Algebraic Query Engine

Jiří Tůma, KA MFF UK
Tomáš Jirotka, GoodData

Počátky

- Firma GoodData založena v Brně roku 2007
  - M. Dovrtěl, později koupena R. Staňkem
- První kontakt s KA MFF v létě 2009
  - Query engine je nespolehlivý a nerozšiřitelný
  - Firma očekávala prudký růst

Co GoodData dělá?

- BI = business intelligence
  - Zpracování velkého množství dat (big data)
  - Integrace velkého množství vstupů
    - Výstupem jsou přehledné reporty pro podporu manažerského rozhodování (graf, tabulka)
- Dodává řešení od začátku do konce
  - Hardware (cloud), software (SaaS)
  - Professional services

Hlavní problémy QE

- Nespolehlivost
  - Stavový - v závislosti na stavu systému vracel různé výsledky jednoho reportu
- Nerozšiřitelnost
  - Omezené schopnosti - suma či průměr
  - Nepřehledný a složitý kód v Perlu
  - Nízká úroveň pokrytí automatickými testy

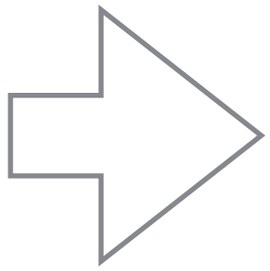
Střet informatiků s matematiky

- Informatické termíny
  - Dataset, hvězda, tabulka, cache, metrika, null,…
  - Nejasně definované, mnohdy přetížené výrazy
- GD abstraktní vrstva znemožňovala přímočarou aplikaci relační algebry

Jak z toho ven?

Příklady

Obrázky

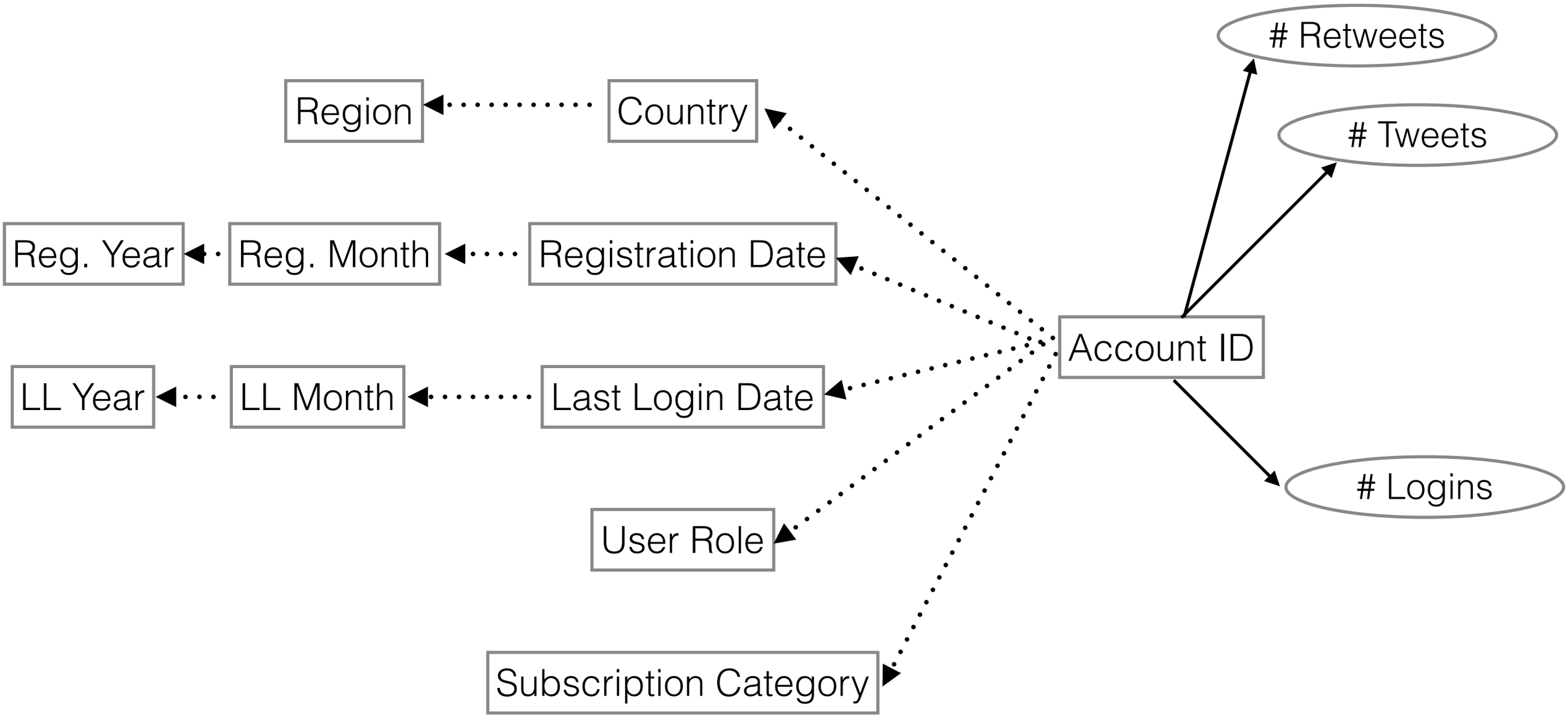


Přesná definice použitých termínů

- Atribut - prvek čum
- Element - instance atributu
- Projekce - uspořádání
- Elementární fakt - funkce do R

Příklad: uživatelé Twitteru

Account ID	Registration Date	Last Login Date	Country	Region	User Role	Subs. Ctg	#Tweets	#Retweets	#Logins
1234	4.3.2013	20.3.2015	CZ	EU	User	Free	23	12	842
452	3.9.2011	22.3.2015	US	NA	Admin	Profi	0	1	912
86	4.11.2014	3.2.2015	FR	EU	User	Free	1	9	23
7680	9.8.2011	31.1.2015	FR	EU	User	Free	8	2	76
6541	11.2.2010	4.3.2015	US	NA	User	Free	90	29	1287
8765	30.9.2011	19.3.2015	AU	APAC	User	Business	312	98	1823
97796	1.2.2013	25.3.2015	FR	EU	User	Free	34	6	479


Region
Country
Reg. Year
Reg. Month
Registration Date
LL Year
LL Month
Last Login Date
User Role
Subscription Category
Account ID
Retweets
Tweets
Logins

Věčná otázka

- Je větší den, nebo rok? A jak kreslit šipky?
- Granularita - zrnitost
- Celkem asi třikrát otočeno
- Projekce!

Datový model

- $L = (A, \leq, F)$ - logický datový model
  - $(A, \leq)$ - uspořádaná množina atributů
  - F_a - množina všech elementárních faktů relevantních k atributu *a*, jejich sjednocení je *F*
- Datová vrstva (DWH)

$$(L,\ \{E_a \mid a \in \mathcal{A}\},\ \{p_{a,b} \mid a, b \in \mathcal{A}, b \leq a\},\ \{h_{a,f} \mid a \in \mathcal{A}, f \in \mathcal{F}_a\})$$

Dotaz

- $q = (a, B, h_{a,f}, \Psi, S)$
  - a - střed hvězdy
  - $h_{a,f}$ - fakt relevantní k a, může být i derivovaný
  - B - antiřetězec nad a, dimenzinalita
  - Ψ - agregační funkce
  - S - selekce, podmnožina E_A