

DĚLAT
DOBRÝ SOFTWARE
NÁS BAVÍ

PROFINIT

B0M33BDT – 2. přednáška

Marek Sušický

2. 10. 2019

Osnova

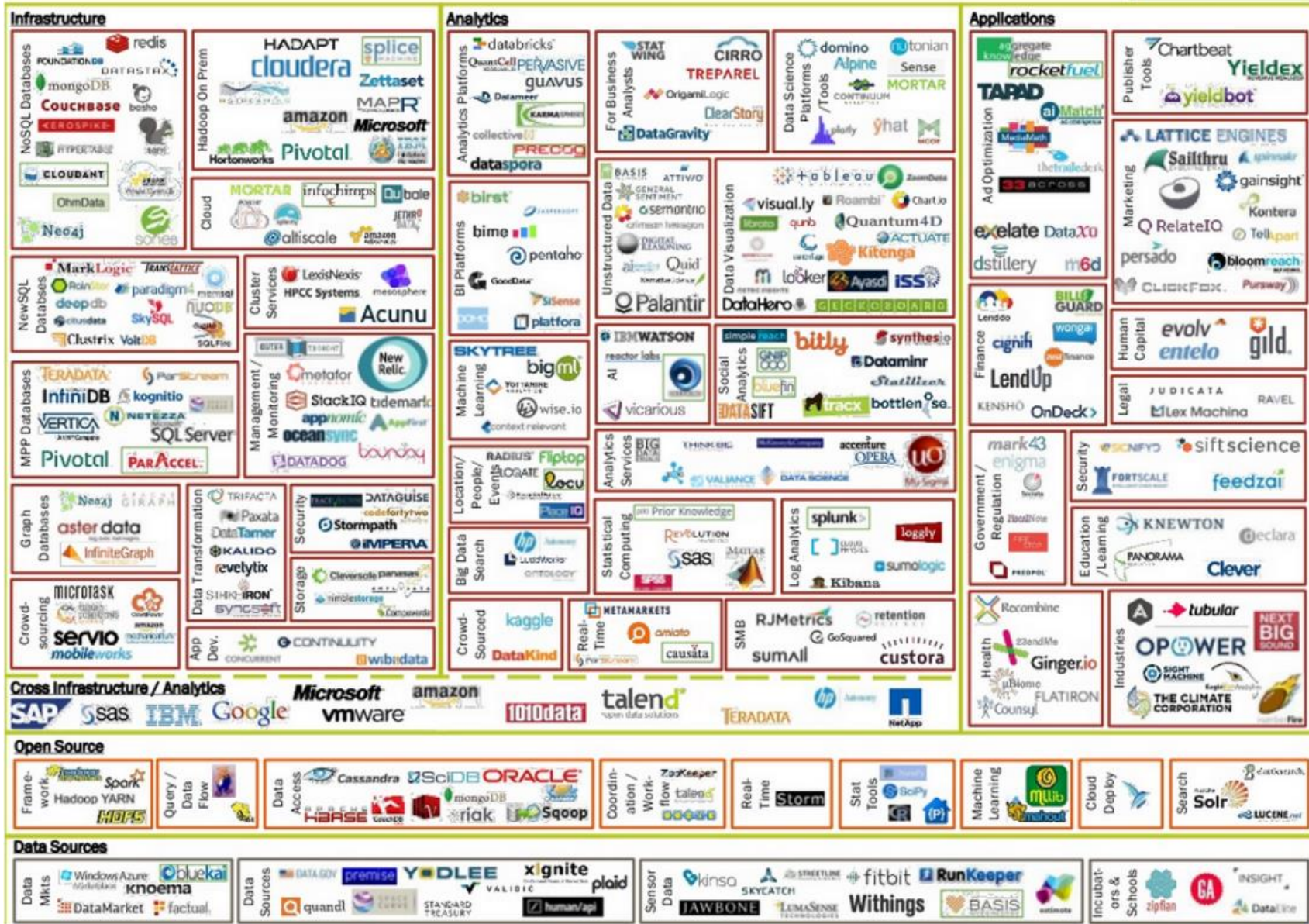
- › Big data a Hadoop
- › Na jakém hardware + sizing
- › Jak vypadá cluster - architektura
- › HDFS
- › Distribuce
- › Komponenty
- › YARN, správa zdrojů

Big data neznamená Hadoop

BIG DATA LANDSCAPE, VERSION 3.0

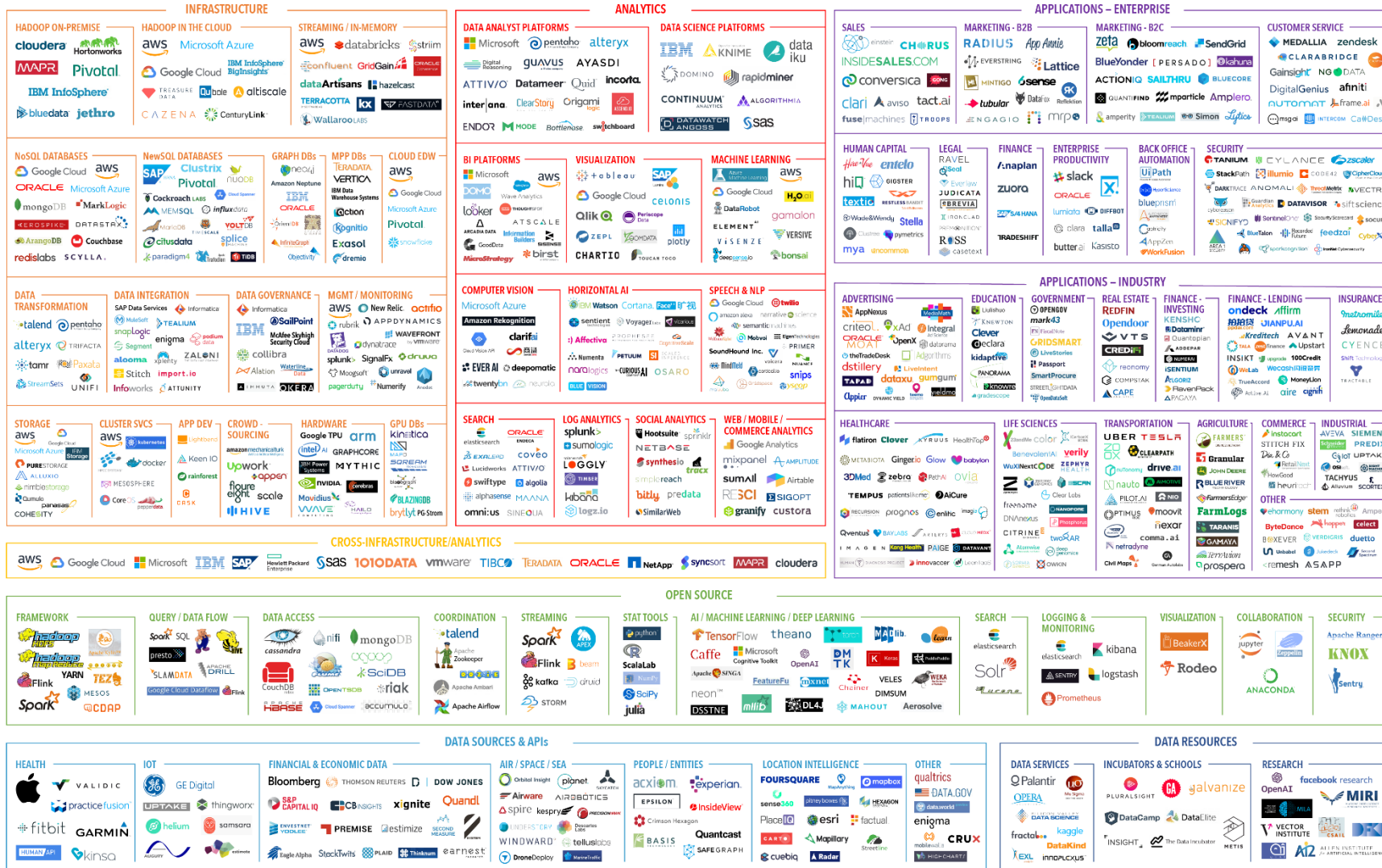
Exited: Acquisition or IPO

PROFIT



PROFIT

BIG DATA & AI LANDSCAPE 2018



Apache Hadoop

- › Wikipedia:
 - Apache Hadoop (pronunciation: /həˈduːp/) is an open-source software framework for **distributed storage** and **distributed processing** of **very large data sets** on computer clusters ***built from commodity hardware***. All the modules in Hadoop are designed with a **fundamental assumption that hardware failures are common** and should be automatically handled by the framework.

- › Commodity hardware
 - stroje za statisíce CZK (ale ne desítky mil.)
 - 2-4 CPU, každé CPU 8-20 jader
 - 256-512 GB RAM, min. 128GB
 - 10-20 2-4-8TB HDD
 - rozhodně ne to, co jako server prodává Alza 😊

Jak vypadá levný HW

All Hosts - Cloudera Manager - Internet Explorer

http://10.171.64.11:7180/cmf/hardware/hosts

All Hosts - Cloudera Manager Hue - Editor

cloudera MANAGER Clusters Hosts Diagnostics Charts Backup Administration Support msusicky

All Hosts

Configuration Add New Hosts to Cluster Re-run Upgrade Wizard Inspect All Hosts

Filters

STATUS 5

- Good Health

CLUSTERS

CORES

COMMISSION STATE

LAST HEARTBEAT

LOAD (1 MINUTE)

LOAD (5 MINUTES)

LOAD (15 MINUTES)

MAINTENANCE MODE

RACK

SERVICES

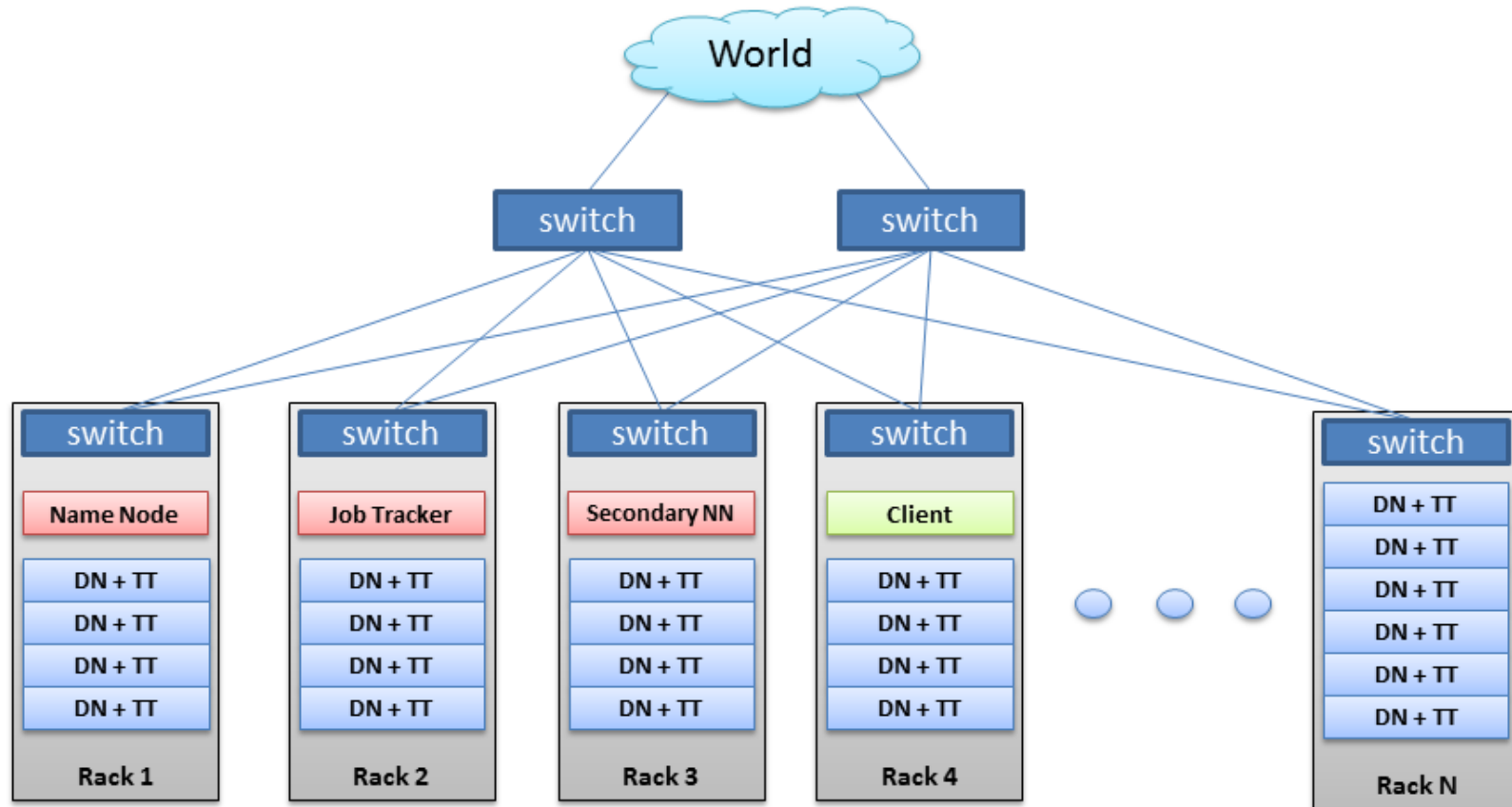
HEALTH TESTS

Actions for Selected

Columns: 9 Selected

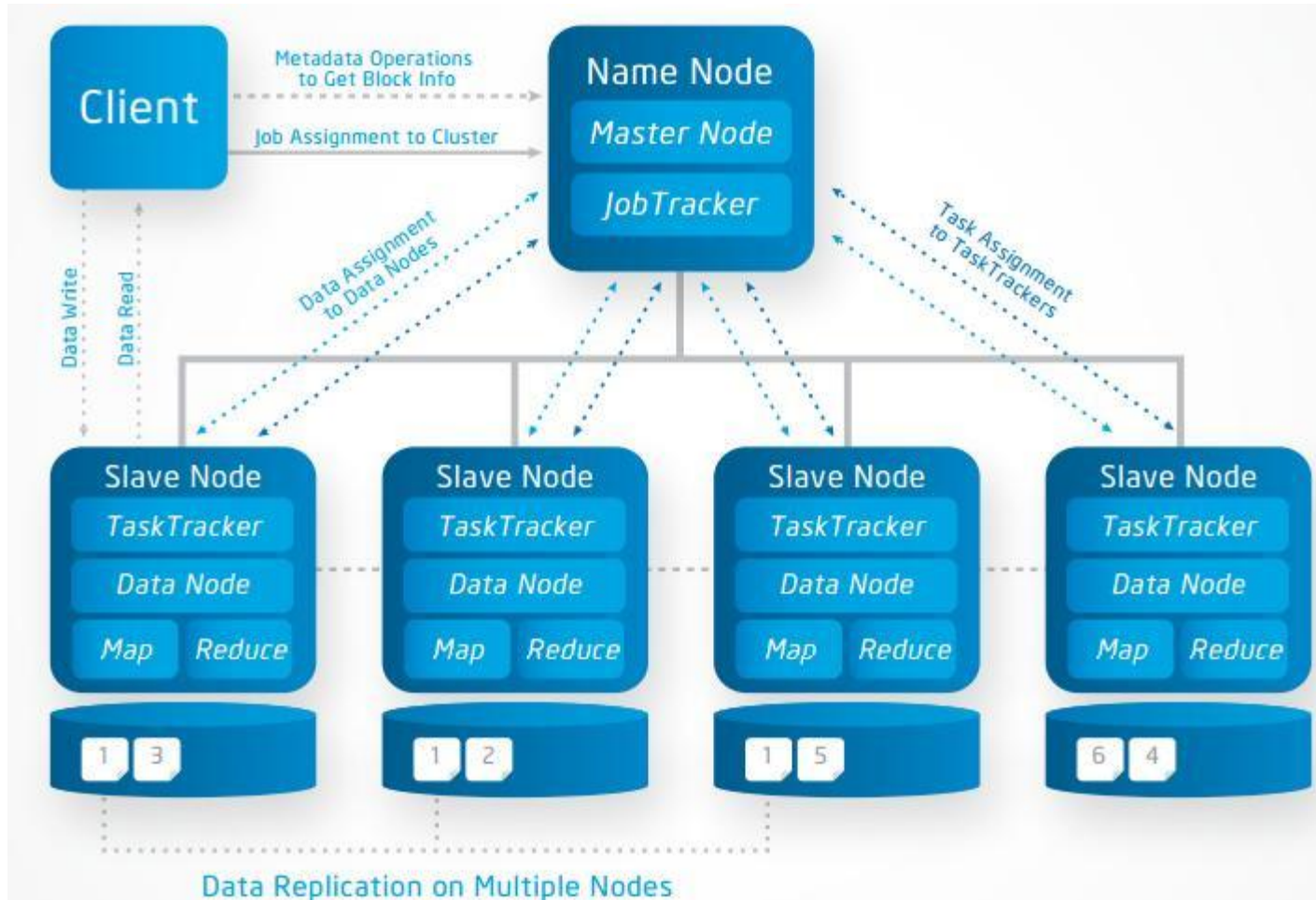
	Status	Name	IP	Roles	Last Heartbeat	Load Average	Disk Usage	Physical Memory	Swap Space
<input type="checkbox"/>	●	dhbbdn05	6.93.1.1	5 Role(s)	7.71s ago	0.00 0.02 0.05	13.6 GiB / 22 TiB	8.6 GiB / 251.8 GiB	0 B / 32 GiB
<input type="checkbox"/>	●	dhbbdn06	6.93.4.1	5 Role(s)	11.64s ago	0.00 0.01 0.05	13.7 GiB / 22 TiB	8.8 GiB / 251.8 GiB	0 B / 32 GiB
<input type="checkbox"/>	●	dhbbdn07	6.93.4.1	4 Role(s)	9.77s ago	0.05 0.04 0.05	13.5 GiB / 22 TiB	7.8 GiB / 251.8 GiB	0 B / 32 GiB
<input type="checkbox"/>	●	dhbfe06	6.93.4.1	16 Role(s)	9.09s ago	0.10 0.22 0.38	29.9 GiB / 141.5 GiB	16.3 GiB / 251.8 GiB	0 B / 32 GiB
<input type="checkbox"/>	●	dhbfe07	6.93.4.1	10 Role(s)	7.42s ago	0.12 0.06 0.05	24 GiB / 353.9 GiB	12.4 GiB / 125.8 GiB	0 B / 32 GiB

Hadoop Cluster

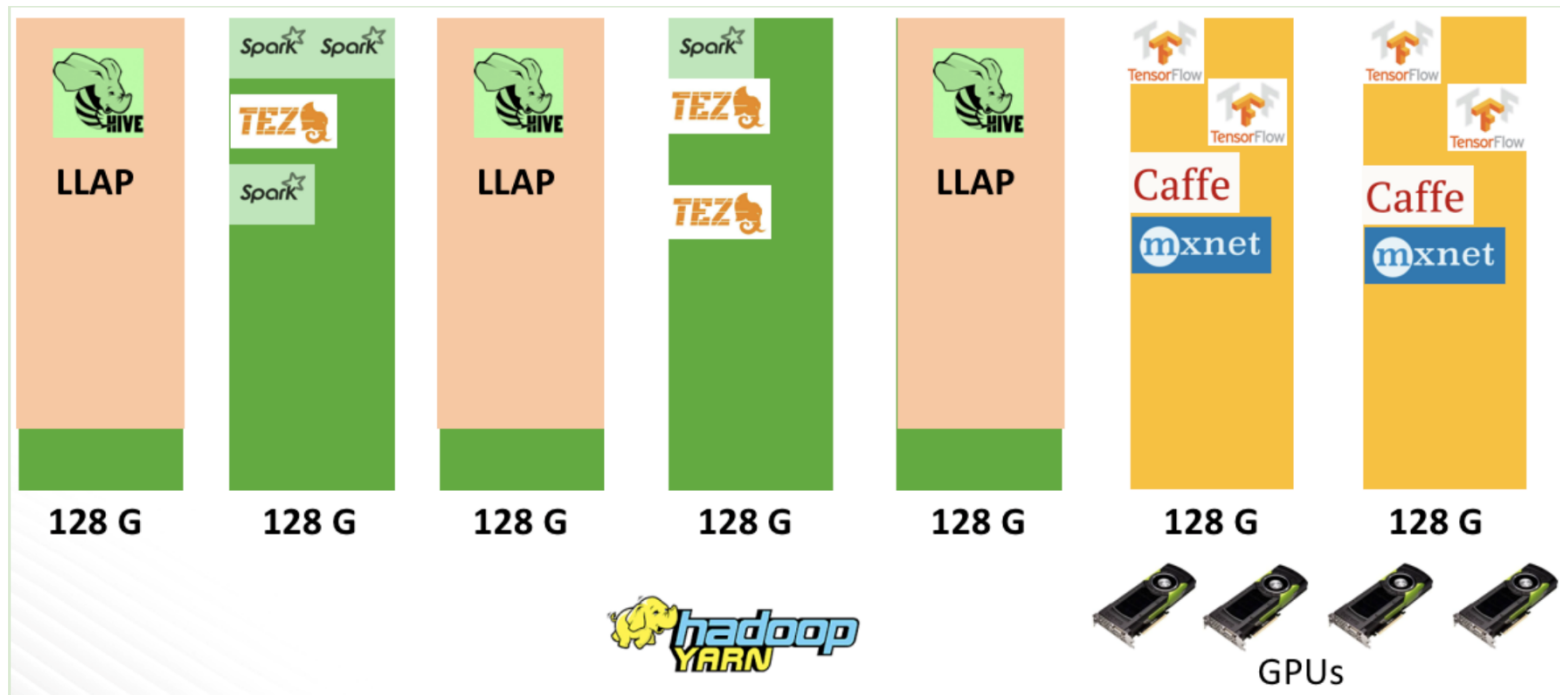


BRAD HEDLUND .com

Hadoop – architektura II



Nově i GPU



With GPU scheduling support, YARN can allocate non-GPU applications (like LLAP / Spark / Tez, etc.) to machines without GPU (left side), and allocate GPU applications (like Tensorflow / Caffe / MXNet) to machines with GPU. (right side)

Sizing

- › Jak postavit Hadoop
 - Jak si ho objednat
 - HDD parametry
 - Přenosová rychlost
 - RAID
 - 0, 1, 1+0, 5, 6, (2,3,4,7)
 - Síťová rychlost
 - SAN/NAS
 - Paměť
 - CPU jádra
 - Obecná doporučení

Sizing

- › Kalkulace HW požadavků
 - Počet nodů
 - Počet disků
 - HW parametry (CPU, RAM, RAID...) typů nodů

Rychlosti čtení dat

- › RAM
 - DDR4 cca 15 GB/s
- › Síť 10 Gbit
 - 1.25 GB/s
- › SSD disk
 - 200-700 MB/s
 - existují „Enterprise level“, které vydrží (garance 5 let)
 - malé kapacity (max 1TB) a hodně drahé
- › HDD 7.2k
 - latence cca 4ms
 - sekvenční čtení 50-100 MB/s
 - velké kapacity (běžně 4TB-8TB) a relativně levné
 - ➔ Hadoop typicky pracuje s úložištěm

Rychlosti čtení dat – HDD

- › **Sekvenční čtení** – cca 100 MB/s za jedním diskem
- › **Random access**
 - velikost bloku ext4 bývá 4kB
 - latence, než disk najde blok cca 4ms
 - max. rychlost čistě náhodného čtení $1/0.004 \cdot 4096 = 1 \text{ MB/s}$

Omezující faktory

- › Příklad: 10 nodů, každý node 12 * 2 TB HDD
 - Rychlost čtení v rámci nodu: $12 \cdot 100 \text{ MB/s} = 1.2 \text{ GB/s}$
 - Rychlost čtení v rámci clusteru: 12 GB/s
- › Omezení:
 - rychlost RAM – 10x větší
 - CPU – čtení nezatěžuje
 - síť – na hraně pro jeden node!
 - sběrnice – pozor na počet disků, musí zvládnout



Sizing

- › Ukázkový příklad k zamyšlení
 - 15GB/5 min
 - Historie 30 dní
 - SLA – 5 sekund pro 85% dotazů
 - 10s pro 100% dotazů
 - Při nesplnění vysoké pokuty

Principy

- › Ukládání velkého množství dat
 - mnoho serverů = nodů [desítky až tisíce]
 - každý node mnoho disků [10-20]
- › Zamezení ztráty dat – výpadek nodu
 - replikace (typicky tři kopie každého souboru)
 - 2 repliky ve stejném racku, třetí replika mimo
- › Rychlost čtení
 - data jsou rozložena v celém clusteru – 1 soubor nemusí být celý v jednom nodu!
 - data jsou replikována – lze paralelně číst na několika nodech bez nutnosti přenosu dat přes síť
 - velké soubory – **výhody sekvenčního čtení**
- › Distribuce výpočtů
 - mnoho nodů, přiřazování výpočetního výkonu

Principy Hadoop – syntéza

- › Maximálně využívat sekvenční čtení
- › Pracovat s velkými soubory, které se čtou sekvenčně
 - ušetří se čas na synchronizaci/orchestraci v rámci distribuovaného systému
- › Čím méně dat se načte, tím rychleji se načtou – využití **komprese**
- › CPU se při čtení fláká, paměť je násobně rychlejší, tj. typicky dekomprese bude rychlejší než IO operace
- › Maximum výpočtů provádět na nodu, kde probíhá čtení dat, přes síť posílat jen co nejvíce agregovaná data
 - síťové rozhraní nebude mít dostatečnou kapacitu
- › **Zásadní omezení – I/O operace**
 - **Tip:** při odhadu, jak dlouho co bude trvat stačí prakticky vždy počítat jen s IO a počtem paralelních čtení/uživatelů (ale neplatí samozřejmě pro ML aplikace)



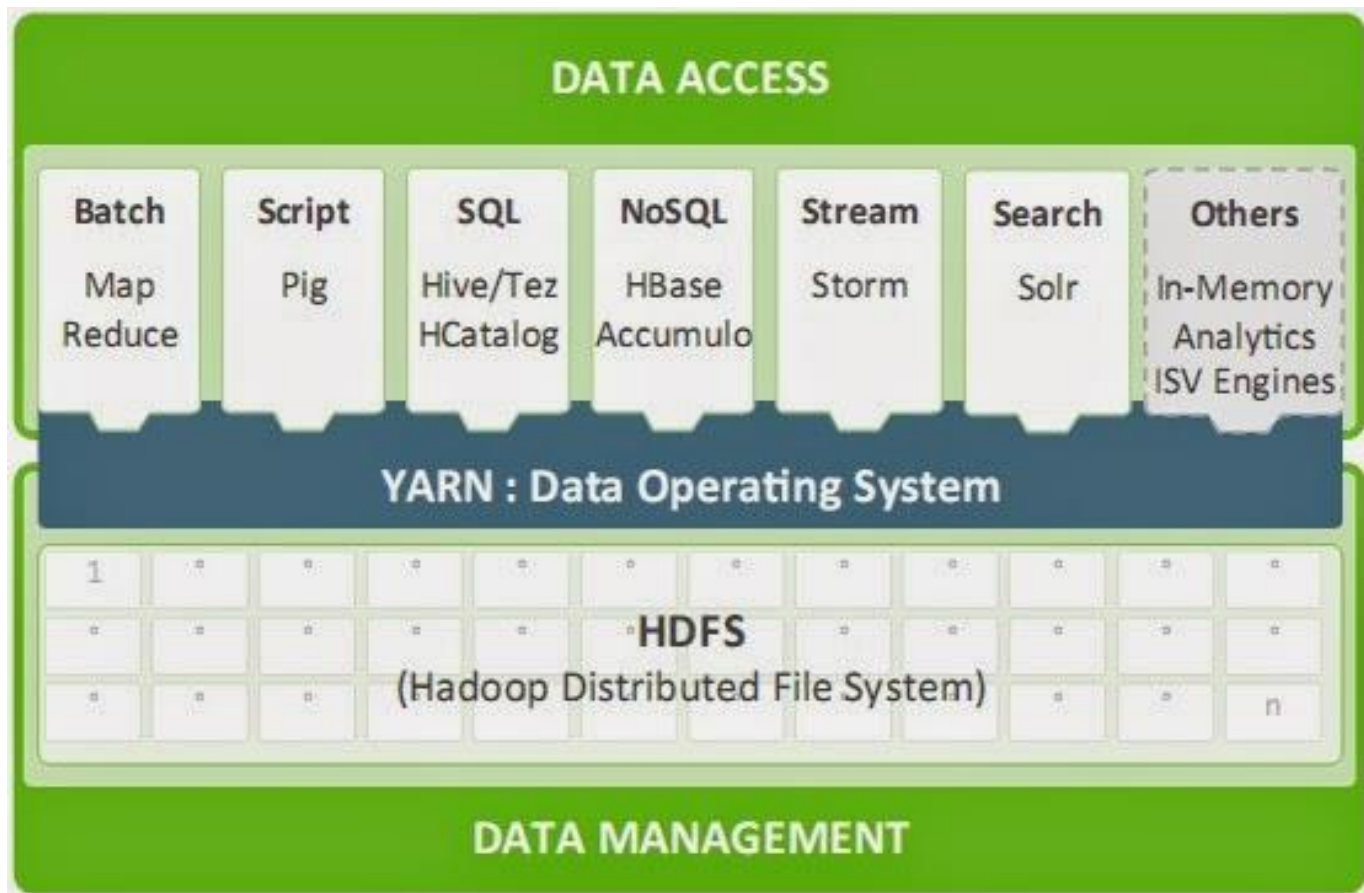
HDFS

HDFS

› HDFS

- NameNode, DataNode
- Replikace
- Operace na souborovém systému
- Bloky, velikost bloku

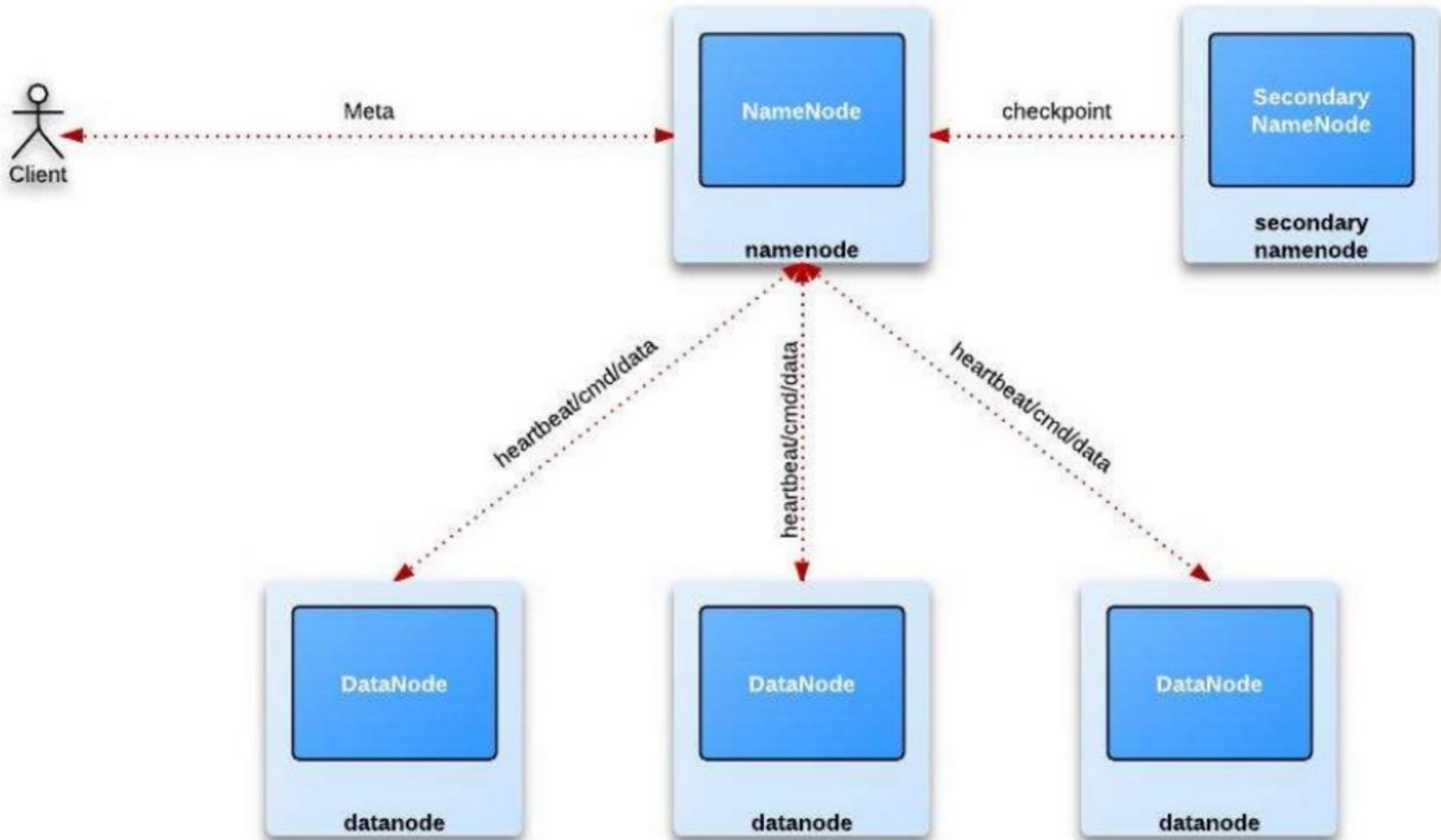
HDFS- architektura



HDFS

- › Hadoop Distributed Filesystem
- › Dobrý pro
 - Velké soubory
 - Streamovaný přístup
- › Špatný pro
 - Spoustu malých souborů
 - Náhodný přístup
 - Nízkolatenční přístup
- › Master-slave design
 - Master – NameNode
 - Slave – DataNode
 - SecondaryNameNode

HDFS



HDFS

- › HDFS soubory jsou rozděleny do bloků
 - Default 64MB/128MB, ale lze změnit
 - Dobré pro velké soubory
 - Děsné pro malé...
- › Replikace
 - Jeden blok může být v nejméně xxx DataNodes
 - Fault tolerant
 - Default hodnota 3

NameNode

- › Metadata filesystemu
 - Kde jsou data
 - V paměti
 - 1GB pro každý milion bloků
- › Ve spojení s
 - Klienty
 - DataNodey
 - SecondaryNameNodey
 - Checkpointing
 - Editlogs a fsimage

DataNode

- › Ukládá Databloky
- › Získává bloky od klientů
- › Získává bloky od ostatních DataNodů
 - Replikace
- › Dostává příkaz delete od NameNode

HDFS filesystem

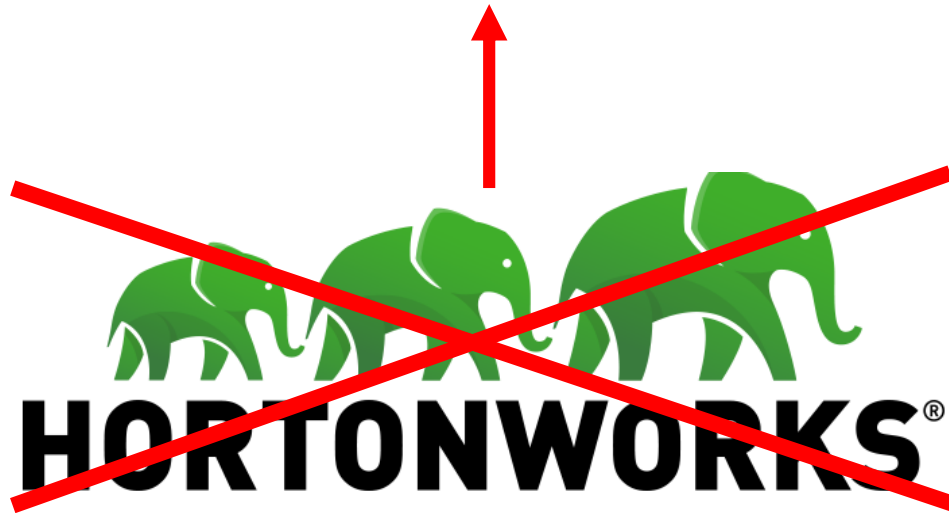
- › put
- › get
- › copyFromLocal
- › ls
- › Rights
 - Chmod
 - Chown
 - Chgrp
- › Další zde
 - <https://hadoop.apache.org/docs/r2.7.1/hadoop-project-dist/hadoop-hdfs/HDFSCommands.html>
- › Vyzkoušíme na cvičení

The background of the slide is a dark gray field filled with a complex, overlapping pattern of translucent, light gray polygons. These polygons vary in size and orientation, creating a sense of depth and movement, similar to a low-poly 3D environment or a crystalline structure. The word "Distribuce" is centered in the middle of the image.

Distribuce

Distribuce

cloudera



MAPR 

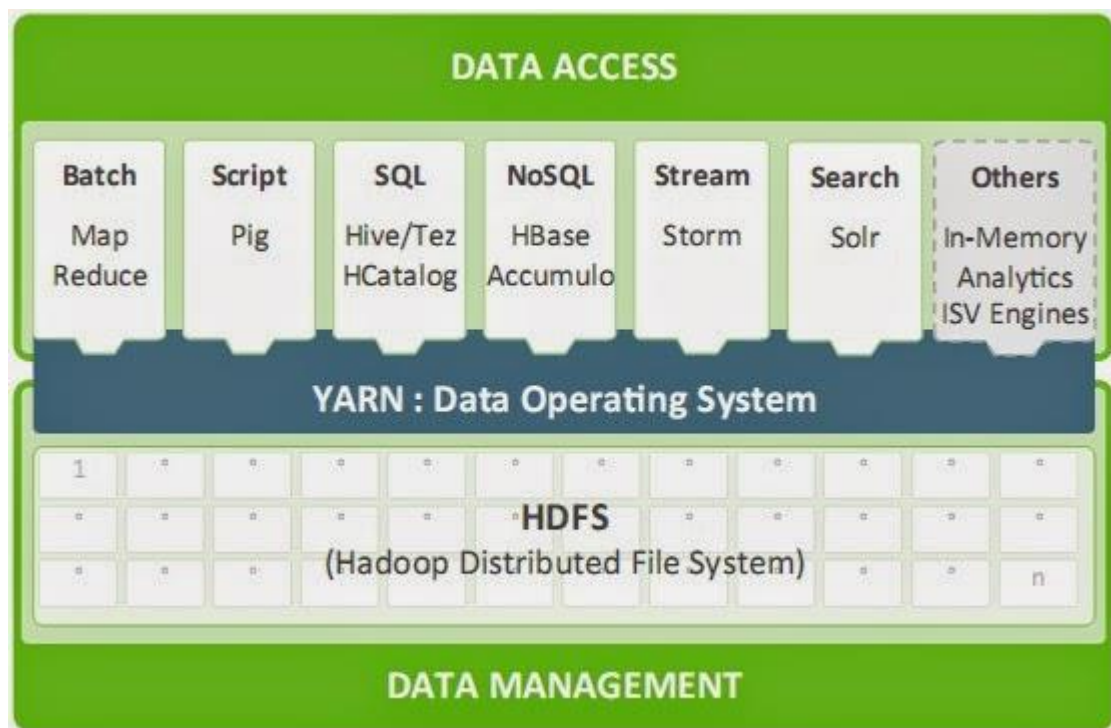
Hotové distribuce

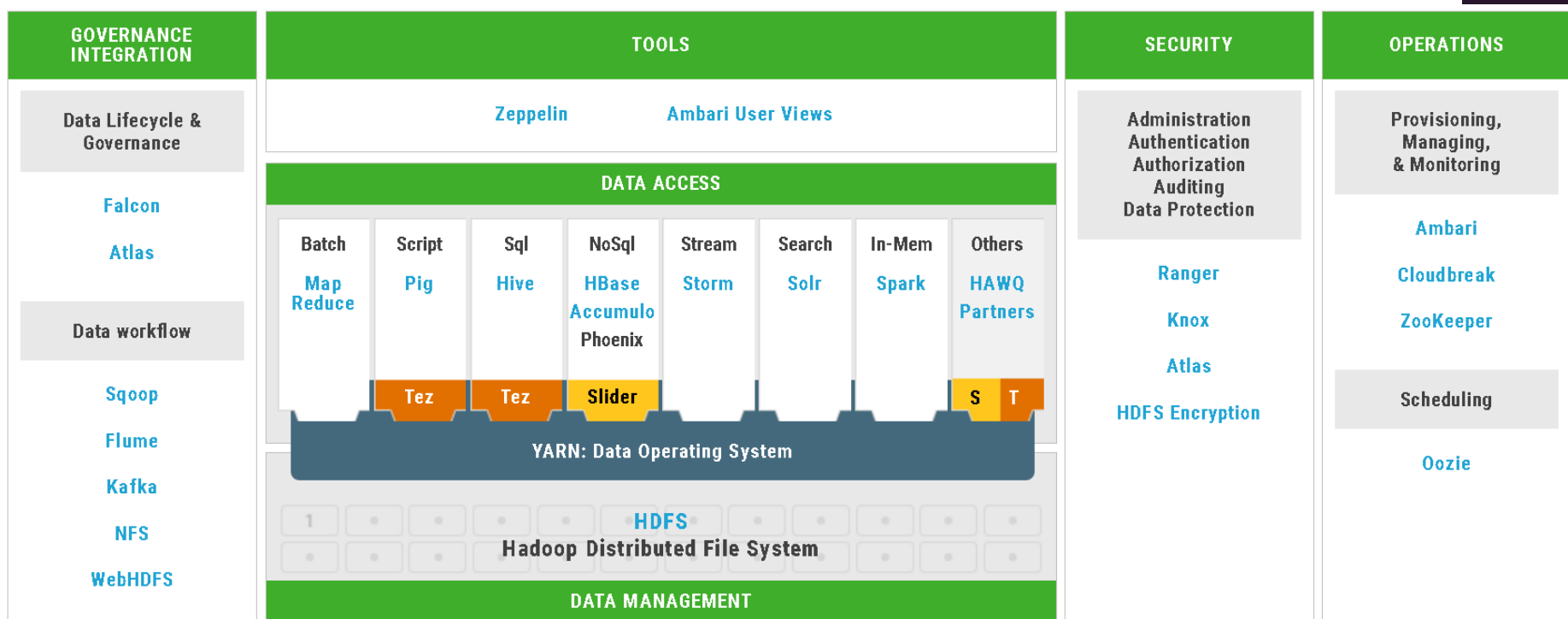
- › Řeší peklo závislostí – jeden update může vyvolat řetězovou vlnu
 - › Nabízejí komerční podporu
 - › Rychlejší reakce ?
 - › Co je zadarmo, je špatné ?!
 - › Proč znovu vymýšlet kolo
-
- › „Musí se k tomu dospět“
 - › “Proč mi tady prodávají Hortonworks?”

The background of the slide is a dark gray field filled with a complex, overlapping pattern of translucent, light gray polygons. These polygons vary in size and orientation, creating a sense of depth and movement, similar to a low-poly 3D environment or a crystalline structure. The word "Komponenty" is centered in the middle of the image in a clean, white, sans-serif font.

Komponenty

Zvěřinec - zjednodušeně





Verze mezi releasy

Ongoing Innovation in Apache																							
HDP 2.5 2H2016	2.7.3	0.16.0	1.2.1+ 2.1*	0.7.0	5.2.2	1.6.2+ 2.0*	0.6.0	0.91.0	1.1.2	4.7.0	1.7.0	1.0.1	0.10.0	0.7.0	1.4.6	1.5.2	0.10.0	2.4.0	1.3.0	3.4.6	4.2.0	0.9.0	0.6.0
HDP 2.4 Mar 2016	2.7.1	0.15.0	1.2.1	0.7.0	5.2.1	1.6.0		0.80.0	1.1.2	4.4.0	1.7.0	0.10.0	0.6.1	0.5.0	1.4.6	1.5.2	0.9.0	2.2.1	1.2.0	3.4.6	4.2.0	0.6.0	0.5.0
HDP 2.3 Oct 2015	2.7.1	0.15.0	1.2.1	0.7.0	5.2.1	1.4.1		0.80.0	1.1.2	4.4.0	1.7.0	0.10.0	0.6.1	0.5.0	1.4.6	1.5.2	0.8.2	2.1.0	1.0.0	3.4.6	4.2.0	0.6.0	0.5.0
HDP 2.2 Dec 2014	2.6.0	0.14.0	0.14.0	0.5.2	4.10.2	1.2.1		0.60.0	0.98.4	4.2.0	1.6.1	0.9.3	0.6.0		1.4.5	1.5.2	0.8.1	2.0.0		3.4.6	4.1.0	0.5.0	0.4.0
HDP 2.1 April 2014	2.4.0	0.12.1	0.13.0	0.4.0	4.7.2				0.98.0	4.0.0	1.5.1	0.9.1	0.5.0		1.4.4	1.4.0		1.5.1		3.4.5	4.0.0	0.4.0	
HDP 2.0 Oct 2013	2.2.0	0.12.0	0.12.0						0.96.1						1.4.4	1.3.1		1.4.4		3.4.5	3.3.2		
	Hadoop & YARN	Pig	Hive	Tez	Solr	Spark	Zeppelin	Slider	HBase	Phoenix	Accumulo	Storm	Falcon	Atlas	Sqoop	Flume	Kafka	Ambari	Cloudbreak	Zookeeper	Oozie	Knox	Ranger
DATA MGMT		DATA ACCESS						GOVERNANCE & INTEGRATION						OPERATIONS			SECURITY						
HORTONWORKS DATA PLATFORM																							

* Spark 1.6.2+ Spark 2.0 – HDP 2.5 support installation of both Spark 1.6.2 and Spark 2.0. Spark 2.0 is Technical Preview within HDP 2.5.
Hive 1.2.1+ Hive 2.1 – Hive 2.1 is Technical Preview within HDP 2.5.

Verze mezi releases

Ongoing Innovation in Apache																									Add on Sku	
HDP 3.0.0 Q3 2018	3.1.0	4.3.1	0.16.0	3.0.0	0.12.0	0.9.1	1.16.0	1.4.7	2.3	0.8.0	2.0.0	5.0.0	1.7.0	1.0.0	1.1.0	1.0.0	1.2.1	1.0.1	2.7.0	3.4.6					6.6.2 ^[4]	
HDP 2.6.5 Q2 2018	2.7.3	4.2.0	0.16.0	1.2.1+ 2.1 ^[3]	0.10.1	0.7.0	1.2.0 ^[6] 1.10 ^[7]	1.4.6	1.6.3+ 2.3	0.7.3	1.1.2	4.7.0	1.7.0	0.12.0	0.7.0	0.8.0	1.1.0	1.0.0	2.6.2	3.4.6	1.5.2	0.10.0	0.90	0.92.0	6.6.2 ^[4]	
HDP 2.6.4 ^[1] Q4 2017	2.7.3	4.2.0	0.16.0	1.2.1+ 2.1 ^[3]	0.10.1	0.7.0	1.2.0 ^[6] 1.10 ^[7]	1.4.6	1.6.3+ 2.2 ^[5]	0.7.3	1.1.2	4.7.0	1.7.0	0.12.0	0.7.0	0.8.0	1.1.0	0.10.1	2.6.1	3.4.6	1.5.2	0.10.0	0.90	0.92.0	5.5.1 ^[4]	
HDP 2.5 Aug 2016	2.7.3	4.2.0	0.16.0	1.2.1+ 2.1 ^[3]		0.7.0	1.2.0 ^[6] 1.6 ^[7]	1.4.6	1.6.2+ 2.0 ^[2]	0.6.0	1.1.2	4.7.0	1.7.0	0.9.0	0.6.0	0.7.0	1.0.1	0.10.0	2.4.0	3.4.6	1.5.2	0.10.0	0.90	0.91.0	5.5.1	
	Hadoop & YARN	Oozie	Pig	Hive	Druid	Tez	Calcite	Sqoop	Spark	Zeppelin	HBase	Phoenix	Accumulo	Knox	Ranger	Atlas	Storm	Kafka	Ambari	Zookeeper	Flume	Falcon	Mahout	Slider	Solr	
	HDP Core		Enterprise Data Warehouse				Data Science			Operational Data Store			Security Governance			Stream Processing		Operations		Removed/Moved Components				HDP Search		
Hortonworks Data Platform																										

[1] HDP 2.6 – Shows current Apache branches being used. Final component version subject to change based on Apache release process.

[2] Spark 1.6.3+ Spark 2.1 – HDP 2.6 supports both Spark 1.6.3 and Spark 2.1 as GA.

[3] Hive 2.1 is GA within HDP 2.6.

[4] Apache Solr is available as an add-on product HDP Search.

[5] Spark 2.2 is GA

Koloběh technologií

- › Nadšení
 - Našel jsem skvělou technologii ! Vyřeší naše problémy.
- › Realita – o několik hodin později
 - Kde je dokumentace?
- › Vystřízlivění – podle povahy o několik hodin, až jednotek dní později
 - Ono to asi vážně nefunguje.
- › Zklamání – o několik zoufalých dnů později
 - Nefungují ani příklady na webu ! Kdo probůh tohle vyvíjí ?
 - Porozhlédneme se tedy jinde...

The background is a dark gray field filled with a complex, overlapping pattern of translucent, light gray polygons. These polygons vary in size and orientation, creating a sense of depth and movement. In the center of the image, the word "YARN" is written in a clean, white, sans-serif font.

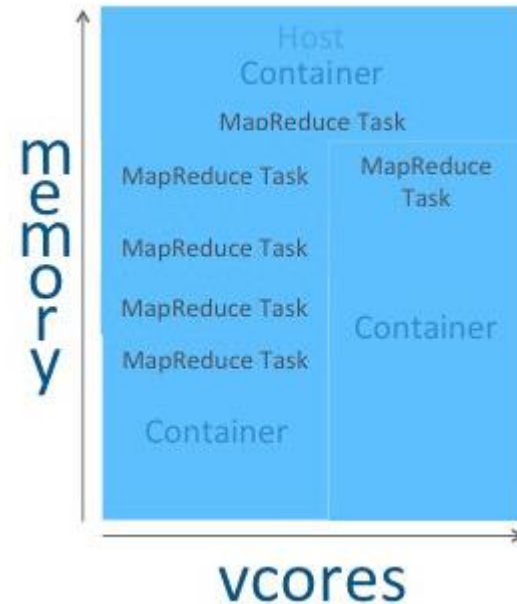
YARN

YARN

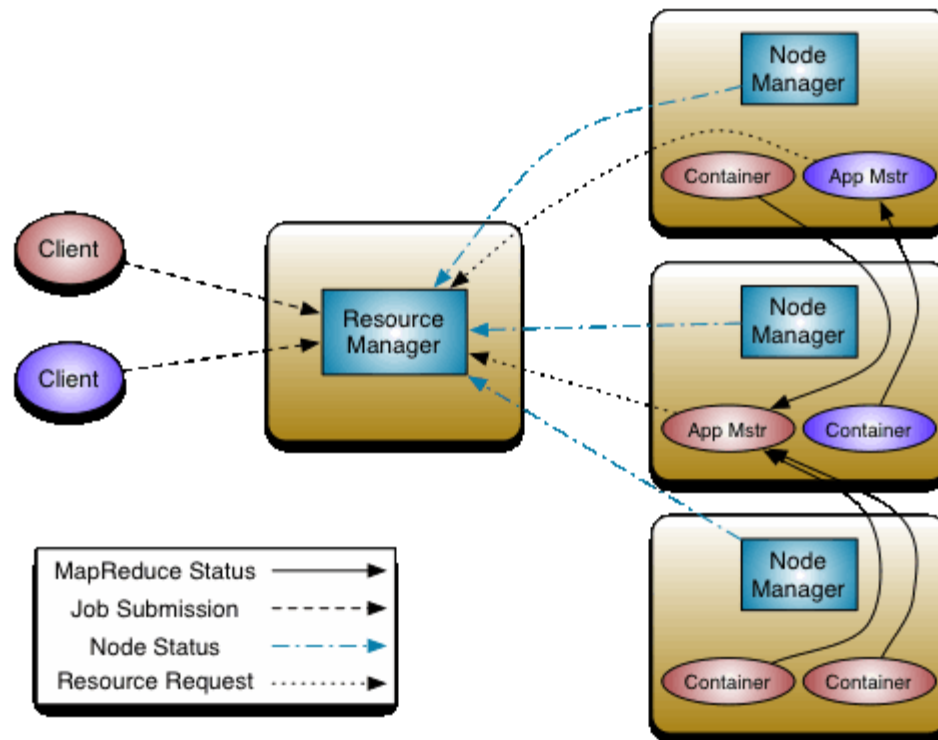
- › **Yet Another Resource Negotiator**
- › tj. plánovač a alokátor zdrojů
 - paměť
 - CPU
 - počet vláken
 - síť...
- › Většinou se využívá transparentně, uživatel o něm moc neví, záleží ale hodně na konfiguraci
- › Ne všechny aplikace YARN využívají, např. Impala má vlastní plánovač
 - každý alokátor by tak měl mít výhradní zdroje...

YARN

- › Application – klientská aplikace
- › Container – zdroje přiřazené aplikaci na konkrétním nodu
- › Resource Manager – globální správce zdrojů pro cluster
- › Node Manager – podřízený správce zdrojů na nodu



YARN



Díky za pozornost

PROFINIT

Profinit, s.r.o.
Tychonova 2, 160 00 Praha 6



Telefon
+ 420 224 316 016



Web
www.profinit.eu



LinkedIn
linkedin.com/company/profinit



Twitter
twitter.com/Profinit_EU