INCAD
A SEARCH TECHNOLOGIES COMPANY

SEARCH & BIG DATA

[& ANALYTICS]

The Expert in the Search Space

NÁSLEDUJÍCÍCH 25 MINUT …

- Proč je letošní prezentace modro-zelená

- Vyhledávání
- a Big data
- Search architektura s využitím Big data

- Co to může přinést …

INCAD
A SEARCH TECHNOLOGIES COMPANY

Od ledna 2015
Center of Search Excellence pro Evropu

SEARCH TECHNOLOGIES

Jsme vedoucí nezávislou IT společností zaměřenou na návrh, implementaci a správu podnikových a big data **vyhledávacích řešení**

SEARCH TECHNOLOGIES

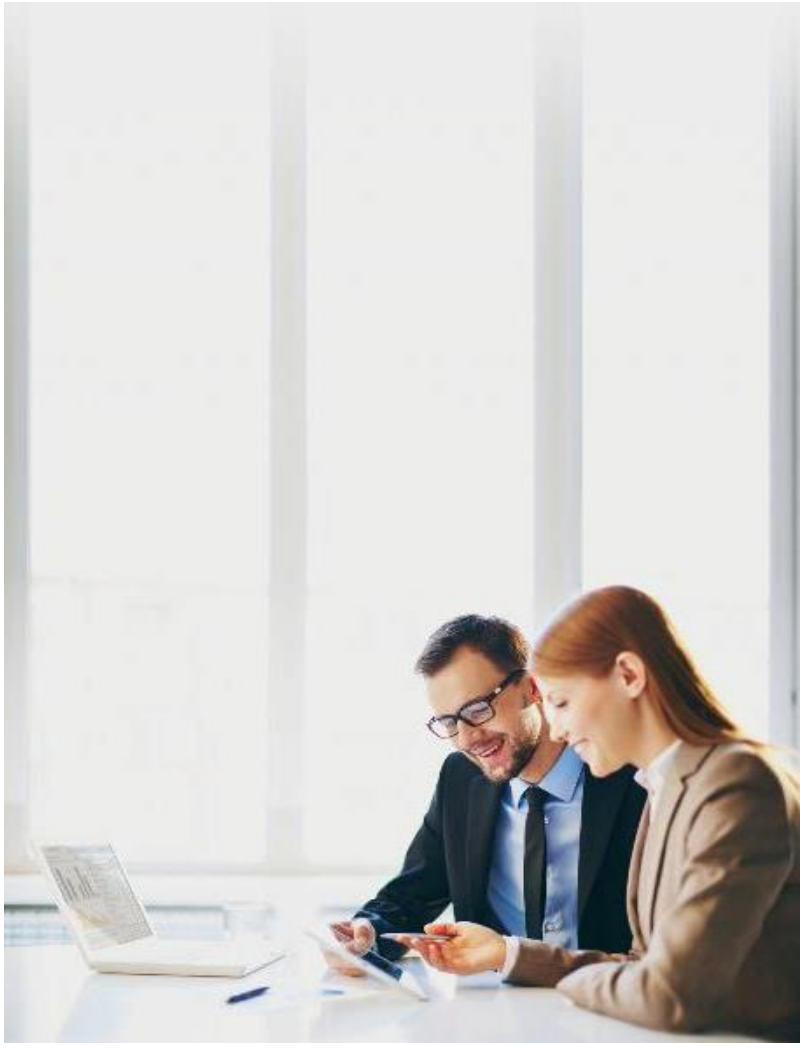
The leading independent solutions company specializing in the **design, implementation** and **management of Search** and **Big Data analytics** applications.

150+ SEARCH & ANALYTICS **EXPERTS**

600+ **CUSTOMERS GLOBALLY**

INCAD.CZ SEARCHTECHNOLOGIES.COM

600+ ZÁKAZNÍKŮ



LIBRARY OF CONGRESS
Reed Business Information
EMC²
Johnson & Johnson
BBC
CISCO
jobsite
TOSHIBA
teleflora
Hitachi Data Systems
WELLS FARGO
Seagate
Chick-fil-A
lenovo
RECKITT BENCKISER
Deloitte.
xerox
PETCO
LORD
LOCKHEED MARTIN
accenture
IBM
CHRISTIE'S
Capital Group Companies
life technologies
nectar
comcast
GPO
EADS ASTRIUM
Unilever
IHS
AARP
The power to make it better.
ESRI
ACTIVISION
Daily Mail
Nintendo
AMERICAN LEGAL Publishing Corporation
PIMCO
Booz | Allen | Hamilton
Ford
imagination at work
ELSEVIER
vmware
Genentech
CPA GLOBAL
RANDOM HOUSE BERTELSMANN
HALLIBURTON
YELLOWPAGES.COM™
Ministry of JUSTICE
QUALCOMM

CO DĚLÁME

- Podnikové vyhledávání (Enterprise Search)
- Data warehouse vyhledávání (BI/Fraud detection)
- E commerce vyhledávání (machine learning, accuracy)
- Search and Match (HR)
- Vyhledávání pro media a nakladatele
- Goverment (portals, archiving, services)

INCAD
A SEARCH TECHNOLOGIES COMPANY

SEARCH & BIG DATA

[& ANALYTICS]

VYHLEDÁVÁNÍ

VYHLEDÁVÁNÍ

- Je výjimečně rychlé
- Je škálovatelné
- Je „schema-free“
- Zvládá velké objemy dat

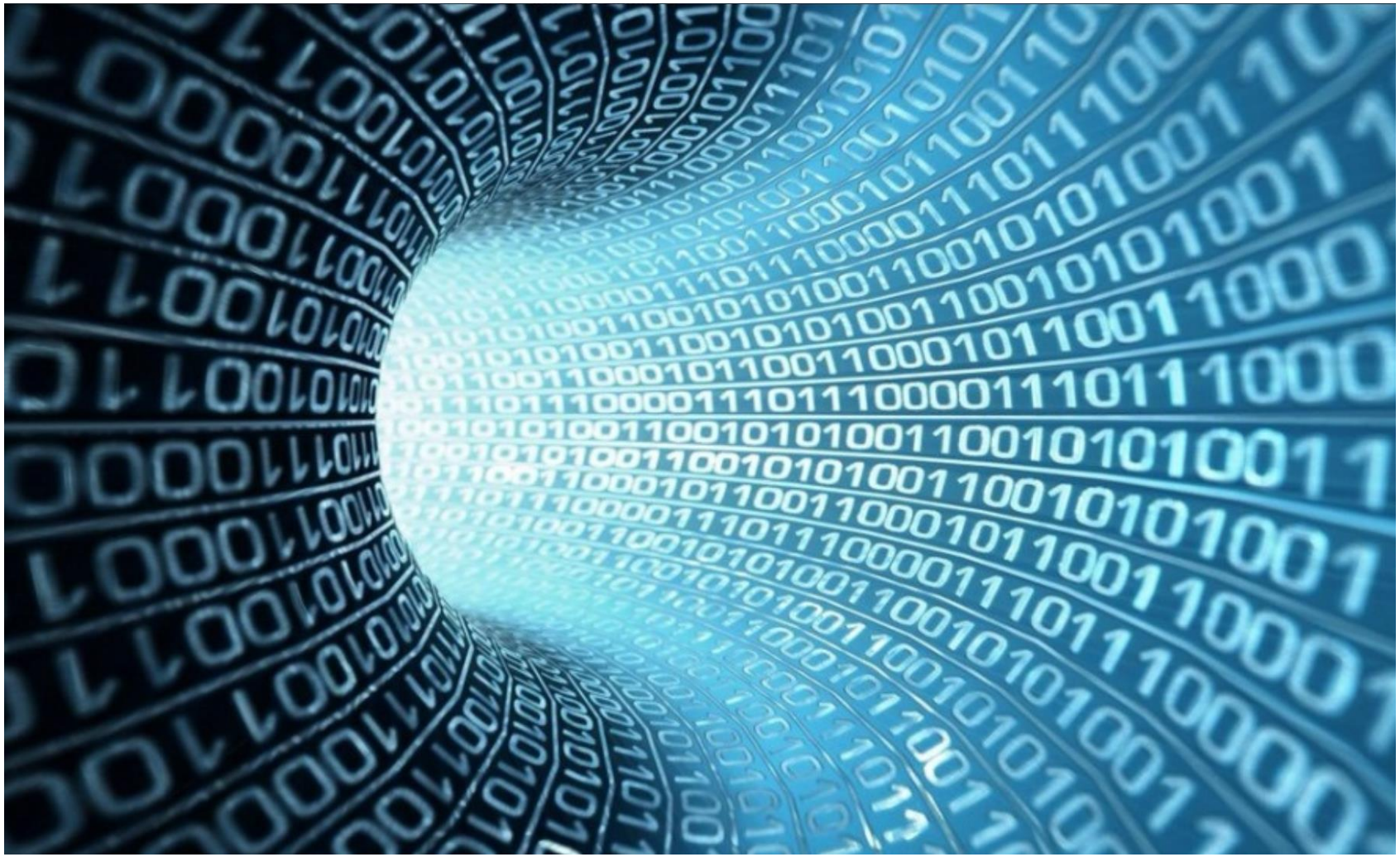
- A je uživatelům známé …

BIG DATA …

"Big data is like teenage sex:
everybody talks about it,
nobody really knows how to do it,
everyone thinks everyone else is doing it,
so everyone claims they are doing it…"

http://www.mediapost.com/publications/article/224120/an-open-letter-to-ibm.html?print

TAKŽE CO VLASTNĚ ?

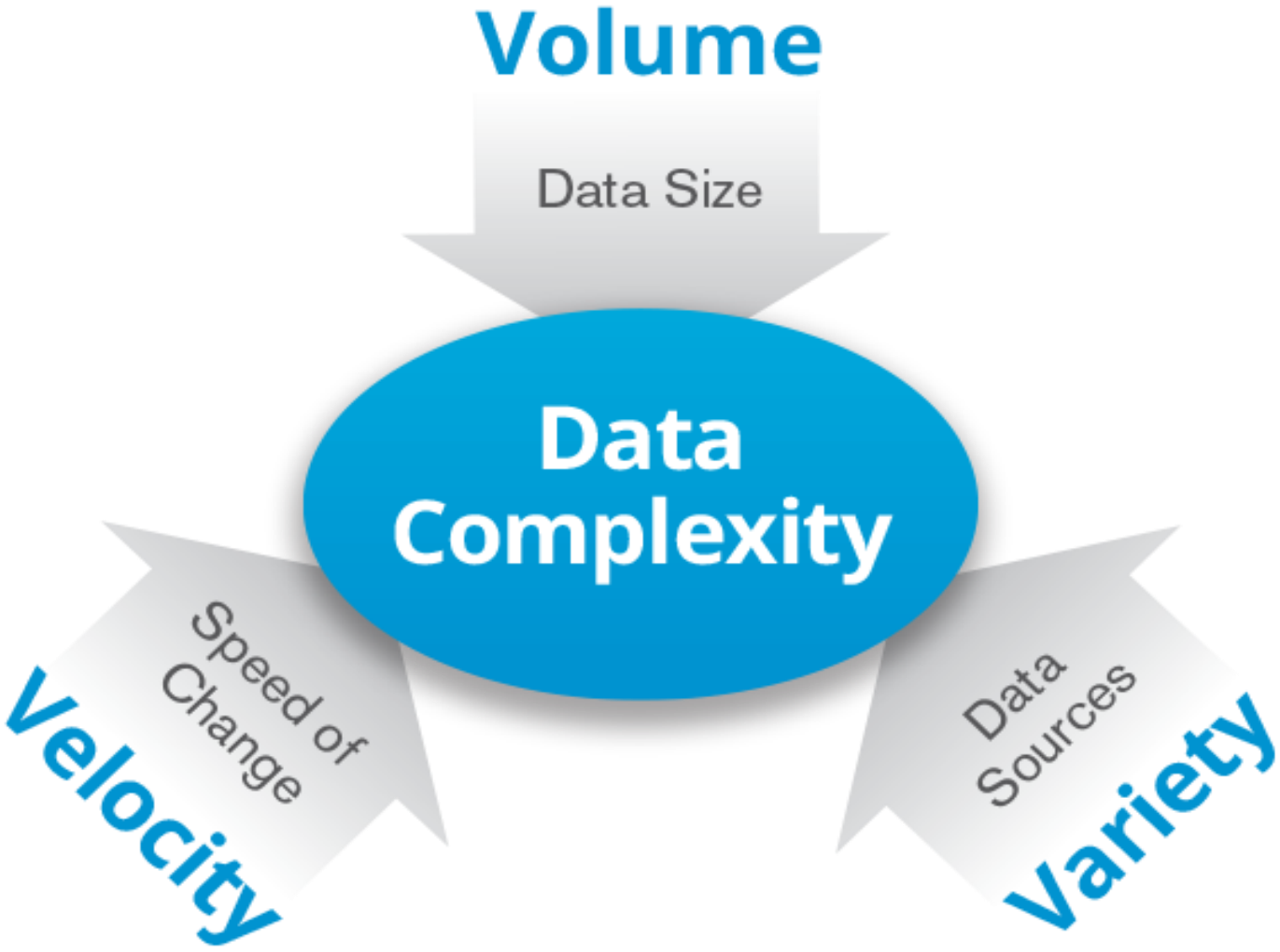


HROMADA DAT

Big data is a broad term for data sets so large or complex that traditional data processing applications are inadequate. Challenges include analysis, capture, data curation, search, sharing, storage, transfer, visualization, and information privacy. The term often refers simply to the use of predictive analytics or other certain advanced methods to extract value from data, and seldom to a particular size of data set. Accuracy in big data may lead to more confident decision making. And better decisions can mean greater operational efficiency, cost reductions and reduced risk.

http://en.wikipedia.org/wiki/Big_data

JAKÝCH DAT?


Volume
Data Size
Data Complexity
Velocity
Speed of Change
Variety
Data Sources

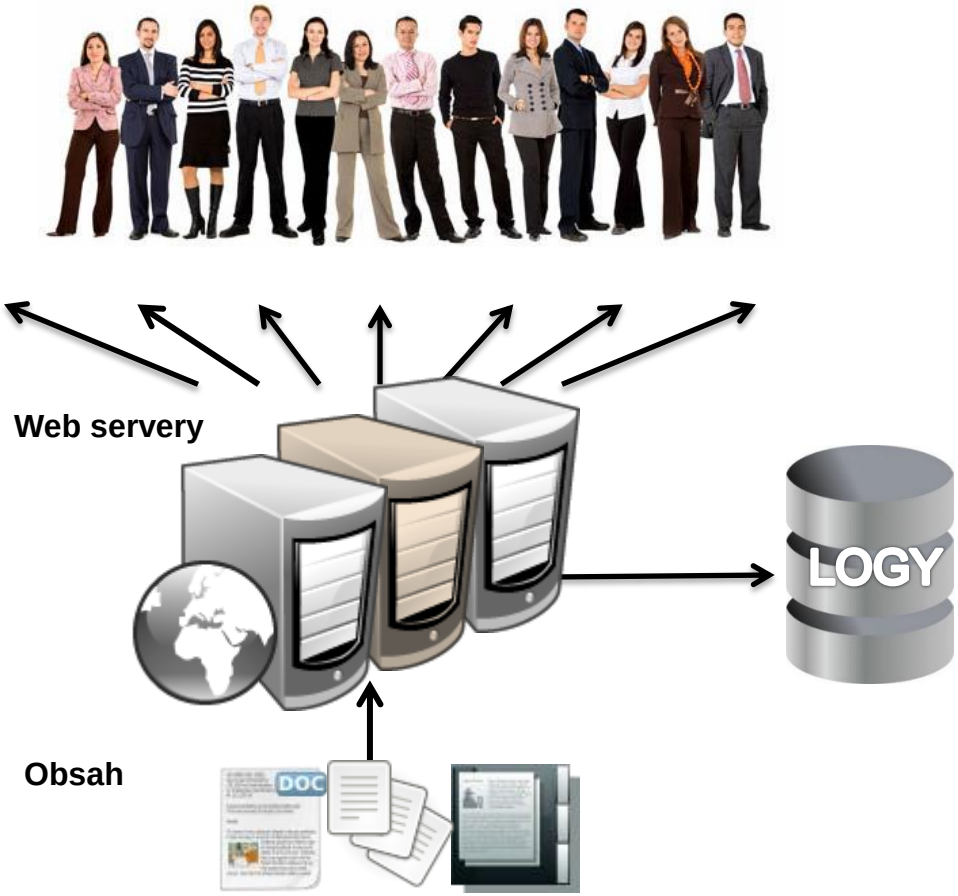
JAKÝCH DAT?

- Příliš na jeden server
  - je fyzicky nemožné je zpracovat na jednom stroji

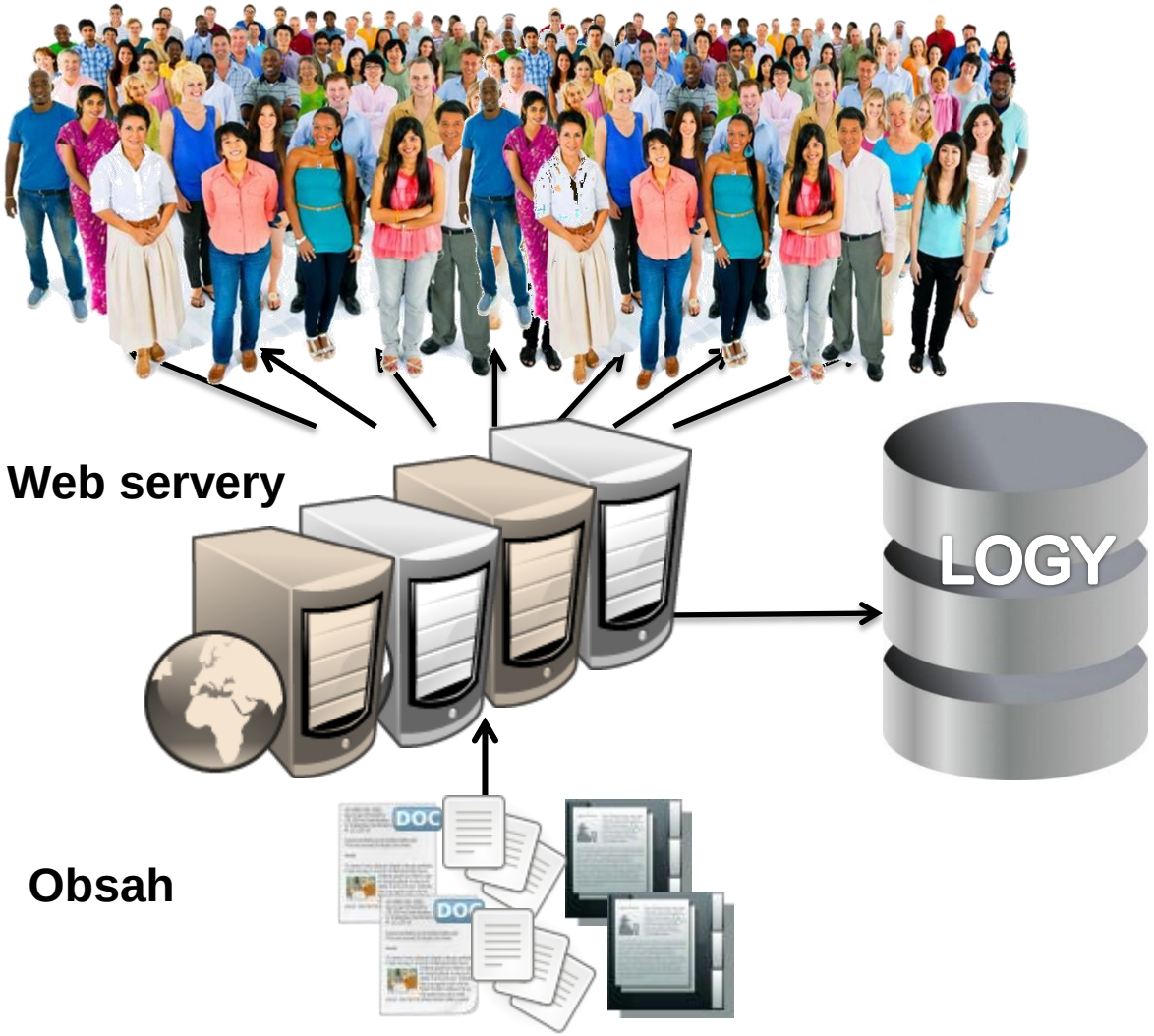
- Agregace dat & Analýza
  - Transformovat datové záznamy nestačí
  - Data je třeba je agregovat / “vyvařit”
  - Dávkové zpracování
  - Dlouho trvající procesy (not real-time)

Spousta dat ≠ “Big Data”

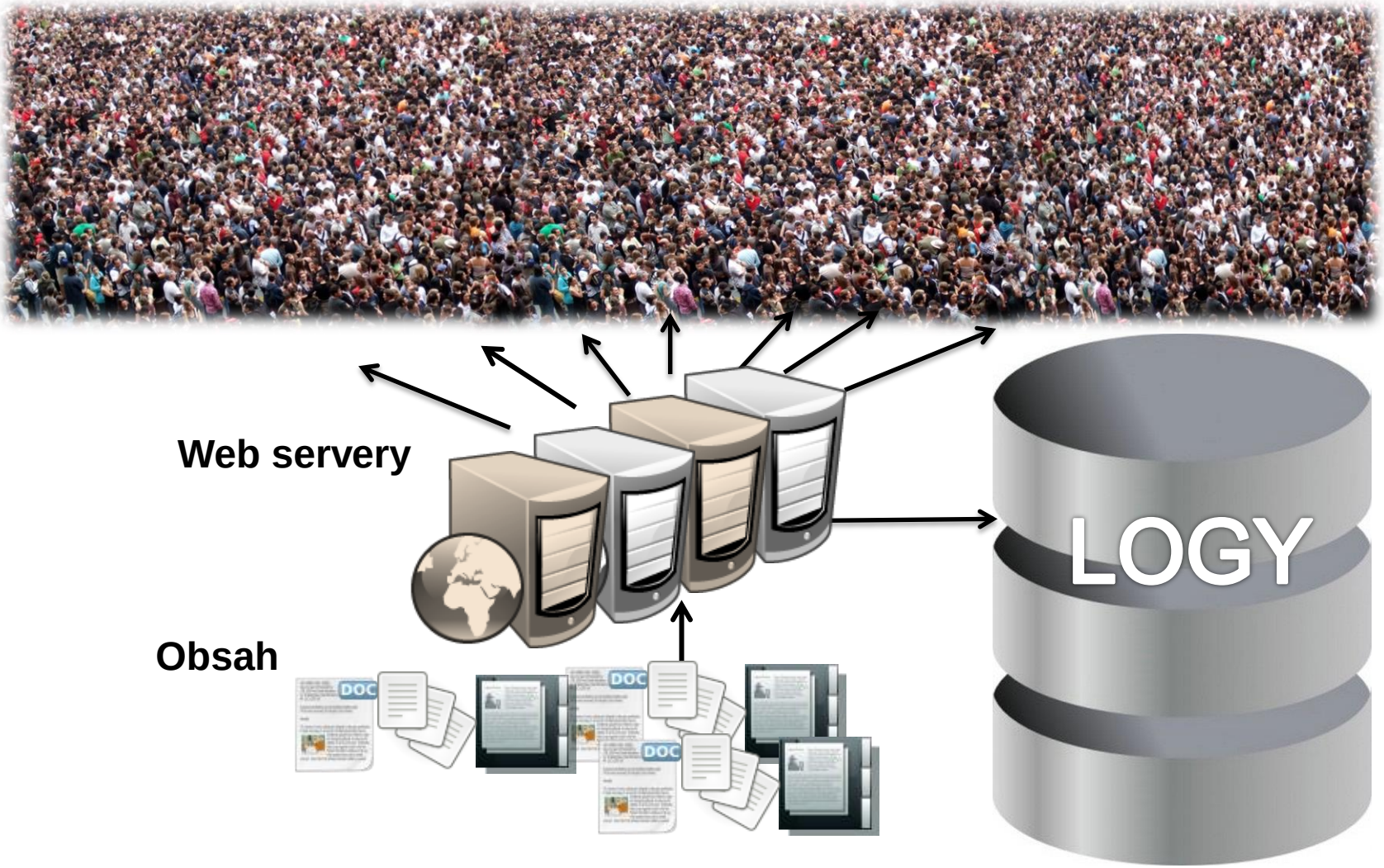
ODKUD SE BEROU ?



ODKUD?



ODKUD?

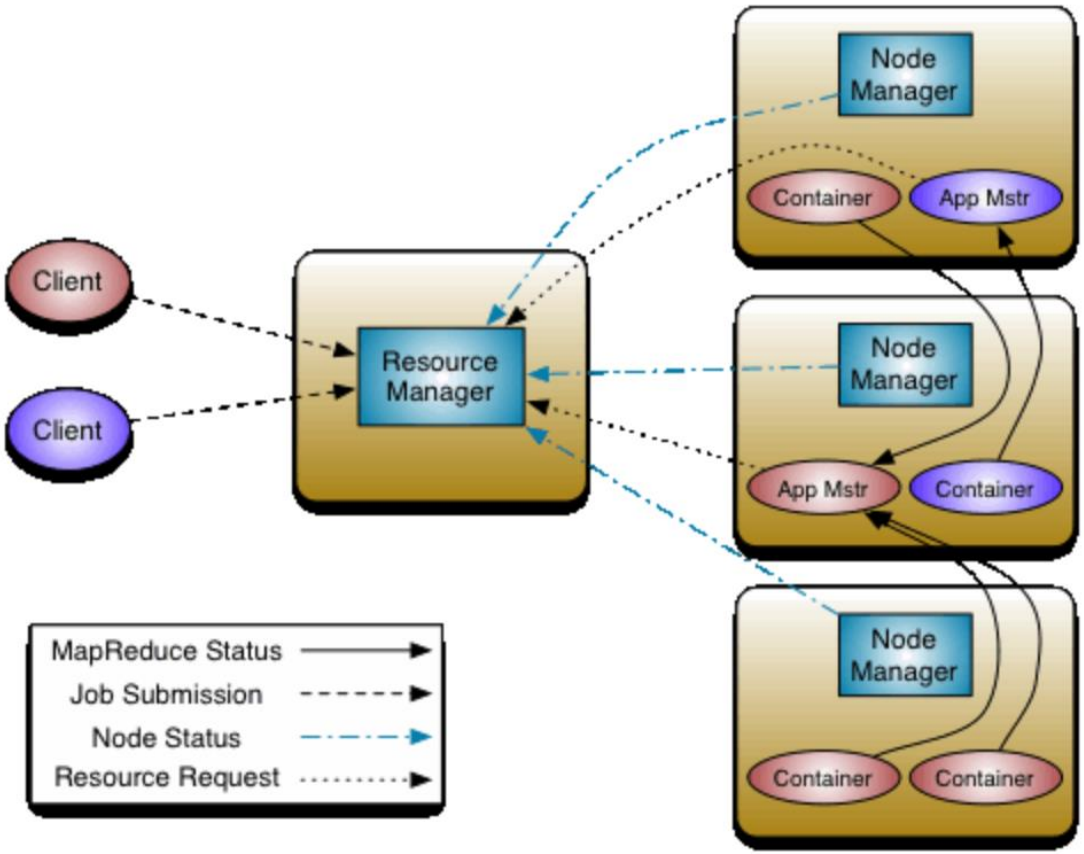


APACHE HADOOP

je open-source projekt vyvíjející software pro spolehlivé, škálovatelné distribuované zpracování.

The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than rely on hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures.

HADOOP MAP REDUCE



http://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html

PROČ HADOOP A SEARCH ?

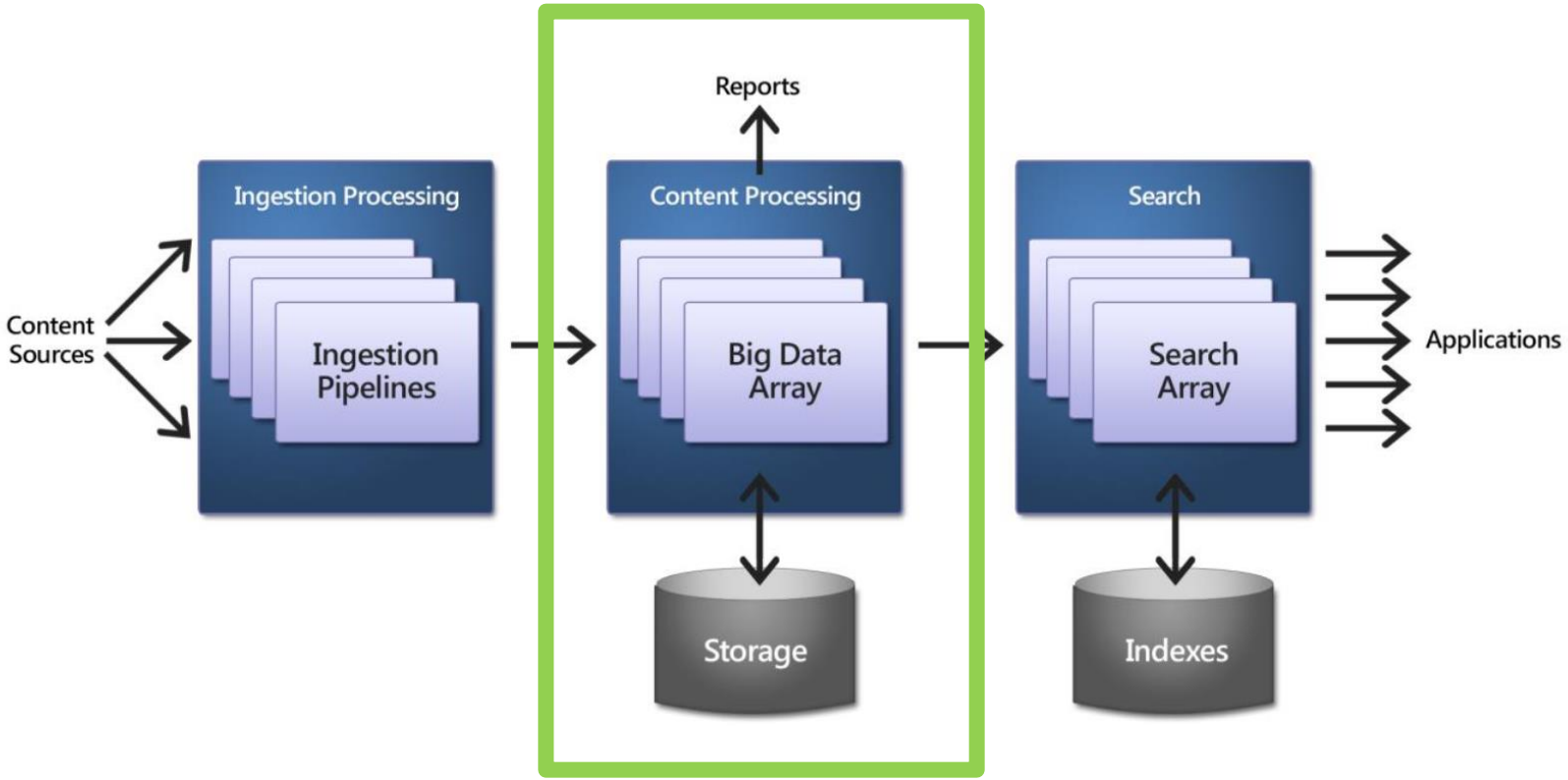
Vyhledávání bude hrát klíčovou roli v příští generaci interaktivních, data využívajících aplikací.

Všichni víme, jak vyhledávat - využití vyhledávací vrstvy k objevování souvislostí v datech je přirozeným postupem.

Řešení, využívající koncept Velkých dat, mohou poskytnout výrazné zlepšení spolehlivosti a flexibility vyhledávacích systémů…

… TO PLATÍ I PRO KNIHOVNY

BIG DATA ARCHITEKTURA PRO VYHLEDÁVÁNÍ

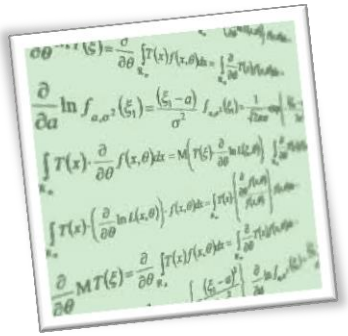


BIG DATA CONTENT PROCESSING

- **Je platformou pro zpracování / normalizaci / obohacení dat a jejich analýzy**

- Udržuje bezpečnou kopii původního textu a metadat každého dokumentu
- Může být využit pro vytváření indexů

CO UMOŽŇUJE BIG DATA SEARCH



HADOOP

CLOUD COMPUTING

MODERNÍ STATISTICKÉ ANALÝZY

MACHINE LEARNING

K ČEMU PŘÍSTUP VYUŽÍVÁME ?

- Search & Match
- Analýza citací (předpokládané a zpětné)
- Latentní sémantické analýzy
- Dokonalejší vyhodnocování (TF/IDF+)
- Detekce přibližných duplicit
- Tagování témat v dokumentech
- Hodnocení výsledků na základě popularity
- Doporučování na základě chování uživatelů
- Doporučené dotazy dle na chování uživatelů
- >>>

ZPRACOVÁNÍ VSTUPŮ

SEARCH → MATCH

Dotaz: Zadán uživatelem

Vstup: Klíčová slova uživatele

Vyhodnocení

Pouze několik klíčových slov

Přesné zadání

Hledání známého

Dotaz: Generovaný automaticky

Vstup: Dokumenty & Data

Normalizované výsledky (%) hodnoty

Řada datových signálů

Semantické hledání

Podobných/související

ZKUŠENOSTI S S&BD

Architecture	Typical re-index rates	Typical time to index 10 million documents
Traditional: Indexing requires re-fetching content from the source	1 to 5 dps*	3 to 16 weeks
Hadoop-based, with a separate, downstream indexer	100 to 300 dps	9 to 28 hours
Hadoop-based, with an integrated indexing capability	1000 to 3000 dps†	1 to 3 hours

dps = documents per second

* For a typical CMS running on a separate sub-net, and downloading primarily MS Office documents and PDFs.

† Distributed indexing, assuming direct access to indexes for up to 6 servers in the Hadoop cluster.

JSME NA INFORU … CO KNIHOVNY ?

- Nesbíráme a zahazujeme data o čtenářích o jejich chování ☹

- Jejich využití může sloužit
  - K optimalizaci fondu
  - K personalizaci služeb
  - Přesnost/správnost hledání
  - Predikci trendů
  - …

ZÁVĚR ?

- Big data přístup může výrazně posunout vyhledávání
- A nyní je lze začít ☺

INCAD
A SEARCH TECHNOLOGIES COMPANY

DOTAZY ?

… DĚKUJI ZA POZORNOST …