

Aktuální otázky provozu datových skladů

PAVEL HNÍK

K čemu slouží datové sklady

- IT podporuje business podniků
- S velikostí podniku se zvyšuje náročnost zpracování dat
- DWH = unifikovaná datová základna pro výstupní analýzy BI řešení.



Agenda

1. Faktory ovlivňující provoz DWH

- Architektura DWH – volba komponent, realizace ETL
- Požadavky businessu – jaká data, v jaké formě, kdy

2. Provoz DWH, rozdělení hlavních činností

- Týmy zabezpečující provoz DWH – pravomoce, odpovědnosti, SLA
- Podpora produkčního prostředí – práce v režimu 24/7
- Podpora rozvoje DWH – ETL, BI, infrastruktura

3. Problémy při provozu DWH, klasifikace chyb

- HW chyby – omezení a selhání jednotlivých komponent
- SW chyby – pády ETL úloh, chyby v datech

4. Shrnutí, prostor pro dotazy

1. Faktory ovlivňující provoz DWH

Architektura DWH

- Volba řešení DWH

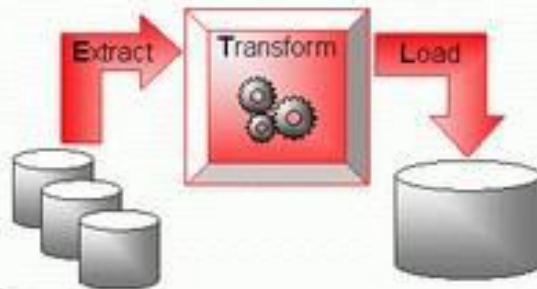
- Rozvoj starého DWH nebo implementace nového
- Dodavatel většinou externí, spolupráce s interním IT týmem
- Velikost rozpočtu

- Volba komponent

- Jaká databáze datového skladu (Oracle vs Teradata)
- Jak nalít data z primárních systémů (Transportní server, ODS)
- Jak realizovat ETL (ETL vs ELT)

E-LT Architecture

Conventional ETL Architecture



Transform in Separate ETL Server

- Proprietary Engine
- Poor Performance
- High Costs
- IBM & Informatica's approach

Next Generation Architecture

“E-LT”



Transform in Existing RDBMS

- Leverage Resources
- Efficient
- High Performance

Benefits

- Optimal Performance & Scalability
- Easier to Manage & Lower Cost

Obrázek 2: ETL vs ELT

How Oracle GoldenGate Works

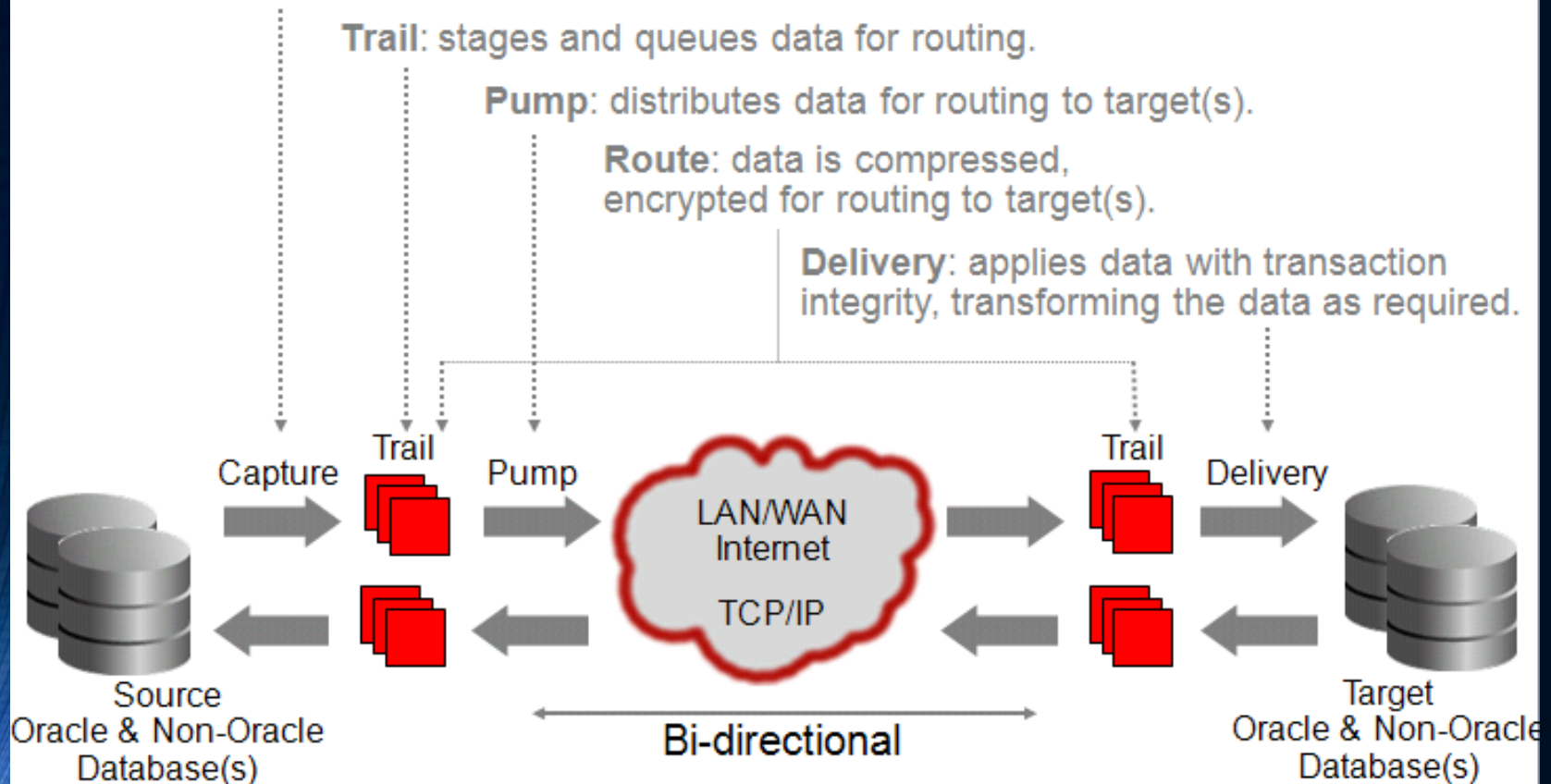
Capture: committed transactions are captured (and can be filtered) as they occur by reading the transaction logs.

Trail: stages and queues data for routing.

Pump: distributes data for routing to target(s).

Route: data is compressed, encrypted for routing to target(s).

Delivery: applies data with transaction integrity, transforming the data as required.



ORACLE

Obrázek 3: Oracle GoldenGate

Požadavky businessu

- Spokojenost businessu je hlavním cílem BI řešení
 - Informace jsou přínosem a přímo nebo nepřímo zvyšují zisk podniku
 - Tím se opodstatňují i další investice do rozvoje DWH
- Jaká data?
 - Data o klientech k vymáhání dluhů nebo data pro zpětné hodnocení kampaní?
- Jak rychle?
 - Kdy budou business okna?

2. Provoz DWH, rozdělení činností

Rozvoj DWH

- Požadavky na změnu dat
 - Release na primárních systémech
 - Změna portfólia produktů
 - Oprava chyb v datech
- Rozvoj infrastruktury
 - Omezení jednotlivých komponent řešení (nedostatek diskového prostoru)
 - Nedostatečná rychlost zpracování (optimalizace)
 - Konec podpory ze strany dodavatelů (časově omezená platnost licencí)

Denní zpracování dat

- ETL proces
 - Objem dat, množství systémů
 - Zpracování dat rozdělené do několika kroků
 - Dohromady několik tisíc malých úloh
 - Parametry úloh:
 - Dopad do loadu v případě chyby
 - Závislost na ostatních úlohách
 - Restartovatelnost
 - Délka zpracování

Týmy zabezpečující provoz DWH

- Několik týmů zodpovědných za jednotlivé oblasti
 - Dodávka dat z primárních systémů
 - Zpracování dat (analýza, vývoj, provoz)
 - Hardware
 - Datová kvalita
- Komunikace mezi týmy, role koordinátorů
- Jaké služby je nutné zabezpečit
 - Smluvní podmínky
 - Jaké jsou SLA (Reakce na kritické HD)

Podpora produkčního prostředí

- Monitorování loadu
 - Rozdělení na denní a noční služby
- Proaktivní vs reaktivní přístup k chybám
 - Vyhodnocování chyb, eskalace
- Nasazování nového kódu
- Provozní reporting (dle smlouvy např. měsíčně)

Monitorování loadu

- Co to vlastně je, kdo jej konkrétně zajišťuje
 - provádí se na TEST i PROD prostředí
- Způsoby monitoringu – vychází z architektury
 - scheduler – interní, externí
 - inteligentní helpdesk
- Pravomoce, odpovědnosti, eskalační proces
 - Provozní problémy
 - problém kvality dat (kdo rozhodne o zahození záznamu)
 - Špatný kód (eskalace na vývoj)

Plocha Office Viewer VPA - Citrix Receiver

PDC-Monitor - Windows Internet Explorer

Soubor Úpravy Zobrazit Oblíbené položky Nástroje nápověda

Oblíbené položky

PDC-Monitor PDC-Monitor x

Running Jobs : **1** Failed Jobs : **0** Environment : PROD Load Date : 17.09.2013 Engine # 0 Task(s) : 20 Task Type : AUTOMA

Filters : Filter is OFF Active Engine ID : 0

Přihlášeny uživatel: Pavel Hnik
Logout

Monitor
Control
Scheduler
Finish jobs
Restart Failed Jobs
Restart Finished Jobs
Block Jobs
Unblock Jobs
Manual Batch
Logs

Stream Statistics		Job Statistics		
Runnable Category	Count	Status	Description	Count
FINISHED	214	2	Finished successfully	200
FORCE_FINISHED		172	Skipped	2168
RUNABLE	1865	R	Force finished failed job	1
RUNNING	1	0	Ready to run	2001
ALL	2681	100	Ready to skip	2702
		1	Running	1
		ALL		10700

Job Details => Status 1
Prev 1
Records 1 to 1 of 1

Job ID	Job Name	N run	Last Update	Status
10709420	POKRP_DYN_PROD_PROD_L1	1/1	19.09.2013 10:25:07	1

Obrázek 4: Externí scheduler



Advanced

Ve frontě	4
Active	261
Uzavřen	11872
Request	6
Internal Solutions	0
Solutions	2

Numara FootPrints 10

Welcome to Numara FootPrints 10

Featuring homepage usability improvements for filtering, dynamically resizing and reordering your homepage list, and seamless integration with asset management.

[Version 10 Interface Video](#)
[Version 9.5 Interface Video](#)
[Learn Everything About Version 10](#)

Global Tickets

There are no Global Tickets at this time

Number	Priority	Title	Status	Assignees	Last Updated
12162	INFO	Testovací kolo T0A Ahoj, prosím o realizaci mimořádného testu minikola T0A včetně nasazení balíčků N00456 a N00457 na serveru Pakárna. Sdíly jic [more...]	Request	Not Assigned	1 hr 40 min ago
12161	ERRO	Job BCKP_DWHPROD_DICTIONARY job_id=10039410 has failed. ... V řešení [more...]	V řešení	Michal Baranek	6 hrs ago
12158	WARN	Job WF_TRFL1_POELPVL_POJ_ELEMENT_PJZ_APO job_id=1069403... v řešení [more...]	V řešení	Michal Baranek...	14 hrs ago
12157	WARN	Job WF_TRFL1_WRK_APORIZ_APO job_id=1069403 is running f... V řešení [more...]	V řešení	Michal Baranek...	17 hrs ago
12156	ERRO	Job BCKP_DWHPROD_DICTIONARY job_id=10039678 has failed. ... Eskalováno na dostupný CPOJ [more...]	V řešení	Michal Baranek...	20 hrs ago
12150	INFO	Testovací kolo T08 Ahoj, prosím o zpracování T08 k testu T08 včetně nasazení následujících balíčků na TEST. M. Endrizal N00457, RLS_20130719_FV_LY_1109.K.Nolubová [more...]	Request	Not Assigned	1 day 20 hrs ago
12137	INFO	Nasazení k: load_date=14.9.2013 Nasazeno na PROD před load_date=14.9.2013 [more...]	Vyřešen	Pavel Hník...	2 days 16 hrs ago
12133	INFO	aa-13.9.2013 bez KOPSML KOPSML proskocane pro vlt. date=13.9.2013 [more...]	Vyřešen	Pavel Hník	3 days 13 hrs ago
12119	WARN	Job WF_HIST_PUKSTAV_POJRIZ_STAV_KDA_25000006 job_id=100... v řešení [more...]	V řešení	Pavel Hník...	5 days ago

Obrázek 5: Inteligentní helpdesk

Podpora rozvoje DWH

- Vývoj nového kódu
 - Business požadavek -> Datový architekt (nacenění, specifikace zadání) -> Analýza (modely) -> vývoj (ETL nástroj/SQL procedury) -> test kolo (provoz/vývoj) -> produkce
- Infrastruktura DEV-TEST-PROD
 - rozdíly proti DEV-PROD (případně PREPROD)
- Rychlost nasazování nového kódu
 - Cca 5-15 balíčků měsíčně (projekty s 35 balíčky týdně)
- Rozdělení činností mezi týmy (spolupráce/komunikační matice)
 - DBA, vývoj, analytický tým (test), provoz
 - Činnosti ostatních týmů může být chápán jako blackbox

3. Problémy při provozu DWH, klasifikace chyb

Příklady chyb v DWH

- Chyby na primárních systémech (špatná vstupní data)
 - Chyby v ETL (nekorektní transformace dat)
 - Selhání HW komponent (selhání diskového pole)
 - Ostatní chyby (plošný výpadek proudu)
-
- Omezení systému (výkon, kapacita)

Čím jsou chyby způsobené

- Chyba v prostředí (nedodání očekávaných dat)
- Špatná analýza/vývoj (provázanost mezi úlohami)
- Chyba při nasazení nového kódu
- Kategorizace chyb dle jejich projevu
 - Pád úlohy
 - „Zombie“ úloha
 - Provozně ok, generující špatná data
- Způsob opravy chyb
 - Fix vs hotfix (standardní postup vs ad-hoc řešení)

Shrnutí

- Jednoduchost/složitost provozování DWH vychází z architektury řešení
- Jak je DWH (potažmo celé BI řešení) vnímáno je závislé na businessu – na něm je také závislá výše investic do DWH
- Provoz zařituje mnoho týmů pro každou konkrétní oblast s odpovídajícími povinnostmi a pravomocemi
- Chyby v DWH, jejich příčiny a dopad
- Zákazník vs dodavatel

Děkuji za pozornost