

# **Vysoká škola ekonomická v Praze**

**Fakulta informatiky a statistiky**

**Katedra informačních technologií**

Student : **Světlana Husáriková**  
Vedoucí bakalářské práce : **Ing. Daniel Rydzi**  
Recenzent bakalářské práce : **Miroslav Líška, DiS.**

**TÉMA BAKALÁŘSKÉ PRÁCE**

**ETL Procesy**

**Rok 2006**

## **Prohlášení**

Prohlašuji, že jsem bakalářskou práci zpracoval(a) samostatně a že jsem uvedl(a) všechny použité prameny a literaturu, ze kterých jsem čerpal(a).

V Praze dne 28.8.2006

.....  
podpis

# Abstrakt

Tato práce se zabývá ETL procesy a snaží se pokrýt jak obecné termíny, tak praktické informace a rady. Postupně vysvětluje jednotlivé fáze celého procesu, související termíny a pojmy a uvádí, kde se lze s ETL procesy nejčastěji setkat. Dále se věnuje návrhu celého ETL procesu, který je zde podrobně rozebrán od počátečních kroků, jako jsou analýzy zdrojových systémů, vytváření metadat a potřebných dokumentů, modelů datových toků, až po definování pravidel pro čištění dat.

V praktické části pak upozorňuje na mnoho běžných úskalí, se kterými se vývojáři ETL procesů setkávají a kterým se lze často vyvarovat, nebo je alespoň minimalizovat.

Následuje výběr několika nejpoužívanějších ETL nástrojů včetně shrnutí jejich hlavní funkcionality, případných výhod a nevýhod, a v jednom z těchto nástrojů, DataStage firmy IBM, je vytvořen ukázkový příklad, který se vypořádává s několika dříve nastíněnými problémy a demonstruje implementaci klasického ETL procesu.

Závěr práce se věnuje několika nejvýznamnějším technologiím, které budou s největší pravděpodobností ovlivňovat vývoj ETL procesů v nejbližších letech.

# Abstract

This study is focused on ETL processes and its goal is to cover general terms, practical information and advice. Particular stage of whole process, related terms and notions are explained step by step and the paper indicates where ETL processes can be most frequently met. Furthermore it follows project of the whole ETL process which is analyzed in detail from initial steps (such as analyses of origin systems, generation of metadata and required documents, bit rate models) to data-cleaning rules definitions.

The practical part highlights many usual issues noticed by ETL developers which can often be avoided or minimized at least.

Following is a selection of most widely used ETL tools including summary of their main features, possible advantages and disadvantages. In one of these tools, IBM DataStage, the sample case is created which deals with previously outlined issues and demonstrates classical ETL process implementation.

The final part is addressed to some most significant technologies which will most probably influence ETL processes and its evolution in the next few years.

# Obsah

Abstrakt.....	3
Abstract.....	4
Obsah.....	5
Úvod.....	7
1 Pojem ETL.....	8
1.1 Základní části ETL procesu.....	8
1.1.1 Získávání dat (Extraction) .....	8
1.1.2 Transformace dat (Transformation).....	9
1.1.3 Ukládání dat (Load).....	9
1.2 ETL jako součást Business Intelligence.....	10
2 Návrh a dokumentace ETL procesu.....	12
2.1 Analýza zdrojových systémů a tvorba metadat.....	13
2.2 Model ETL procesů a datových toků.....	15
2.3 Analýzy dat, data profiling .....	16
3 Úskalí ETL procesů.....	19
3.1 Rozdílná terminologie.....	19
3.2 Různé formáty dat.....	19
3.3 Chybějící data.....	19
3.4 Různá měřítka.....	20
3.5 Nejednoznačné údaje.....	20
3.6 Vícenásobné položky.....	20
3.7 Nedodržená referenční integrita.....	20
3.8 Duplicita dat.....	21
3.9 Vícenásobné hierarchie.....	21
3.10 Náhodné chyby .....	21
3.11 Odlišná kódování, jazyky.....	21
3.12 Nedostatečná testovací data.....	21
4 ETL nástroje.....	22
4.1 Moderní ETL nástroje.....	22
4.2 Volba správného ETL nástroje.....	23
4.3 Významné ETL nástroje.....	24
4.3.1 WebSphere DataStage TX .....	25
4.3.2 Informatica PowerCenter 8.....	26
4.3.3 Oracle Warehouse Builder.....	26
4.3.4 Enterprise ETL Server.....	27
4.3.5 Data Transformation Services (DTS).....	28
4.3.6 Data Integrator.....	28
4.3.7 Sunopsis Data Conductor .....	29
4.3.8 Podrobnější popis nástroje DataStage (IBM).....	30
4.4 Návratnost investic do nástrojů ETL.....	34
5 Praktický příklad ETL procesu.....	35
6 Budoucnost ETL.....	40
6.1 ETL, ELT nebo ETLT? .....	40
6.2 Real-time datové sklady.....	41

6.3 Datové gridy - ETL architektura příští generace.....	42
6.4 Budoucí vývoj ETL trhu.....	42
Závěr.....	44
Použitá literatura.....	45
Seznam použitých termínů a zkratk.....	46
Seznam obrázků.....	48

# Úvod

Datové sklady se staly základní datovou vrstvou, na které jsou v dnešní době budovány Business Intelligence systémy. Dříve byly datové sklady doménou především velkých společností, tato doba je však již pryč a dnes nasazují datové sklady i středně velké a menší firmy. Budování datových skladů a tržišť je tedy úkol, před kterým IT organizace stojí stále častěji, a podle odhadů specializovaných firem lze očekávat, že tento trend bude v blízké budoucnosti nadále pokračovat.

Jednou z nejdůležitějších částí budování datového skladu jsou procesy ETL, jejichž úkolem je v souladu s definovanými požadavky integrovat velká množství různorodých dat z rozličných zdrojů a platform do jednotného prostředí datového skladu.

Cílem této práce je poskytnout ucelený a aktuální pohled právě na oblast ETL procesů, kterým se zatím česká odborná literatura příliš nevěnuje. To byl také jeden z důvodů, vedle mého zájmu o datové sklady, proč jsem si toto téma zvolila.

Práce vysvětluje jednotlivé pojmy, popisuje základní architekturu a procesy, zabývá se běžnými problémy, se kterými se lze setkat během implementace, a poskytuje stručný přehled o nejvýznamnějších ETL nástrojích, předpokládaném vývoji a trendech v nejbližší budoucnosti. Obsahuje také krátký ukázkový příklad, který demonstruje realizaci ETL procesu v nástroji DataStage firmy IBM.

Doufám, že tato práce bude přínosem všem, kteří se již setkali s pojmem Business Intelligence a mají zájem proniknout hlouběji do jedné z jeho součástí - ETL. V práci jsem vycházela především ze zahraničních materiálů umístěných na internetu a praktických zkušeností osob pracujících několik let přímo v oblasti ETL v České republice.

# 1 Pojem ETL

Zkratka ETL je složena ze slov Extraction, Transformation a Load, což volně přeloženo znamená získávání, modifikace a uložení. Jedná se o sadu procesů, na jejichž základě jsou data získávána z mnoha různorodých databází, aplikací a systémů, měněna a upravována na základě stanovených pravidel a nakonec uložena do cílových systémů, což nejčastěji bývají datové sklady, datová tržiště, ale stejně tak to mohou být klasické relační databáze či textové soubory.

S termínem ETL se většinou setkáme u datových skladů, další nejčastější použití nalezneme při migraci dat mezi aplikacemi (80% datové sklady, 9% migrace dat, 11% ostatní).

Přestože ETL část bývá v počátcích projektů datových skladů často podceňována, zabírá návrh a následná implementace více než polovinu (50%-80%) času a zdrojů celého vývoje a špatně navržené ETL procesy jsou pak velmi nákladné na údržbu, aktualizace a změny. Proto je velmi důležité věnovat ETL procesům dostatek času již na začátku celého projektu a zvolit správně technologie, postupy a nástroje, které budou použity.

## 1.1 Základní části ETL procesu

### 1.1.1 Získávání dat (Extraction)

V první fázi ETL procesu je třeba získat potřebná data ze zdrojových systémů. Často se jedná o zcela odlišné zdroje dat, ať už jde o datové sklady, relační databáze, textové soubory v různých formátech, XML soubory, COBOL, ERP, CRM aplikace apod. Každý systém může navíc používat odlišnou organizaci dat a různé datové formáty či být umístěn na rozdílných platformách. V této fázi je klíčové, aby se podařilo získat všechna data ze zdrojových systémů v požadovaném čase, což může být z různých důvodů problematické. Ponecháme-li stranou úplnou nedostupnost zdrojového systému, může být problémem například to, že-li tento systém sám provozován v režimu kritické dostupnosti, takže masivní čtení z tohoto systému by představovalo nepřipustnou zátěž a je tedy třeba hledat jiná řešení – zálohové systémy, čtení pouze modifikovaných dat apod. Dalším běžným problémem bývají velké objemy dat, které mohou být v řádech desítek GB denně.



### 1.1.2 Transformace dat (Transformation)

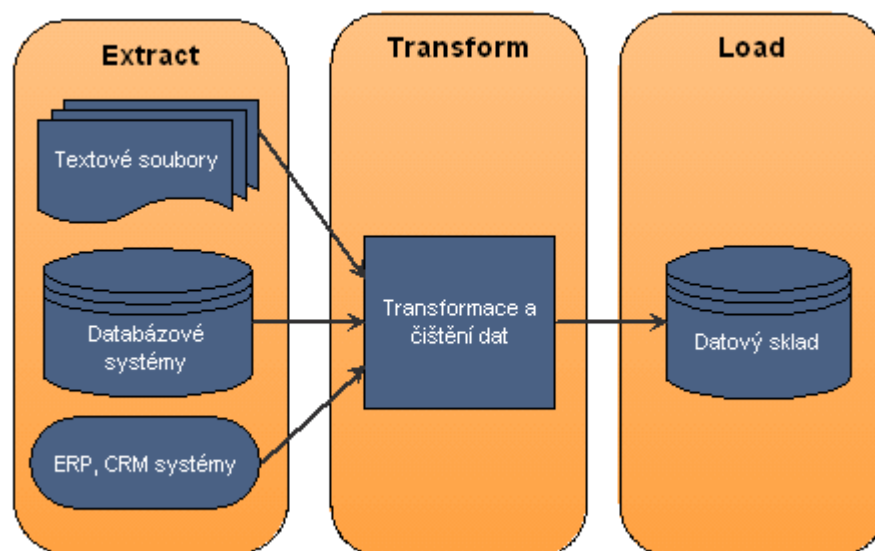
Druhou fází jsou transformace a čištění dat. Na začátku je třeba specifikovat obchodní pravidla, která musí definovat sami uživatelé. Tato pravidla říkají, jak zacházet s daty, jejich nekonzistencí, jakým způsobem řešit chybějící data či chyby v nich. V praxi se jedná o pravidla definující referenční integritu, která nám popisuje vazby mezi daty, pravidla pro zacházení s duplicitami či chybějícími daty, dále pravidla pro doménovou integritu, která definují, jakých hodnot či typů hodnot mohou data nabývat, a také pravidla týkající se dat jako takových. Nakonec se musí stanovit pravidla pro mapování zdrojových dat na cílová.

Příklad běžně používaných transformací:

- výpočet nové hodnoty (finální cena = původní cena \* (1-sleva))
- součet dat z více řádků (celkový počet zaměstnanců v pobočce)
- spojení dat z různých zdrojů (jména zaměstnanců, kteří pracují v místě, kde se vyvíjí určitý produkt)
- rozdělení dat do více polí
- sjednocení formátů a měřítek

### 1.1.3 Ukládání dat (Load)

V poslední fázi ETL procesu je potřeba získaná, upravená a pročištěná data uložit do cílového systému (nejčastěji datového skladu). Náročnost tohoto ukládání je různorodá a záleží jak na povaze zdrojových dat, tak na požadavcích na cílový systém. Některé datové sklady pouze přepisují staré informace novými, jindy je třeba dělat inkrementální aktualizace. Existují také více komplexní systémy, které udržují historii a audit všech změn v datech.



Obr. 1 - Základní části ETL procesu

## 1.2 ETL jako součást Business Intelligence

Cílem Business Intelligence nástrojů je podpora konkurenceschopnosti firmy, poskytnutí možnosti efektivně třídit, upravovat a zpracovávat data na informace a znalosti, tak aby mohly být použity jako základ pro analytické výstupy, dolování dat, sledování trendů, vývojových řad. Jinými slovy, aby poskytovaly kvalitní informace pro podporu plánování strategických rozhodnutí firmy.

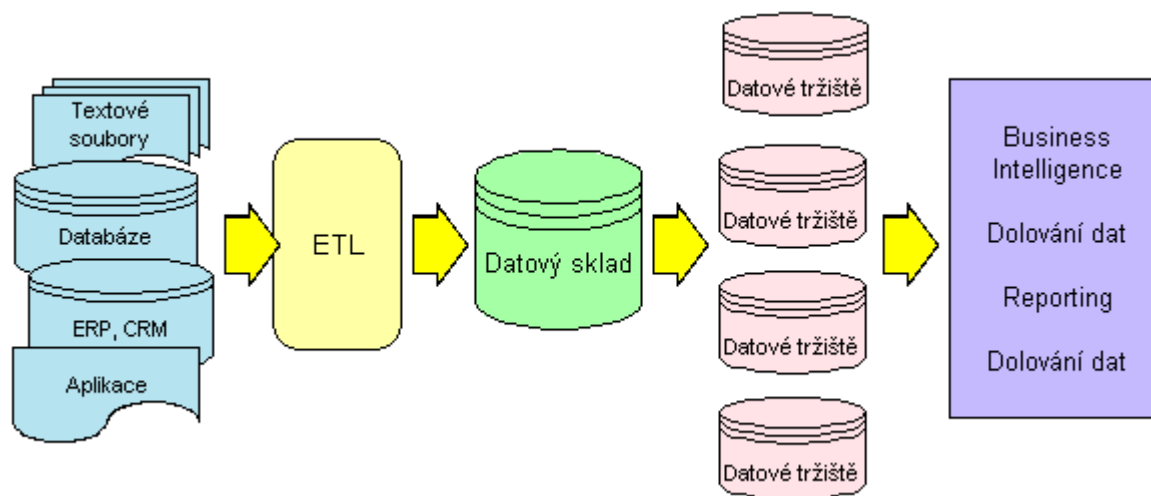
Aby bylo možné tohoto dosáhnout, je potřeba nejprve prostřednictvím ETL procesů a nástrojů získat a vhodným způsobem uspořádat, sjednotit a pročistit data ze zdrojových provozních informačních systémů, včetně údajů archivních a historických, a uložit je v prostředí datového skladu.

Odhady renomovaných společností říkají, že převážnou příčinou (více než 50%) neúspěšnosti BI projektů jsou nekvalitní data, což je většinou dáno podceněním implementace ETL procesů, jejichž součástí je i profilování a čištění dat pro zvýšení jejich kvality.

Klasické Business Intelligence řešení se skládá z:

- ETL nástroje pro získání, transformaci a uložení dat
- vlastní technologie uložení dat (datový sklad či provozní datový sklad - ODS)
- nástrojů pro tvorbu datových tržišť a vlastních datových tržišť

- nástrojů pro hlubší analýzy, prezentaci a dolování dat atd.



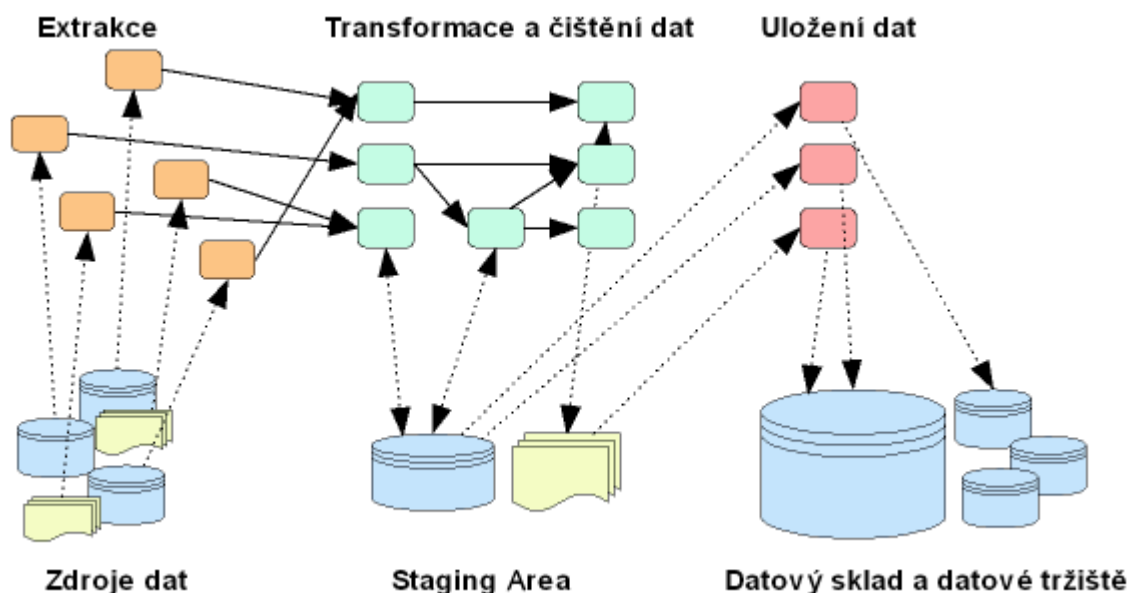
Obr. 2 - ETL v Bussines Intelligence

Dobré Business Intelligence řešení by mělo svým uživatelům poskytnout dostatečně kvalitní data a současně i nástroje, pomocí kterých budou schopni sami uživatelé realizovat své požadavky a přetvářet uložená data na hodnotné informace.

## 2 Návrh a dokumentace ETL procesu

Jednotlivé fáze ETL procesu zahrnují:

- identifikaci potřebných informací na zdrojových systémech
- extrakce těchto informací
- úpravy dat v jednotný formát
- mapování dat
- čištění výsledných dat na základě technických a obchodních pravidel
- uložení dat do datového skladu či datových trhů



Obr. 3 - Toky dat v ETL

Na obrázku výše je zobrazena základní struktura toků dat v ETL procesu. Na spodní straně jsou zobrazeny úložiště dat během celého procesu. Vlevo se vyskytují původní zdroje dat, což jsou většinou textové soubory či databáze. Data z těchto zdrojů jsou extrahována a následně uložena do Staging Area úložiště, kde jsou transformována a čištěna před uložením do cílového datového skladu, který je zobrazen vpravo. V horní části jsou pak zobrazeny jednotlivé procesy, šipky označují datové toky.

**Staging Area** je místo, kde ETL procesy pracují s daty. Většinou se jedná o databáze či pracovní

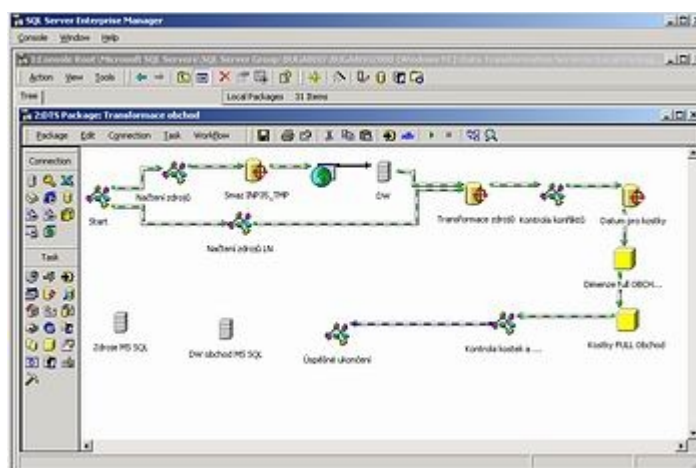
soubory ukládané na samostatném dedikovaném serveru, či vlastním ETL serveru. Jde tedy o fyzickou datovou vrstvu mezi zdrojovými a cílovými systémy. Staging Area může být centralizovaná nebo decentralizovaná. Ve většině případů se používá centralizovaná, kdy jsou veškerá data transformována na jednom místě.

## 2.1 Analýza zdrojových systémů a tvorba metadat

V prvních fázích návrhu ETL procesu je třeba učinit několik kroků. Prvním je zahrnout veškeré požadavky koncových uživatelů a vytvořit na jejich základě obchodní pravidla. Dalším pak vytvořit modely datových toků a ETL procesů, počínaje analýzou struktury a obsahu existujících datových zdrojů až po vytvoření podkladů pro mapování dat do modelu datového skladu. Všechny tyto informace nazýváme metadaty.

**Metadata**, tj. data popisující jiná data, jsou v ETL např. názvy tabulek, sloupců, jejich datové typy, struktury, mapovací pravidla, konverzní pravidla, transformace atd. Každý zdrojový systém stejně jako cílový tedy obsahuje svá vlastní metadata, která je potřeba zpřístupnit různým skupinám uživatelů či vývojářů. Ve většině případů tak vzniká problém, jelikož mnoho systémů je proprietárních (uzavřených) a nedovolí zvnějšku přímo pracovat s metadaty. Řešením je centrální metadata repository (repozitář metadat), které některé ETL nástroje přímo obsahují, nebo existují speciální aplikace. Centrální správa metadat mimo jiné výrazně redukuje chybovost a nepřesnosti při transformacích, jelikož vývojář nezadává metadata pokaždé znovu, ale pouze využívá ta již dříve uložená v repository.

Mezi metadata repository aplikace patří např.: IBM Websphere MetaStage (dříve Ascential), ASG Rochade společnosti ASG Software Solutions, SAS Metadata Server, OracleAS Metadata Repository, PowerMart Repository od firmy Informatica či Warehouse Control Center od firmy Intellidex.



Obr. 4 - Metadata repository

Metadata související s procesy ETL jsou:

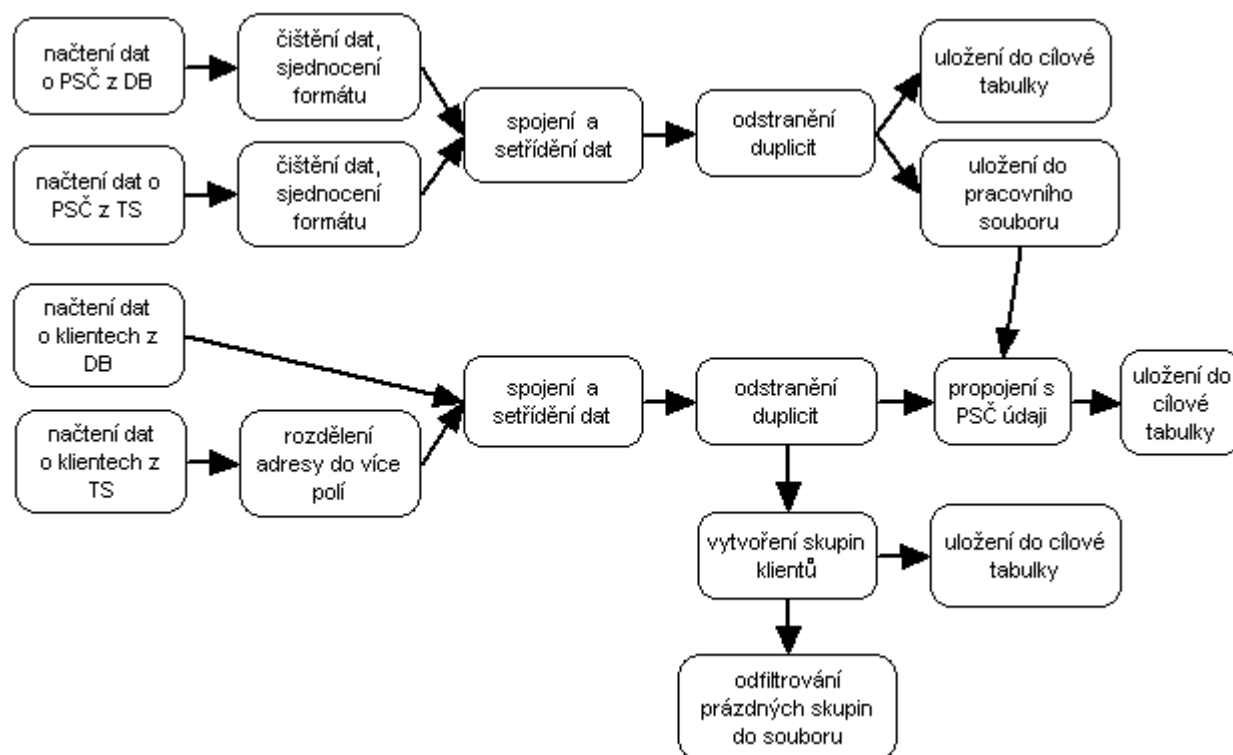
- obchodní metadata (obchodní pravidla, definice, termíny, zkratky, obchodní výpočty)
- technická metadata (definující zdrojové a cílové systémy, struktury jejich tabulek a sloupců, vztahů a závislostí, mapovací pravidla, konverzní pravidla)
- provozní metadata (informace o běhu provozních aplikací, jejich frekvenci, počty řádků, statistiky)
- projektová metadata (dokumenty apod.)

Typickým příkladem technických metadat v ETL procesech jsou:

- definice atributů zdrojových i cílových systémů
- mapování zdrojových atributů na cílové
- konverze datových typů
- konverze datových formátů
- konverze souborových formátů (oddělovače apod.)
- nastavení defaultních hodnot
- změny jazykových kódování
- logika spojování atributů z více zdrojových systémů

## 2.2 Model ETL procesů a datových toků

Následujícím krokem v návrhu ETL procesů je vytvoření diagramu či modelu ETL procesů. Tento diagram zpravidla vytváří hlavní ETL vývojář společně s databázovým administrátorem a analytikem pro datovou kvalitu. Důvodem pro vznik tohoto diagramu je zobrazení závislosti všech procesů mezi sebou, mezi extrakčními utilitami, transformacemi, pracovními soubory a tabulkami, procesy pro ošetřování chyb, aktivitami pro ověřování dat a procesy pro ukládání dat do cílového systému.



Obr. 5 - Diagram ETL procesů a datových toků

Načítání dat – často zde existují vzájemné závislosti mezi několika zdrojovými soubory a databázemi, ze kterých jsou data brána. Například aktualizovat tabulku klientů musíme před zpracováním souboru obchodních transakcí. Tyto závislosti musí být pochopeny, protože mohou ovlivnit načasování a běh sekvencí ETL.

Řazení a spojování dat – velmi často je potřeba data seřadit podle určitého sloupce či více sloupců, tak aby mohla být spojena s jinými daty ještě předtím, než začne vlastní transformace. Seřazení dat může také v mnoha případech značně zvýšit rychlost jejich ukládání do databáze (firmy jako

Oracle, Sybase, Informix i Microsoft doporučují řadit data před jejich vložení, které pak může být až 10x rychlejší).

Transformace – většina dat se transformuje z mnoha důvodů. Je důležité prozkoumat nejvhodnější čas pro transformaci. Proto změny aplikovatelné na všechna zdrojová data, jako například sjednocení datových typů a formátů, by měly být provedeny na začátku ETL procesu. Transformace specifické pro cílové databáze, jako jsou agregace a sumarizace, probíhají naopak až na konci ETL procesu.

Pracovní soubory a tabulky - řazení, spojování a transformace vyžadují většinou poměrně mnoho pracovního místa k uložení dočasných výsledků do souboru či databáze. Tyto pracovní soubory a tabulky mohou být často i větší než původní zdrojová data.

Práce s chybnými daty – pokud dojde k chybě při transformaci dat (například pokud data neodpovídají požadovanému typu či formátu) jsou taková data na základě předem daných obchodních pravidel buď akceptována, nahrazena defaultními hodnotami, nebo zamítnuta a uložena do souboru.

Aktivity pro kontrolu dat – každý program, který manipuluje s daty by měl provádět kontrolní a srovnávací součty. Nejjednodušší kontrolou je porovnání počtu záznamů na vstupu a výstupu, pokud chceme kvalitnější kontrolu, provedeme součty specifických částí dat nebo částek na zdrojových i cílových datech (například celkový objem transakcí).

Ukládání dat – je nezbytné určit pořadí jednotlivých ukládání dat. Některé tabulky je třeba ukládat v pevně daném pořadí (například kvůli referenčním integritám je nutné nejdříve nahrát zákazníky a pak teprve faktury), jiné mohou být naopak plněny souběžně, což má pochopitelně pozitivní dopad na rychlost celého procesu.

## 2.3 Analýzy dat, data profiling

Pro úspěch ETL procesů je klíčové dodání správných, spolehlivých, korektních a kompletních informací. Údaje z různých zdrojových systémů mají často rozdílnou kvalitu. Aby bylo možné tato data seskupit do jednoho důvěryhodného zdroje informací, je součástí ETL procesů i proces čištění dat. Vlastní proces čištění probíhá na základě systémových, doménových a obchodních pravidel, která musí data splňovat.



Proces hledání a definování těchto pravidel se nazývá data profiling. Během data profilingu se analyzují data a hledají se nekompletní, nepřesné nebo dvojznačné údaje, ověřují se vztahy mezi nimi a zjišťuje se referenční a doménová integrita.

### Běžné postupy v data profilingu

Pro zjišťování, zda může sloupec obsahovat prázdnou hodnotu, se vyhledá počet prázdných hodnot ve sloupci. Pokud je taková hodnota nalezena a není jasně definované pravidlo pro její nahrazení defaultní hodnotou, musí být sloupec nullable<sup>1</sup>.

Pro zjišťování unikátnosti dat a definování unikátních klíčů se porovnají hodnoty sloupce v distinctu<sup>2</sup> a bez něj.

Odhalování překlepů probíhá vyhledáním unikátních hodnot a sad stejných hodnot a jejich srovnáváním.

Ověřování vazeb referenční integrity lze aplikovat pouze u sloupců neumožňujících prázdné hodnoty. Analyzuje se, zda všechna data z jednoho sloupce nejsou podmnožinou druhého a naopak.

Údaje, které obsahují chyby se nazývají špinavá data. Mezi klasické příklady špinavých dat patří:

- data porušující pravidla doménové a referenční integrity (duplicity ve sloupcích, kde jsou potřeba unikátní data, prázdné hodnoty v polích, kde je hodnota vyžadována, chybné vzájemné vazby mezi tabulkami)
- chybné údaje v místech, kde jsou definována obchodní a jiná pravidla a rozsahy (např. záporný věk či nesmyslné rodné číslo)
- překlepy
- data v jiných polích než mají být
- používání zkratk a zkrácených tvarů (jednou je zkrácené slovo použito a podruhé ne - např. Václavské nám. a Václavské náměstí - data jsou pak vnímána jako dva rozdílné údaje)
- různé pořadí dat v polích (Jan Novák a Novák Jan)

Jedním z finálních kroků v návrhu je vytvoření dokumentu, který specifikuje transformace a mapování zdrojových dat na cílová. Takový dokument by měl obsahovat všechny vstupní i cílové tabulky, sloupce a jejich datové typy a délky, defaultní hodnoty a formáty a transformační logiku pro každý sloupec, pak je základem pro práci ETL vývojářů.

<sup>1</sup> Sloupec, který je označen jako nullable, může obsahovat prázdné hodnoty (takzvané NULL hodnoty).

<sup>2</sup> Distinct je klíčové slovo jazyka SQL, které slouží k odstranění duplicitních hodnot při práci s daty.

Cílová tabulka		Zdrojová tabulka					Pravidla pro mapování
	Jméno sloupce	Jméno tabulky	Jméno sloupce	Datový formát	Klíč	Může být null	
1	Source system code	UACC_2	AC2SSC	char (8)	A	N	1:1
2	Account Number	UACC_2	AC2AN	char (8)	A	N	1:1, pokud je hodnota sloupce AC2AGC prázdná, ignoruj celý záznam
3	Account group code	UACC_2	AC2AGC	char (3)		A	1:1, pokud je NULL, pak nahraď hodnotou z AC2AGTC
4	Account group description	UACC_2	AC2AGD	char (35)		A	Pokud je AC2AGC = "A", pak dej "CZ" jinak "EU"
5	Account group type code	UACC_2	AC2AGTC	char (15)		A	Pokud je AC2AGC = "A", pak dej "CZ" account group", jinak "EU account group"

Tabulka 1 - Specifikace mapování zdrojových dat na cílová

## 3 Úskalí ETL procesů

Datové objemy, se kterými firmy pracují, se stále zvyšují. Velká množství dat a potřeba jimi naplnit datové sklady v co nejkratším možném čase patří mezi standardní výzvy, se kterými se vývojáři setkávají.

Při samotných transformacích se však v ETL procesech vyskytují stále se opakující problémy, které je potřeba řešit.

### 3.1 Rozdílná terminologie

Když jsou dodávána z více zdrojů data, která mají sice stejný název, ale obsahují jiné informace, nebo naopak existují různé názvy pro stejný typ informací, pak je třeba provést sjednocení terminologie. V praxi se především u velkých systémů může stát, že některým uživatelům schází určitý atribut a použijí místo něj jiný, který nevyužívají. Jiní uživatelé však tento atribut používají správně a dojde tedy k tomu, že pokaždé obsahuje hodnoty s jiným významem.

### 3.2 Různé formáty dat

Dalším častým problémem bývají různé formáty pro datumy, čísla, rodná čísla, PSČ apod. U datumů jde nejčastěji o pořadí dne, měsíce a roku, používají se i různé oddělovače, u měny se můžeme setkat zase s tím, že je k oddělení desetinné hodnoty použita v nějakém systému čárka a v jiném zase tečka. Tyto problémy je nutné odhalit v počátečních analýzách a formáty sjednotit na začátku transformací.

### 3.3 Chybějící data

Stává se, že v primárních systémech chybějí potřebná data. Zvlášť podstatným údajem pro datové sklady jsou datumy, jelikož prakticky každý datový sklad má časovou dimenzi. V některých případech lze tato data doplnit z jiných zdrojů, pokud ne, je potřeba definovat pravidla pro jejich doplnění. Když není žádná jiná možnost, lze neúplná data vhodně označit a uložit je v datovém skladu.

### 3.4 Různá měřítka

Pokud se například slučují data ze zdrojových systémů z různých zemí, dochází k tomu, že jsou používány jiné měny, váhové i mírové jednotky apod. Zde je nutné převést veškerá data při transformaci na jednotné měny a jednotky. Dalším případem mohou být třeba různá měřítka, kdy jednou jsou částky uloženy v tisících, podruhé v milionech apod.

### 3.5 Nejednoznačné údaje

Poměrně častým problémem je nejednoznačnost údajů či číselníků. Jedná se o případy, kdy je jeden a ten samý údaj uložen různými způsoby. Například je-li pro jeden typ zboží nadefinováno více označení či kódů, což mohlo vzniknout tak, že bylo zboží zadáno v několika různých prodejnách, kde bylo přijato a bylo třeba jej do systému rychle zavést. Toto je pak nezbytné při transformacích sjednotit.

### 3.6 Vícenásobné položky

Některá data bývají ve zdrojovém systému sloučena a je třeba je pro potřeby datového skladu rozdělit na jednotlivé části. Jedná se například o pole s adresou, kde mohou být sloučené údaje jako ulice, číslo popisné, město, PSČ a stát, každý údaj je zde zapsaný na novém řádku. U dalšího záznamu však může být pořadí těchto údajů jiné, například PSČ je uvedeno před městem, nebo město před ulicí apod. Vzhledem k objemům dat, se kterými datové sklady pracují, není možné toto dělat ručně a je nutné tyto transformace zautomatizovat. Je pak ale potřeba počítat s tím, že nová data nemusí být stoprocentně správná, proto pokud se jedná o zásadní údaje, měly by se použít dodatečné pokročilejší metody pro čišťení dat (např. porovnávání s databázemi PSČ, států, měst apod.).

### 3.7 Nedodržená referenční integrita

Může se stát, že ve zdrojových systémech nejsou dodrženy či uplatněny dostatečná pravidla pro referenční integritu a díky tomu jsou porušeny vazby mezi jednotlivými tabulkami, které jsou klíčové pro datový sklad. V takovém případě je třeba stanovit pravidla, podle kterých se chybějící data doplní nebo reportují, aby je bylo možné opravit již ve zdrojových systémech.

### 3.8 Duplicita dat

Duplicitní hodnoty mohou výrazným způsobem zhoršit kvalitu dat v datovém skladu. V případě, že máme více zdrojových systémů, je pravděpodobné, že některá data budou uložena ve více z nich, a pak musí být definováno, která mají přednost, či jakým způsobem se finální hodnoty odvodí.

### 3.9 Vícenásobné hierarchie

Dimenze datového skladu mohou mít vícenásobné hierarchie. Klasickým případem je časová dimenze, která má obvykle hierarchii den-měsíc-rok, ale také může mít hierarchii například den- týden-fiskální měsíc-rok apod. Chybné zacházení s těmito hierarchiemi pak má za následek špatné plnění datového skladu.

### 3.10 Náhodné chyby

Jedná se především o překlepy, různé pořadí dat v jednom atributu (například prohazování jména a příjmení), občasné používání zkratk atd.

### 3.11 Odlišná kódování, jazyky

Zdrojové systémy mohou používat různá kódování a je tedy nezbytné je sjednotit do kódování, které bylo určeno pro datový sklad. Tzn. že je potřeba si stanovit pravidla, při kterých jsou speciální znaky typické pro daný jazyk nahrazovány jinými znaky, popřípadě zcela vypuštěny. Problém nastává především v případě, pokud jsou tyto znaky použity ve sloupcích, které slouží jako referenční klíče pro jiné tabulky. Může pak dojít k tomu, že po převedení speciálních znaků získáme hodnotu, která již v tabulce je, a tím pádem dojde k porušení unikátnosti hodnot ve sloupci.

### 3.12 Nedostatečná testovací data

Pokud nemáme žádná testovací data, nebo jen malé vzorky nedostatečně kvalitní, může se při spuštění ETL procesů s ostrými daty snadno stát, že nastanou takové chyby v datech, na které nejsou definována žádná pravidla, že dojde k razantnímu zhoršení výsledné kvality dat. Proto abychom byli schopni připravit kvalitní pravidla pro čištění a transformace dat, je potřeba mít i dostatečná testovací data ještě před zahájením implementace ETL procesů.

## 4 ETL nástroje

Na trhu je dostupné velké množství ETL nástrojů, které se ve své funkcionalitě značně liší. Výběr správného nástroje proto může být poměrně komplikovaný proces a pro kvalitní a spolehlivé ETL řešení je taková volba klíčová, jelikož na rozdíl od jiných částí procesu je dodatečná výměna ETL nástroje velmi komplikovaná.

Obecně existují dvě možnosti, jak řešit ETL implementaci. Buď vytvořit vlastní program, který bude data extrahovat, transformovat i ukládat, a to s použitím standardních nástrojů a programovacích jazyků jako jsou COBOL, C, PERL, PL/SQL, nebo použít nějaký na trhu dostupný ETL nástroj. Vytvoření vlastního ETL programu může být výhodné v případě, kdy se jedná o jednoduché transformace, zdrojové systémy běží na jedné stejné platformě a jsou k dispozici programátoři z vlastních zdrojů. Pak je vývoj ETL aplikace finančně méně náročný, jelikož není potřeba kupovat žádný drahý software, ani není třeba speciálně vyškolených odborníků. Ruční psaní ETL procesů však s sebou nese tyto klasické problémy:

- údržba a správa ručně psaných programů založených většinou na skriptech není flexibilní a je značně obtížná. Ruční psaní kódů je více náchylné na chyby než grafické modelování procesů v moderních ETL nástrojích.
- zdrojové systémy jsou často různé a je tedy nutné napsat pro každý z nich vlastní procesy. Pokud se tyto systémy časem změní, je následná údržba a aktualizace procesů čím dál komplikovanější.
- neexistuje většinou žádné centrální skladiště metadat, což vede ke zvýšenému riziku, že dojde k rozdílu u datových typů v různých částech procesu.
- individuálně tvořené ETL programy bývají často pomalejší, zvláště čím jsou složitější. Moderní nástroje dokáží využívat paralelní zpracování, multithreading a jsou vysoce škálovatelné, což lze ručně naprogramovat velmi obtížně.

### 4.1 Moderní ETL nástroje

Mezi hlavní přednosti moderních ETL nástrojů patří především:

- integrovaná správa metadat

- podpora víceuživatelské práce
- správa verzí
- znovupoužitelnost procesů a objektů
- otevřená architektura
- konektivita k mnoha typům zdrojů dat
- integrace s reportingem a analýzou dat
- škálovatelnost
- technická podpora, vývojářská fóra apod.

Podle Stevena R. Meyera, ředitele pro implementaci datových skladů firmy Quareo, se ETL nástroje dají rozdělit do následujících kategorií:

**EtL nástroje** – malé “t” ve zkratce znamená, že je zde velmi slabá transformační část, nebo zcela chybí. Jedná se tedy spíše o nástroje pro migraci dat a ne klasické ETL nástroje.

**eTL či ETl nástroje** - malé “e” nebo “l” značí, že se jedná o nástroje, které jsou úzce spjaty s určitým specifickým vstupem či výstupem, ať už se jedná o textové soubory či konkrétní databázový systém. Poskytují však robustní transformační funkce. Může se jednat například o nástroje poskytované přímo výrobcem určité databáze.

**eTl nástroje** - velké písmeno “T” říká, že se jedná o nástroje s kvalitní a výkonnou transformační částí, které ale postrádají dobrou konektivitu k různým datovým zdrojům a formátům.

**ETL nástroje** - jedná se o komplexní ETL nástroje poskytující kombinaci dobré konektivity i výkonných transformačních funkcí. Většinou však bývají tyto nástroje podstatně dražší než nástroje z ostatních výše uvedených kategorií. Pro velké a složité projekty nebo takové projekty, které pracují s velkými objemy dat, bývají tyto nástroje většinou jediným opravdovým řešením umožňujícím úspěch ETL implementace.

## 4.2 Volba správného ETL nástroje

ETL nástrojů jsou v dnešní době již stovky, od robustních řešení, která zvládají obrovské objemy dat, mnoho datových zdrojů, podporují práci mnoha vývojářů najedou apod., až po malé aplikace, které vykonávají třeba pouze jednu část z ETL procesů (např. jen transformace), popřípadě jsou specializované na určitou oblast (např. pojišťovnictví).

Pro každý projekt může být vhodný jiný ETL nástroj, proto je důležité jeho volbu nepodcenit. Obecná rada je, že pokud ETL procesy pracují s běžnými transformacemi a problémy, měly by vyhovovat i na trhu běžně používané ETL nástroje.

Zde je několik klíčových bodů, které je třeba brát při výběru ETL nástroje v úvahu:

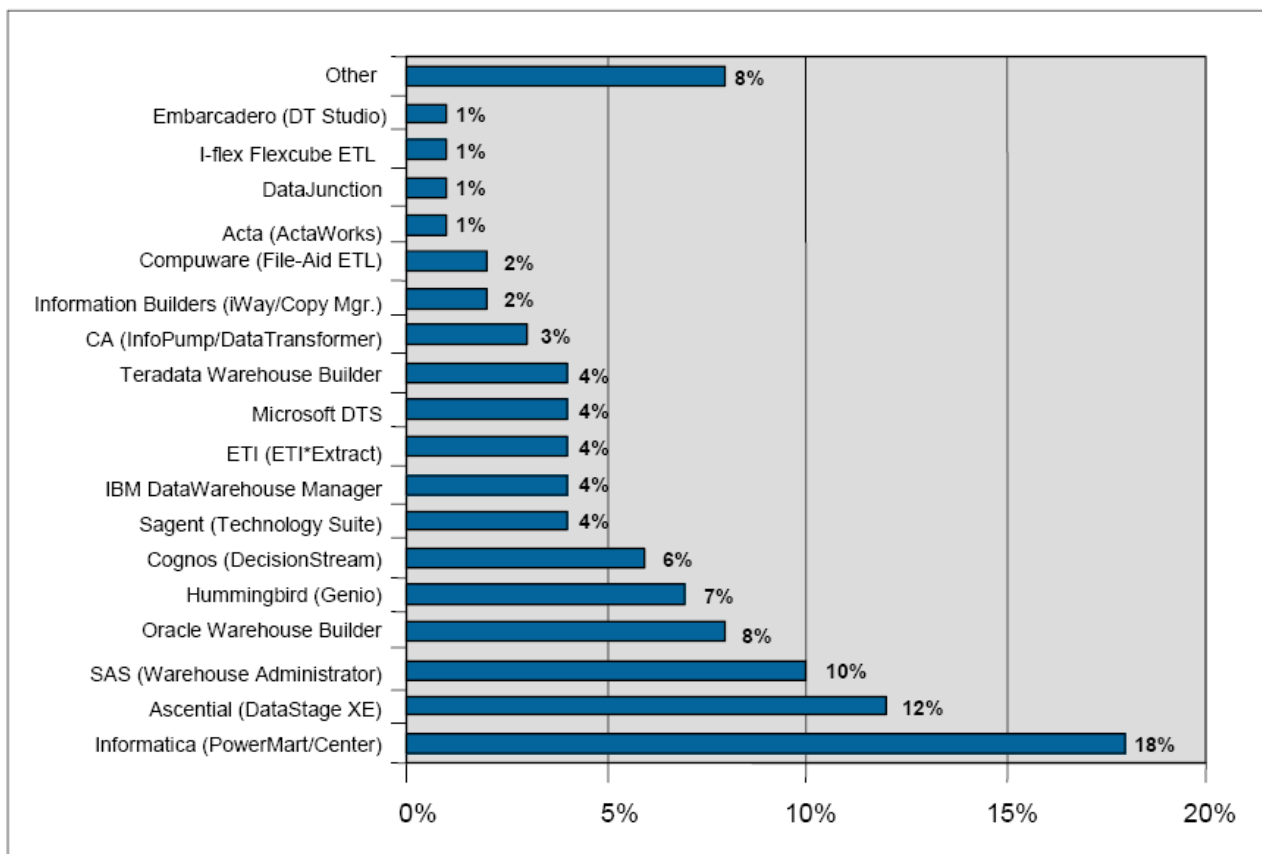
- je obvyklé, že zdrojové systémy používají různé databáze a aplikace, a je tedy potřeba, aby ETL nástroj dokázal sám přistupovat ke všem typům zdrojových dat nejlépe nativně. Vývojář se pak nemusí o konektivitu starat, zvolí jen odpovídající komponentu.
- zdrojové databáze často disponují speciálními nástroji (bulk loadery) pro rychlé manipulace (ukládání a čtení) s velkým množstvím dat, které jsou mnohonásobně rychlejší než tradiční přístupy. Je potřeba, aby tyto nástroje byly podporovány.
- ETL nástroj by měl být schopen generovat a udržovat metadata v centrálním úložišti, jelikož v ETL procesech je velké množství metadat - datové definice zdrojových a cílových systémů, datové modely, pravidla pro transformace a mapování, obchodní pravidla, statistiky apod.
- standardní přístup v ETL procesech využívá dávkový mód, kdy musí extrakce proběhnout v pevně daném vymezeném čase. ETL nástroj by měl být dobře škálovatelný, schopný zvládat paralelní zpracování toků dat a optimálně využívat hardwarových prostředků.
- cena daného ETL řešení. Zde se nejedná jen o cenu ETL nástroje, ale i o další výdaje s ním spojené. Ať už je to cena za hardware, který nástroj k práci potřebuje, cena za implementaci a následnou údržbu systému, technickou podporu, licenční poplatky či v neposlední řadě různá školení. Také je vhodné ověřit, že je pro danou technologii dostatek odborníků na trhu a případně jestli je možné vyškolit vlastní odborníky. Dále je dobré zvážit, zda zvolený ETL nástroj neobsahuje zbytečně mnoho nástrojů, utilit a vlastností, které nejsou v projektu využitelné a přesto jsou zahrnuty v ceně, jako jsou například nástroje pro čištění dat, reporting, data mining apod.

## 4.3 Významné ETL nástroje

Nejpoužívanějšími nástroji v implementacích robustních ETL projektů jsou produkty společností Informatica a Ascential (dnes IBM). Toto potvrzuje i rozsáhlá analýza ETL nástrojů, kterou provedla renomovaná firma Forrester Research, Inc. Zaměřila se především na funkcionalitu



v oblastech rozšiřitelnosti, konektivity a podpory možnosti spolupráce více vývojářů na projektu. Výsledkem je, že Informatica a Datastage jsou technologicky nejkvalitnějšími produkty a skutečnými vůdci v ETL implementacích. Za nimi jsou pak s určitým odstupem produkty firem Oracle, Hummingbird, Business Object, Data Mirror a SAS.



Source: Giga Information Group

**Obr. 6 - Podíly významnějších ETL nástrojů na trhu (rok 2002)**

Obrázek převzat od firmy Giga Information Group, rok 2002

#### 4.3.1 WebSphere DataStage TX

Výrobce: IBM, dříve Ascential (spojení v roce 2005)

Aktuální verze: 8.0

WebSphere DataStage je hlavní částí balíku IBM® WebSphere® Data Integration Suite a je integrována s jedněmi z nejlepších nástrojů na zvyšování kvality, čištění a profilování dat.

Hlavní vlastnosti WebSphere DataStage:

- vysoká konektivita k široké oblasti aplikací, databází a externích informačních zdrojů, textových souborů, XML atd.
- kompletní vnitřní knihovna s více než 300 funkcemi
- k programování lze použít DataStage BASIC
- maximální propustnost dat, využívá architekturu paralelního zpracování (parallel processing)
- obsahuje nástroje pro vývoj, deployment a údržbu bez potřeby ručního kódování
- běží na různých platformách
- škálovatelnost na vysoké úrovni, využívající huby a klastry serveru pro multithreading
- kvalitní podpora víceuživatelského přístupu, verzování objektů, projektů, správa repository
- podpora data grids
- jedná se o dražší nástroj, který u velkých konfigurací stojí i více než 1 mil. dolarů

#### 4.3.2 Informatica PowerCenter 8

Výrobce: Informatica

Aktuální verze: verze 8

Vlastnosti Informatica PowerCenter 8:

- rozsáhlá konektivita k mnoha typům zdrojů (např. strukturované, nestrukturované a částečně strukturované údaje, data relační, souborová, z mainframů, reportové systémy)
- podpora data grids
- optimalizace decentralizace zpracování
- kvalitní podpora znovupoužitelnosti definic
- podpora programování v Javě
- velmi dobrá škálovatelnost využívající stejně jako DataStage huby a klastry serverů
- dobré možnosti pro víceuživatelský přístup, správa verzí objektů, projektů, metadat
- jedná se o dražší nástroj, který u velkých konfigurací stojí i nad 1 mil. dolarů

#### 4.3.3 Oracle Warehouse Builder

Výrobce: Oracle

Aktuální verze: 10g Release 2

Oracle Warehouse Builder je ETL nástroj integrovaný v rámci Oracle Enterprise, vhodný pro střední a menší implementace. Je zaměřený především na Oracle databáze.

Základní vlastnosti:

- jako zdroj dat se umí připojit ke standardním databázím, XML souborům, SAPu, Oracle EBusiness Suite , PeopleSoftu
- používá Oracle Data Server pro svůj ETL engine, což je vysoce škálovatelná databáze
- pro ETL transformace je generován PL/SQL kód
- designer ETL je v jazyce JAVA, tzn. že je platformově nezávislý
- obsahuje procesy pro čištění dat, datový audit, relační a dimensionální modelování
- správa a importy metadat
- podporuje správu verzování objektů, víceuživatelský přístup
- ceny se pohybují v závislosti na velikosti konfigurace od desítek tisíc dolarů až nad 1 mil. dolarů

#### **4.3.4 Enterprise ETL Server**

Výrobce: SAS

Aktuální verze: 9

Tento nástroj je úzce integrovaný s ostatními produkty SAS Institute. Mezi jeho základní vlastnosti patří:

- kvalitní podpora konektivity pro získávání dat z různých systémů
- transformace, čištění dat a load, vytváření datových skladů, tržišť nebo BI a analytických datových úložišť
- metadata jsou zachycována a dokumentována během datové integrace a transformačního procesu
- transformace mohou běžet na libovolné platformě s jakýmkoliv datovým zdrojem
- je zde více než 300 předdefinovaných transformací
- součástí je wizard na generování transformací či Java plugin pro návrh šablon pro znovupoužitelné transformace, které jsou registrovány v metadatech
- transformace jsou spustitelné různými způsoby, což umožňuje jednoduché znovupoužití

v různých projektech i technologických prostředích

- je zde pouze slabá podpora víceuživatelského přístupu a správy verzí
- je to jedno z nejdražších řešení, ceny se pohybují často vysoko nad 1 mil. dolarů

#### 4.3.5 Data Transformation Services (DTS)

Výrobce: Microsoft

Aktuální verze: 2000 SP3

ETL procesy se vytvářejí přímo v prostředí Microsoft SQL serveru. Hlavní vlastnosti DTS jsou:

- konektivita je poměrně špatná, podpora je přes ODBC a OLE DB, dále pro textové soubory a nativně pouze Microsoft SQL server
- datové pumpy jsou tvořeny balíčky, balíček obsahuje posloupnost kroků od připojení k datovým zdrojům, přes transformaci až po cílové systémy
- balíčky je možné spouštět paralelně
- vše je přístupné přes COM rozhraní, lze tedy používat k transformacím programovací jazyky jako např. Visual Basic, Visual C++, Delphi apod.
- balíčky lze ukládat formou zdrojových kódů pro Visual Basic
- víceuživatelská práce není příliš dobře podporovaná, chybí funkce pro check-in a check-out
- jedná se o levnější řešení vhodné pro menší projekty, ceny se pohybují v desítkách tisíc dolarů

Microsoft nahrazuje DTS novější technologií Server Integration Services (SSIS).

#### 4.3.6 Data Integrator

Výrobce: Business Objects

Aktuální verze: Data Integrator XI (version 11.0.2.5)

Mezi hlavní rysy tohoto nástroje patří:

- škálovatelnost je velmi dobrá, podobně jako DataStage a Informatica zvládá paralelní zpracování prostřednictvím hubů a serverových klastrů
- konektivita je také velmi dobrá, je zde podporováno mnoho různých zdrojů nativně jak pro čtení, tak pro zápis včetně aplikací SAP, Siebel, PeopleSoft, JDE a Oracle, podporuje i ODBC,

textové soubory, XML atd.

- Metadata manager sbírá BI, ETL a jiná metadata, která lze analyzovat a vytvářet z nich metadata reporty, které poskytují informace o datové kvalitě
- data quality firewall – kvalitní data profiling umožňuje vytvořit pravidla pro data, která jsou pak filtrována mezi zdrojovým a cílovým systémem (čistění dat)
- knihovna transformací obsahuje i operace s XML soubory, kontingenčními tabulkami, pomalu měnícími se dimenzemi, čistěním dat, validací apod.
- víceuživatelský přístup, verzování objektů
- integrace s BI produkty od stejné firmy
- paralelní zpracování, distribuované zpracování, real-time data přesuny
- cenově se jedná o levnější řešení, ceny se pohybují v závislosti na velikosti konfigurace v desítkách tisíc dolarů

#### 4.3.7 Sunopsis Data Conductor

Výrobce: Sunopsis

Aktuální verze: 4.1

Sunopsis Data Conductor má následující významné znaky:

- k ETL používá přístup business-rules-driven, který odděluje obchodní pravidla od vlastní implementace
- používá architekturu E-LT
- nepotřebuje vlastní samostatný server a engine, místo toho využívá sílu RDBMS engine
- obsahuje data consistency firewall – nástroj, který automaticky detekuje chybná data vložení do cílové aplikace, což je prováděno na základě pravidel pro kvalitu dat
- používá bulk data transformace (ne řádek po řádku), které jsou založeny přímo na možnostech RDBMS a výrazně zvyšují výkonnost transformací
- je v jazyce Java, tzn. že je platformově nezávislý
- konektivita ke všem RDBMS systémům přes JDBC včetně všech hlavních DW platforem (Teradata, IBM DB2, Oracle, Sybase ICQ, Netezza) a další technologie jako flat soubory, XML, LDAP, ERP apod.
- dobrá škálovatelnost založená na možnostech RDBMS systémů

- chybí zde podpora víceuživatelské práce a je zde jen slabá podpora pro verzování
- jedná se o cenově průměrný produkt, ceny se pohybují v závislosti na velikosti konfigurace v desetitísících dolarů

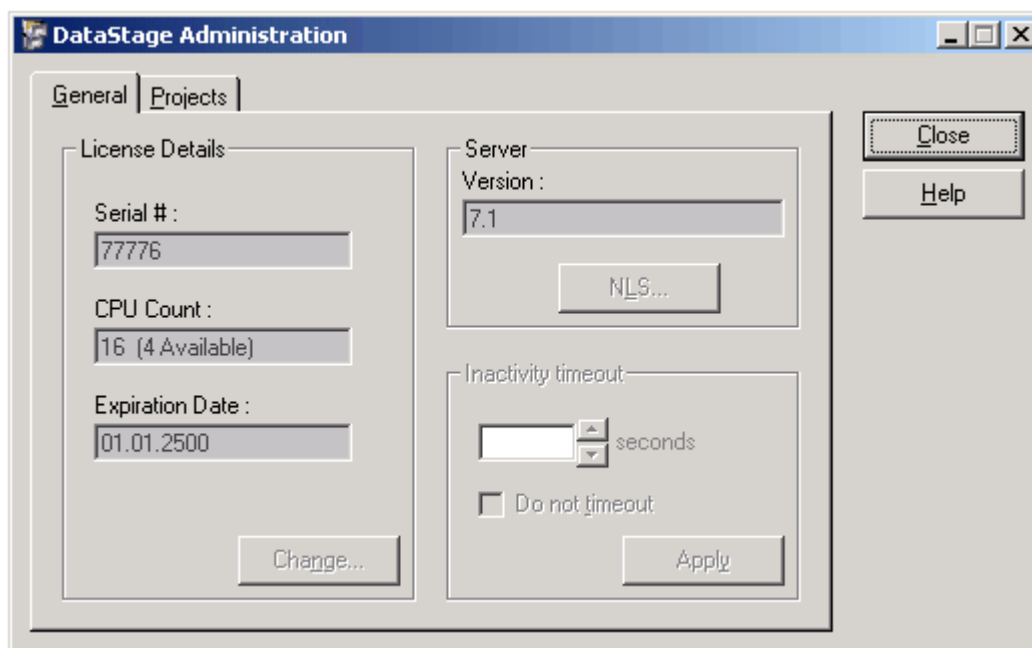
#### 4.3.8 Podrobnější popis nástroje DataStage (IBM)

DataStage se skládá z několika aplikací a nástrojů, které umožňují kompletní vývoj, správu, logování, kontrolu a spouštění ETL procesů, vše v přehledném grafickém prostředí.

Základní klientské aplikace DataStage jsou:

- DataStage Administrator (slouží k nastavení projektů)
- DataStage Designer (vývojová část, kde se navrhují ETL procesy)
- DataStage Director (nástroj na spouštění a logování ETL procesů)
- DataStage Manager (správa skladu objektů, metadat)

#### DataStage Administrator



Obr. 7 - DataStage Administrator

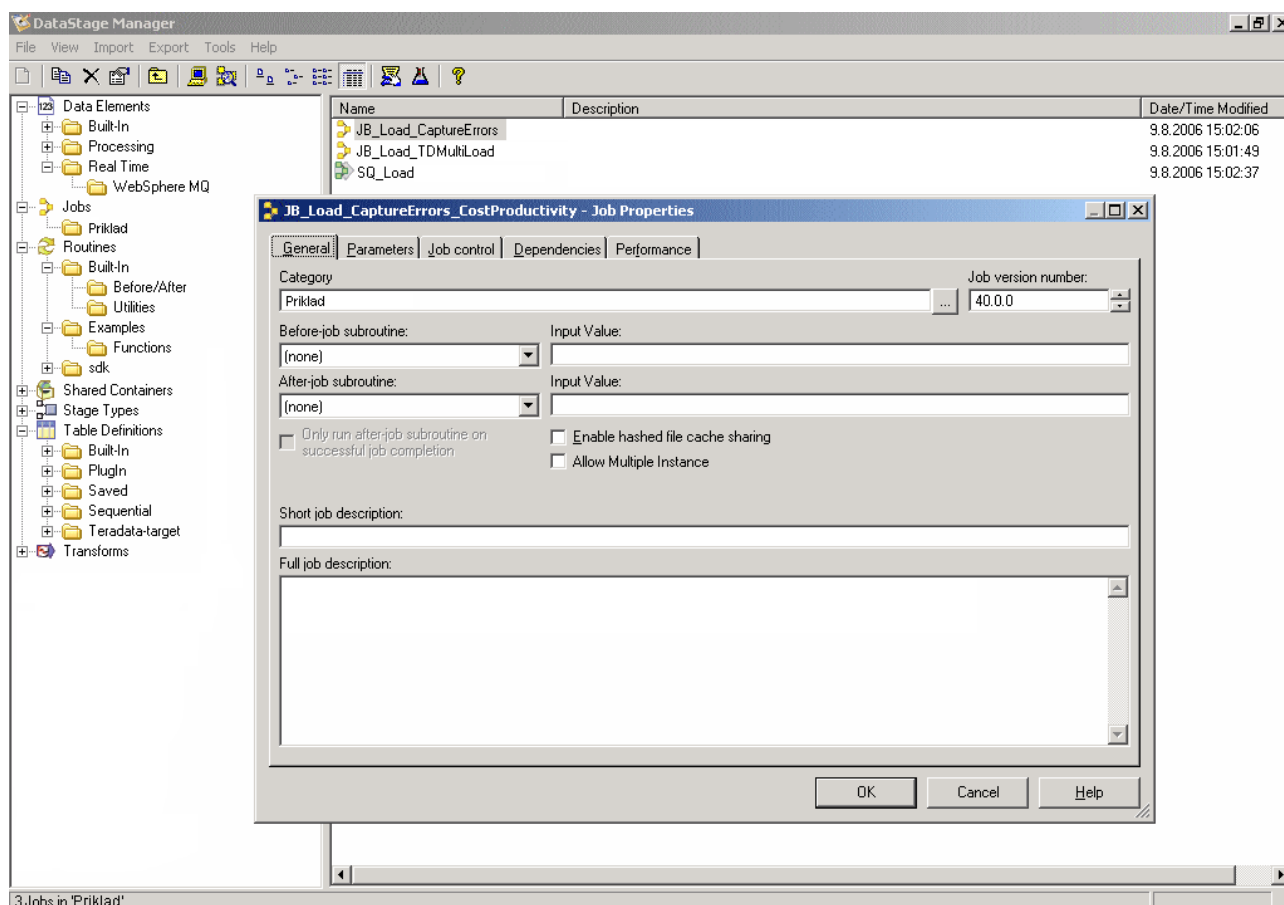
Tento nástroj slouží k přímému přístupu k serveru DataStage, nastavují se zde serverové parametry, spravují projekty, nastavují defaultní vlastnosti projektů, práva přístupů k jednotlivým modulům

a v neposlední řadě se zde provádějí optimalizace výkonu. Práva a uživatelské skupiny vycházejí z pravidel operačního systému, na kterém DataStage server běží (Windows NT, Unix atd.).

### DataStage Manager

DataStage Manager se používá ke správě skladu jobů, rutin, metadat, transformací a dalších objektů. Umožňuje import metadat z mnoha různých zdrojů a jejich sdílení mezi více servery DataStage. Prostřednictvím tohoto nástroje se provádí přesun projektů mezi prostředími, umožňuje exportovat jak jednotlivé objekty či jejich skupiny, tak celé projekty.

Na obrázku v levé části je zobrazena stromová struktura celého skladu objektů, rutin, metadat, komponent a transformací. V popředí je otevřeno okno s detaily jobu.

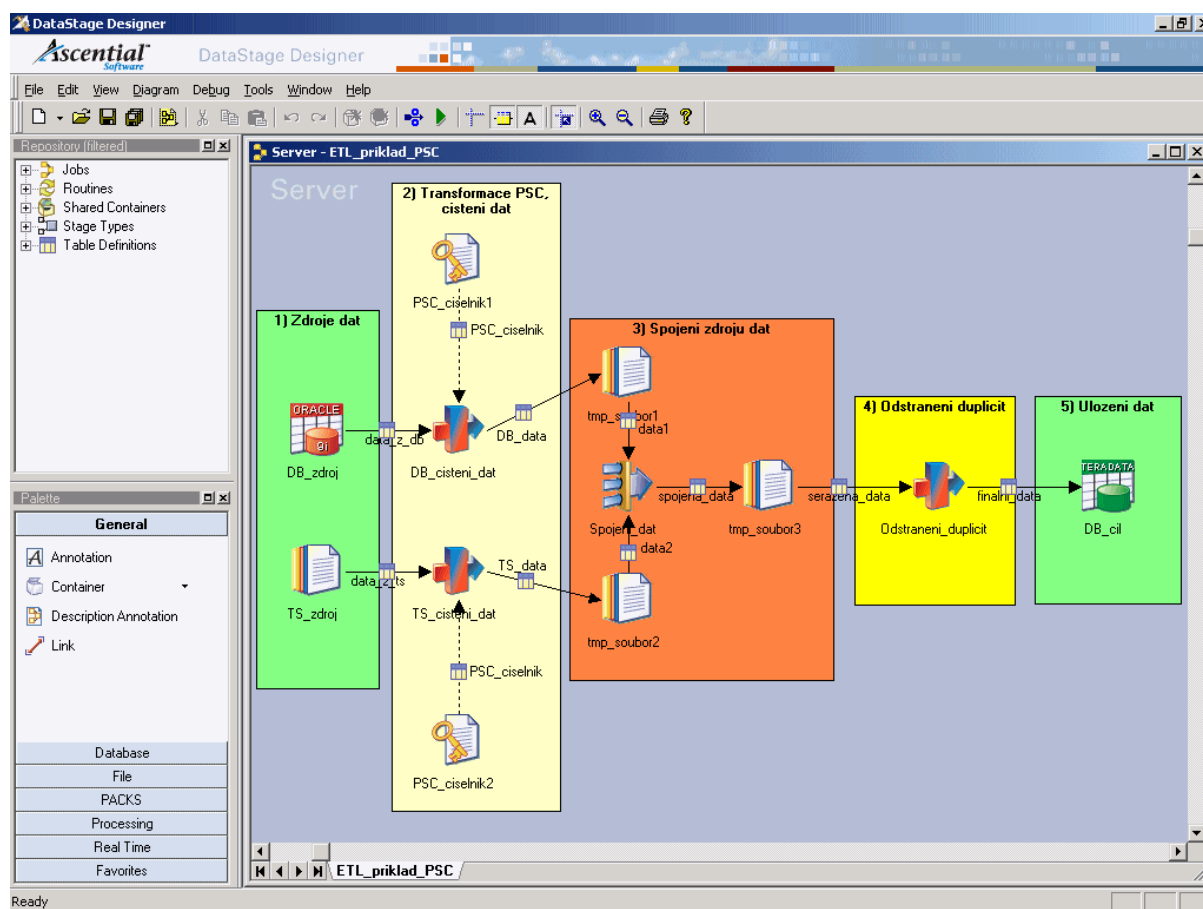


Obr. 8 - DataStage Manager

## DataStage Designer

Designer slouží k vytváření vlastních ETL procesů v grafickém prostředí prostřednictvím pokládání grafických ikon, které zastupují jednotlivé ETL komponenty, a jejich následném propojování, čímž se určují datové toky. Jednotlivé části takto sestavených transformací se nazývají joby a seskupují se do sekvencí. V jobech jsou definovány zdroje, transformace i cíle. V případě, že se jedná o Enterprise Edition, jsou k dispozici i speciální paralelní joby, které jsou optimalizované pro paralelní zpracování a disponují více komponentami. Mezi základními komponentami, dostupnými pro obě verze, jsou např.: komponenty pro přímý přístup k různým databázím, přístup k textovým a hashed souborům (speciální typ souborů určený především pro lookupy), bulk loadery, agregátor, FTP plugin, spojení řádků či sloupců, řazení, rozdělení a spojení datových toků, transformer (hlavní komponenta pro definování transformací dat) atd.

V horní levé části obrázku se vyskytuje sklad objektů a metadat, pod ním pak paleta komponent, které se používají pro vlastní transformace, a uprostřed je zobrazen vlastní job definující tok dat, transformace atd.



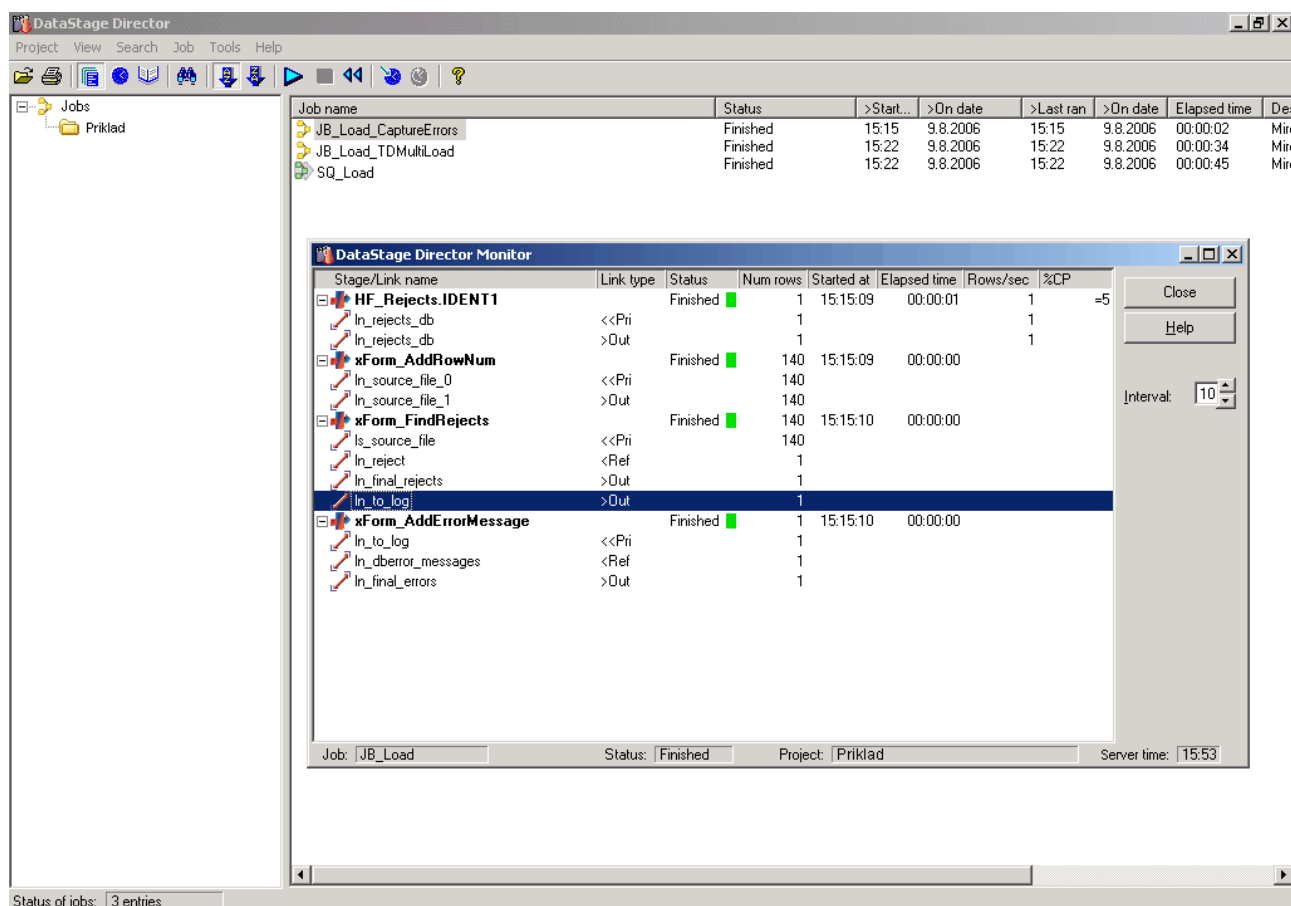
Obr. 9 - DataStage Designer



## DataStage Director

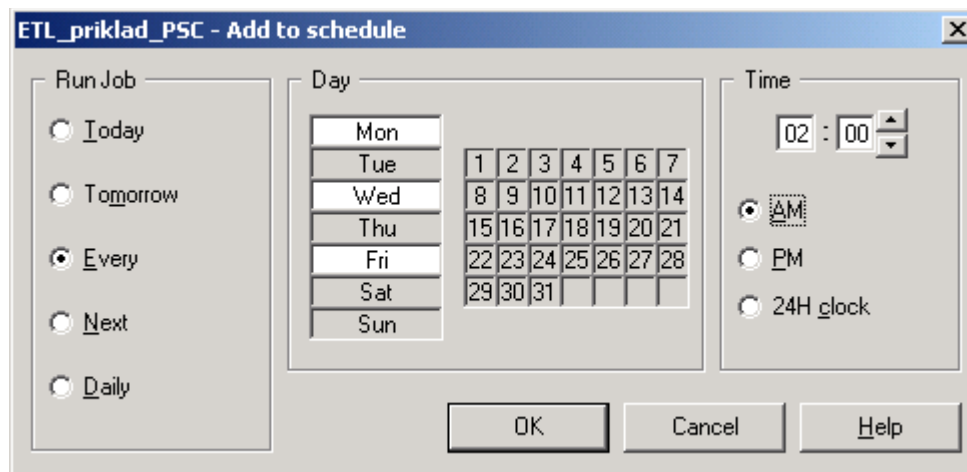
Prostřednictvím tohoto nástroje lze joby a sekvence spouštět, zastavovat, validovat, restartovat, monitorovat a sledovat jejich jednotlivé kroky v podrobném logu, který je automaticky vytvářen. Tento log obsahuje jak interní zprávy informující o tocích dat, tak zprávy systému i chybové kódy, které jsou vráceny prostřednictvím jednotlivých komponent. K naplánování automatického spouštění slouží plánovač, který umožňuje vytvořit kalendář spouštění jak pro jednotlivé joby, tak pro sekvence, a pro každý plán lze nastavovat jiné vstupní parametry.

Následující obrázek zobrazuje monitorování běhu jobu včetně jednotlivých prvků a datových toků, jejich propustnost a počty zpracovaných řádků.



Obr. 10 - DataStage Director

Takto vypadá nastavení automatického spouštění jobu. Bude se spouštět každé pondělí, středu a pátek ve dvě hodiny ráno.



Obr. 11 - Plánovač v nástroji DataStage Director

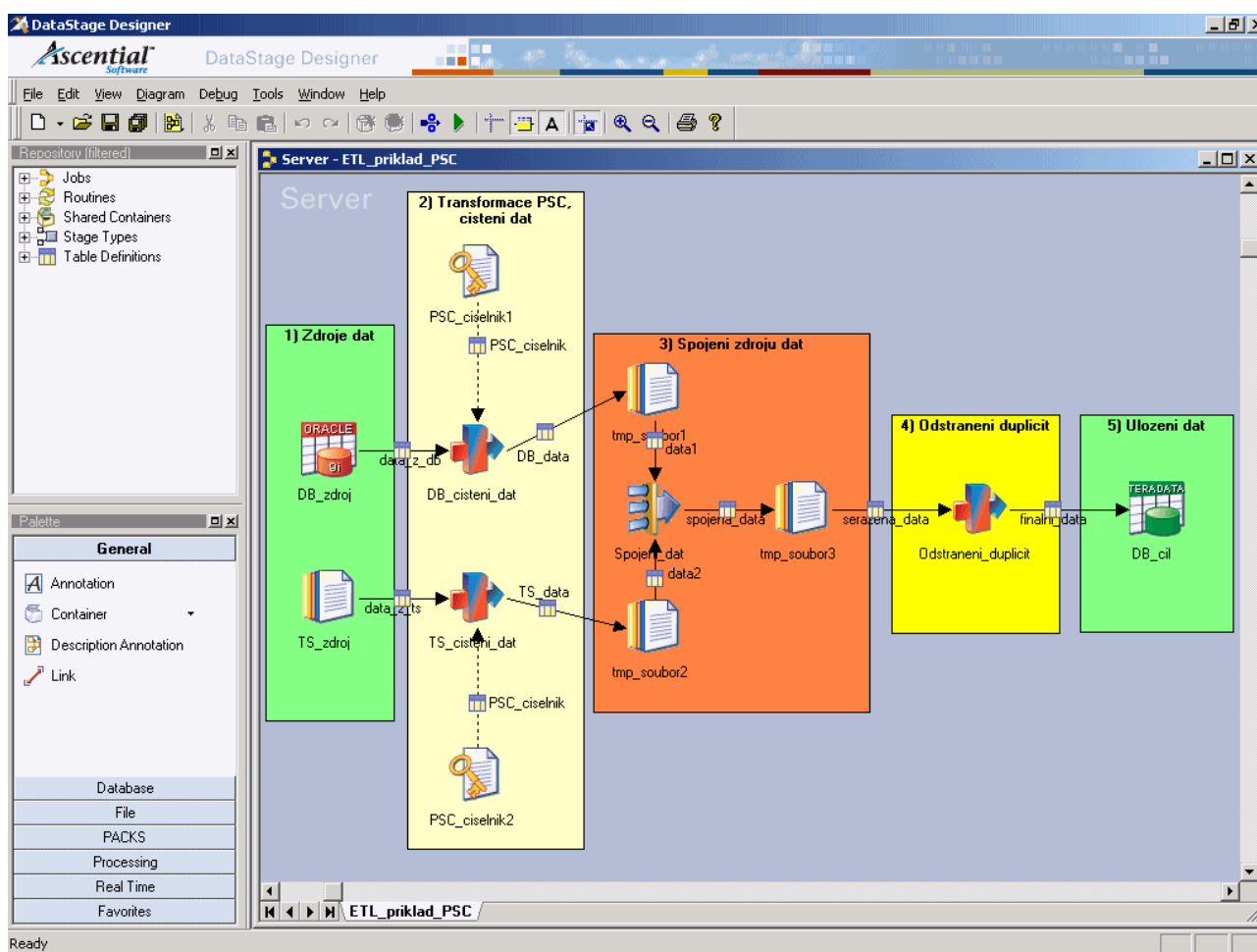
## 4.4 Návratnost investic do nástrojů ETL

Analýza návratnosti investic do ETL procesů je obtížná a liší se pro každou implementaci od jedné firmy ke druhé. Jedním ze základních rozdílů, který výši investice ovlivňuje zásadním způsobem, je současný stav IT firmy a její postoj k těmto technologiím. Základem je samozřejmě to, k čemu má ETL proces sloužit. Zda ke sjednocování a převodu dat mezi aplikacemi, při přechodu na nové systémy (například při fúzi či akvizici firem), nebo k plnění datových skladů. Dá se říci, že čím více je zdrojových systémů, čím jsou variabilnější a čím větší jsou objemy zpracovávaných dat, tím spíše se vyplatí profesionální ETL řešení.

## 5 Praktický příklad ETL procesu

V tomto příkladu bych chtěla ukázat jednoduchý návrh ETL transformace s použitím nástroje DataStage verze 7.1. od získání dat, přes transformaci, jednoduché čištění dat, až po jejich uložení do cíle.

Budu vycházet ze dvou zdrojových systémů obsahujících databáze klientů, které budou sloučeny do jednoho systému. Pro ukázkou jsem zvolila část, ve které je plněn číselník měst a jejich poštovních směrovacích čísel z těchto dvou zdrojových systémů. Oba systémy obsahují pouze neúplné číselníky, tzn. některé záznamy jsou pouze v jednom z nich, jiné mohou být v obou. Tyto duplicity je třeba eliminovat a pravidlo pro eliminaci zní, že přednost mají záznamy ze systému číslo 1.

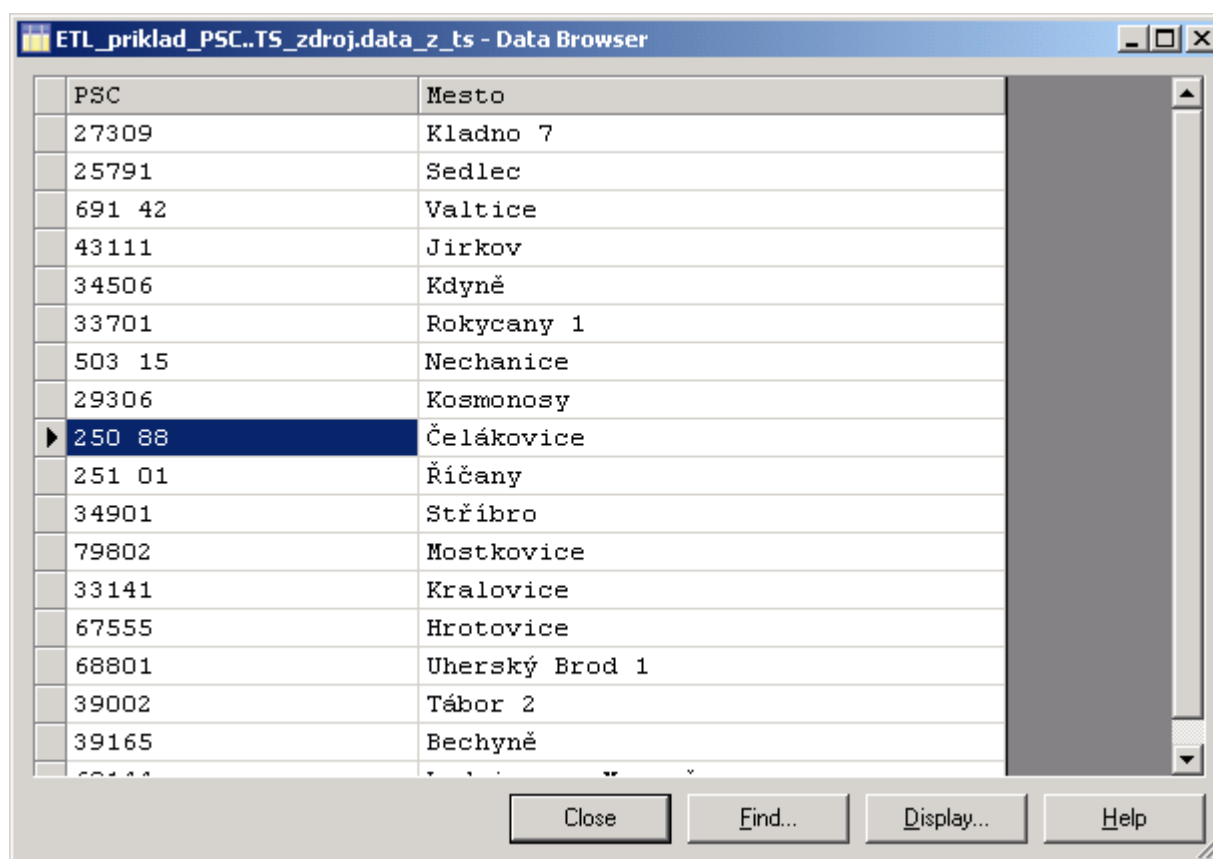


Obr. 12 - Výsledný job zobrazující proces transformace

**Krok 1) Zdroje dat**

Máme dva zdroje dat. Prvním je tabulka v Oracle databázi, druhým textový soubor. Nahrávat začneme data současně z obou zdrojů. Používáme zde následující dvě komponenty (stage):

- Oracle komponenta umožňující nativní připojení k databázi Oracle. V této komponentě definujeme připojovací údaje a SQL dotaz, pomocí kterého získáme data. Také jsou zde definovány datové typy nahrávaných dat.
- Flat File komponenta umožňující načtení dat z textového souboru uloženého na serveru DataStage. Stačí nadefinovat metadata popisující data uložená v souboru (názvy sloupců a datové typy), formát souboru (oddělovače, zda první řádek obsahuje názvy sloupců apod.) a cestu k souboru.

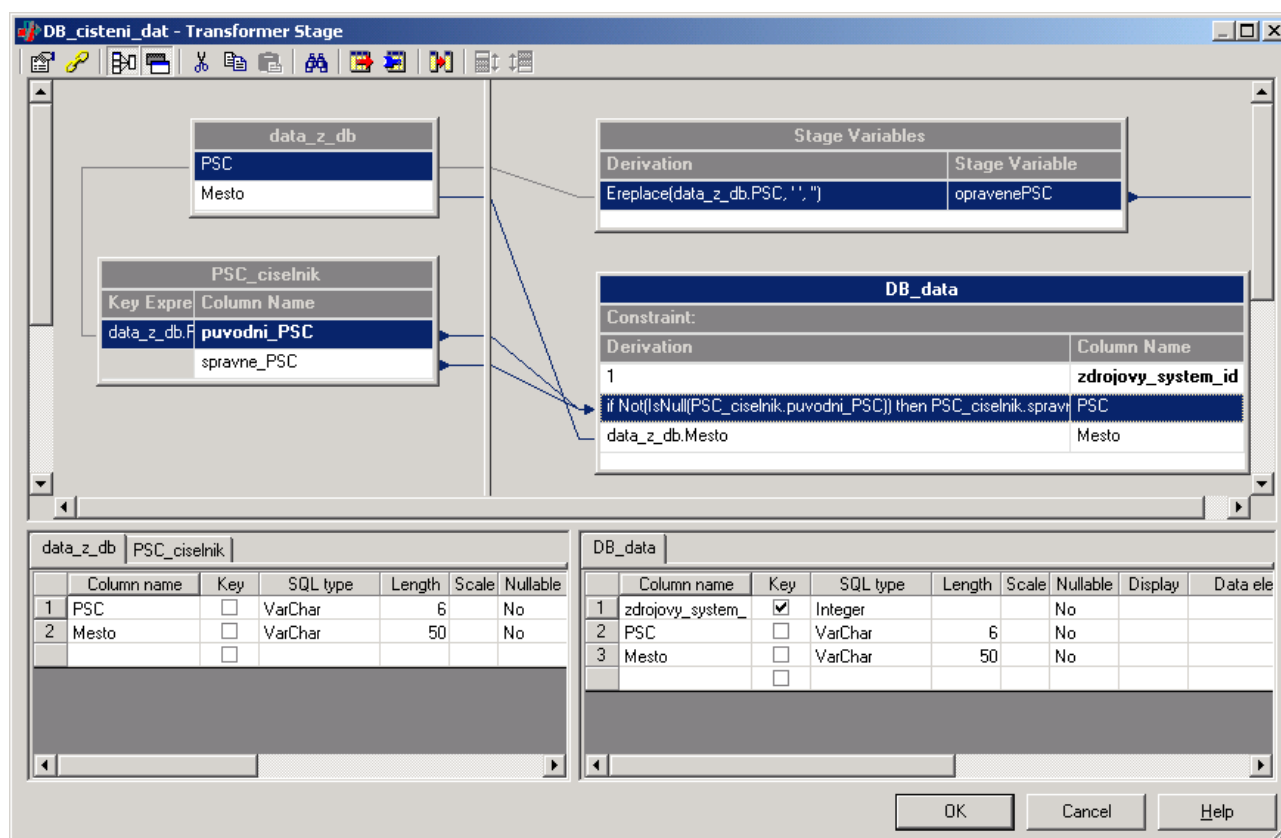


PSC	Mesto
27309	Kladno 7
25791	Sedlec
691 42	Valtice
43111	Jirkov
34506	Kdyně
33701	Rokycany 1
503 15	Nechanice
29306	Kosmonosy
250 88	Čelákovice
251 01	Říčany
34901	Stříbro
79802	Mostkovice
33141	Kralovice
67555	Hrotovice
68801	Uherský Brod 1
39002	Tábor 2
39165	Bechyně

Obr. 13 - Ukázka zdrojových dat v textovém souboru

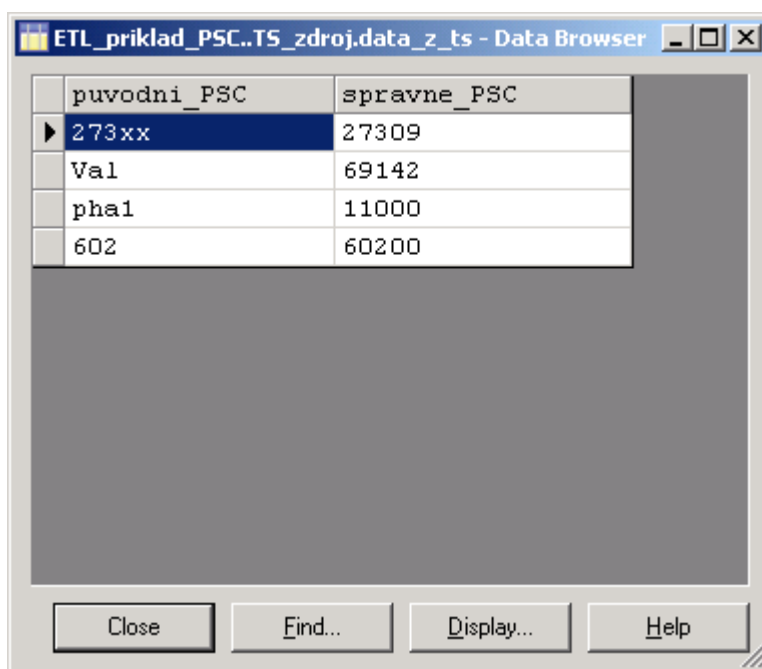
## Krok 2) Transformace PSČ, čištění dat

Jelikož víme, že v obou systémech se používají PSČ ve formátu "12345" a také ve tvaru s mezerou mezi 3. a 4. číslem tzn. "123 45", chceme tento formát sjednotit tak, že eliminujeme mezeru (výsledek ukládáme do Stage proměnné "opravenePSC").



Obr. 14 - Vlastnosti Transformer stage - definování transformace dat

Z původní analýzy zdrojových dat víme, že PSČ ve zdrojových systémech jsou občas chybná (například místo PSČ 11000 ( PSČ pro Prahu 1) je zde uvedena zkratka pha1 apod.). Proto jsme si vytvořili soubor, který obsahuje v jednom sloupci tato chybná PSČ a ve druhém jejich správnou náhradu. Tento soubor jsme připojili jako lookup a pokud nalezneme ve zdrojových datech některé z chybných PSČ, nahradíme je správnými.



puvodni_PSC	spravne_PSC
273xx	27309
Val	69142
pha1	11000
602	60200

Obr. 15 - Ukázka dat ze souboru pro nahrazování chybných dat

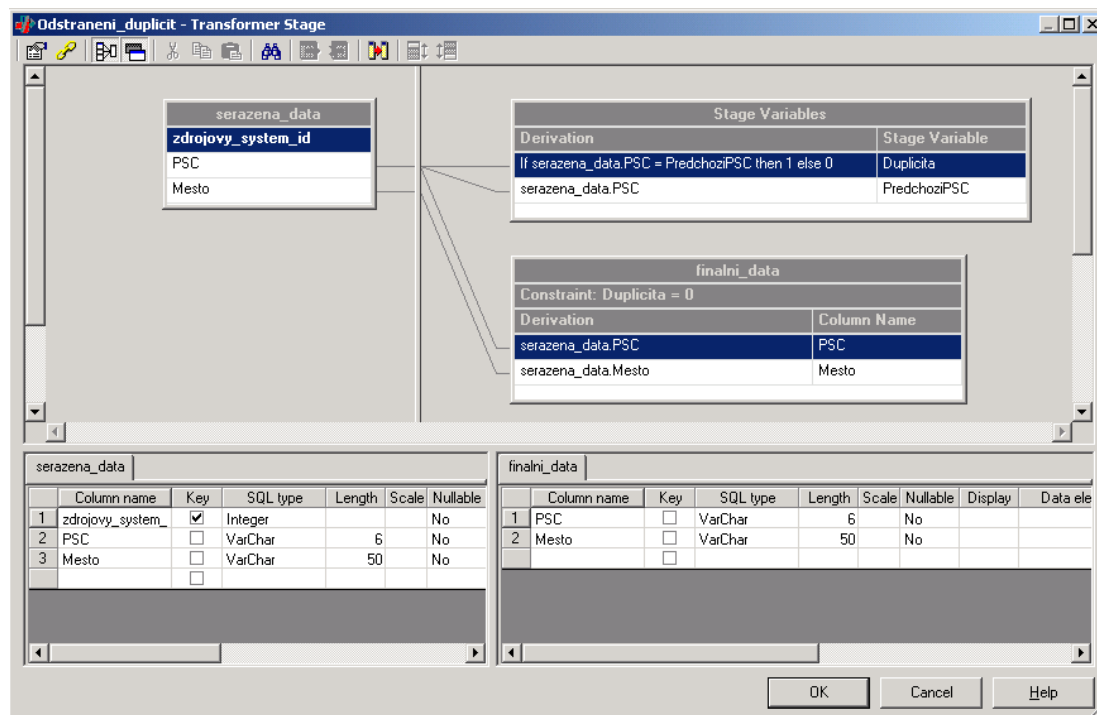
Nakonec ještě přidáme k datům sloupec zdrojovy\_system\_id, abychom i po spojení dat byli schopni rozlišit, ze kterého systému data pocházejí, a mohli tak při eliminování duplicit dostat pravidlu, že přednost mají údaje z Oracle databáze.

### Krok 3) Spojení zdrojů dat

Nyní máme oddělené dva toky dat, ve kterých jsou již opravená PSČ. Abychom mohli oba datové toky spojit, je potřeba data nejprve uložit do pracovních souborů a následně spojit prostřednictvím prvku Collector. V tomto prvku jsme také nadefinovali pravidlo, podle kterého se spojené záznamy seřadí, abychom byli schopni eliminovat duplicity.

### Krok 4) Odstranění duplicit

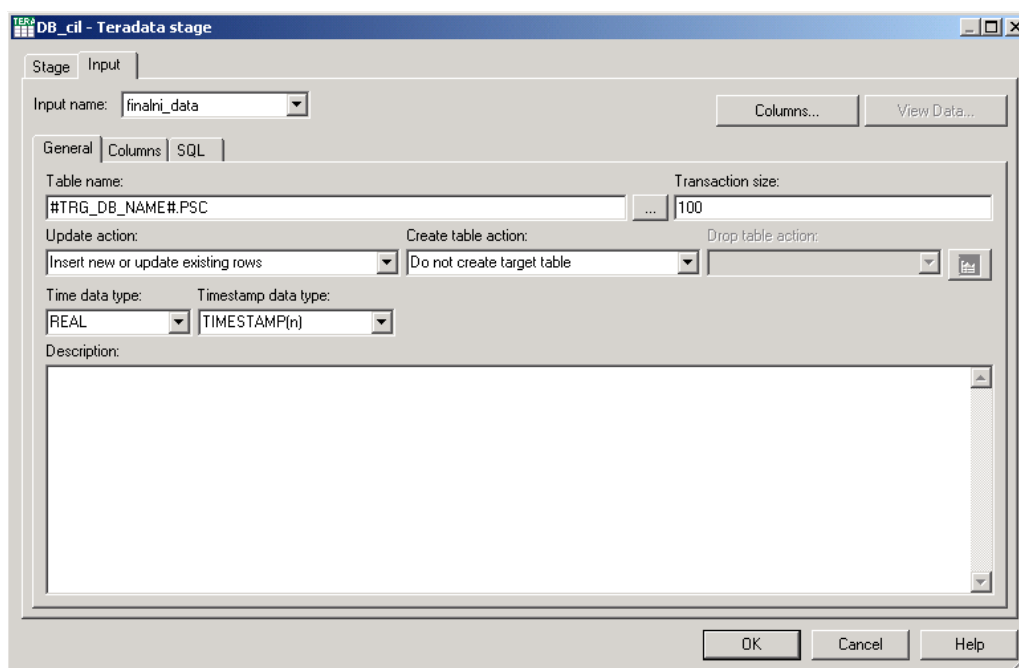
Použitý prvek v této fázi je Transformer, což je základní prvek pro provádění transformací v DataStage. Vstupní data přicházejí seřazena podle PSČ a čísla zdrojového systému. Zavedli jsme tedy Stage proměnnou “PredchoziPSC”, kam ukládáme právě zpracované PSČ a proměnnou “Duplicita”, ve které testujeme, zda nové zpracovávané PSČ není stejné, jako PSČ uložené v proměnné “PredchoziPSC”. Pokud ano, je proměnná “Duplicita” nastavena na hodnotu 1, v opačném případě na 0. Nakonec přidáme omezení pro tok dat (constrain) pro výstupní linku “finalni\_data”, která říká, že řádek dat může pokračovat pouze v případě, že Duplicita = 0.



Obr. 16 - Vlastnosti Transformer stage - odstranění duplicit

### Krok 5) Uložení dat

Zde je použita komponenta pro nativní připojení do databáze Teradata. V ní definujeme parametry pro připojení k databázi, názvy sloupců a datové typy cílové tabulky PSČ a způsob, jakým budou data do tabulky ukládána.



Obr. 17 - Nastavení parametrů pro cílovou databázi

## 6 Budoucnost ETL

### 6.1 ETL, ELT nebo ETLT?

Tradiční ETL nástroje pracují tak, že nejprve extrahují data z různých zdrojů, transformují je ve svém vlastním ETL serveru (vzdáleném serveru jak zdrojovým datům, tak cílové databázi) a následně transformovaná data uloží do cílového datového skladu. Data jsou tedy přenesena po síti dvakrát – jednou ze zdroje na ETL server, podruhé z ETL serveru do cílového datového skladu. Pokud je navíc při transformacích potřeba porovnávat transformovaná data s cílovými (například pro kontrolu referenční integrity), pak se musí data z cíle nejprve stáhnout, což způsobuje zpomalení a zvýšení síťového provozu.

Naproti tomu ELT architektura výše zmíněné problémy nemá. Základní rozdíl je ten, že se data nahrají do cílové databáze, kde se teprve transformují pomocí klasických SQL dotazů, které jsou generovány přímo pro daný cílový databázový systém, a tím plně využívá výkonu a škálovatelnosti cílového RDBMS. Tzn. není potřeba vlastní ETL server a odpadá i vícenásobné přenášení dat po síti či problémy s lookupy.

Další významný rozdíl je ten, že ETL proces zpracovává data řádek po řádku, ale ELT dělá transformace prostřednictvím SQL hromadně.

ELT nástroje se objevovaly již od roku 1990, ale nejpoužívanější databáze jako Oracle, DB2 a Sybase nepodporovaly dostatečně bohaté sady SQL operátorů potřebné pro zvládnutí komplexních transformací dat. V posledním desetiletí však dodavatelé databází zvýšili funkcionalitu SQL jazyka (například přidáním ranking a windowing funkcí (min over partition, max over partition, lead, lag, rank)) a umožňují tak efektivní zvládnutí komplexních transformací a agregací u velkých dat.

Samozřejmě i ELT nástroje mají své problémy, pro různé druhy transformací či práci s jinými zdroji dat než databázemi jsou vhodnější ETL nástroje, proto optimální cesta je umožnit vývojářům rozhodnout, které řešení je pro dané transformace vhodnější, a umožnit jim v jednom nástroji pracovat s oběma možnostmi, tzn. cosi jako ETLT přístup.

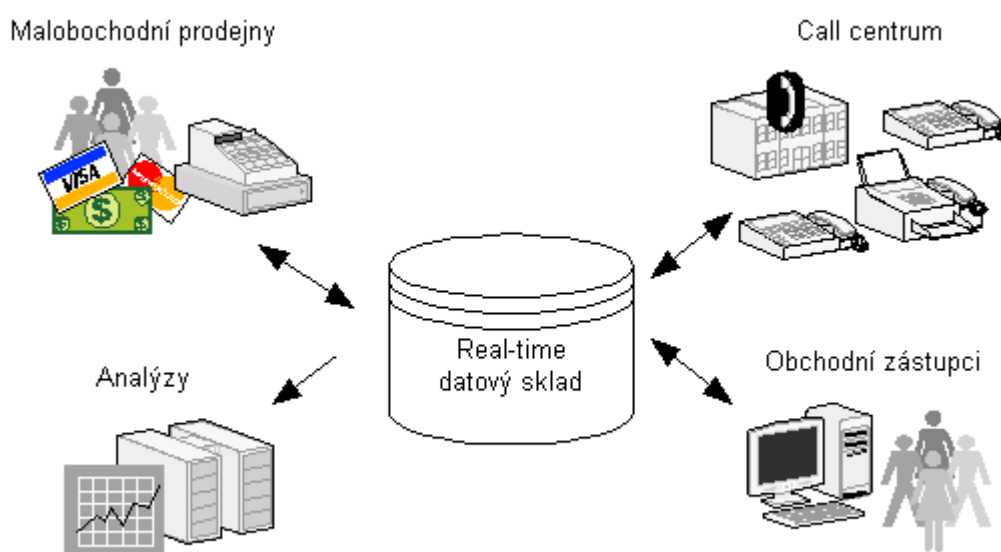


## 6.2 Real-time datové sklady

Real-time datové sklady jsou považovány za další krok v evoluci datových skladů a ETL procesy v tom hrají základní roli.

Firmy jsou díky datům uloženým v datových skladech schopné provádět dolování dat, detailní analýzy a získávat důležité znalosti o trhu či zákaznících. Tyto informace by významným způsobem ovlivnily fungování celé firmy, pokud by byly zpřístupněny do prodejen, call center, obchodním zástupcům - každému, kdo je přímo ve styku se zákazníky.

Jelikož však klasický dávkový ETL proces sám data zpracovává často celé hodiny, nastává zde problém s aktuálností dat, která nesmí být starší než pár minut, či vteřin. Zde je potřeba jiné řešení - real-time datový sklad.



Obr. 18 - Real-time datový sklad

Real-time datový sklad obsahuje kompletní aktuální data a je synchronizován se zdrojovými systémy, které tato data poskytují. U near-real-time datových skladů existuje určité zpoždění.

Pro ETL nástroje to tedy znamená, že plnění cílového datového skladu musí probíhat během běžné pracovní doby bez omezení přístupu uživatelů ke zdrojovým systémům. Je tedy potřeba dosáhnout

následujících podmínek:

- snížit na minimum čas potřebný k získání nových a změněných dat ze zdrojových systémů
- zredukovat čas potřebný k čištění, transformaci a uložení dat
- co nejvíce snížit čas potřebný na přepočítání agregací

Pro tyto účely je potřeba upravit provozní zdrojové systémy, tak aby bylo možné z nich brát data bez přílišného zatížení systému. To lze provést například tak, že se změny v datech, ke kterým dojde během dne, ukládají do paralelních tabulek. Jakmile se pak rozběhne ETL proces, čerstvě nahrané tabulky se jen přehodí do produkce a tabulky se starými daty jsou naopak z produkce odstraněny. V takovém případě se jedná spíše o near-real-time datový sklad, což je pro většinu případů dostatečné řešení. Pokud je skutečně vyžadován přístup real-time, pak se obvykle upraví provozní systémy, tak aby se měněná data zároveň zapisovala i do cílového datového skladu. Zde je samozřejmě klíčová volba správného ETL nástroje, který má pro zvolené postupy nejlepší podporu a poskytuje nejvyšší výkon.

### 6.3 Datové gridy - ETL architektura příští generace

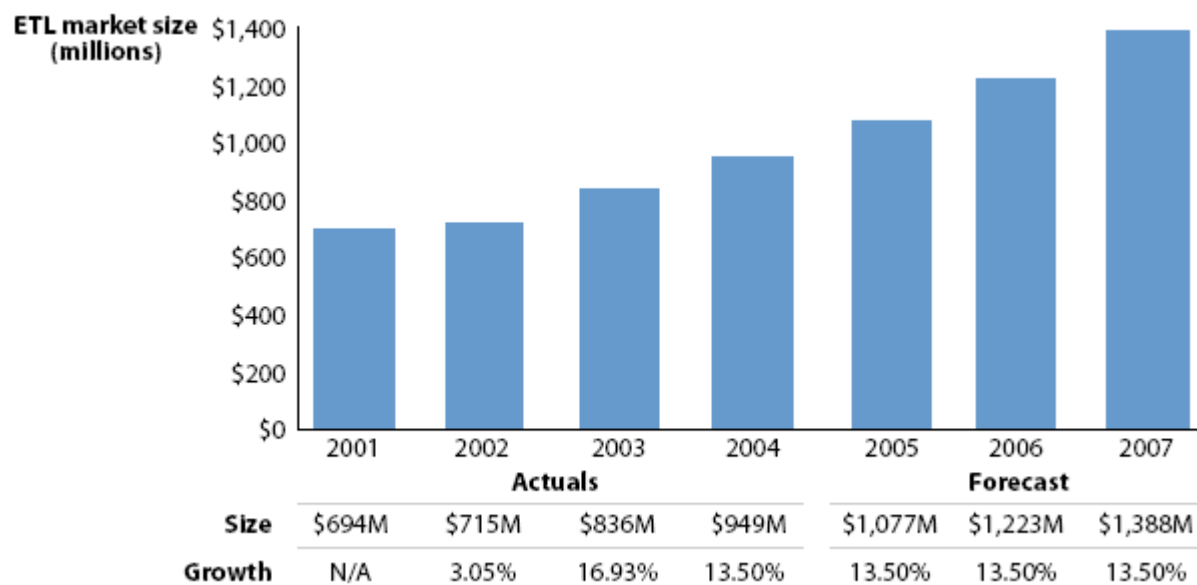
Grid computing je architektura, která poskytuje mechanismy pro přístup, slučování a řízení infrastruktury počítačů v síti. Tzn. cílem je sdílet různé IT zdroje, které se pak chovají jako jeden systém, pro jednotlivé procesy a výpočty, tak aby jeden výpočet mohl být rozdělen mezi jednotlivé počítače v síti.

Architektura datových gridů umožňuje přistupovat k informacím z různorodých zdrojů a kombinovat je jako součást jednoho celku dat. Například aby mohla firma provádět analýzy pro dolování dat na více různých databázích společně, aniž by musela posílat velká množství dat po síti (a nejprve data sloučit v jedné databázi).

Tato technologie je považována za příští generaci ETL architektury (a nejen v ETL) a na jejím vývoji pracuje mnoho velkých firem jako IBM, Microsoft, Oracle, HP, SUN atd.

### 6.4 Budoucí vývoj ETL trhu

Následující graf pochází z výsledků výzkumu firmy Forrester Research, Inc a ukazuje, že růst ETL trhu je posledních několik let stabilní a vykazuje růst nad 10% ročně, což je mu předpovídáno i na následující rok.



Obr. 19 - Vývoj ETL trhu

Obrázek převzat od firmy Forrester Research, Inc ([www.forrester.com](http://www.forrester.com))

Také na otázku položenou firmám, zda plánují používat ETL nástroje více či méně než doted', odpovědělo plných 79%, že se je chystají využívat více.

Dá se tedy předpokládat, že trh ETL v následujících období nečeká veliký skok či propad, ale udrží si podobné tempo růstu jako doposud.

## Závěr

Ráda bych nyní shrnula veškeré poznatky získané během vytváření této práce. Přestože o ETL neslycháme příliš často, je základním stavebním kamenem při budování datových skladů, migracích dat a vytváření datových základů pro BI aplikace. Jejich implementace dokonce v rámci takového projektu obvykle zabírá nadpoloviční většinu času a prostředků.

Stejně jako se neustále zvyšují objemy dat, které firmy produkují při svých běžných aktivitách, zvyšují se i nároky na aktuálnost a objemy dat v datových skladech, což klade stále vyšší požadavky na ETL řešení a kvalitu, výkonnost a škálovatelnost ETL nástrojů.

Přestože ETL implementace nejsou ve světě žádnou novinkou, u nás je to poměrně mladá oblast a můžeme se s ní setkat pouze u opravdu velkých podniků, jakými jsou banky, pojišťovny, mezinárodní společnosti apod.

Jelikož u nás zatím nejsou ETL procesy příliš rozšířené, nezabývá se jimi ani odborná literatura a pokud se chceme dozvědět něco více, narazíme maximálně na velmi odborné články či diskuze zaměřené na specifickou oblast ETL.

Přínosem této práce je úvod do problematiky ETL procesů, od vysvětlení důležitých pojmů, shrnutí nejčastějších problémů při vlastních implementacích, přes přehled hlavních nástrojů až po praktický příklad vlastního ETL procesu v jednom z těchto nástrojů. Tato práce může být prospěšným studijním materiálem jak pro začínající ETL vývojáře, tak pro managery, kteří mají přijít s ETL procesy teprve do styku.

## Použitá literatura

- [1] Lacko, L.: Databáze: datové sklady, OLAP a dolování dat. 1.vyd. Brno: Computer Press 2003. 488 s. ISBN 80-7226-969-0.
- [2] Humphries, M. a kol.: Data warehousing – návrh a implementace. 1.vyd. Praha: Computer Press 2001. 257 s. ISBN 80-7226-560-1.
- [3] Články ze zpravodajského portálu SystémOnLine, <http://www.systemonline.cz/>
- [4] <http://www.sunopsis.com/>
- [5] <http://www.forrester.com/>
- [6] <http://www.dbazine.com/>
- [7] <http://datawarehouse.ittoolbox.com/>
- [8] <http://www.adastracorp.com/>
- [9] <http://www.ibm.com/>
- [10] <http://www.informatica.com/>
- [11] <http://www.oracle.com/>
- [12] <http://www.sas.com/>
- [13] <http://www.microsoft.com/>
- [14] <http://www.businessobjects.com/>
- [15] <http://www.etlguru.com/>
- [16] <http://www.learndatamodeling.com/>

# Seznam použitých termínů a zkratek

**BI** (Business Intelligence) - podnikové zpravodajství; sada technologií a aplikací pro sběr, dolování, dotazování, analýzy a reporting dat, které slouží jako podpora firmám pro zkvalitnění obchodních rozhodnutí

**Čistění dat** - proces opravování chyb ve zdrojových datech před jejich uložením do datového skladu

**CRM** (Customer Relationship Management) - systém pro řízení vztahů se zákazníky, v rámci kterého se sbírají veškerá obchodní data týkající se zákazníků a následně se využívají pro zkvalitnění individuálního přístupu k zákazníkům

**Data profiling** - proces analýzy dat, tvorby metadat a pravidel

**Data Staging Area, Staging Area** - prostředí (obvykle databáze či soubory) které je používáno při operacích s extrahovanými zdrojovými daty před jejich uložením do cílového systému

**Datové gridy** - technologie umožňující společnou práci s daty uloženými v různých systémech na různých platformách nezávisle na jejich umístění

**DTS** (Data Transformation Services) - ETL nástroj od firmy Microsoft, který je součástí MSSQL serveru

**DW** (Data Warehouse) - datový sklad, nebo-li databáze obsahující kolekce historických dat pro analýzy

**ELT** (Extract, Load and Transform) - označení pro ETL proces, kdy se nejdříve data extrahují ze zdrojového systému, následně uloží do cílové databáze a pak se teprve provádějí transformace dat

**ERP** (Enterprise Resource Planning) - informační systém, ve kterém jsou integrovány všechny důležité obchodní a výrobní aplikace podniku

**ETL** (Extract, Transform and Load) - sada procesů pro migraci a transformaci dat, kdy se data získají ze zdroje, poté upraví (transformují) a nakonec uloží do cílového systému

**JDBC** (Java DataBase Connectivity) - standard, který definuje přístup k databázím pro aplikace napsané v jazyce Java

**LDAP** (Lightweight Directory Access Protocol) - protokol který umožňuje přístup k adresářovým službám, což jsou databáze, které se používají pro správu adresářových položek (jako jsou např. kontakty)

**Mapování** - definování vztahů mezi zdrojovými a cílovými daty

**Metadata** - data popisující jiná data (jako jsou datové typy, struktury, obchodní pravidla atd.)

**Metadata repository** - úložiště metadat, tzn. místo (obvykle to bývá databáze), kam se ukládají různé typy metadat. Většinou jsou tato metadata sdílená pro více uživatelů či systémů.

**Multithreading** - technologie pro paralelní běh více procesů (vláken) v jednom systému

**ODBC** (Open Database Connectivity) - standardizované rozhraní pro přístup k relačním databázím od firmy Microsoft

**ODS** (Operational Data Store) - provozní datový sklad, který obsahuje repliky dat z provozních systémů, většinou za období 1-2 měsíců

**OLE DB** (Database Object Linking and Embedding) - standard datového rozhraní pro přístup k datům

**PL/SQL** (Procedural Language/Structured Query Language) - objektově orientovaný programovací jazyk vyvinutý společností Oracle, který rozšiřuje vlastnosti jazyka SQL

**RDBMS** (Relational Database Management System) - databázový systém, který umožňuje přístup k normalizovaným datům uloženým v tabulkách. Tabulky obsahují sloupce a řádky a mohou být vzájemně spojovány.

**SQL** (Structured Query Language) - dotazovací jazyk pro manipulaci s daty v relačních databázích

**XML** (eXtensible Markup Language) - značkovací jazyk, který umožňuje uspořádat data do struktur pomocí značek

# Seznam obrázků

Obr. 1 - Základní části ETL procesu.....	10
Obr. 2 - ETL v Bussines Intelligence.....	11
Obr. 3 - Toky dat v ETL.....	12
Obr. 4 - Metadata repository.....	14
Obr. 5 - Diagram ETL procesů a datových toků.....	15
Tabulka 1 - Specifikace mapování zdrojových dat na cílová.....	18
Obr. 6 - Podíly významnějších ETL nástrojů na trhu (rok 2002).....	25
Obr. 7 - DataStage Administrator.....	30
Obr. 8 - DataStage Manager.....	31
Obr. 9 - DataStage Designer.....	32
Obr. 10 - DataStage Director.....	33
Obr. 11 - Plánovač v nástroji DataStage Director.....	34
Obr. 12 - Výsledný job zobrazující proces transformace.....	35
Obr. 13 - Ukázka zdrojových dat v textovém souboru.....	36
Obr. 14 - Vlastnosti Transformer stage - definování transformace dat.....	37
Obr. 15 - Ukázka dat ze souboru pro nahrazování chybných dat.....	38
Obr. 16 - Vlastnosti Transformer stage - odstranění duplicit.....	39
Obr. 17 - Nastavení parametrů pro cílovou databázi.....	39
Obr. 18 - Real-time datový sklad.....	41
Obr. 19 - Vývoj ETL trhu.....	43