Data a BI v pojišťovnictví

4. 11. 2022

Martin Branda

Kooperativa
VIENNA INSURANCE GROUP

Pro život, jaký je

4.6.3 Newtonův-Raphsonův algoritmus

Abychom mohli aplikovat Newtonův-Raphsonův algoritmus, je nutné spočíst **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

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

můžeme **Hessovu matici** zapsat ve tvaru

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

Označíme vektor prvních parciálních derivací logaritmické věrohodnostní funkce

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

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 $\left\|\hat{\beta}^{(k)}-\hat{\beta}^{(k-1)}\right\|<\varepsilon$:

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 $\nabla^{(k)}$ a $\hat{\mathbf{H}}^{(k)}$.

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 $y=0,1,2,\ldots$ 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

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


AI

Created by MB in wordart.com

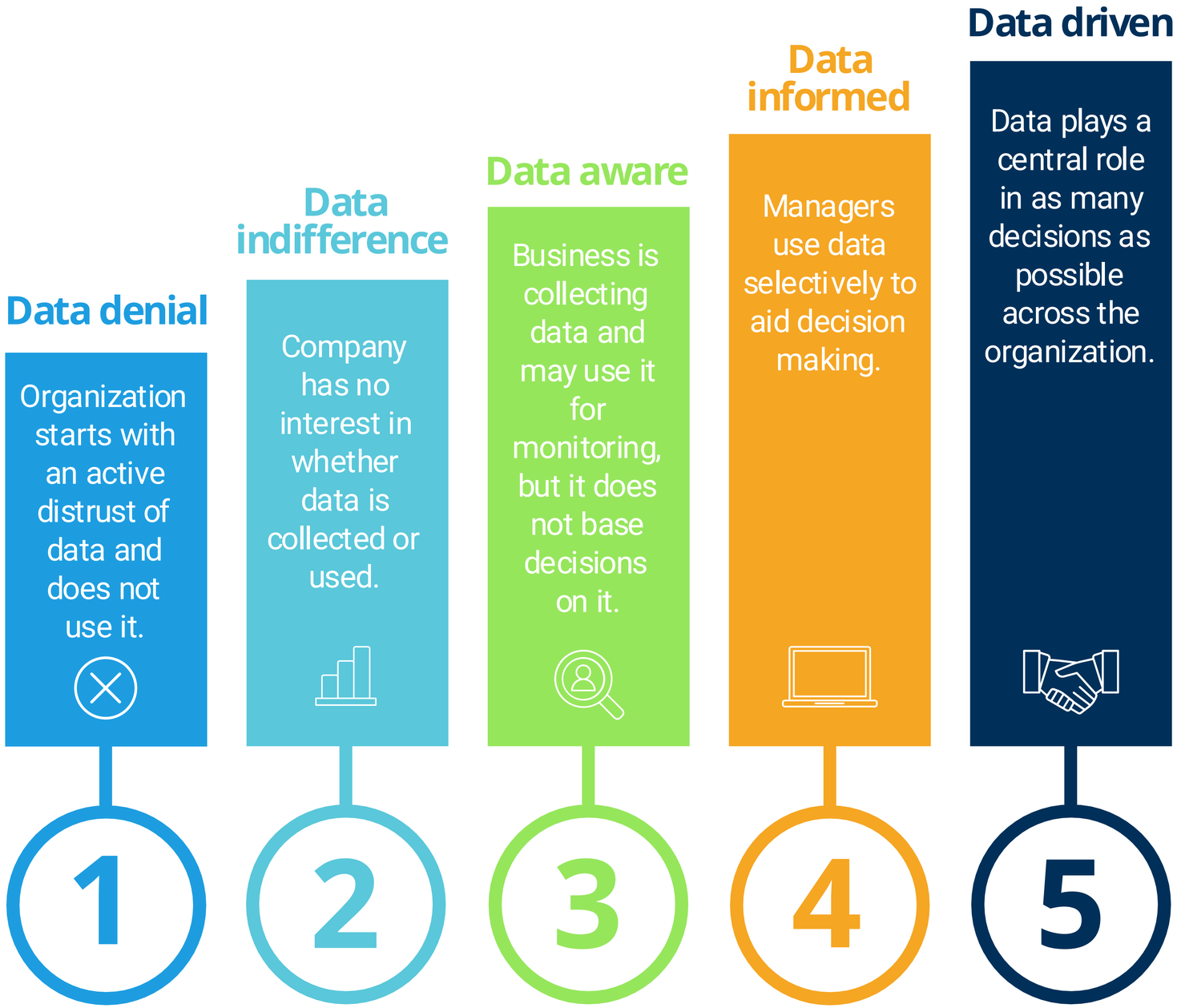
LIDÉ
VS.
TECHNOLOGIE

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.



https://www.smartsheet.com/data-driven-decision-making-management

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

Data/datový model

Business „co-proč-jak firma dělá“

„Techniky potřebné pro moji práci“

Strukturovaná/nestrukturovaná data

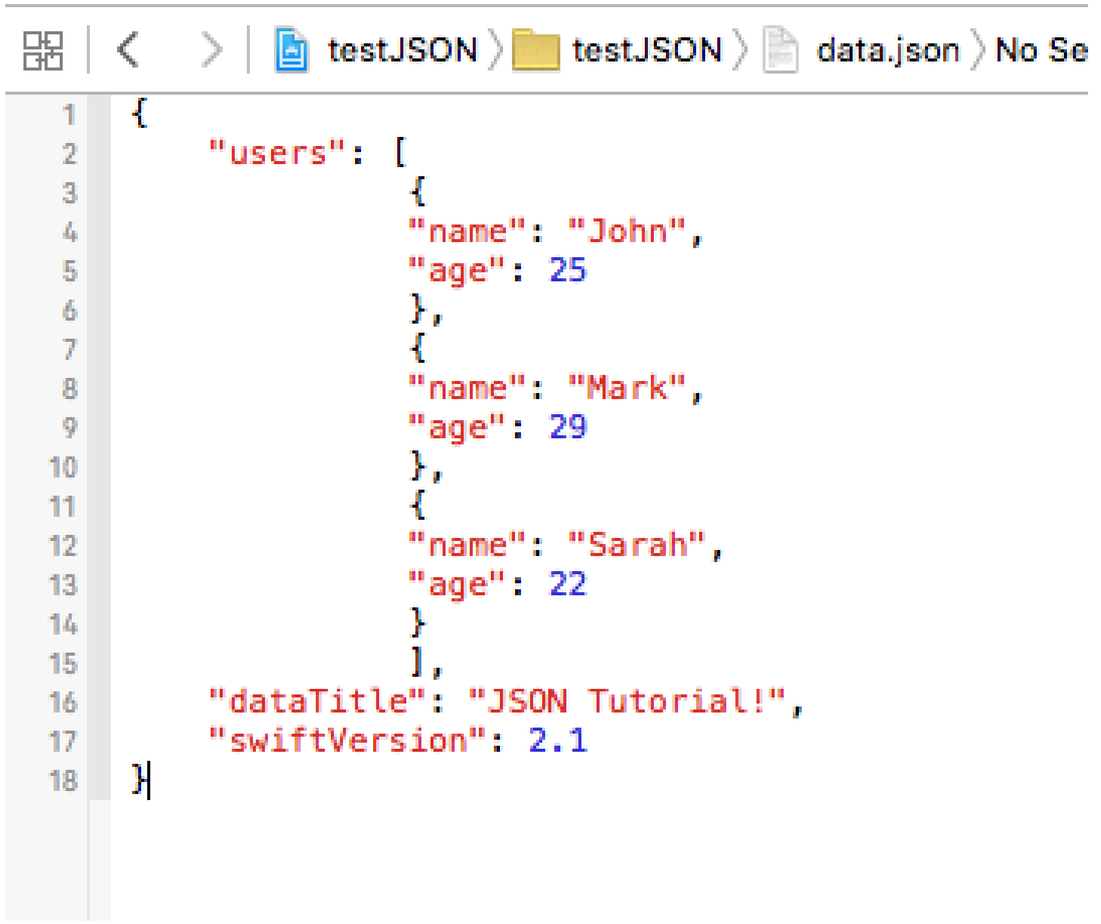
- **Strukturovaná data** – je znám datový model, datové typy apod., např. data v DWH

- **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
<?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>
```



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)

Microsoft SQL Server

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.

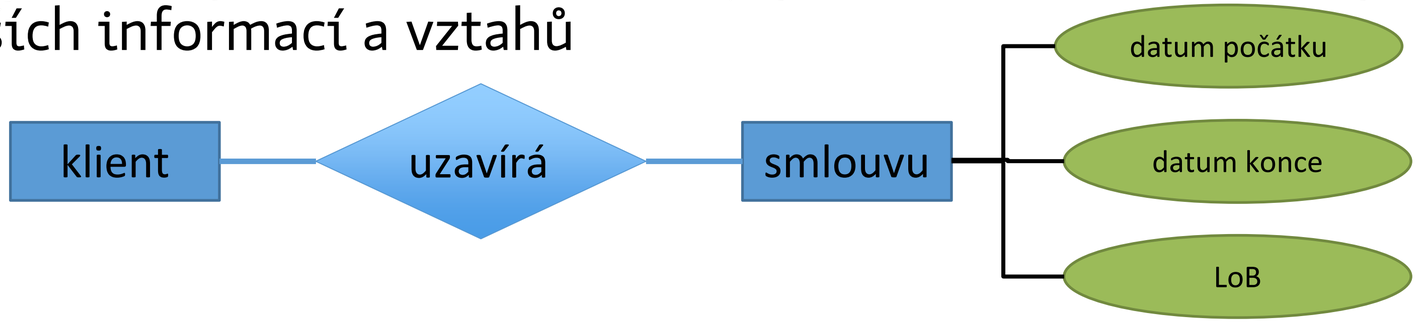
Extract – Load – Transform (ELT)

… 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.

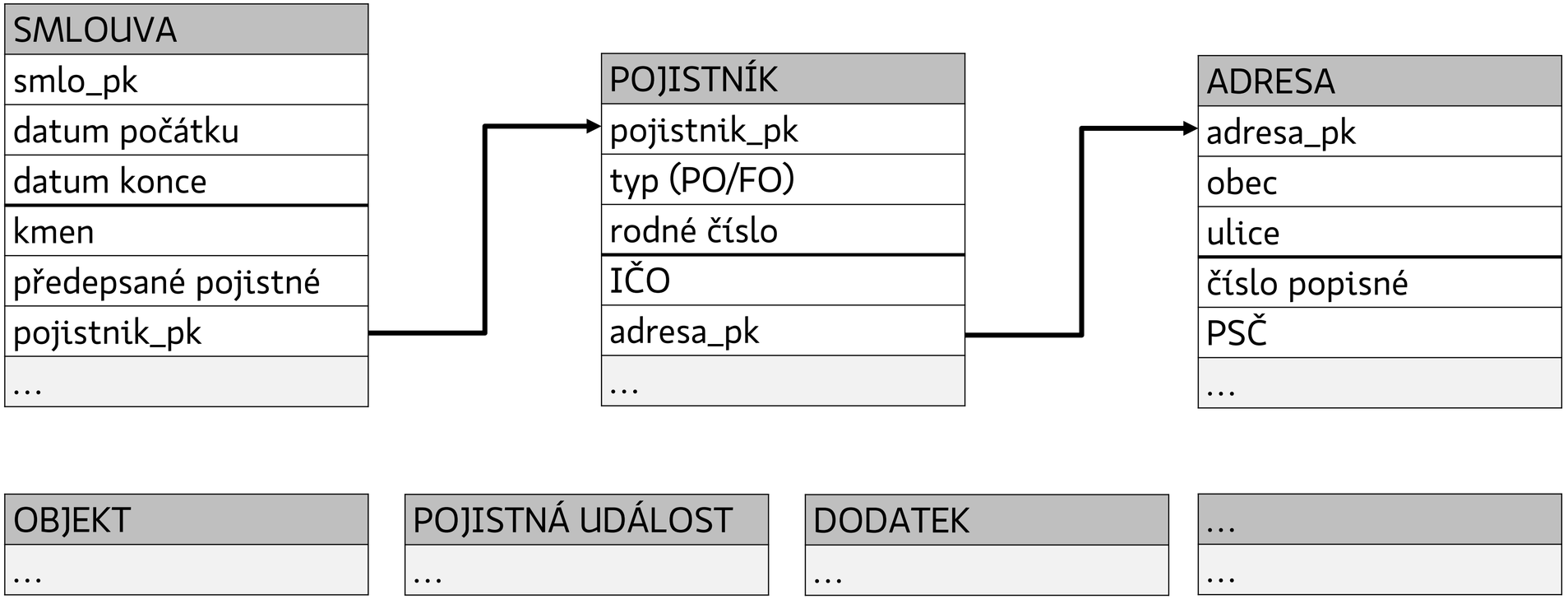
- Tři úrovně:
  - **Model reality** (konceptuální úroveň) – nejvyšší míra abstrakce – vymezení nejdůležitějších informací a vztahů



  - **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.rodné_číslo
        , b.IČO
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.pojistnik_pk
        , b.typ
        , b.rodné_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

Google Cloud

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

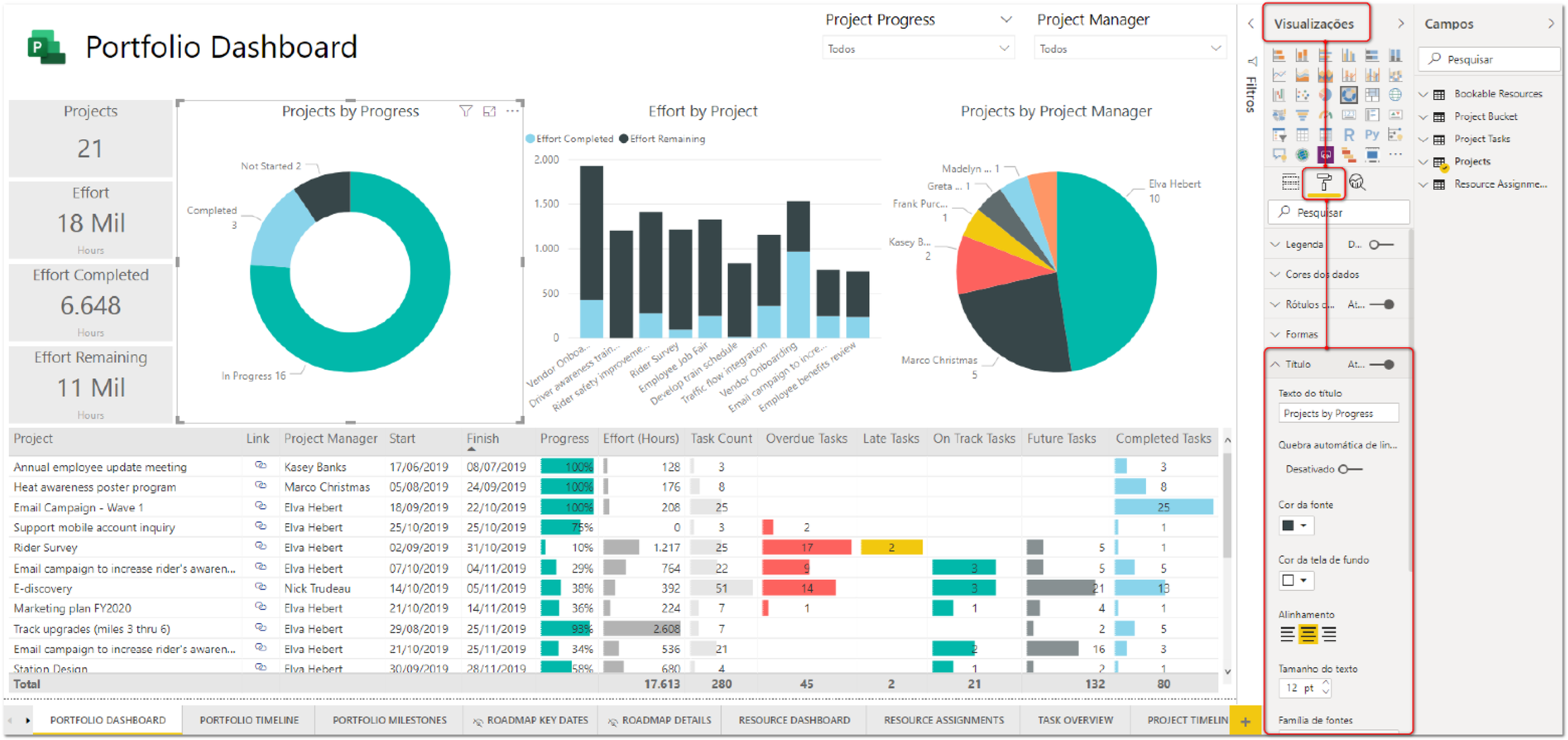
Příklad: Parametrizace škod pojistným matematikem VS. uzavření smlouvy obchodníkem. Kdo potřebuje vyšší dostupnost?

BI nástroje pro reporting

Příprava reportů a dashboa

Source: Gartner (February 2019)

BI nástroje pro reporting

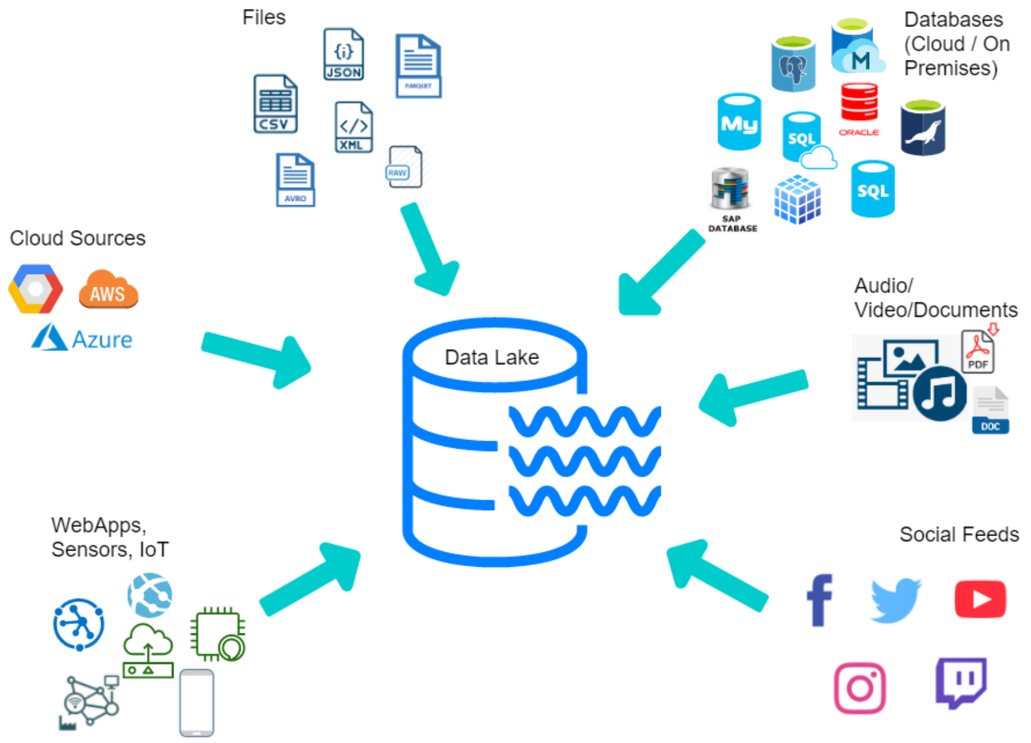


Datalake

- 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

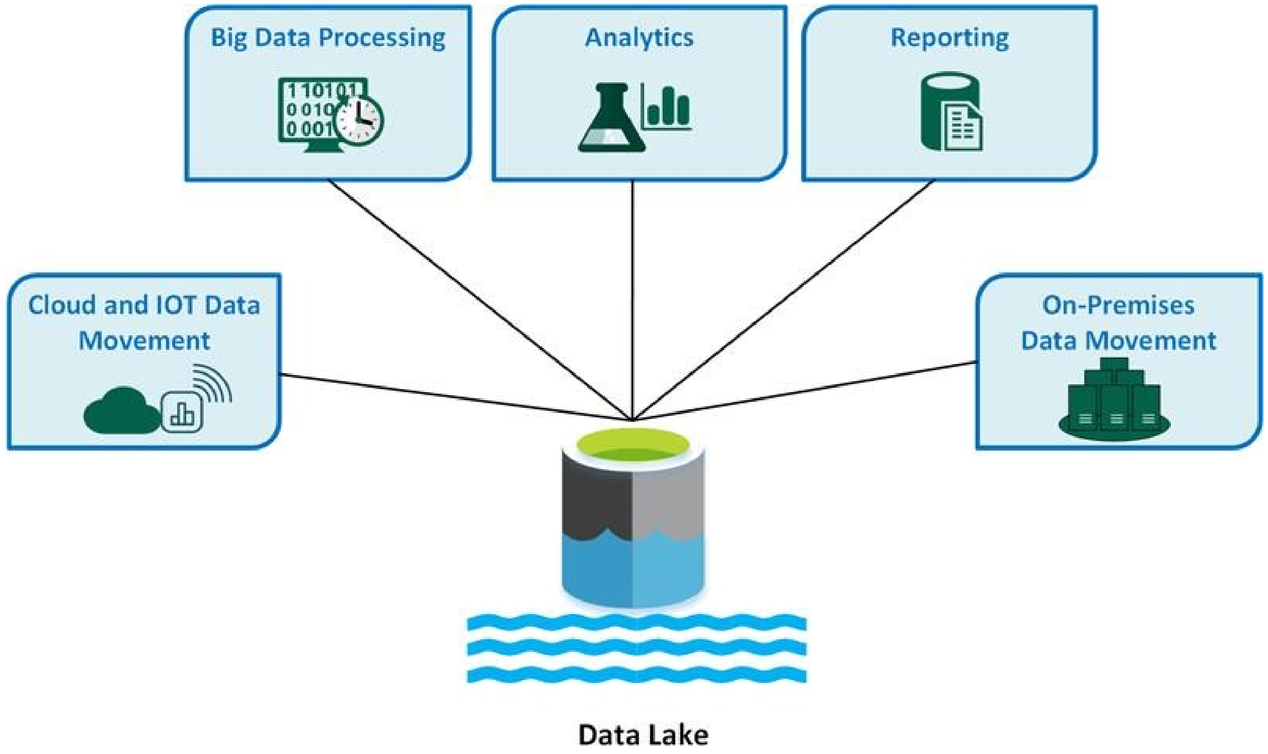


Datalake

- 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



https://learn.microsoft.com/cs-cz/azure/architecture/data-guide/scenarios/data-lake

Datová kvalita

- Dostupnost
- Přesnost
- Úplnost
- Konzistence

Komunita uživatelů dat

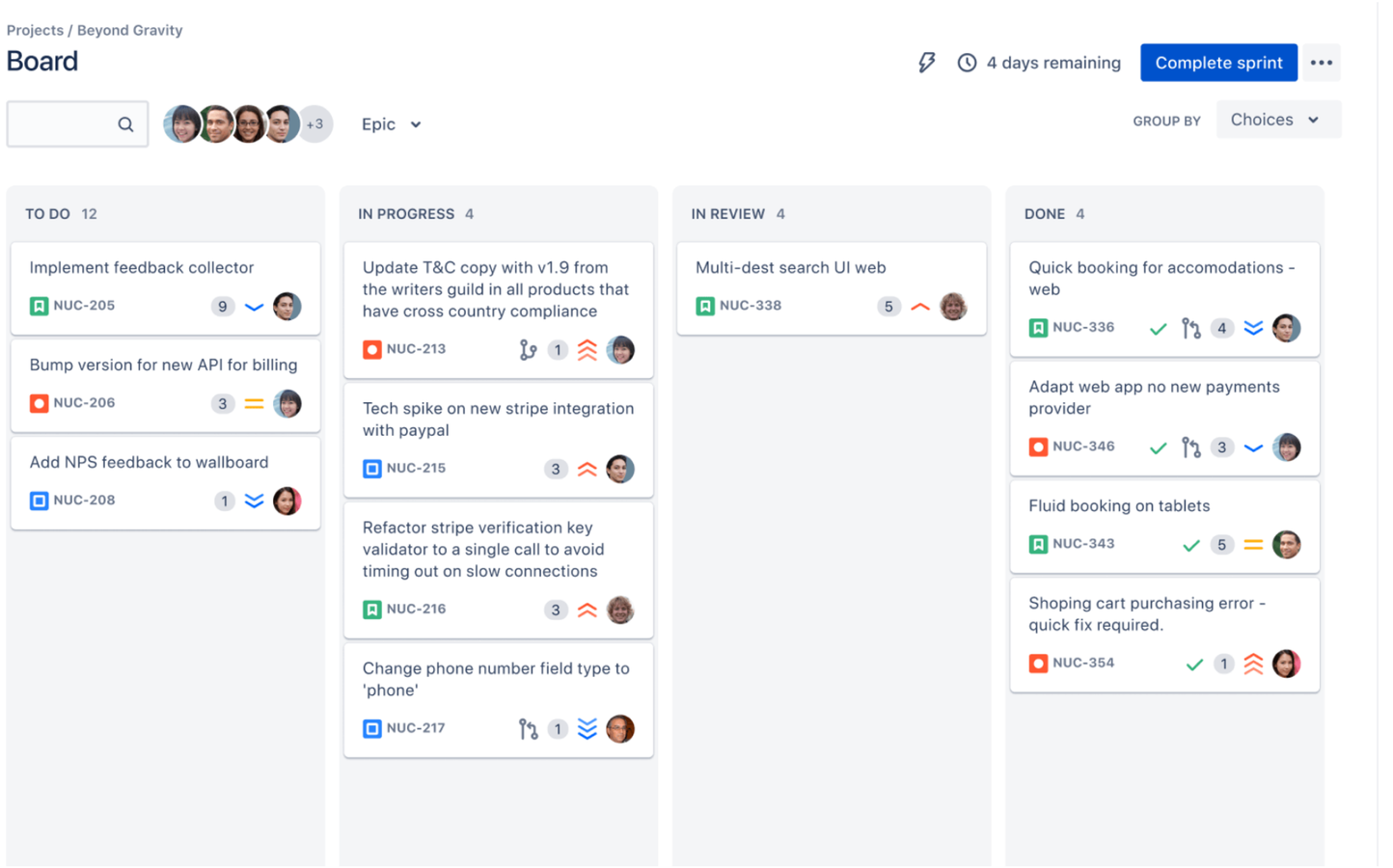
- Útvary pracující s daty mají své zástupce
- Ti se pravidelně setkávají a sdílejí vzájemně informace
- Následně jsou tyto informace sdíleny v rámci útvarů

Co se stane, když něco z výše uvedeného nefunguje?

Tiketový systém

Jira

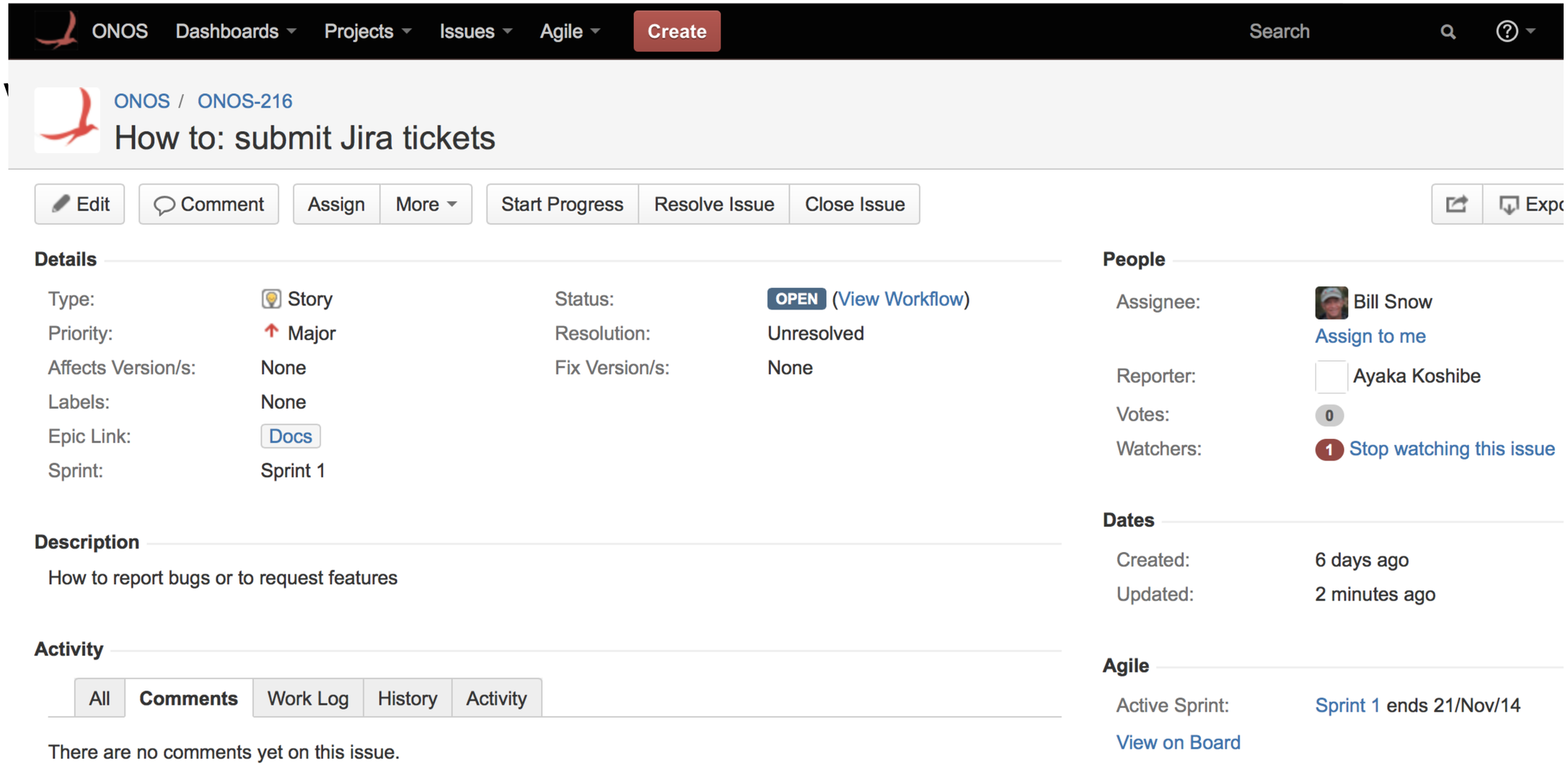
Všechny požadavky na změny, úpravy, opravy chyb se zadávají jako tikety.



https://www.atlassian.com/software/jira/features/scrum-boards

Tiketový systém

Jira



https://wiki.onosproject.org/display/

Další ne-nedůležitá témata

GitLab

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

Nakonec dopadneme takto ... :)