

SAS a analýza velkých dát

eFOCUS

III. Technológie a softvéry na spracovanie veľkých dát

Tomáš Lencz, SAS Slovakia, s. r. o.



Big data, Hadoop... takto?



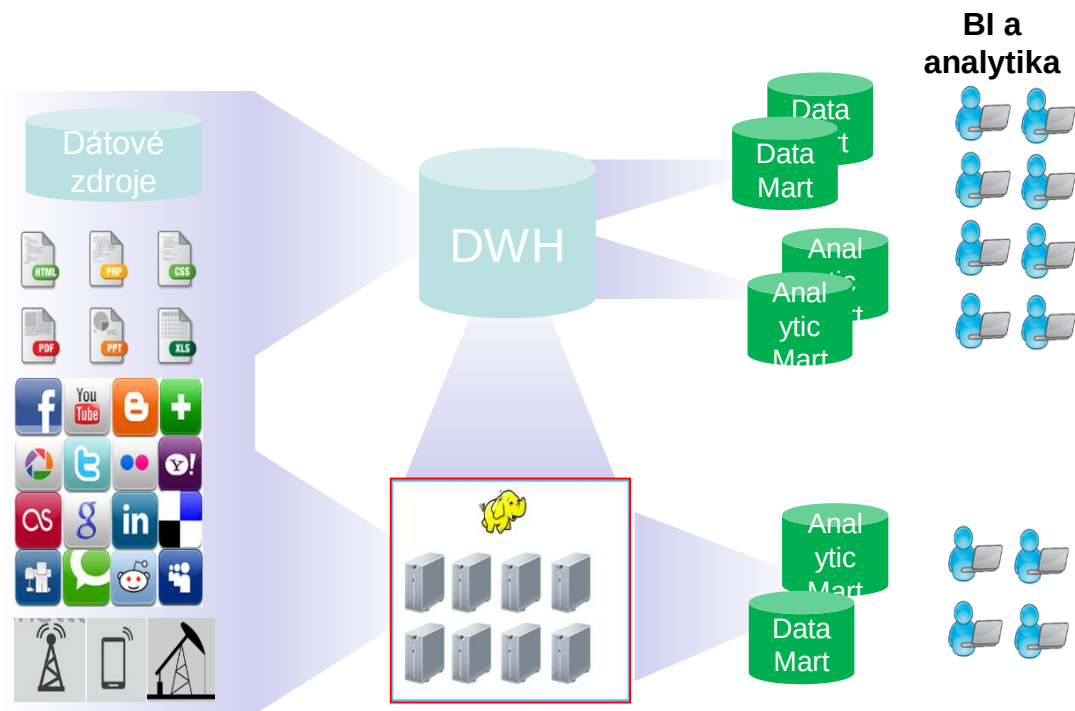
- **Dátové jazero** či **dátová skládka**?
 - Hadoop výhoda: „Schema on Read“
 - Data management!
- Využitie Hadoopu cez databázový prístup cez Hive? Alebo Impala?
- MapReduce ako podpora interaktívneho BI a analytiky?
- Prerobenie ETL existujúcich DWH?
- Hlavné argumenty pre Hadoop - webové logy a Internet of Things?

Source: GettyImages

Prínosy analýzy Big data

- **Získať užitočné informácie** z rozsiahlych dát, najlepšie **interaktívne**
 - Už teraz veľa obchodných, zákazníckych a prevádzkových dát (+ nové data + archív)
 - Sú potrebné **viacprechodové** analytické algoritmy pre dáta a text - **SQL/HiveQL, MapReduce nestačia**
- **Lievik na dáta** – príklad z analytiky:
 - Prediktívny model na 1000+ premenných
 - Nižší výkon môže viesť k nutnosti použiť vzorky dát
 - Na výsledný skóring stačí malá časť premenných

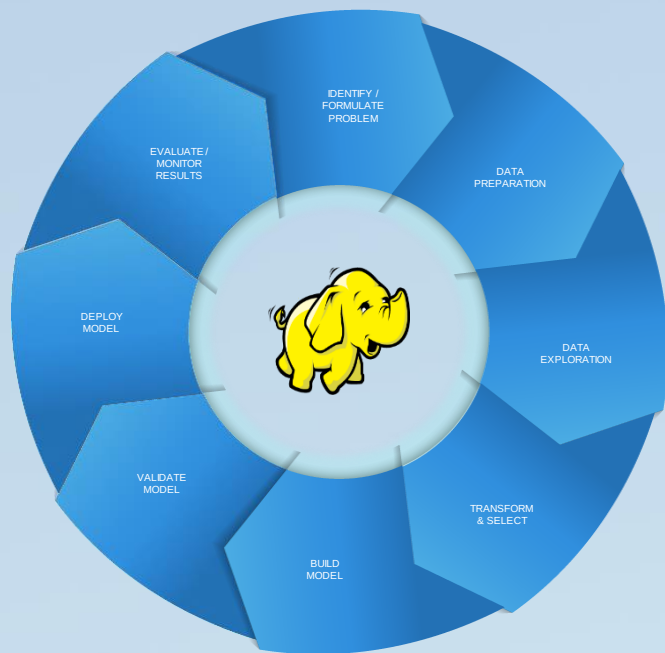
Hadoop treba mať



*Schéma spolupráce
DWH a Hadoopu*

- Pred – vedľa DWH – namiesto DWH?
- DWH pravdepodobne nebude v dohľadnom čase nahradený pre štatutárny reporting či kontroľovanie
- **Ostatné je vec jasných cieľov a dobrej architektúry**

Čo nesmie chýbať



*Životný cyklus pre
Hadoop*

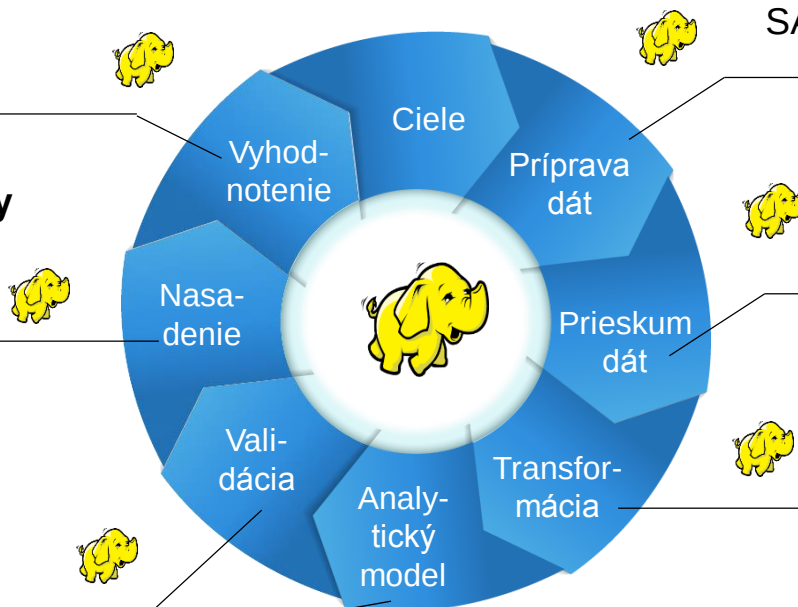
- Pokrytie celého **životného cyklu**
- Transparentnosť pre užívateľov
- Data management
 - Metadáta
 - Profilovanie dátovej kvality
 - Párovanie rôznorodých dát
- In-memory analytika (dáta, text) ako kontrapunkt k MapReduce
- Pre IoT – event stream processing ako vstupný filter

SAS unikátne pokrytie životného cyklu

Interaktívny vizuálny nástroj SAS Visual Analytics **(in memory)**

Skórovacie akcelerátory pre **Hadoop aj relačné databázy**

„High Performance“ verzie existujúcich vizuálnych analytických nástrojov **(in memory)**



Vizuálne nástroje pre analytikov Enterprise Guide a Enterprise Miner (Hive, Impala, priamy prístup) SAS Event stream processing

Vizuálna interaktívna analýza dát a textu **„in-memory“** - Visual Analytics, Visual Statistics

Vizuálne nástroje pre ETL a **dátovú kvalitu** - DI Studio, Data Loader for Hadoop (beta)

Ďakujeme za pozornost

