

Data a BI v pojišťovnictví

4. 11. 2022

Martin Branda



Pro život, jaký je

4.6.3 Newtonův-Raphsonův algoritmus

Abychom mohli aplikovat Newtonův-Raphsonův algoritmus, je nutné spočítat druhé parciální derivace logaritmické věrohodnostní funkce:

$$\frac{\partial}{\partial \beta_{j'}} \left(\frac{\partial l}{\partial \beta_j} \right) = \sum_{i=1}^n \frac{\partial}{\partial \mu_i} \left(\frac{(Y_i - \mu_i) X_{ij}}{g'(\mu_i) \varphi V(\mu_i)} \right) \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_{j'}}$$

kde

$$\begin{aligned} \frac{\partial}{\partial \mu_i} \left(\frac{(Y_i - \mu_i)}{g'(\mu_i) \varphi V(\mu_i)} \right) &= \frac{-1}{g'(\mu_i) \varphi V(\mu_i)} - (Y_i - \mu_i) \frac{g''(\mu_i) V(\mu_i) + g'(\mu_i) V'(\mu_i)}{(g'(\mu_i))^2 \varphi (V(\mu_i))^2}, \\ \frac{\partial \mu_i}{\partial \eta_i} &= \frac{1}{g'(\mu_i)}, \\ \frac{\partial \eta_i}{\partial \beta_{j'}} &= X_{ij'}. \end{aligned}$$

Definujeme-li diagonální matici

$$\mathbf{V} = \text{diag} \left\{ \frac{-1}{(g'(\mu_i))^2 \varphi V(\mu_i)} - (Y_i - \mu_i) \frac{g''(\mu_i) V(\mu_i) + g'(\mu_i) V'(\mu_i)}{(g'(\mu_i))^3 \varphi (V(\mu_i))^2} \right\},$$

můžeme Hessovu matici zapsat ve tvaru

$$\mathbf{H} = \mathbf{X}^T \mathbf{V} \mathbf{X}.$$

Označíme vektor prvních parciálních derivací logaritmické věrohodnostní funkce

$$\nabla^T = \left(\frac{\partial l}{\partial \beta_1}, \dots, \frac{\partial l}{\partial \beta_m} \right).$$

Algoritmus: Počáteční odhady $\hat{\mu}_i^{(0)} = Y_i$, $\hat{\nabla}^{(0)}$ a $\hat{\mathbf{H}}^{(0)}$. Pro $k \geq 1$ opakuj následující kroky, dokud není splněno kritérium konvergence $\|\hat{\beta}^{(k)} - \hat{\beta}^{(k-1)}\| < \varepsilon$:

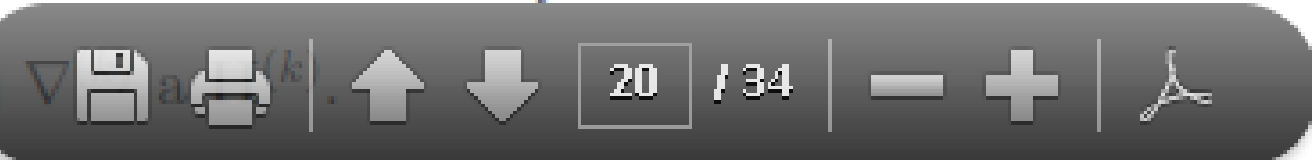
1. Spočti nový odhad parametrů

$$\hat{\beta}^{(k)} = \hat{\beta}^{(k-1)} - (\mathbf{H}^{(k-1)})^{-1} \nabla^{(k-1)}$$

2. Spočti

$$\hat{\mu}_i^{(k)} = g^{-1}(\mathbf{x}_i^T \hat{\beta}^{(k)}).$$

3. Aktualizuj



4.1.2 Gamma rozdělení

Značíme $Y \sim \Gamma(a, p)$: Pro $0 < y < \infty$ můžeme hustotu vyjádřit jako

$$\begin{aligned} f(y; a, p) &= \frac{a^p}{\Gamma(p)} y^{p-1} \exp\{-ay\} \\ &= \exp\{(p-1) \log y - ay + p \log a - \log \Gamma(p)\} \\ &= \exp\left\{ \frac{y(-a/p) + \log a/p}{1/p} \right. \\ &\quad \left. + p \log p - \log \Gamma(p) + (p-1) \log y \right\} \end{aligned}$$

kde $\theta = -a/p$, $\varphi = 1/p$, $b(\theta) = -\log(-\theta)$. Potom dostaneme

- $\mathbb{E}Y = b'(\theta) = -1/\theta = p/a = \mu$,

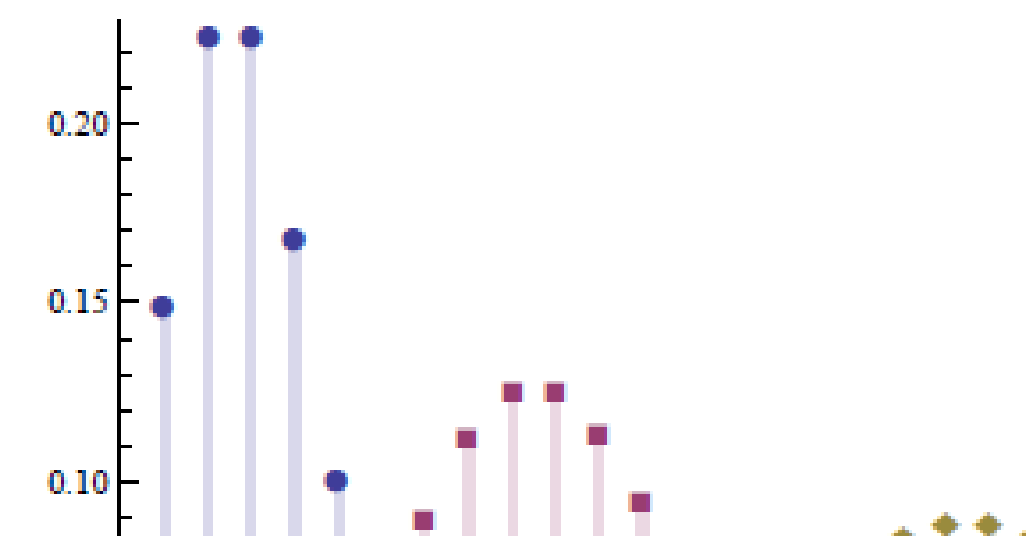
4.1.4 Poissonovo rozdělení

Značíme $Y \sim Po(\lambda)$: Pro $y = 0, 1, 2, \dots$ můžeme hustotu vyjádřit jako

$$\begin{aligned} f(y; \lambda) &= \frac{\lambda^y e^{-\lambda}}{y!} \\ &= \exp\left\{ \frac{y \log \lambda - \lambda}{1} - \log y! \right\}, \end{aligned}$$

kde $\theta = \log \lambda$, $b(\theta) = e^\theta$ a $\varphi = 1$. Potom dostaneme

- $\mathbb{E}Y = b'(\theta) = e^\theta = \lambda$,
- $\text{var}(Y) = \varphi b''(\theta) = e^\theta = \lambda$, tj. rozptyl závisí na střední hodnotě $V(\mu) = \mu$.



~~AI~~



LIDÉ
vs.
TECHNOLOGIE



LIDÉ & TECHNOLOGIE

STRATEGIE

„data-driven company“

Business Intelligence (BI)

Veškeré technologie, aplikace, znalosti a postupy
používané k pochopení fungování a řízení
obchodní společnosti
s pomocí dat.

Data denial

Organization starts with an active distrust of data and does not use it.



1

Data indifference

Company has no interest in whether data is collected or used.



2

Data aware

Business is collecting data and may use it for monitoring, but it does not base decisions on it.



3

Data informed

Managers use data selectively to aid decision making.



4

Data driven

Data plays a central role in as many decisions as possible across the organization.



5

Uspořádání ve VIG ČR

- ▶ Úsek dat a analytiky

- ▶ DWH

- ▶ Reporting a kampaně

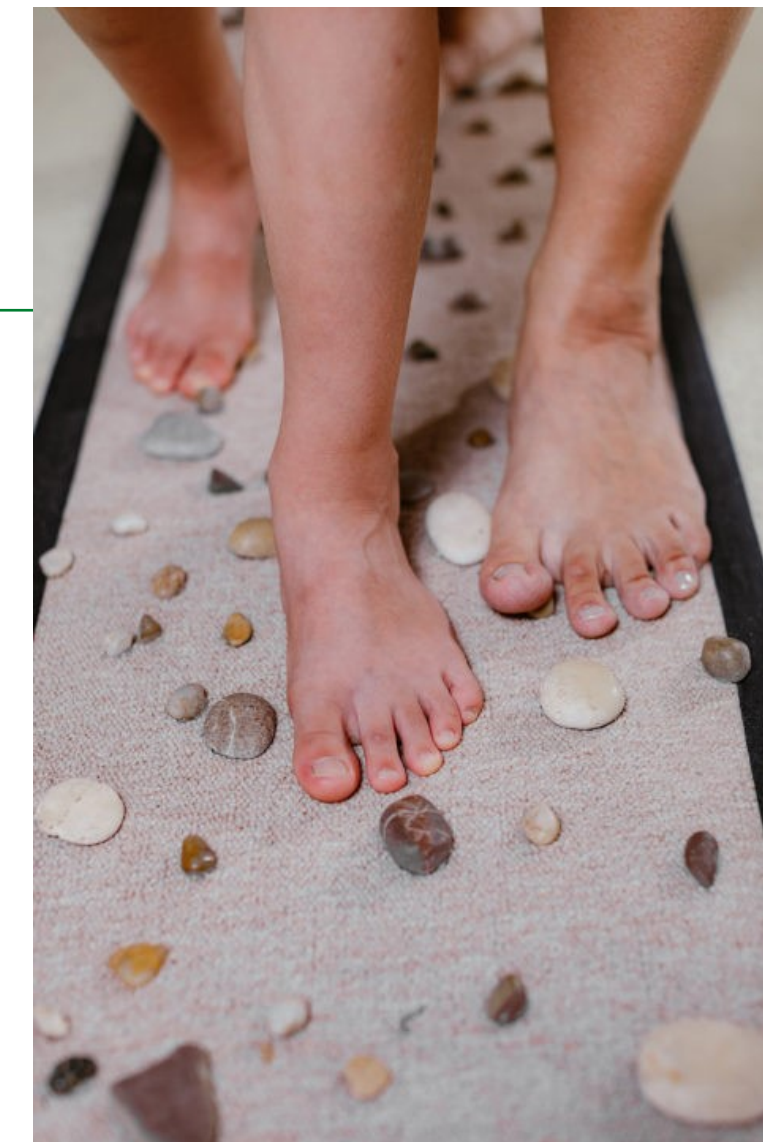
- ▶ Datová kvalita

- ▶ Pokročilá analytika

... úsek Kooperativy sdílený v rámci VIG ČR s ČPP

... vznikl 2017

Tři nohy pro rovnováhu



Business
„co-proč-jak
firma dělá“



Data/datový
model



„Techniky potřebné
pro moji práci“

Strukturovaná/nestrukturovaná data

- **Strukturovaná data** – je znám datový model, datové typy apod., např. data v DWH
- **Částečně strukturovaná data** – mají určitou danou strukturu, ale nejsou přímo analyticky zpracovatelná, např. XML, html, JSON
- **Nestrukturovaná data** – neznáme předem datový model nebo nejsou přesně organizovány, např. log ze systému

XML & JSON, Batch & Stream

```
<?xml version="1.0" encoding="UTF-8"?>
- <EmployeeData>
  - <employee id="34594">
    <firstName>Heather</firstName>
    <lastName>Banks</lastName>
    <hireDate>1/19/1998</hireDate>
    <deptCode>BB001</deptCode>
    <salary>72000</salary>
  </employee>
  - <employee id="34593">
    <firstName>Tina</firstName>
    <lastName>Young</lastName>
    <hireDate>4/1/2010</hireDate>
    <deptCode>BB001</deptCode>
    <salary>65000</salary>
  </employee>
</EmployeeData>
```

```
testJSON > testJSON > data.json > No Se
1 {
2   "users": [
3     {
4       "name": "John",
5       "age": 25
6     },
7     {
8       "name": "Mark",
9       "age": 29
10    },
11    {
12      "name": "Sarah",
13      "age": 22
14    }
15  ],
16  "dataTitle": "JSON Tutorial!",
17  "swiftVersion": 2.1
18 }
```

Batch / stream processing – zpracování dávky / záznamu

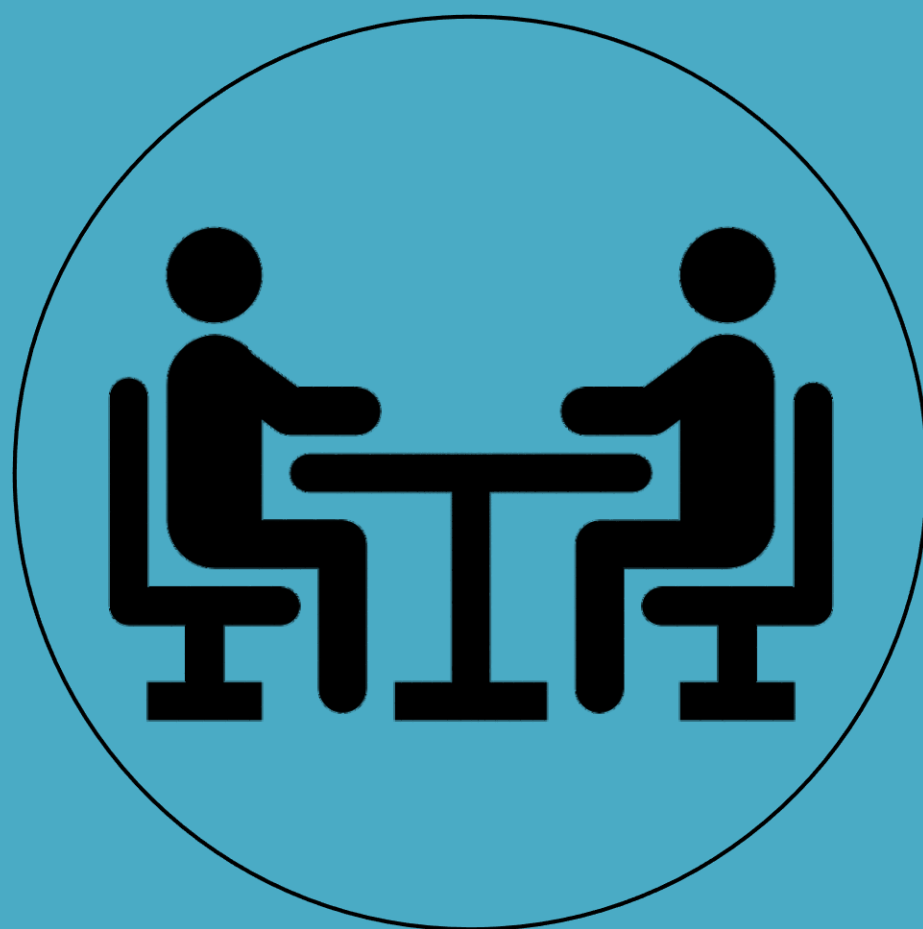
Příklad - úkol



ÚKOL: Vypočítej a porovnej *škodní frekvenci* na smlouvách povinného ručení za retail, flotily a leasing v letech 2020 a 2021.

- VYBERTE SPRÁVNÉ ŘEŠENÍ:
 1. Připojím se k DWH, najdu tabulku pojištění vozidel, nahodím editor SQL, spočítám si nějak expozice smluv, připojím si pojistné události, všechno sečtu a podělím, následně překopíruji do excelu, hezky obarvím a pošlu Outlookem šéfovi. *
 2. V katalogu reportů najdu dashboard, který škodní frekvence obsahuje, ověřím si definici v datovém slovníku, na šéfa pošlu pouze link na reporting server.

* Zesvodněné škody, co regres, co nulové škody, jen uzavřené, rozhodné ... ?



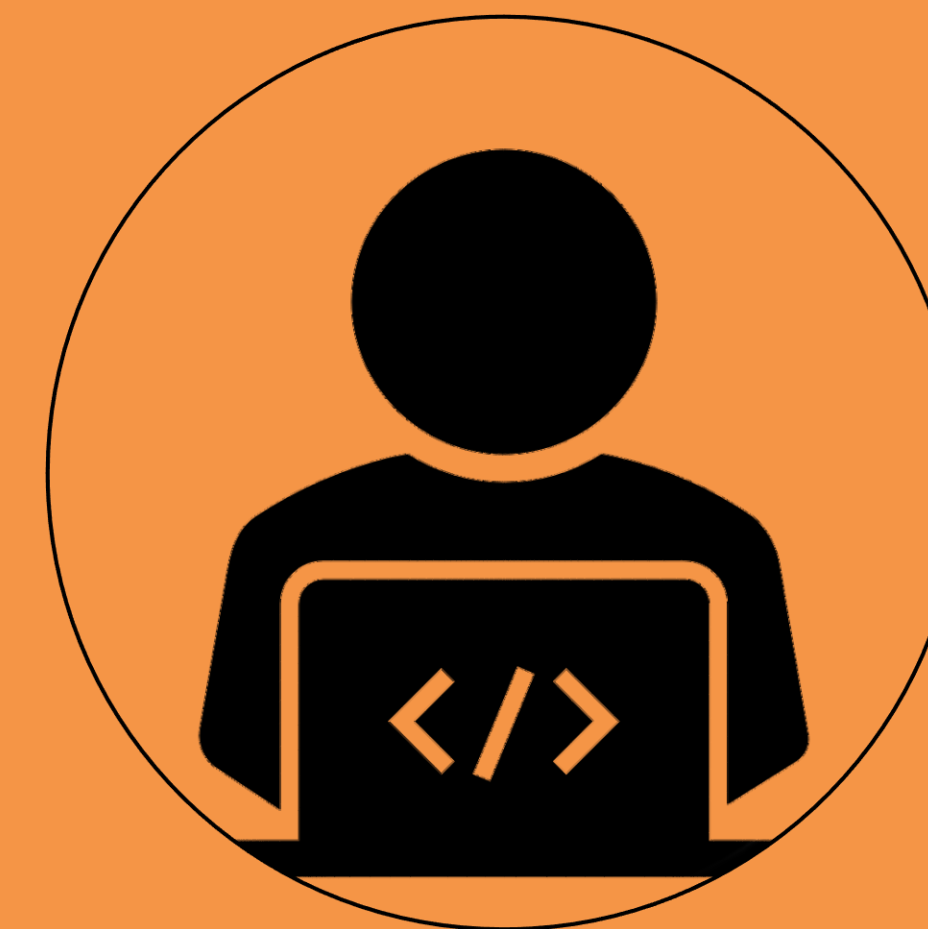
Zdroje dat

- ▶ Obchod
- ▶ Provoz
- ▶ Účetnictví
- ▶ Likvidace
- ▶ Kontaktní centrum
- ▶ ...



BI

- ▶ DWH
- ▶ Datalake
- ▶ Reporting
- ▶ ...



Uživatelé dat

- ▶ Analytika
- ▶ Fraud
- ▶ Campaign management
- ▶ Pojistná matematika
- ▶ ...





Back office

- ▶ IT
- ▶ DWH/BI
- ▶ Kontroling
- ▶ Ekonomické
- ▶ Pojistná matematika



Middle office

- ▶ Produkty (ŽP, MAJ, POV+HAV)



Front office

- ▶ Obchod
- ▶ Likvidace
- ▶ Kontaktní centrum



Data Warehouse (DWH, datový sklad)

DWH

- Je centralizované úložiště pro strukturovaná i nestrukturovaná data
- Spojuje velké množství dat z různých zdrojů
- Může obsahovat nejen aktuální data, ale i historizovaná
- Umožňuje analytické dotazování nad rozsáhlými soubory dat

Data Warehouse (DWH, datový sklad)

Vrstvy DWH

- L0 (stage): kopie dat ze zdrojových systémů
- L1 (target): rozpad, transformace, historizace dat
- L2 (datamarty): agregace dat z L1 pro analytické zpracování
- L3 (OLAP): mnohorozměrné uspořádání velkého množství dat pro analytiku
- Sandbox („pískoviště“): oddělený prostor pro „hraní“ s daty a analytiku

Extract – Transform – Load (ETL)

- 1. E:** Extrakce dat z jednoho či více zdrojů dat zahrnující relační databáze, strukturovaná i nestrukturovaná data, interní či externí zdroje dat.
- 2. T:** Aplikace různých pravidel na data před jejich načtením zahrnující např.:
 - Čištění
 - Výběr sloupců
 - Sjednocení hodnot se stejným významem
 - Výpočet nových hodnot (odvození data narození a pohlaví z rodného čísla)
 - Agregace
- 3. L:** Nahrání dat do cílového systému, např. DWH. Obvykle automatizované načtení. Může jít o přehrání dat nebo přidání dat k již existujícím.

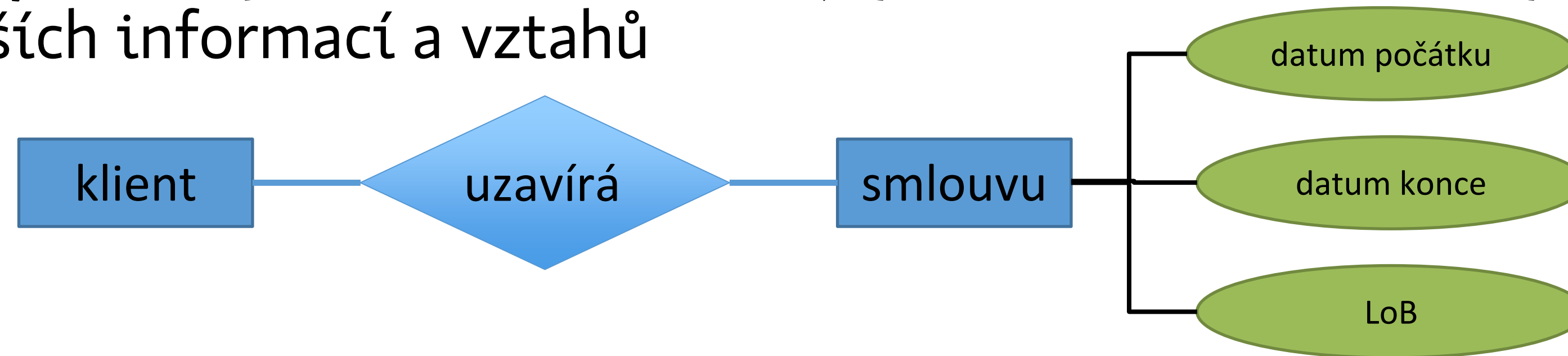
Extract – Load – Transform (ELT)

... data jsou nejprve stažena do cílového úložiště a až poté probíhají transformace.

Datový model/datové modelování

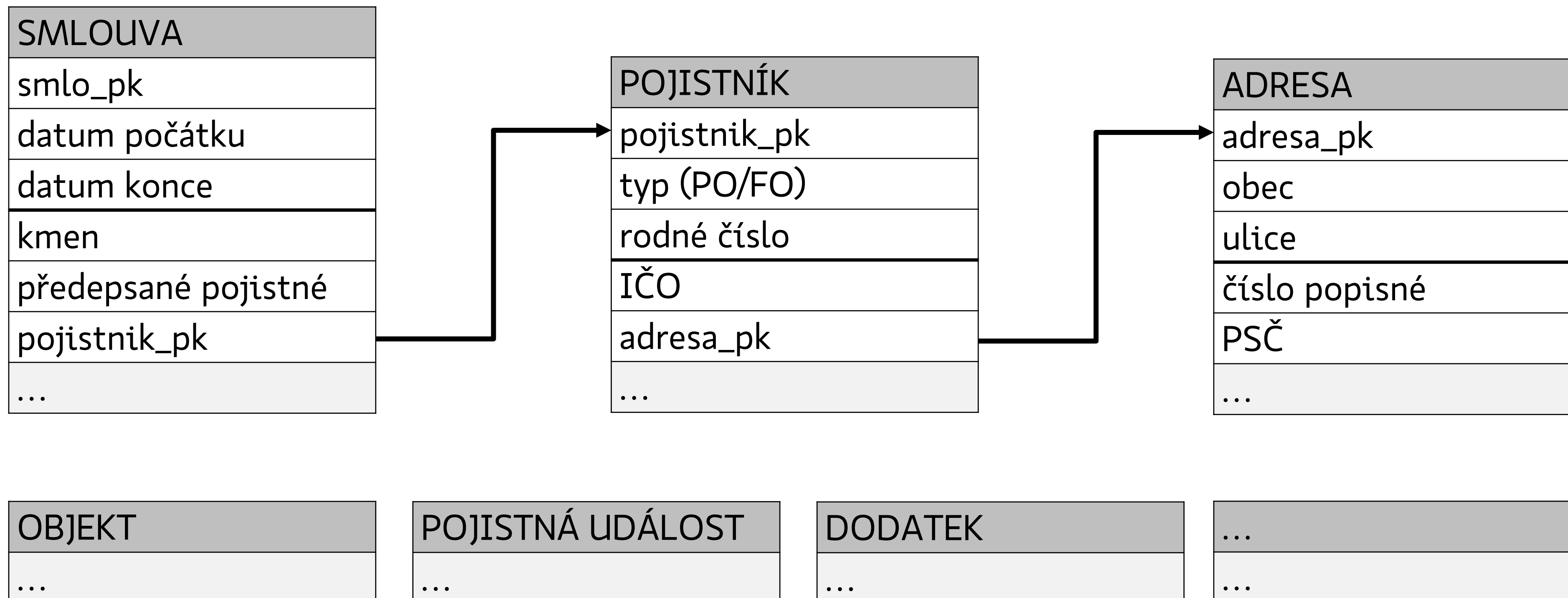
Definice a popis formátu a struktury dat v informačních systémech, určení vzájemných vztahů jednotlivých datových elementů.

- Tři úrovně:
 - **Model reality** (konceptuální úroveň) – nejvyšší míra abstrakce – vymezení nejdůležitějších informací a vztahů



- **Technologický model** (technologická úroveň) – střední míra abstrakce – způsob uchování dat, např. relační vs. objektová databáze
- **Implementační model** (fyzická úroveň) – nejnižší míra abstrakce – implementace v konkrétním databázovém systému

Relační databáze - příklad



SQL (Structured Query Language)

- Standardizovaný strukturovaný dotazovací jazyk pro práci s relačními databázemi
- Příkazy pro
 - Definici dat (CREATE, ALTER, DROP, ...)
 - Řízení dat (GRANT, REVOKE, COMMIT, ...)
 - Manipulaci s daty (SELECT, INSERT, GROUP BY, ...)
 - Další příkazy

SQL - příklad

```
CREATE TABLE work.smlouva_klient
AS SELECT a.smlo_pk
        , a.kmen
        , b.pojistnik_pk
        , b.typ
        , b.rodne_cislo
        , b.ICO
FROM L2.smlouva a
LEFT JOIN L2.pojistnik b
ON a.pojistnik_pk = b.pojistnik_pk
```


SQL - příklad

```
SELECT a.smlo_pk  
      , a.kmen  
      , b.pojistnik_pk  
      , b.typ  
      , b.rodne_číslo  
      , b.IČO  
INTO #smlouva_klient  
FROM L2.smlouva a  
LEFT JOIN L2.pojistník b  
ON a.pojistnik_pk = b.pojistnik_pk
```

SQL - příklad

```
SELECT a.smlo_pk
      , a.kmen
      , b.typ
      , b.rodné_číslo
      , concatenate(c.ulice, ' ', c.číslo_popisné, ' ', c.obec)
INTO #smlouva_klient_adresa
FROM L2.smlouva a
LEFT JOIN L2.pojistník b
ON a.pojistnik_pk = b.pojistnik_pk
LEFT JOIN L2.adresa c
ON b.adresa_pk = c.adresa_pk
WHERE a.kmen = 'ZIV'
```



On Premise / Cloud / Hybridní architektura

- On Premise („on-prem“) - software je nainstalován lokálně na vašich serverech nebo počítačích. Licenční poplatek většinou bývá jednorázový, ale musíte se starat o správu a údržbu HW.
- Cloud - software je hostován centrálně a licence jsou na subscription bázi. Cloud aplikace jsou většinou přístupné prostřednictvím lehkého klienta.
- ?! dostupnost / náklady / bezpečnost / údržba



Dostupnost serveru/cloudu

- SLA (Service-level agreement) – „smlouva o úrovni služeb“ = přesnější definice produktu, který je zákazníkovi poskytován, např. v IT
- Maximální výpadek služeb (z 365 dní)
 - 98% dostupnost – výpadek nepřesáhne 7 dní, 7 hodin a 19 minut
 - 99% dostupnost – výpadek nepřesáhne 3 dny 15 hodin a 40 minut
 - 99,5% dostupnost – výpadek nepřesáhne 1 den 19 hodin a 50 minut
 - 99,9% dostupnost – výpadek nepřesáhne 8 hodin a 46 minut
 - 99,99% dostupnost – výpadek nepřesáhne 52 minut a 35 sekund



Příklad: Parametrizace škod pojistným matematikem VS. uzavření smlouvy obchodníkem. Kdo potřebuje vyšší dostupnost?

BI nástroje pro reporting

Příprava reportů a dashboardů



Source: Gartner (February 2019)

BI nástroje pro reporting

Portfolio Dashboard

Project Progress: Todos
Project Manager: Todos

Projects: 21

Effort: 18 Mil Hours

Effort Completed: 6.648 Hours

Effort Remaining: 11 Mil Hours

Projects by Progress

Effort by Project

Projects by Project Manager

Project	Link	Project Manager	Start	Finish	Progress	Effort (Hours)	Task Count	Overdue Tasks	Late Tasks	On Track Tasks	Future Tasks	Completed Tasks
Annual employee update meeting	Link	Kasey Banks	17/06/2019	08/07/2019	100%	128	3					3
Heat awareness poster program	Link	Marco Christmas	05/08/2019	24/09/2019	100%	176	8					8
Email Campaign - Wave 1	Link	Elva Hebert	18/09/2019	22/10/2019	100%	208	25					25
Support mobile account inquiry	Link	Elva Hebert	25/10/2019	25/10/2019	75%	0	3	2				1
Rider Survey	Link	Elva Hebert	02/09/2019	31/10/2019	10%	1.217	25	17	2		5	1
Email campaign to increase rider's awaren...	Link	Elva Hebert	07/10/2019	04/11/2019	29%	764	22	9		3	5	5
E-discovery	Link	Nick Trudeau	14/10/2019	05/11/2019	38%	392	51	14		3	21	13
Marketing plan FY2020	Link	Elva Hebert	21/10/2019	14/11/2019	36%	224	7	1		1	4	1
Track upgrades (miles 3 thru 6)	Link	Elva Hebert	29/08/2019	25/11/2019	93%	2.608	7				2	5
Email campaign to increase rider's awaren...	Link	Elva Hebert	21/10/2019	25/11/2019	34%	536	21			2	16	3
Station Design	Link	Elva Hebert	30/09/2019	28/11/2019	58%	680	4			1	2	1
Total						17.613	280	45	2	21	132	80

Visualizações

Filtros

Campos

Pesquisar

- Bookable Resources
- Project Bucket
- Project Tasks
- Projects
- Resource Assignme...

Pesquisar

Legenda: D...

Cores dos dados

Rótulos: At...

Formas

Titulo: At...

Texto do título: Projects by Progress

Quebra automática de lin...: Desativado

Cor da fonte: [dropdown]

Cor da tela de fundo: [dropdown]

Alinhamento: [options]

Tamanho do texto: 12 pt

Família de fontes: [dropdown]

PORTFOLIO DASHBOARD
PORTFOLIO TIMELINE
PORTFOLIO MILESTONES
ROADMAP KEY DATES
ROADMAP DETAILS
RESOURCE DASHBOARD
RESOURCE ASSIGNMENTS
TASK OVERVIEW
PROJECT TIMELIN

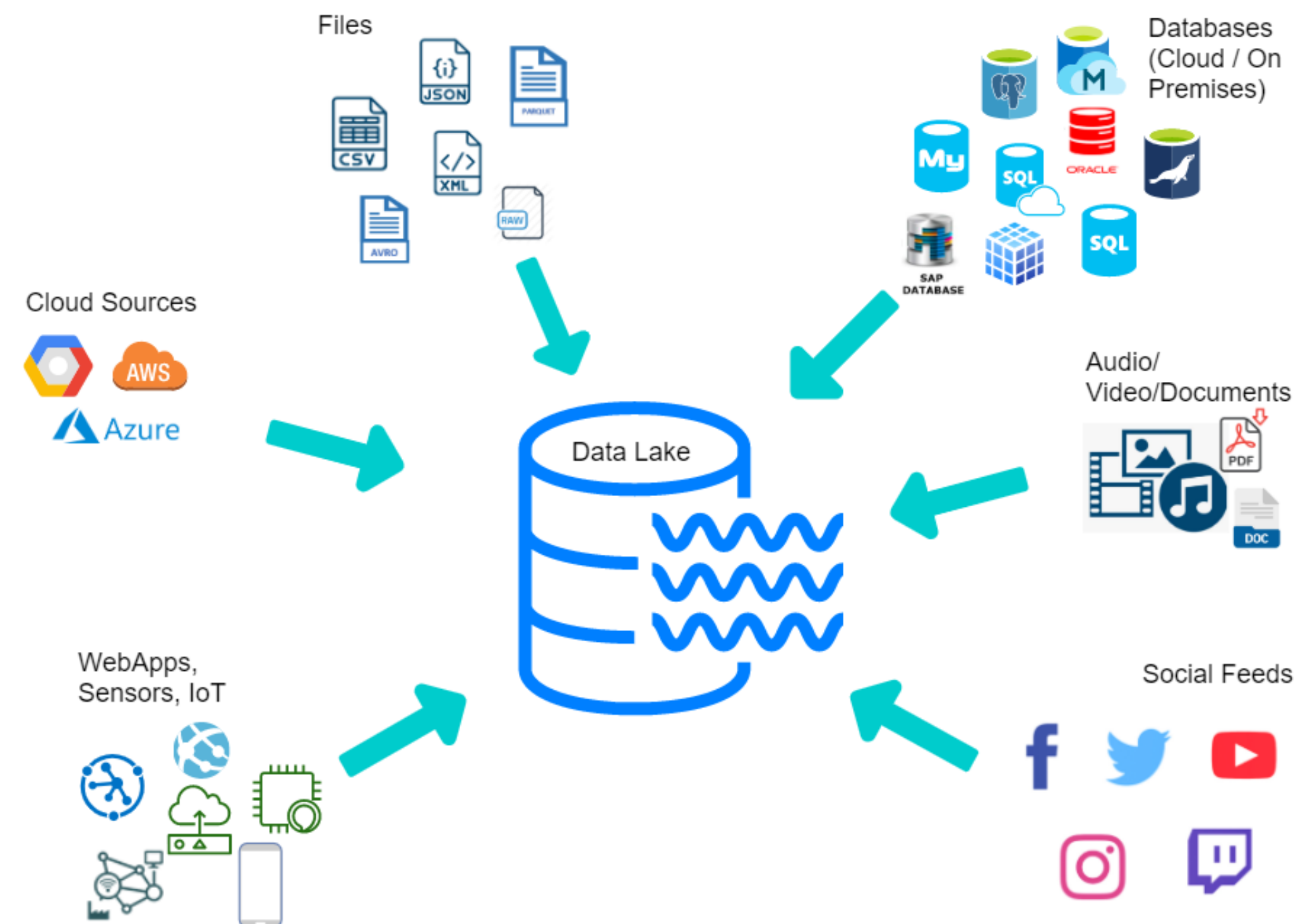
Datalake



- Uložení, zpracování a analýza (nejen velkých) dat
- Data uložena v jejich přirozeném „hrubém“ formátu, například soubory z interních i externích souborů, data ze sociálních sítí, streamovaná data ze senzorů, obrázky, audio, video, ...

= „vše, co nepatří do DWH“

- Nesmí být datové smetiště
- Relační databáze i NoSQL úložiště
- On-prem nebo v cloudu



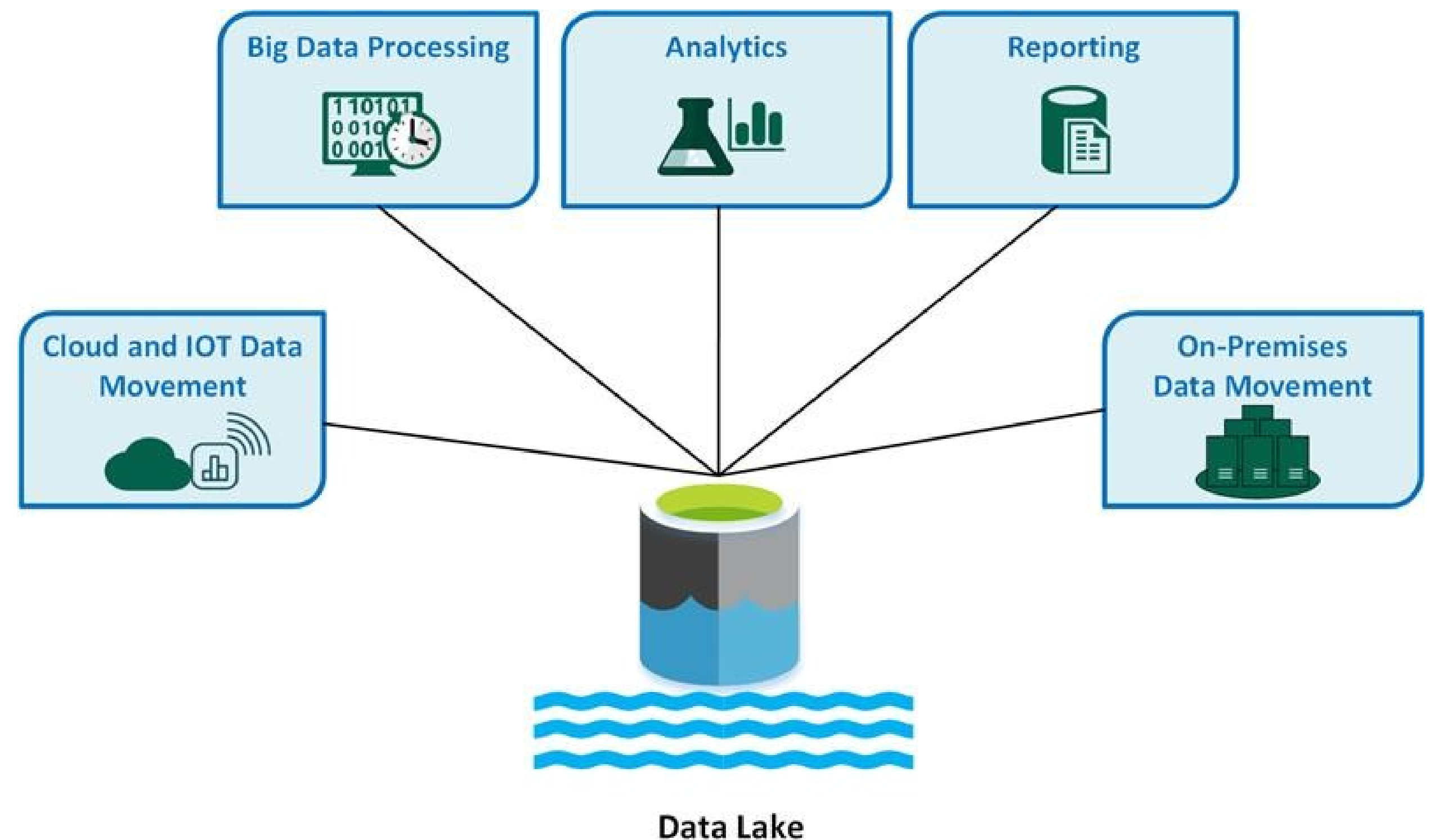
Datalake



- Uložení, zpracování a analýza (nejen velkých) dat
- Data uložena v jejich přirozeném „hrubém“ formátu, například soubory z interních i externích souborů, data ze sociálních sítí, streamovaná data ze senzorů, text, obrázky, audio, video, ...

... „vše, co nepatří do DWH“

- Nesmí být datové smetiště
- Relační databáze i NoSQL úložiště
- On-prem nebo v cloudu



Datová kvalita

- Dostupnost
- Přesnost
- Úplnost
- Konzistence

Komunita uživatelů dat

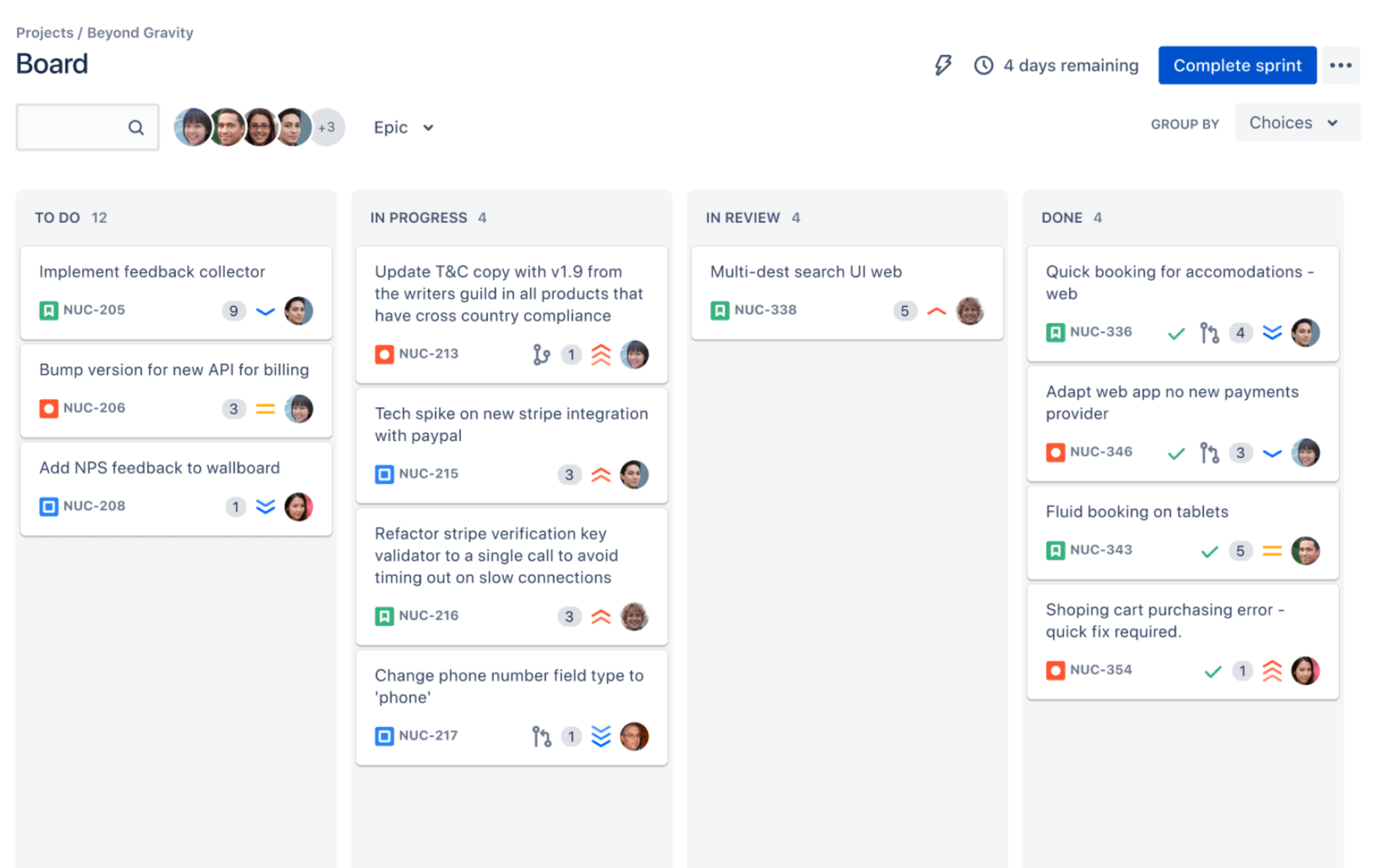
- Útvary pracující s daty mají své zástupce
- Ti se pravidelně setkávají a sdílejí vzájemně informace
- Následně jsou tyto informace sdíleny v rámci útvarů



Co se stane, když něco z výše uvedeného nefunguje?

Tiketový systém

Všechny požadavky na změny, úpravy, opravy chyb se zadávají jako tikety.



Projects / Beyond Gravity
Board

⚡ ⌚ 4 days remaining Complete sprint ...

SEARCH [] [] [] [] +3 Epic ▾ GROUP BY Choices ▾

Column	Count	Item	ID	Priority	Assignee
TO DO	12	Implement feedback collector	NUC-205	9	Low
		Bump version for new API for billing	NUC-206	3	Medium
		Add NPS feedback to wallboard	NUC-208	1	Low
IN PROGRESS	4	Update T&C copy with v1.9 from the writers guild in all products that have cross country compliance	NUC-213	1	High
		Tech spike on new stripe integration with paypal	NUC-215	3	Medium
		Refactor stripe verification key validator to a single call to avoid timing out on slow connections	NUC-216	3	Medium
		Change phone number field type to 'phone'	NUC-217	1	Low
IN REVIEW	4	Multi-dest search UI web	NUC-338	5	High
DONE	4	Quick booking for accomodations - web	NUC-336	4	Low
		Adapt web app no new payments provider	NUC-346	3	Medium
		Fluid booking on tablets	NUC-343	5	Medium
		Shoping cart purchasing error - quick fix required.	NUC-354	1	High

Tiketový systém

ONOS Dashboards ▾ Projects ▾ Issues ▾ Agile ▾ **Create** Search 🔍 ?


ONOS / ONOS-216
How to: submit Jira tickets


Edit Comment Assign More ▾ Start Progress Resolve Issue Close Issue Export


Details


Type:	📖 Story	Status:	OPEN (View Workflow)
Priority:	↑ Major	Resolution:	Unresolved
Affects Version/s:	None	Fix Version/s:	None
Labels:	None		
Epic Link:	Docs		
Sprint:	Sprint 1		

People

Assignee:  **Bill Snow**
[Assign to me](#)

Reporter:  Ayaka Koshibe

Votes:  0

Watchers:  [Stop watching this issue](#)

Dates

Created: 6 days ago
Updated: 2 minutes ago

Agile

Active Sprint: [Sprint 1](#) ends 21/Nov/14
[View on Board](#)

Description

How to report bugs or to request features

Activity

All **Comments** Work Log History Activity

There are no comments yet on this issue.

Další ne-nedůležitá témata



- Uživatelské přístupy a „citlivá“ data
- Číselníky a jejich správa
- Chování uživatelů na DWH
- Správa kódů (GIT)
- ...

Nakonec dopadneme takto ... :)

