

Základy Data Vault

Jak vytěžit maximum z vašich dat

Kateřina Saryčeva, Ing

Ness Czech

katerina.saryceva@ness.com

 [linkedin](#)



Cíl přednášky

- Představit základní koncept Data Vault
- Vysvětlit proč dnes o něj roste zájem
- Upozornit na potenciální potíže a jak je obejít

Agenda

- Koncepty, na které jsme si již zvykli
- Základní principy Data Vault
- Výhody a nevýhody
- Příklad ze života a doporučení
- Diskuze

Přístupy k architektuře DWH

...a modelování

Ralph Kimball

Business Dimensional Lifecycle

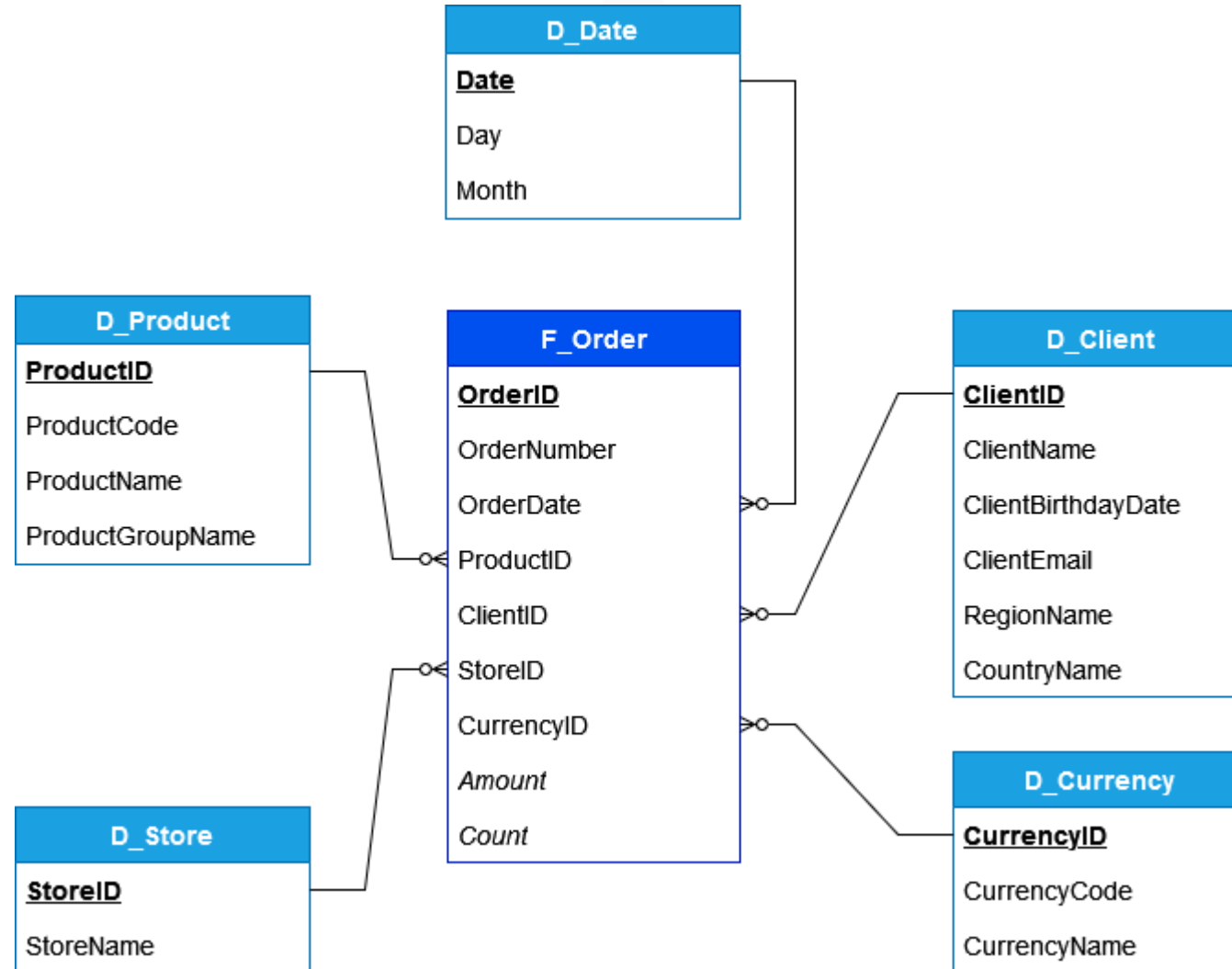
postup shora dolů

- + srozumitelné pro odběratele dat
- + optimalizované pro čtení dat a tvorbu reportů
- duplikace dat
- Chyby při aktualizaci dat
- Potíže při provedení změn ve strukturách



STAR SCHEMA

Star schema

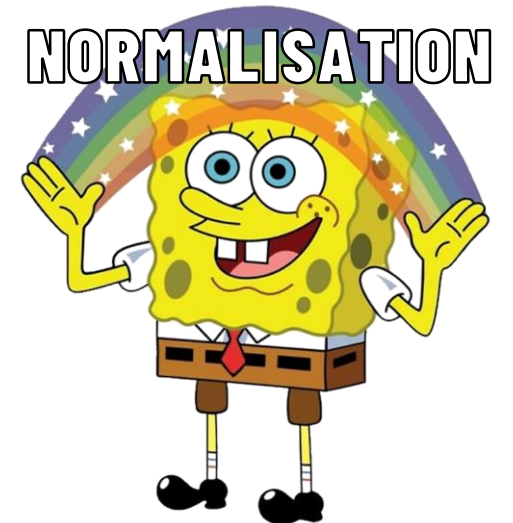


Bill Inmon

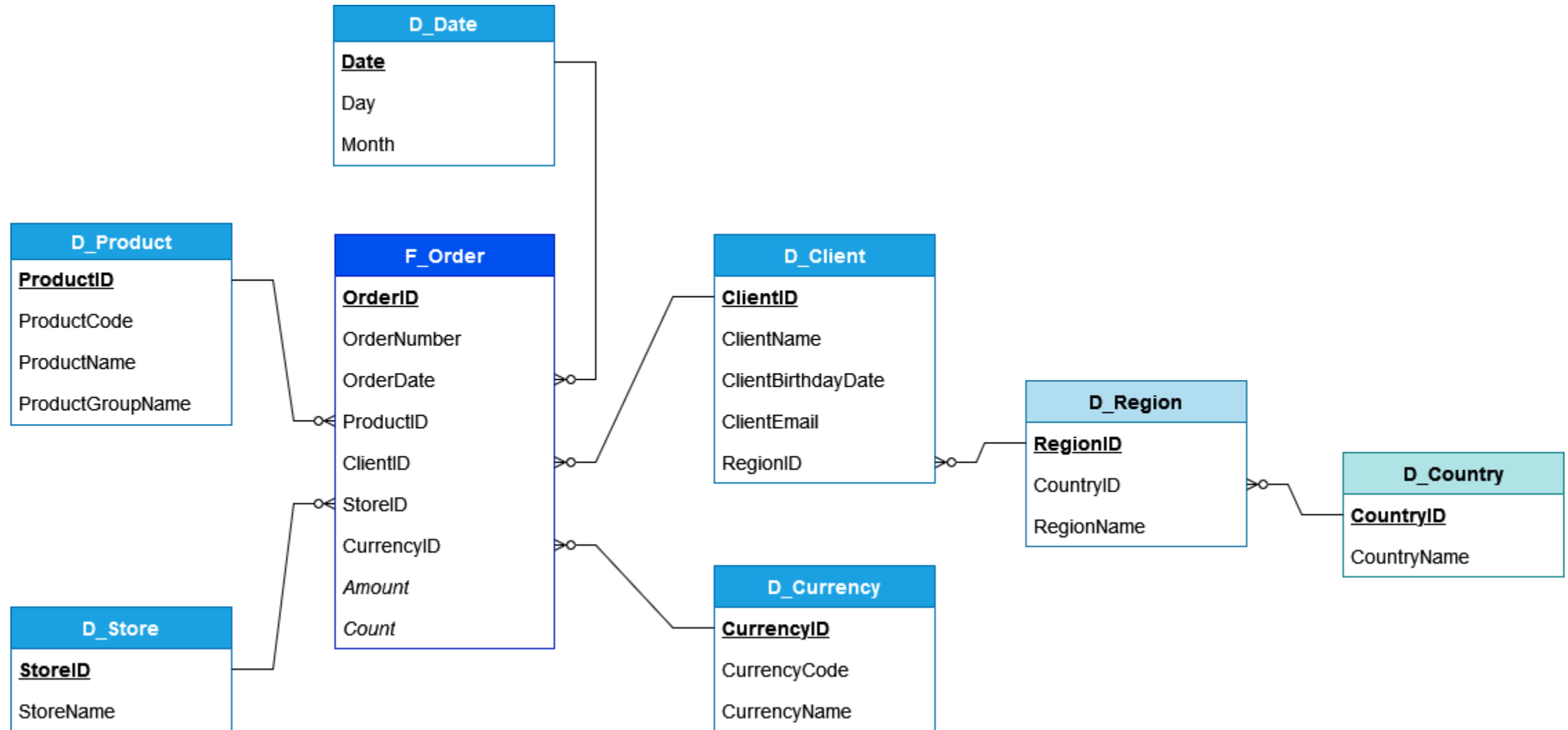
Enterprise Data Warehouse

postup zdola nahoru

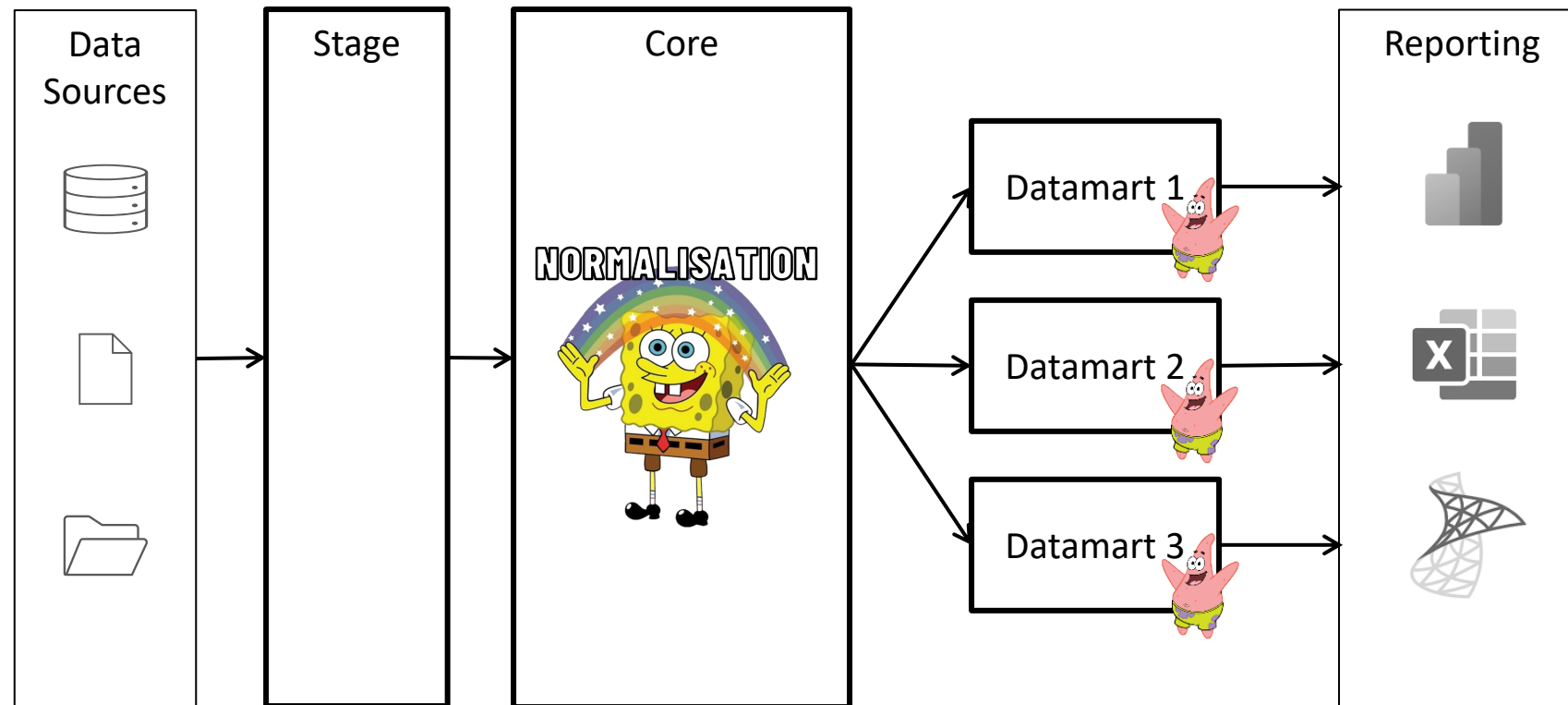
- + místo jedné pravdy pro celou firmu
- + deduplikace dat, méně chyb při aktualizacích
- náročný na zdroje a čas
- komplexní, ale flexibilní



Snowflake schema



„The best of...”



Tlak na zrychlení a optimalizaci

- Drahý výpočetní výkon
- Velký nárůst dat
- Hodně zdrojových systémů
- Hledání jedné pravdy trvá
...a ještě k tomu se pravda mění

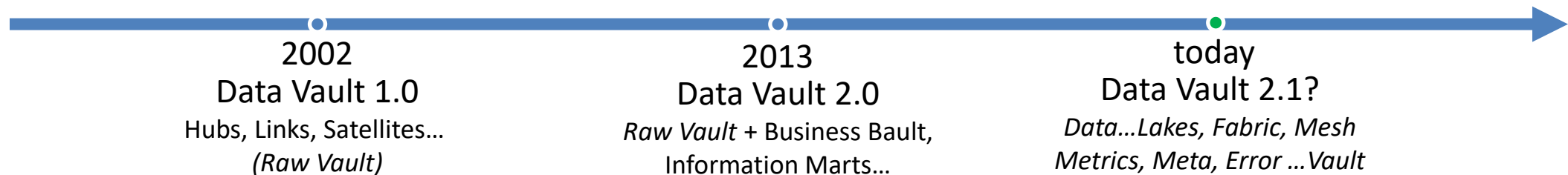
Data Vault

Dan Linstedt, Michael Olschimke

Data Vault

všechna data jsou relevantní

- + jednoduché rozšiřování, vkládání dat
- + zapadá do moderních tendencí
- + předpokládá se změnami řešení
- hodně komplexní
- málo zkušených odborníků



Jak zrychlit DWH

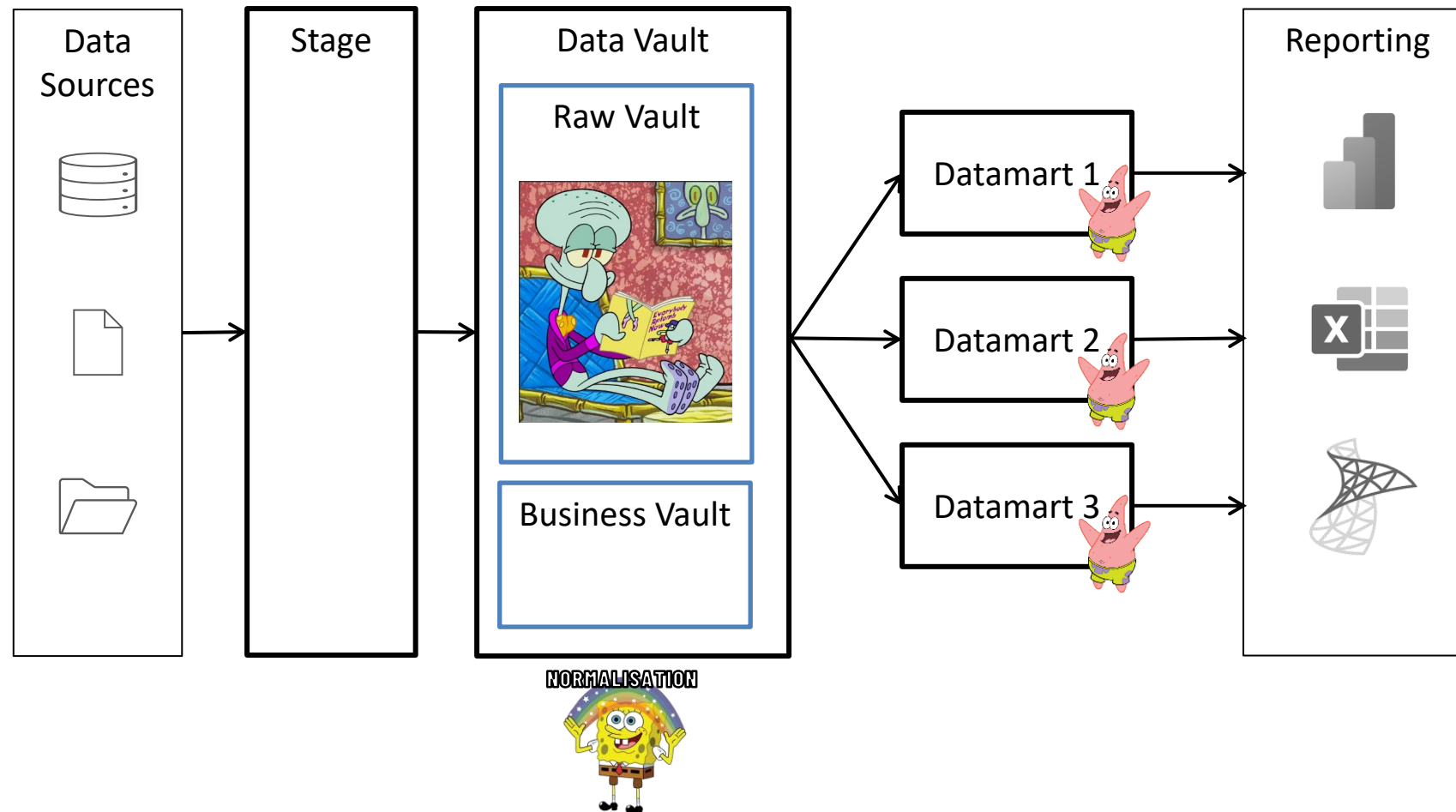
- Zaplatit větší výpočetní výkon
- Snížit množství dat v dávce
- Načítat několik dávek paralelně



Proč Data Vault?

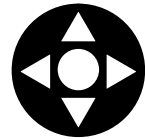
- Vysoká paralelizaci loadu
- Oddělení historizace dat od business pravidel
- Můžeme historizovat, jakmile máme metadata
- Kombinovatelný s jinými technologiemi a přístupy (agile, medalion architecture...)
- Z pohledu business pořád v mártu pořád jsou fakta a dimenze

Data Vault architecture



Raw Vault

standardizované uložení dat a historizace
bez business logiky



Hub

- Byznys klíč entity
OrderNumber,
CustomerName+CustomerBirthdayDate...



Link

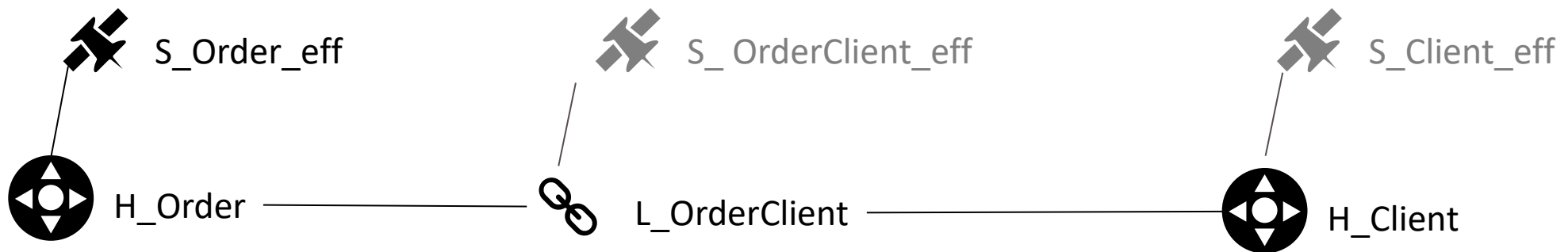
- Byznys klíč vazby mezi entity
OrderNumber+ CustomerName+CustomerBirthdayDate



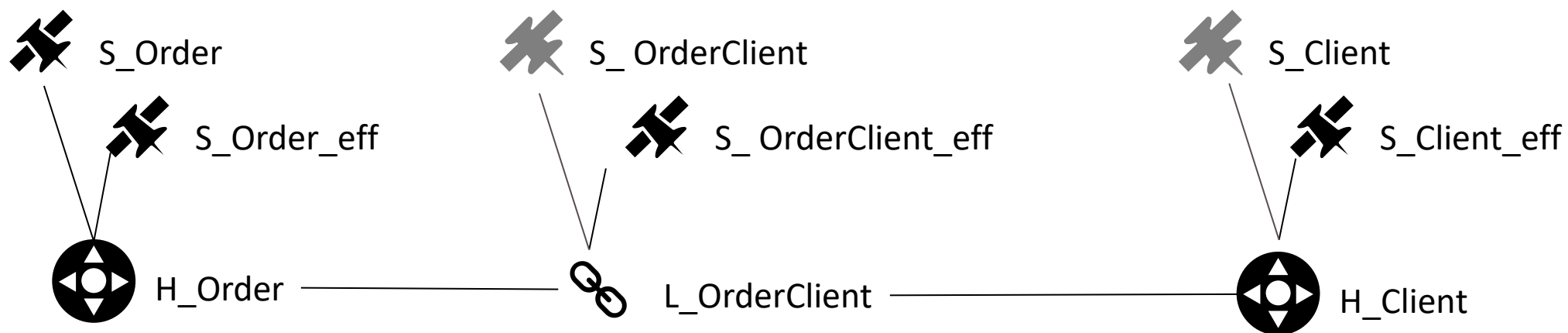
Satelite

- Hash klíč hubu/linku a atributy *co se mění v čase*
OrderStatus, ClientEmail...

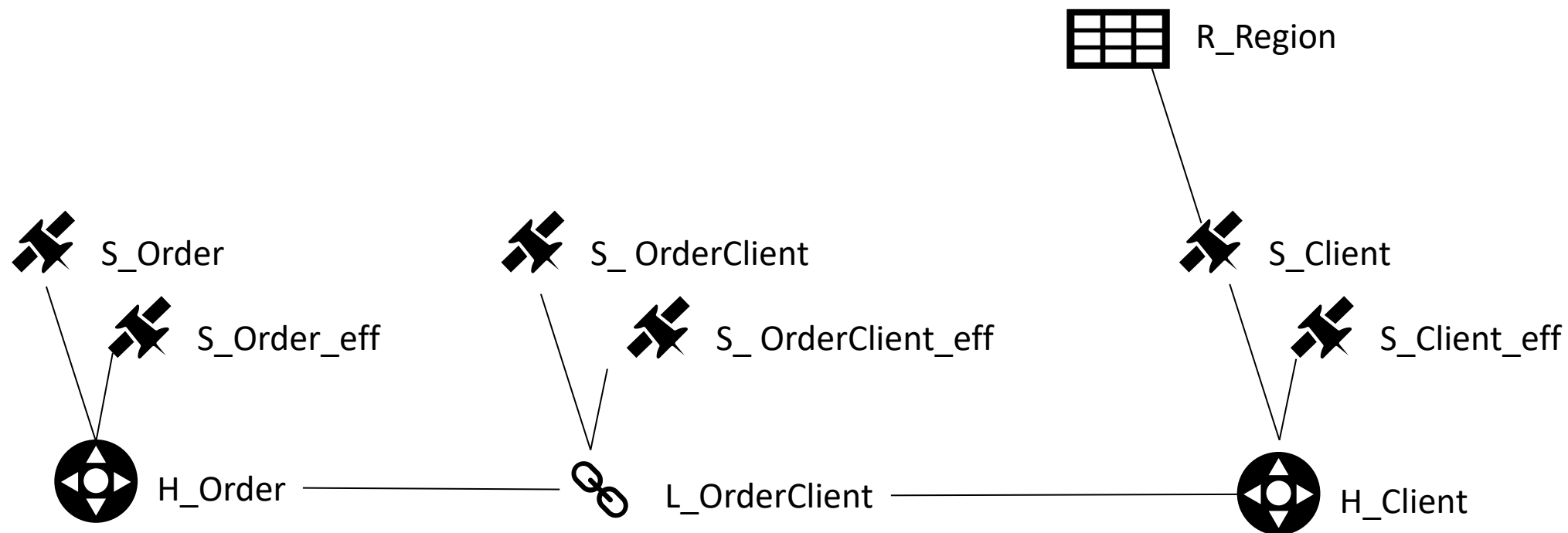
Raw Vault tabulky



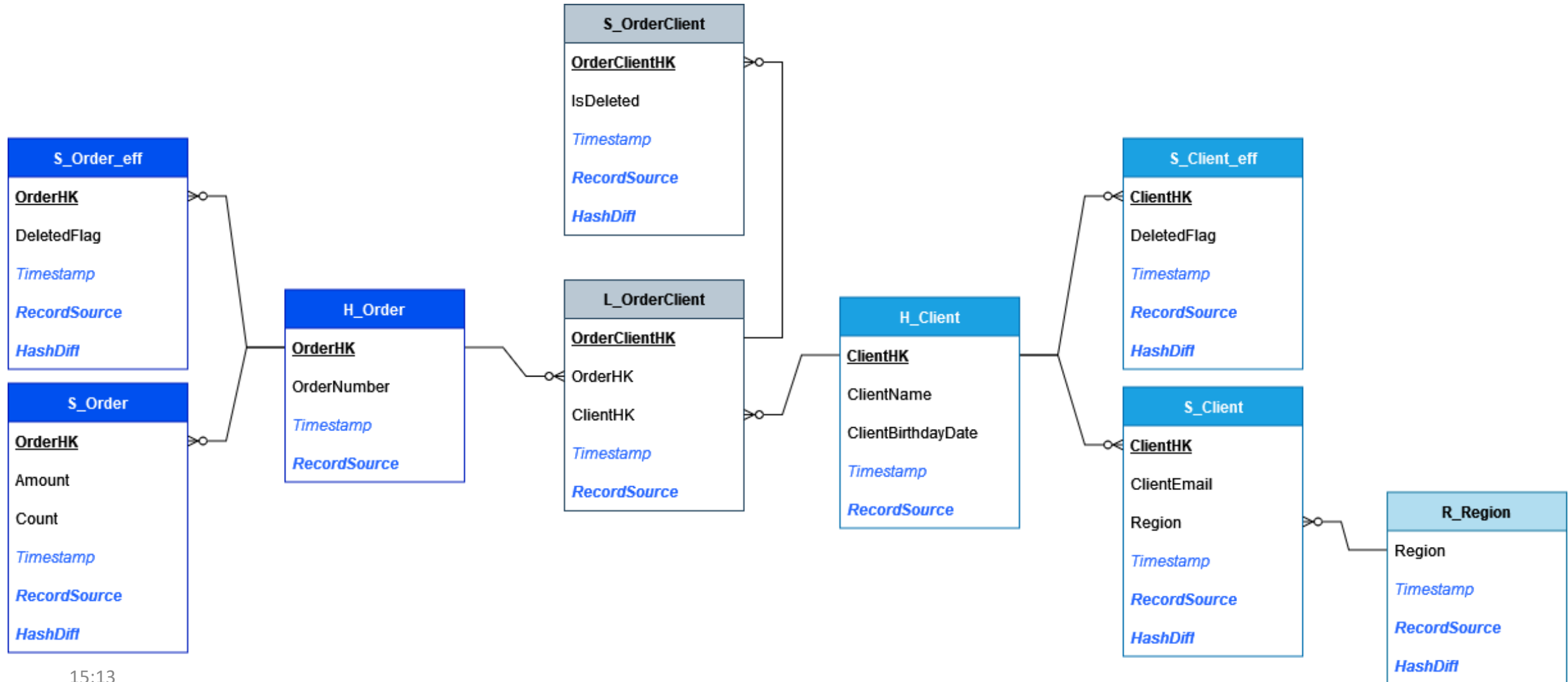
Raw Vault tabulky



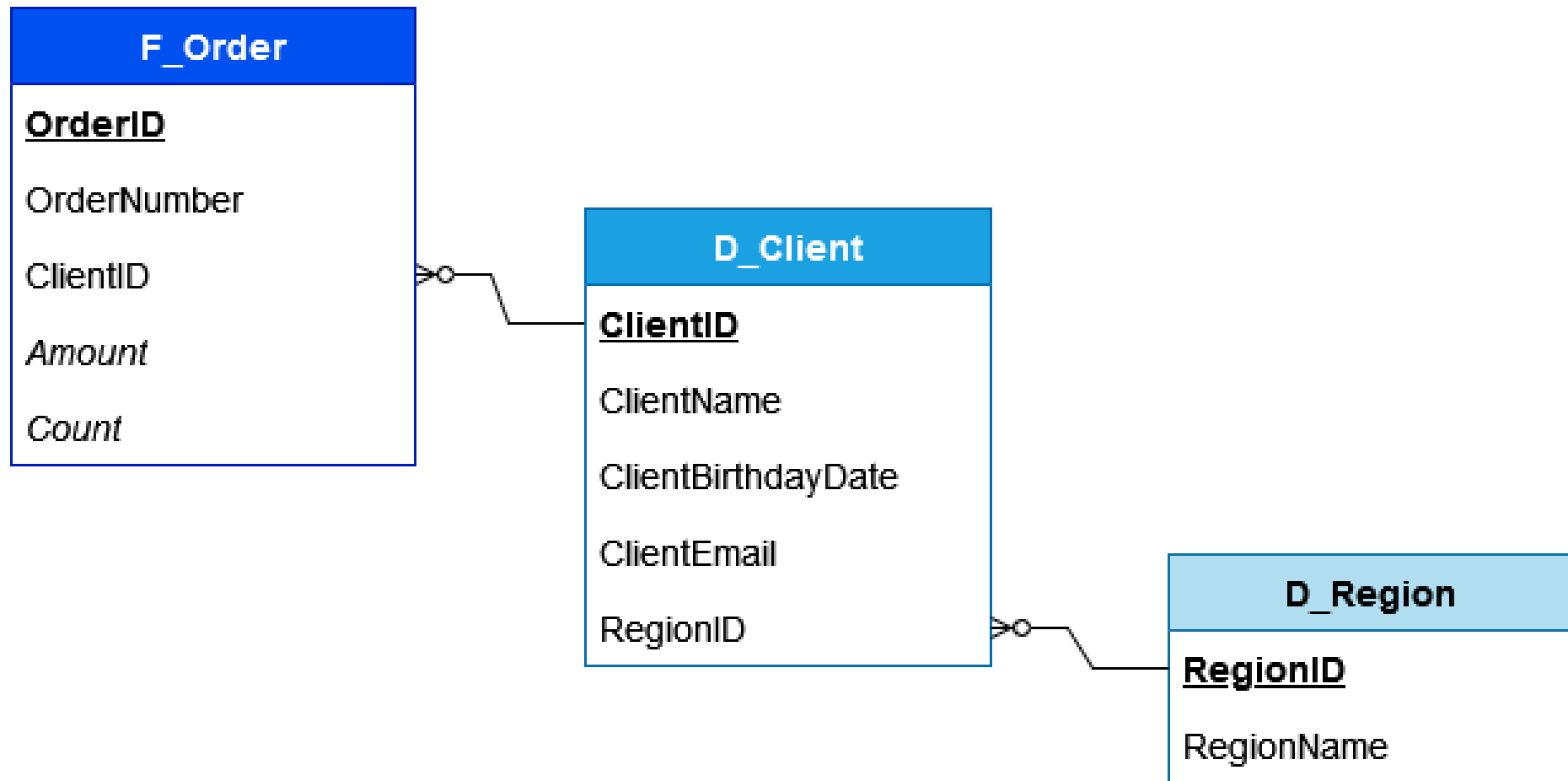
Raw Vault tabulky



Raw Vault



3. NF



Business Vault

- optimalizace joinů při náčtu mártů
- centralizované business výpočty

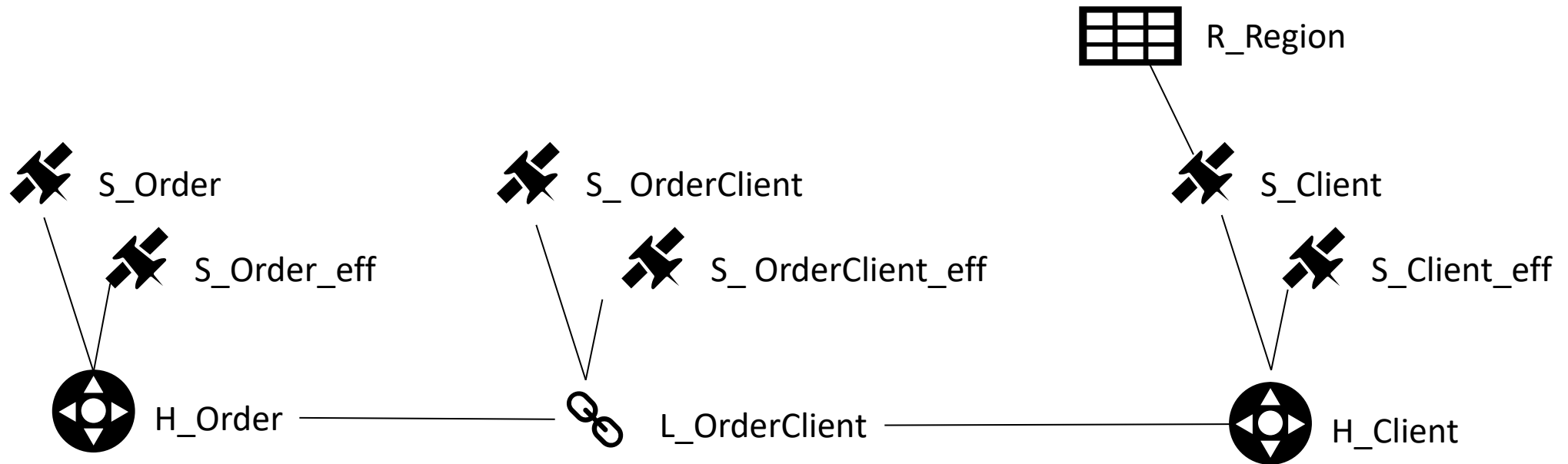
Point in Time

- Odkazy na klíče satelitů jednoho Hubu/Linku v čase

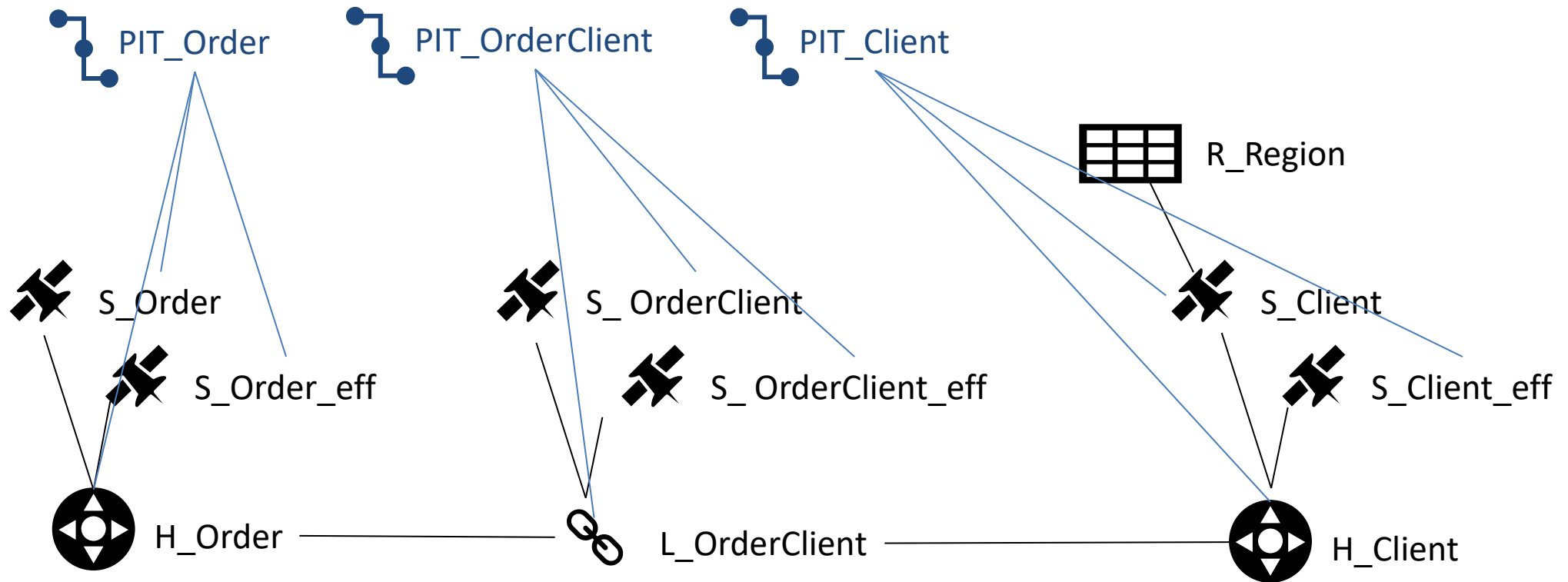
Bridge

- Odkazy na klíče satelitů přes několik Hubů/Linků

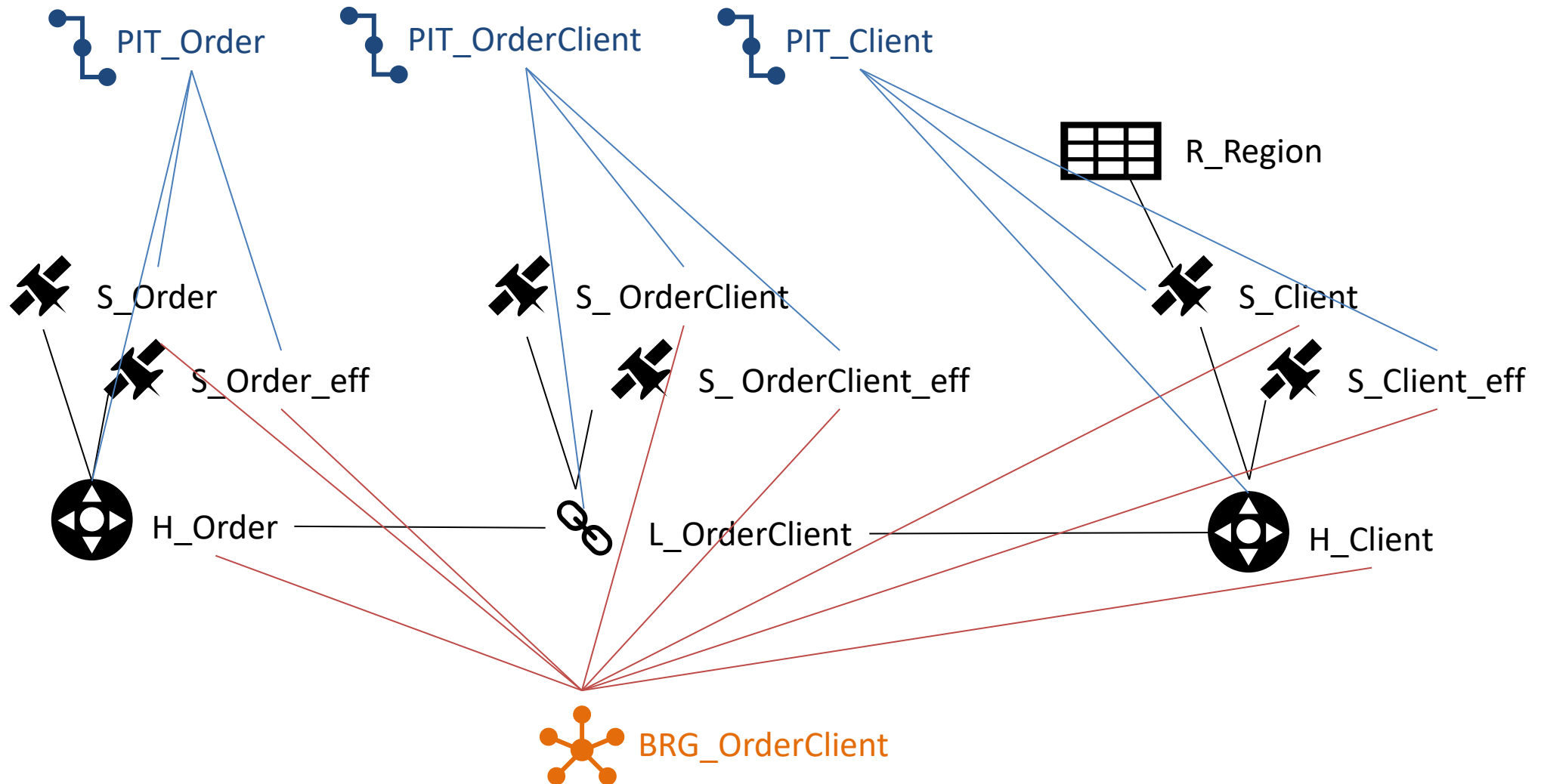
Raw Vault



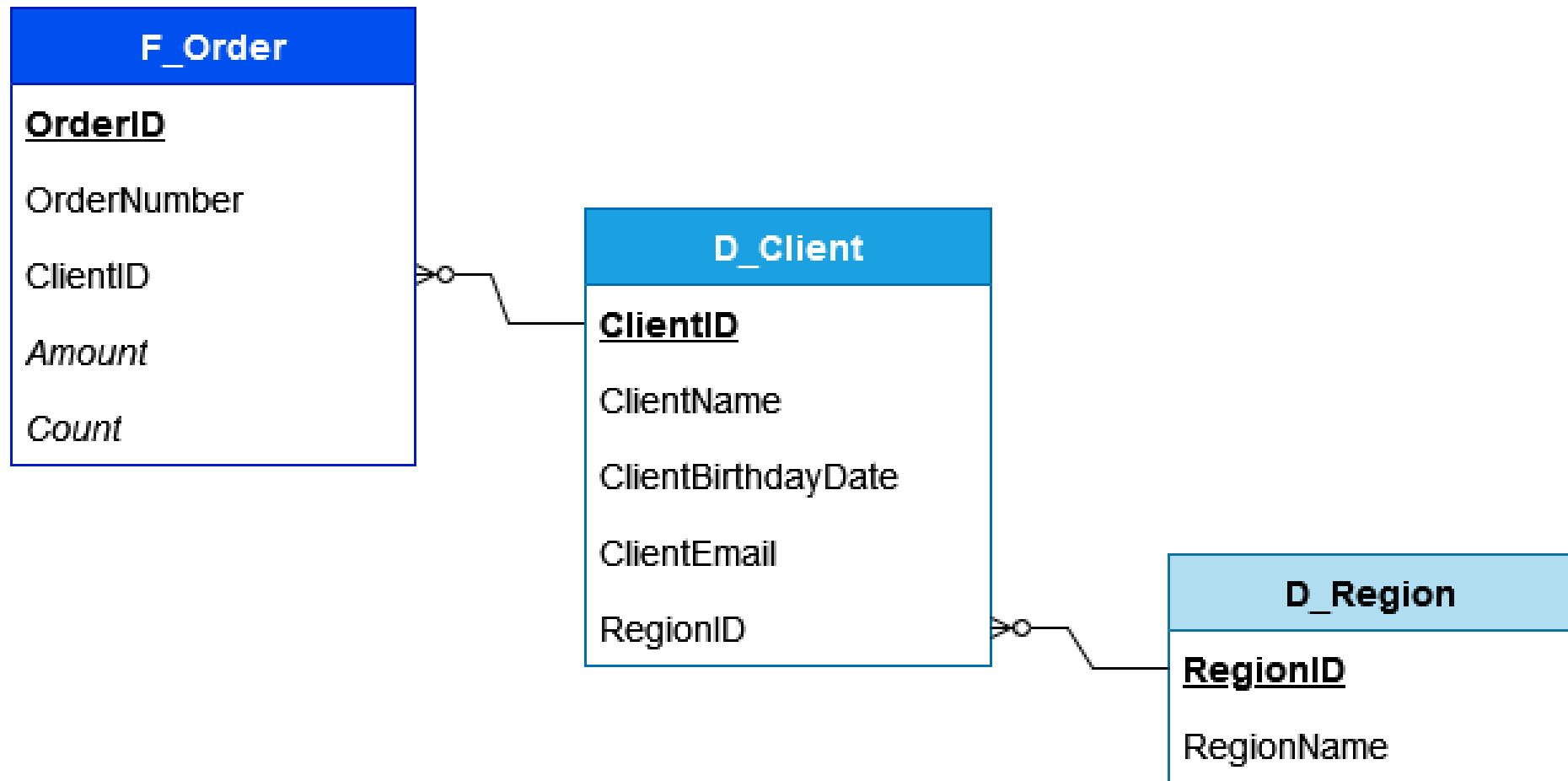
Business Vault



Business Vault



3. NF



Klíče

D_Client	
<u>ClientID</u>	
ClientName	
ClientBirthdayDate	
ClientEmail	
RegionName	

Klíče

nejsou sekvenční!

- Business klíč
 - Deterministický
 - Srozumitelný pro uživatele (rodné číslo, kombinace textových polí...)
- Hash klíč
 - Jen „formátovaný“ business klíč
 - Není 100% unikátní, ale docela ano
 - nejde o šifrování a zabezpečení dat
 - Hash funkci je potřeba zvolit sám

Doporučení pro HK

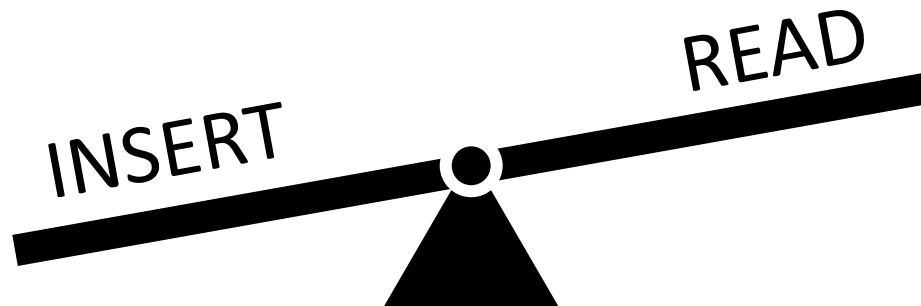
- Nahradit NULL za prázdný řetězec
- Převést všechno na string
- Převést na velká písmena
- Smazat netisknutelné znaky
- Oddělovat částí složeného klíče

```
BK = ClientName | ClientBirthdate
```

```
HK = SHA1(upper(concat(  
trim(isnull(cast(ClientName as varchar(50)), '')), '|',  
trim(isnull(cast(ClientBirthdate as varchar(50)), '')), '|',  
)))
```

Metodika

- Detailně propracovaná, ale umožňuje určitou volnost
- Stanoví povinná a doporučená metadata
- Obsahuje patterny loadu tabulek
- Doporučení pro Stage (generovat HK...)
- Doporučení pro mártý (views-marts, pokud stačí výkon...)



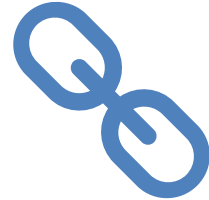
6.2 Field Naming Conventions

Attribute Type	Prefix or Suffix
Record Source	R, RSRC, RS
Sequence ID	SEQ, SEQ
Date Time Stamps	DTS, DTM
Date Stamps	DS, DT
Time Stamps	TMS, TS, TM
Load Date Time Stamps	LDT, LDDTS, LDTM
User Watch Fields	USR, U
Occurrence Numbers	OCC, OCN, ONUM
Load End Date Time Stamps	LEDTS
Sub Sequence	SSQN, SSQ, SUBSQ
Applied Dates	APD, ADT, APDT
Hash Keys	HK, HashKey, HKEY
Hash Differences	HD, HashDiff, HDIFF, HDF

6.1 Entity Naming Conventions

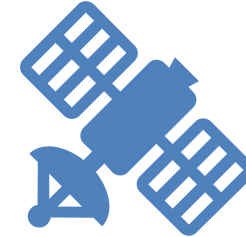
Entity Type	Prefix or Suffix
Hub	H, HUB, HB
Link	L, LINK, LNK
Satellite	S, SAT
Stage	STG
Hierarchical Links	HL, HLNK, HLINK
Same-As Links	SAL, SALNK, SLNK
Point-in-Time	PIT, PT
Bridge	B, BRDG, BRG
Business Hub	BH, BHUB
Business Link	BL, BLNK, BLINK
Business Satellite	BS, BSAT, BST
View	V
View Dimension	VDIM, VD
View Fact	VF, VFCT
Fact	FCT, FACT, F
Dimension	D, DIM
Report Collection	RPT, RC

Speciální druhy tabulek



Link

- Same-as Link
- Hierarchical Link
- Non-Historization Link
- Exploration (throw-away) Link

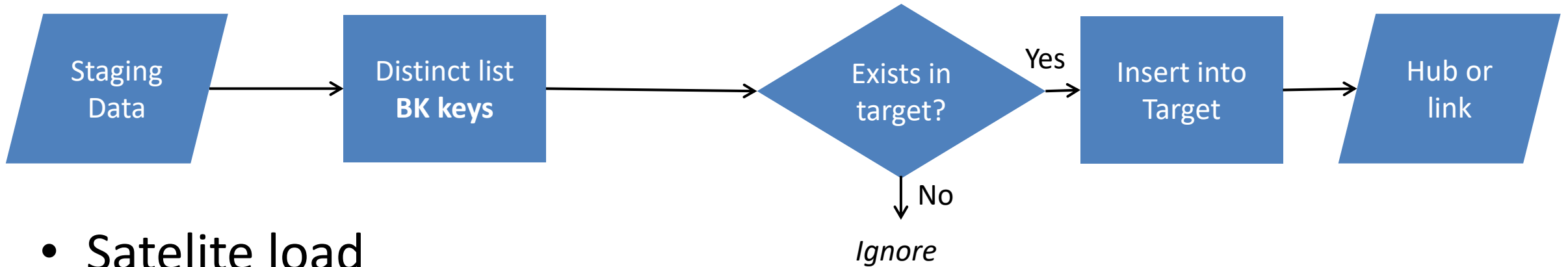


Satelite

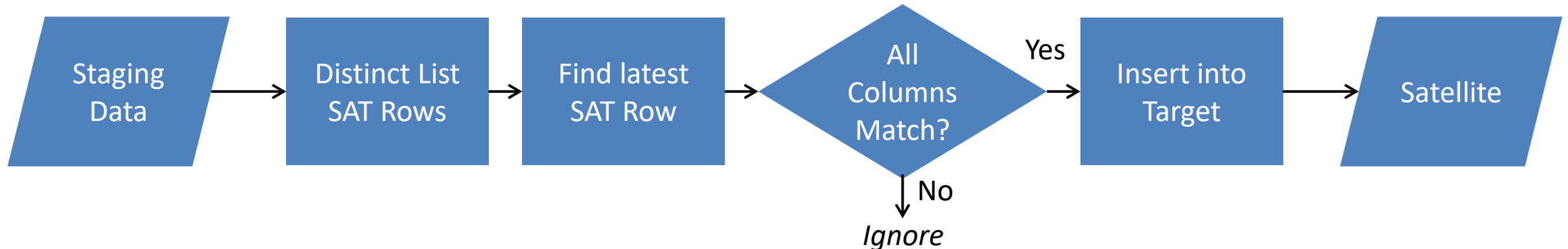
- Effectivity Satellite
- Multi-Active Satellite
- System driven Satellite
- Overloaded Satellites
- Status Tracking Satellites
- Record Tracking Satellites
- Computed Satellites

Load template

- Hub/LINK load



- Satelite load



Reálný příklad

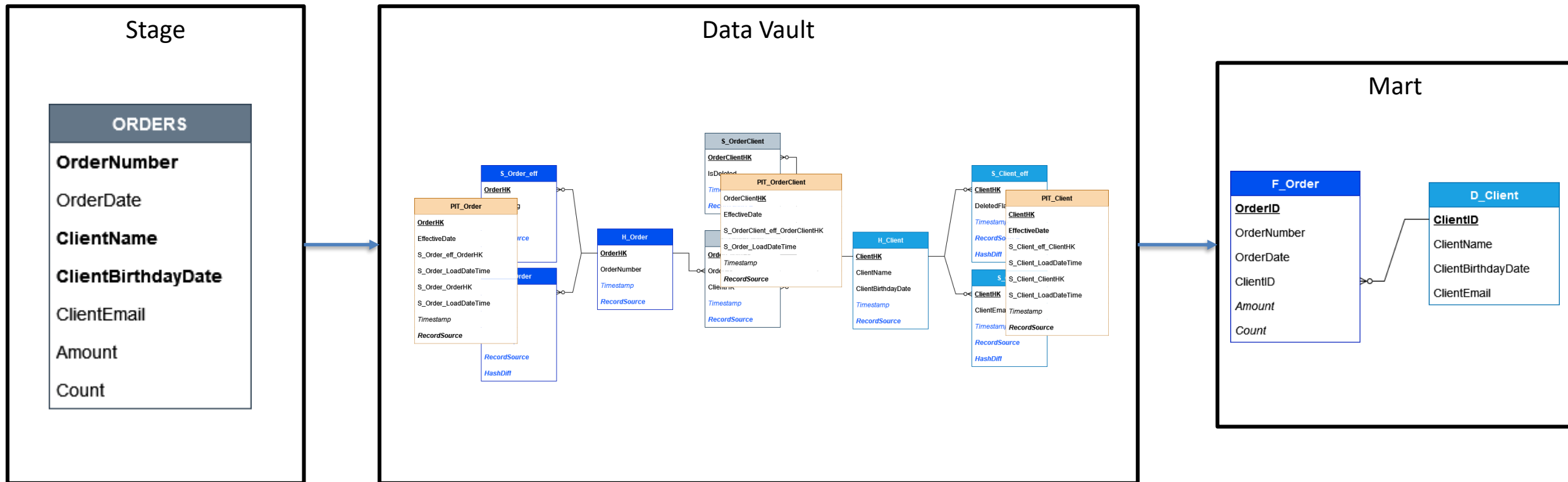
DWH v číslech

- Něco mezi DV 1.0 a 2.0
- Počet tabulek Raw Vault... **930**
 - Star schema největšího mártu... **110** (fakta 35)
- Paralelních vláken loadu... **12**
- Jedna procedura měla... **462** joiny

Raw Vault tables

SAT 460	REF 300	LNK 160	HUB 100
---------	---------	---------	---------

Příklad



Jak se neutopit v tabulkách

- **Automatizace** je nezbytná
- **Generovat z metadat** všechno co jde
- Raw DV tabulky **nejsou** pohodlné **pro lidi, ale pro počítač**
- Metodika **počítá s dočasnými** a nekompletními **řešeními**
- Dva přístupy pro provedení business výpočtů:
Information Marts a Business Vault
- **Business dokumentace** má stejné problémy jako všude

Data Vault 2.1

- **Data Lakes** (definition, discussion, where they fit in the architecture)
- **Data Fabric** (definition, discussion, where they fit in the architecture)
- **Data Mesh** (same as above), how DV Methodology aligns – as per Cindi's presentation at the Conference. Including teaming, distributed ownership, development & release
- **Terminology** shifts: Landing Zone, File Zone, Transient Zone, Enterprise Memory Zone, Information Zone (and a bunch more)
- **Operational Data Vault** (discussion, introduction, definition)
- **Metrics Vault** (discussion, introduction, definition)
- **Meta Vault** (discussion, introduction, definition)
- **Logical & Conceptual Data Vault models** (definition, discussion, split from Physical implementation)
- **Error Vault** (discussion, introduction, definition)
- Deeper dive into Taxonomies and Ontologies to support **Agile Ways of Working**
- Taxonomies and Ontologies to support **Text Integration, Image, Video Audio** with preprocessing and hooks into analytics.
- and more...

Microsoft Fabric



Loading & Staging

RAW Data



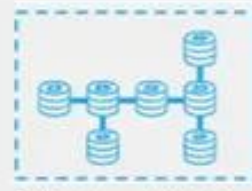
RDV (Raw Data Vault)



RAW DV only hard rules are allowed



BDV (Business Data Vault)



Soft business rules (physical or virtual)



Information Delivery



Information Marts (physical or virtual)

Dotazy

Kateřina Saryčeva, Ing

Certified: Power BI Data Analyst Associate | Power Automate RPA Developer Associate

katerina.saryceva@ness.com

 [linkedin](#)