solutia
Idea & Solution Together

intel ORACLE COMPUTERWORLD

Big Data

Martin Stufi

Zaměření	**Big Data, IoT, Cloud, Integrace, Architektura IS**

Certifikace

- Certified Big Data Hadoop Developer
- PRINCE2 Foundation
- TOGAF 9 Certified
- ITIL v3
- Další

Martin Stufi

Zkušenosti

- 15+ let
- Solution Architekt
- Program Manažer
- Vývoj SW
- Bezpečnost

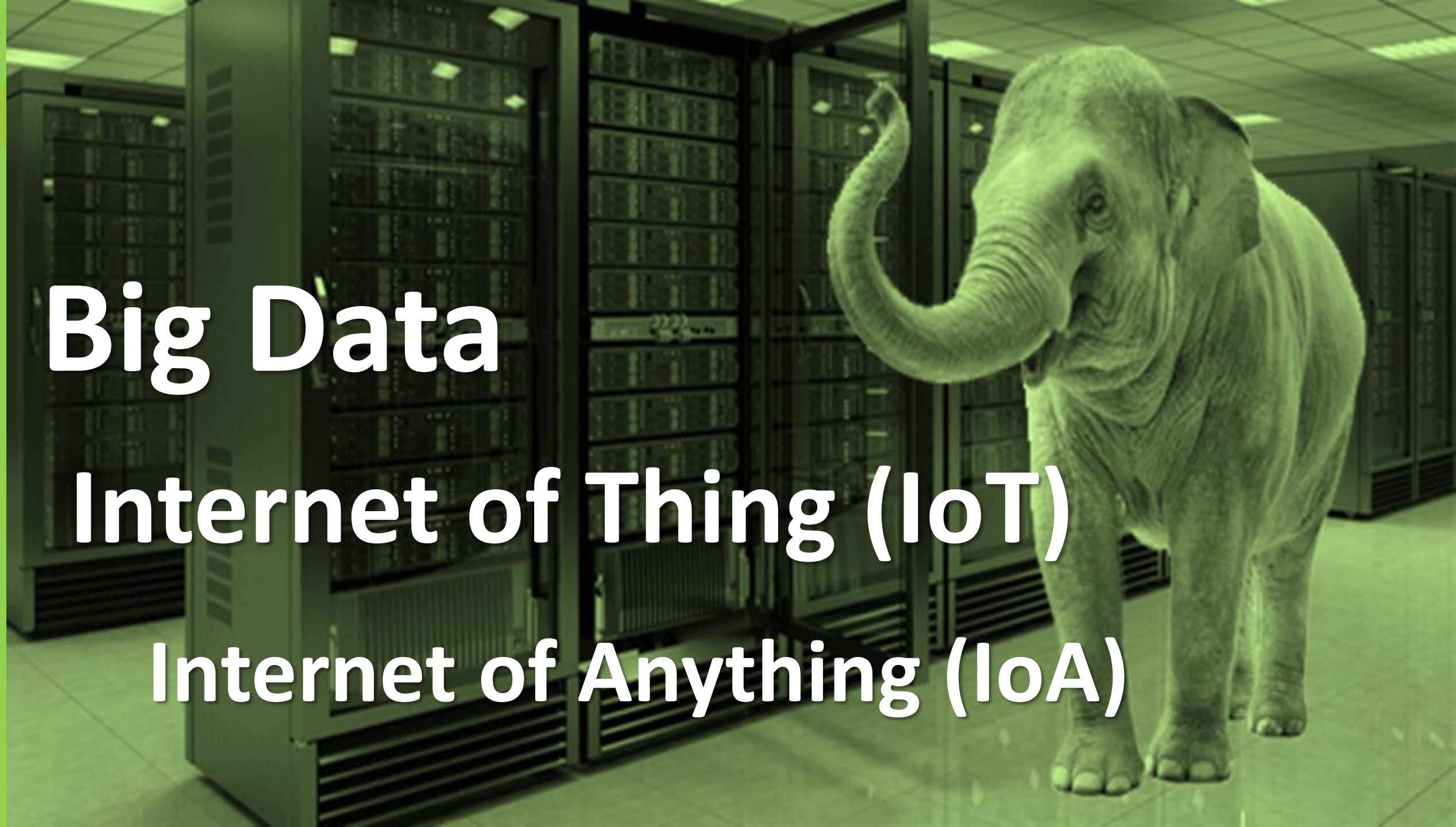
Pozice

- CEO Solutia, s.r.o.
- Student PhD se zaměřením na Big Data

Zaměření

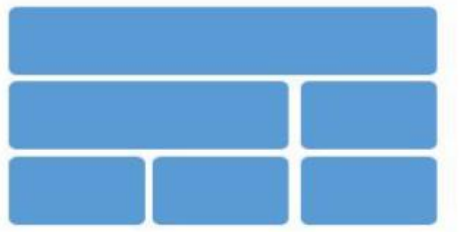
- Big Data Architektura
- Vývoj Big Data aplikací
- Hadoop, Pig Latin, Hive, Java, Scala, R
- Map Reduce, YARN, Spark

https://www.linkedin.com/in/martinstufi martin.stufi@solutia.cz @stufim


Big Data
Internet of Thing (IoT)
Internet of Anything (IoA)

Big Data

- Stejně jako Cloud, Big Data je matoucí pojem pro většinu lidí. I když by to mohlo být další velká věc, „módní slovo ve své kůži", není jednoduché dát definitivní odpověď na to, co to je? Pokud jste náhodou zmínil o **Big Data** můžete také být podrobeny otázkám jako: **"Je to nástroj, produkt?"**, **"Je zpracování velkých objemů dat pouze pro velké podniky?"**, **„Kolik to stoji?"**, **„Jaká je to technologie?"**, **„Jak se licencuje?"**, **„Čemu to slouží?"** A mnoho dalších otázek ve stejném duchu.



Strukturována data

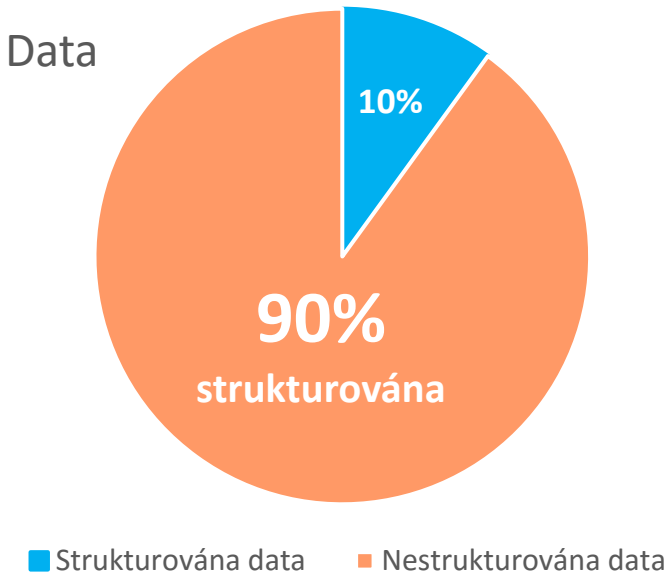
Nestrukturována data

Big Data Analýza

Konkurenční výhoda

Potřeba zpracování velkého objemu dat

- Věděli jste, že podle odhadu byla vytvořena **v posledních 3 letech** přibližně **90% světových digitálních dat**?
- Formát dat: 10% strukturovaných dat a **90% nestrukturovaných dat**
- Nestrukturována data lze obtížně analyzovat?
- Generování vs. Analýza a vizualizace dat (80% x 20%) ➔ (20% x 80%)



Zdroj: Oracle

„Internet of Anything“ (IoA) - Tsunami přichází

- Celosvětový objem dat se v minulosti **zdvojnásobil za posledních 100 let**
- Dnes, **Každé 2 roky se nám zdvojnásobí objem dat**!!!
- Záplava těchto dat je řízená „Internet of Anything“ (Internet, Mobilní zařízení, logy, geokoordinaty, „Machine data“, Sensory apod.) ➔ **Exponenciální nárůst dat**

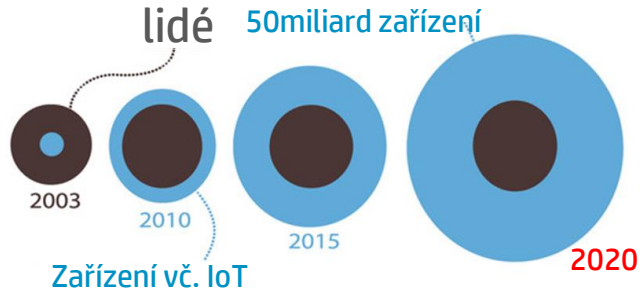
Průzkum podle Forbsu:

- 59% podniků se domnívá, že data a analytické nástroje jsou "životně důležité" na provoz svých organizací, s dalším 29% ji považovaly za "velmi důležité".
- 69% uvádí, že existuje obchodní důvod pro investice do hledání nových způsobů, jak vytvořit přidanou hodnotu prostřednictvím datových projektů.
- 83% říkají, že analytika dat stávajících služeb vytváří produkty výhodnějším.
- **Ale 48% cítí, že jejich organizace v minulosti, nedokázala využít příležitostí těžit z jejich dat.**

Internet of Things (IoT)

- Rychlost a záplava dat se rozumí, že digitální svět bude růst ze 4ZB dat na 44ZB během této dekády.
- 1.7 megabajtů nových informací bude vytvořena každou sekundu pro každého člověka na planetě, třetina z nich procházející z Cloudu(1)
- 1.5 miliard měsíčních aktivních uživatelů Facebook reprezentuje generování, 265 miliard fotografií, 62 milionů písniček hrané 22B krát
- Gartner uvádí, že 32% podniků, které podnikl digitální transformace businessu tvrdí, že jejich podniky jsou nyní „Digital Business Company(2)"

Evoluce datových platforem



Zdroje:
1.EMC
2. http://www.gartner.com/technology/research/digital-business/

"Big Data age" - Volume, Variety, Velocity, ...

Zdroj: http://www.mathepic.com/big-data/cambrian-period-big-data
Rok 2015

amazon web services™

twitter

YouTube

Chevron

NETFLIX

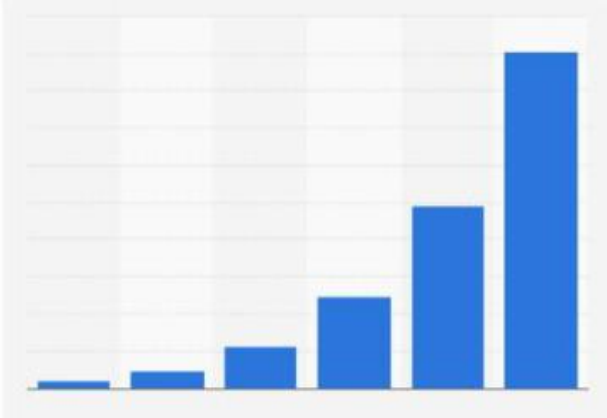
LinkedIn

J.P.Morgan

Co je Big Data?

- Podle Gartner Inc.:

"Big Data is **high-volume**, **high-velocity** and/or **high-variety** information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation."



Objem/**Volume**

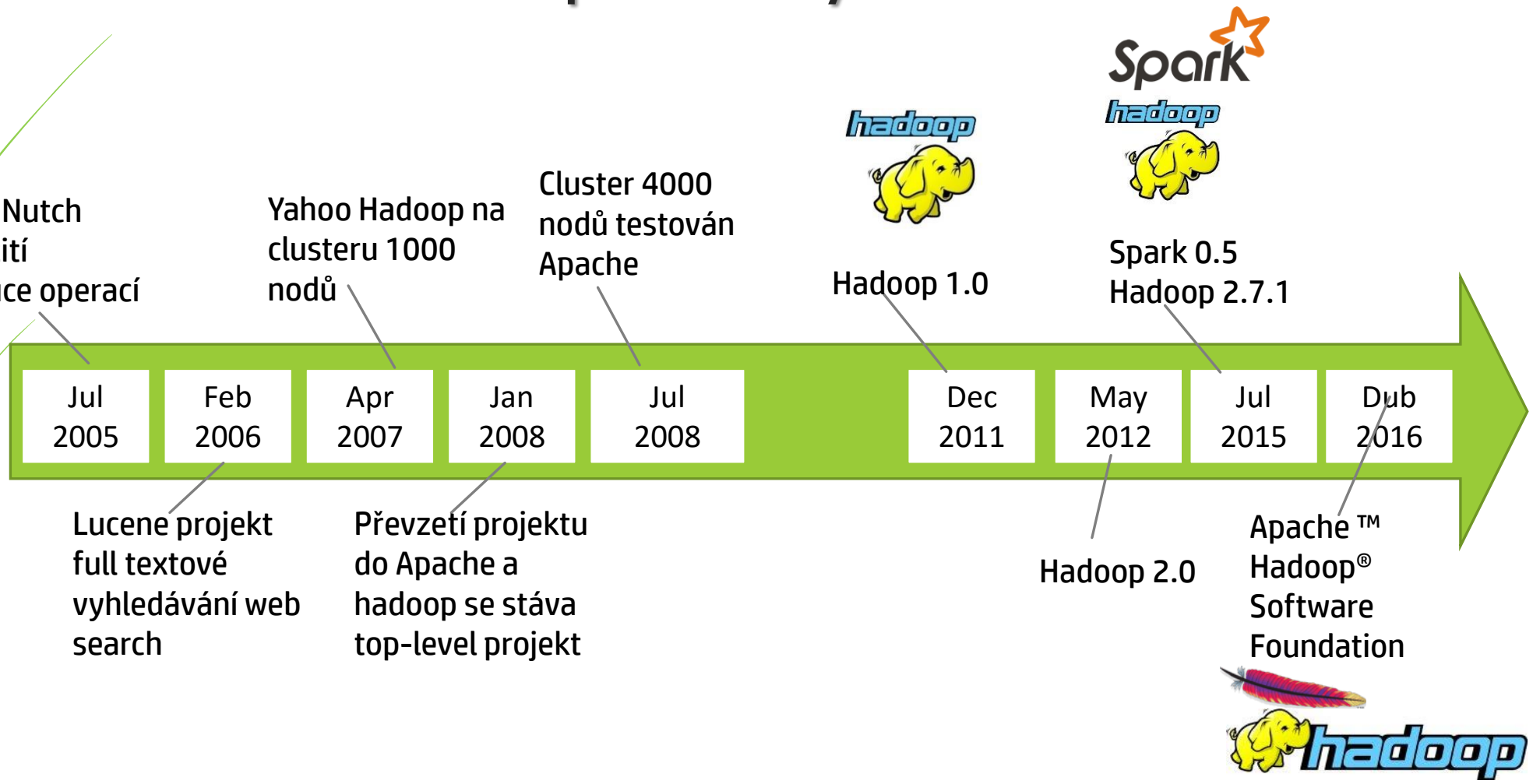
Složitost/**Variety**

Tempo růstu/**Velocity**

Charakteristiky Big Data -7V

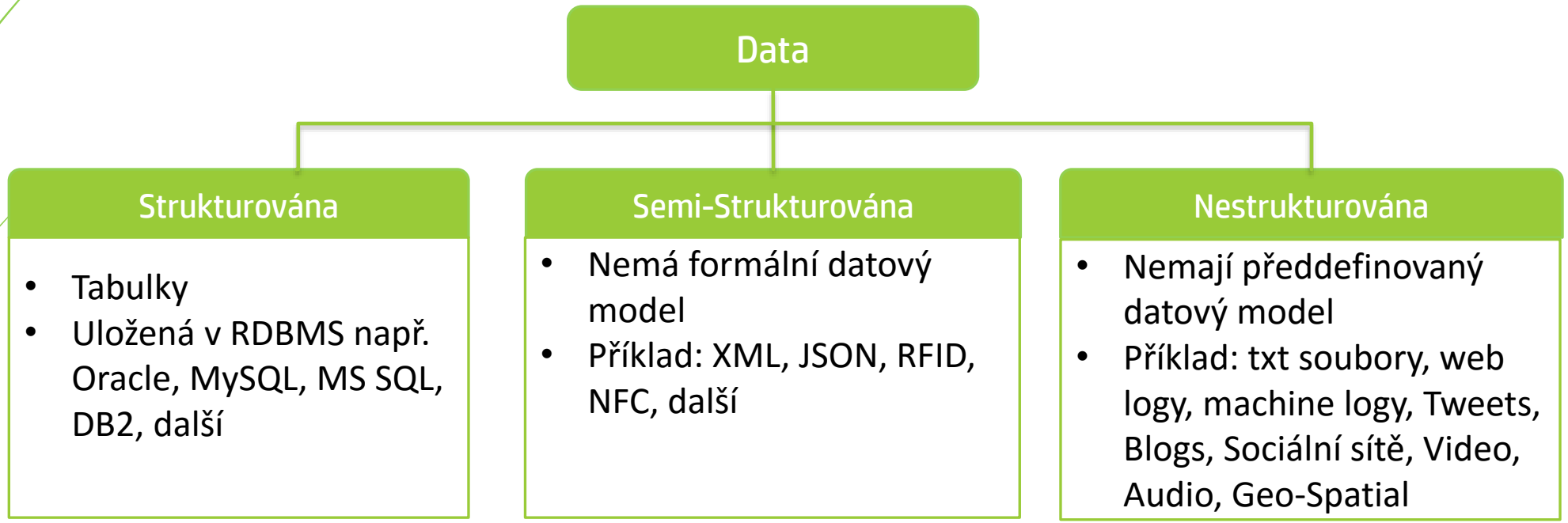
Variety	Správa komplexních dat v různých strukturách, od relačních dat do logů a neupraveného textu.
Velocity	Správa datových proudů a objemných dat s vysokou rychlostí.
Volume	Správa obrovské množství dat od TB na ZB. Data Accuracy .
Veracity	Věrohodnost řídí přesnost dat a jejich analýzu zejména v automatizovaném procesu rozhodování.
Variability	Neustále měnící se data (Big Data)
Visualisation	Generování složitých grafů, které mohou zahrnovat různé změny dat, přičemž stále data zůstávají srozumitelná a čitelná.
Value	Hodnota dat je v analýze, z dat se vytvářejí informace, následně znalostí

Historie a hadoop milníky



Typy Dat

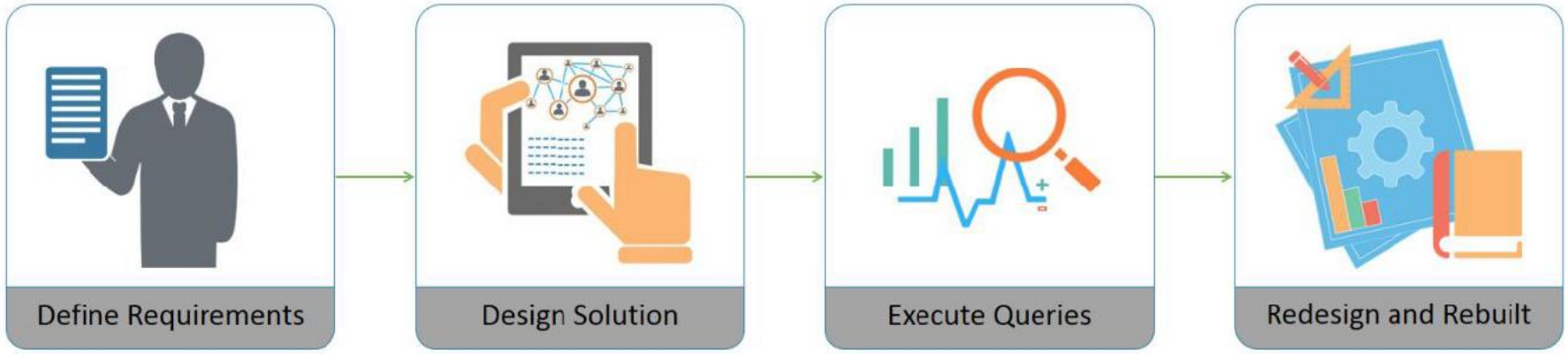
- Primární typy dat



Vývoj analytických platforem

Systém	Období	Význam
Podpora rozhodování	1970 - 1985	Systémy pro podpory rozhodování
Exekutivní podpora	1980 - 1990	Zaměřit se na analýzu dat ze strany vedoucích pracovníků
OLAP (Online Analytical Processing)	1990 – 2000	Analýza více datových tabulek (multidimensional)
Business Intelligence	1989 – 2005	Nástroje pro podporu rozhodování na základě dat, s důrazem reportovací nástroje a dashboardy
Analytické systémy	2005 – 2010	Zaměření na statistické a matematické analýzy pro rozhodování
Big Data	**2010 – aktuálně**	**Zaměření na velký objem dat, nestrukturována data a rychlé měnicí se data**

Tradiční IT Analytický systém



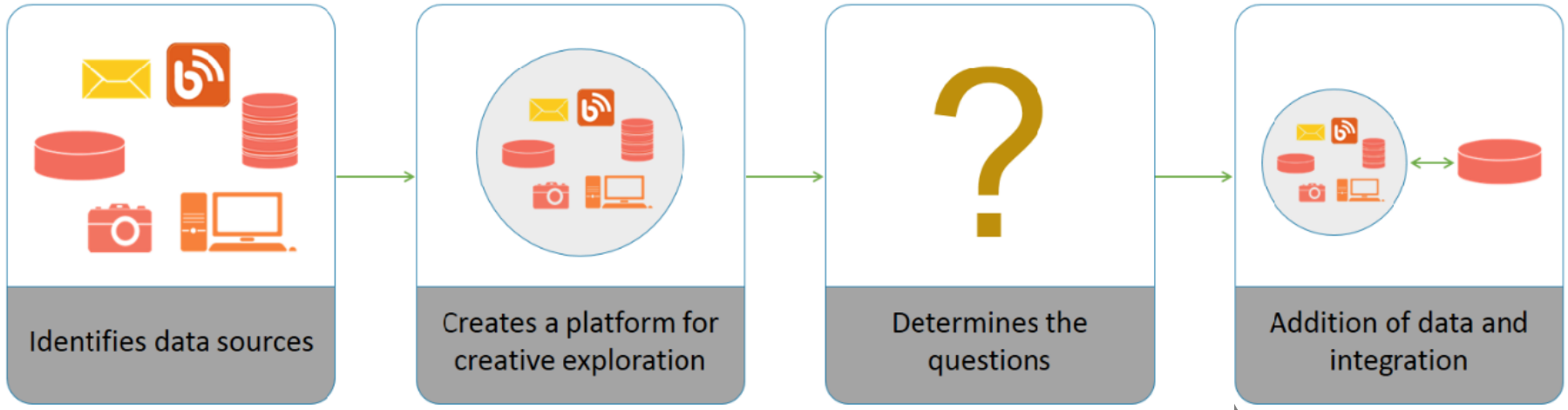
Definice požadavků

Business tým definuje otázky před IT vývojem
Business tým definuje datové zdroje a struktury dat

Výzvy

Požadavky jsou iterativní a rychlé se měnící
Zdroje dat se neustále mění

Big Data analýza

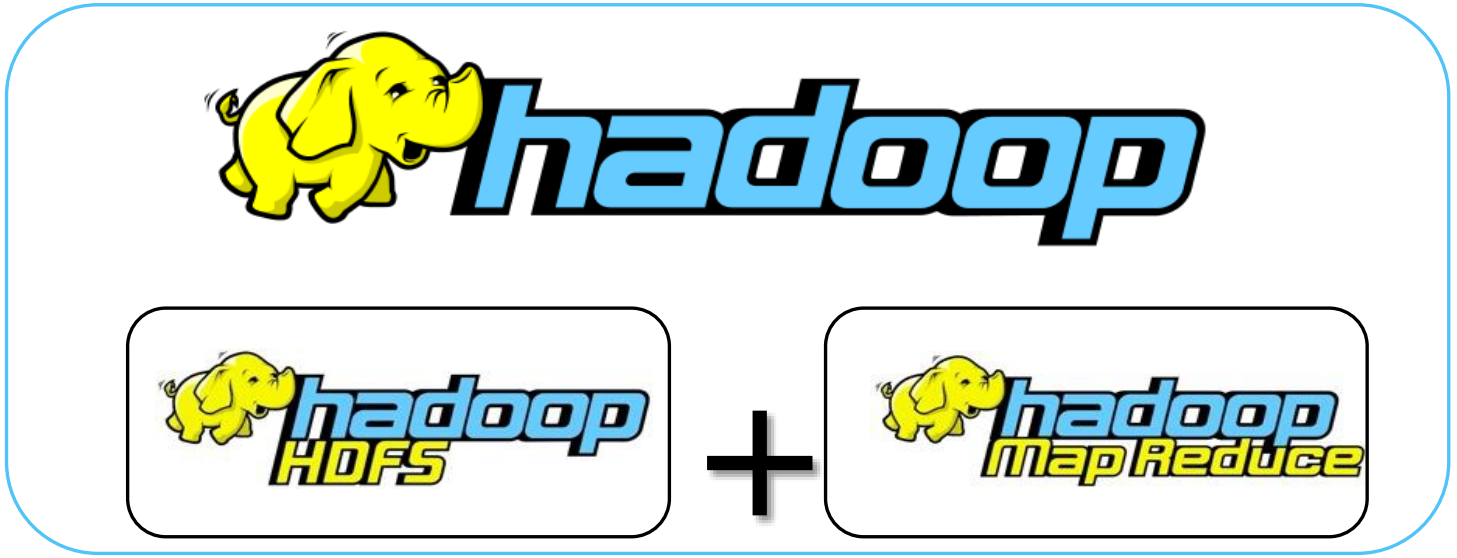


Definice požadavků	Business tým definuje datové zdroje Členové týmu definují hypotézy
Výzvy	Technologie by měla umožnit explorativní analýzu Integrace systémů a zdrojových dat dle potřeby

Hadoop Architektura

Co je Hadoop a jeho klíčové komponenty

HDFS je distribuovaný souborový systém, který poskytuje rychlý přístup k datům napříč Hadoop clusteru. Stejně jako u ostatních Hadoop souvisejících technologií, HDFS je klíčovým nástrojem pro správu jednotlivých komponent pro zpracování velkých objemů dat a podporu při zpracování v rámci analytických aplikací.



Regulární FS vs HDFS

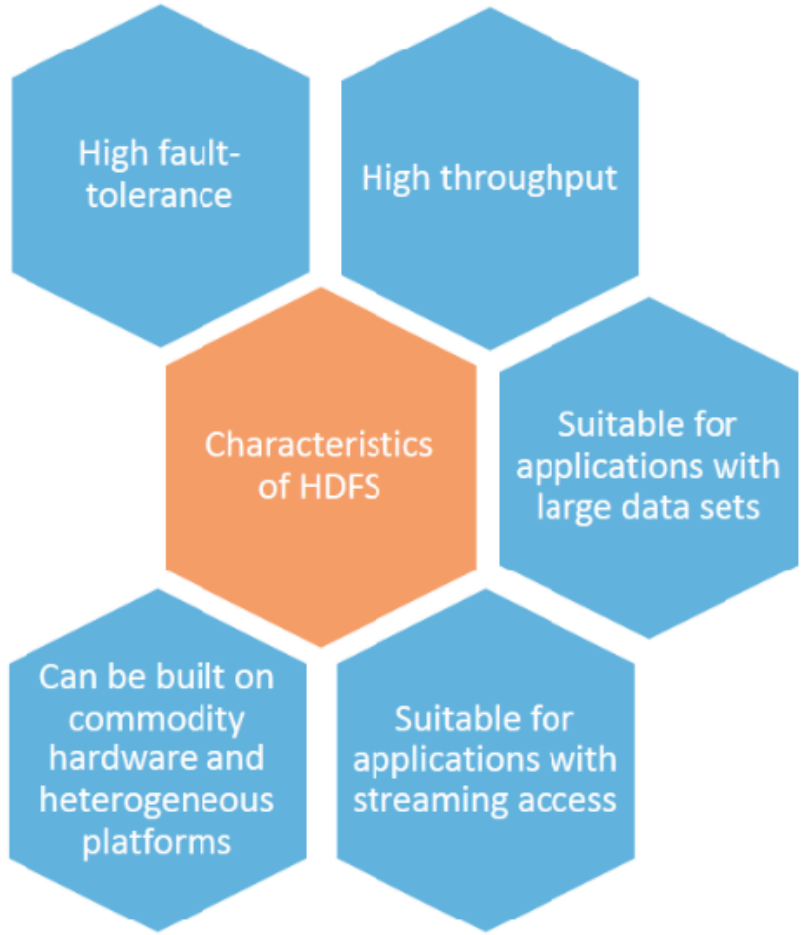
Tradiční File Systém	HDFS
Malý datový blok, přibližně 51 bajtů	Datový blok je 64MB (hadoop – verze 1.x) Datový blok je 128MB (hadoop verze 128MB)
Přístup většímu objemu dat vytváří I/O problémy (seek operation)	Čtení velkého množství dat (single seek)

Tradiční File Systém vs HDFS

Tradiční File Systém	HDFS
Malý datový blok, přibližně 51 bajtů	Datový blok je 64MB (hadoop – verze 1.x) Datový blok je 128MB (hadoop – verze 2.x)
Přístup většímu objemu dat vytváří I/O problémy (multi seek operation)	Čtení velkého množství dat (single seek)

HDFS Charakteristiky

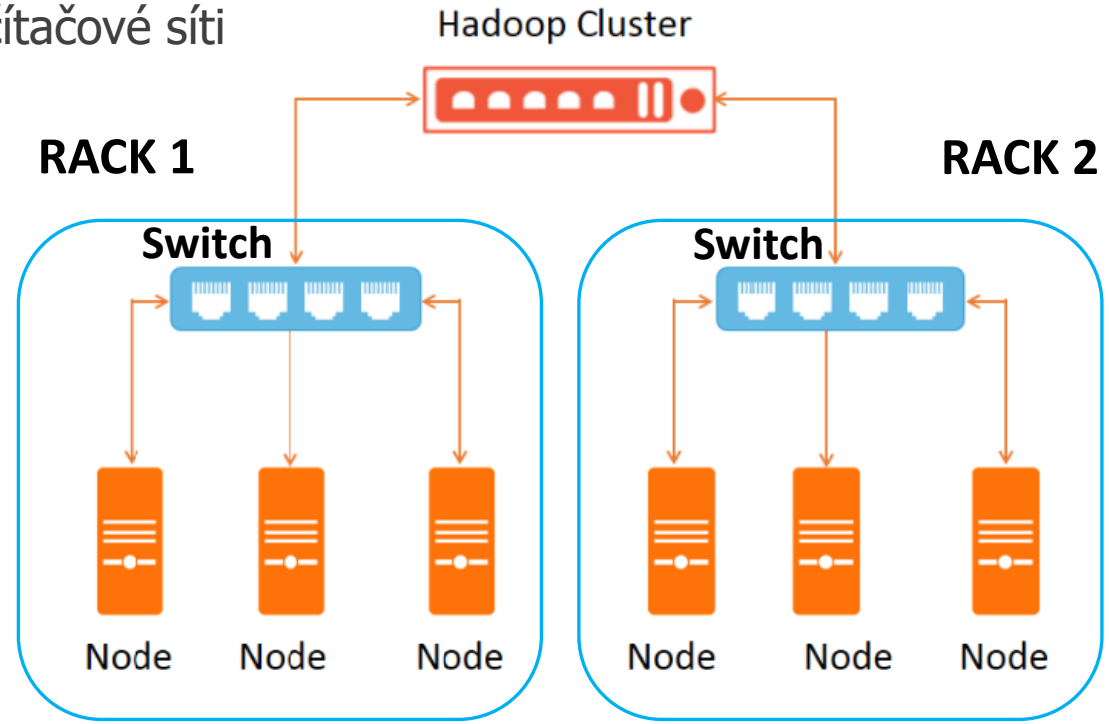
- Vysoká **dostupnost**
- Vysoká **propustnost**
- Vyhovuje aplikacím s **velkými data sety**
- Vyhovuje aplikacím využívající **streaming da**
- Lze ho vybudovat na **komoditním HW**
- Na **heterogenních** platformách



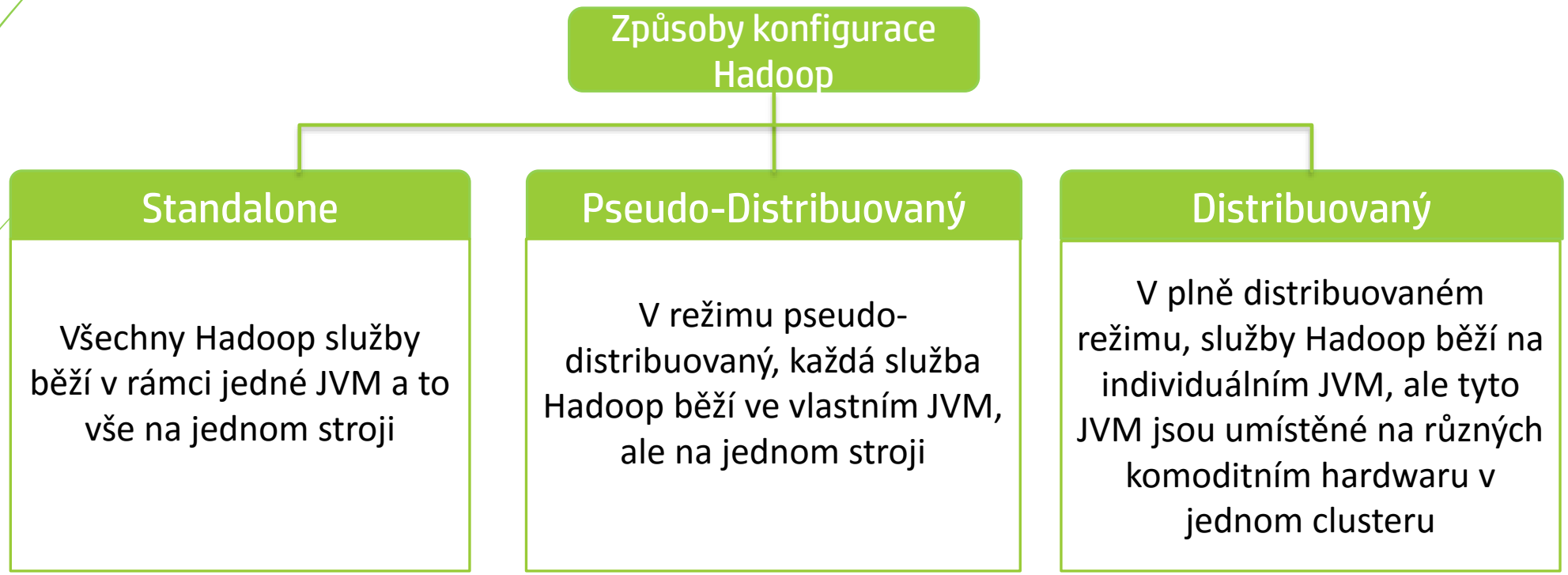
Proč Hadoop

Klíčové pojmy a Hadoop Cluster

- Komoditní HW – PC, která používáme pro tvorbu clusteru
- Cluster – Interkonece počítačového systému v počítačové síti
- Node – Komoditní server připojen v počítačové síti
- Uplink Rack ➔ Node: 3-4Gb/s
- Uplink Rack ➔ Rack: 1Gb/s
- Max. počet nodů v racku 30-40

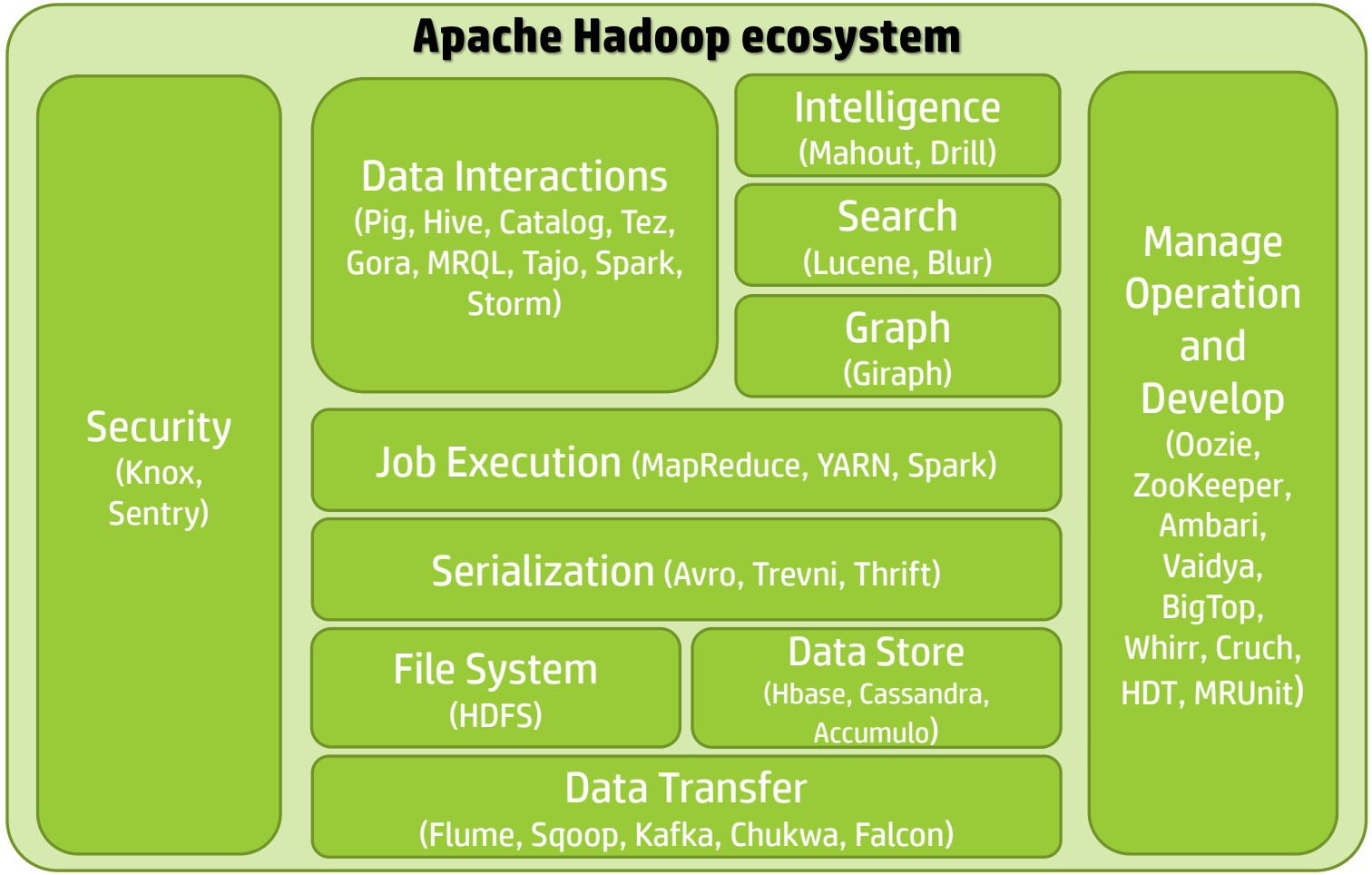


Hadoop konfigurace

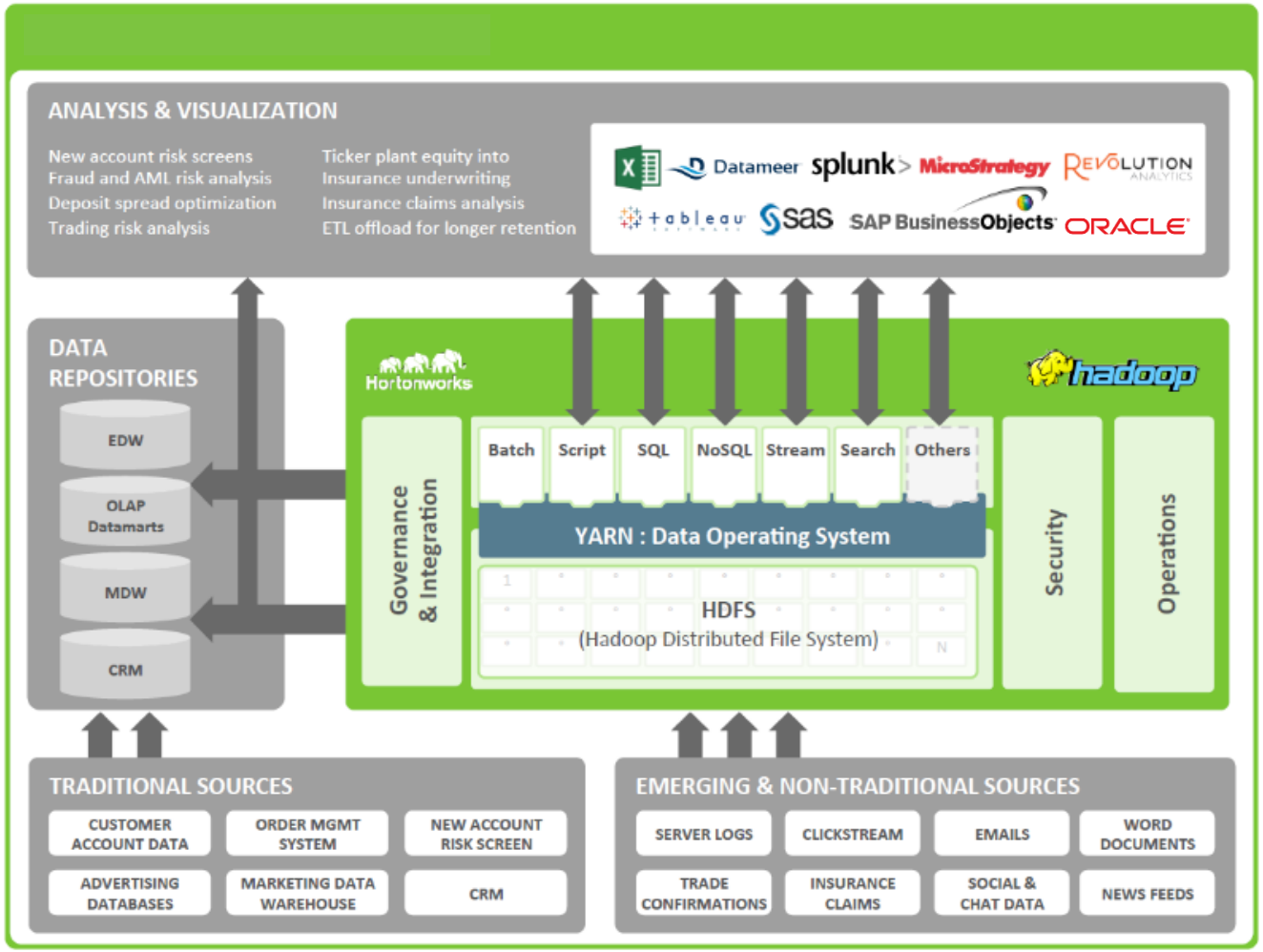


Apache Hadoop ekosystém

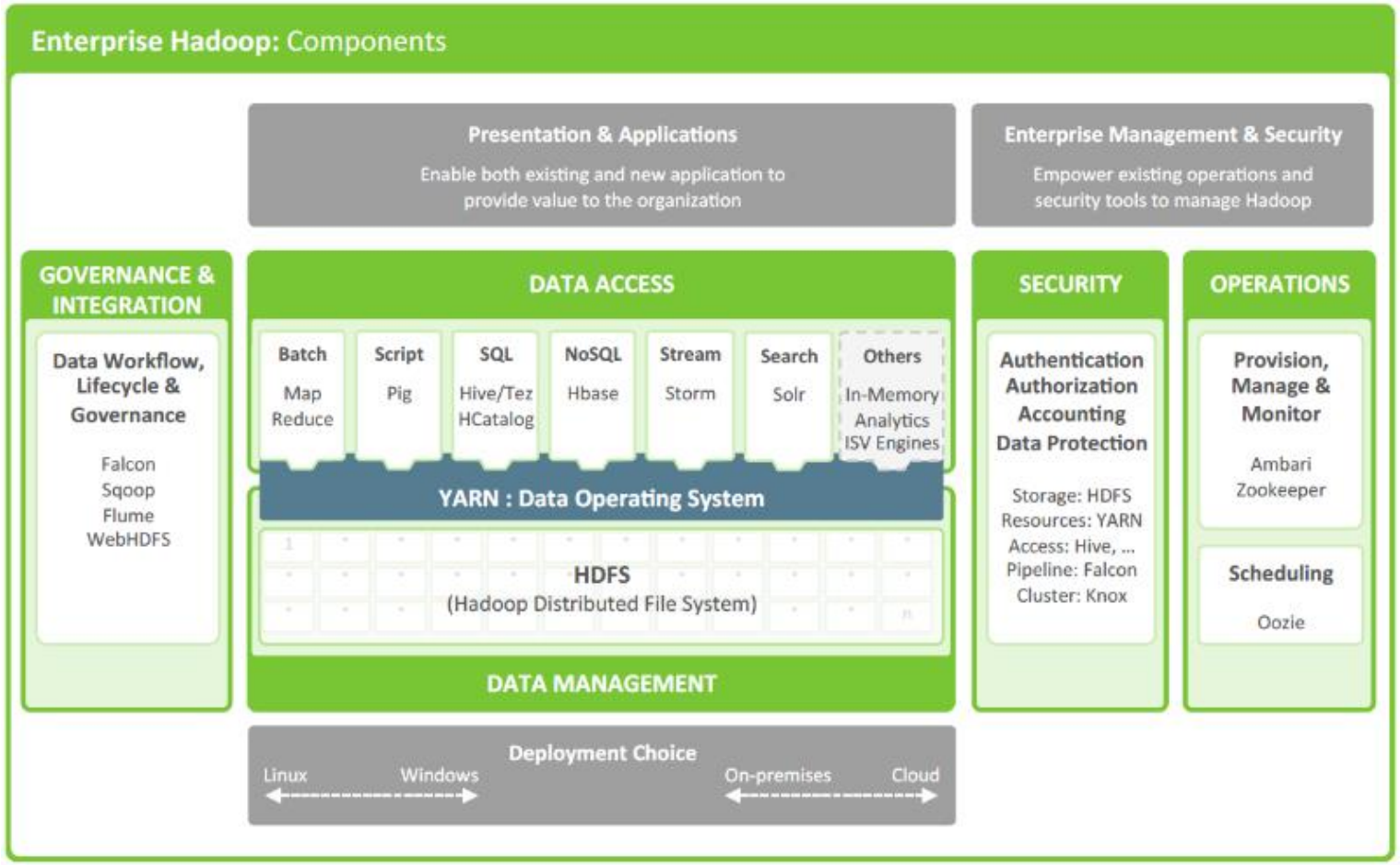
- File System
- Data Store
- Serialization
- Job Execution
- Work Management
- Developement
- Operation
- Security
- Data transfer
- Data interactions
- Analytics and Intelligence
- Search
- Graph processing



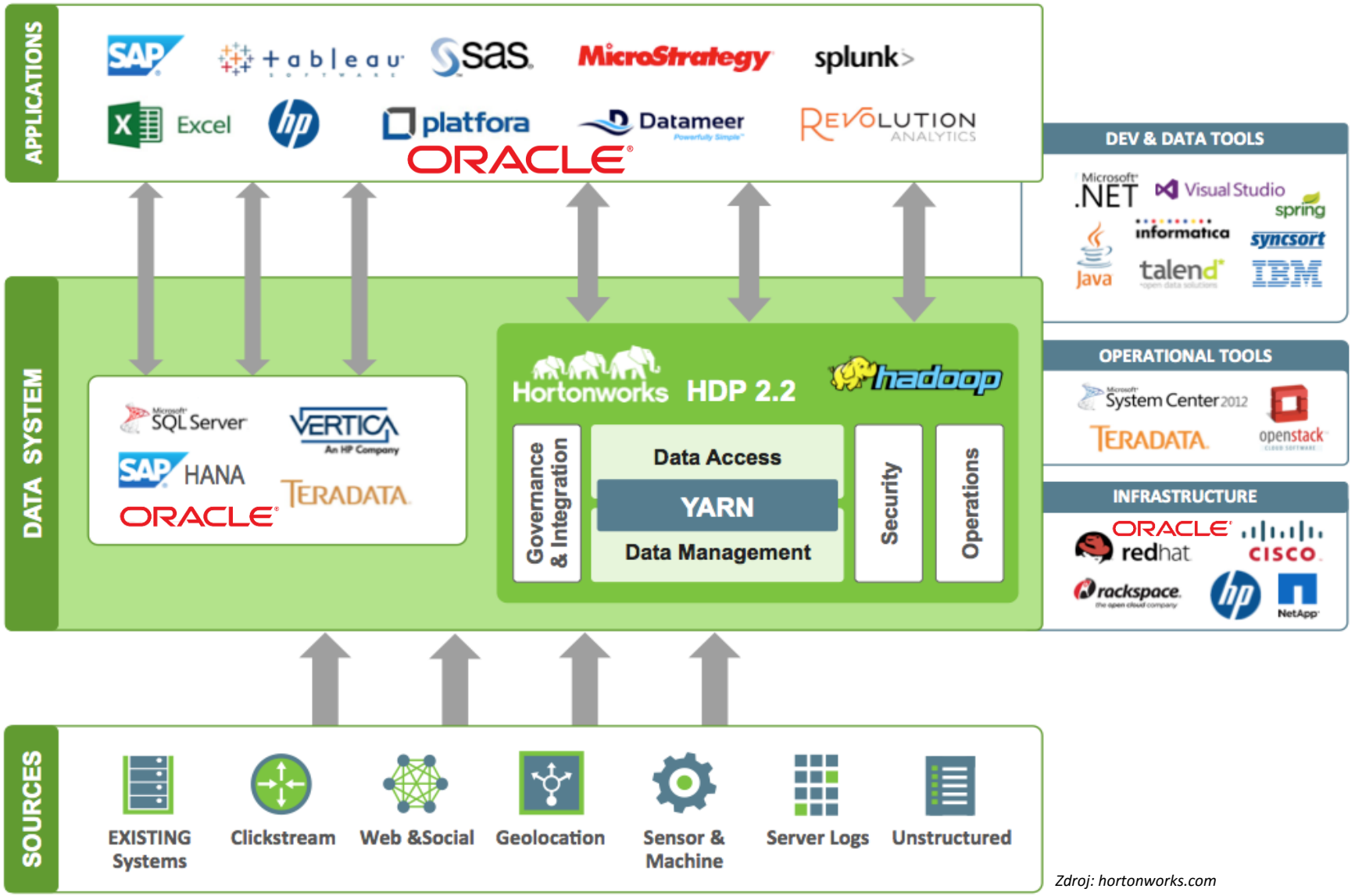
Tradiční systémy + Big Data architektura



Enterprise Hadoop komponenty



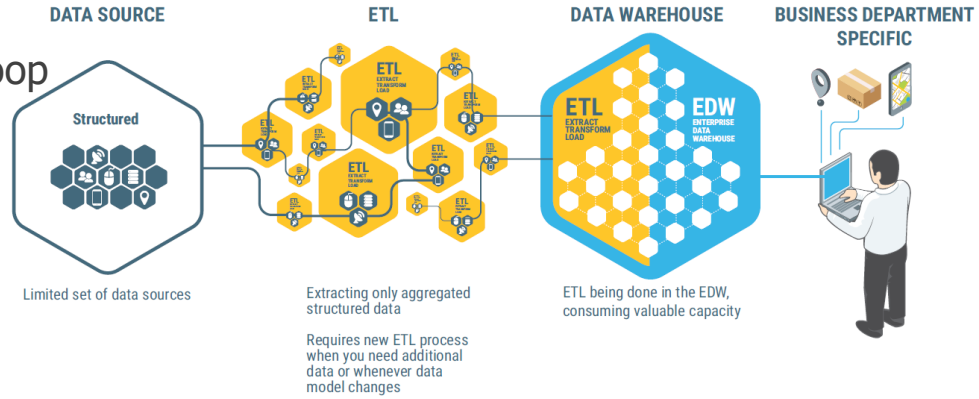
Integrace Hadoop v moderní datové arch.

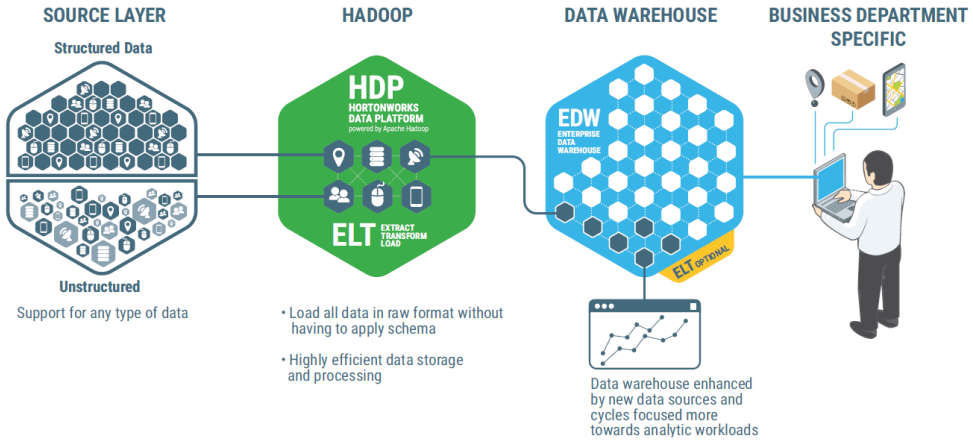


Zdroj: hortonworks.com

ETL proces před a po zavedením Hadoop

- **Archive Data**
  - Move cold or rearly used data into Hadoop
  - Active archive to Hadoop
  - Expand the amount of history that can be maintained
- **Onboard Data**
  - Implement ETL processes in Hadoop
  - More efficient ETL
  - Reduce cost data movement
- **Enrich the value of your EDW**
  - Refine new data source by Hadoop
  - Web, Machine data
  - Fuel business and expand opportunity



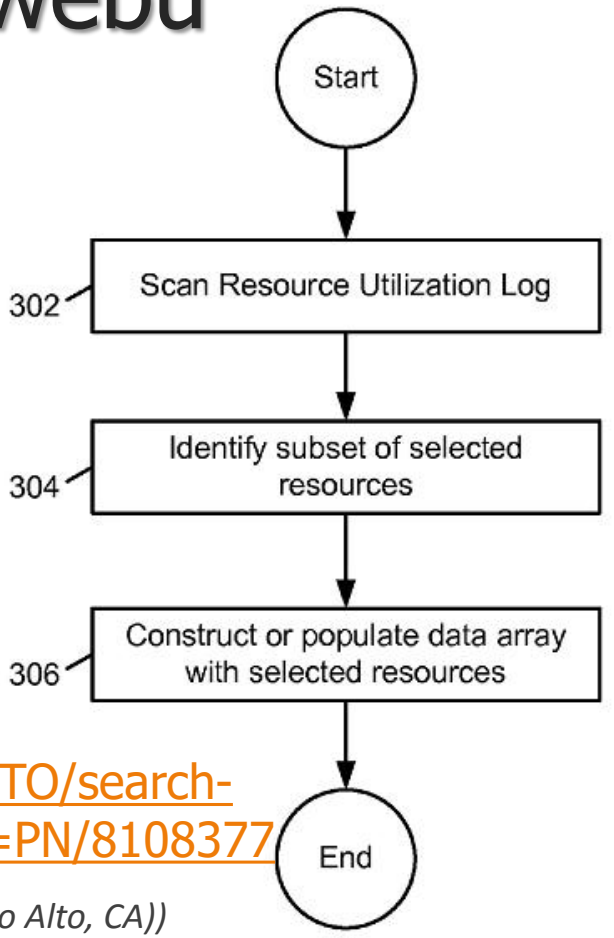


Use cases

- Po konzultaci zmínit se o pár zajímavých:

Facebook predikce obsahu na webu

- Chytré soc. sítě využívají Hadoop a Hive v řízení predikce
- Zdroje který zahrnují Java skripty, styly apod. na
  Základě Map/Reduce a dalších výpočetních algoritmů
- Jsou to distribuovaných systémů
- Analyzují miliardy zápisy v protokolech zdrojů



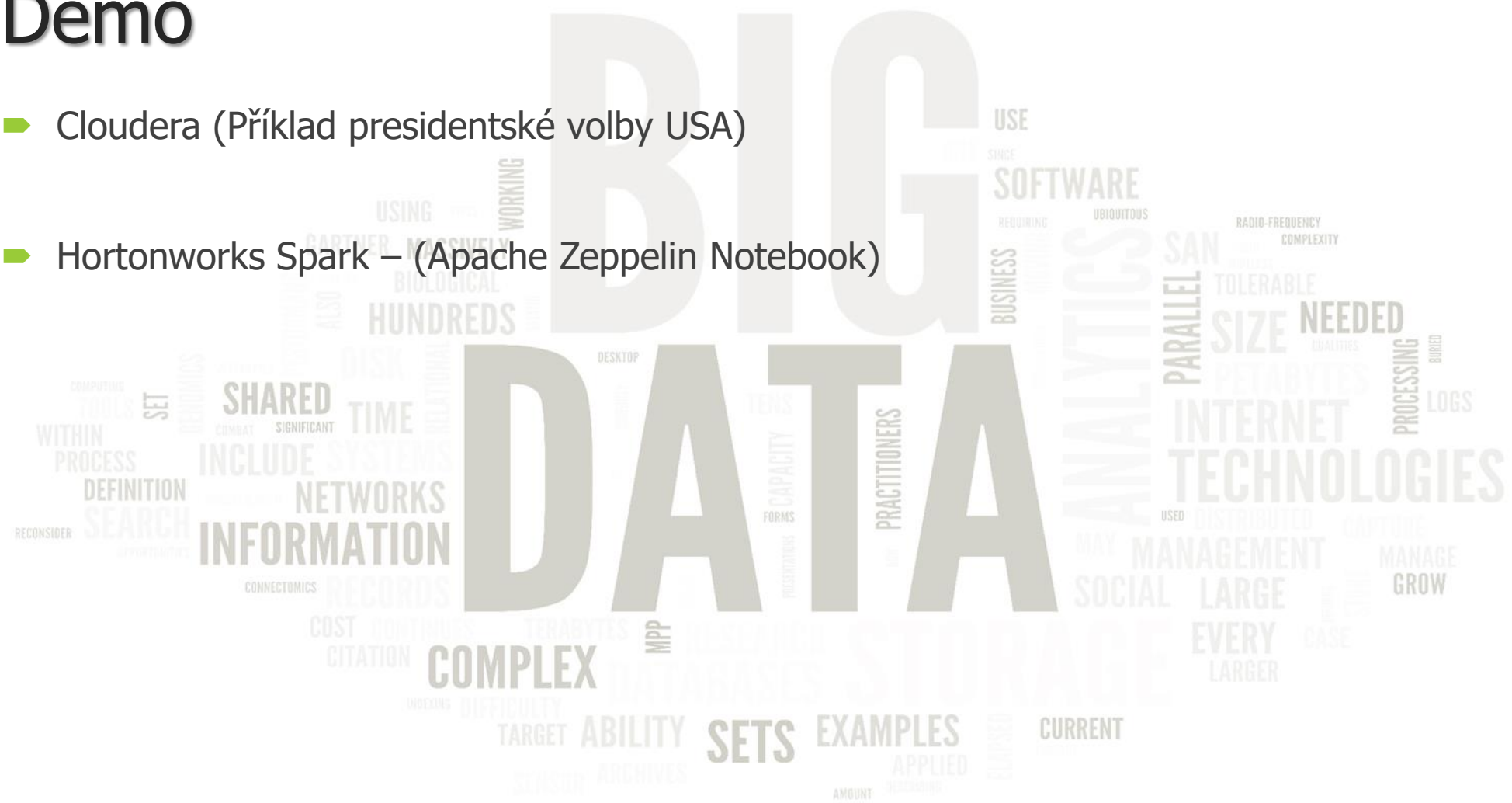
Zdroj:

- http://patft.uspto.gov/netacgi/nph-Parser?Sect2=PTO1&Sect2=HITOFF&p=1&u=/netahtml/PTO/search-bool.html&r=1&f=G&l=50&d=PALL&RefSrch=yes&Query=PN/8108377

(Inventors: Jiang; Changhao, Wei; Xiaoliang; Assignee: Facebook, Inc. (Palo Alto, CA))

Demo

- Cloudera (Příklad presidentské volby USA)
- Hortonworks Spark – (Apache Zeppelin Notebook)



Děkuji za pozornost!

Now that Hadoop has become more commonplace, two types of users have emerged.

The first are people "who find a problem they cannot solve any other way," Cutting said.
As an example, Cutting cited a credit card company with a data warehouse that could only store 90 days' worth of information. Hadoop allowed the company to pool five years' worth of data. Analysis revealed patterns of credit card fraud that could not be detected within the shorter time limit.
The second type of user will apply Hadoop to solve a problem in a way that had not been technically possible before.

Doug Cutting, February 2016, InformationWeek

Martin Stufi
martin.stufi@solutia.cz