
ADASTRA

ADASTRA

# Soumrak klasických datových skladů

Data Management Workshop 2019
9/10/2019 Praha

Martin Bém

# Stručně o mě

- Data & DW architekt
- Příležitostný Data Engineer
- V Adastře od roku 2007
- 12+ let praxe v DW a BI
- 25+ projektů v ČR a v zahraničí
- Zaměření na bankovnictví a finanční služby
- Organizátor meetupu PDM

- Správa partnerství
  - Oracle (včetně Oracle Cloud)
  - Amazon Web Services
  - Google Cloud
- Nejoblíbenější technologie
  - Oracle Exadata Database Machine
  - Oracle Autonomous Database
  - SAP PowerDesigner
  - Google BigQuery
  - Amazon S3

**Martin Bém**

Senior Consultant

martin.bem@adastragrp.com

+420 603 505 247

https://www.linkedin.com/in/martinbem/

Data Warehousing | Data Modeling | Big Data | Data Governance | Data Management | Master Data Management | Reference Data Management | Data Quality | Operational Data Store | Metadata Driven Development | Cloud

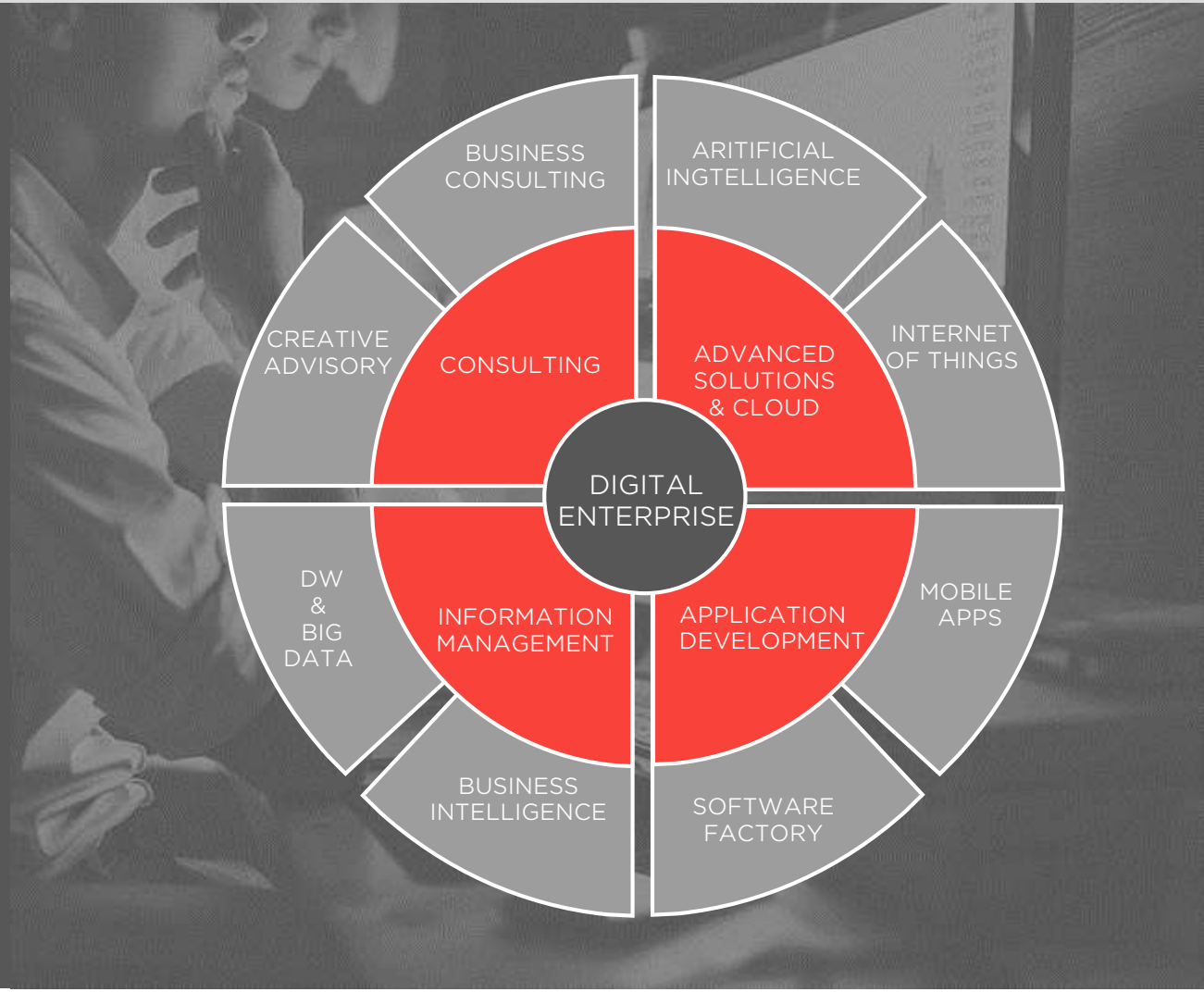
# Data spojují všechno & všechny

## Náš cíl

Prostřednictvím chytrých datových řešení přispívat k rozvoji byznysu našich zákazníků.

## Naše motto

Jsme Váš partner pro digitalizaci




Adastra se umístila na 9. místě TOP 100 ICT společností v ČR za rok 2018 dle časopisu CIO Business World.

# ADASTRA není už dávno jen Data Warehousing

## ADASTRA GROUP

**ADASTRA**

Mezinárodní konzultační společnost, která dodává funkční odvětvová řešení usnadňující přechod do digitální éry.

**ADASTRA BUSINESS CONSULTING**

Poradenství v oblasti řízení rizik, prodeje a optimalizaci procesů.

**BLINDSPOT.AI**

Softwarová řešení v oblasti strojového učení a aplikované umělé inteligence.

**ataccama**

Vývoj SW pro řízení datové kvality, MDM, reference data management, Big Data a Data Governance.

**ADASTRA.ONE**

Vývoj mobilních aplikací a dodávky mobilních řešení pro velké spektrum klientů.

**acamar**

IT bodyshopping pro společnosti zejména z bankovnictví, telca a pojišťovnictví včetně Adastry.

**ph**

Kreativní agentura se silným technologickým know-how.

**instarea**

Řešení pro monetizaci big dat.

# Adastra Group ve světě

**3,5**

**Tržby 3,5 mld. Kč**
**V roce 2018**

**2000+**

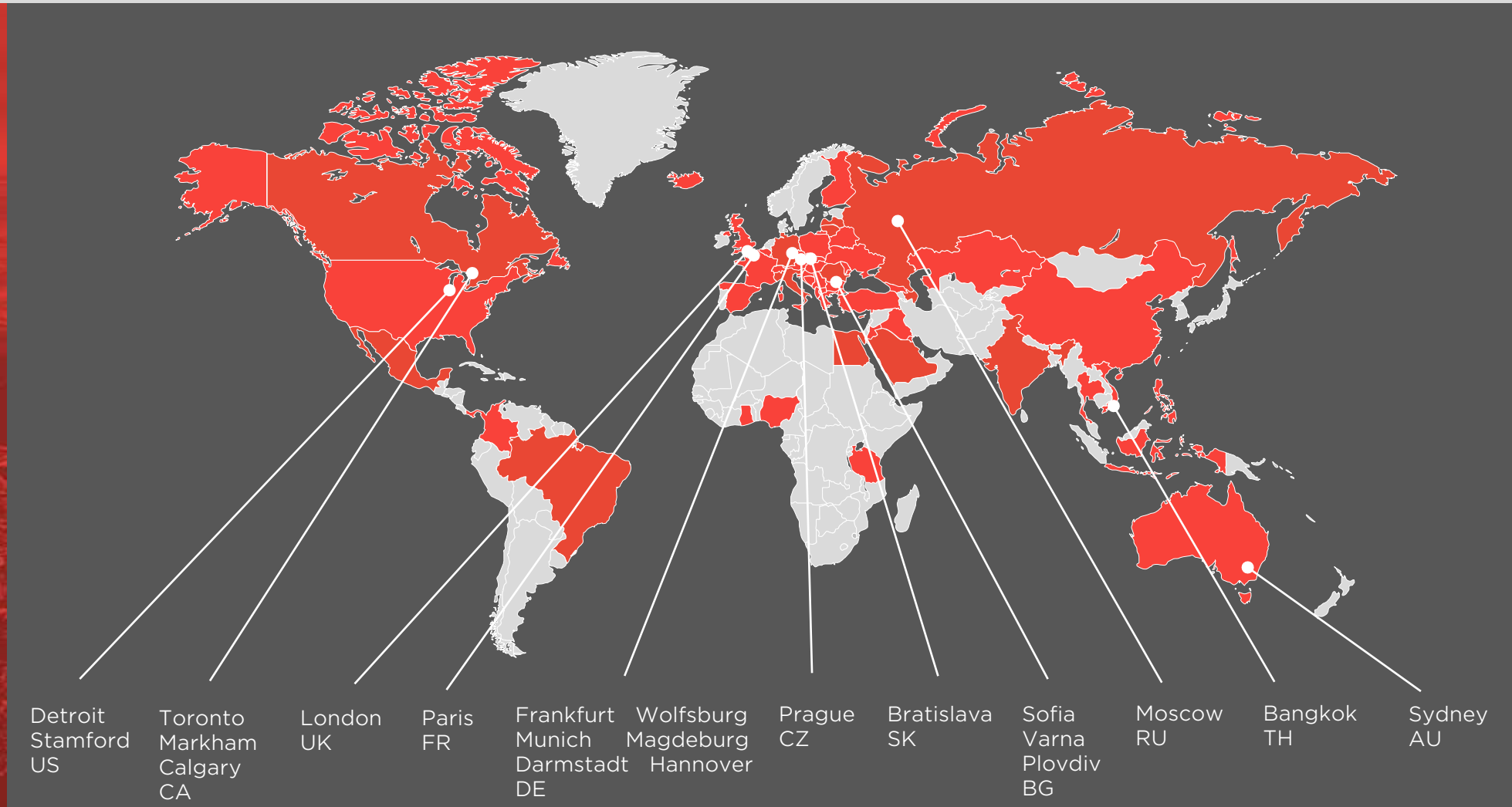
**Více než 2000**
**profesionálů**

**370+**

**370+ projektů**
**ve více než 46 zemí**

**21**

**21 poboček**
**v 11 zemích**



# Několik naších oblíbených referencí

## Banking

ČSOB · KB · Equa bank · MONETA Money Bank · Česká spořitelna

## Finance

BNP Paribas · PPF · ČSOB Pojišťovna · Kooperativa Vienna Insurance Group · mastercard

## Manufacturing & Retail

TESCO · ŠKODA · OMV · Tipsport · IKEA · Mountfield

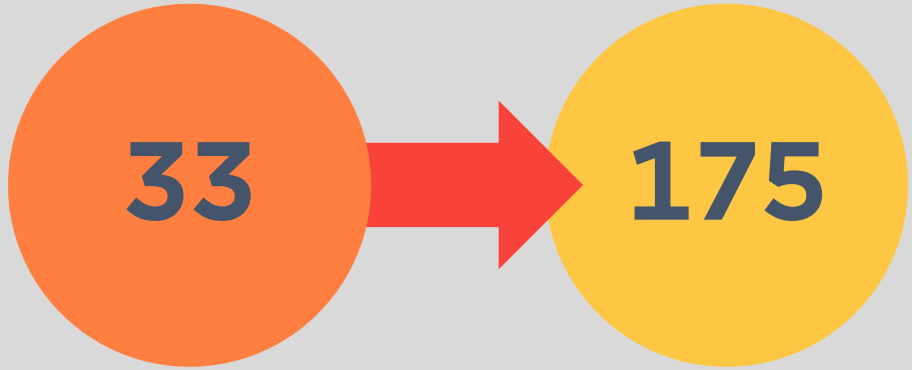
## Government

Český statistický úřad · Celní správa České republiky · Generální finanční ředitelství · ČSSZ · Česká televize

## Telecoms & Utilities

vodafone · O2 · T · orange · E · upc · telenor

# Data Boom




**The Global Datasphere will grow from 33 Zettabytes in 2018 to 175 Zettabytes by 2025**

Note: Zetta is 21 zeros (000 000 000 000 000 000 000).

Source: Data Age 2025 by IDC

# Stručná historie Data Warehousingu a Business Intelligence

## Pravěk

Kontrolovaný chaos
Zrod Best Practice
Manualní scripting
Primitivní relační analytika

1985 - 1995

## Antika

Titani: Kimball vs. Inmon
Dospívání Best practices
Enterprise Data Warehouse
ETL
OLAP
Reference Data Management
Klasická relační analytika

1995 – 2005

## Středověk

**Klasický Data Warehouse**
Evoluce Hub-and-Spoke
Data Governance
Master Data Management
Metadata-Driven Development
ELT
Data Vault
Data Mining
DW Appliance
Columnar DB
In-memory DB
Úsvit Hadoopu
Analytika nestrukturovaných dat

2005 - 2015

## Novověk

Cloud
Automatizace
Logical Data Warehouse
Extended Data Warehouse
Data Lake
Polyglot Architecture
Kappa
Lambda
Databus
Data Pipeline
Real-time
Big Data ETL
Open Source Analytics
Big Data Analytics
Self-service BI & ETL
Data Science
Machine Learning & AI
Hadoop bez Hadoopu
Stream Analytics
Data Management Platform
Autonomous Technologies
Compute & Storage Model
Serverless

2015 - 2025

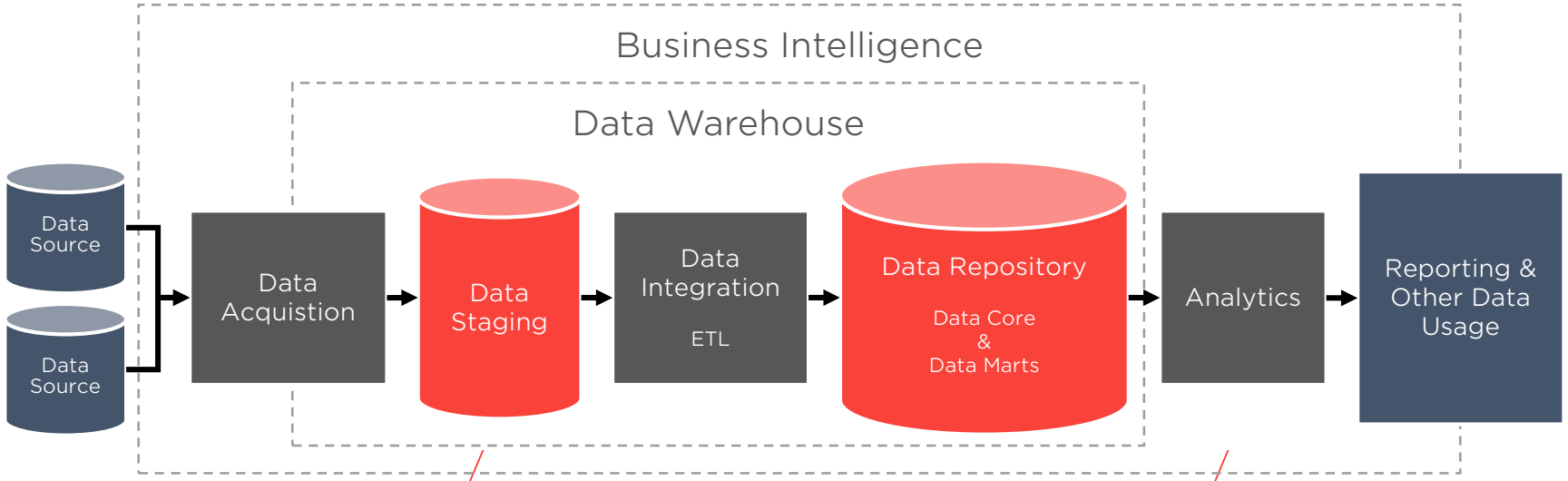
## Future?

2025 - ∞

# Jak vypadá klasický datový sklad (DW)?

- Klíčová datová platforma posledních desetiletí
- Datový systém používaný pro reporting a datové analýzy. Klíčová komponenta Business Intelligence. Integruje data z jednoho nebo více zdrojů
- Velké množina dat, kterou společnost ukládá s spravuje pro podporu rozhodování
- Starý ale velmi vyspělý koncept v nových podmínkách

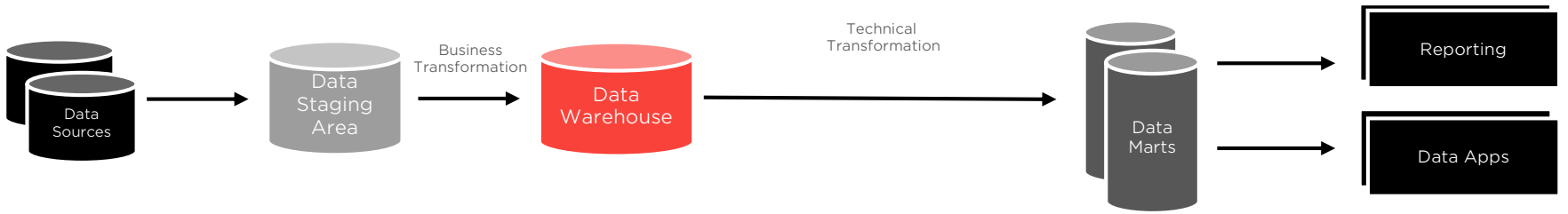
- Základní vlastnosti
  - Database (typicky RDBMS)
  - Subject Orientation
  - Data Integration
  - Historie
  - Stabilní struktura
- Dávkové zpracování
- Významné zpoždění mezi vznikem dat a jejich dostupností pro analytiku
- DW, DWH, MIS, ADS, ADW, EDW, DP



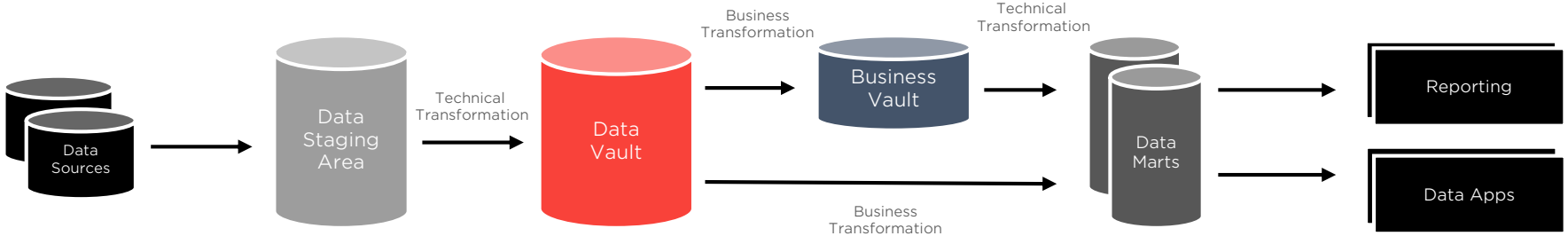
# Základní varianty klasického DW (HUB-AND-SPOKE)

## Ralph Kimball
Data Warehouse Bus (DW)
Bottom-Up

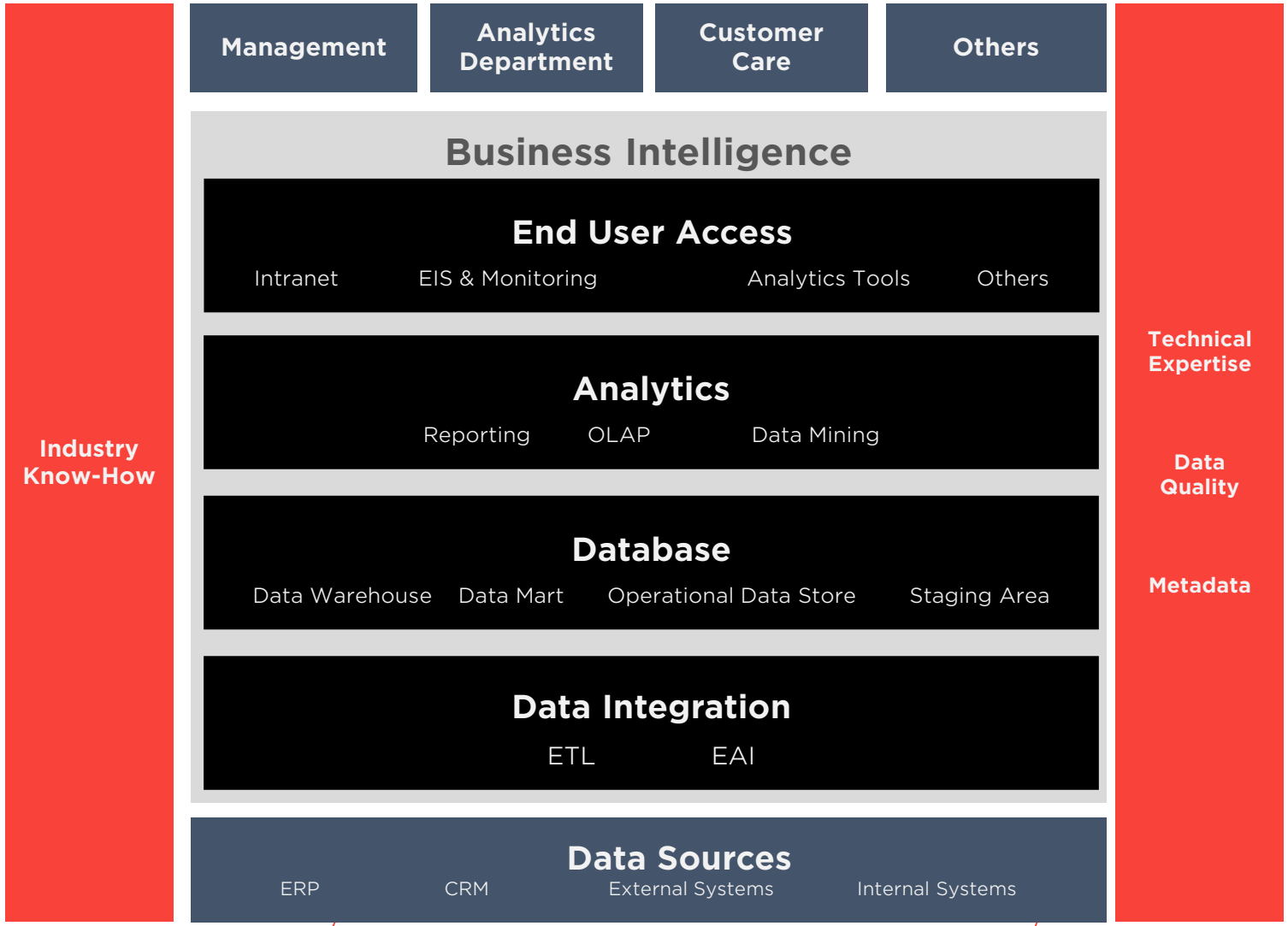

## Bill Inmon
Enterprise Data Warehouse (EDW)
Top-Down




## Dan Linstedt
Data Vault (DV)
Top-Down



# Adastra IM Reference Architecture 2005



# Proč klasický datový sklad nestačí?

## Potřeby Businessu

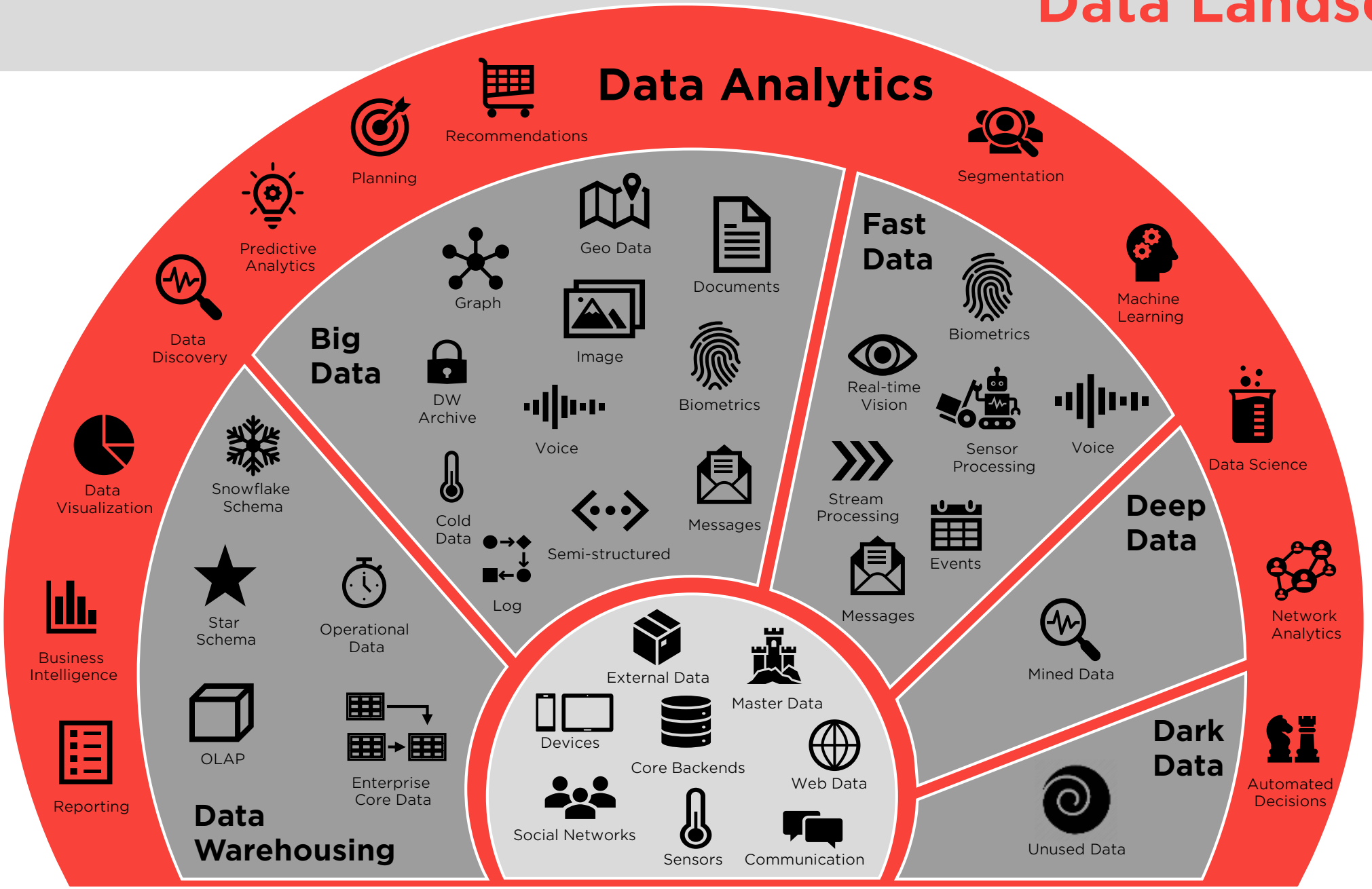
- Důvěryhodná data (klidně i raw data)
- Rostoucí tržby a zisk
- Kvalitnější Customer Experience
- Lepší produkty a služby
- Pohled 360 stupňů na business aktivity
- Digitální transformace
- Agilita
- Automatizace
- Rychlejší reakce na nové obchodní výzvy
- Rozhodování v reálném čase řízené daty
- Kvalitní a rychlá predikce blízké budoucnosti
- Chytřejší a inteligentnější business
- Monitorování externího prostředí

VS.

## Problémy klasického datového skladu

- Přílišný důraz na dokonalost (single version of truth)
- Pouze strukturovaná statická data
- Dávkové zpracování a vysoká datová latence
- Utavení rostoucím objemem dat
- Poptávka Businessu překračuje kapacitu i rozpočet IT
- Neúplná datová základna (pouze strukturovaná data)
- Rostoucí provozní režie
- Chybí real-time pohled na data
- Nelze snadno škálovat
- Omezená pokročilá analytika
- Vysoké TCO často spojené s vendor lock-in
- Zastaralá governance and security
- Malý důraz na self-service
- Obvykle nekompatibilní s DataOps
- Relační datové zdroje místo datových objektů
- Technologie včerejška řeší problémy dneška

# Data Landscape



# Nahradí klasický datový sklad třeba Data Lake?

**James Dixon, CTO, Pentaho, 14. 10. 2010**

„If you think of a datamart as a store of bottled water – cleansed and packaged and structured for easy consumption – the data lake is a large body of water in a more natural state. The contents of the data lake stream in from a source to fill the lake, and various users of the lake can come to examine, dive in, or take samples.“

Link

Rychle a snadno postavitelný

Lze zpřístupnit všechna data

Velmi náročný v dlouhodobém horizontu

Může se změnit v datovou bažinu

Koncept datové technologie, nikoliv konkrétní technologie

Masivní snadno dostupné datové uložiště  postavené na Big Data technologiích

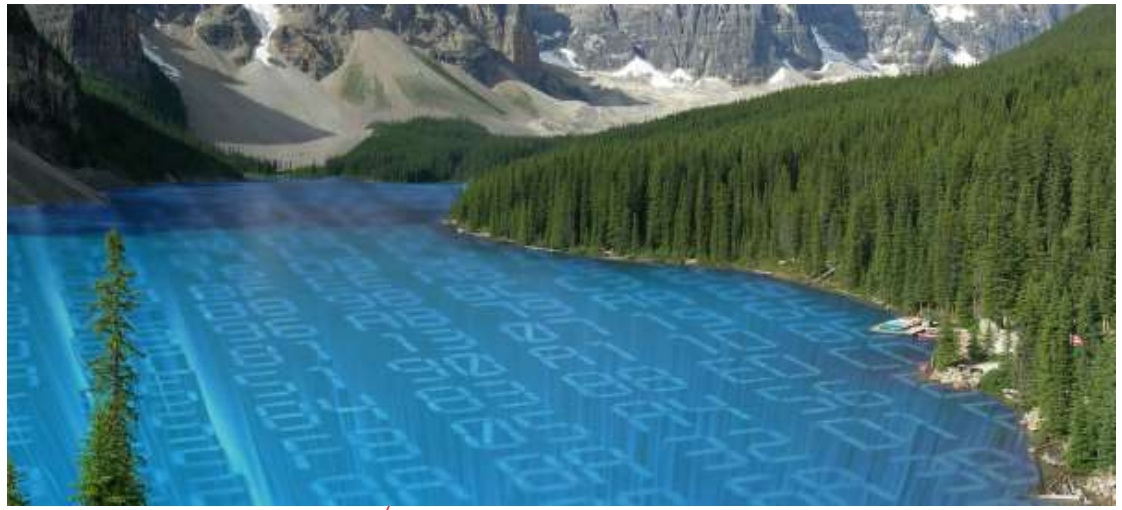
Ukládání všech dat v původní formě pro pozdější další využití

Snadná a rychlá dostupnost pro všechny

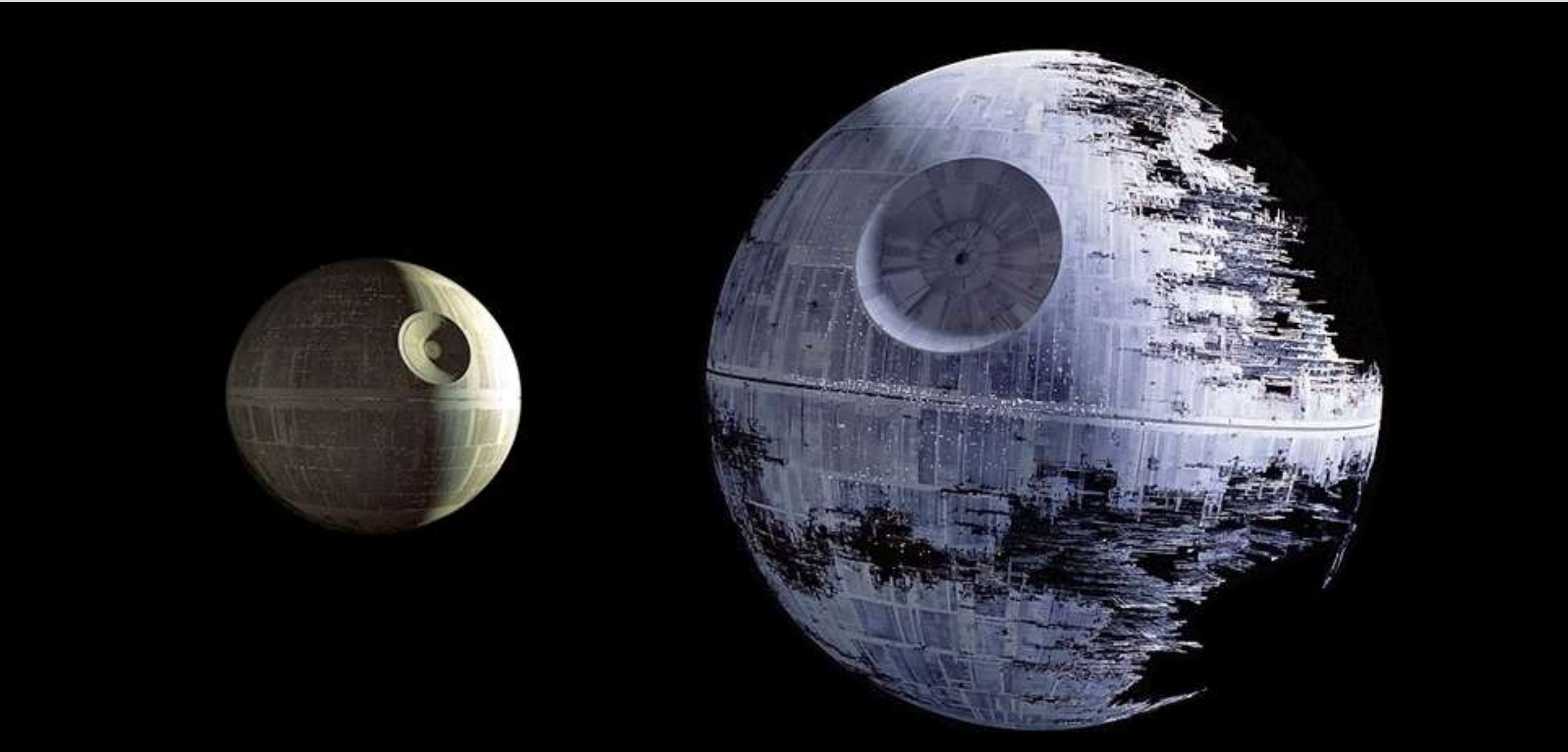
„Brute force“ eliminace datových sil

Velmi kontroverzní téma i poli Big Dat

Vyžaduje velmi kvalifikované uživatele



# Sklad vs. Lake: Různé technologie ale stejný výsledek



# DWH vs. Data Lake vs. Moderní datové platformy

| | Traditional Data Warehouse (DW) | Data Lake (DL) | Extended Data Warehouse (XDW) / Lambda / Kappa |
|---|---|---|---|
| **Data** | Structured | Structured & Semi-Structured & Unstructured | Structured & Semi-Structured & Unstructured |
| **Data Processing** | Processed | Raw | Processed & Raw |
| **Data Streaming** | Data At-rest | Data At-rest | Data At-rest & Data In-motion |
| **Data Schema** | Schema-on-write | Schema-on-read | Schema-on-write & Schema-on-read |
| **Data Model** | Relational | Object-based | Relational & Object-based |
| **Data History** | Hierarchically archived | No hierarchy | Hierarchically archived & No hierarchy |
| **Agility** | Fixed configuration | Reconfigured anytime as needed | Fixed configuration<br>Reconfigured anytime as needed |
| **Security** | Mature | Maturing | Mature |
| **Primary Users** | Data analysists &<br>Business professionals | Data Scientists | Data analysists & Business professionals &<br>Data scientists |
| **Technology** | RDBMS | NoSQL DBMS<br>Hadoop<br>Other distributed storages | RDBMS<br>NoSQL DBMS<br>Hadoop<br>Other distributed storages |
| **Agility** | Low | High | Medium |
| **Added Value** | Medium | Medium | High |
| **Cost** | High | Low, but later high | Medium |

# Tajná přísada: Data Governance

Data Governance is a collection of practices and processes which help to ensure the formal management of data assets within an organization.

## Why

### Concepts

- Vision & Mission & Strategy
- Guiding Principles
- Organization & Roles
- Business Rules
- Activities
- Scope
- Benefits & Goals

## What

### Components

- Data Architecture
- Data Quality
- Data Integration
- Operations
- Security
- RDM & MDM
- Metadata
- Data Platform & BI

## How

### Tools

- CASE
- Enteprise Metadata Repository
- Data Quality Tools
- QA Framework
- Workflow & Orchestration
- IDE
- Audit Log
- Resource Management
- RDBMS
- NoSQL
- Hadoop
- Integration tools
- Monitoring
- Source Code Repository
- Testing Tools
- Others

# Moderní datová platforma musí umět hodně věcí

| Data Ingest | Data Integration | Data Management | Architecture | Data Model | Database | Data Repository | Data Usage | Governance | Deployment |
|---|---|---|---|---|---|---|---|---|---|
| Data Loading | ETL | Key Management | Hub & Spoke | E-R Model | RDBMS | Data Warehouse | Business Intelligence | Data Governance | On-premise |
| Data Replication | Cleansing | Time Variance | Polyglot | Multidimensional Model | Multidimensional DB | Data Mart | Reporting | Data Catalog | Cloud |
| Change Data Capture | Aggregation | Data Latency | Lambda | Star Schema | Graph DBMS | Sandbox | Data Discovery | Metadata | Hybrid Cloud |
| Stream Processing | Standardization | Audit | Kappa / Databus | Snowflake Schema | Columnar DBMS | Data Lake | Data Science | Reference Data Management | Multi-Cloud |
| Manual Inputs | Reconciliation | Date Tiering | Big Data Fabric | Document Store | Object Store | Operational Data Store | Machine Learning | Master Data Management | |
| Data API | Orchestration | Data Retention | | Graph Data Model | In-memory | Data Archive | Data Quering | Data Literacy | |
| Legacy | Automation | Data Security | | Key-Value | NoSQL | Master Data Repository | | TCO Management | |
| | Data Pipelines | | | Column Family | Distributed File System | File Repository | | | |

ANY DATA
ANY LATENCY
ANY FORMAT
ANY HISTORY
ANY GRANULARITY
ANY CONSOLIDATION
ANY ANALYTICS
ANYWHERE
ANYTIME

# Evoluce datové architektury

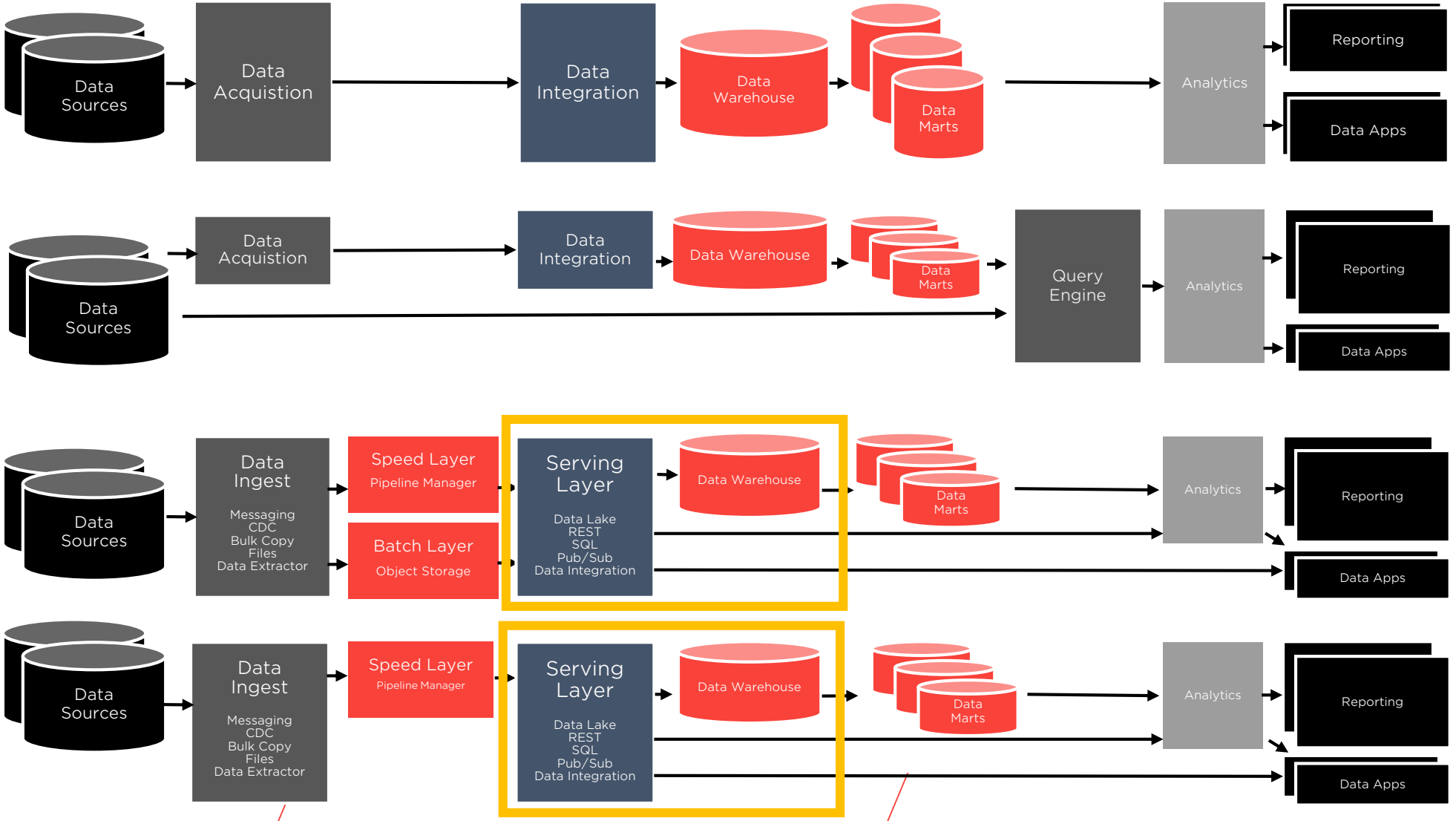
## Hub-and-Spoke
Data Warehouse

## Polyglot
Data Federation
Data Virtualization
Logical Data Warehouse

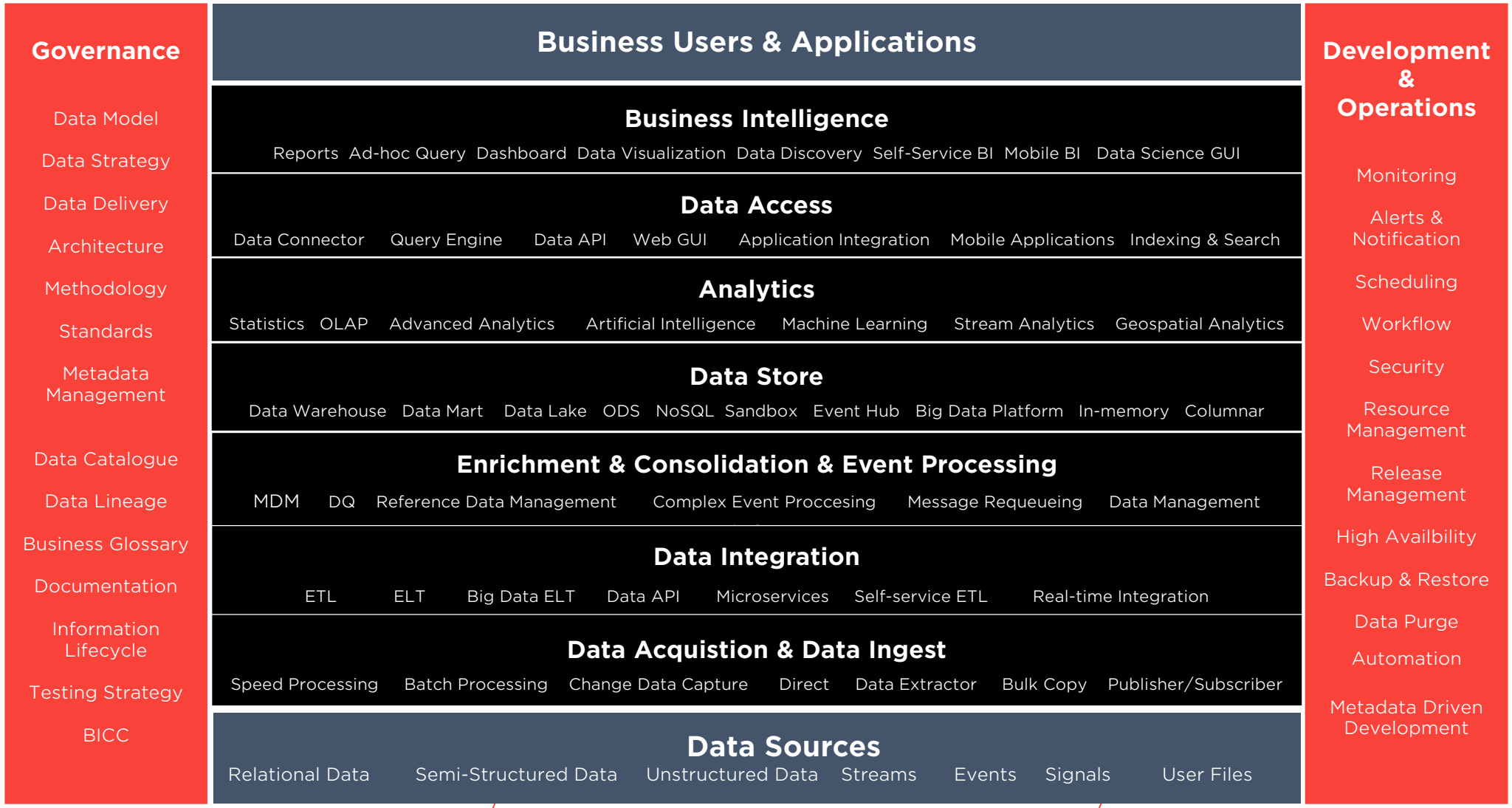
## Lambda

## Kappa
Databus

EVOLUCE



# Adastra IM Reference Architecture 2019



# Shrnutí soumraku klasických datových skladů


## Klasický datový sklad (Hub-and-Spoke)

- Minulost (Data Warehouse zombie)
- Konsolidovaná relační data at-rest
- Dávkové zpracování, vysoká latence dat
- Nezvládá opravdu velké objemy dat
- Omezení možnosti analytiky / Nízká agilita
- Obvykle On-premise / Komerční software

REST IN PEACE

## Extended Data Warehouse / Lambda / Kappa

- Budoucnost (datový sklad jen jedna z komponent)
- Libovolná data at-rest & data in-motion (raw & konsolidace)
- Různé frekvence zpracování dle potřeby včetně real-time
- Škálovatelnost / Kontejnerizace / Oddělené Computing a Storage
- Pokročilá analytika / DataOps / Self-service / Vysoká agilita
- Cloud / Serverless / Komerční software + Open Source

# Nemáte občas z Vašeho klasického datové skladu stejný pocit?☺

# Kontakty

**Martin Bém**

martin.bem@adastragrp.com

+420 603 505 247

https://www.linkedin.com/in/martinbem/

**Adastra Czech Republic**

Karolinská 654/2

186 00 Praha 8

www.adastra.cz

infocz@adastragrp.com