IBM

InfoSphere™ software

Trusted Information

Information Management software

DataStage - QualityStage

zkušenosti s DB2

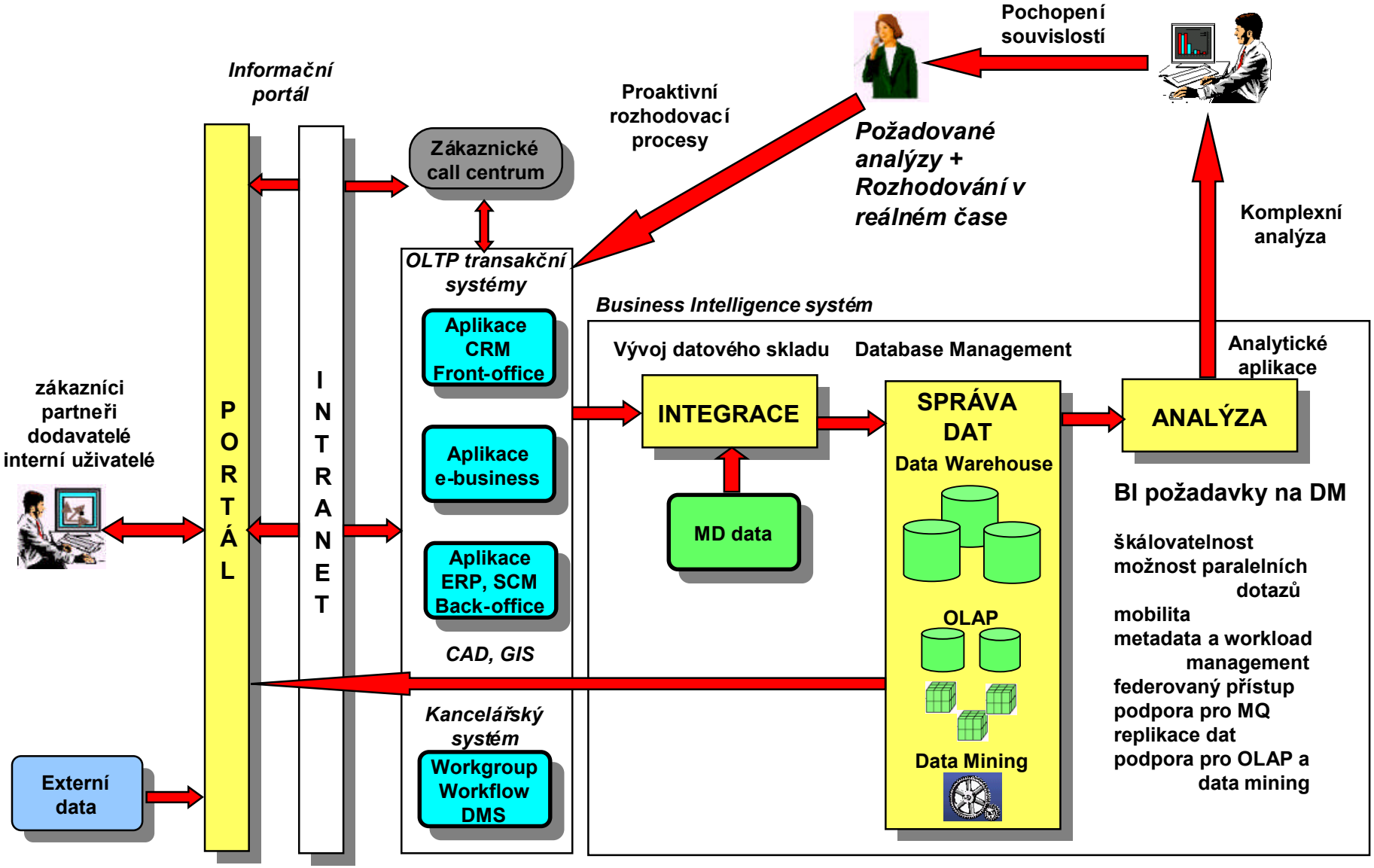
Jiří Veselý
jvesely@mfservis.cz

CIDUG 22.9.2009

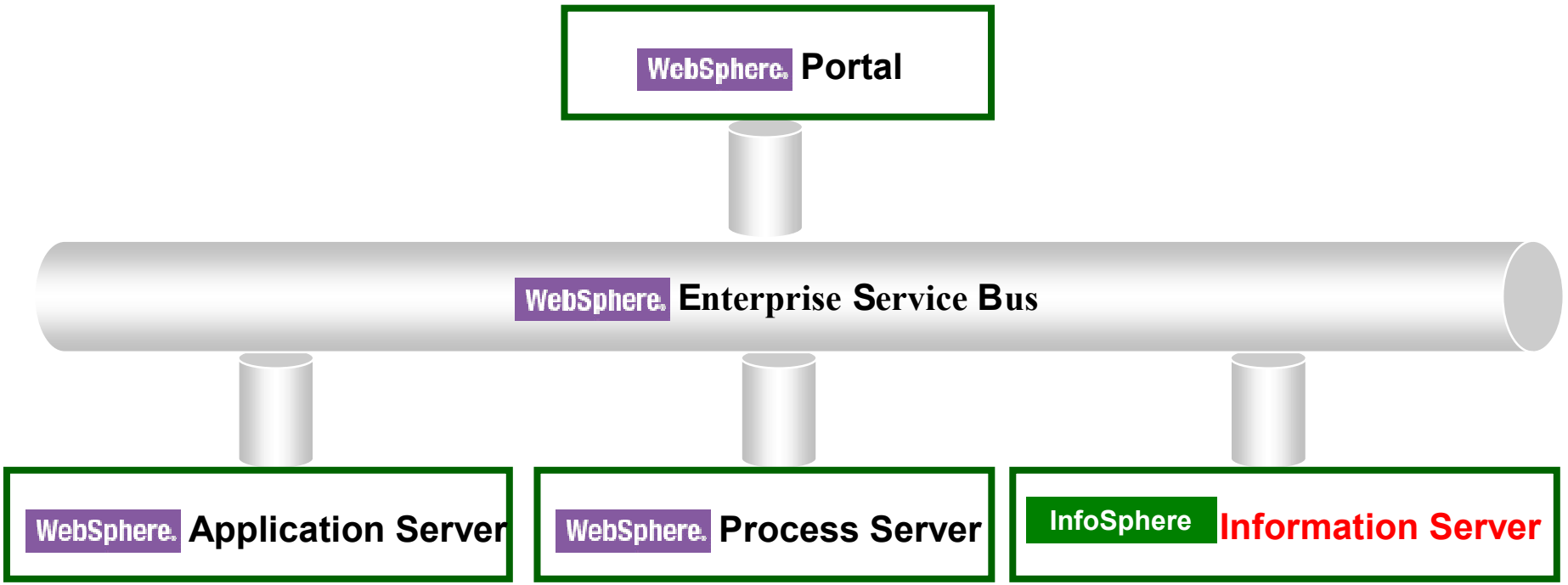
Obsah prezentace

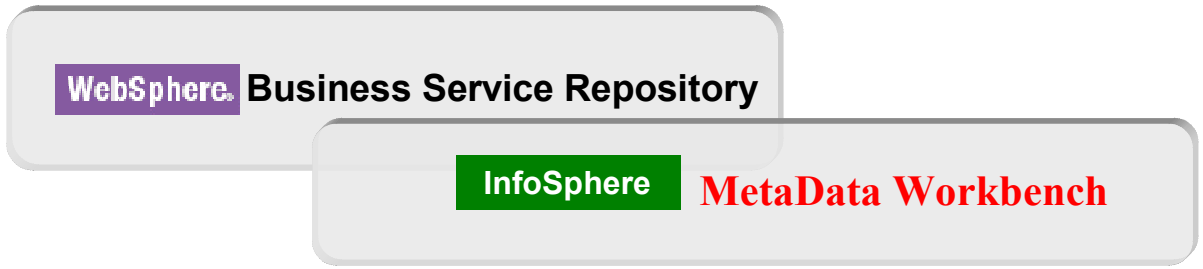
1. Integrace dat
2. InfoSphere Information Server
3. DB2, podpora analytických aplikací

IT struktura organizace



Integrace informací a procesů v SOA - pojetí IBM





integrace sdílením metadat

Obecný postup integrace dat

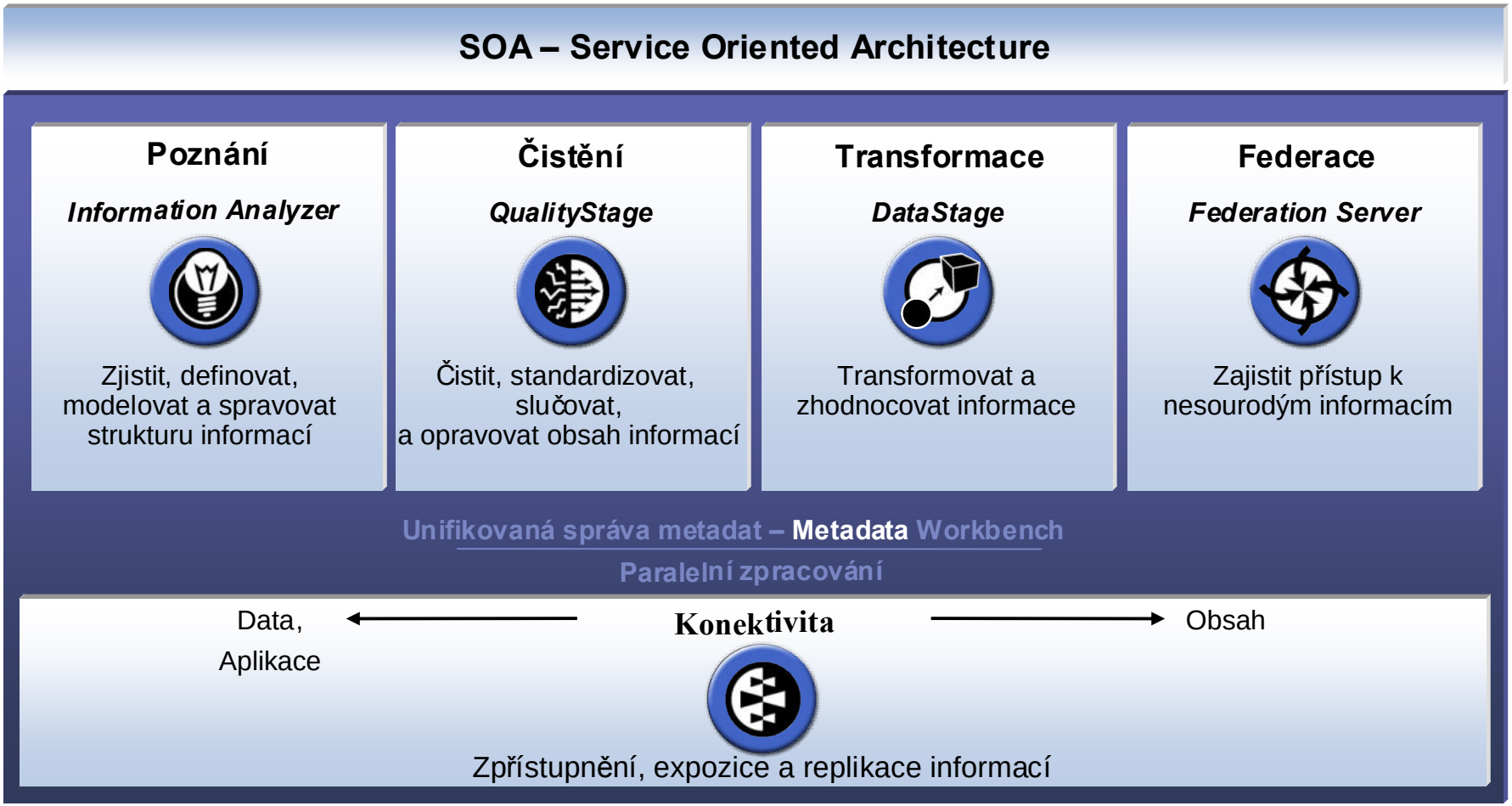
Požadavky :

- *Rychle a jednoduše přesunout informace (data) z několika zdrojů do datového skladu*
- *Vytvořit strukturu integrovaných dat vhodnou pro podporu rozhodovacích procesů*
- *Cílem jsou konzolidovaná a kvalitní data*

Postup :

- **Poznání dat**
  - ✓ znalost datových struktur a jejich vzájemných vztahů
- **Vyčištění dat**
  - ✓ nemít pochybnost o kvalitě dat
- **Transformace dat**
  - ✓ přesun dat do nového konzolidovaného úložiště – datového skladu
- **Datová federace** *(volitelná část)*
  - ✓ přístup k datům nezávisle na struktuře uložení

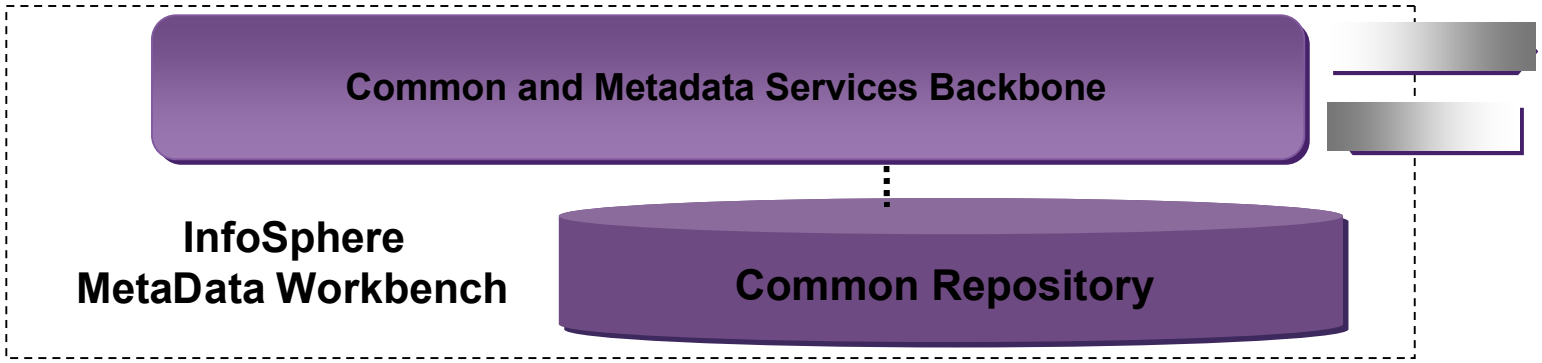
IBM InfoSphere Information Server



Co to jsou metadata ?

Metadata umožňují dát věcem obsah a smysl.

InfoSphere MetaData Workbench



- MetaData Workbench

  - **Architektura jednotné repozitory pro datovou integraci**
    - obecná repozitory pro jednotlivé technologie InfoSphere Information Serveru
    - založena na otevřeném Eclipse Modeling Frameworku

  - **Přímé využití metadat v jednotlivých nástrojích IIS**
    - dynamická a sdílená metadata
    - využití metadat napříč platformami
    - odpadá nutnost importu / exportu mezi nástroji

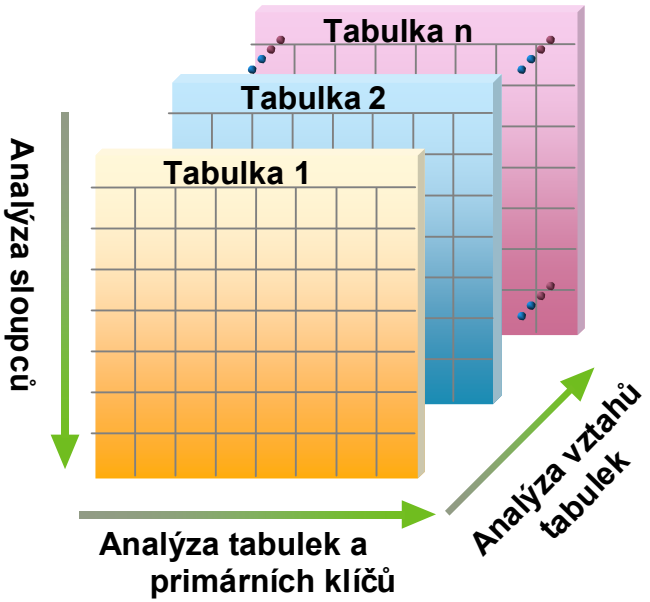
InfoSphere Information Analyzer

Analýza zdrojových systémů

- *umožňuje podrobné poznání zdrojových dat*
  - přehled datových struktur, jejich tabulek a sloupců
  - analýza sloupců tabulek
  - analýza primárních klíčů
  - analýza cizích klíčů
  - analýza obsahu, vazeb a vztahů mezi tabulkami
  - hodnocení zadaných pravidel a omezení

Analýza sloupců tabulek

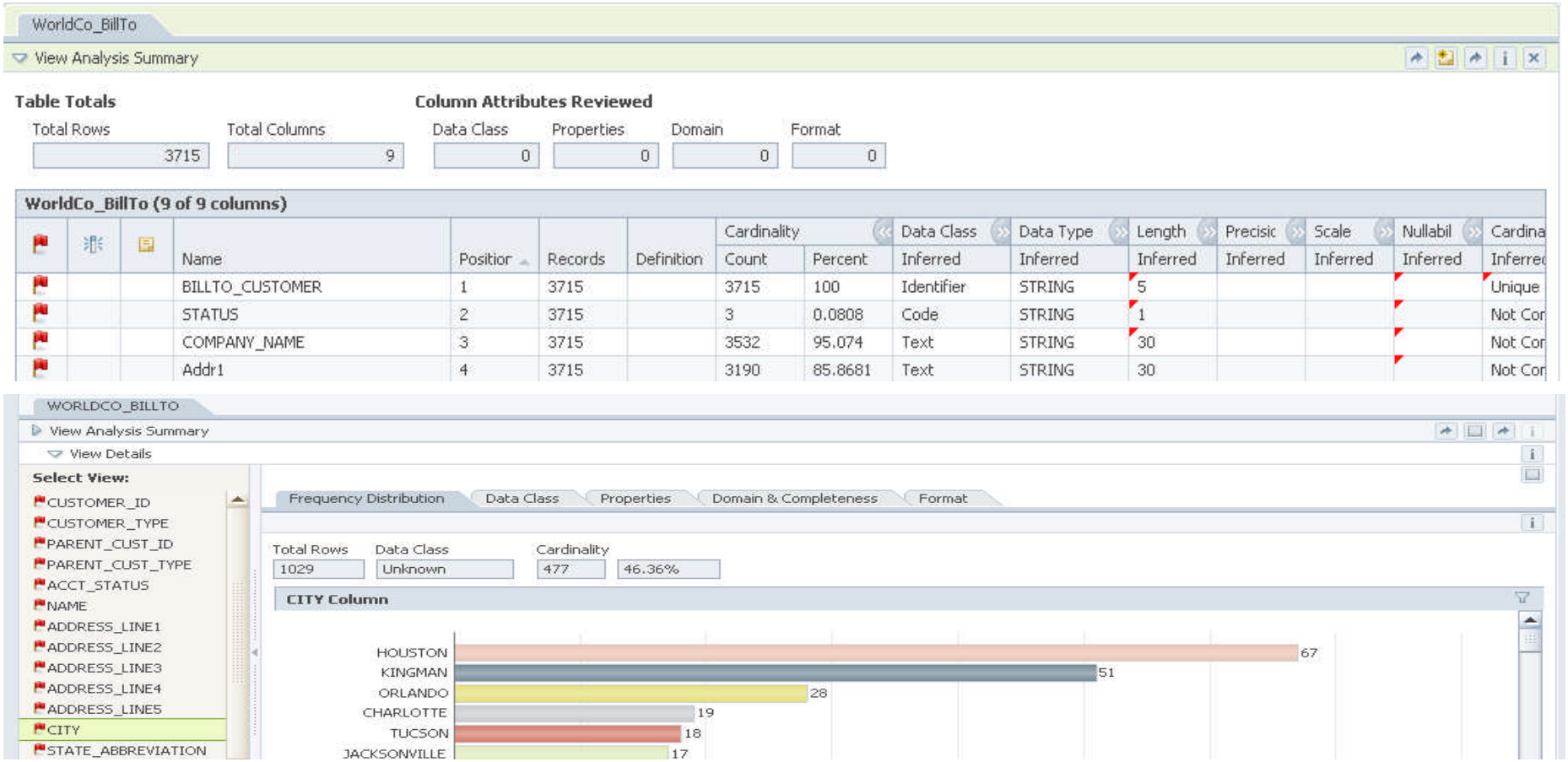
- *předmětem analýzy a reportů je zjištění parametrů :*
  - počet rozdílných hodnot nebo četnost jejich výskytu
  - počet prázdných hodnot, null hodnot
  - minimum, maximum a průměr numerických hodnot
  - základní datové typy včetně rozdílných formátů data - času
  - minimum, maximum a průměrná délka
  - přesnost a rozptyl numerických hodnot



InfoSphere Information Analyzer

Výsledky analýzy sloupců tabulek

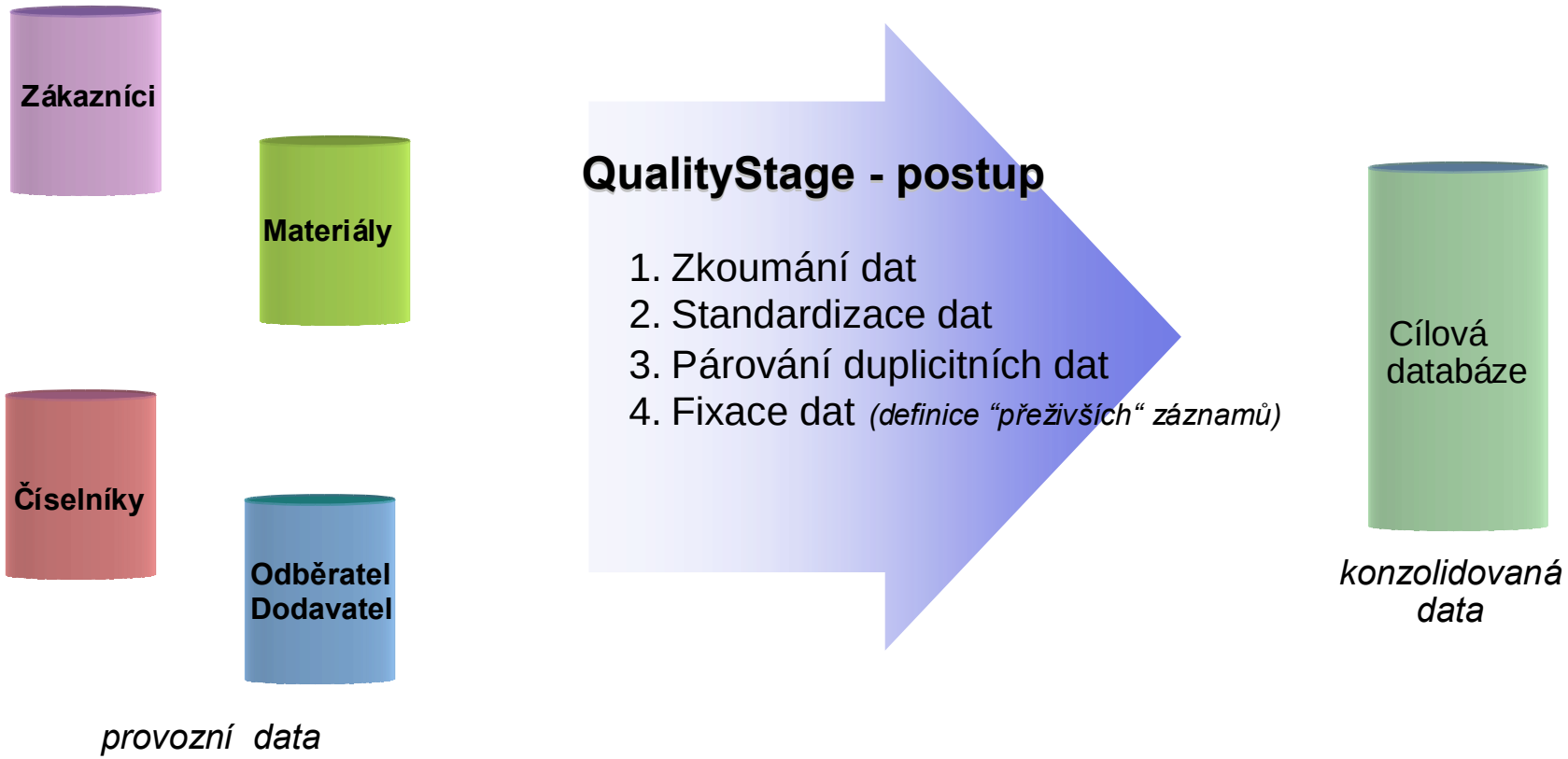
- odvození klasifikace druhu dat sloupce (identifikátor, kód,…) z obsahu
- odvození vlastností sloupce (typ dat, délka,…) z obsahu
- generování rozložení četnosti všech hodnot sloupce
- uživatel může akceptovat nebo odmítnout generované výsledky



InfoSphere QualityStage

Základní problémy a chyby obsahu dat řešené pomocí QualityStage :

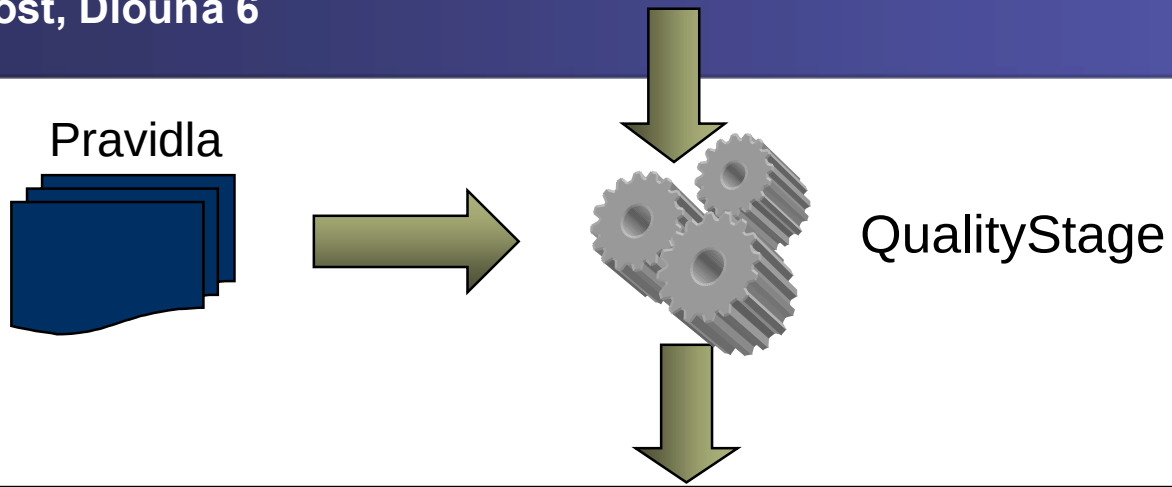
- chyby vzniklé při vkládání dat
- duplicity
- párování záznamů v různých zdrojích dat



InfoSphere QualityStage

Zkoumání a standardizace dat

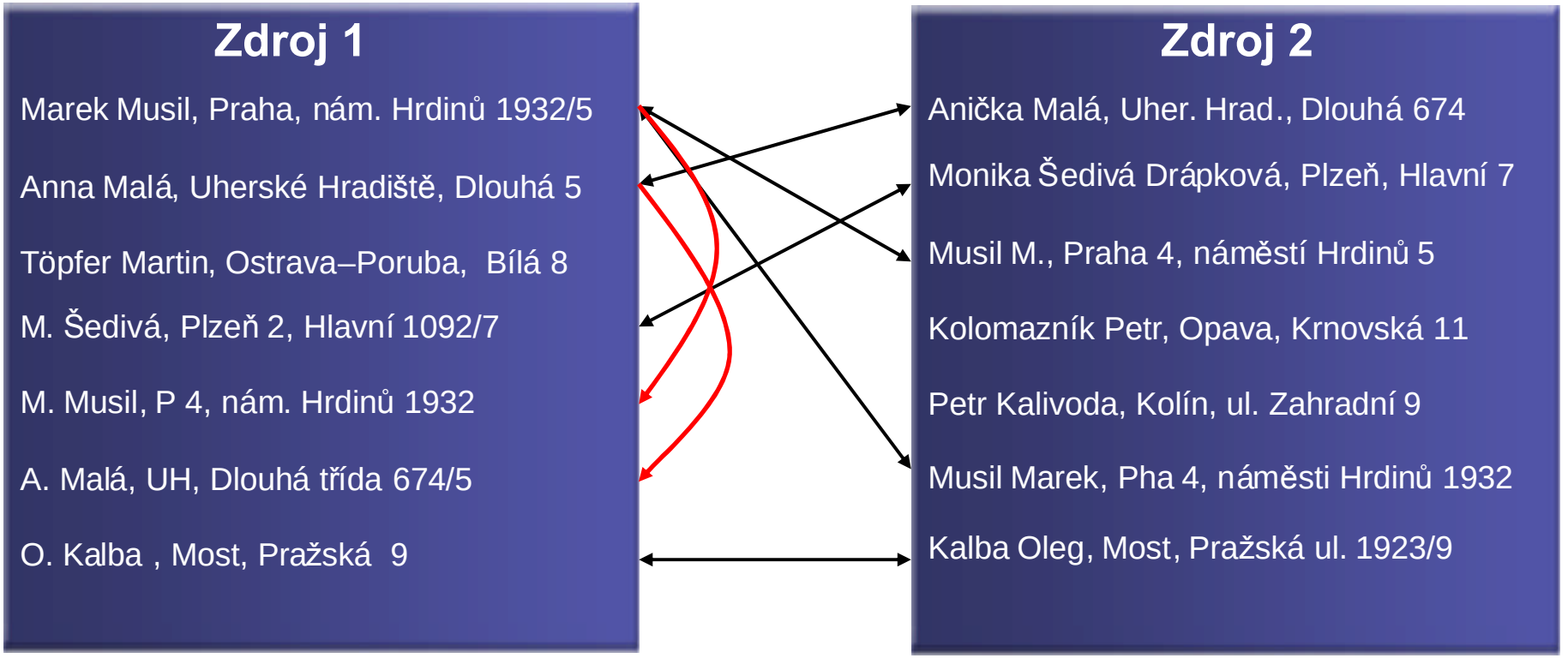
278 01 Kralupy n. Vlt., ul. Čechova 494/2
150 00 Pha, Štefánikova 32
643 00 St.Město u UH, nám Hrdinů 145
Most, Dlouhá 6



PSČ	Obec	Typ	Jméno ulice	č.p.	č.e.
278 01	Kralupy nad Vltavou	Ulice	Čechova	494	2
150 00	Praha	Ulice	Štefánikova		32
643 00	Staré Město u Uherského Hradiště	Náměstí	Náměstí Hrdinů		145
410 00	Most	Ulice	Dlouhá		6

InfoSphere QualityStage

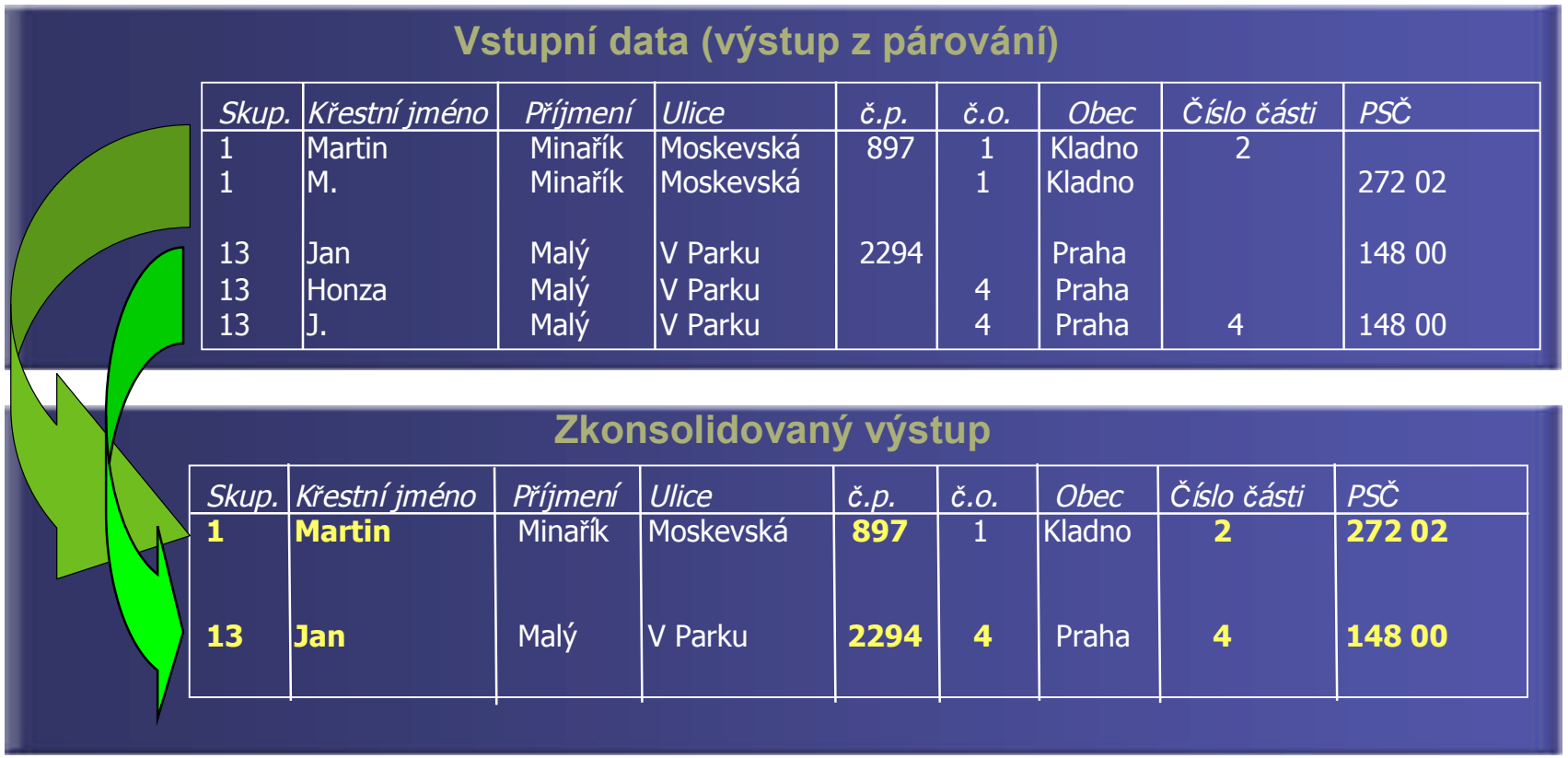
Párování záznamů



InfoSphere QualityStage

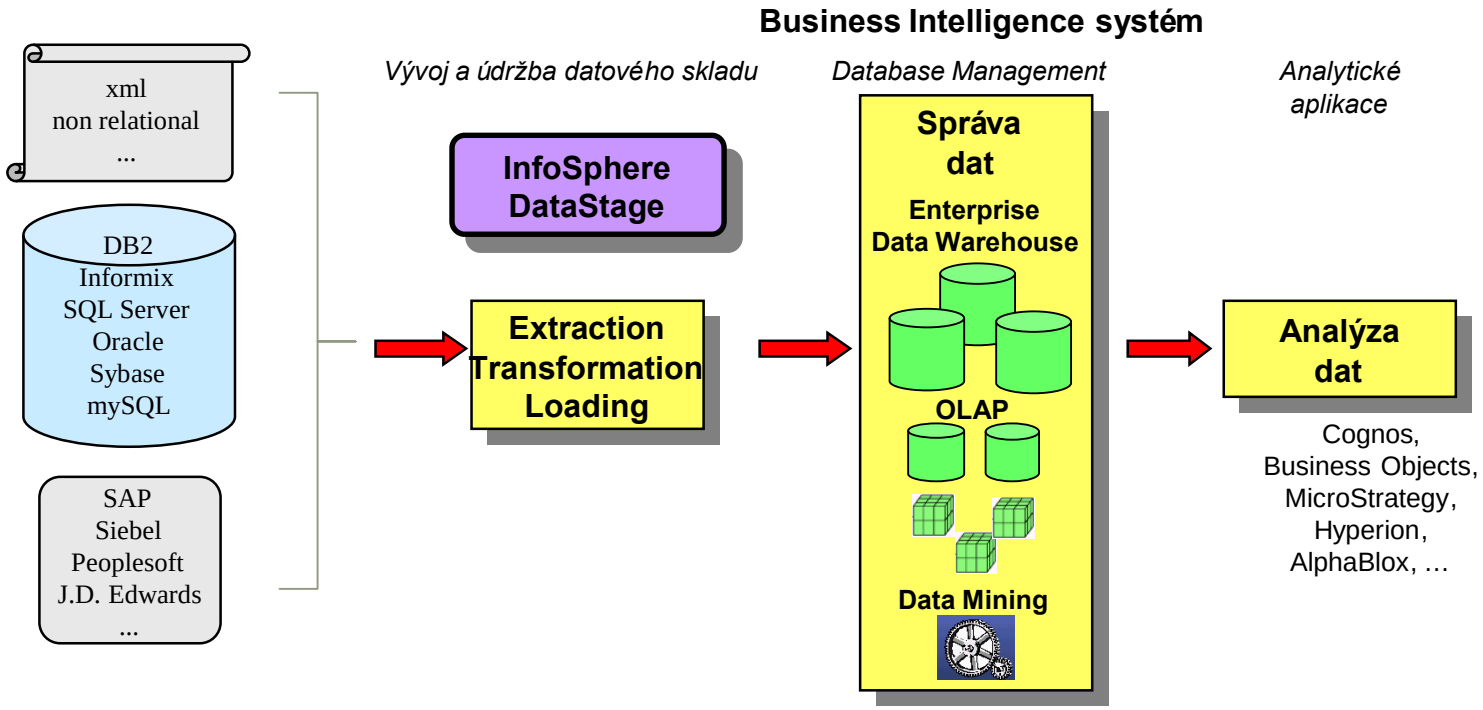
Fixace dat

- *Definice platných záznamů při redundanci*
  - automatická fixace záznamů
    - volba kritérií
      - nejblíže standardní hodnotě
      - záznam s nejvyšší četností výskytu
      - …
  - manuální fixace záznamů
    - manuální výběr platných vět
- Záznamy označené jako neplatné jsou zrušeny a všechny jejich relace jsou přepojeny k platnému záznamu



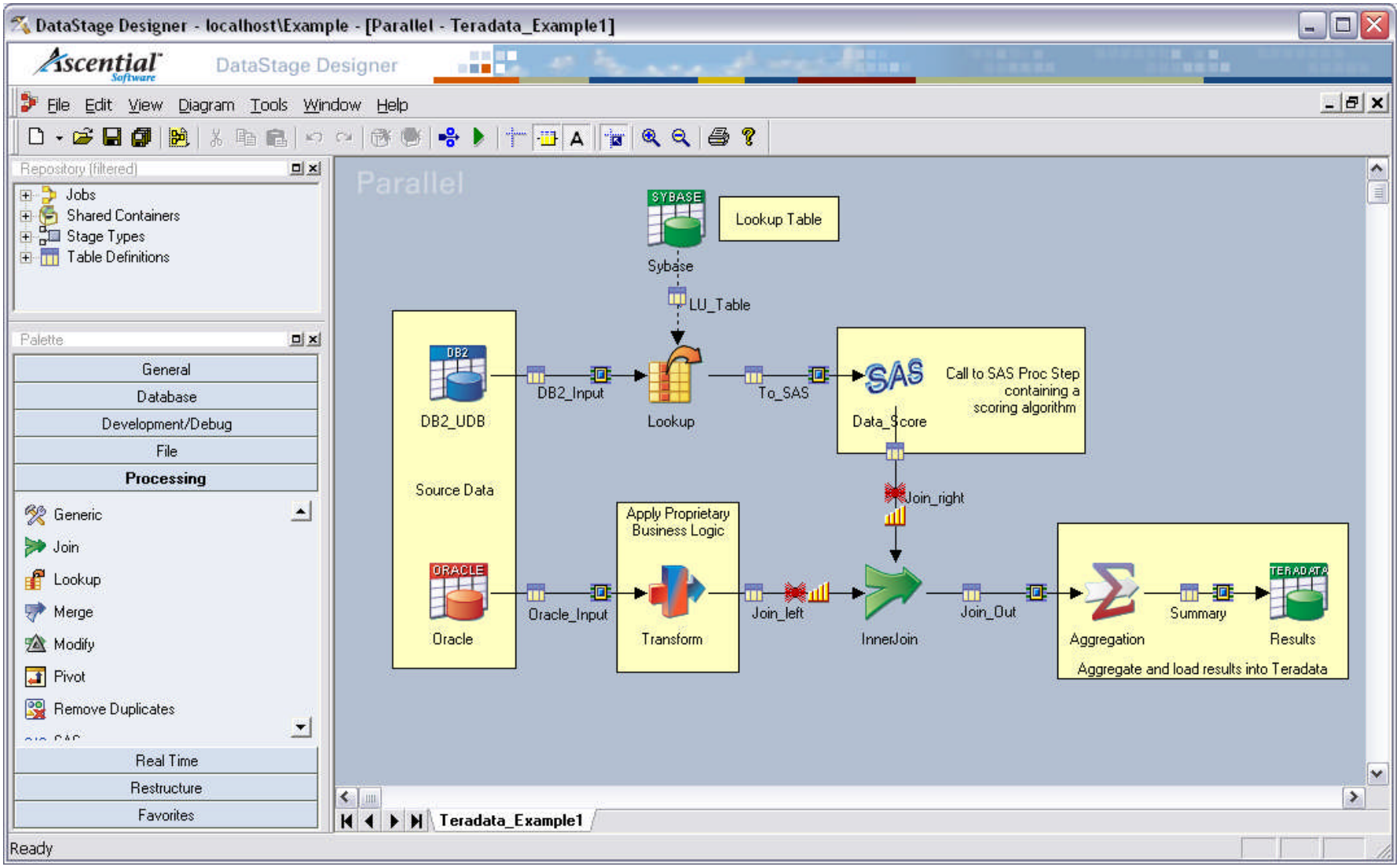
InfoSphere DataStage

- *hlavní komponenta InfoSphere Information Serveru*
  - provádí výběr, transformaci a ukládání dat (ETL)
- *integrace v rámci IBM SW*
  - podporuje SOA a WebSphere MQ
  - podporuje Cubing Services (definice, tvorba, uložení a distribuce OLAP metadat v DB2)



InfoSphere DataStage

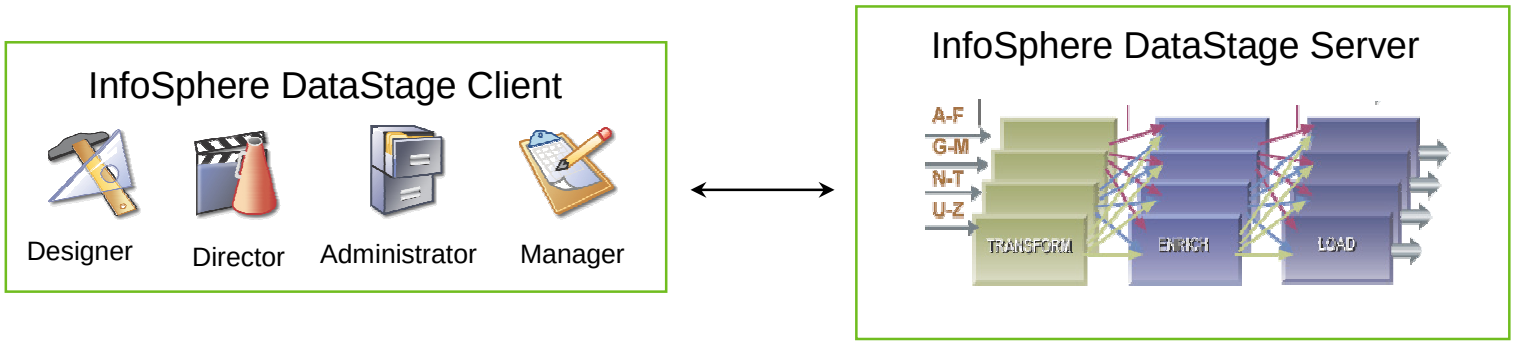
Schéma DataStage Designer



InfoSphere DataStage

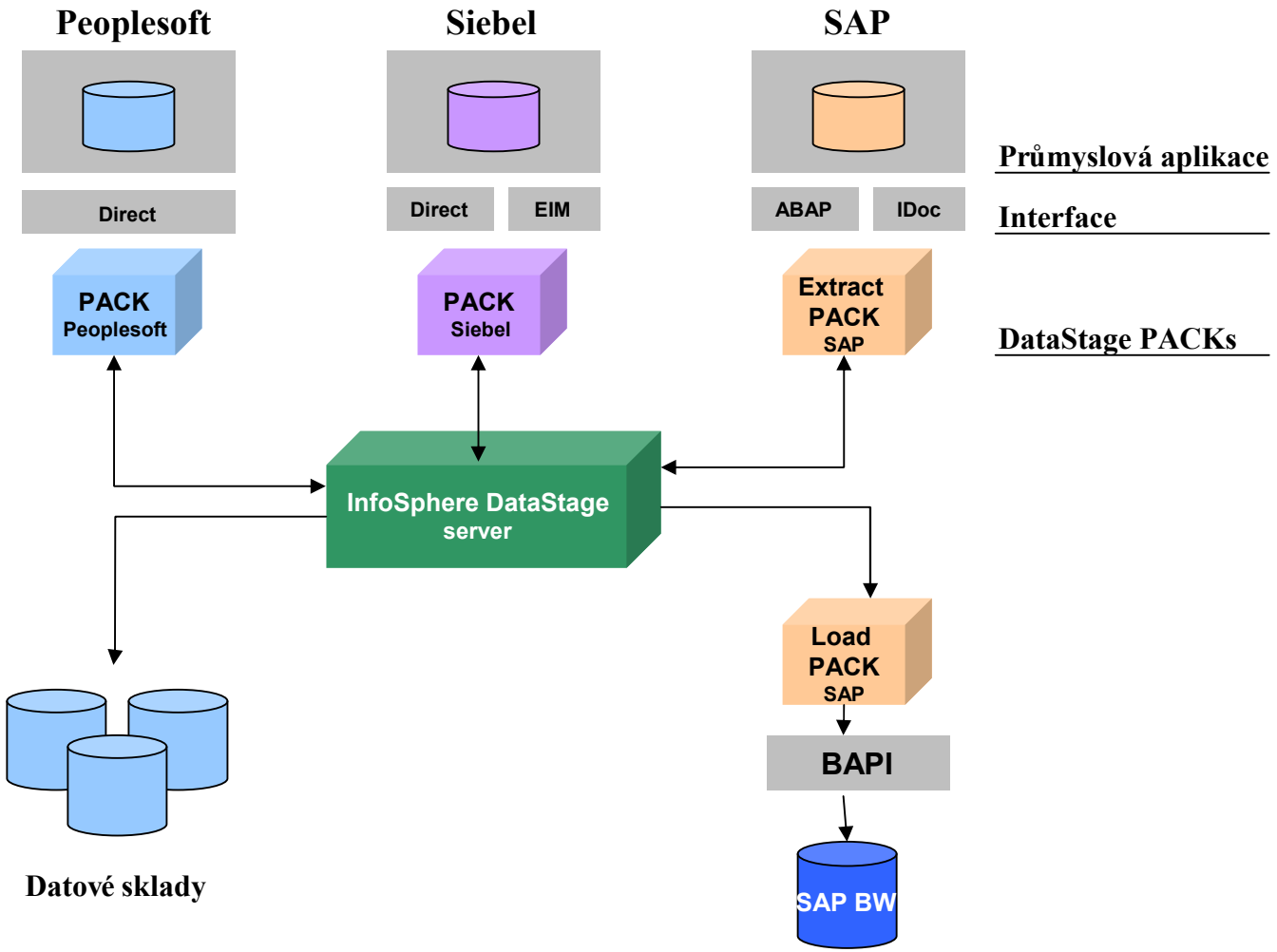
Charakteristika

- Integruje data v širokém rozsahu průmyslových datových zdrojů
- Aplikuje pravidla validace dat
- Transformuje velké objemy dat využitím paralelního zpracování
- Provádí komplexní transformace dat
- Řídí vícenásobné integrační procesy
- Poskytuje přímou konektivitu k průmyslovým aplikacím SAP, Siebel, Peoplesoft,…
- Využívá metadata pro analýzu a údržbu procesu integrace dat
- Pracuje v reálném čase, dávkovém režimu nebo jako Web Service
  (všechny transformační joby, čistící joby i federované dotazy mohou být jednak využívány jako webové služby, ale také navíc mohou webové služby využívat)
- Linux, Windows, AIX, Solaris, HP-UX, zSeries



InfoSphere DataStage

Integrace dat průmyslových aplikací

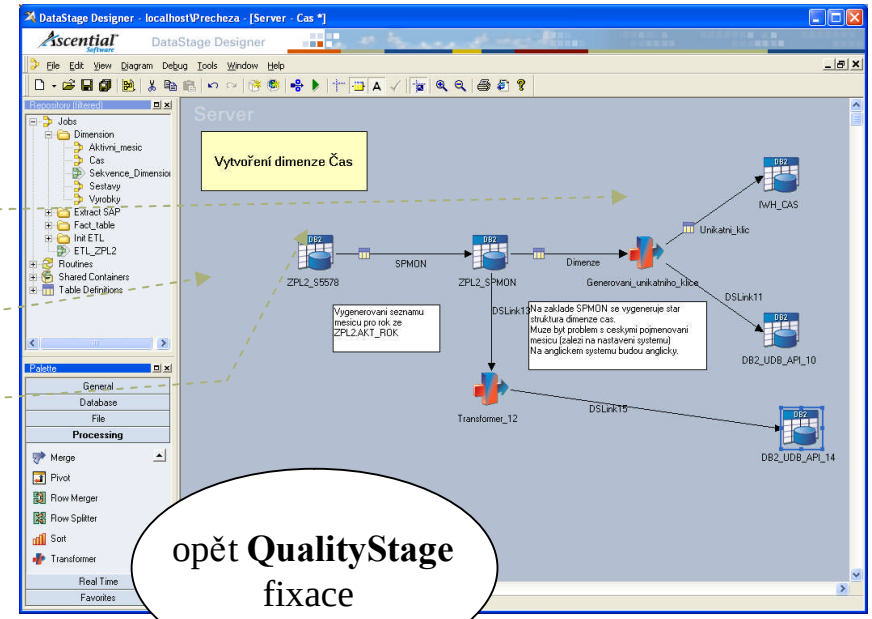


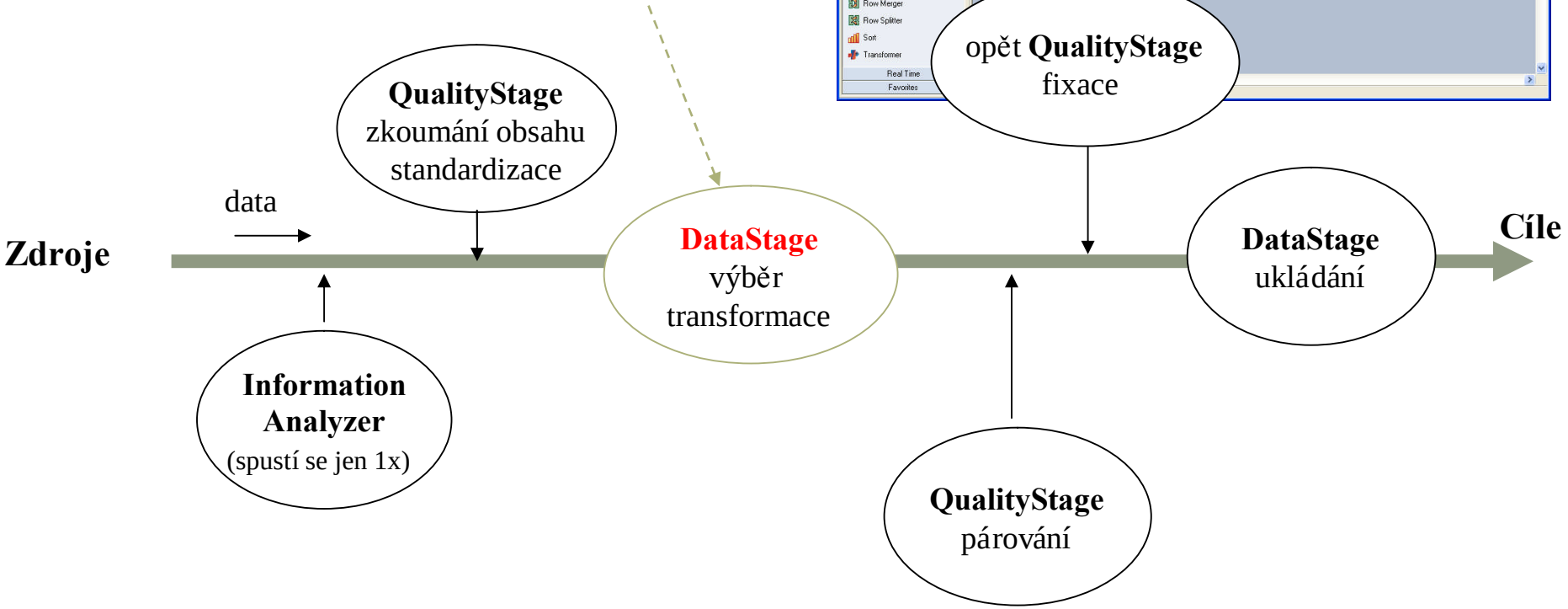
InfoSphere DataStage

Vývoj

- **Datastage Designer**
- **Transformační a další bloky (300+)**
- **Projekt**
  - úloha (Job), libovolný počet úloh
  - krok (Stage), libovolný počet kroků
- **Job Sequencer: návrh sítí závislých úloh**

Job





DB2 – podpora analytických aplikací (1)

• Portabilita, škálovatelnost, vysoká dostupnost

- DB2 pro Unix, Linux, Windows, zLinux, z/OS
- DB2 HADR
- architektura „nothing-shared“, > 10 let zkušeností s DPF v datových skladech velikosti terabytů
- SQL optimalizátor, ‚Design Advisor‘ – pro konfiguraci, pro optimalizaci datového modelu ...
- SAP používá DB2 pro vývoj a provoz BW jako strategickou databázi

• Federovaný přístup

- InfoSphere Federation Server (založen na DB2 DataJoiner, DB2 Relational Connect) umožňuje integraci dat z heterogenní datové základny
- konektory pro Informix a DB2 součástí licence DB2

• Uložení dat

- pureXML : nativní uložení XML, dotazovací jazyky: SQL + XPath + XQuery
- komprese dat : řádky, indexy, XML, temporary tabulky; redukuje kapacitu storage až o 70 %, nižší náklady na diskový prostor, vyšší výkon I/O operací

• Budoucnost BI aplikací

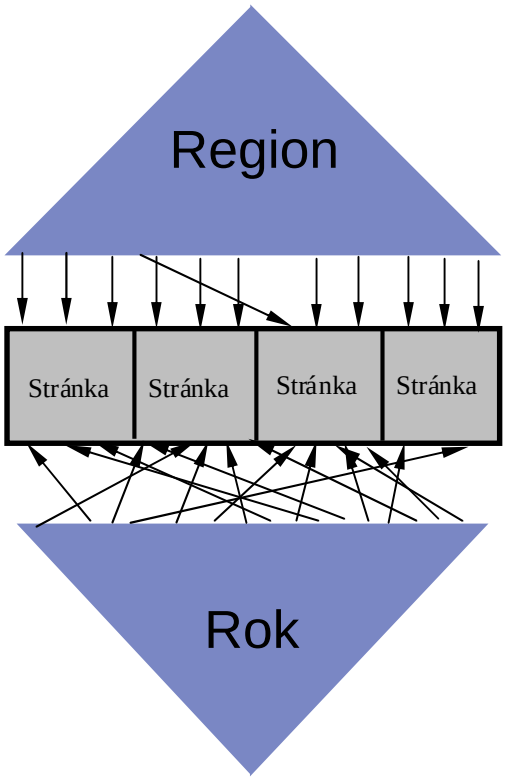
- zdroj dat v jednom relačním db stroji pro DW, OLAP i Data Mining
- výhody : jeden jazyk - standardní SQL;
  odstranění redundance dat;
  jeden typ datového modelu;
  ladění výkonu jednoho db stroje

DB2 – podpora analytických aplikací (2)

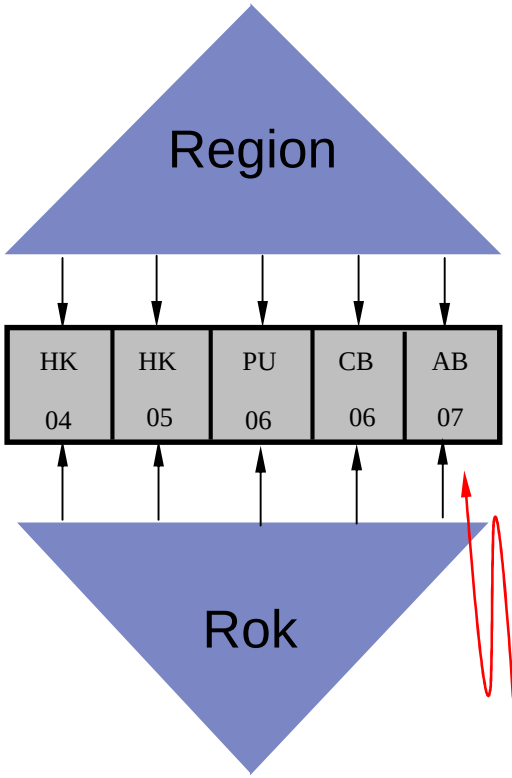
- **OLAP metadata**
  - informace o struktuře dat v datovém skladu uložených v tabulkách faktů a hierarchii všech dimenzí
  - metadata umožňují mapování multidimenzionálního modelu do relační struktury
- **DB2 Cubing Services**
  - nástroj, který provádí definici, uložení a distribuci OLAP metadat a vkládá je do katalogu DB2
- **Sdílení metadat**
  - sdílení metadat přes „bridge“ pro Cognos, Business Objects, MicroStrategy, ...
  - přímé sdílení metadat pro QMF for Windows a Office Connect
- **MQT - optimalizace a zásadní zrychlení OLAP přístupů**
  - agregace a součty v hierarchii dimenzí jsou předzpracovány v DB2 objektech - MQT (Materialized Query Tables)
  - optimizer DB2 je schopen přepsat dotaz SQL tak, aby směřoval proti MQT místo základním tabulkám
  - Cubing Services na základě popisu metadat generuje skripty pro tvorbu MQT
- **MDC – Multi Dimensional Clustering**

MDC (Multi Dimensional Clustering)

- tradiční indexy se odkazuji na záznamy
- tradiční tabulky jsou organizovány po stránkách
- tradiční tabulky mohou mít jen **jeden** cluster index
- clustering pouze v jedné dimenzi; ani tento clustering není garantován (neúčinný poté co je vyčerpán volný prostor stránky)
- přístup přes cluster index redukuje počet stránek, které mají být načteny



DB2 v7 bez MDC



DB2 od **v8 s MDC**

- MDC indexy se odkazují na bloky
- MDC tabulky jsou organizovány po blocích
- clustering ve více dimenzích
- každý řádek v bloku má stejnou hodnou MDC dimenze
- MDC tabulky mohou mít jak MDC tak i normální indexy

*všechny řádky v tomto bloku jsou z regionu **AB** a z roku **2007***

```
CREATE TABLE MDC1 (
Date DATE, Region CHAR(2), Product VARCHAR(10),
RokMesic generated as INTEGER(Date)/100, ... )
ORGANIZE BY DIMENSIONS (RokMesic, Region, Product)
```

DB2 MDC - výhody

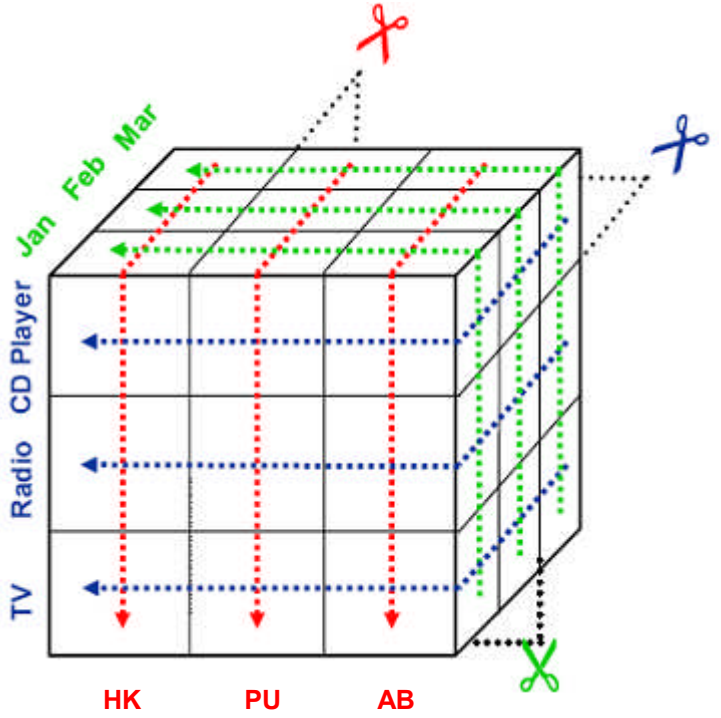
Výhody

- nižší počet I/O operací
- vyšší výkon dotazů
- rychlé rušení MDC bloků
- automatický a garantovaný clustering
- menší indexy založené na blocích
- součást ceny OEM SAP DB2

MDC - unikátní vlastnost DB2 pro MD datový model SAP NetWeaver BI

- podpora pro PSA, objekty DataStore, InfoCubes a Aggregates
- uživatelské rozhraní pro definici dimenzí MDC pro tabulky faktů InfoCube a objekty DataStore
- nástroj pro konverzi existujících objektů InfoCubes/ DataStore do MDC

SELECT vynos FROM table WHERE produkt=TV AND region=HK



DB2 komprese řádků a indexů

Tabulka bez komprese

ID	*Jméno*	*Příjmení*	*Město*	*Stát*	*PSČ*
3108	**Pavel**	**Králík**	**Pardubice**	**Česká republika**	**53002**
7749	**Jiří**	**Králík**	**Pardubice**	**Česká republika**	**53002**

Adresář

01	**Králík**
02	**Pardubice, Česká republika, 53002**

Tabulka s kompresí

3108	**Pavel**	**01**	**02**
7749	**Jiří**	**01**	**02**

- **Zkušenosti z referencí**
  - ✓ tabulky komprimované o 40-70 %
  - ✓ rychlejší dotazy

Příklad porovnání TCO

Prostředí SAP	*Bez komprese*	*S kompresí 40 %*
SAP ERP produkční	**350 GB**	**210 GB**
SAP ERP standby	**350 GB**	**210 GB**
SAP ERP testovací	**100 GB**	**60 GB**
SAP BI produkční	**1250 GB**	**750 GB**
SAP BI standy	**1250 GB**	**750 GB**
SAP BI testovací	**250 GB**	**150 GB**
Celkem	3.550 GB	2.130 GB

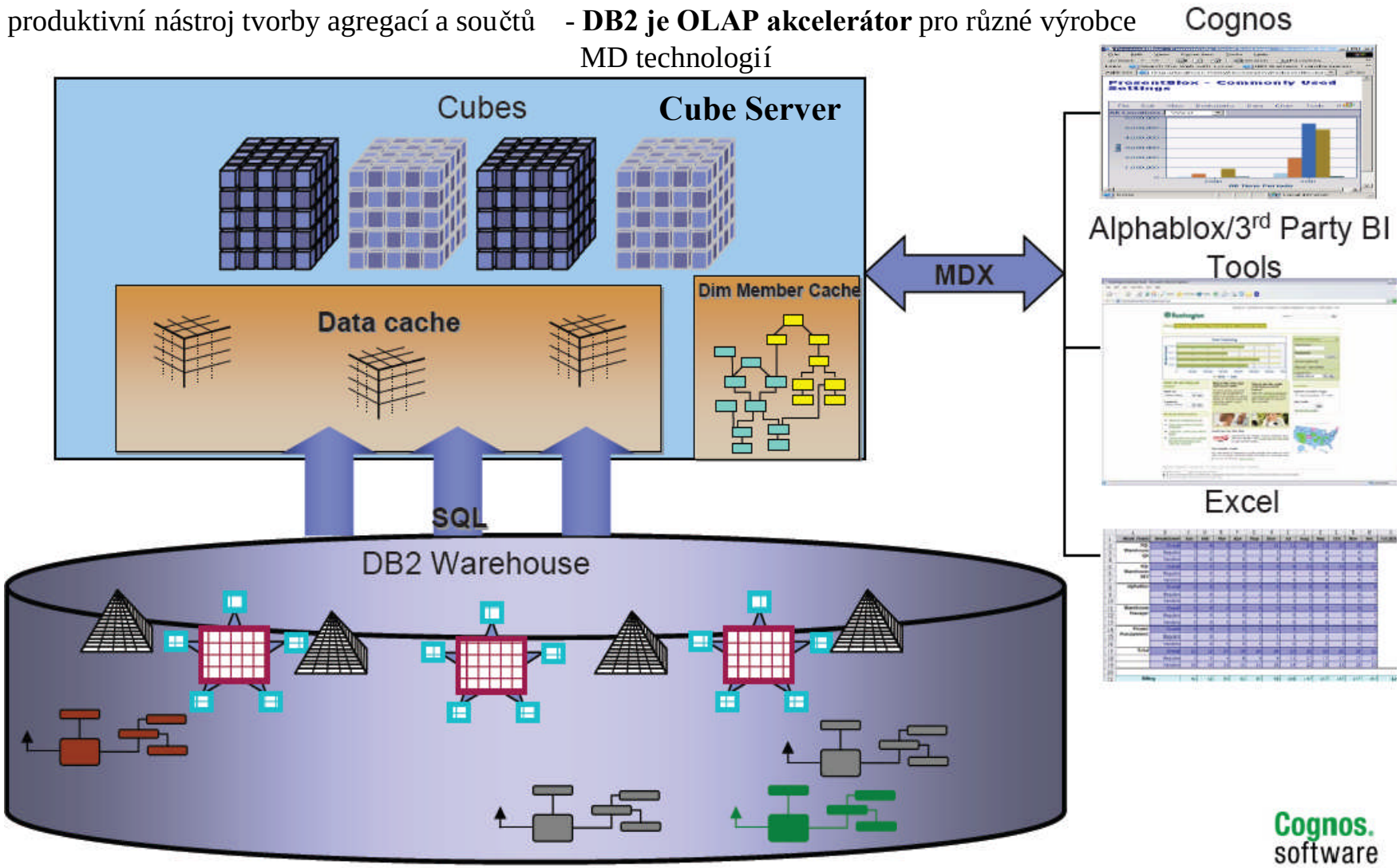
- Úspory diskového prostoru = 1.42 TB
  - ✓ 1 GB zabezpečeného prostoru na disku = cca 60 € p.a.
  - ✓ úspora = 85.200 € p.a.

- **DB2 komprese – standard pro SAP NW BI**
  - tabulky faktů (E a F)
  - tabulky ODS a PSA.

DB2 – Cubing Services

► Význam pro technologie OLAP

- vysoký výkon plnění MD struktur (kostek)
- produktivní nástroj tvorby agregací a součtů
- vysoký výkon relačních dotazů v hybridním OLAPu
- **DB2 je OLAP akcelerátor** pro různé výrobce MD technologií




konec prezentace