Vývoj databází a jeho reflexe v konferencích DATASEM, DATAKON a Data a znalosti v letech 1981 - 2016

J. Pokorný
MFF UK, Praha

Obsah

- Úvod
- DATAKON a jeho historie
- Historie DB v ČSR
- Kategorie příspěvků a jejich počty
- Vstup do 3. tisíciletí
- Závěr – aneb jak dál s DB

Úvod

- Motto:

  “Isn’t it ironic that in 2016 a non-skilled user can find a web page from Google’s untold petabytes of data in millisecond time, but a highly trained SQL expert can’t do the same thing in a relational database one billionth the size?.

  Jim Starkey (zakladatel NuoDB)

Úvod

- Databázové systémy – technologie
  - jak efektivně ukládat data na vnějších pamětech,
  - jak efektivně formulovat dotazy na takto uloženými daty.
- Databázové systémy – problémy návrhu, aplikace v IS, teorie

DATASEM a jeho historie

Samostatné konference u nás:

- Od r. 1981 – seminář DATASEM
  - O jazyku SEQUEL již v r. 1978 na konferenci SOFSEM
- Od r. 1986 – Moderní databáze
  - Další konference: Objekty, Systémová integrace, EurOpen (od r. 1990)
- Od r. 1995 – konference DATASEM
- Od r. 2001 – 2014 pod názvem DATAKON

DATASEM a jeho historie

- Od r. 2001 – existuje konference Znalosti
  - Pozn.: problematika znalostí v průniku DATAKON a Znalosti (viz články Jirků, P. Bartoše, P. Hájka v letech 1985, 1986)
- 2015 integrace – DATAKONu a Znalostí s novým názvem Data a znalosti
- Publikované historické přehledy: v r. 2000, 2005
- Poznámka: příkladem jiné úspěšné národní IT konference je EurOpen.cz. V říjnu 2016 to byla 47.!

Historie databází v Československu

- Svět: 1965 – CODASYL a DBTG – specifikace pojmů SŘBD, DBS, databázové schéma atd. + hierarchický model; 1970 - relační model dat
- první knihu o DB od J. C. Date z r. 1976 jsme viděli poprvé v ruském vydání někdy v r. 1977
- překlad knihy Database systems (D. C. Tsichritzis, F. H. Lochovsky) z r. 1977 vyšla v ČSR v překladu až v r. 1987
- první česká kniha o DBS od JP v r. 1992
- Zde: od 70. let IDMS, IDS + domácí produkty

Historie databází v Československu

- Svět: 1965 –CODASYL a DBTG – specifikace pojmů SŘBD, DBS, databázové schéma atd. + hierarchický model; 1970 - relační model dat
- první knihu o DB od J. C. Date z r. 1976 jsme viděli poprvé v ruském vydání někdy v r. 1977
- překlad knihy Database systems (D. C. Tsichritzis, F. H. Lochovsky) z r. 1977 vyšla v ČSR v překladu až v r. 1987
- první česká kniha o DBS od JP v r. 1992
- Zde: od 70. let IDMS, IDS + domácí produkty

SESAM, SOFIS

Historie databází v Československu

- Konceptuální modelování - doma populární (viz DATAKON ’81 a ´82.
  - vznik databázového modelu HIT (prezentován i na VLDB 1985)
  - funkcionální přístup – alternativa, navíc typovaný lambda-kalkul jako základ funkcionálních databázových jazyků
- Paralelně se rozvíjí funkční analýza.

Kategorie příspěvků a jejich počty

Kategorie	1981-2000	2001-2005	2006-2015
DB modely	32	13	7
Ontologie	NULL	NULL	3
NoSQL, Big Data	NULL	NULL	7
SŘBD cizí	29	1	0
SŘBD domácí	17	0	0
Distribuované SŘBD	22	0	4
Teorie databází	7	4	2
Architektury DBS	20	6	3
Projektování IS	70	12	14
Dotazovací jazyky	19	9	1
Textové databáze, Zpracování textu na Webu	20	5	14

Kategorie příspěvků a jejich počty

Kategorie	1981-2000	2001-2005	2006-2015
Sítě, Internet	7	NULL	NULL
Sítě	NULL	3	0
Web, XML, Open Data	NULL	16	26
Fyzické datové struktury, provoz DBS	15	9	1
Umělá inteligence	34	9	NULL
Dolování dat, Analytika	NULL	NULL	13
Aplikace	15	12	16
Přehledové příspěvky (Tutoriály)	12	0	27
Bezpečnost	11	15	10
Řízení IS/ITC	17	3	10
Ostatní	20	6	24
Celkem	367	123	182

Vstup do 3. tisíciletí

- XML databáze (EurOpen, DATAKON 2004, kniha 2008)

XML DB: současný dstav a perspektivy (Pokorný, 2004)
XML DB (Mlýnková, Pokorný, Richta, Toman, K., Toman, V, Grada 2008)

Vstup do 3. tisíciletí

- Směrem k velkým datům (DATAKON 2011, 2014, kniha 2015)

Zvané přednášky:
NoSQL DB (Pokorný, 2011)
BigData (Pokorný, 2014), (Slabý, 2014)
OpenData, Linked Data (Chlapek, Kučera, Nečaský, 2014)

Kniha: BigData a NOSQL DB (Holubová, Kosek, Minařík, Novák, Grada 2015)

Vstup do 3. tisíciletí

- Nové databázové architektury
  - NoSQL
    - nerelační,
    - horizontálně škálovatelné
    - sdílení ničeho,
    - podpora replikací
    - obvykle žádné schéma,
    - jednoduché dotazovací jazyky,
    - ACID vs. CAP

Vstup do 3. tisíciletí

- Big Data Management Systems (BDMS)
  - ASTERIX,
  - Oracle Big Data Appliance kombinuje v SQL Hadoop a NoSQL v jeden dotaz SQL.
- NewSQL databáze (od r. 2011)
  - jsou navržený pro horizontální škálování na strojích v režimu sdílení-ničeho,
  - garantují ACID vlastnosti ,
  - aplikace na nich interagují s databází primárně přes SQL (včetně spojení),

Vstup do 3. tisíciletí

- používají pro řízení souběžného zpracování protokol bez zamykání,
- poskytují vyšší výkon než tradiční relační DB

- Další přístupy v NewSQL
  - SQL on Hadoop (např. Big SQL – IBM)
  - relační SŘBD na Hadoop (např. HP Vertica se sloupcovou relační DB a Hadoop konektory $\Rightarrow$ SQL dotazy přímo nad Hadoop
  - Možnost vzdálených dotazů nad Hadoop zasláním do relačního SŘBD a následnou integrací odpovědi s SQL (viz Oracle Big Data Appliance)
  - Další Hadoop-relační hybrid: HadoopDB

Nové databázové architektury v posledních 15 letech

Mílník	Kategorie	Subkategorie	Reprezentanti
2009	NoSQL	klíč-hodnota	Redis
		sloupcově-orient.	Cassandra
		dokumentově-orient.	MongoDB
		grafové databáze	Neo4j
2005	BDMS	1. generace	Hadoop software stack
2010		2. generace	Asterix software stack
2011	NewSQL	obecné	NuoDB, VoltDB, Clustrix
		hybridy Google	Spanner
		Hadoop-relační	Vertica , HadoopDB
		SQL-on-Hadoop	Hive, BigSQL
	Ostatní	NoSQL s ACID	FoundationDB, MarkLogic , OrientDB

Význačnost NoSQL ve světě databází (podle DB-Engines Ranking*, říjen 2016)

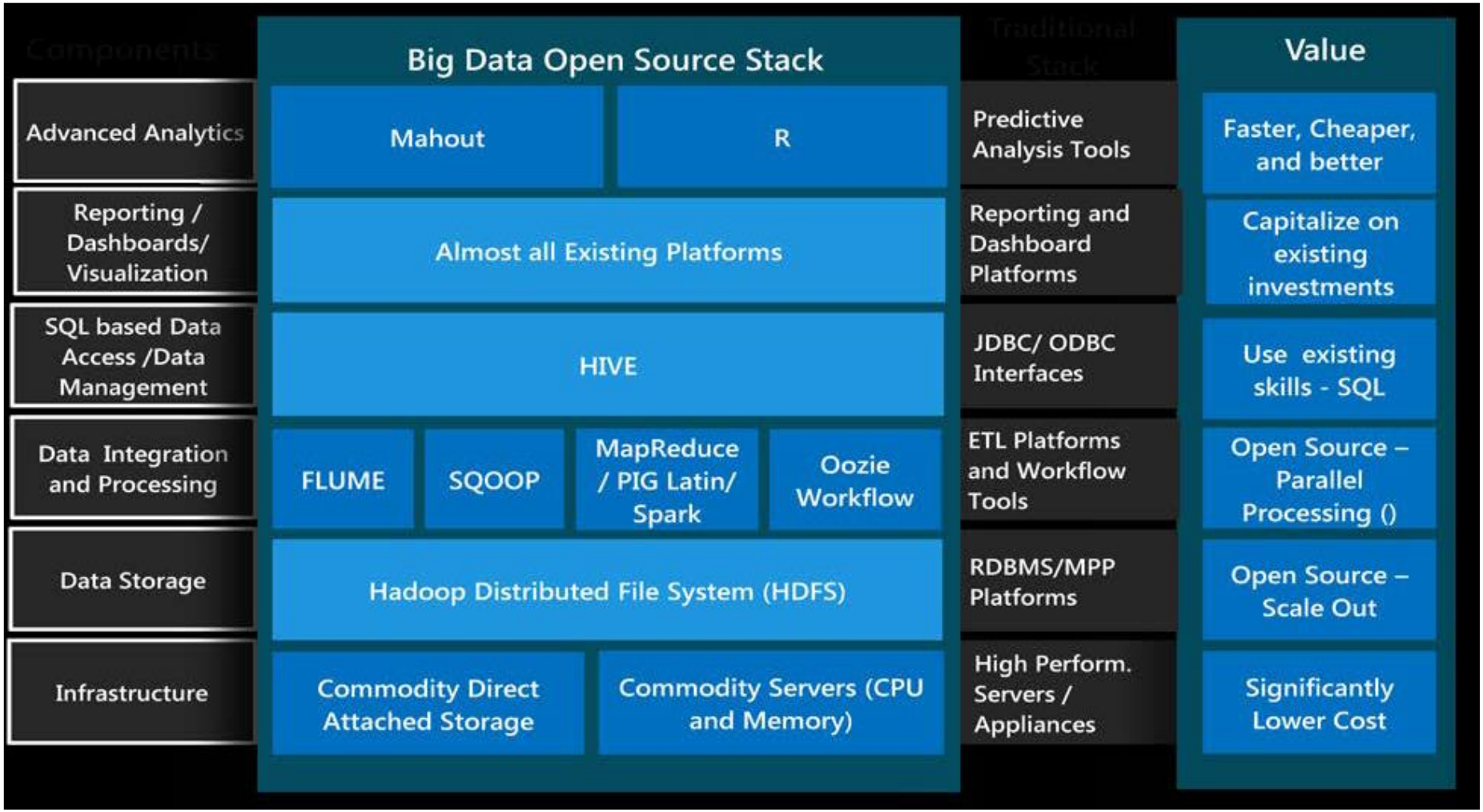
Pořadí	SŘBD	Databázový model	Skóre
1	ORACLE	relační	1417.10
2	MySQL	relační	1362.65
3	Microsoft SQL Server	relační	1214.18
4	MongoDB	dokumentově-orient.	318.80
5	PostgreSQL	relační	318.69
6	DB2	relační	180.56
7	Cassandra	sloupcově-orient.	135.06
8	Microsoft ACCESS	relační	124.68
9	Redis	klíč-hodnota	109.54
10	SQLite	relační	108.57

*http://db-engines.com/en/ranking

Závěr – aneb jak dál s DB

- klíčové problémy jsou v rozhodnutích s NoSQL:
  - volba správného produktu
  - návrh vhodné architektury pro danou třídu aplikací, integrace s dalšími systémy
    - Stefan Edlich navrhuje zvolit NoSQL DB po zodpovězení okolo 70 otázek v 6 kategoriích + zhotovení prototypu
- nové alternativy pro analýzu Big Dat s NewSQL:
  - Big Analytics (např. ClustrixDB (pro TP a analytiku v reálném čase)
  - In-memory databáze (např. SAP HANA podporuje jak OLTP tak OLAP, H-store, VoltDB)

Příklad architektury
Open Source Big Data Stack



Závěr – aneb jak dál s DB

- Extreme Big Data (EBD) (do YBajtů (10*24))
  - Problémy:
    - nejen velikost, hlavně integrace;
    - jak sestavit výslednou architekturu.
- Malá data, velké algoritmy (Ullman - IDEAS, 2016):

Př.: podobné prvky, aniž by testoval $\binom{n}{2}$ párů prvků (což pro $n=10^6$ vede ke trilionu porovnání).

LSH používá *k* kapes (buckets) a hašovací funkce, kterými prvky zahašuje v čase O(*kn*). Zajistí, že 2 podobné prvky jsou v jedné kapse. $\Rightarrow$ přibližné vyhledávání.

Závěr – aneb jak dál s DB

- Důsledky velkého množství dat
  - Utichá debata o konsistenci vs. dostupnosti (viz CAP teorém)
  - Nelze zjistit, co přesně skutečně existuje (uživatel je schopen prohlédnout jen několik prvních stránek výsledků)
- Dedukce vs. indukce
  - Věda založená na datech je výhradně založena na induktivním uvažování.

    Anderson, Ch.: The end of theory: The data Deluge Makes the Scientific Method Obsolete. Wired (2008)
- Síla Big Dat
  - nejen symbolicky reprezentují svět, ale rovněž ho produkují.
  - vyhledávací algoritmy na síti např. ovlivňují volby

    Epstein, Robertson:The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. Proc Natl Acad Sci U S A. (2015)

Závěr – předposlední poznámka

- Některé výzvy pro komunitu z Beckman Report on Database Research* (2016):
  - Databázové vzdělávání: je dnes odtrženo od reality. Je třeba předělat DB sylaby (DB v paměti, DB klíč-hodnota, zpracování proudů dat, MapReduce, …)
  - Data science: cíl – transformovat (velká) data do znalostí. Kvalifikace – nejen datařina, ale i BI, počítačové systémy, matematika, statistika, strojové učení, optimalizace.
  - Kultura výzkumu: zvyšující se důraz na citace vs. výzkumný impakt. Práce na velkých projektech, vývoji koncového software, sdílení velkých kolekcí dat atd., trvá déle než řešení speciálních konkrétních problémů. Vina je i na PC konferencí.

*Abadi, D. et al: The Beckman Report on Database Research, CACM, 59, 2, 2016

Závěr – poslední poznámka

- Klasický citát anglického básníka T.S. Eliota:

  Kde je moudrost?
  Ztracena ve znalostech.
  Kde jsou znalosti?
  Ztraceny v informacích.

A databázista J. Celko pokračuje:

  Kde jsou informace?
  Ztraceny v datech.
  Kde jsou data?
  Ztracena v databázích.

A autoři JP a MV v knize o databázích v r. 2013

  Kde jsou databáze?
  Ztraceny v cloudu.