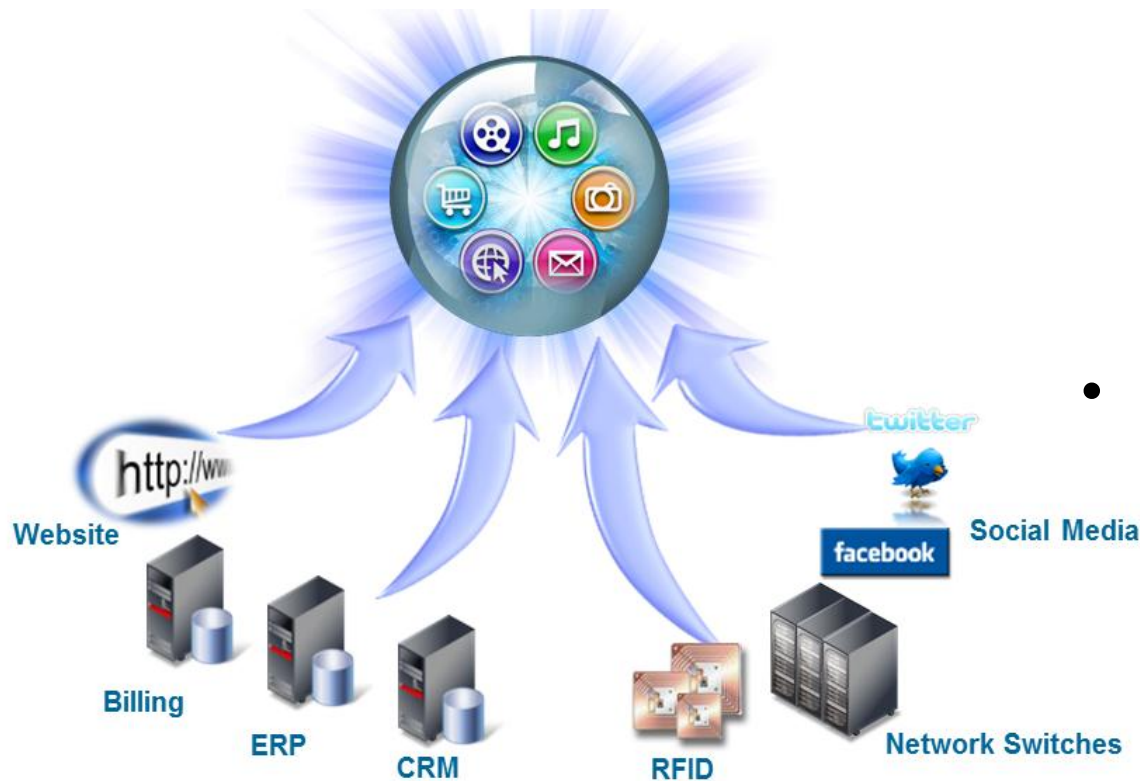


BIG DATA je oveľa viac ako Hadoop

*e*FOCUS



Analýza všech dostupných dat?



- **Big data = ?**

- = Buzzword?
- = Hadoop?

- **Hadoop**

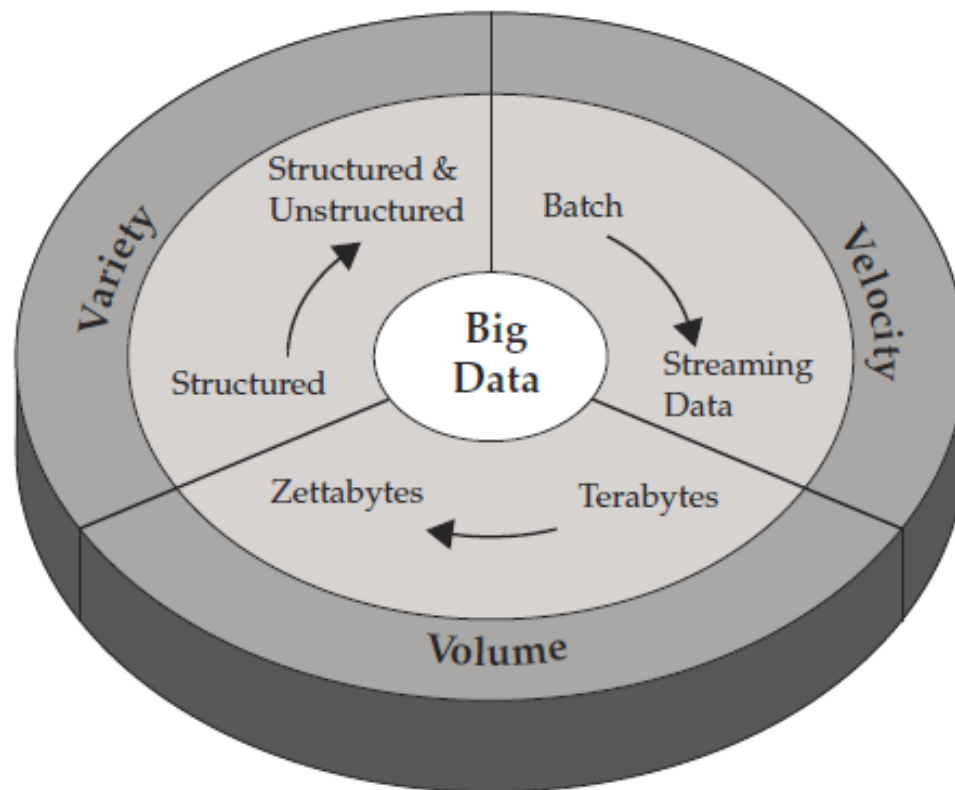


- Jen ke zpracování **nestrukturovaných dat?**



Mentální posun ⇔ něco za něco

Big data = 3 + 1 “V”



- Čtvrté „V“: VERACITY

- Důvěryhodnost
- Odstranění šumů, ...



1.) Konsolidovaný pohled na data



*Rozhraní pro koncové uživatele
připomínající **PORTÁL***

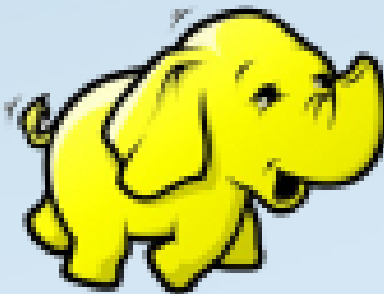
- Federativní vyhledávání
- Data zůstávají na místě
- Odlehčené “MDM”
- Data
 - Strukturovaná
 - Nestrukturovaná
 - Vnitrofiremní (i v Hadoopu)
 - Externí
 - Sociální sítě, diskuzní fóra, ...

2.) Analýza „velkých surových dat“



Archaický obrázek Hadoop clusteru

- **Hadoop**
 - Apache open source software framework
 - Inspirován Googlem
 - Google FS
 - Google MapReduce
- **HDFS**
- **Map/Reduce**



2.) Analýza „velkých surových dat“

- **HDFS <= Velké objemy dat**
 - Jak data v HDFS analyzovat?
 1. Konverze do relační podoby
 2. Využití speciálních technologií pracujících s HDFS



3. *Použití standardních technologií + **nová rozhraní***
 - Zejména pro strukturovaná relační data
 - Střízlivé uvažování je ale na místě

2.) Analýza „velkých surových dat“

Tradiční přístup

Opakovatelné analýzy

Business

- Ví přesně, co chce



IT

- připraví infrastrukturu,
- datové struktury,
- ...



Např.:

- Pravidelné měsíční reporty
- Zákaznické dotazníky,
- ...

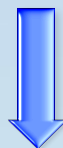
Big data přístup

Iterativní přístup, zkoumání



IT

- připraví platformu poskytující kreativní zkoumání možností



Business

- Zkoumá, co může být skutečně využito
- Definuje úlohy / zadání pro IT dle tradičního přístupu



2.) Analýza „velkých surových dat“

- **Hadoop** má mnoho distribucí

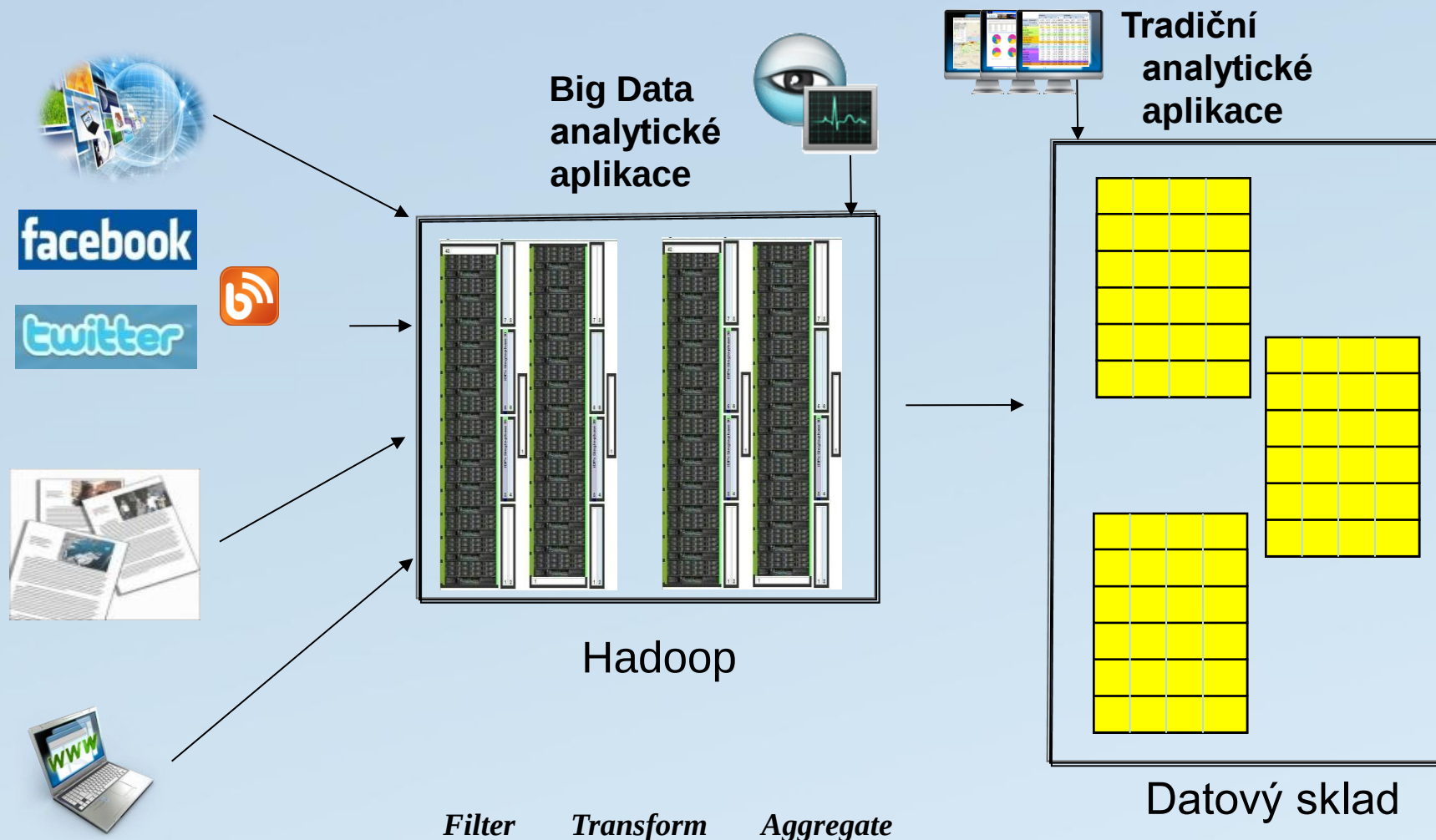


a mnoho dalších ...

- Co je specifického na **IBM** distribuci?
 - **Odstínění koncových uživatelů od složitosti** platformy HADOOP
 - Pro business uživatele:
 - Analytické nástroje pro prvotní zkoumání „velkých surových dat“
 - Pro IT uživatele:
 - Snadná instalace & správa systému, vylepšení open-source standardu, konektivita na běžné systémy, textová analytika, integrace
 - Komfortní vývojové prostředí
 - Nový programovací jazyk pro rychlý vývoj aplikací - JAQL

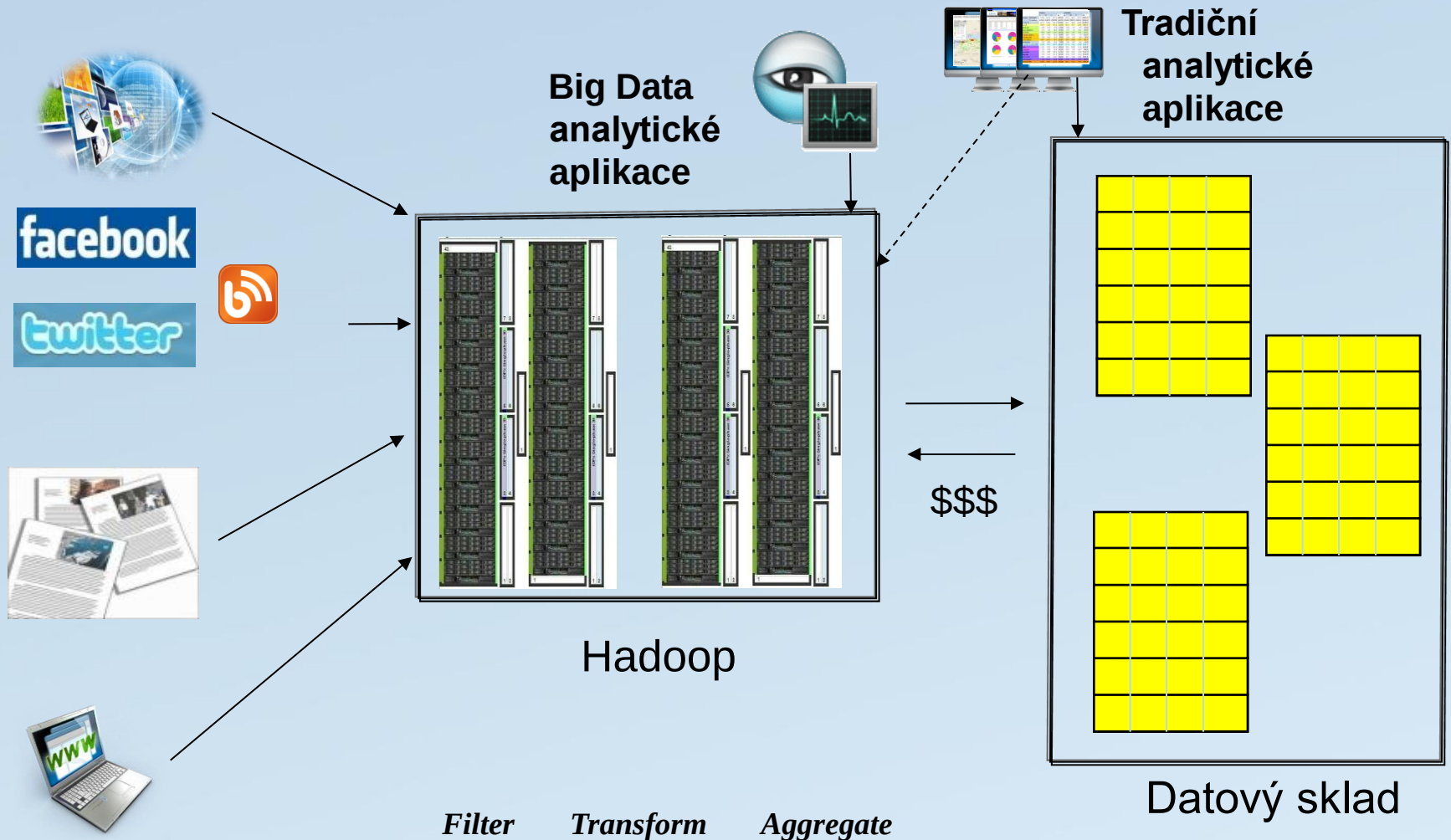
2.) Analýza „velkých surových dat“

- Standardní pozicování: Hadoop vs. datový sklad



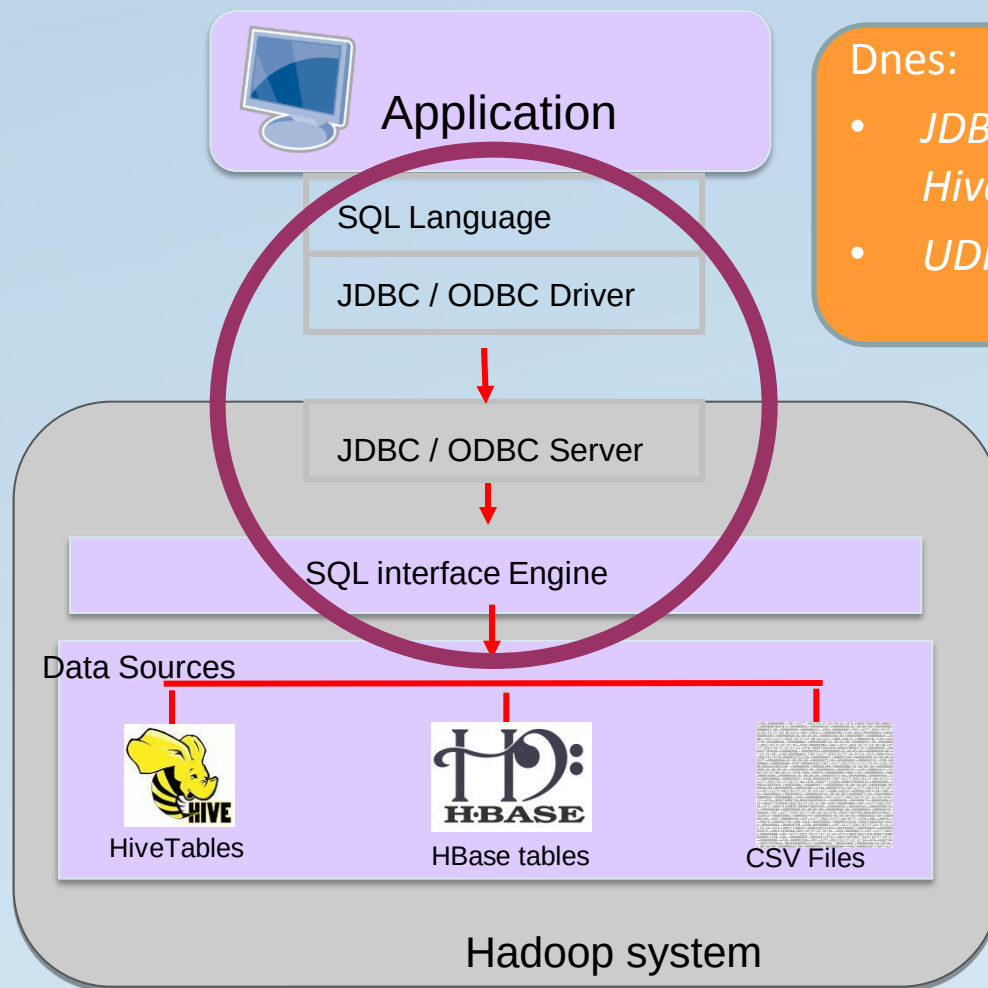
3.) Snížení nákladů díky Hadoopu

- Alternativní pozicování: Hadoop vs. datový sklad



3.) Snížení nákladů díky Hadoopu

- Velmi blízká budoucnost



Dnes:

- *JDBC for Hive*
- *UDF*

- Podpora SQL 92
 - Částečně i SQL 2011
- Přístup ke všem relačním datům v Hadoop
- Podpora ODBC & JDBC

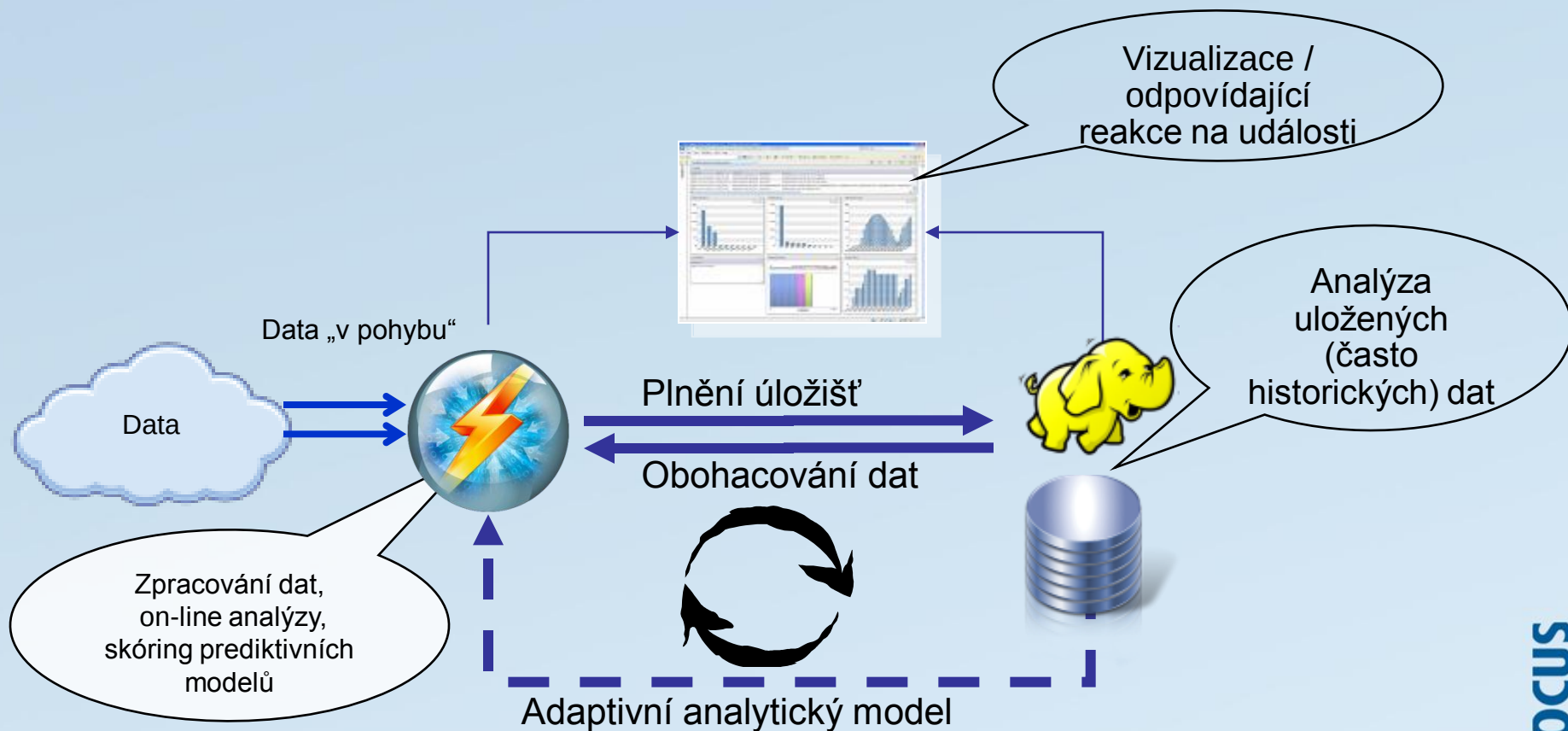
4.) Zjednodušení datových skladů

- Dnes jsou **velmi složité**
 - Díky různým (**často pochybným**) požadavkům
- Použití standardních OLTP databázových technologií
 - Neodpovídají současným analytickým potřebám
 - Zásadní poždavek: Flexibilita a jednoduchost
 - Dnes je realizace poždavků často v řádech měsíců
 - Kompletně ve správě IT => **Posun směrem k business uživatelům**
- Big data do DWH => ještě větší složitost

Analytik funguje často **iterativně**
=> Obrovský tlak na **ad-hoc** dotazy / analýzy

5.) Zpracování dat v pohybu

- Jedním z „V“ z Big data problematiky je „Velocity“
- **Dávkové zpracování dat nemusí stačit**



IBM BIG DATA platforma

Klíčové oblasti

1.) Konsolidovaný pohled na data



InfoSphere
Data Explorer

2.) Analýza „velkých surových dat“



InfoSphere
BigInsights

3.) Snížení nákladů díky Hadoopu

4.) Zjednodušení datových skladů



PureData
System for Analytics

5.) Zpracování dat v pohybu



InfoSphere
Streams

- Integrace
- Využitelnost předchozí práce

Ďakujeme za pozornost

