - BI
Test7, Testovac
Favorites
Browse
Comments
Home
mzdy
File
View
Edit in Power BI Desktop
Explore
Refresh
CASTKA podle kategorie SOUCAST a Rok
CASTKA podle kategorie POHLAVI
CASTKA podle kategorie ZARAZENI
Filtry
Filtry na všech stránkách
Rok je (Vše)
Mesic je (Vše)
SOUCAST je (Vše)
STR_NAKLADOVE je (Vše)
TYP_AKCE je (Vše)
AKCE je (Vše)
KPL je (Vše)
SLM_KOD_NAZEV je (Vše)
ZARAZENI
Platnost dat
Rekapitulace mezd - grafy
Hitparada Top N = 10
Rekapitulace mezd
Rekapitulace mezd po NS
Rekapitulace mezd po TA


ads-olap.is.cvut.cz/reports/powerbi/mzdy
ČVUT
ADS - BI
Test7, Testovac
Favorites
Browse
Comments
Home
mzdy
File
View
Edit in Power BI Desktop
Explore
Refresh
CASTKA podle kategorie TYP_AKCE a Rok
CASTKA podle kategorie POHLAVI
CASTKA podle kategorie ZARAZENI
Filtry
Filtry na všech stránkách
Rok je (Vše)
Mesic je (Vše)
SOUCAST je (Vše)
STR_NAKLADOVE je (Vše)
TYP_AKCE je (Vše)
AKCE je (Vše)
KPL je (Vše)
SLM_KOD_NAZEV je (Vše)
ZARAZENI
Platnost dat
Rekapitulace mezd - grafy
Hitparada Top N = 10
Rekapitulace mezd
Rekapitulace mezd po NS
Rekapitulace mezd po TA


ads-olap.is.cvut.cz/reports/powerbi/mzdy
ČVUT
ADS - BI
Test7, Testovac
Favorites
Browse
Comments
Home
mzdy
File
View
Edit in Power BI Desktop
Explore
Refresh
CASTKA podle kategorie Rok a SOUCAST
CASTKA podle kategorie Rok a TYP_AKCE
CASTKA podle kategorie Rok a KPL
CASTKA podle kategorie Rok a SOUCAST
Filtry
Filtry na všech stránkách
Rok je (Vše)
Mesic je (Vše)
SOUCAST je (Vše)
STR_NAKLADOVE je (Vše)
TYP_AKCE je (Vše)
AKCE je (Vše)
KPL je (Vše)
SLM_KOD_NAZEV je (Vše)
ZARAZENI
Platnost dat
Rekapitulace mezd - grafy
Hitparada Top N = 10
Rekapitulace mezd
Rekapitulace mezd po NS
Rekapitulace mezd po TA