# Dátový sklad DWH/BI NBS

Predmetom pripravovaného verejného obstarávania je dodanie centrálneho celobankového dátového skladu (NBS DWH). Požadované riešenie novej dátovej integrácie je navrhnuté ako jeden dátový ekosystém v kombinácii dvoch platforiem:

* DataLake
* Relačná časť

Dodávateľ by mal vyvinúť/dodať nové komponenty dátového skladu, ktoré budú:

* Integrovať dáta z existujúcich informačných systémov NBS a z externých systémov
* Poskytovať operatívny reporting, meranie výkonnosti a podporu analytických procesov nad dátami pre potreby jednotlivých organizačných zložiek spoločnosti.
* Podporovať prevádzku systému
* Umožnia samostatný rozvoj NBS DWH pracovníkmi IT a ďalších organizačných zložiek spoločnosti

Ďalej by mal dodávateľ v maximálnej miere vyžiť výsledky projektu Pilot DWH, ktorý mal za úlohu otestovať možnosti niektorých technológií. V súčasnosti zahŕňa asi 5% rozsahu dát NBS. V zmysle schválenej dátovej stratégie NBS je odporúčané tento pilot prepoužiť.

* Technická infraštruktúra – Oracle DB, MS SQL/TOLAP a Power BI
* Know-how pri extrakcii dát zo ŠZP a RBUZ
* ETL nástroj pre NBS DWH bol vybratý Oracle Data Integrator ako výsledok interného výberu

## Rozsah dodávky

Súčasťou dodávky uchádzača sú nasledovné komponenty:

1. Detailná analýza požiadaviek
2. Návrh dátového modelu na základe detailnej analýzy business požiadaviek
3. Integrácia na zdrojové systémy, vrátane pravidelného nahrávania dát do dátového skladu
4. Návrh kapacitného plánu
5. Inštalácia SW zabezpečenia
6. Implementácia dátového modelu
7. Implementácia reportov a analýz
8. Nastavenie BI nástroja
9. Podpora a prevádzka informačného systému dátový sklad (súčasťou dodávky uchádzača je aj prevádzkový manuál)
10. Prostredie pre podporu správu, riadenie a monitoring aktualizačných úloh
11. Prostredie pre správu číselníkov, mostíkov, referenčných dát a hierarchií jednotlivých dimenzií
12. Prostredie pre správu, spúšťanie a vyhodnocovanie dátovej kvality
13. Prostredie pre správu prístupov (DB a reportingová úroveň) na základe rolí
14. Prostredie pre správu dokumentácie k DWH a iných metadát (o modely, o reportoch, o mapovaní zdroj-cieľ)
15. Projektové riadenie zastrešujúce dodávku dátového skladu
16. Technická dokumentácia implementovaného riešenia
17. Zaškolenie obsluhy a používateľov dátového skladu

## Realizácia projektu

Obstarávateľ predpokladá realizáciu projektu do 36 mesiacov odo dňa nadobudnutia účinnosti zmluvy. Projekt bude mať štyri definované fázy dodávky s jednotlivými iteráciami (business výstup) s postupným pribúdaním funkcionality. Rozfázovanie dodávky je navrhnuté v projektovom pláne (Príloha A) a uchádzač je povinný toto rozfázovanie odborne pripomienkovať.

## Požiadavky na riešenie NBS DWH



## Východisková situácia

Východiskový stav je ďalej popísaný z pohľadu procesov spracovania dát, hlavných systémov využívaných v banke v súvislosti so spracovaním dát. Zvláštna pozornosť je venovaná popisu aktuálnej architektúry systémov a nástrojov používaných pre prácu s integrovanými dátami, ich reportingom a analýzou. Dôležitým faktorom pre budovanie cieľového riešenia je potom existujúce pilotné riešenie dátového skladu, ktorého niektoré komponenty sú k dispozícii pre ďalšie využitie.

### Procesy a systémy pre spracovanie dát

NBS je pre výkon svojej činnosti v prevažujúcej miere závislá na množstve externých dát zbieraných od subjektov na finančnom a nefinančnom trhu a externých poskytovateľov ekonomických informácií, a to na mikroekonomickej úrovni (dáta o konkrétnych subjektoch a finančných inštrumentoch), ako aj makroekonomickej úrovni (makro dáta – agregované ukazovatele o ekonomickej situácii a trendoch v jednotlivých sektoroch trhu). Tieto dáta sú zbierané, analyzované, čiastočne sa odvodzujú nové dáta ako produkt makroekonomických prognóz a výskumov. Niektoré dáta sú ďalej zdieľané v rámci medzinárodných inštitúcií (najmä ESAs, v rámci Eurosystému alebo Jednotného mechanizmu dohľadu) alebo aj s verejnosťou. Menšia časť dát vzniká primárne v NBS ako produkt vlastných aktivít NBS, napr. registre o subjektoch na trhu v rámci dohľadu a ďalšie dáta v rámci obchodných procesov, riadenia peňažnej hotovosti alebo finančného riadenia banky. Preto ďalej popisujeme spracovanie dát v týchto oblastiach, ktoré majú osobitný prístup k spracovaniu dát:

* Produkcia štatistických dát
* Hlavné procesy produkujúce dáta
* Analytické procesy
* Podporné procesy

#### Produkcia štatistických dát

Proces zberu a produkcie štatistických dát je v NBS riešený primárne v systéme **ŠZP** (Štatistický zberový portál), pričom je za tento proces zodpovedný predovšetkým Odbor štatistiky. Systém ŠZP je zákazkové riešenie od externého dodávateľa na mieru pre potreby NBS.

Štatistické dáta sú však zbierané aj formou e-mailových príloh, systémom RBÚZ a RS, ktoré sú vyčlenené z pôsobnosti odboru štatistiky – viď ďalej.

Pre účely dohľadu nad finančným sprostredkovaním a finančným poradenstvom sa pre zber dát využíva aj systém **RegFAP**, ktorého vlastníkom je odbor ODK a je vyvíjaný na odbore OFI.

Zber dát pre niektoré agendy je však stále riešený aj manuálne cez email a dáta v **MS Excel**.

ŠZP umožňuje definovanie používateľských, štandardných výstupov s exportom do CSV a XLSX pre ďalšie spracovanie (typicky pre zabezpečenie lokálnych požiadaviek) a automatizovanú výmenu niektorých dát s ESMA (XML štandard) a EBA/EIOPA (v štandarde XBRL).

Zber detailných dát pre AnaCredit a ich výmena s ECB sú riešené samostatným systémom **RBUZ/RS** (Register bankových úverov a záruk / Register subjektov). Tento systém bol vytvorený pre účely dohľadu, nie štatistiky, ale z povahy pravidelného spracovania dát od vykazujúcich subjektov sleduje rovnaký proces. Z pohľadu ECB je AnaCredit tiež štatistický systém, no v NBS nebol takto implementovaný.



Obrázok 1: Oblasť spracovania dát – Produkcia štatistických dát

Uvedené riešenia pre zber dát od vykazujúcich subjektov a ich ďalšie spracovanie však nezahŕňajú nástroje pre reporting, analýzy a systematickú distribúciu výstupov. Za týmto účelom vyvinul odbor štatistiky pre svoje vlastné potreby riešenie Apoštol s analytickým nástrojom Qlik Sense, ktoré malo za cieľ suplovať neexistujúcu analytickú platformu (resp. DWH/BI).

Existujú používatelia, ktorí pristupujú k dátam priamo cez SQL, R resp. Python.

V banke prebiehajú alebo sa plánujú tieto súvisiace aktivity:

* Moderná štatistika – konsolidácia agendy zberu dát na OST do systému ŠZP
* Upgrade ŠZP (vyššia flexibilita definovania štatistických výkazov a výstupov, zlepšenie výstupného rozhrania pre dátovú integráciu a ďalšie...)
* Zber detailných dát o úveroch fyzických osôb a vytvorenie riešenia pre register fyzických osôb
* ŠZP ESAs - presun časti funkcionality štatistického procesu zameraného na výkazy vo formáte XBRL zo ŠZP do nového dedikovaného systému ŠZP ESAs

#### Ďalšie procesy produkujúce dáta

Niektoré dáta vznikajú vnútri NBS pri výkone niektorých hlavných činností, ktoré nespadajú do oblasti produkcie štatistických dát, ako sú:

* Dohľad (zodpovednosť za vedenie rôznych registrov a údaje o individuálnych procesoch/konaniach so subjektmi)
* Bankové operácie (údaje o vykonaných obchodoch – systém IBFO/WSS)
* Riadenie rizika (analytické systémy Riskhouse, Credit Manager a Risk Manager)
* Riadenie hotovosti (údaje o stave a obehu hotovosti – systém EZO)
* Platobné systémy (údaje o transakciách – systémy SIPS, Target2)

Pre každú agendu existuje špecializovaný informačný systém, ktorý poskytuje aj operatívne výstupy. V prípade využívania spoločných informačných systémov na úrovni Európskej únie (IMAS pre procesy dohľadu nad významnými bankami, Target2 – zúčtovanie veľkých platieb) sú k dispozícii súčasne aj analytické reporty a výstupy pre uspokojenie dátových potrieb relevantných odborov.



Obrázok 2: Oblasť spracovania dát – Hlavné procesy produkujúce dáta

Dáta jednotlivých odborov (hlavných procesných oblastí) sú spracúvané a využívané viac menej pre vlastné potreby každého príslušného odboru a sú zdieľané iba obmedzene. Výnimkou sú spolupracujúce odbory Bankových operácií a Riadenia rizika, ktoré majú vytvorený vlastný analytický systém Riskhouse, ktorý napĺňa požiadavky týchto odborov bez potreby ďalšej konsolidácie s inými systémami.

Napriek malému prieniku týchto dátových oblastí prináša neexistencia jednotnej integrovanej (resp. analytickej) platformy komplikácie užívateľom niektorých odborov v oblastiach:

* Manuálna a netransparentná príprava dát pre reporting (mimo existujúcich analytických platforiem)
* Chýbajúci nástroj pre tvorbu vlastných reportov a analýz
* Duplicitné spracovanie externých dát
* Neznalosť a nemožnosť analyzovať dáta v kontexte dát z iných odborov
* Chýbajúce prepojenie registrov a IDAS/ASDR s vykazovanými údajmi dohliadaných subjektov z oblasti štatistiky

Aktuálne strategické aktivity NBS v tejto oblasti sú:

* ASDR – Implementácia jednotného agendového systému pre dohľad a reguláciu (proces manažment dohľadu), ktorý má nahradiť IDAS
* DMS / ECM – Implementácia integrovaného systému pre správu dokumentov (DMS) a Enterprise Content Managementu

#### Analytické procesy

Osobitná skupina procesov z pohľadu spracovania dát sú:

* Finančná stabilita
* Menová politika
* Výskum

Tieto procesy sú konzumentom dát publikovaných z procesu štatistiky a ďalej množstva externých ekonomických dát a zdrojov, ktoré sú predmetom rôznych analýz, modelovaní a predikcií. V každom prípade sú tieto procesy životne závislé na integrovaných dátach a súčasne poskytujú nové dáta, respektíve obohacujú agregované štatistické dáta o odvodené ukazovatele a predikcie.



Obrázok 3: Oblasť spracovania dát – Analytické proces

Dnes sú činnosti príslušných útvarov závislé na manuálnej príprave dát cez MS Excel a štatistických nástrojoch.

#### Podporné činnosti

Medzi podporné činnosti sú v NBS radené procesy:

* účtovníctvo a finančné riadenie (SAP FINU/HRO + BW)
* hospodársko-prevádzkové činnosti a služby (SAP FINU/HRO + BW)
* interný audit
* legislatívno-právna agenda
* IT
* ochrana a bezpečnosť
* medzinárodné vzťahy
* komunikácia
* riadenie ľudských zdrojov
* agenda riadiacich orgánov

V rámci týchto procesov vzniká iba obmedzené množstvo údajov, ktoré prakticky nemajú využitie v hlavných procesoch banky.



Obrázok 4: Oblasť spracovania dát – Podporné procesy

Aktuálne strategické aktivity NBS v tejto oblasti sú:

* Migrácia SAP FINU/HRO na technológiu S4HANA

### Súčasná architektúra aplikačnej vrstvy práce s dátami

Nasledujúci obrázok znázorňuje relevantnú časť súčasnej architektúry aplikácií a systémov v NBS, pričom tieto sú zoskupené do logických skupín podľa účelu.



Obrázok 5: Aplikačná vrstva – súčasný stav

* **Externé zdroje** – zdroje dát a informácií využívané pri analýzach a reportingu. Vo väčšine prípadov sú dáta z týchto zdrojov získavané manuálne, prípadne dochádza k výmene dát cez zdieľané úložisko.
* **Primárne systémy** – aplikačné systémy a registre prevádzkované bankou pre podporu primárnych a podporných procesov a úloh. Ich dáta sú čiastočne integrované v rôznych lokálnych databázach (*viď Integračná vrstva*), resp. sú exportované do Excelov pre ďalší reporting a analýzy. Medzi zdrojové systémy z tohto pohľadu zahŕňame aj Makroekonomickú databázu, ktorá je iba zdieľaným úložiskom časových radov v MS Excel.
* **Integračná vrstva** – riešenia pre integráciu dát NBS, ktoré vznikli v priebehu histórie pre podporu individuálnych potrieb jednotlivých odborov NBS; zahŕňajú ETL procesy pre plnenie integrovanej dátovej základne a prípadné data marty. V NBS sa využívajú databázové systémy Oracle (DWH Pilot, SAP BW) a Microsoft SQL Server (Apoštol databáza, Riskhouse, AŠI) + ETL systém MS Integration Services.
* **Prezentačná / Analytická vrstva** – obsahuje nástroje a služby Business Intelligence, ktoré sprostredkovávajú prístup k dátam pre koncových užívateľov (Qlik Sense v prípade Apoštol, MS Analysis Services Tabular Model pre DWH Pilot, resp. MS Reporting Services pre Riskhouse)
* **Aplikačná vrstva BI** – predstavuje koncové klientske nástroje pre prístup k integrovaným dátam

#### Externé zdroje

Rôzne odbory využívajú nasledujúce externé zdroje:

Štátna správa

| Označenie | Názov | Komentár k používaným dátam |
| --- | --- | --- |
| ŠÚ SR | Štatistický úrad SR | vybrané datasety podľa požiadaviek odborov https://data.statistics.sk/api/html/help-sk.html |
| KaPor | Katastrálny portál |  |
| FR SR | Finančné riaditeľstvo SR | súbory údajov z daňových výkazov a z registrov daňových subjektov |
| Sociálna poisťovňa | Sociálna poisťovňa | v súčasnosti veľmi obmedzený prístup k dátam, vo vyjednávaní je dohodnutie prístupu k širším anonymizovaným dátam  (počty zamestnancov vo firmách, mikrodáta o poistencoch a poberateľoch dávok) |
| Finančná správa | Dáta z Finančnej správy na základe zmluvy | mesačné agregáty, daňové priznania PO a FO k dani z príjmov – mikrodáta - ročné, daňové priznania DPH mikrodáta – mesačné a kvartálne |
| Register trestov | oversi.gov.sk |  |
| Štátna pokladnica |  | dáta o plnení verejných rozpočtov, ručne sťahované a integrované v AŠI databáze |
| Ardal | Agentúra pre riadenie dlhu a likvidity | Výmena údajov na základe zmluvy cez SFTP úložisko |
| Eurofondy | Systém ITMS 2014+ | OEM využíva 1 dataset cez API prevolávané z Matlab a odkladané do lokálnej databázy |
| SILC | Survey of Income and Living Conditions | Dáta prieskumu o vývoji príjmov a životných podmienok domácností od ŠÚ SR – manuálny vstup |
| Vzorka podnikov | Zjednodušené finančné výkazy pre vzorku cca 5000 podnikov | CSV súbor posielaný zo ŠÚ SR cez SFTP |
| eKasa | Dáta o tržbách podnikov z Finančnej správy | Atomické anonymizované dáta; dnes prenos emailom |
| Počty sporiteľov | Štatistika MPSVaR o počtoch klientov a objeme príspevkov | za každú DSS polročne - manuálny vstup |

Tabuľka 5: Prehľad externých zdrojov – štátna správa

Verejné registre

|  |  |  |
| --- | --- | --- |
| Označenie | Názov | Komentár k používaným dátam |
| RÚZ | Register účtovných závierok | 1x/mesiac automatický download do RS |
| RPO | Register právnických osôb | primárny zdroj Registra subjektov (denná synchronizácia); využívané aj na jednotlivých odboroch |
| RO | Register organizácií | zdroj Registra subjektov |
| Gleif | GLEIF - Global Legal Entity Identifier Foundation | zdroj Registra subjektov |
| RPVS | Register partnerov verejného sektora | aktuálne využívané pri AML validáciách protistrán <https://rpvs.gov.sk/OpenData/swagger/ui/index> |
| ŽR SR | Živnostenský register |  |

Tabuľka 6: Prehľad externých zdrojov – verejné registre

EÚ

| Označenie | Názov | Komentár k používaným dátam |
| --- | --- | --- |
| ECB SDW | ECB Statistical Data Warehouse | všetky datasety kde figuruje krajina "SK" |
| KL | Kurzový lístok ECB |  |
| ESMA | European Securities and Market Authority | údaje zo systémov TREM, FIRDS, FITRS, EMIR a SFTR (XML ISO 200022) + shortselling, Registre ESMA (XML dáta) – veľké dáta, v súčasnosti prebieha problematické ručné spracovanie |
| CSDB | Centralised Securities Database - databáza cenných papierov | 4x CSV export manuálne stiahnutý 1x/mesiac do Apoštol |
| Eurostat | Štatistický úrad EÚ | štatistiky na európskej úrovni, ktoré umožňujú porovnávanie medzi krajinami a regiónmi |
| SHSDB | Securities Holdings Statistics Database - databáza držby cenných papierov | dáta zo SHSDB 1x za mesiac stiahnuté ručne cez Lotus Notes a nahraté do IS Apoštolu |
| ECB/FIS |  |  |

Tabuľka 7: Prehľad externých zdrojov – EÚ

Zdroje ekonomických informácií

| Označenie | Názov | Komentár k používaným dátam |
| --- | --- | --- |
| Bloomberg | Bloomberg | prístup cez Excel plug-in  očakáva sa možnosť flexibilného rozširovania požiadaviek na preberanie údajov z Bloomberg (API) |
| Bisnode | Bisnode - poskytovateľ digitálnych obchodných informácií | závierky vo formáte CSV cez http; 1x/mesiac do platformy Apoštol; ďalej údaje o konkurzoch a iné podnikové mikrodáta |
| CRIBIS | CRIBIS - Register informácií o firmách, majetkové a personálne prepojenia |  |
| Reuters / Refinitiv | Najväčší poskytovateľ multimediálneho spravodajstva na svete | Používaný pre monitoring spravodajstva.  nie sú k dispozícii konkrétne požiadavky na integráciu |
| Tradeweb | elektronické mimoburzové trhy na obchodovanie s produktmi s pevným výnosom, ETF a derivátmi |  |
| Macrobond | Integrátor finančných a ekonomických dát | Ad-hoc dáta pre potreby ekonomických analýz pre OEM |
| CRN | Ceny rezidenčných nehnuteľností | XLS, štruktúrované dáta, cca 100 záznamov,  manuálne poskytované / kvartálne |
| United Classifieds | Spoločnosť poskytujúca online inzerciu (nehnuteľnosti) | Dáta o cenách nehnuteľností na sekundárnom trhu, štruktúrované dáta v CSV |

Tabuľka 8: Prehľad externých zdrojov – zdroje ekonomických informácií

#### Primárne systémy

Hlavné interné systémy a dátové zdroje relevantné pre dátovú integráciu sú nasledujúce:

| Označenie | Názov | Hlavná funkcia |
| --- | --- | --- |
| ŠZP | Štatistický zberový portál | Centrálny systém pre výkazy a hlásenia od subjektov finančného trhu pre štatistické účely a výkon funkcie dohľadu - core systém pre povinné výkazníctvo, zber dát od cca 7000 subjektov |
| ŠZP ESAS (Fujitsu XBRL) | Štatistický zberový portál (European Supervisory Authorities) | Nový systém pre výkazníctvo subjektov vo formáte XBRL (EIOPA, EBA) – v súčasnosti riešené samostatným modulom ŠZP |
| RBÚZ | Register bankových úverov a záruk | Lokálna implementácia AnaCredit - zber údajov o bankových úveroch a zárukách – prepojený s Registrom subjektov a systémom AnaCredit |
| RS | Register subjektov | Master systém pre subjekty, ktorých úvery sú vykazované cez RBÚZ (právnické osoby a podnikatelia) – založený na dátach ŠÚSR, RPO, RO a ďalších zdrojoch |
| Subjekty 2.0 | Subjekty 2.0 | Systém pre správu statických údajov o subjektoch finančného trhu pre účely dohľadu |
| REGFAP | Register finančných agentov a finančných poradcov | Systém pre evidenciu dohliadaných subjektov a pre zber dát a hlásení o ich činnosti. Evidenčné informácie sa plánujú integrovať do Subjekty 2.0, zber dát do ŠZP |
| CERI | CERI - Centrálna evidencia regulovaných informácií | Databáza regulovaných informácií, ktorými sú emitenti cenných papierov - neštruktúrované informácie - prospekty k emisiám cenných papierov, oznámenia, výročné správy, pre reporting sa požadujú iba metadáta o dokumentoch pre štatistiky |
| Register OFV |  | Odborná spôsobilosť finančných agentov a poradcov - register osôb (cca 70.000 FO), ktoré absolvovali OFV alebo odbornú skúšku podľa zákona o finančnom sprostredkovaní |
| Skúšky |  | Databáza otázok na skúšky, cvičné testy, ich výsledky, termíny skúšok |
| SIPS | Platobný systém SIPS | Retailový platobný systém určený na spracovanie klientskych platieb prevádzkovaný Národnou bankou Slovenska |
| SAP FINU/HRO | Finančné účtovníctvo, hospodárenie a rozpočet, platobný styk | Implementácia na báze SAP s kompletnou kustomizáciou pre NBS (moduly SAP ECC v6.0 - účtovné jadro, SAP BCA - Bank Customer Accounts modul pre správu klientskych účtov, SAP FI, CO, IM, MM, SD, PO) |
| IDAS | Agendový systém dohľadu | Lotus Notes aplikácia pre podporu agendy dohľadu ODK, OPD a ODB pre menej významné banky |
| ASDR | Agendový systém dohľadu a regulácie | Tiež označovaný ako ISD (Informačný systém dohľadu) - *Systém vo vývoji* |
| IBFO/WSS | Investičné bankovníctvo a finančné obchodovanie | Kustomizovaná implementácia Wall Street Suite (WSS) pre podporu obchodných procesov banky – t.j. realizáciu a vysporiadanie obchodných operácií realizovaných na OBO a OVO, vrátane ich účtovania na OFR. |
| Makroek. DB | Makroekonomická databáza | Konsolidované dáta pre tvorbu modelov na oddelení makroekonomických analýz OEM - spoločné úložisko časových radov. Kópia sa publikuje na webe NBS pre verejný prístup. |
| Credit Manager |  | Aplikácie pre riadenie rizík na odbore ORR |
| Risk Manager |  |
| EZO | Euro zásoby a obeh | Funkcie pre správu emisií bankoviek a mincí, evidenciu stavu hotovosti v ekonomike a na účtoch NBS, peňažné toky medzi NBS a tretími stranami.  Interný vývoj |
| HFCS | Zisťovanie o financiách a spotrebe domácností | Dáta prieskumu o vývoji životných podmienok domácností – manuálny vstup - cca 6000 domácností |
| BLS | Bank Lending Survey | Dotazník o vývoji ponuky a dopytu na trhu úverov – manuálny vstup |

Tabuľka 9: Prehľad interných zdrojov – primárne systémy

V NBS sa využívajú tiež niektoré ďalšie systémy pre podporu primárnych procesov, ktoré ale neboli identifikované ako relevantné pre dátovú integráciu:

* **Target2-SK** – lokálny komponent systému Target2 prevádzkovaného Eurosystémom na spracovanie veľkých platieb, externý Target2 poskytuje tiež reportingovú službu CRSS postačujúcu pre potreby OPS
* **IMAS (**SSM Information Management System) – spoločná európska procesná platforma pre podporu dohľadových procesov vo významných finančných inštitúciách
* **IRDA** (IMAS Data Reporting and Analytics) - reportingová služba nad dátami IMAS

#### Integračná vrstva

Rôzne odbory využívajú rôzne platformy pre spracovanie dát z rôznych zdrojov. Dlhodobým cieľom je migrácia reportingu na zjednotenú platformu DWH a Data Lake.

| Označenie | Názov |
| --- | --- |
| Apoštol databáza | Analytická platforma odboru štatistiky – predstavuje databázu integrovaných dát (**MS SQL Server**), kde hlavným zdrojovým systémom je **ŠZP** a k nemu sa nahrávajú najrôznejšie ďalšie registre a manuálne vstupy. Nad touto databázou je sprevádzkovaný in-memory BI nástroj Qlik Sense, v ktorom je tiež implementované množstvo integračnej a výpočtovej logiky. |
| DWH Pilot | Hlavným cieľom pilotného projektu DWH/BI bolo vybudovanie infraštruktúry DWH/BI (**Oracle + MS Analysis Services + MS Power BI**) a vytvorenie riešenia pre základné procesy spojené so spracovaním dát DWH a analytickými výstupmi v podobe BI riešenia. Projekt, ktorý začal v 03/2020, sa zameral na spracovanie vybraných dát zo systémov **RBÚZ, RS a ŠZP**.  Pilotný projekt DWH/BI vytvoril základ pre budúce riešenie, rešpektuje štandardné prístupy pre budovanie dátových skladov a môže byť využitý ako základ pre ďalší rozvoj DWH. |
| Riskhouse | Riskhouse predstavuje vlastný DWH systém odboru ORR (**MS SQL Server)**, ktorého vývoj začal v roku 2012 a ktorý bol navrhnutý s klasickým prístupom a architektúrou pre budovanie dátových skladov na podporu procesov zverených odboru ORR. Hlavným zdrojom dát je systém **IBFO/WSS**, ktorý poskytuje iba slabú analytickú podporu. Predpokladá sa, že riešenie Riskhouse bude v strednodobom horizontu zachované, vzhľadom k tomu, že implementuje množstvo špecifickej biznisovej logiky pre účely riadenia rizík. |
| SAP BW | SAP Business Warehouse v 7.3 predstavuje komodizovaný dátový sklad (**Oracle**), ktorý je dodávaný ako samostatný modul balíka podnikových aplikácií SAP. V NBS je nasadený SAP BW najmä pre reportingové potreby OFR, pričom jeho kontrolingová časť je kustomizovaná pre vykazovanie podľa ESCB metodiky COMCO.  V súvislosti s plánovaným prechodom na S4HANA bude treba zvážiť alternatívny prechod na BW4HANA, alebo pokrytie požiadaviek jednotným DWH systémom. |
| AŠI | Databáza (MS SQL Server) o subjektoch verejnej správy, vytvorená najmä ako dátová základňa pre analytické spracovanie **údajov zo Štátnej pokladnice** vnútri odboru OEM, oddelenia fiškálnych a štrukturálnych analýz. Databáza nebola koncipovaná ako DWH. |
| Makroek. databáza | Úložisko v **MS Excel**, v ktorom sú konsolidované dáta pre tvorbu modelov na oddelení makroekonomických analýz OEM. Databáza nebola koncipovaná ako DWH, ale spoločné úložisko časových radov.  Pre prístup sa využívajú rôzne nástroje (Matlab, EViews, WinSolve, BEAR, R Studio, Python), pomocou ktorých sa aj generujú výstupy pre management a analytické správy. Databáza sa tiež kopíruje na web NBS pre prístup zo strany verejnosti. |

Tabuľka 10: Prehľad integračných platforiem v NBS

#### Prezentačná / Analytická vrstva

Pre prístup k dátam v jednotlivých integračných platformách sa využívajú rôzne nástroje:

| Prezentačná / Analytická vrstva | Integračná vrstva | Názov |
| --- | --- | --- |
| Qlik Sense | Apoštol databáza | Qlik Sense je self-service BI systém s in-memory analytikou a schopnosťami stavať analytické riešenia nad rôznymi dátovými zdrojmi. Aj v rámci riešenia Apoštol sa silne využíva skriptovanie pre transformácie dát a dodatočné integrovanie externých a manuálnych vstupov.  Prípravu analytických priestorov a dashboardov zabezpečuje odbor OST. |
| MS Analysis Services | DWH Pilot | Dáta z data martov pilotného DWH sú spracúvané do tabulárnych modelov na MS SQL Serveri (in-memory OLAP), ku ktorým sa pristupuje cez MS Power BI klienta.  Spracovanie dát zabezpečuje OIT. |
| MS Reporting Services | Riskhouse | Prístup k dátam v integrovanej databáze Riskhouse je zabezpečený cez MS Reporting Services, ktoré sa využívajú pre operatívny reporting pre bežných užívateľov s obmedzenými možnosťami ďalšej manipulácie.  Reporty pripravuje a dokumentuje odbor ORR, ktorý sa tiež stará o celé riešenie Riskhouse. |

Tabuľka 11: Analytické vrstvy v NBS

#### Aplikačná vrstva BI

Koncové nástroje pre prístup k dátam, reportom a dashboardom boli zavedené v súvislosti s podkladovými analytickými systémami:

| Prezentačná / Analytická vrstva | Integračná / Analytická  vrstva | Názov |
| --- | --- | --- |
| Browser (webový prehliadač) | Apoštol databáza / Qlik Sense | Analytické dashboardy sa otvárajú cez tenkého klienta (webový prehliadač) |
| MS Power BI  (+ Power BI Report Server) | DWH Pilot / MS Analysis Services | Dáta tabulárnych modelov sú otvárané z klienta MS Power BI, vytvárané reporty sú publikované do MS Power BI Report Servera. |
| Browser (webový prehliadač) | Riskhouse / MS Reporting Services | Prístup k reportom publikovaným na MS Reporting Services cez webový prehliadač. |
| MS Excel  (Business Explorer add-in) | SAP BW | Prístup k dátam je realizovaný cez Business Explorer (Excel plug-in), ktorý umožňuje získanie dát zo SAP BW do Excelu, ale bez možnosti ďalšieho formátovania výstupov a tvorby grafov. Tieto výstupy sa robia následne v MS Excel ručne. |
| Štatistické nástroje | AŠI | Pre prístup sa využívajú rôzne nástroje (Matlab, EViews, WinSolve, BEAR, R Studio, Python), pomocou ktorých sa aj generujú výstupy pre management a analytické správy (vo formáte pdf a MS Excel s tabuľkami a grafmi) |
| Makroek. databáza |

Tabuľka 12: Aplikačná vrstva BI v NBS

### Pilot DWH

Hlavným cieľom pilotného projektu DWH/BI bolo vybudovanie infraštruktúry DWH/BI, zadefinovanie pojmov a základného data governence. Hlavným cieľom bolo otestovanie rôznych technológií a ich vhodnosť pre prostredie NBS. V neposlednom rade išlo o vytvorenie riešenia pre základné procesy spojené so spracovaním dát DWH a analytickými výstupmi v podobe BI riešenia.

Projekt, ktorý začal v 03/2020, sa zameral na spracovanie vybraných dát zo systémov ŠZP, RBUZ a RS. V súčasnosti zahŕňa asi 5% rozsahu dát NBS.

#### Logická architektúra „DWH Pilot“

Diagram

Description automatically generated

Obrázok 6: DWH – Logická architektúra

Pre účely DWH bola zvolená 3-vrstvová architektúra, ktorá pozostáva zo Stage vrstvy (L0), Integračnej (L1) a Prezentačnej vrstvy (PL). L1 a PL sú modelované prístupom R. Kimball, pričom v L1 vrstve sú dáta perzistované a PL vrstva obsahuje dáta perzistované (oblasť ŠZP) ako aj pohľady nad L1 vrstvou.

V rámci PL vrstvy boli logicky vytvorené viaceré star schémy nad objektami L1 vrstvy. SSAS tabulárny model obsahuje objekty štrukturálne rovnaké ako sú objekty v PL vrstve a pomocou technického účtu si dáta preberá 1:1 z PL vrstvy. PL vrstva slúži aj pre zabezpečenie prístupov, či už riadkovú ako aj stĺpcovú.

#### Integrované dáta

Zdrojové systémy NBS Pilot sú:

* ŠZP (**Štatistický zberový portál**) - Centrálny systém pre výkazy a hlásenia od subjektov finančného trhu pre štatistické účely a výkon funkcie dohľadu - core systém pre povinné výkazníctvo, zber dát od cca 7000 subjektov. Z toho integrované časti sú:
  + Menová a finančná štatistika
  + Cenné papiere
  + Bilančná trieda
* RBÚZ (**Register bankových úverov a záruk**) - Lokálna implementácia AnaCredit - zber údajov o bankových úveroch a zárukách – prepojený s Registrom subjektov a systémom AnaCredit. Do DWH sa prenášajú:
  + Údaje nástroja
  + Údaje protistrana – nástroj
  + Údaje spoločných záväzkov
  + Celkom cca 50% atribútov.
* RS (**Register subjektov**) - Master systém pre subjekty, ktorých úvery sú vykazované cez RBÚZ (právnické osoby a podnikatelia) – založený na dátach ŠÚSR, RPO, RO a ďalších zdrojoch. Do DWH sa preberajú všetky autoritatívne záznamy.

Tieto zdrojové systémy z pohľadu komplexnosti poskytnutých informácií celej NBS boli v projekte „NBS DWH Pilot“ integrované na cca 20%.

#### Dátové oblasti L1 vrstvy

L1 vrstva obsahuje tieto star schémy

| Zdroj | Star schéma | Obsah |
| --- | --- | --- |
| RBUZ | Inštrument a finančné dáta | Statické a finančné údaje o inštrumentoch |
| RBUZ/RS | Subjekty | Agregované finančné údaje z RBUZ na úrovni subjektu vystupujúceho v roli debtor (dlžník), informácie o subjektoch. |
| ŠZP | Trieda MFŠ | Menová a finančná štatistika dohľadaných subjektov. Star schéma je organizovaná vo forme key-value, kde key atribút určuje významovosť hodnoty |
| ŠZP | Trieda bilancie | Bilančné výkazy dohľadaných subjektov. Star schéma je organizovaná vo forme key-value, kde key atribút určuje významovosť hodnoty |
| ŠZP | Trieda Cenné papiere | Výkazy cenných papierov dohľadaných subjektov. Star schéma je organizovaná vo forme key-value, kde key atribút určuje významovosť hodnoty |
| DWH (MFŠ, RBUZ/RS) | Konsolidovaný model MFŠ-RBÚZ | Konsolidácia údajov zo štatistickej triedy MFŠ a RBUZ/RS. Porovnanie nominálnej hodnoty vykázanej ako agregát v ŠZP a napočítaného údaju z granulárnych dát RBUZ/RS podľa BSI metodiky (ECB). |
| RBUZ/RS | Prepojenie subjektu a inštrumentu | Prepojenie inštrumentu na subjekt vystupujúci v konkretnej roli (debtor, creditor...), údaje spoločných záväzkov a nominálnej hodnoty napočítanej podľa BSI štatistiky (ECB). |

Tabuľka 13: Star-schémy DWH – vrstvy L1

#### Technická architektúra „DWH Pilot“

Graphical user interface

Description automatically generated

Obrázok 7: DWH – Technická architektúra

Ako základ pre DWH bola zvolená databáza od firmy Oracle. Pre BI riešenie bolo zvolené Microsoft SQL Server Analysis Services s reporting platformou Power BI.

V rámci DWH existujú 2 prostredia:

1. DEV používané ako na vývoj tak aj testovanie a
2. samotné produkčné prostredie.

Každé prostredie má samostatnú databázu a samostatný hardware.

Spracovanie zdrojových dát prebieha nasledujúcim spôsobom:

1. Dáta sa extrahujú cez databázovú linku Oracle->Oracle cez extrakčné pohľady, pričom z nich si berie full množinu dát.
2. Dáta sú prenesené do temporárnych tabuliek v DWH.
3. V L0 schéme sú však za daný deň uložené full snímky ako aj inkrementálne snímky, ktoré sú tvorené porovnaním 2 po sebe idúcich full snímkov. Tabuľky v L0 schéme sú organizované podľa zdrojového systému a samotného extrakčného pohľadu. Výnimku tvorí hlavná tabuľka výkazov ŠZP, ktorá obsahuje cez 1 mld záznamov. Z tejto tabuľky sú dáta ukladané do osobitných, štrukturálne rovnakých tabuliek podľa samotných tried. Nedochádza k transformácii dát zo zdrojových systémoch.
4. Ako posledný krok beží vloženie všetkých chýbajúcich záznamov tak, aby inkrementálne snímky zachovali referenčnú integritu.
5. Integrácia dát do L1 vrstvy je zabezpečená pomocou PL/SQL procedúr, ktorých úlohou je:

* Vygenerovanie umelých kľúčov pre príslušnú tabuľku
* Transformovanie dát do perzistentných LD tabuliek, ktoré sú obrazom tabuliek L1.
* Historizácia dát z LD tabuliek do L1 v závislosti na typu tabuľky

Na spracovanie dát v rámci DWH (microscheduling) sa používa custom workflow nástroj vyvinutý dodávateľom pilota.

Na spracovanie dát do tabulárnych modelov SSAS sa používa SQL Server Integration Services.

Dostupná dokumentácia k pilotnému DWH/BI

* Základné princípy DWH/BI - Teoretický materiál
* Funkčný návrh DWH/BI - Logický a funkčný návrh riešenia PL vrstvy, popisuje zvolenú metodiku modelovania podľa R.Kimballa, popis star schém (datamartov), fyzický návrh modelu PL vrstvy
* Architektúra DWH/BI - Návrh integrácie okolitých systémov, ETL architektúra a workflow manažment, BI a reporting
* Technické požiadavky na zabezpečenie DWH/BI - Popis hardware a software požiadaviek na zabezpečenie serverových komponentov DWH a BI, špecifikácia prostredí
* Menné konvencie
* Technické riešenie DWH - Detailný technický popis jednotlivých vrstiev DWH/BI, ETL objektov, logovaní a zabezpečenie DWH/BI

#### HW špecifikácia NBS DWH pilot

| HW komponent | Typ | Parametre |
| --- | --- | --- |
| DB server | Výrobca/ verzia | Oracle 19c EE + Partitioning option |
| Operačný systém | Oracle Linux 8 (RedHat) |
| RAM (Typ, počet) | 512GB, DB má dedikované SGA/PGA na úrovni 50GB/40GB |
| CPU (Typ, počet) | 1xCPU Xeon Gold 6244 (8 cores), 3,6 GHz; 1 slot ešte voľný |
| Súčasná utilizácia CPU | v maxime 30% |
| Veľkosť DB | 2TB (dáta a indexy) |
| Diskové pole (Typ HDD, počet, veľkosť) | 14x 3TB SSD lokálne disky, zapojené do RAID6, cez PCI RAID radič, je možné zriadiť prístup aj na centrálne diskové pole 3PAR, lebo server má 2x 16 Gbit FC karty |
| LAN na diskové pole (rýchlosť, počet) | - |
| LAN do siete (rýchlosť, počet) | 4x 10GBit |
| BI server | Výrobca/ verzia | MS SQL Server 2019 Enterprise |
| Operačný systém | Windows Server 2022 Data Center edícia. |
| RAM | 32 GB |
| CPU | 4 vCPU |
| Diskové pole | 3PAR |
| Aplikačný server | Výrobca/ verzia | Apache Tomcat/9.0.56 |
| Operačný systém | Red Hat Enterprise Linux 7.9 |
| RAM | 10 GB |
| CPU | 6 vCPU, Intel(R) Xeon(R) Gold 6134 CPU @ 3.20GHz |
| HDD | 25 GB |

**Poznámka:** na HW DB nebeží len DWH Pilot DB, ale ďalšie produkčné databázy NBS. Súčasné vyťaženie CPU je na úrovni 30%. Pilot DWH obsahuje cca 2TB dát.

#### Reporting

Projekt pilota zaviedol BI infraštruktúru pre reporting postavenú na nástroju MS Power BI a MS Power BI Report Server.

## Technické požiadavky

### Úvod

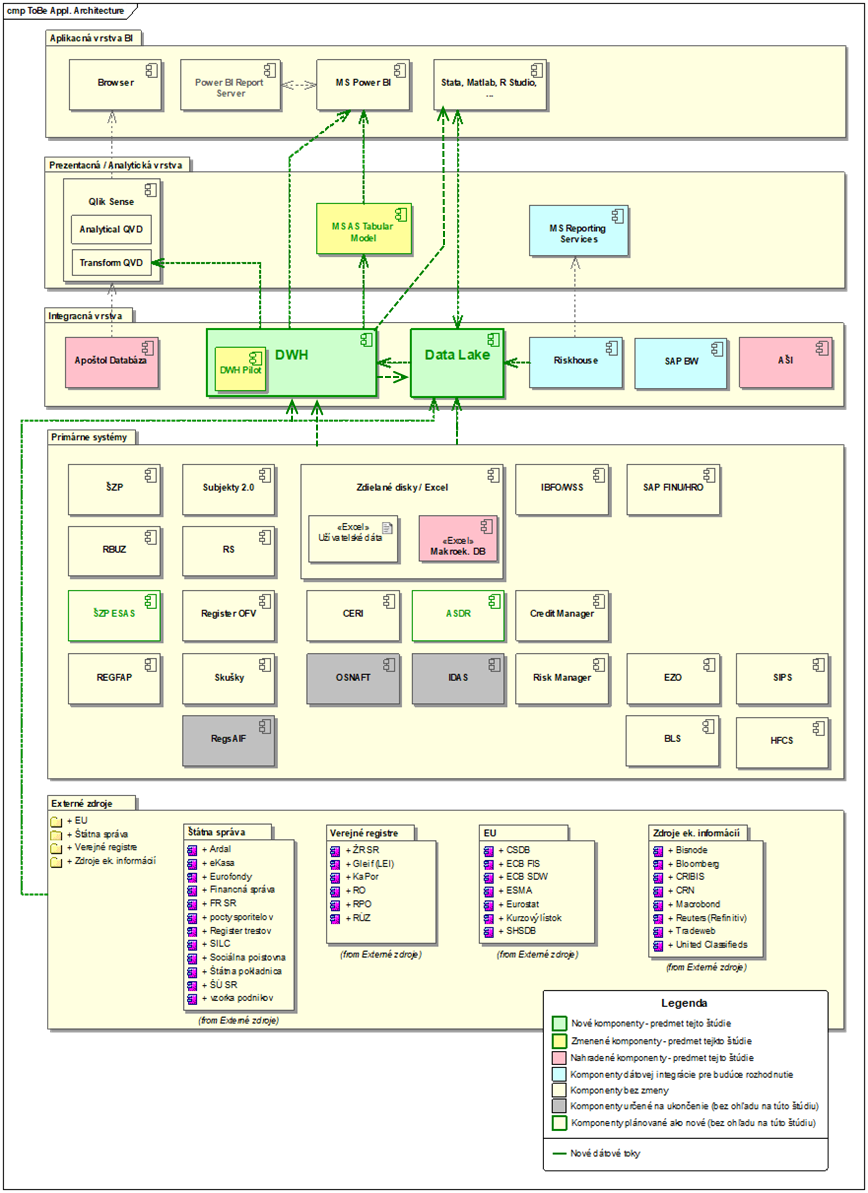
Táto kapitola obsahuje technické požiadavky na nový systém NBS DWH. Požiadavky sú rozdelené do jednotlivých oblastí ako:

1. Architektúra (logická, fyzická a aplikačná)
2. Popis vlastností jednotlivých komponentov na úrovni relačnej databázy a DataLake
3. Popis aplikačných vlastností podporných aplikácií DWH.

Rovnako sa táto kapitola zaoberá bezpečnostnými požiadavkami a požiadavkami na spustenie systému (iniciálny load).

### Požadovaná architektúra aplikačnej vrstvy práce s dátami

Plánované zmeny zachytáva nasledujúci diagram, ktorý vychádza zo schválenej dátovej stratégie bankovou radou NBS v Apríli 2022:



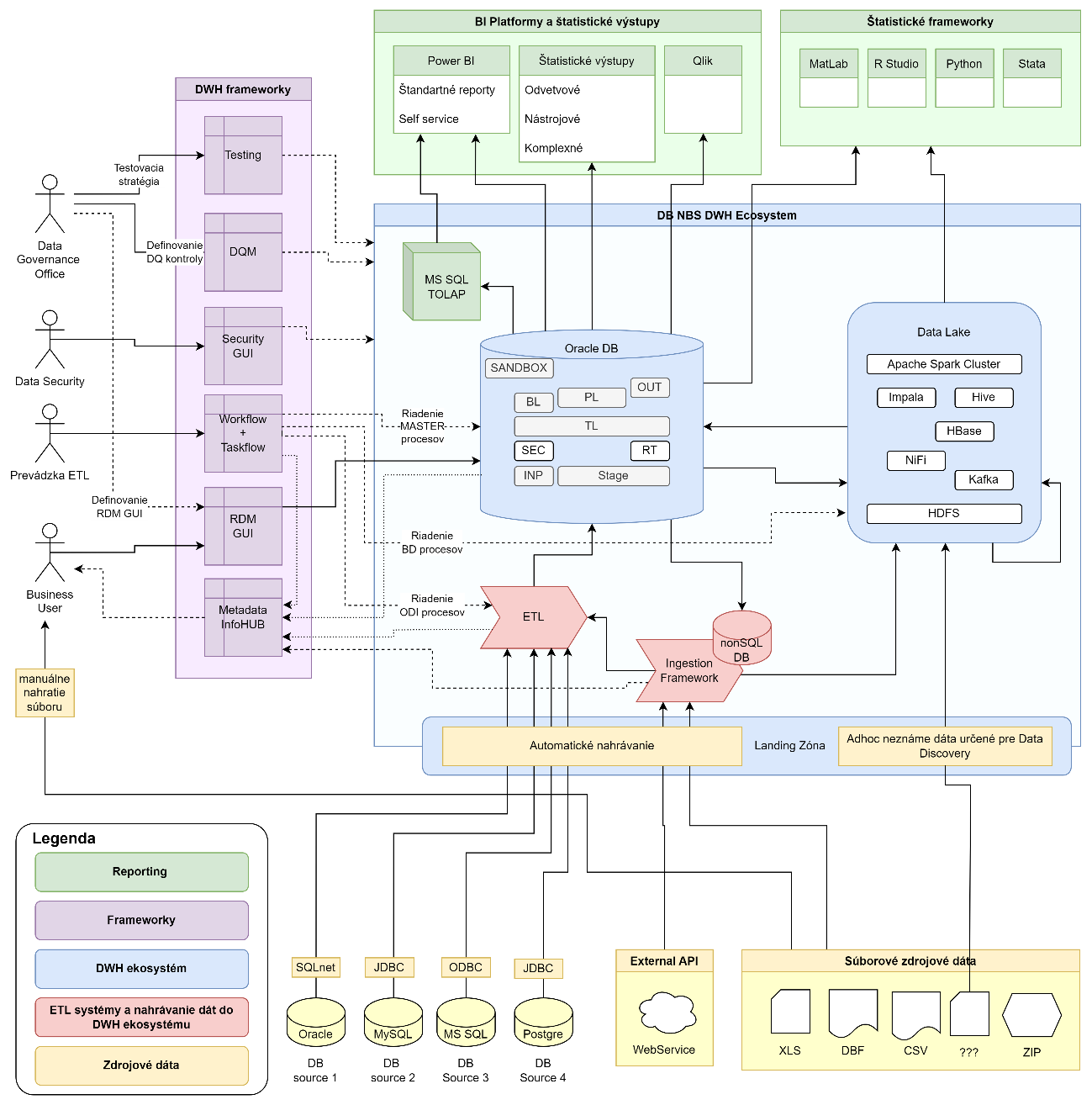
Obrázok 8: Aplikačná vrstva – budúci stav

### Logická architektúra

Riešenie novej dátovej integrácie je navrhnuté ako jeden dátový ekosystém v kombinácii dvoch platforiem:

* Data Lake
* Relačná časť, ktorý sa skladá z:
  + Oracle DB, ktorá rieši problematiku dát ako je atomicita, konzistencia, izolovanosť, trvanlivosť, či unikátnosť
  + MS SQL DB s in-memory T-OLAP podporou pre PowerBI reporting.

Každá z týchto platforiem nie je osobitnou aplikáciou, ale súborom technológií, nástrojov a na mieru naprogramovaných procesov a integrácií, ktoré vo svojom celku plnia úlohu dátovej integrácie a spracovania dát pre účely reportingu a pokročilých analýz. Z tohto dôvodu sú v ďalšom texte obe platformy ďalej dekomponované na jednotlivé dátové vrstvy, dátové toky, funkčné bloky a použité technológie.



Obrázok 9: Logická architektúra nového riešenia

Predmetom riešenia sú tieto komponenty architektúry:

* **Relačná časť** – Dátový sklad, jeho dátové vrstvy, dátové toky, metadátové nastavenia a podporné frameworky
* **DL** – Data Lake, jeho dátové vrstvy, dátové toky, metadátové nastavenia a podporné frameworky
* **Spoločné frameworky** – aplikácie a nástroje, ktoré nie sú súčasťou samotných platforiem dátovej integrácie, ale využívajú sa pre celkové riadenie spracovania dát a podporu data governance procesov.

Nasledujúce časti aplikačnej architektúry sú dotknuté týmto riešením:

* **BI platformy** – existujúce BI aplikácie, ktoré budú v rámci projektu pripojené na DWH a Data Lake pre čítanie dát.
* **Štatistický framework** – existujúce portfólio štatistických nástrojov, pre ktoré budú zadefinované mechanizmy prístupu k DWH a DataLake
* **Zdrojové systémy s priamym prístupom** – reprezentuje všetky systémy, ktoré budú poskytovať priamy prístup ku svojím dátam, resp. cez vyhradené databázové rozhranie
* **Externé systémy poskytujúce extrakty** - reprezentuje všetky systémy, ktoré budú poskytovať extrakt svojich dát, ktorý bude vymieňaný s Data Landing zónou cez SFTP server alebo extranet Sharepoint.
* **Zdrojové systémy poskytujúce služby** - reprezentuje všetky systémy (interné aj externé), ktoré budú poskytovať svoje dáta prostredníctvom vystavenej služby (REST API, SOAP).

### Fyzická architektúra

Navrhnutý datawarehouse bude pozostávať z rozširovateľných a otvorených komponentov. Všetky logické komponenty sú prevádzkované formou „On-Premise“ v dátových centrách NBS ako fyzické či virtuálne servery.

Fyzicky bude infraštruktúra rozvrhnutá nasledovne na 4 skupiny:

* Relačná časť
* DataLake
* Aplikačný server pre podporné aplikácie
* BI oblasť

Nasledujúca tabuľka obsahuje súpis hardvérových požiadaviek.

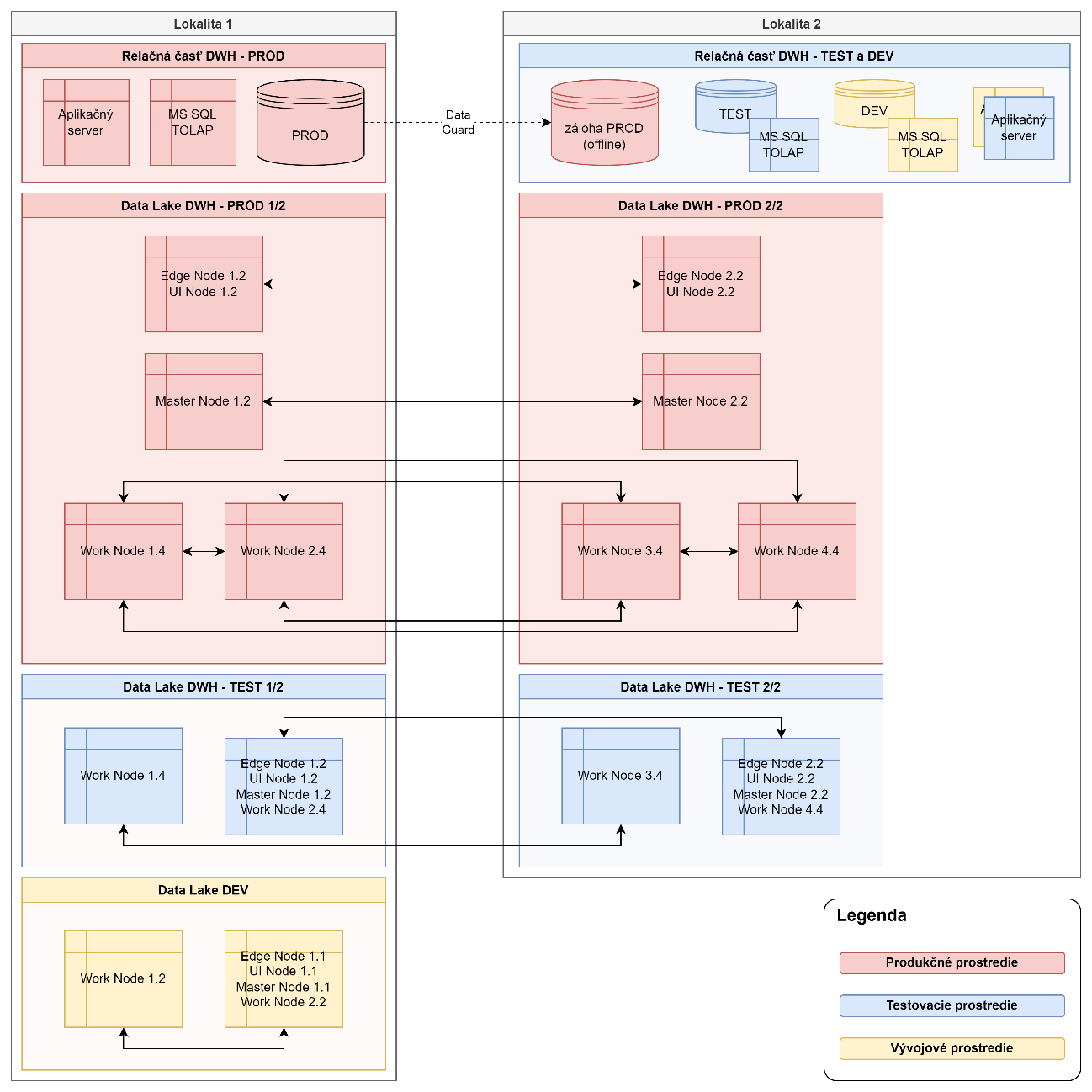
| Platforma | Uzly | Navrhované riešenie |
| --- | --- | --- |
| Relačná časť | PROD-DB  TEST-DB  DEV-DB | DWH NBS bude nasadený na 2 samostatné servery v oddelených lokáciách s rovnakým výkonom.  Nová dohodnutá konfigurácia:   * 2x fyzický server, (na 2) a o 512GB RAM (na 1TB).   + CPU: 1x X86 CPU 8 cores s potencionálnym rozšírením o ďalší CPU   + RAM: 512 GB RAM s potencionálnym rozšírením na 1 TB   + SSD: 4x3,2TB v RAID s potencionálnym rozšírením * Na serveroch bude inštalovaná Oracle 19c EE + Partitioning option * Pre zabezpečenie vysokej dostupnosti bude vytvorená standby produkčná inštancia cez Oracle Data Guard na vývojový server. * Na serveri budú ďalej bežať RunTime ETL nástroja Oracle Data Integrator. |
| Data Lake | Worker Node  Master Node  UI Node  Edge Node | Pre Data Lake bude potrebné vytvoriť celkovú infraštruktúru cez 3 prostredia. V každom prostredí bude vytvorená sieť výpočtových uzlov, na ktorých pobežia Big Data services.  Vzhľadom na predpokladanú veľkosť dát v NBS s funkciou vysokej dostupnosti sa očakáva vybudovanie infraštruktúry vo virtuálnom prostredí NBS. Architektúra Data Lake má charakter vysokej dostupnosti a nebude potrebné budovať standby inštanciu, ale použije sa roztiahnutie produkčných uzlov cez obe lokality.  Produkčné uzly:   * 4x Worker Node (2+2 v lokalitách)   + CPU: X86 8 cores   + RAM: 64GB   + Disk: 20 TB * 2x Master Node (1+1 v lokalitách)   + CPU: X86 4 cores   + RAM: 32GB   + Disk: 1,5 TB * 2x Edge Node+UI Node (1+1 v lokalitách)   + CPU: X86 4 cores   + RAM: 32GB   + Disk: 1,5 TB   Test uzly:   * 2x Worker Node (1+1 v lokalitách)   + CPU: X86 8 cores   + RAM: 64 GB   + SSD: 480 GB pre OS   + Disk: 4 TB s možnosťou pridať ďalšie disky * 2x Master Node+ Edge Node+UI Node (1+1 v lokalitách)   + CPU: X86 4 cores   + RAM: 32GB   + Disk: 2,5 TB   Vývojové uzly:   * 1x Worker Node (1 v lokality)   + CPU: X86 4 cores   + RAM: 32 GB   + Disk: 3 TB s možnosťou pridať ďalšie disky * 1x Master Node+ Edge Node+UI Node (1 v lokalite)   + CPU: X86 2 cores   + RAM: 32GB   + DIsk: 3 TB |
| Aplikačný server |  | Aplikačný server na beh aplikácií (frameworkov DWH). Tento komponent je podobne ako BI server tvorená virtualizačnou platformou.   * Apache Tomcat/9.0.56 aplikačný server na Red Hat Enterprise Linux prostredí * CPU: 4 vCPU * RAM: 10 GB * Disk: 25 GB   Vysoká dostupnosť (failover) je riešená na úrovni virtualizačnej platformy. |
| BI | PROD-BI  DEV-BI  TEST-BI | BI servery sú/budú virtualizované na platforme VMware ESX, virtualizačná platforma je v správe IT NBS. Výkon je možné prideliť podľa potreby.   * MS SQL Server 2019 Enterprise * Windows Server 2022 Data Center edícia. * CPU: 4 vCPU * RAM: 32 GB   Vysoká dostupnosť (failover) je riešená na úrovni virtualizačnej platformy. Budú dve platformy:   1. PROD-DB (produkčný BI server) 2. DEV-DB (vývojový BI server) + TEST-BI (testovací BI server) |
| PC | | PC | Reprezentuje pracovné stanice koncových užívateľov  Inštalované komponenty podľa potreby.   * MS Power BI Pro – pre vývojára reportov a analytických dashboardov * MS Visual Studio, resp. SQL Server Data Tools - pre vývojára ETL pre spracovanie tabulárnych modelov v MS SQL Server Analysis Services * Oracle client pre vývojára ETL, resp. reportov * Oracle SQL developer pre vývojára ETL * Oracle Data Integrátor pre vývojára ETL |

Tabuľka 14: Prehľad HW komponent riešenia

NBS preferuje využitie už vytvorenej infraštruktúry, ale táto infraštruktúra by mala byť dodávateľom cieľového riešenia posúdená z pohľadu kapacity a dostupnosti s prípadným dopadom na SW licencie.

Dodávateľ túto infraštruktúru zreviduje a navrhne dodatočné hardware komponenty pre prevádzkovanie cieľového riešenia v 3 prostrediach, vrátane platformy DataLake.

Z hľadiska vytvorenia vysokej dostupnosti bude fyzická architektúra nastavená nasledovne:

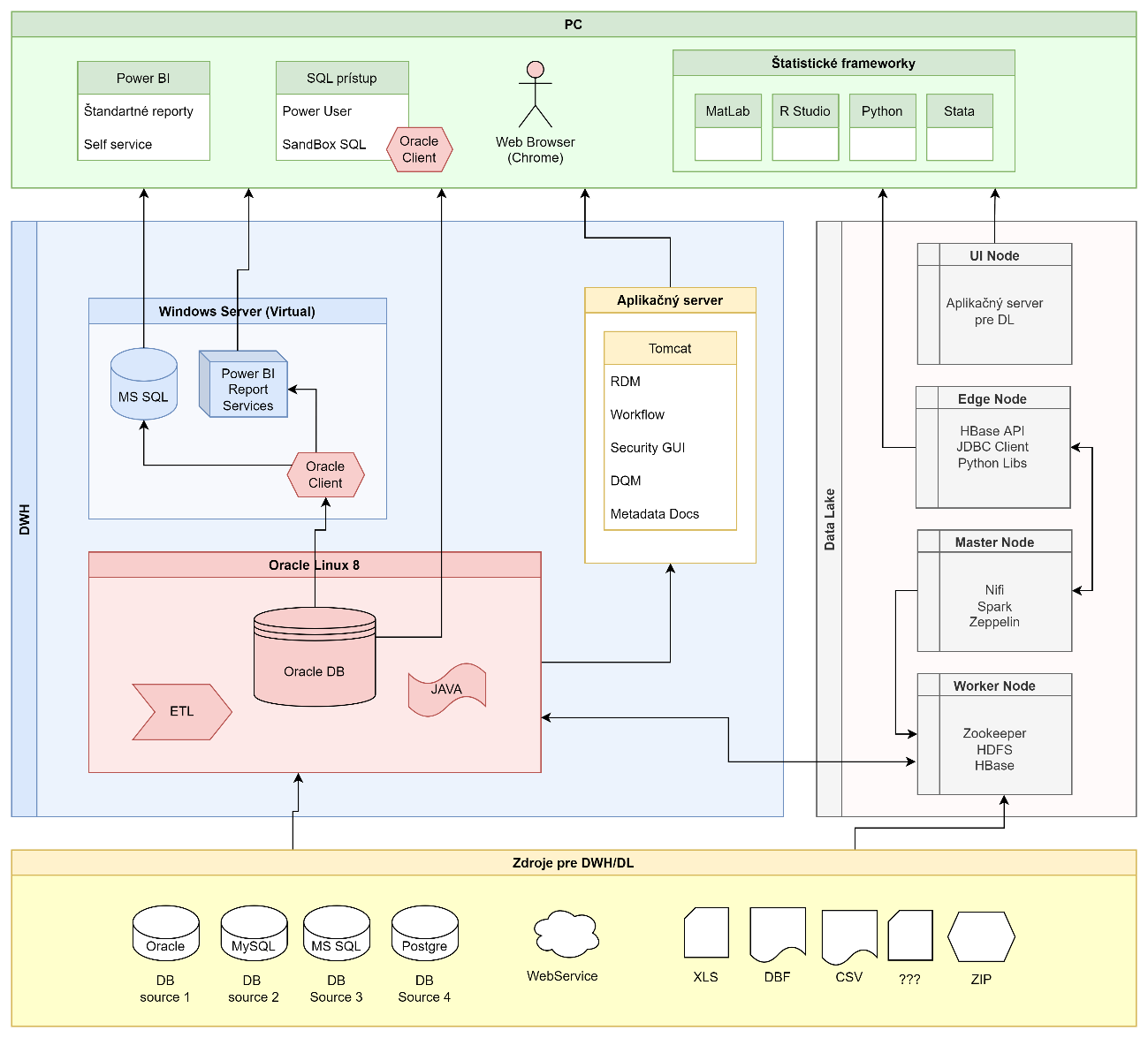


Obrázok 10: Vysoká dostupnosť nového riešenia

### Aplikačná architektúra

#### Úvod

Nasledujúci infraštruktúrny diagram popisuje navrhovanú novú aplikačnú architektúru NBS DWH z  hľadiska inštalácie jednotlivých produktov.



Obrázok 11: Aplikačná architektúra nového riešenia

Navrhované riešenie tvorené:

1. relačným DWH a BI platformou bude prevádzkovaná s využitím existujúcej aplikačnej infraštruktúry pilotného DWH/BI.
2. Data Lake, bude potrebné vybudovanie novej aplikačnej infraštruktúry.

V prípade DWH je základným software komponentom databázový systém a systém pre ETL. NBS preferuje využiť existujúcu investíciu do pilotného DWH, ktorý je postavený na:

* Databáza – Oracle DB v edícii Enterprise
* Reporting – MS SQL Server Analysis Services a MS Power BI

Navyše je v tomto projekte požadované vytvorenie Data Lake prostredia ako oblasť pre uloženie dát pre data discovery.

Jednotlivé komponenty riešenia budú vyžadovať nasledujúci software, a to minimálne v 3 oddelených prostrediach (vývoj, test, produkcia) v súlade s detailnými požiadavkami:

| Platforma | Software | Navrhované riešenie |
| --- | --- | --- |
| DWH | Operačný systém | Výber operačného systému je podriadený voľbe databázy, prípadne podkladovej hardware infraštruktúry.  V prípade existujúceho pilotného DWH sa využíva **Oracle Linux 8** (RedHat distribúcia). |
| Databáza | Existujúci pilotný DWH je postavený na relačnej databáze **Oracle 19c Enterprise Edition + Partitioning Option**. |
| Doplnkové časti DB | Súčasné nastavenie DB a vybrané opcie (**Partitioning Option**) sú dostatočné a neplánuje sa úprava. |
| Data Lake | Operačný systém | Výber operačného systému pre Data Lake bude podriadený voľbe Big Data platformy, ktorá musí mať pre daný operačný systém podporu.  NBS preferuje **Red Hat Linux**. |
| Big Data platforma | Výber Big Data platformy je predmetom výberového konania. |
| Platform Management | Výber Platform Management systému je predmetom výberového konania. |
| ETL systém | SQL ETL | Ako ETL framework je preferovaný nástroj **Oracle Data Integrator** |
| Big Data ETL | Výber Big Data platformy je predmetom výberového konania. Podmienka NBS je použitie Enterprise riešenia pred čistým OpenSource. |
| Ingestion framework | Projekt NBS DWH nebude o veľkosti spracovávaných dát, ale o počte rôznorodých dátových vstupov a vzorov. Denne/mesačne budú prichádzať stovky rôzne štruktúrovaných súborov, ktoré sa počas životnosti môžu a budú meniť. Preto požadujeme dodať v projekte systém na metadátové riadené spracovanie súborov. Bližšie v kapitolách 3.2.13. |
| BI | Operačný systém | Výber operačného systému je podriadený voľbe BI servera, prípadne podkladovej hardware infraštruktúry.  V prípade existujúceho pilotného BI servera, ktorý je postavený na technológiách Microsoft a virtualizačnej platforme VMware ESX, sa využíva **Windows Server 2022 Data Center edícia**.  Aktuálne je zalicencovaná celá virtualizačná platforma, prípadné rozšírenie kapacity BI je preto bez dopadu na potrebu ďalších licencií OS. |
| BI server | NBS využíva **MS SQL Server 2019 Enterprise** pre BI riešenie, konkrétne **Analysis Services** (tabulárne modely) pre implementáciu dátových kociek a **MS Power BI Report Server** pre publikáciu a zdieľanie reportov a analýz.  Na odbore štatistiky je to ďalej **Qlik Sense** – banka však nepredpokladá ďalšie rozširovanie tohto systému mimo odbor štatistiky.  Tabulárne modely v pilotnom režime boli vyvíjané na odboroch OFI a OST.  Aktuálne je zalicencované a prevádzkované 1 produkčné a 1 vývojové prostredie (2x 4 jadra). |
| BI klient | NBS využíva **MS Power BI** (On-premise) pre tvorbu reportov a plánuje tento nástroj naďalej využívať ako jednotnú aplikáciu pre reporting a analýzy.  Aktuálne je k dispozícii 13 licencií MS Power BI Pro. |
| Štatistický framework | Štatistické nástroje používané analytickými oddeleniami sú rôzne a používajú sa vždy podľa konkrétneho účelu. Tento štatistický framework zahŕňa nástroje Matlab, R studio, Eviews, Stata, jazyky SQL, Python, Scala a pod.  Predmetom projektu nie je dodávka ďalších nástrojov alebo licencií. |
| AI/ML | Nástroje pre umelú inteligenciu a Machine Learning sa zatiaľ nepoužívajú. Platforma Data Lake by však mala vytvoriť predpoklad pre ich budúce nasadenie.  Predmetom projektu nie je dodávka licencií, resp. infraštruktúry pre AI/ML. |
| Aplikačné frameworky | Workflow Manager | Pre orchestráciu spracovania cez rôzne platformy bude zavedený jednotný nástroj.  Výber konkrétneho nástroja je súčasťou výberového konania. |
| Dátové modelovanie | Nástroj pre dátové modelovanie je potrebný pre zabezpečenie celého procesu vývoja dátovej integračnej platformy.  Výber konkrétneho nástroja pre dátové modelovanie je súčasťou výberového konania. |
| RDM  DQM  Metadata Management | Nástroje Data Governance pre správu referenčných dát (RDM), Data Quality Management (DQM), a Metadata Management ako sú popísané v kapitolách 3.2.14  Výber konkrétnych nástrojov je súčasťou výberového konania. |

Tabuľka 15: Software komponenty riešenia

#### Relačná databáza

Existujúci pilotný DWH je postavený na relačnej databáze Oracle 19c Enterprise Edition + Partitioning Option. Návrh nového NBS DWH počíta s prepoužitím už existujúcich komponentov.

**Oracle DB** prináša robustnosť a širokú škálu funkcionality pre DWH riešenia. DB je optimalizovaná pre spracovanie veľkého objemu dát pomocou hromadných operácií, s využitím partícií a kompresie je možné efektívne využívať diskový priestor a znížiť nároky na diskové pole, a to ako z pohľadu kapacity, tak aj z pohľadu výkonnosti. Medzi ďalšie výhody Oracle DB patrí možnosť využívať bitmapové a funkčné indexy, ktoré spolu s DB paralelizmom umožňujú skracovať nielen ELT čas, ale aj zbiehanie užívateľských reportov. Vo svete DWH sa veľmi často využívajú analytické funkcie, ktoré umožňujú jednoducho vytvárať aj veľmi komplexné reporty bez potreby ukladania medzivýsledkov v DB. Z pohľadu security je veľmi užitočnou vlastnosťou Virtual Private Database, ktorá umožňuje riadiť prístup nielen na úrovni objektov, ale aj na úrovni riadkov a atribútov. Oracle Database Resource Manager umožňuje administrátorom riadiť prideľovanie zdrojov (CPU a IO operácií) pre jednotlivých užívateľov a procesy. Pre komplexné monitorovanie výkonu a optimálne využívanie DB slúži Oracle Enteprise Manager spolu s Automatic Workload Repository, ktoré poskytujú kompletný prehľad o využívaní zdrojov jednotlivými procesmi. Oracle DB v prípade potreby umožňuje jednoducho rozširovať funkcionalitu o ďalšie možnosti ako je napríklad: In-memory, OLAP DML, datamining.

**Aplikačné frameworky**, ktoré majú pomôcť zrýchliť vývoj DWH, zjednodušiť správu a spríjemniť technické riešenie pre netechnických používateľov. Z aplikačného hľadiska sú to (detailný popis v kapitolách 3.2.13 a 3.2.14):

* Nástroje pre podporu prevádzky
  + Workflow Management - Prostredie pre podporu správu, riadenie a monitoring aktualizačných úloh
  + Testovací framework
  + Prostredia na podporu nahrávania manuálnych dát – selfservice pre používateľov
  + Ingestion framework - Prostredie na podporu automatického nahrávania veľkého počtu rôznych súborov
  + Prostredie pre správu prístupov (DB a reportingová úroveň)
  + Prostredie na podporu riadenia dátovej retencie
* Nástroje pre Data Governance
  + Reference Data Management
  + Data Quality Management
  + Metadata management
  + Dátové modelovanie

**ETL framework**

ETL (Extraction – Transformation – Load) framework je základným subsystémom pre integráciu dát v DWH a jeho voľba do značnej miery určuje budúcnosť ďalšieho rozvoja a správy integračnej platformy.

ETL framework by mal najmä spĺňať tieto požiadavky:

* Poskytovať štandardizované prostredie pre vývoj, prevádzku a ďalší rozvoj DWH a integrácie dátových zdrojov a spracovanie dát
* Štandardizovať prístup k vývoju ETL kódu a minimalizovať technické chyby v implementácii
* Podporovať kolaboratívny prístup k vývoju, umožňovať spoluprácu vývojárov a vývojových tímov
* Poskytovať systematickú podporu pre verziovanie kódu a release management
* Poskytovať metadáta pre technickú data lineage
* Poskytnúť prostredie na aplikovanie templateov a prednastavených ETL častí. Vyžaduje sa podpora a prístup low-code, všetky parametre mapovania ETL musia byť uložené v metadátach. Samotné vytvorenie ETL kódu do SQL nastane kompiláciou contextového nastavenia topológie infraštruktúry, zvolenej technológie a znalostných templateov.
* ETL framework by mal mať API na hromadné nastavenie metadát.
* Poskytnúť možnosť vymeniť template ETL kódu bez nutnosti zmeny nastavenia samotného mapovanie, napr. pri ladení ETL.
* Ako nástroj vývoja spracovania dát by mal byť jednoducho použiteľný a na slovenskom trhu by mal byť dostatok kvalifikovaných IT špecialistov

Na základe hore uvedených požiadaviek je odporúčaný nástroj Oracle Data Integrátor.

Oracle Data Integrátor (ODI) umožňuje jednoducho a rýchlo integrovať údaje z rôznych technológií. Knowledge Module (KM), ktoré sa využívajú pri vytváraní rozhraní prinášajú na rozdiel od tradičných ETL nástrojov, možnosť využiť natívne vlastnosti danej technológie pri spracovaní údajov v jednotlivých krokoch extrakcie, transformácie a nahrávaní. Široké možnosti definovania fyzickej a logickej topológie a ich prepojení pomocou kontextu dávajú veľkú voľnosť pri zmene infraštruktúry bez potreby zásadných zásahov do ELT procesu. Napriek tomu, že KM sú dodávané spolu s produktom ODI je ich možné upravovať podľa potreby konkrétnej riešenej úlohy. Prostredníctvom ODI API je možné vytvárať mapovania ako aj aktualizovať ODI metadáta (topológiu, security, ODI model, atď.). Táto možnosť sa využíva predovšetkým pri hromadných zmenách, prostredníctvom API je možná hromadná zmena KM spolu s vygenerovaním ODI scenárov a ich nasadením do produkčného repozitára. Pri správnom používaní ODI je veľkou výhodou možnosť robiť impakt analýzu pri zmenách na zdrojových systémoch a rovnako pri dohľadaní zdroja, ktorý sa používa na aktualizáciu DWH objektov. Okrem toho, že ODI priamo podporuje vytváranie impakt analýz, zdokumentované ODI metadáta slúži na generovanie online ETL dokumentácie v rámci Data Governance stratégie.

#### Data Lake

**Data Lake** je súbor aplikačných komponentov, ktoré majú používateľovi priniesť :

* Distribuovaný súborový systém umožňujúci spoľahlivo uchovávať dáta v rôznych formátoch
* Nástroje umožňujúce vytváranie databáz, tabuliek, partícií a iných komponentov
* Distribuovaný výpočtový framework umožňujúci spracovanie nad dátami
* Grafický interface pre potreby správy platformy
* Granulárne riadenie prístupov do platformy auditnými funkciami
* Správa zdrojov platformy
* Nástroje umožňujúce prístup k dátam použitím tradičného SQL jazyka ako aj programovacích jazykov (R, Python, Scala, ...)

**Správa DataLake**

Data Lake platforma je zložená zo súboru technológií, ktoré medzi sebou komunikujú a zabezpečujú bezproblémový chod platformy. Každá technológia, ktorá je súčasťou platformy zastrešuje inú funkcionalitu, ktorá je od Data Lake očakávaná. Tieto technológie sú zväčša inštalované na viaceré lokácie z dôvodu zabezpečenia vysokej dostupnosti a robustnosti platformy.

Pre efektívne spravovanie platformy je vyžadované, aby existoval nástroj s grafickým rozhraním, ktorý ponúka možnosti na spravovanie a monitorovanie platformy. V prípade, že by nebol využitý nástroj na správu platformy, je nutné všetky zmeny vykonávať manuálne alebo polo-manuálne prostredníctvom zmien v konfiguračných súboroch, ktoré sa môžu veľa krát nachádzať na rôznych serveroch. Práve z tohto dôvodu vznikli nástroje, ktoré umožňujú jednoduchú správu Data Lake platformy, či už platforma pozostáva z troch serverov alebo z desiatok/stoviek.

Pomocou takéhoto nástroja bude možné jednoduchým spôsobom konfigurovať a sledovať stav platformy. Požiadavky na nástroj :

* Grafické rozhranie
* Podpora pridávania a modifikovania Data Lake služieb
* Podpora reštartovania služieb, prípadne celej platformy
* Podpora real-time monitorovania stavu platformy
* Podpora analýzy historických dáta z monitorovania behu v podobe grafov a metrík
* Podpora konfigurácie custom alertov na základe aktuálneho stavu platformy
* Možnosť upgradovať platformu priamo z nástroja
* Podpora správy užívateľov

**Storage (Distribuovaný súborový systém)**

Hlavným pilierom Data Lake platformy bude distribuovaný súborový systém, ktorý zabezpečuje robustné a spoľahlivé riešenie na ukladanie a následné procesovanie veľkého množstva dát. Distribuované súborové systémy Data Lake platformy budú optimalizované na čítanie a zápis veľkého množstva dát v akomkoľvek formáte.

Tieto technológie zväčša využívajú master/slave architektúru. Master server uchováva jednotlivé metadáta o súboroch a na slave serveroch sú rozdistribuované jednotlivé dátové bloky. Vysoká dostupnosť dát bude zabezpečená replikáciou metadát ako aj jednotlivých dátových blokov naprieč servermi Data Lake platformy. Replikácia dát zabezpečuje bezproblémový chod platformy aj v prípade výpadku viacerých diskov, poprípade celých serverov.

**ETL framework (Data Lake)**

ETL (Extraction – Transformation – Load) framework je základným subsystémom pre integráciu dát v Data Lake platforme obdobne ako pre DWH. ETL frameworky umožňujú jednoduchým spôsobom:

* integrovať externé dáta do platformy - štruktúrované, polo-štruktúrované aj neštruktúrované dáta (csv, xml, json, logy, pdf, obrázky …)
* efektívne spracovanie dát do finálnych štruktúr
* export dát do ďalších systémov.

Moderné Data Lake platformy ponúkajú rôzne prístupy riešenia ETL požiadaviek. V rámci budovania DataLake je potrebné zabezpečiť nasledujúce potreby:

* možnosť čítania dát z externých databáz pomocou štandardných JDBC/ODBC konektorov, SFTP serverov, distribuovaných file-systémov (HDFS, S3, ..)
* distribuované spracovanie dát pomocou štandardných SQL príkazov a syntaxe, ako aj možnosť použitia programovacích jazykov (Python, Java, Scala, R). Technológia by mala umožniť aj kombináciu oboch prístupov (SQL + programovací jazyk) v jednom ETL.
* možnosť exportovať napočítané dáta do štandardných relačných databáz ako aj NoSQL databáz. Taktiež možnosť exportovať dáta priamo do DataLake platformy či už vo forme súborov alebo vo forme dát v tabuľkách.

**Orchestrácia ETL úloh**

Každá úloha musí spustená práve v správnom čase a práve v presnej závislosti na iné úlohy. O tieto úlohy sa stará systém na orchestráciu. Je vyžadované:

* Každá úloha ETL frameworku DataLake-u musí logovať svoj stav do dohodnutej DB tabuľky
* Úlohy ETL DL bude možné zavolať cez WebService z Oracle prostredia aplikačného frameworku pre zabezpečenie globálneho riadenia. T.j. je vyžadované aby makro scheduller bežal na relačnej časti DWH. ETL DL bude riadený vzdialene ako ucelené bloky. Ich podčasti je možné kontrolovať mimo DWH makroscheduller, ale všetky časti bude možné skontrolovať z jedného miesta.
* ETL DL nebude plánované samostatne, ale len ako jeden celok DWH (relačná + DL časť)

**JDBC prístup**

Jednou z najdôležitejších požiadaviek na Data Lake platformu je možnosť napojenia štatistických a BI nástrojov na dáta, ktoré sú súčasťou platformy. Najštandardnejším spôsobom, ktorý je využívaný v BI a štatistických nástrojoch je práve spôsob dotazovania pomocou JDBC a ODBC technológií.

DataLake platforma musí poskytovať spôsob napojenia pomocou týchto technológií.

Požiadavky na JDBC/ODBC:

* Možnosť dotazovania pomocou JDBC/ODBC spojenia
* Možnosť granulárnej konfigurácie oprávnení pre JDBC/ODBC spojenia
  + Pre technických aj business používateľov
  + obmedzenia na čítanie/zápis len pre vybrané objekty
* Dotaz definovaný v JDBC/ODBC spojení musí byť distribuovane spracovaný DataLake platformou.

**Import súborových dát priamo do Data Lake**

Data Lake platforma by mala poskytovať rozhranie na ukladanie a čítanie súborov z distribuovaného súborového systému prostredníctvom grafického rozhrania. Týmto spôsobom je možné pridávať, poprípade kontrolovať súbory ktoré sa nachádzajú na distribuovanom súborovom systéme.

Požiadavky:

* Možnosť čítania súborov na distribuovanom súborovom systéme
* Možnosť zápisu dát na distribuovaný súborový systém
* Možnosť granulárnej konfigurácie oprávnení pre manipuláciu so súbormi na úrovni súborov a adresárov.

Pomocou priameho prístupu na distribuovaný súborový systém je možné nahrávať do platformy dátové zdroje, pre ktoré neexistuje automatizovaná integrácia (tento prístup je možné použiť aj v prípade manuálnych korekcií dát).

**Resource Management**

Data Lake platforma by mala poskytovať Resource Management službu, ktorá by zastrešovala prerozdelovanie dostupného výpočtového výkonu data lake platfromy medzi technické aplikácie a business užívateľov. Je nutné si uvedomiť že aplikácie ktoré budú využívať výpočtový výkon Data Lake platformy sú súčasťou ETL spracovania alebo sú súčasť biznisových procesov (machine learning, ad-hoc štatistika, ...). Resource Management služba by mala byť schopná zabezpečiť možnosť definovať prioritu a dostupné zdroje (CPU + RAM) pre rozdielne ETL aplikácie a rozdielnych business užívateľov (poprípade skupinu užívateľov). Takýmito pravidlami je možné zabezpečiť vyváženosť prerozdeľovania výpočtového výkonu Data Lake platformy medzi rozdielne aplikácie.

V prípade že by služba Resource Managementu nebola dostupná pre Data Lake platformu mohli by nastávať situácie využívania celkového výkonu paltformy v prospech jednej (alebo viac) aplikácie, pričom ostatné aplikácie by neboli schopné alokovať potrebné zdroje..

Resource manager Data Lake platformy by mal byť schopný zabezpečiť:

* Možnosť konfigurácie resource group na úroveň užívateľov, skupiny užívateľov a skupiny technických aplikácií
  + Minimálne a maximálne zdroje.
* Efektívne prerozdeľovanie dostupných zdrojov
* Monitoring dostupných a využitých zdrojov

**Security**

Data Lake platforma spĺňať bezpečnostné štandardy, ktoré zabezpečia že dáta uchovávané a spracúvané v Data Lake platforme budú dostupné len užívateľom a aplikáciám s prislúchajúcimi právami.

Nároky na bezpečnosť prostredia Data Lake platformy:

* Centrálne riadené
* Možnosť granulárnej konfigurácie oprávnení
  + databázy, tabuľky, stĺpce
  + adresáre a súbory prislúchajúceho distribuovaného súborového systému
* Grafické rozhranie na správu prístupových oprávnení
* Auditovanie údajov (všetky aktivity na Data Lake musia byť auditované a dostupné v reálnom čase)
  + Prístupy a modifikácie tabuliek a databáz
  + Prístupy a modifikácie na súborovom systéme
  + Spúšťanie distribuovaného spracovania

### Metodika vývoja

Pri vývoji DWH musia byť dodržané metodické postupy aby návrh DWH bol:

* **Nezávislý na platforme** - Navrhnutý dátový model by mal byť v čo najväčšej miere nezávislý od technológie, na ktorej bude DWH prevádzkovaný. Dátový model by mal byť prenositeľný medzi jednotlivými platformami (MS SQL, Oracle, PostgreSQL, atď.) a nemal by sa spoliehať na špecifiká vybraného databázového systému. Rozdiely plynúce zo syntaxe resp. samotného jazyka DDL jednotlivých platforiem nie sú považované za systémové špecifikum. Súčasť dodávky však budú tvoriť DDL definície iba pre dodanú platformu. V rámci projektu NBS DWH sa prepoužije technológia z Pilota DWH.
* **Nezávislý od zdrojových systémov** - Navrhnutý dátový model by mal byť čo najmenej závislý od zdrojových systémov. Predpokladá sa, že v dlhodobom horizonte, v čase životnosti nového dátového skladu dôjde k výmene niektorých business systémov.
* **Flexibilný** - Dátový sklad by mal byť z dlhodobého hľadiska rozšíriteľný na nové okruhy dát podľa budúcich požiadaviek zákazníka ako napr. nové typy produktov, zmena procesov atď.
* **Časová nezávislosť dát** - Navrhnutý dátový model musí zabezpečiť ako aktuálny, tak aj historický pohľad na dáta v DWH na základe výberu časového intervalu. Dátová retencia bude dohodnutá počas projektu. Predpokladaná dĺžka histórie je 10 rokov, granularita dát minimálne na úrovni kalendárneho dňa.
* **Aktuálnosť** - Dáta v dátovom sklade by mali byť aktualizované minimálne raz denne. V DWH by sa mal nachádzať stav dát s platnosťou minimálne D-1. Avšak môžu existovať definované typy dát, kde bude potrebné navrhnúť a použiť aj nižšiu frekvenciu aktualizácie (cyklické near-online spracovanie).
* **Interaktívny a inkremetálny vývoj** - Ktorákoľvek úloha v rámci projektu, alebo po ňom môže byť predmetom iterácie. Či iterovať alebo nie, rovnako ako počet iterácií, toto rozhodnutie je rôzne od prípadu k prípadu. Úlohy môžu byť iterované za účelom zvýšenia kvality výstupov na požadovanú úroveň, za účelom získania dostatočnej miery detailu alebo vylepšenia a rozšírenia výstupov na základe spätnej väzby používateľa.

Dodávateľ bude preukázateľne disponovať skúsenosťami v oblasti dátového modelovania a návrhu dátových modelov pre dátové sklady. V rámci projektu dodávateľ navrhne a zdokumentuje metodiku tvorby a ďalšej správy dátových modelov relevantných pre projekt, súvisiace design štandardy, názvové konvencie a prevádzkový model, tj. akým spôsobom budú dátové modely designované, aké role na ich tvorbe budú participovať a popíše i súvisiace procesy. Tieto štandardy prevádzkový model budú odsúhlasené s Data Governance kanceláriou a následne aplikované počas realizácie projektu.

V rámci projektu budú zavedené minimálne tieto dátové modely a súvisiace výstupy:

* Konceptuálny dátový model – dokumentujúci biznisové koncepty a pojmy podľa definície banky
* Fyzické dátové modely:
  + Data Lake
    - Primárna stage
    - Discovery vrstva
  + DWH
    - Transakčné repozitory (L1 – Integračná vrstva)
    - Prezentačná vrstva (L2 - Analytická vrstva)

Dátové toky a transformácie dát budú systematicky a metadátovo popísané tak, aby bolo možné tieto metadáta využiť pre navigáciu v dátových tokoch, zisťovanie dátových zdrojov pre cieľové atribúty a analýzu dopadov pri zmenách v dátových zdrojoch na cieľové štruktúry.

Nasledujúce princípy návrhu dátového modelu musia byť počas projektu dodržané a považujeme ich za kritické pre konzistentnosť, integritu, otvorenosť a výkonnosť DWH riešenia ako celku.

* Model transakčnej a prezentačnej vrstvy sú integrované a spoločné. Jedna entita je mapovaná do jednej tabuľky bez ohľadu na počet inštancií danej entity v zdrojových systémoch.
* Dimenzie v prezentačnej vrstve sú denormalizované a zdieľané a využívané cez všetky business oblasti. Existuje iba jedna dimenzia zákazníkov, iba jedna dimenzia umiestnení, atď. Zdieľané dimenzie podporujú krížové analýzy cez jednotlivé business oddelenia v spoločnosti.
* Prezentačná vrstva pozostáva zo star schém, bez snowflake. Inými slovami faktová tabuľka je spojená priamo s dimenziou nie je nutný viacnásobný join.
* Všetky dimenzie a referenčné tabuľky sú prepojené výlučne pomocou cudzích kľúčov. Cudzie kľúče chránia dátový sklad pred zmenami v zdrojových systémoch a umožňujú integráciu viacerých zdrojov do jednej tabuľky v DWH.
* Všetky dimenzie, prepojovacie lookup tabuľky (v prípade potreby aj faktové tabuľky) obsahujú aj prirodzené kľúče a identifikátor zdrojového systému. To znamená, že záznam v DWH môže byť dohľadaný v zdrojovom systéme.
* Multitenant – globalizácia a efektívne využívanie zdrojov vedie k situáciám, kedy v jednom zdrojovom systéme sa prevádzkujú údaje viacerých spoločností. V tomto prípade nestačí v DWH rozlišovať z ktorého zdrojového systému prišiel do DWH daný záznam, ale je potrebné riešiť kto „vlastní“ daný záznam v zdrojovom systéme. V DWH je potrebné sledovať vlastníka („tenanta“) daného záznamu. Táto informácia sa používa v určitých typoch reportov, kde je potrebné rozdeliť údaje o jednom zákazníkovi z pohľadu daného vlastníka napriek tomu, že ide o jednu a tú istú právnickú alebo fyzickú osobu.
* V prípade potreby budú údaje v DWH unifikované. Bude dohodnuté, ktoré dáta a ako budú unifikované. Prístup k unifikácií využíva dve rozdielne techniky:
  + Riadková unifikácia, napríklad v prípade zákazníkov. V tomto prípade je jedna informácia v DWH reprezentovaná jedným záznamom bez ohľadu na počet záznamov v zdrojových systémoch. Údaje sú unifikované pomocou rôznych čistiacich a unifikačných procedúr.
  + Druhý prístup je unifikácia pomocou hierarchií. Tento prístup sa využíva v prípade dimenzií (napr. produkty). V tomto prípade má každý produkt v DWH jeden záznam rovnako ako v zdrojovom systéme. Nad takouto dimenziou je vytvorená hierarchia, ktorá v reportoch zgrupuje identické produkty do jedného záznamu.
* Všetky riadky v tabuľke majú zhodnú granularitu a význam. To znamená, že nie je akceptovateľné miešať detailné a agregované údaje alebo údaje s rôznou granularitou.
* Integritná konzistencia (primárne, unikátne a cudzie kľúče) bude definovaná pre všetky tabuľky v transakčnej a prezentačnej vrstve. Avšak pre veľké tabuľky bude integrita dát kontrolovaná iba v ODI, nebude kontrolovaná na úrovni DB, z dôvodu výkonnostných problémov ELT.
* Pre veľké tabuľky je nutné využiť Oracle DB partície, aby sa uľahčil ELT proces, zvládol nárast historických údajov a zlepšila výkonnosť ako užívateľských dotazov, tak aj ELT nahrávania dát.
* Jedným zo základných princípov pri budovaní DWH je, že budovanie histórie v transakčnej vrstve DWH je s prihliadnutím, že údaje sa v DWH nikdy fyzicky nemažú. Počas projektu bude stanovená retenčná doba jednotlivých dát.

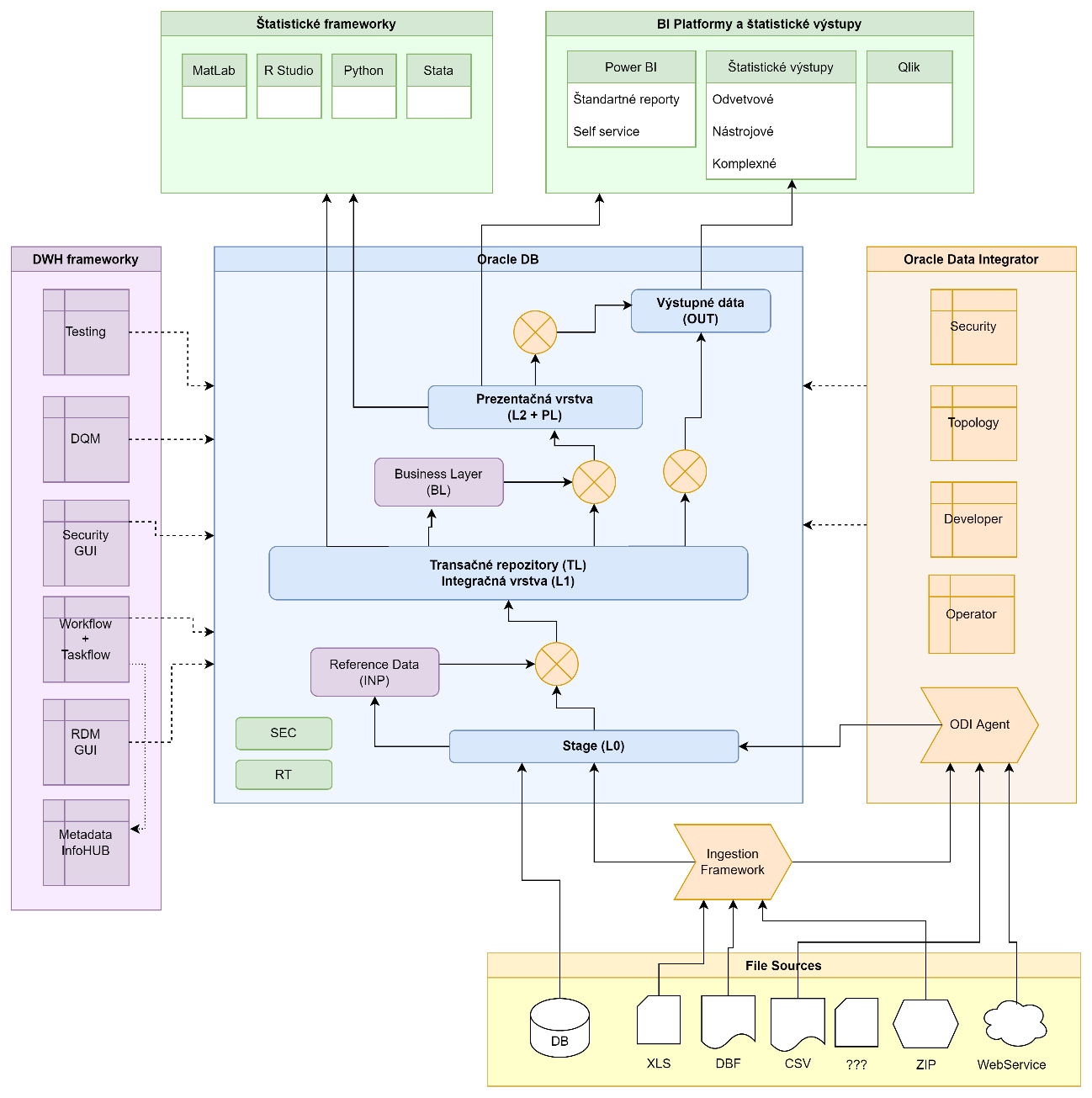
Dátový sklad na prezentačnej úrovni bude budovaný ako tzv. dimenzionálny dátový sklad, ktorý obsahuje dva základné typy objektov:

* Fakty, t.j. numerické hodnoty, ktoré sú sledované. Fakty sú spravidla aditívne, resp. semiaditívne hodnoty, ktoré je možné sčítať resp. spriemerovať podľa požadovaných kritérií. Fakty sú zoskupené do tzv. faktových tabuliek.
* Dimenzie, t.j. kritériá, na základe ktorých sú fakty členené. Príkladom je čas, nákladové stredisko, organizačná štruktúra atď. Dimenzie väčšinou obsahujú veľké množstvo atribútov a úrovní, ktoré prvok na dimenzii charakterizujú. Z úrovní je možné tvoriť hierarchie a prirodzeným spôsobom tak prechádzať z agregovanej hodnoty (vyššia úroveň) na detailnejšie dáta (nižšia úroveň). Dimenzie v dimenzionálnom dátovom sklade musia byť unifikované pre zdieľanie medzi rôznymi faktovými tabuľkami. Jedine tak je možné prechádzať („Drill Accross“) medzi faktami z rôznych faktových tabuliek - t.j. na základe rovnakých hodnôt spoločných dimenzií.

### Relačná časť NBS DWH

#### Úvod

Dátový sklad (DWH) bude logicky členený na niekoľko dátových vrstiev, pričom každá bude plniť špecifickú úlohu. Vrstvy sú navrhnuté v súlade so zavedenými prístupmi pre vnútorný návrh DWH.



Preto je požadované v  návrhu DWH dodržať zásady spísané v nasledujúcich podkapitolách.

#### Stage oblasť/L0

Vstupná vrstva dátového skladu, ktorá obsahuje dáta v štruktúre zdrojových systémov, resp. ich výstupných interfejsov. Účel tejto vrstvy je:

* Poskytnúť obraz vzájomne konzistentných zdrojových dát pre spracovanie v DWH na spoločnej technologickej platforme
* Štandardizovať dátové typy (napr. dátumy a číselné údaje)
* Poskytnúť krátkodobú cache pre prípadné dohľadávanie chýb a podozrivých stavov v dátach

Dáta môžu byť prenášané do STAGE vrstvy vo forme kompletných kópií zdrojových dát (napr. kmeňové dáta a referenčné dáta) alebo zmenových prírastkov (nové transakcie, resp. záznamy, kde došlo k zmene od posledného spracovania). Pre každý dátový interface (zdrojový systém, je jedno či ide o databázu, súbor, alebo záložku napr. v XLS) je vyžadované aby vznikol dokument tzv. Dohoda o extrakcii (Interface agreement). Kde bude popísané:

* Forma interface (DB, súbor, WebService)
* Veľkosť zdrojového systému a predpokladaný denný/mesačný inkrement
* Spôsob prenosu dát (dblink, FTP, JDBC, ODBC, http, ...)
* Umiestnenie systému (IP adresa, Linka, ...)

Názvy tabuliek a stĺpcov by mali reflektovať názvy v zdrojových systémoch. Počas projektu bude dohodnutá názvová konvencia, ktorá stanoví štandardy pre názvy tabuliek (a ich prefixy/sufixy), spôsob pomenovania business stĺpcov a názvy technických stĺpcov.

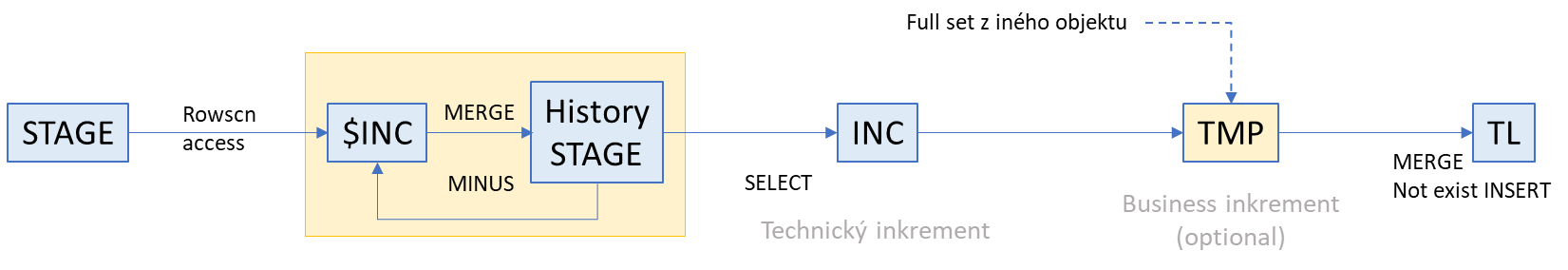
#### Princípy zachytávania zmien (Change Data Capture)

Prístup k zachytávaniu zmien údajov bude implementovaný spôsobom, ktorý minimalizuje množstvo údajov, ktoré majú byť spracovávané a taktiež zmeny riadkov, ktoré sú potrebné pre SCD. Budú použité niektoré z nasledujúcich techník, ktoré sú plne podporované ODI.

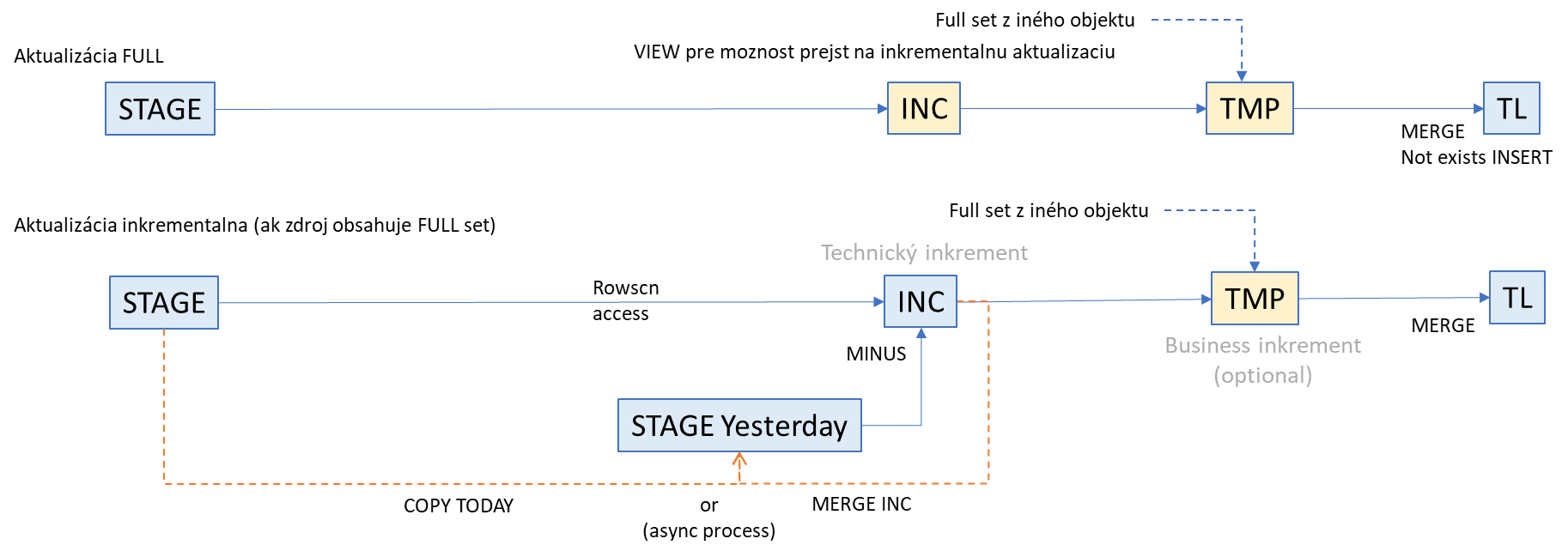
* Priamo systém poskytuje interface s potrebnými dátami na extrakciu.
  + Dáta budú zo systému priamo extrahované a uložené do STAGE. Toto uloženie bude označené metadátovou značkou, kedy dáta boli prenesené pre spätnú kontrolu kvality prípravy extraktu na zdroji – identifikácia prípadných duplicít, alebo opravných setov dát.
* Systém neposkytuje priamo dáta.
  + Pre malé a stredne veľké tabuľky budú sa robiť full extrakty zo zdrojových systémov. Tieto full extrakty sa porovnajú s údajmi, ktoré sa už nachádzajú v DWH buď množinovou operáciou MINUS alebo pomocou FULL OUTER JOIN na identifikáciu zmien.
  + Ak zdrojové tabuľky obsahujú spoľahlivú časovú známku, budú extrahované zo zdrojového systému riadky, ktoré majú časovú známku novšiu ako je predchádzajúca extrahovaná časová známka. Prenesené dáta musia byť označené metadátovou značkou.
  + V prípade, že extrakcia založená na časových známkach nie je možná, pripadá do úvahy extrakcia s využitím DB trigrov. Vyžadujú si úpravy na strane zdrojových systémov (vytvorenie DB trigrov – ako je MVIEW a podobne) a preto nepredpokladáme využitie tejto metódy v NBS DWH.

Pre účely riadenia zápisu zistených zmien na zdrojovej strane sa vyžaduje použitie nasledujúcej schémy ETL procesy:

**1. Pri použití historizovaného STAGE/L0**



**2. Pri použití klasického nehistorizovaného STAGE/L0**



#### Reference Data Oblasť (INP)

Účel tejto vrstvy je:

* Udržať konzistenciu dát na úrovni číselníkov
* Vstupná brána pre používateľov cez aplikačné GUI pre doplnenie dát do DWH
* Poskytnúť cez REST API tieto číselníky ostatných odberateľom v NBS

Táto oblasť preto obsahuje číselníky a čistiace, či mostíkové tabuľky pre:

* mapovanie zdrojových hodnôt na DW kľúče,
* doplnenie hierarchií a údržbu master dimenzií,
* pridávať ďalšie informácie importom súborov cez aplikačné GUI

#### Transakčné repozitory (L1 – Integračná vrstva)

Táto oblasť je základ dátového skladu, kde sa uchovávajú údaje na najnižšom detaile ako je možné. V transakčnej vrstve sa tvorí jedna pravda na jednom mieste, ktorá generuje históriu a sleduje zmeny v čase. Pri aktualizácii sú povolené výlučne technické transformácie a business transformácie, ktoré nie sú ovplyvňované business rozhodnutiami. Táto vrstva je striktne navrhovaná ako tretia normálna forma (metodika Billa Inmona) a vytvára deliacu čiaru medzi ďalšími komponentmi dátové skladu. Ak je potrebné zmeniť alebo doplniť ďalší zdrojový systém, upraví sa len prislúchajúca časť ETL v tejto vrstve. Všetky ostatné časti DWH sa nemenia. Navyše táto vrstva vytvára predpoklady na integráciu a normalizáciu údajov. Keďže transakčná vrstva zohľadňuje všetky relácie medzi objektmi, je možné vytvárať komplexné analýzy a dotazy, ktoré sú nad rámec dimenzionálneho sveta.

Pri návrhu transakčnej časti DWH je nutné v súčasnosti prihliadať aj na nové poznatky a metodiky, ktoré dopĺňajú a rozširujú skôr spomenuté prístupy modelovania. Pomáhajú zrýchliť, zjednodušiť ETL procesy napríklad použiť aj princípy modelovania metódy Data Vault 2.0[[1]](#footnote-1) .

Ako nadstavba nad týmto transakčným repozitárom je vytvorená prezentačná vrstva v zmysle dimenzionálneho modelu ako ju navrhuje Ralph Kimball pre vytváranie reportov a analýz.

Dátový model zachytáva konkrétne koncepty, entity, ich vzťahy a klasifikácie nezávisle na tom, ako sú zodpovedajúce dáta organizované v zdrojových systémoch. Vlastníkom dátového modelu je NBS.

T.j. cieľom tejto vrstvy dát je:

* Poskytnúť aktuálny aj historický pohľad na dáta, resp. ich historické verzie k akémukoľvek bodu v čase (použitie umelých kľúčov, vymedzenie platnosti záznamov – tzv. SCD2 mechanizmus u „pomaly sa meniacich“ dimenzií a väzobných tabuliek, resp. transakčné a snímkovacie tabuľky pre „rýchlo sa meniace údaje“)
* Organizovať dáta zrozumiteľne podľa business entít (entitno-relačný model), t.j. dáta z rôznych systémov zodpovedajúce jednej entite sú uložené spoločne
* Štandardizovať, resp. unifikovať dáta (vynútenie jednotného spôsobu zápisu údajov s rovnakým významom)
* Nahradiť neznáme, prípadne nekvalitné dáta dohodnutým spôsobom (default hodnoty a ošetrenie tzv. „late arriving“ údajov)
* Konsolidovať klasifikácie dát z rôznych zdrojov pre spoločný reporting (okrem uloženia pôvodných referenčných hodnôt to prebieha ako premapovanie cez konsolidované číselníky pomocou dát uložených v „Reference Data Oblasti“).
* Konsolidovať dátové entity podľa konceptuálneho návrhu dátového skladu. Ako sú napr.:
  + údaje o subjektoch (právnických aj fyzických osobách), ktoré môžu byť reprezentované v rôznych zdrojových systémoch pomocou rôznych identifikátorov a kódov – ich prepojenie bude implementované cez dohodnutý algoritmus tak, aby bolo možné prepojiť údaje súvisiace s konkrétnym výskytom subjektu pochádzajúcich z rôznych zdrojov.
  + údaje o entitách typu cenné papier/finančný inštrument.

#### Business oblasť

Podobne ako oblasť DWH nazývaná „referenčné dáta“, ktorá obsahuje dimenzie a číselníky pre doplnenie ETL procesu počas normalizácie a konsolidácie zo stage do transakčnej oblasti, tak oblasť nazývaná „business oblasť“ obsahuje mostíky, šablóny a prioritizačné tabuľky pre riadenie (ak je to potrebné, nie každé ETL to potrebuje) spracovanie z transakčnej oblasti do prezentačnej vrstvy.

#### Prezentačná vrstva (PL + L2 - Analytická vrstva)

Prezentačná vrstva je replikovaná časť transakčného repozitára, vytvorená na základe špeciálnej „business“ požiadavky. Zvyčajne sa navrhuje metódami Ralpha Kimballa a má dimenzionálny charakter, t.j. táto vrstva DWH poskytuje pohľad na dáta v štruktúre, ktorá je vhodná pre konkrétne analytické resp. reportovacie účely. Organizácia dát je podriadená konkrétnym požiadavkám užívateľov a BI aplikácií.

Na rozdiel od technických transformácii sa v priebehu ELT procesu z transakčnej do prezentačnej vrstvy využívajú predovšetkým business transformácie, ktoré sa v čase môžu zmeniť. V prípade zmeny business definície je možné podľa požiadavky prepočítať spätne časť alebo celú históriu na základe údajov uchovaných v transakčnej vrstve. V ELT proces sa využívajú špeciálne transformačné techniky (na základe definícií uložených v business vrstve DWH), ktoré umožňujú údaje transformovať na základe používateľských mostíkov, šablón a priorít.

Účelom tejto vrstvy je teda:

* Organizovať dáta optimálne pre konkrétne analytické a reportingové požiadavky do star-schém s potrebnou denormalizáciou a prípadnou agregáciou
* Poskytovať dodatočné vypočítané/odvodené dáta, ktoré sú špecifické pre danú analytickú potrebu.

#### Výstupná vrstva (OUT)

Výstupná vrstva je obdobou Prezentačnej vrstvy, dáta sú však organizované do dátových štruktúr, ktoré sú požadované na výstupe odoberajúcim systémom, resp. externým užívateľom (open data, ECB, iné inštitúcie, Off loading do Data Lake a pod.). Táto štruktúra teda nemusí spĺňať princípy pre modelovanie analytických dát pre reporting, je podriadená špecifickým požiadavkám následného použitia, resp. zverejnenia.

Z tejto vrstvy sa generujú výstupné extrakty pre distribúciu (disemináciu).

Podobne ako stage vrstva je definovaná dokumentom „Interface Agreement“ (Dohoda o extrakcii), tak OUT vrstva musí byť definovaná a dohodnutá cez dokument „Dohoda o poskytnutí dát z DWH“. Podobne ako dokument „Dohoda o extrakcii“ musí byť dokument popisujúci OUT vrstvu podpísaný oboma stranami dátovej komunikácie.

Dáta v OUT vrstve sú tvorené ETL procesmi z transakčnej vrstvy, ale aj z dát prezentačných datamartov.

#### DWH SandBox

DWH SandBox je dohodnuté miesto tvorené špeciálnym používateľom, ktorý má právo:

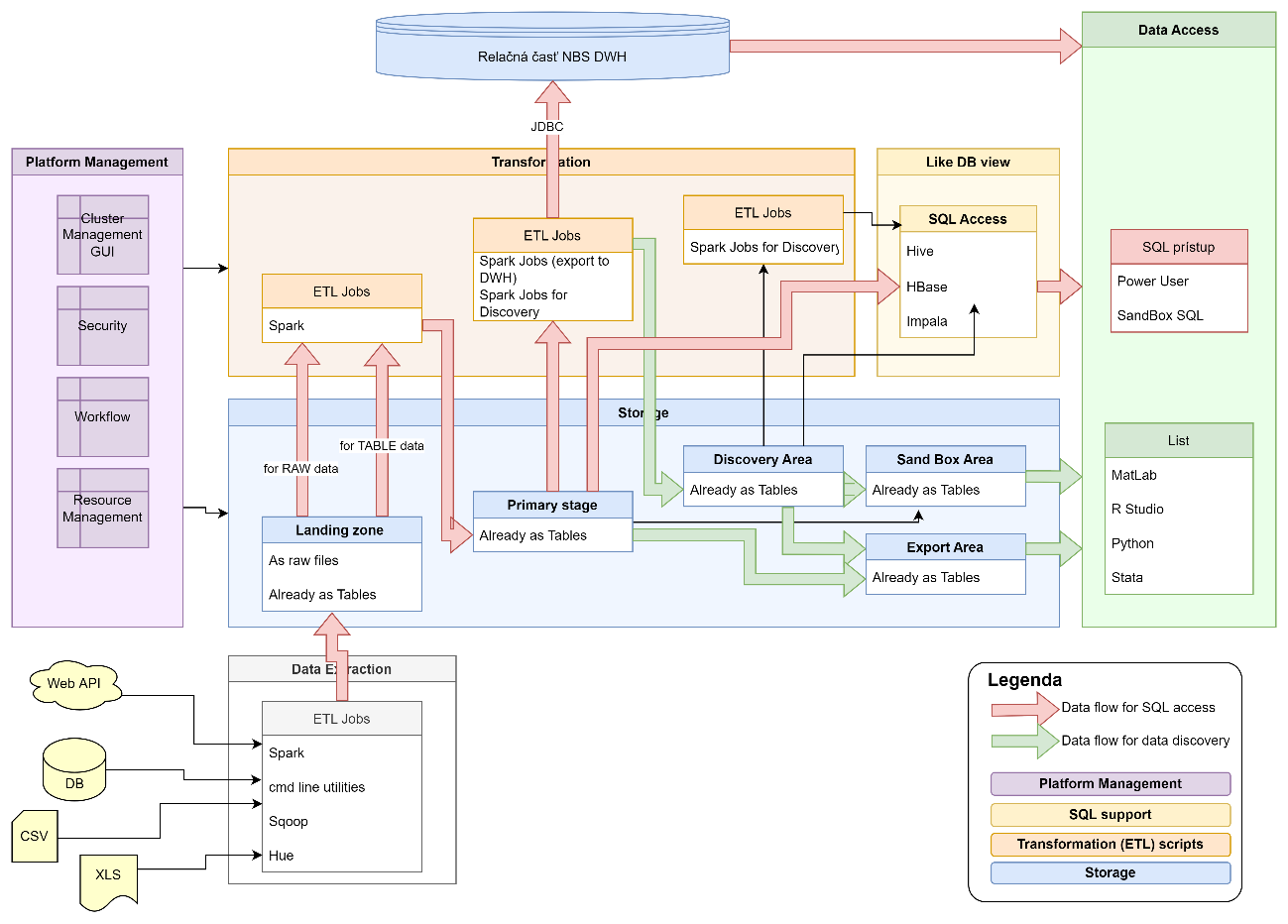
1. čítať dohodnuté dáta z DWH/DL
2. vytvárať nové tabuľky a pohľady
3. napĺňať tieto tabuľky
4. vymazávať dáta a celé objekty
5. pre pokročilých používateľov aj vytvárať procedúry a iné programy, ktoré môžu slúžiť na spustenie časti ETL a napr. pre overenie určitej hypotézy.

Sandbox nemá slúžiť na pravidelné prevádzkovanie časti spracovania DWH/DL. Pokiaľ je používateľ spokojný z pokusom je veľmi dôležité zapojiť IT do finálnej optimalizácie a zaradenia do pravidelného (denného) behu vytvorenej novej časti spracovania.

### DataLake časť NBS DWH

#### Úvod

Podobne ako pri návrhu relačnej časti DWH musia tieto predpoklady platiť pri vývoji Data Lake. Nasledujúce podkapitoly popisujú úlohu DataLake platformy v navrhovanej architektúre.



Dátové vrstvy DataLake budú logicky rozdelené na niekoľko vrstiev tak ako v prípade logického členenia dátových vrstiev DWH. Dátové vrstvy DataLake platformy môžu kopírovať aj DWH model.

#### Landing zóna

Dáta nachádzajúce sa v tejto oblasti obsahujú 1:1 informácie zo zdrojových systémov. Dáta sa môžu nachádzať buď vo forme tabuliek (štruktúrované dáta exportované zo zdrojových DB) alebo vo forme RAW súborov rôznych formátov pripravených na ďalšie spracovanie. Dáta môžu prichádzať buď inkrementálne alebo vo forme celkových snímok.

DataLake by mal poskytovať mechanizmy na import dát z nasledujúcich typov zdrojových systémov:

* Databázové (za použitia JDBC/ODBC)
* Súborové (možnosť stiahnuť dáta z externého SFTP, interného SFTP , alebo interného Sharepoint SSL úložiska (časť dát z externých inštitúcii, kde sa momentálne využíva protokol WebDAV sa presunie na zabezpečené Sharepoint SSL úložisko).
* API + WEB (možnosť importovať dáta dostupné cez API rozhranie (SOAP/REST), alebo dát dostupných za použitia parsovania public webu)

Podobne musí DataLake umožňovať priamo zapisovať dáta do Landing zóny iným ETL nástrojom.

#### Primárna stage

Dáta nachádzajúce sa v tejto oblasti obsahujú informácie, ktoré boli spracované z Landing vrstvy. Dáta v tejto vrstve majú vopred určenú štruktúru na základe biznisových a technických požiadaviek (vo forme tabuliek/partícií alebo dátových súborov). Dáta v tejto vrstve už nemusia obsahovať všetky informácie, ktoré boli exportované zo zdrojových systémov. Typickým príkladom v prostredí NBS môžu byť rozsiahle XML súbory, ktoré sú exportované zo zdrojových systémov, pričom len malé množstvo atribútov je následne procesovaných vo vyšších vrstvách. Pred vstupom dát do Primárnej stage vrstvy sú dáta kontrolované na dátovú kvalitu a transformované do finálnej podoby. Čiže hlavné transformácie dát spracúvaného obdobia zo zdrojových štruktúr do štruktúr Landing integračnej vrstvy sú:

* Štandardizácie / unifikácie dát
* Technických kontrol dátovej kvality
* Ošetrenia default hodnôt
* Výpočet spoločných odvodených údajov
* Historizácie dát v Primárnej stage

Dáta z primárnej stage vrstvy môžu byť exportované do iných systémov pre potreby ďalšieho spracovania.

#### Discovery

Dáta nachádzajúce sa v Discovery vrstve obsahujú spracované dáta pripravené na použitie pre analytické a štatistické potreby. Tieto dáta sú obohacované a transformované na základe biznisových a technických požiadaviek ETL procesmi, ktoré zabezpečujú potrebnú kvalitu dát.

Transformácia dát spracúvaného obdobia zo zdrojových štruktúr Primárnej stage do Discovery analytickej vrstvy prebieha najmä z dôvodu:

* Denormalizácie dát
* Dodatočných výpočtov
* Optimalizácie pre reporting (napr. materializácie výpočtov a joinov a agregácií)
* Historizácie dát v Discovery vrstve podľa konkrétnej analytickej potreby

Štatistické a analytické nástroje využívajú primárne dáta z tejto vrstvy na vytváranie analýz a reportov.

#### DL Sandbox

Data Lake platforma by mala umožňovať dátovým vedcom vytváranie “sandboxov“ pre potreby importovania a manipulácie aj väčšieho množstva dát. Výhodou takýchto “sandboxov“ je možnosť spájania externých dát s dátami, ktoré sú dostupné v Data Lake platforme bez nutnosti zapojenia IT. Platforma by mala poskytovať jednoduché rozhranie, ktoré by dátovým vedcom umožňovalo import externých dát v podobe súborov a možnosť nad takto importovanými súbormi vytvárať databázové tabuľky, poprípade používať tieto súbory pre účely analýz za použitia programovacích/štatistických jazykov. Data Lake platforma by mala umožňovať pri analyzovaní takto importovaných súborov využívať celkový výpočtový potenciál platformy.

Dáta v sandboxoch by mali byt logicky oddelené a nemali by byť súčasťou dát v Discovery vrstve.

### Informačná bezpečnosť

#### Autentifikácia používateľov

Systém musí zabezpečiť jednoznačné overenie používateľa pred akýmkoľvek prístupom. Požiadavky na autentifikáciu:

* Každý používateľ musí mať jedinečný prístupový identifikátor/užívateľské meno (UID)
* Každý užívateľ sa musí autentifikovať minimálne pomocou UID a hesla.
* Heslá musia byť v systéme uložené tak aby boli nedešifrovateľné
* Systém by mal ponúkať možnosť napojenia na SSO riešenie

#### Prístup k dátam

Systém musí poskytnúť možnosť na riadenie prístupov na všetkých úrovniach technickej architektúry. Aplikované postupy a princípy riadenia oprávnení používateľov budú zdokumentované v relevantnej dokumentácii riešenia.

Prístup k dátam musí byť definovateľný:

* Na úrovni rolí jednotlivých používateľov minimálne pomocou riadenia prístupu k jednotlivým entitám dátového modelu pre analytické role
* Riadenie prístupu k záznamom (riadková bezpečnosť)
* Riadenie prístupu k informačným typom (stĺpcová bezpečnosť) najmä z dôvodu zavedenia chránenia citlivých dát
* Osobitné oddelenie manažmentu rolí a prideľovania oprávnení užívateľom

#### Logovanie používateľských aktivít

Systém by mal zabezpečiť záznam všetkých používateľských aktivít ohľadne prístupu k dátam. Zo záznamov by malo byť spätne identifikovateľné kto, kedy a k akým dátam pristupoval. Záznam by mal obsahovať minimálne:

* Dátum a čas operácie
* Identifikátor používateľa
* Názov a adresa pracovnej stanice užívateľa
* Vykonaná operácia a/alebo dátová požiadavka (napr. kópia SQL príkazu)
* Úspech/neúspech vykonanej operácie

Systém musí zachytávať minimálne nasledovné udalosti:

* Prihlásenie/odhlásenie užívateľa
* Používateľský prístup k dátam
* Zmena dát

### Iniciálny load

Pre každú tabuľku DWH v transakčnej vrstve bude zabezpečený tzv. iniciálny load - prvotné nahratie kompletného obrazu dát zo zdrojového systému. To môže obsahovať aj historické položky číselníkov a kmeňových dát, ak existujú, prípadne historické transakcie a snímky, ak ich daný systém udržiava v produkčných tabuľkách (veľmi často teda ide napr. o historické transakcie - k nim musia byť dohraté aj relevantné referenčné a kmeňové dáta, aj keď dnes už nemusia byť aktívne - dlhopis, platobná karta, subjekt).

Zvyčajne sa nerekonštruujú historické stavy statických, resp. pomaly sa meniacich dát, ak tieto zmeny neobsahuje aj samotný zdrojový systém vo svojich základných zdrojových štruktúrach.

### Migrácia dát z pilotného DWH

Nie je vyžadovaná migrácie dát z pilotného DWH do nového riešenie.

### Metodika fyzického vývoja a inštalácie na produkciu

Z dôvodu zabezpečenia bezpečnej prevádzky a ochrany údajov v DWH budú požité na vývoj, testovanie a samotnú prevádzku 3 prostredia:

* DEV – development prostredie, jeho charakteristiky:
  + málo dát,
  + anonymizované, alebo vybratá reprezentatívna vzorka, ktorá ale neporušuje predpisy ochrany údajov
  + vykonanie systém testov
* TEST – testovacie prostredie, jeho charakteristiky:
  + väčšie množstvo dát pre vykonanie určitých performance testov
  + vykonanie integračných testov
  + úvodné UAT testy business používateľov
* PROD – produkčné prostredie
  + záverečné UAT testy business používateľov
  + bežná prevádzka

Vývoj bude prebiehať v nasledujúcom predpise:

* DEV prostredie je plne R/W. Vývoj bude prebiehať pre ETL/Workflow len nastavením metadát. Nie je dovolené používať PL/SQL procedúry pre ETL úlohy. Všetky aplikačné PL/SQL procedúry musia byť uložené ako metadáta (anonymné bloky) vo workflow frameworku, kde majú svoju historickú verziu a metadáta dôvodu vzniku.
* TEST prostredie je plne R/W ale v koordinácii s DEV. Prostredia musia byť synchronizované pravidelne na metadátovej úrovni. ETL/Workflow, ktorý je vo vývoji bude mať metaznačku.
* PROD prostredie je čisté R/O prostredie pre metadáta. Prenos z TEST do PROD sa vykoná cez inštalačný balíček. Zápis do metadát ETL/Workflow je možný len počas riešenia produkčného incidentu 1. stupňa (hot fix). Pri iných riešeniach incidentov bude použitý štandardný postup DEV->TEST->PROD.

### Nástroje na podporu prevádzky

#### Workflow Management – Data Orchestration

Kvalitné riadenie a správa aktualizácie dátového skladu je jedným z kritických ukazovateľov úspešnosti implementácie dátového skladu. Len aktualizácia, ktorá dáva presný obraz o svojom behu dáva dôveru business používateľovi, že dáta uložené v dátovom sklade sú korektné a pripravené spracovanie v reportoch. Preto takýto framework musí umožňovať:

* Nastavenie kalendára a času spustenia jednotlivých úloh
* Riadenie závislostí (dynamické zapínanie a vypínanie)
* Error handling – znovu spustenie, nastavenie chybovej vetvy behu
* Manuálne spustenie úlohy
* Nastavenie cyklického spracovania dát
* Riadenie heterogénneho prostredia
* Logovanie behu úloh
* Analytiku nad logmi behov jednotlivých úloh
* Nastavenie používateľských práv a prístupov

Framework je možné rozdeliť do viacerých logických celkov:

1. Editor matadát
   1. Vytváranie úloh
   2. Prioritizácia úloh
   3. Rozdelenie úloh do vertikálnych a horizontálnych skupín
   4. Definovanie závislostí medzi úlohami v heterogénnom prostredí (PL/SQL, ODI, JAVA, WebService, ShellScript a iné)
   5. Možnosť definovania anonymného bloku SQL namiesto zavolania externej úlohy
   6. Definovanie kalendára
   7. Verzionovanie zmien
2. RunTime
   1. Spúšťanie úloh podľa kalendára, závislostí a stavov skupín
   2. Automatická reštartovateľnosť úlohy
   3. Automatická prioritizácia úloh podľa výsledku behu
3. Monitoring
   1. GUI podporujúci admin konzolu pre bežný dohľad - Celkový pohľad na end-to-end spracovanie z jedného miesta
   2. GUI podporujúci zjednodušenú konzolu pre kontrolný dohľad, napr. vybraným business používateľom
   3. Manuálna reštartovateľnosť úloh v prípade poruchy – Schopnosť nadviazať na už vykonané úlohy
   4. Notifikácie v prípade dobehnutia alebo chybových stavov spracovania
   5. Analýza štatistík o spúšťaní jednotlivých úloh, sledovanie kritickej cesty a dohodnutej SLA o dobehnutí ETL úloh
4. procedúry pre logovanie a samotné logovacie tabuľky
   1. Popísané API na jednotné logovanie
   2. Integrované API (VIEW) pre čítanie ODI logov

Riadenie spracovania dát v prostredí NBS bude riešené jedným spoločným nástrojom, ktorý bude dostatočne technologicky zdatný designovať, spúšťať a administrovať tok spracovania cez rôzne technické platformy. Táto požiadavka zabezpečí, že jednotlivé kroky dátového spracovania bude možné vykonávať na rozdielnych platformách, čo bude mať za dôsledok efektívnejšiu správu zdrojov nad platformami.

#### Testovací framework

Testovací framework predstavuje funkcie pre definíciu a automatizované spúšťanie a vyhodnocovanie testovacích scenárov cez rôzne prostredia. Typicky slúži na automatizované (alebo opakované) vykonávanie rôznych testovacích prípadov v rámci SDLC procesu, napr.:

* kontroly stavu prostredia
* funkčné testy (správne plnenie cieľových štruktúr)
* regresné testy

Framework by mal poskytovať minimálne tieto funkcie:

* repozitár testovacích skriptov (SQL)
* mechanizmus pre automatizované spúšťanie zvolených testovacích skriptov
* logovanie výsledkov testov a ich sprístupnenie pre vyhodnotenie (SQL, resp. v reportoch cez BI nástroj)

Je možné tento framework spojiť s DQM podporou, ale potom sa vyžaduje aby nastavenie metadát DQM a jeho exekúcia, prípadne reporting boli jasne, zrozumiteľné a ľahko oddeliteľné od zadaných Testovacích scenárov.

#### Manuálne vstupy – selfservice pre používateľov

Táto požiadavka rozširuje požiadavku pre GUI RDM pre “Manuálne vstupy statických dát malého rozsahu” a funkcionalitu vytvorenia importného medzikroku na kontrolu dátovej kvality importovaných dát na základe zadefinovaných pravidiel.

Proces manuálneho importu používateľom bude následne vyzerať (navyše okrem štandardnej RDM funkcionality):

* Importované dáta zo známeho a predom dohodnutého formátu sú transformované (rozšírené o mapované dohodnuté dimenzie) do medzipamäte
* GUI zobrazí stav importu s tým, že chybné dáta sa zobrazia červenou farbou.
* GUI nemusí mať podporu na opravu týchto dát
* Po kontrole a správnom nastavení všetkých riadkov sa importný balík označí metaznačkou v STAGE vrstve DWH
* Importovacie GUI musí obsahovať sumárny report (súbor, kedy, stav, počet chýb, atď.) pre sledovanie importnej činnosti priamo používateľom

#### Ingestion framework - nahrávania veľkého počtu rôznych súborov

Súčasťou DWH/DL platformy by mal byť Ingestion framework, ktorý by primárne zabezpečoval importovanie externých zdrojov do Stage Zóny platformy. Externé zdroje sú dostupné prostredníctvom API rozhraní, parsovania public  webových stránok alebo sťahovania dát zo vzdialených SFTP serverov mimo NBS infraštruktúru. Spracovanie Ingestion frameworku by malo byť riadené prostredníctvom Workflow managera.

Požiadavky na Ingestion Framework :

* Možnosť pridávania nových externých zdrojov do framework
  + API, parsovanie stránok, SFTP, Kafka, JDBC
* Možnosť metadátovo definovať parser čítača pre známe formáty ako je CSV, AWRO, JSON , XML, DBF a podobne
* Možnosť priamo čítať ZIP na vstupe
* Možnosť metadátovo definovať mapovania zdroj-cieľ
* Možnosť priamo počas spracovania dát tieto dáta obohacovať o iné dáta definovanou reláciou na mapovacie dimenzie.
* Spúšťaný z centralizovaného Workflow mangera.
* Schopnosť importovať len inkrementy dát pričom Ingestion framework by mal byť zodpovedný za udržiavanie kompletného prehľadu o už spracovaných dávkach. T.j. musí mať mechanizmy ako spracovať „práve raz a všetky“ vstupné zdroje. Toto spracovanie musí byť kontrolované, auditované a reportovateľné – interface, beh, čas, počet spracovaných vstupných riadkov, počet výstupných, počet chybových, počet zahodených z dôvodu duplicita a atd.
* Udržiavať si metadáta o externých zdrojoch (API kľúče, mená, heslá, tokeny, logy, aktuálny stav spracovania + v prípade inkrementálneho spracovania všetky potrebné informácie potrebné pre stiahnutie ďalšieho inkrementu, počty exportovaných záznamov pre každé spustenie)
* Schopnosť spracované dáta uložiť cez JDBC do STAGE DWH, alebo cez HDFS konektor priamo do zóny DL.
* Schopnosť robiť tzv. multi-insert, t.j. keď vstupný tok dát obsahuje viac rôznych dátových formátov, tak dodaný framework tieto formáty bude vedieť rozoznať a uložiť do rôznych cieľových objektov
* Schopnosť odložiť chybné záznamy počas spracovania do iného dohodnutého objektu na ďalšiu analýzu. Túto schopnosť bude možné nastaviť metadátovo.
* Schopnosť identifikovať duplicity v dohodnutom spracovávanom okne a v rámci jedného streamu. Kontrolu duplicít bude možné nastaviť metadátovo.

#### Security GUI (relačná časť DWH)

Framework pre správu prístupov je kritický komponent, ktorý umožňuje riadiť viditeľnosť dát a celkové práva používateľov DWH systému. Framework musí umožňovať:

* Definícia role (skupiny) používateľov a priradenie reálneho používateľa z DB
* Definícia prístupovej role (objektová skupina) a priradenie reálnej role z DB
* Definícia riadkovej bezpečnosti objektu
* Definícia stĺpcovej bezpečnosti objektu
* Priradenie riadkovej a stĺpcovej bezpečnosti objektu k prístupovej role (objektová skupina)
* Priradenie role prístupovej role (objektová skupina) na rolu (skupinu) používateľov

Framework nesmie umožňovať z bezpečnostných dôvodov:

* Definícia používateľa
* Definícia role

Framework musí vedieť na požiadanie vygenerovať požadované VPD politiky pre jednotlivých používateľov a vytvoriť požadovanú stĺpcovú bezpečnosť.

#### Riadenie dátovej retencie

Dátový sklad umožňuje uložiť dáta v histórii podľa požiadaviek používateľov, zvyčajne v nekonečnej. Avšak musí existovať retenčná politika napr. z dôvodu ochrany investícií do IT infraštruktúry, alebo z nejakých legislatívnych dôvodov.

Je požadované aby v rámci projektu NBS DWH bol dodaný framework na:

* automatické pridávanie, presúvanie, defragmentovanie, odpájanie definovaných dátových particií.
* dátová retencia bude definovaná metadátami pre každý objekt zvlášť.
* Spustenie frameworku bude riadiť workflow manažér
* Framework musí vedieť zalogovať svoj stav do riadiacej konzoly

### Nástroje pre Data Governance

Súčasťou dodávky budú aj aplikácie a nástroje na podporu novo vznikajúcich procesov Data Governance, ktoré budú zahŕňať aktivity, procesy a potrebné výstupy pre:

* Reference Data Management (RDM) – riadenie referenčných dát
* Data Quality Management (DQM) – riadenie dátovej kvality
* Metadata Management – správa metadát
* Dátové modelovanie a správu dátových modelov

Tieto nástroje budú dodané a nasadené v kontextu implementácie platforiem Data Lake a DWH, ale rozsah ich využitia by nemal byť v budúcnosti obmedzený iba na tieto platformy. Tieto nástroje by mali byť schopné aplikovať potrebné koncepty a procesy aj v rámci ďalších systémov NBS.

#### RDM – Reference Data Management

Pre klasifikáciu údajov v DWH/BI je možné vychádzať z číselníkov a klasifikácií, ktoré sú poskytované zdrojovými systémami ako súčasť jeho natívneho dátového modelu.

Pre reporting však vznikajú aj požiadavky na triedenie dát podľa odvodených charakteristík, prípadne unifikovaných klasifikácií, ktorých kódovníky nie sú v zdrojových systémoch k dispozícii, resp. zdrojové systémy neposkytujú potrebné údaje (napr. užívateľské hierarchie, súčtové prvky, príznaky, deskriptívne názvy a pod.).

Na správu týchto rôznych číselníkov a sú základom pre kvalitný reporting, je nevyhnutné mať viac ako menej rôznych formulárov, ktoré podporia:

* Správa referenčných dát a hierarchií údajov.
  + číselníky / kódovníky
  + klasifikácie a iné malé dimenzie
  + definície hierarchie
  + lookup tabuľky pre konsolidáciu číselníkov a ich obohatenie o dodatočné informácie (popisy, príznaky)
* Správa konverzných mostíkov (napr. aj pre potreby ETL – transformácia technických transakcií na business definície pre niektoré hviezdicové schémy)
  + parametre (zoznamy, podmnožiny hodnôt) používané v rozhodovacej logike spracovania dát
  + rozhodovacie tabuľky pre odvodenie dát a aplikáciu obchodných pravidiel pri spracovaní dát v DWH
* Udržovanie nastavení DWH metadát – napr. spôsob distribúcie niektorých automatických reportov cez EMIAL a podobne.
* Manuálne vstupy statických dát malého rozsahu, ktoré nie sú primárne uložené a obhospodarované v inom systéme

Tento systém by mal poskytovať minimálne tieto funkcie a nástroje pre správu referenčných dát:

* umožniť používateľom (odborným aj technickým) modifikovať referenčné dáta cez jednoduchý GUI formou selfservice.
* riadiť prístup k jednotlivým referenčným dátam s autentifikáciou a autorizáciou užívateľov k jednotlivým referenčným dátam
* auditovať všetky akcie a verzovať zmeny v záznamoch s časovými značkami
* umožňovať definíciu nových referenčných dát s validáciami vstupných dát (minimálne na doménovú a referenčnú integritu)
* sprístupniť dáta pre ETL proces cez definované rozhranie
* poskytovať online služby pre ostatné aplikačné systémy, ktoré chcú svoje referenčné dáta delegovať na RDM

Riešenie RDM bude koncipované ako „s úložiskom priamo v DWH“ s možnosťou batchovej či okamžitej synchronizácie do DWH/Data Lake. Rovnako musí prostredie RDM poskytnúť REST API pre prístup k dátam iným aplikačným systémom.

#### Data Quality Management

Systém pre DQM (Prostredie pre správu, spúšťanie a vyhodnocovanie dátovej kvality) bude riešiť nasledujúcu funkcionalitu:

* Konfigurovať pravidlá pre meranie dátovej kvality (DQ indikátory)
* Organizovať pravidlá DQ do skupín a plánov ich spúšťania
* Spúšťať a plánovať spúšťanie vybraných skupín indikátorov
* Monitorovať výsledky meraní cez reporty a grafy
* Možnosť zasielať výsledky kontrol na email - generovať podklady pre Data Quality Incident
* odkladanie chybných záznamov do špeciálnych tabuliek pre následnú kontrolu a riešenie incidentov
* zastavenie ETL procesu pri prekročení hraničného počtu chýb
* zakomponovanie spustenia DQ kontroly do procesu ETL – ako samostatný krok spracovania

Tento systém bude pripojený na DWH, nad ktorým bude prebiehať meranie indikátorov kvality cez rôzne dátové vrstvy. Riešenie by malo obsahovať komponenty pre podporu riadenia dátovej kvality na úrovni:

* Technických kontrol – štandardné / generické kontroly vyplývajúce z dátového modelu
* Business kontrol – užívateľsky definované parametre kvality na úrovni atribútov (data quality indikátory)

**Technické kontroly** by mali zahŕňať:

* Dátové typy
* Obmedzenia (Constraints) – povinné hodnoty, unikátnosť, referenčná integrita
* Doménové kontroly – validácia na prípustné hodnoty
* Kvalita zbehnutia SCD

Spúšťanie technických kontrol by malo byť integrálnou súčasťou ETL procesu.

**Business kontroly** môžu byť voľne definované zo strany užívateľov (resp. Data Stewardov) vo forme tzv. *indikátorov dátovej kvality (DQI)*, napríklad:

* hraničné (podozrivé) hodnoty atribútov
* neočakávané kombinácie hodnôt súvisiacich atribútov
* neočakávané vzory v dátach a pod.

Tieto indikátory môžu mať komplexnú definíciu, a teda môžu byť definované vo forme SQL dotazu.

#### Metadata Management

V procesoch Data Governance vzniká množstvo metadát, ktoré popisujú vecnú aj technickú stránku spracúvaných a publikovaných dát, ktoré je potrebné efektívne sprístupniť širšej základni používateľov a analytikov, a uľahčiť tak ich navigáciu v dostupných dátach.

Budúci systém integrácie dát by mal byť doplnený o riešenie pre správu metadát, ktoré bude poskytovať NBS nástroje pres správu biznis katalógov, data lineage, data katalog a navigáciu v nich.

Medzi tieto metadáta patria:

* Business slovník pojmov
* Dátové katalógy – popisujúce dostupné dáta, zdroje, Data Lake, DWH
* Ďalšie katalógy, napr. KPI katalóg, DQI katalóg, katalóg reportov
* Relácie medzi pojmami a dátovými modelmi resp. dátovými katalógmi
* Údaje o dostupnosti a kvalite dát

Nástroj pre manažment metadát bude umožňovať tieto funkcie:

* Správa a audit Business slovníka
* Správa a audit Dátového katalógu s popisom dát v dátových zdrojoch, DWH a Data Lake
* Automatické (na spustenie) naplnenie metadát z dostupných metadát DB, ODI, Workflow Managera, DQ nastavení a podobne. Vyžaduje sa minimalizácia manuálnej správy informácií.
* Možnosť vytvárať vlastné katalógy podľa potreby Data Governance a prispôsobovať ich obsah
* Funkcionalita data lineage a impact analýzy nad objektami zaznamenanými v katalógoch (business pojmy, KPI, DQI) a importovanými metadátami o objektoch
* Funkcia fulltextového vyhľadávania v rámci všetkých atribútov katalógov a importovaných metadát
* Portál s webovým frontend-om pre navigáciu v metadátach

#### Dátové modelovanie

Nástroj pre dátové modelovanie by mal podporovať vývojárov pri analýze, designe, implementácii, testovaní a nasadzovaní softwarových riešení. Okrem toho plní dokumentačnú funkciu, je zdrojom metadát dokumentujúcich dátové štruktúry a závislosti pre business data lineage (závislosti dát DWH na ultimatívnych zdrojových štruktúrach). V jednotlivých fázach SDLC sa využíva nasledujúcim spôsobom:

* + Analýza – reverse-engineering a dokumentácia dátových zdrojov, dopadová analýza zmien v zdrojových štruktúrach na štruktúry v DL/DWH, komunikačný prostriedok so zadávateľom
  + Design – návrh a dokumentácia nových dátových štruktúr, definícia logiky mapovania na dátové zdroje, generovanie časti dokumentácie
  + Implementácia – generovanie kódu pre tvorbu a modifikáciu databázových štruktúr, exportovanie metadát pre generovanie niektorých transformácií a technických kontrol pri spracovaní dát v DL/DWH
  + Testovanie – zdroj informácií pre testerov k dátovému modelu a očakávanej transformačnej logike
  + Nasadzovanie - generovanie kódu a metadát pre nasadenie, sledovanie stavu vývoja a deploymentu štruktúr cez prostredia

Voľba konkrétneho nástroja pre dátové modelovanie tak môže výrazne ovplyvniť spôsob a kvalitu implementácie. Prosíme odporučiť nástroj na podporu hore napísaného procesu. Bude zakúpený mimo výberu tohto projektu.

## Business požiadavky

V rámci dodávky projektu požadujeme implementáciu end2end reportingového nástroja, ktorý bude podporovať analytické procesy.

Na základe pred-analýzy boli identifikované kľúčové oblasti pre:

1. Dátovú integráciu zdrojových systémov
2. Reporting a analýzy
3. Výstupné dátové oblasti

### Zdrojové (vstupné) systémy

Tabuľka nižšie obsahuje zoznam zdrojových systémov, ktoré je potrebné integrovať do nového NBS DWH. Tabuľka obsahuje systémy označené ako:

1. Interné – systém, ktorý nesie dáta požadované pre integráciu do DWH a je v internej sieti (infraštruktúre) NBS. Dáta sú uložené buď v databázach, alebo v extraktoch vo forme súborov.
2. Externé – systém, ktorý obsahuje dáta požadované pre integráciu a je mimo infraštruktúry NBS. K dátam sa môže pristupovať rôzne cez API (Web Service) až po súborové rozhrania (HTTPS/SFTP a iné.)
3. Manuálne dátové vstupy - dáta sú dáta získavané z neštruktúrovaných dátových zdrojov ako je napr. e-mail, web, ročenka, výročné správy, ktoré je možné ukladať v štruktúrovanej podobe do strojovo spracovateľných formátov, avšak ide o výsledok manuálnej práce spracovateľov týchto dát. V prehľade nižšie je rámcový výpis takýchto zdrojov rozdelených podľa účelu použitia v aktuálne dostupnom formáte. Predpokladom pre automatizáciu bude aj vytvorenie rozhrania, cez ktoré sa budú rôznorodé manuálne vstupy do procesu integrovať. Momentálne takéto rozhranie neexistuje a každé spracovanie používa prispôsobené logiky pre prácu s manuálnymi vstupmi a potrebné dáta extrahuje cez vlastný dekóder. Pozri kapitolu 3.2.13

Význam ďalších atribútov tabuľky je nasledovný:

* Fáza ... predpokladaná fáza dodávky dátovej integrácie do DWH (pozri kapitolu 3.4.5)
* Dôvod integrácie:
  + BI ... dáta z označeného systému pôjdu primárne použiť v BI reportingu
  + OUT ... dáta z takto označeného systému sú potrebné pre výstupy z DWH na odosielanie mimo NBS v presne definovanej štruktúre
* Databázový systém
  + DB ... dáta uložené v NBS databáze (Oracle, MS SQL, MySQL)
  + Súbor ... lokálne uložené súbor (CSV, MS Excel) na FTP úložisku alebo SharePoint v prostredí NBS
  + WebAPI ... Existencia WebAPI na externom serveri mimo NBS
  + Lotus Notes ... dokumentačná databáza LN
* Objem dát (GB) .. predpokladaný objem dát v Giga Bytoch (1.000.000.000 bytov)
* Počet objektov ... tabuliek / súborov / API endpointov

| Názov systému | Popis systému | Interný /externý | Fáza | Dôvod integrácie | Databázový systém | Objem dát (GB) | Počet objektov |
| --- | --- | --- | --- | --- | --- | --- | --- |
| ARDAL | Agentúra pre riadenie dlhu a likvidity - údaje o úveroch, vkladoch, CP, derivátoch, repo obchodoch (stavy a transakcie); (pre spracovanie GFS, BoP, FA) | externý | 2 | OUT | Súbor (XLSX) | 1 | 2 |
| ASDR | Agendový systém dohľadu a regulácie - ISD (Informačný systém dohľadu) | interný | 4 | BI | Oracle DB | 9 | 100 |
| BISNODE | Databáza firiem - kvantitatívne a kvalitatívne informácie o subjektoch | externý | 1 | BI, OUT | WEB API (CSV) | 1,3 | 14 |
| BLMBRG | Bloomberg - obchodná platform | externý | 2 | BI | WEB API (JSON/XML) | 1 | 50 |
| CERI | Centrálna evidencia regulovaných informácií | interný | 2 | BI | MySQL DB (Tabuľky/JSON) | 1 | 3 |
| CRIBIS | CRIBIS - databáza firiem, majetkové a personálne prepojenia | externý | 3 | BI | REST API | 1 | 1 |
| CSDB | Centralised Securities Database | interný | 2 | BI, OUT | WEB API (XML) | 170 | 4 |
| ECB SDW | European Central Bank Statistical Data Warehouse - všetky datasety kde figuruje krajina "SK" | externý | 3 | BI | Súbor (CSV/XML) | 1 | 50 |
| ESKN | Katastrálny úrad (nehnuteľnosti, listy vlastníctva, vlastníci, prepojenia medzi dátami) | externý | 2 | BI | Súbor (DBF) | 2 | 15 |
| ESMA | European Securities and Market Authority | externý | 1 | BI | Súbor (XML) | 500 | 5 |
| Eurofondy | Exporty dát z čerpania eurofondov | externý | 3 | BI | WEB API (JSON) | 50 | 39 |
| Finančná správy SR | Tabuľky Finančnej správy SR pre spracovanie BoP | externý | 2 | OUT | Súbor (XLSX) | 0,5 | 2 |
| FUJITSU ŠZP ESAS | ŠZP ESAS (XBRL Fujitsu) | interný | 1 | BI | Súbor (XBRL) | 80 | 100 |
| FUJITSU ŠZP ESAS | ŠZP ESAS (XBRL Fujitsu) | interný | 1 | BI, OUT | DB – Oracle (Tabuľky/BLOB obsahujú XML v ZIPe) | 1 | 100 |
| IDAS | Agendový systém dohľadu | interný | 3 | BI | Súbor (CSV) | 55 | 10 |
| KL | Kurzový lístok ECB | externý | 1 | BI, OUT | WEB API (XML) | 1 | 2 |
| LEI kódy / GLEIF |  | externý | 1 | BI | WEB API (XML) | 0,05 | 1 |
| Makroek. DB | Makroekonomická databáza | interný | 2 | BI | Súbor (XLS/CSV) | 5 | 8 |
| manuálne vstupy | Konfiguračné metadáta pre všetky spracovania | interný | 2 | OUT | Súbor (XLS/CSV) | 1 | 60 |
| manuálne vstupy | Týždenné štatistické ukazovatele | interný | 2 | OUT | Súbor (XLS/CSV) | 8 | 100 |
| manuálne vstupy | Ročné štatistické ukazovatele | interný | 2 | OUT | Súbor (XLS/CSV) | 2 | 20 |
| manuálne vstupy | Manuálne výkazy a kompilácie údajov NBS | interný | 2 | OUT | Súbor (XLS/CSV) | 3 | 40 |
| Ministerstvo financií SR | Tabuľky MF SR pre spracovanie BoP, FA | externý | 2 | OUT | Súbor (XLSX) | 1 | 3 |
| oversi.gov.sk | Register trestov | externý | 4 | BI | WEB API (-) | 1 | 1 |
| RBUZ | Register bankových úverov a záruk | interný | 1 | BI | DB – Oracle (Tabuľky) | 30,6 | 24 |
| REGFAP | Register finančných agentov a finančných poradcov | interný | 1 | BI | DB – MySQL (Tabuľky) | 1 | 30 |
| Register OFV |  | interný | 2 | BI | DB – Oracle (Tabuľky) | 0,1 | 2 |
| RO | Register organizácií (pre spracovanie MFI BSI, FA) | externý | 2 | OUT | Súbor (CSV) | 0,3 | 2 |
| RPO | Register právnických osôb | externý | 1 | BI | WEB API (XML) | 2 | 10 |
| RS | Register subjektov | interný | 1 | BI | DB – Oracle (Tabuľky) | 60 | 18 |
| RUZ | Register účtovných závierok | externý | 1 | BI | WEB API (JSON) | 1 | 6 |
| SHSDB | Securities Holdings Statistics Database | interný | 2 | BI, OUT | WEB API (XML) | 170 | 4 |
| Skúšky | Databáza otázok na skúšky, cvičné testy, ich výsledky, termíny skúšok | interný | 2 | BI | DB – MySQL (Tabuľky) | 0,1 | 3 |
| Sociálna poisťovňa | Tabuľky Sociálnej poisťovne pre spracovanie BoP | externý | 2 | BI, OUT | Súbor (XLSX) | 0,2 | 4 |
| SU SR | Číselníky a registre poskytované štatistickým úradom SR (pre spracovanie GFS, IC, BoP, FA a iné) | externý | 1 | BI, OUT | Súbor (JSON, XML (štandard SDMX-ML), XLSX, CSV, ACCDB) | 1 | 30 |
| Subjekty 2.0 | Subjekty 2.0 | interný | 1 | BI | DB – MySQL (Tabuľky/JSON) | 0,4 | 4 |
| Štátna pokladnica | Tabuľky Štátnej pokladnice pre spracovanie GFS | externý | 2 | OUT | Súbor (XLSX) | 5 | 5 |
| ŠZP | Štatistický zberový portál | interný | 1 | BI | DB – Oracle (Tabuľky) | 250 | 100 |

V rámci dátovej architektúry NBS existujú aj ďalšie databázy a importy z externých zdrojov, ktoré plánujeme počas projektu integrovať. Je ich 28 v rôznom rozsahu náročnosti integrácie. Detaily sú v prílohe „Zdrojové systémy NBS DWH“. Niektoré z týchto systémov plánujeme integrovať priamo do relačnej časti DWH (registre, číselníky) a niektoré dátové zdroje plánujeme požadovať uložiť len do DaraLake Stage vrstvy pre Data discovery účely.

### Reportingová oblasť

Integrácia dátových zdrojov do konsolidovanej dátovej základne bude hlavným predpokladom pre vybudovanie analytickej dátovej vrstvy - datamarty. Sú požadované vytvoriť aj ukážkové reporty, ale predpokladáme že ďalšie reporty a analýzy si NBS vytvorí na základe naučených vedomostí z projektu.

Tieto datamarty budú slúžiť za účelom:

* „self-service“ vytvárania analytických výstupov (produktov):
  + pre podporu manažérskeho rozhodovania,
  + pre podporu strategických činností banky – dohľad, regulácia, štatistika, ochrana spotrebiteľa,
  + pre podporu operatívneho rozhodovania.
* zdieľania analytických údajov a prehľadov v interaktívnej grafickej podobe naprieč bankou,
* ukladanie a historizáciu analytických údajov,
* prepojenie individuálnych a agregovaných údajov (a ich rekonciliácií),
* vytvorenie konsolidovanej informácie o subjektoch finančného trhu.

V rámci požadovaného projektu je zadefinovaných týchto 11 analytických oblastí:

| **ID** | **Označenie datamartu** | **Fáza** | **Popis datamartu** | **Rámcový popis požadovaných prvkov** |
| --- | --- | --- | --- | --- |
| 1 | REPORTS | 1 | * Datamarty pre štatistické údaje vyzbierané cez výkazy portálu ŠZP. | * časové rady ekonomických položiek, medzimesačné zmeny objemov súvahových položiek podľa rôznych ekonomických členení, výpočet vážených priemerov, etc. |
| 2 | MFI\_ESAS\_ANACREDIT | 1 | * Datamarty pre údaje o bankových úveroch a ich protistranách. * Datamart pre konsolidáciu agregovaných údajov ŠZP (menová a finančná štatistika) a individuálnych údajov RBÚZ. * Datamart na rekonciliáciu údajov ESAS (COREP a FINREP) na RBÚZ, resp. kosolidované porovnanie týchto dát | * časové rady ekonomických veličín (nominálna hodnota, priemerná úroková sadzba, etc.), prehľady podľa subjektu, nástroja, prijatého zabezpečenia, rizika protistrany, etc. konsolidované porovnanie nominálnych hodnôt agregátov menovej štatistiky a individuálnych nástrojov s možnosťou drill-down analýzy v zmysle BSI metodiky časové rady položiek, zmeny medzi referenčnými dátumami položiek (objemy a percentá), poradia subjektov |
| 5 | AML | 1 | * Datamart pre spracovanie dát AML pre risk assessment. | * výpočet AML rizikových indexov z výkazov ŠZP |
| 6 | EXR | 1 | * Datamart pre kurzový lístok | * prehľad kurzového lístka, časové rady výmenných kurzov |
| 4 | SECURITIES | 2 | * Datamart pre konsolidáciu údajov o cenných papieroch CSDB a vykazovaných údajoch o cenných papieroch ŠZP | * konsolidované porovnanie atribútov cenných papierov z výkazov ŠZP voči databáze CSDB a SHSDB |
| 11 | ESMA | 2 | * Datamart pre cenné papiere Datamart pre ESMA dáta | * Obsah reportov približne obdobne ako v publikáciách ESMA:  https://www.esma.europa.eu/file/122189/download?token=kTi4gzPC https://www.esma.europa.eu/file/122190/download?token=Cd0YfTB2 |
| 16 | MACROP\_\_\_1 | 2 | * Datamart pre monitoring fin. inštitúcií + CBD * Datamart pre proticyklický smerovník | * Reporty zamerané na základný monitoring a analýzu rizík predovšetkým v bankovom sektore v členení podľa jednotlivých oblastí - ziskovosť, kapitálová primeranosť, kvalita aktív, štruktúra bilancie, likvidita a pod. Možnosť uskutočňovať medzinárodné porovnania indikátorov pomocou údajov z databázy ECB SDW. * Report na výpočet konsolidovaných bankových údajov (CBD), ktoré sa zasielajú na pravidelnej báze do ECB. * Reporty zamerané na základný monitoring a analýzu rizík v nebankových sektoroch v členení podľa jednotlivých sektorov - DSS, DDS, KI, OCP. Report ma kvantifikáciu trhových rizík a výpočty stresového testovania pre fin, inštitúcie. |
| 19 | MACROP\_\_\_3 | 2 | * Datamart pre podnikový sektor | * Na komplexné pokrytie monitoringu podnikového sektora (primárne na účely finančnej stability) |
| 9 | ESAS | 3 | * Datamart pre dohľad nad kapitálovým trhom * Datamart pre dohľad nad platobnými službami * Datamart pre risk-based supervison * Datamart pre dohľad nad poisťovníctvom * Datamart pre údaje bankového dohľadu v portáli ESAs | * súvaha, výsledovka, analýza 1 subjektu, trhová analýza, zloženie portfólia aktív, výkonnosť, časové rady, medziročné zmeny, vážené priemery, definovanie vhodného metapopisu na integrované dáta, risk dashboard, SREP skóre, kapitálová primeranosť, pákový efekt, systémové riziko. * Sektory: centrálny depozitár, burza, obchodníci s cennými papiermi, správcovské spoločnosti a ich fondy, limity pre fondy časový pohľad na vývoj vybraných atribútov (napr. fin. ukazovateľov) 1 subjektu vrátane agregovaného výstupu za sektor vo forme tabuľky príp. grafu vrátane medzimesačného a medziročného porovnania; * analýza 1 subjektu v podobe tabuľky (zahŕňajúca údaje z viacerých zdrojov od finančných ukazovateľov, rôzne kvalitatívne zdroje napr. údaje o osobách, konaniach...) * risk dashboard zahrňujúci analýzu časových radov, prehľad osôb, etc. časový pohľad na portfólio aktív v tabuľkových a grafických prezentáciách (statické tabuľky, kontingenčné tabuľky, koláčové grafy, etc.), medzimesačné, medziročné vývojové indexy časové rady položiek, zmeny medzi referenčnými dátumami položiek (objemy a percentá), poradia subjektov |
| 17 | MACROP\_\_\_2 | 3 | * Datamart pre databázu fin. aktív držaných fin. inštitúciami | * Reporty zamerané na základný monitoring a analýzu rizík v nebankových sektoroch v členení podľa jednotlivých sektorov - DSS, DDS, KI, OCP. * Report ma kvantifikáciu trhových rizík a výpočty stresového testovania pre fin, inštitúcie. |
| 14 | ASDR | 4 | * Datamart pre účely ASDR * Datamart pre dohľad nad kapitálovým trhom * Datamart pre dohľad nad kapitálovým trhom | * Spájanie údajov (kvalitatívne, kvantitatívne a procesné údaje) pre účely dohľadu, ktoré má NBS k dispozícií o danom subjekte podľa jednotlivých sektorov. * Early warnings, vykazovanie načas, KPIs, benchmarkové porovnania. * Prehľady subjektov podľa sektorov, pôsobenie osôb vo funkciách, konflikty záujmov, nedôverihodnosť, odborná spôsobilosť... |

V nasledujúcej tabuľke je kvantifikácia komplexnosti dátových oblastí pre BI reporting. Každá oblasť je definovaná požadovanou dátovou integráciou, predpokladanou náročnosťou výpočtu jednotlivých KPI a predpokladaným rozsahom faktových tabuliek. Počet a rozdelenie dátových oblastí sa môže meniť (ale nie zásadne) pri návrhu realizovaného riešenia.

| **ID** | **Označenie datamartu** | **Počet dátových zdrojov na integráciu** | **Odhad počtu DIM tabuliek** | **Odhad počtu Faktových tabuliek** | **Odhad počtu atribútov** | **Odhad počtu metrík** | **Zložitosť výpočtu metrík** | **Počet vzorových reportov** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | REPORTS | 3 | 100 | 50 | 400 | 50 | Stredná | 10 |
| 2 | MFI\_ESAS\_ANACREDIT | 4 | 67 | 3 | 270 | 30 | Nízka | 5 |
| 5 | AML | 1 | 14 | 1 | 80 | 50 | Vysoká | 2 |
| 6 | EXR | 1 | 3 | 2 | 10 | 0 | Nízka | 1 |
| 4 | SECURITIES | 4 | 18 | 2 | 100 | 10 | Nízka | 2 |
| 11 | ESMA | 8 | N/A | 25 | 1000 | 100 | Vysoká | 10 |
| 16 | MACROP\_\_\_1 | 4 | N/A | N/A | N/A | 100 | Vysoká | 5 |
| 19 | MACROP\_\_\_3 | 6 | N/A | 10 | 100 | 50 | Nízka | 4 |
| 9 | ESAS | 13 | 75 | 25 | 500 | 100 | Stredná | 10 |
| 17 | MACROP\_\_\_2 | 5 | 75 | 25 | 500 | 10 | Vysoká | 5 |
| 14 | ASDR | 12 | N/A | N/A | N/A | N/A | N/A | N/A |
|  | 11 datamartov |  | 352 | 143 | 2960 | 500 |  | 54 |

Poznámka 1: Odhady komplexnosti jednotlivých DataMartov vychádza z výsledkov Pilotu DWH a metodiky vývoja podľa Ralpha Kimbala

Poznámka 2: Detailná tabuľka je v prílohe „DataMarty NBS DWH“ .

### Výstupné dátové oblasti

Druhá veľká časť výstupov NBS DWH sú výstupov dát v presne definovanej štruktúre pre potreby odosielania týchto dát mimo NBS.

Cieľom tejto časti je rámcový popis požiadavky na automatizáciu procesov diseminácie v štatistickom spracovaní dát z pohľadu:

* integrácie požadovaných dát z identifikovaných dátových zdrojov (referencia na tabuľku) do databáz DWH/DL,
* vytvorenia dátových modelov pre účely štatistického spracovania dát v prostredí databáz DWH/DL a nástrojov BI,
* návrhu transformačných a výpočtových procedúr vo vrstvách DWH/DL a nástrojov BI
* prípravy a archivácie výstupov v exportných formátoch (GESMES/TS / SDMX/ML),
* automatizácie kontrol dátovej kvality výstupov.

Výstupy zo štatistického spracovania je možné kategorizovať do nasledovných oblastí:

* odvetvové
* nástrojové
* komplexné

V rámci odvetvových a nástrojových oblastí spracovania je požadované zaviesť automatizované a semi-automatizované procesy, a generovanie výstupných formátov (GESMES/TS / SDMX/ML / TSV/CSV) podľa špecifikácie konkrétneho spracovania. V súčasnej dobe sú využívané dostupné self-service nástroje, pričom logiky spracovania sú decentralizované v rámci týchto nástrojov (VBA, self-service BI, SQL, manuálna práca MS EXCEL). Rámcové ohraničenie rozsahu a štruktúry dátových zdrojov, nad ktorými sa budú vykonávať výpočtové operácie sú uvedené v úvode tejto kapitoly.

V prípade výstupov označených ako komplexné, nie je vzhľadom na povahu a rozsah možné zabezpečenie plne automatizovaného „end-to-end“ spracovania a to najmä kvôli častej potrebe manuálnych korekcií a ad-hoc intervenciám v procese kompilácie výstupov zo strany zostavovateľov štatistík. V tomto prípade bude požadované minimálne zabezpečenie aktuálnych a dostupných dát zo všetkých dátových zdrojov pre efektívne spracovanie výstupov prostredníctvom na to určených „self-service“ nástrojov.

Oblasti štatistického spracovania sa pri tvorbe výstupných súborových formátov vyznačujú rôznorodosťou potrebných dátových zdrojov v kombinácií s ich vysokou dostupnosťou. Tieto dátové zdroje možno rozčleniť na:

* manuálne dátové vstupy (dáta, ktoré nemožno získať z interných alebo externých systémov a sú výsledkami práce expertov danej domény),
* dáta z interných systémov banky (interné databázy, portály, súbory),
* dáta z externých inštitúcii (štatistické údaje v rôznych prenosových formátoch, ktorých rozsah a spôsob výmeny je dohodnutý v bilaterálnych zmluvách).

Zoznam **štatistických domén**, v ktorých je požadované zvýšenie úrovne automatizácie, sa nachádza v tabuľke nižšie:

| Výstup | Doména | Typ | | | Spôsob spracovania dát v DWH | | | | Počet dátových zdrojov |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Komplexné | Sektorové / odvetvové | Nástrojové | príprava dátových zdrojov | Transformácia dát | Tvorba OUT | Nastavenie DQ |
| DM.001 | Government Finance Statistics | X |  |  | X |  |  |  | 24 |
| DM.002 | MFI Interest Rate Statistics |  | X |  | X | X | X | X | 2 |
| DM.003 | MFI Balance Sheet Items |  | X |  | X | X | X | X | 4 |
| DM.004 | Securities Issues |  |  | X | X |  |  |  | 9 |
| DM.005 | Securities Holdings SHSE |  |  | X | X | X | X | X | 2 |
| DM.006 | Securities Holdings SHSS |  |  | X | X |  |  |  | 4 |
| DM.007 | Investment Funds |  | X |  | X | X | X | X | 3 |
| DM.008 | Factoring, Leasing and Consumer Credit Companies |  | X |  | X | X | X | X | 1 |
| DM.009 | Financial Accounts | X |  |  | X |  |  |  | 8 |
| DM.010 | Insurance Corporations, (sektorovo / odvetvové) |  | X |  | X | X | X | X | 8 |
| DM.011 | Pensions Funds |  | X |  | X | X | X | X | 6 |
| DM.012 | Balance of Payments | X |  |  | X |  |  |  | 13 |
| DM.013 | Payment statistics |  |  | X | X | X | X | X | 3 |
| DM.014 | Payment institutions and institutions of electronic money |  | X |  | X | X | X | X | 9 |
| DM.015 | Bank for the Accounts of Companies Harmonised |  | X |  | X |  |  |  | 5 |

V DWH databáze sú požadované **vytvorenie v OUT vrstve** príslušné tabuľky v štruktúre požadovanej pre jednotlivé výstupy. Výstupy budú generované ako 1:1 tieto tabuľky so zmenou do požadovaného formátu:

|  | Doména | Cieľ | Výmenný formát |
| --- | --- | --- | --- |
| OP.001 | MFI Interest Rate Statistics | ECB EXDI | XML (SDMX-ML 3.0) |
| OP.002 | MFI Interest Rate Statistics | WEB NBS | CSV |
| OP.003 | MFI Balance Sheet Items | ECB EXDI | XML (SDMX-ML 3.0) |
| OP.004 | MFI Balance Sheet Items | WEB NBS | CSV |
| OP.005 | Securities Holdings SHSE | ECB EXDI | CSV |
| OP.006 | Investment Funds | ECB EXDI | XML (SDMX-ML 3.0) |
| OP.007 | Investment Funds | WEB NBS | CSV |
| OP.008 | Factoring, Leasing and Consumer Credit Companies | WEB NBS | CSV |
| OP.009 | Factoring, Leasing and Consumer Credit Companies | ECB EXDI | XML (SDMX-ML 3.0) |
| OP.010 | Inssurance Corporations | ECB EXDI | XML (SDMX-ML 3.0) |
| OP.011 | Inssurance Corporations | EIOPA | XML (XBRL) |
| OP.012 | Pension Funds | ECB EXDI | XML (SDMX-ML 3.0) |
| OP.013 | Pension Funds | EIOPA | XML (XBRL) |
| OP.014 | Payment Statistics | ECB EXDI | XML (SDMX-ML 3.0) |
| OP.015 | Payment Statistics | EBA | XML (SDMX-ML 3.0) |

Spracovanie **vyžaduje historizáciu zmien** vstupných a aj výstupných dát vrátane archivácie výmenných formátov vo frekvencii, v akej sú zo zdrojových systémov zbierané a v cieľovom systéme spracovávané.

**Transformačné logiky jednotlivých spracovaní** sú v súčasnosti decentralizované a umiestnené v rôznych komponentoch v závislosti od použitých nástrojov. Jedným zo základných predpokladov automatizácie je zmapovanie a centralizácia týchto logík do prostredia, v ktorom sa bude spracovanie vykonávať. Tieto operácie sa opakujú a možno zhrnúť do nasledovných oblastí:

|  | Operácia | Popis |
| --- | --- | --- |
| a. | Preklad a mapovanie | Aplikovanie pravidiel pre preklad prvkov číselníkov, atribútov riadkov, stĺpcov zo vstupného na požadovaný výstupný formát.  Napr.: Preklad adresy dátového bodu národného dátového modelu na adresu dátového bodu dátového modelu ECB (typ 1:N, N:1). |
| b. | Obohacovanie metadátami | Denormalizácia, pripájanie a kombinovanie s inými dátovými zdrojmi na základe mapovacích pravidiel.  Napr.: Pripojenie atribútov cenného papiera z databázy cenných papierov na základe jednoznačného ID cenného papiera k vykázaným údajom z databázy IS ŠZP. |
| c. | Filtrovanie | Výber relevantných objektov (záznamy, ich atribúty) na základe filtrovacích pravidiel. |
| d. | Agregácia a sumarizácia | Výpočet (aj viackrokový) dátových bodov v požadovanej granularite aplikovaním sumarizačných výpočtov.  Napr.: Výpočet váženého priemeru na základe predpočítanej váhy a hodnoty, súčet za podmienky, počet, etc. |
| e. | Formátovanie | Príprava predpočítaných dát do finálnej štruktúry výstupného formátu – súboru.  Napr.: Formátovanie vypočítaných dátových bodov do formátu XML v štandarde SDMX-ML 3.0. |
| f. | Matematické operácie | Aplikovanie matematických výpočtov na dáta podľa stanovených pravidiel a metodík.  Napr.: Výpočet odhadov, prenásobenie koeficientom. |

## Požadované výstupy projektu (tabuľka po fázach)

Nasledujúce kapitoly obsahujú minimálny zoznam dodaných projektových výstupov.

#### Analýza

| Výstup | Formát | Popis |
| --- | --- | --- |
| Špecifikácia požiadaviek  (Requirements Traceability Matrix) | Dokument | Tabuľka priorít projektu, ktorá sumárne popisuje dohodnutý rozsah spôsobom „Must“, „Should“, „Could“, „Won‘t“.  Tabuľka definuje priority a hlavné ciele. |
| Opis Architektúry  (Projekt Blueprint) | Dokument | Dokument popisujúci architektúru a príslušné rozhrania v rámci implementácie riešenia, rovnako obsahujúci analýzu vplyvu implementácie projektu na NBS. Tento dokument môže byť predmetom aktualizácie v ďalších fázach projektu, ale iba v kontexte pridávania úrovne detailu. Dokument pokrýva detailné rozpracovanie nasledujúcich oblastí:   * Logická architektúra riešenia (RDBMS, integračná platforma, API, workflow, riadenie dátovej kvality, metadáta). * Fyzická architektúra riešenia, HW a SW architektúra, topológia, popis vývoja, test a produkčného prostredia vrátane configuration managment. * Požiadavky na dostupnosť (service level expectations). * High availability stratégia. * Back-up a recovery procesy. * Stratégia historizácia a archivácie dát. * Skupiny používateľov, matica a riadenie prístupových práv, security. |
| Analýza zdrojových systémov  (Interface Agreement) | Dokument | Analýza zdrojových systémov definuje (popisuje) hlavný komunikačné cesty zo zdrojových systémov do DWH. Dokument taktiež popisuje základný subjektový model zdrojového systému. |

#### Design

Táto fáza do detailu rozpracováva výstupy predchádzajúcej projektovej fáze, t.j. do detailu popisuje funkčné a nefunkčné požiadavky na riešenie. Jej obsahom sú tieto oblasti (výstupy):

| Výstup | Formát | Popis |
| --- | --- | --- |
| *Návrh akvizície a konverzie dát*  (Interface Agreement) | Dokument | * Rozhrania zdrojových a cieľových systémov, pre každý systém špecifikácia rozhrania, typu (text file, db link, API, service access atď.), typu aktualizácie (full vs. prírastkový), frekvencie aktualizácie (on-line, batch, frekvencia) a obsahu. * Architektúra dátových tokov, vrátane zoznamu a detailného popisu dátových tokov a frekvencie, závislostí, detailný návrh workflow a job scheduling managementu, vrátane notifikácie užívateľov o prebehnutej aktualizácii dát. * Inicializačný load, vrátane popisu zdrojových systémov, rozhraní, požadovaných a dostupných entít, požadovanej a dostupnej histórie, očakávaných objemov a prístupov automatizácie. * Data refresh strategy – detailný popis postupov aktualizácie dát vo všetkých vrstvách DWH a pre všetky typy dát (číselníky, dimenzie, fakty, odvodené atribúty atd.). * Procesy dátovej kvality pre data acquisition – error handling, error notification, error recovery. |
| Funkčný detailný design – Logický model a mapovanie | Dokument | * definuje logický dátový model poskytujúci detailnú definíciu biznis dátového modelu. Obsahuje entity, atribúty, vzájomné vzťahy, domény, unikátne a alternatívne identifikátory. Atribúty a domény sú opísané formátmi, povolenými hodnotami alebo rozsahmi hodnôt. * definuje hlavné predpoklady pre akvizíciu dát, mapovanie medzi zdrojovými dátami a cieľovými objektmi logickej databázy, mapovacie pravidlá a logiku, ktoré sú potrebné na vytvorenie konverzií a rozhraní v rámci riešenia DWH. Účelom dokumentu je poskytnúť potrebné informácie pre vývojový tím na vytváranie správnej extrakčnej (load) a transformačnej logiky |
| Stratégia prechodu do produkcie | Dokument | high-level opis krokov uvedenia predmetného riešenia do prevádzky. |
| Návrh prístupu k dátam | Dokument |  |
| ETL vývojové štandardy | Dokument | Detaily o využívaní ETL nástroja pre projekt. Je kritickým elementom pre samotný ETL vývoj nakoľko definuje nevyhnutné štandardy. Dokument adresuje nižšie uvedené oblasti pre ODI (Oracle Data Integrator) v kontexte štandardov pre vývoj a prístup do ETL managera:   * Vývojové cykly * Používanie verzionovania * Definícia vykonateľného „jobu“ * Definícia štandardov a pravidiel * Menné konvencie * Používanie Workflow frameworku pre definíciu závislostí |
| Prístup k testovaniu riešenia | Dokument |  |

#### Implementácia

|  |  |  |
| --- | --- | --- |
| Výstup | Formát | Popis |
| Fyzický model | SW |  |
| ETL komponenty | SW |  |
| Školenie | Školenie |  |
| Testovacie scenáre |  |  |
| Prístup k prechodu do produkcie | Dokument |  |
| Návrh iniciálneho nahratia dát |  |  |

#### Nábeh do produkcie

|  |  |  |
| --- | --- | --- |
| Výstup | Formát | Popis |
| Inštalácia na produkcii |  |  |
| Vykonanie iniciálneho loadu |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
| Finálna akceptácia |  |  |

## Prílohy

### A – Projektový plán



### B – Zdrojové systémy NBS DWH



### C – DataMarty NBS DWH



1. Daniel Linstedt, Michael Olschimke (2015). Building a Scalable Data Warehouse with Data Vault 2.0. Morgan Kaufmann. ISBN: 978-0-12-802510-9 [↑](#footnote-ref-1)