

DĚLAT
DOBRÝ SOFTWARE
NÁS BAVÍ

PROFINIT

B0M33BDT Architektury a bezpečnost

Marek Sušický

4. prosinec 2019

Osnova

- › Něco ze života
- › Architektury
 - Hadoop
 - Lambda
 - Kappa
 - Zetta
- › Security a dopady do architektury

Jak vypadá Hadoop?

› Yahoo



Jak vypadá Hadoop?

› Facebook



Jak vypadá Hadoop?

› Google



Jak vypadá Hadoop?



Několik otázek

- › Jaká je rychlost světla v optickém kabelu?
- › Jaká je akceptovatelná latence pro telefonní hovor?
- › Kolik událostí za sekundu zvládnou konvenční velké databáze?
- › Kolik stojí malý clusteřík? (5x 2x 10core, 256GB RAM, 10x2TB HDD)

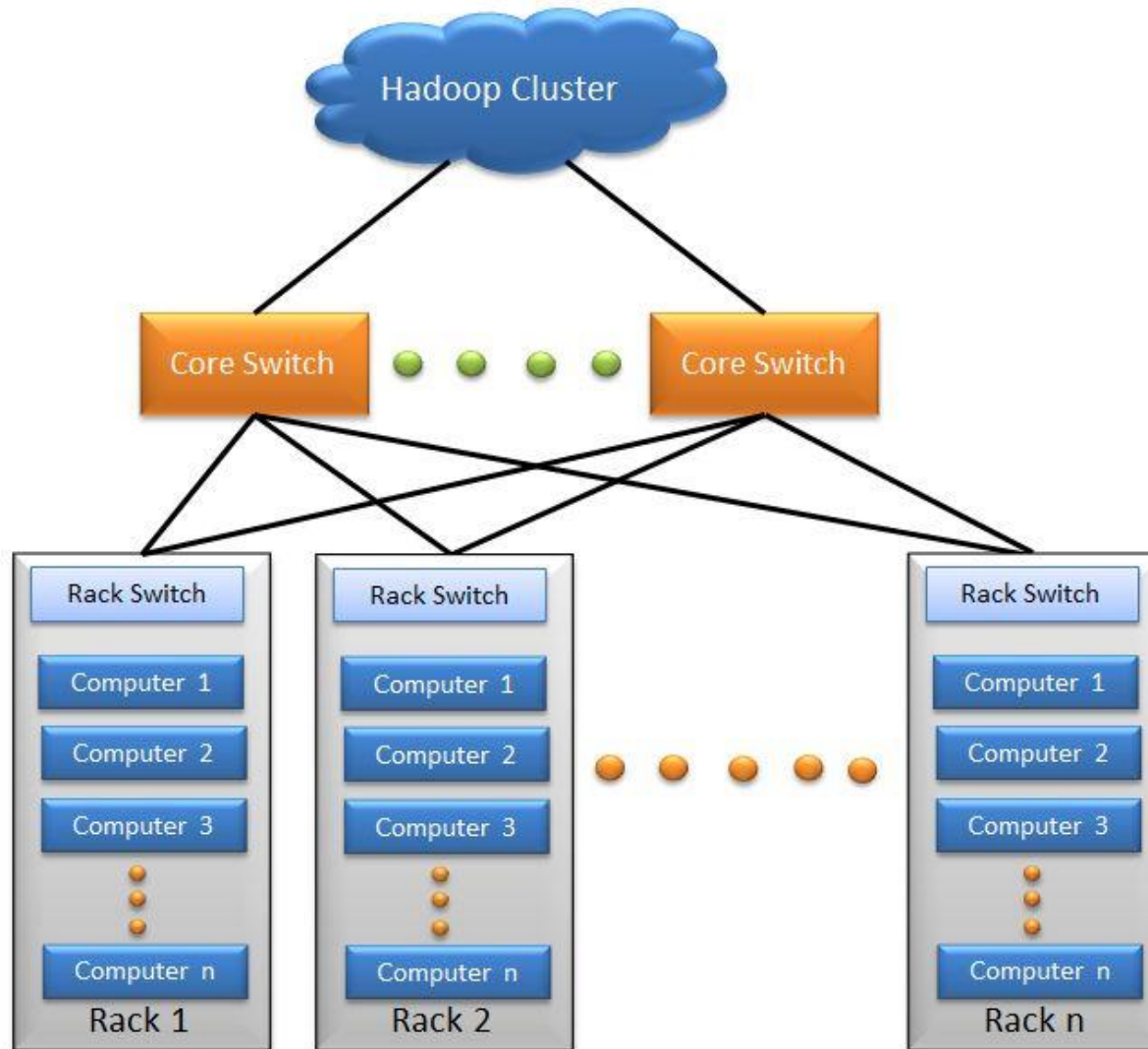
Několik otázek

- › Jaká je rychlost světla v optickém kabelu?
 - 200 000km/s
- › Jaká je akceptovatelná latence pro telefonní hovor?
 - 50ms
- › Kolik událostí za sekundu zvládnou konvenční velké databáze?
 - Cca 10 000
- › Kolik stojí malý clusteřík? (5x 2x 10core, 256GB RAM, 10x2TB HDD)
 - Pod 5 M

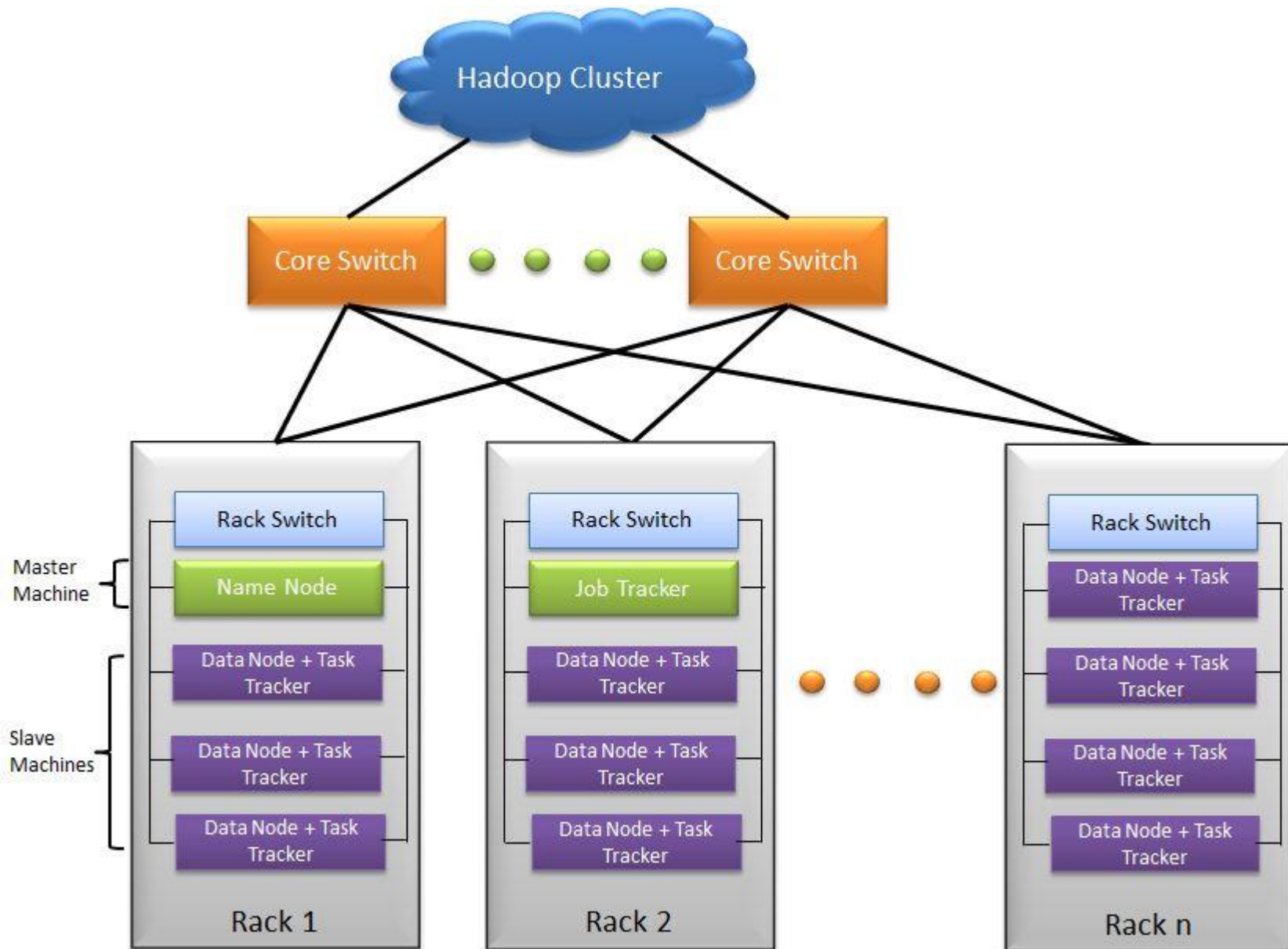
The background of the image is a dark gray field filled with a complex, overlapping pattern of translucent, light gray geometric shapes. These shapes, which include various polygons and rectangles, are layered in a way that creates a sense of depth and movement, resembling a crystalline or architectural structure. The word "Architektury" is centered in the middle of the image in a white, sans-serif font.

Architektury

Hadoop

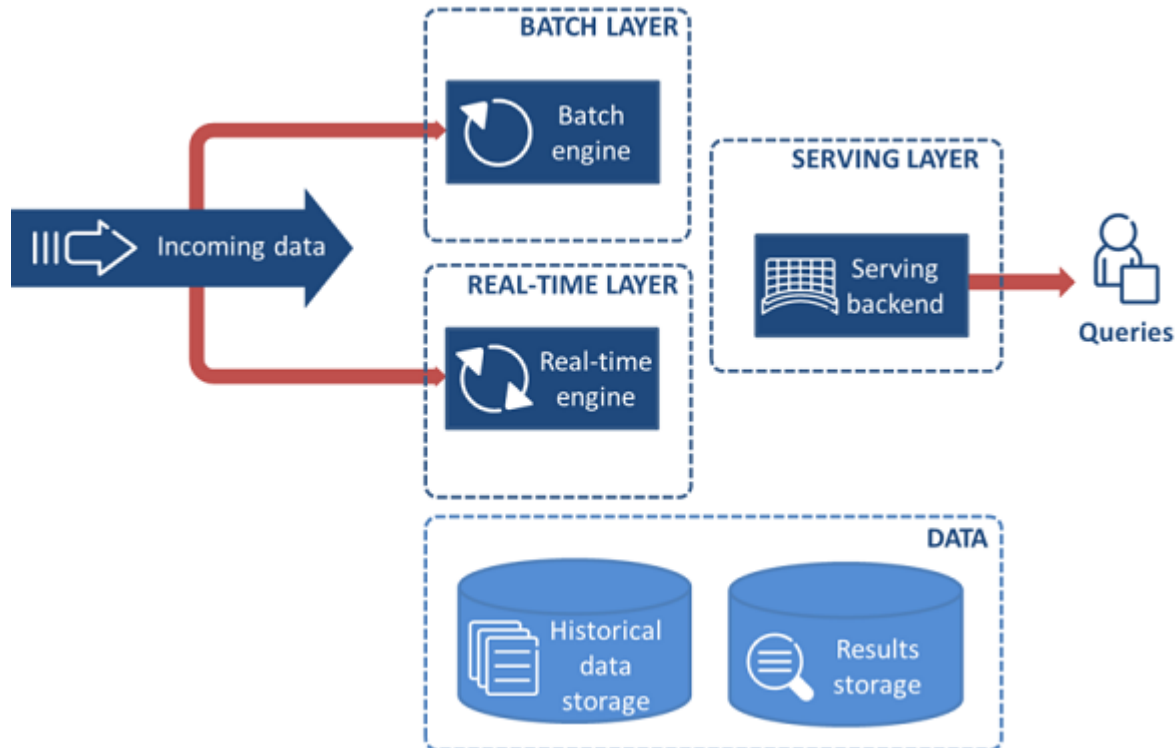


Hadoop



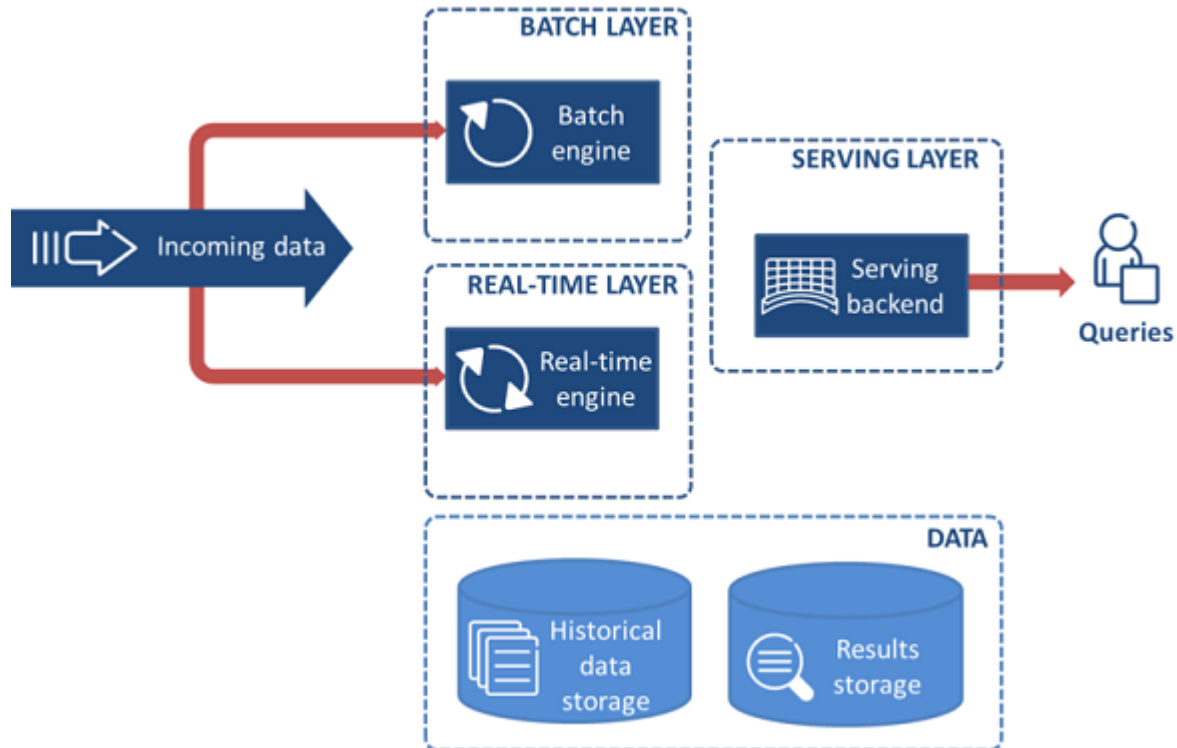
Lambda

- › Z Apache Storm
- › Nathan Marz, 2011
- › <http://nathanmarz.com/blog/how-to-beat-the-cap-theorem.html>
- › Yahoo, Netflix



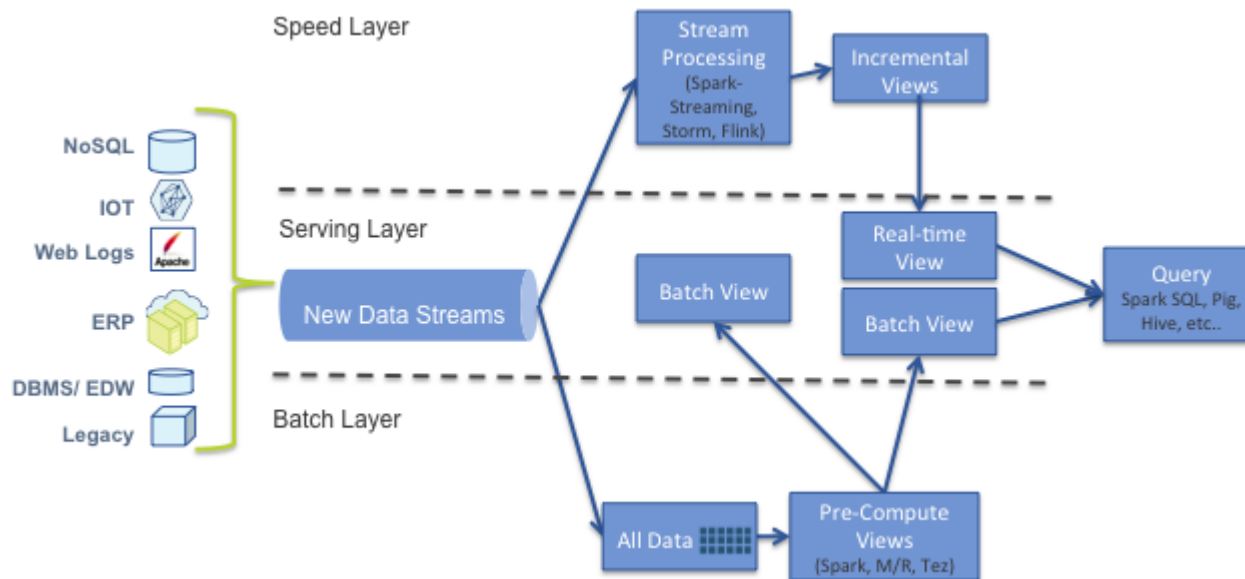
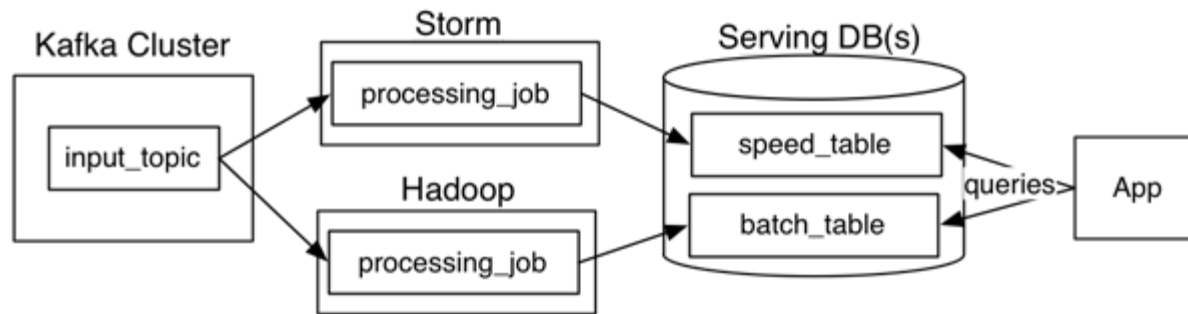
Lambda

› 4 vrstvy



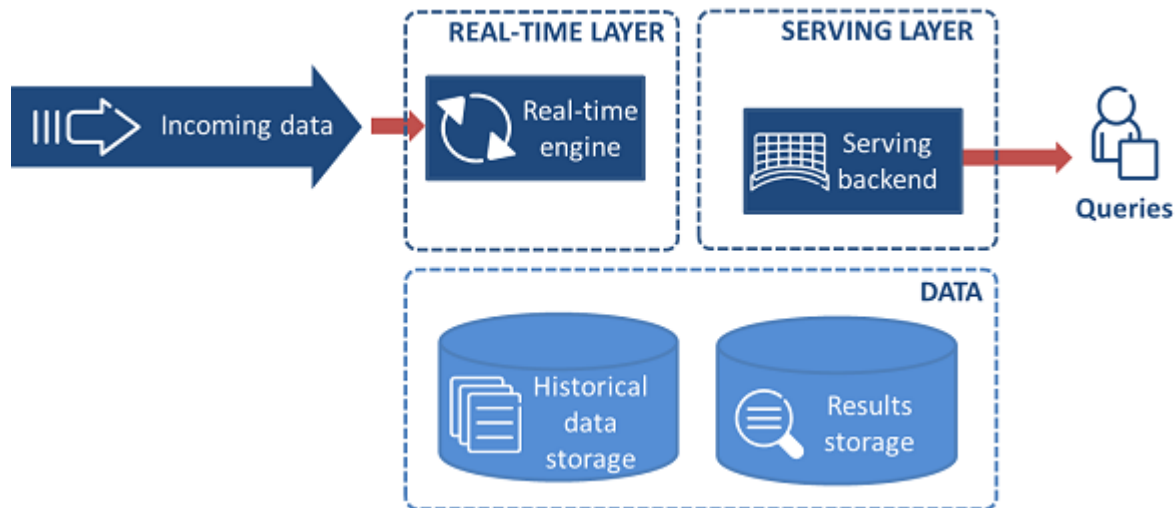
Lambda

› Konkrétní technologie



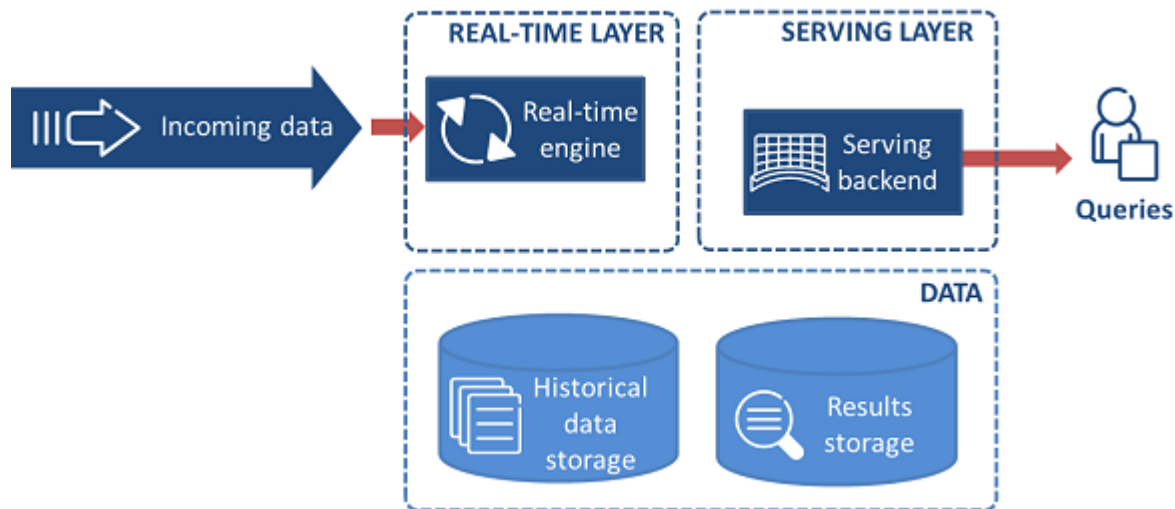
Kappa

- › 2014 Jay Kreps – LinkedIn
- › <https://www.oreilly.com/ideas/questioning-the-lambda-architecture>



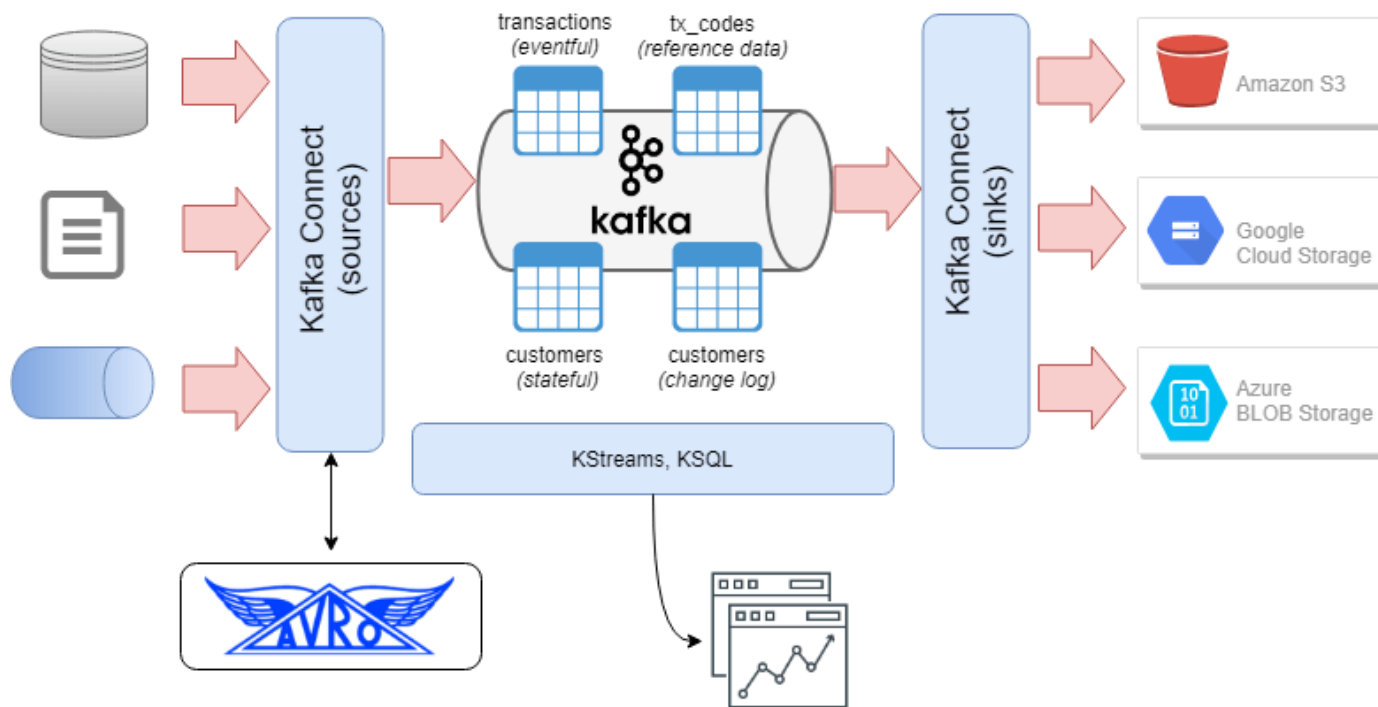
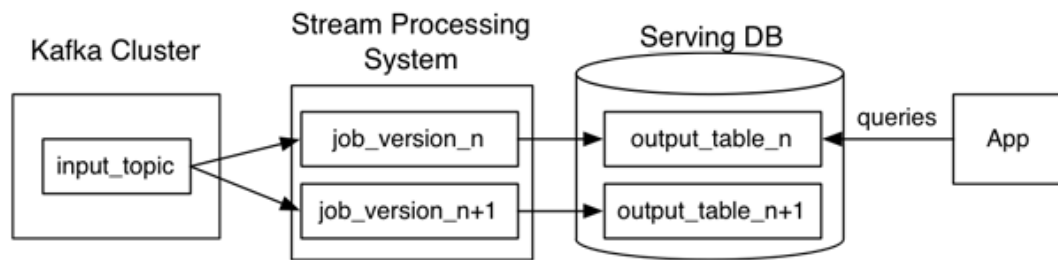
Kappa

- › 3 vrstvy – odstranění batch vrstvy
- › Lze použít dlouhou retenci
- › Problém se stavem – microbatche?



Kappa

› Konkrétní technologie

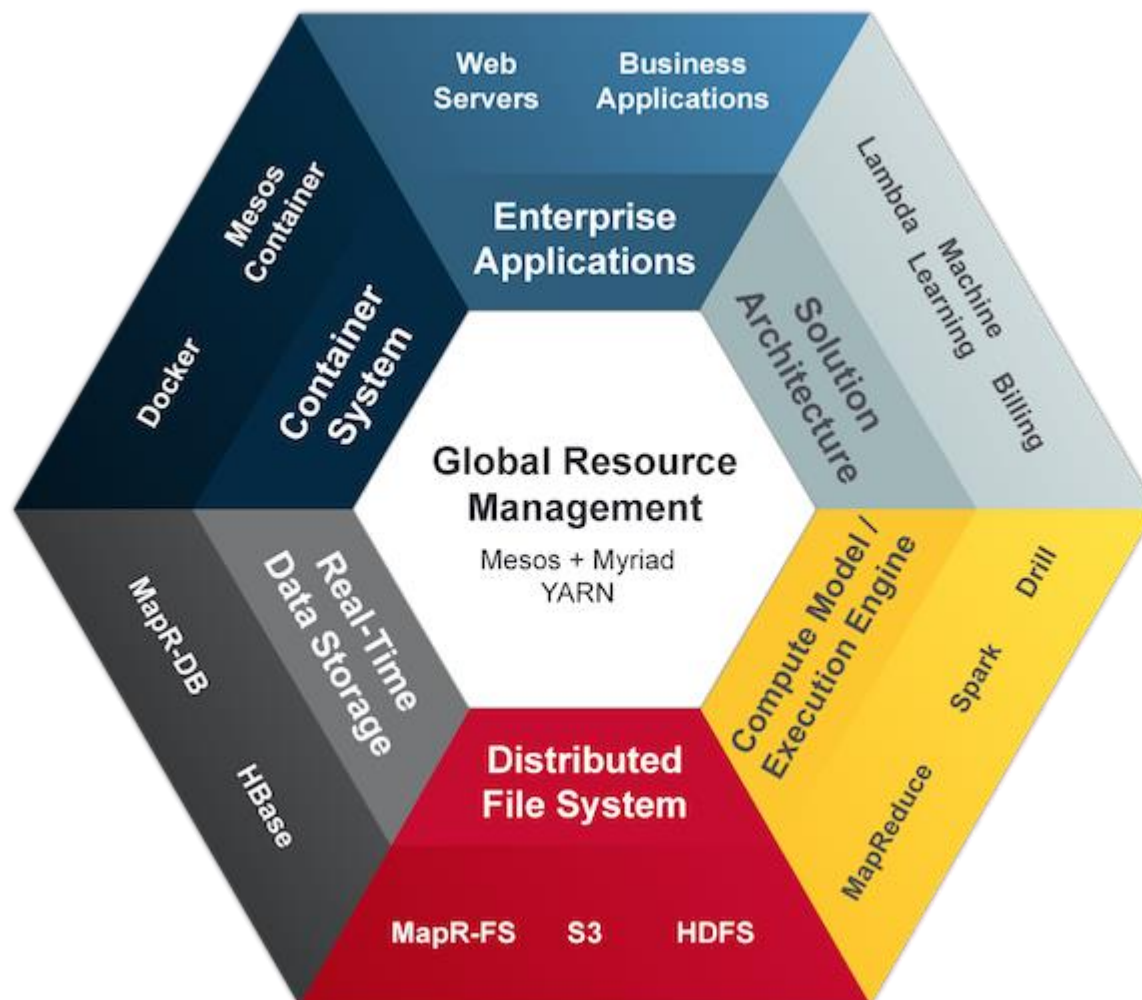


Kappa

- › Tooly
- › Log data store
 - Kafka
- › Streaming computation systems
 - Samza
 - Storm
 - Kafka Streams
 - Flink

Zeta

- › Jim Scott – MapR, 2015
- › (Zeta je 6 číslo řecké abecedy), data-centric



Zeta

› Google Zeta



Zeta

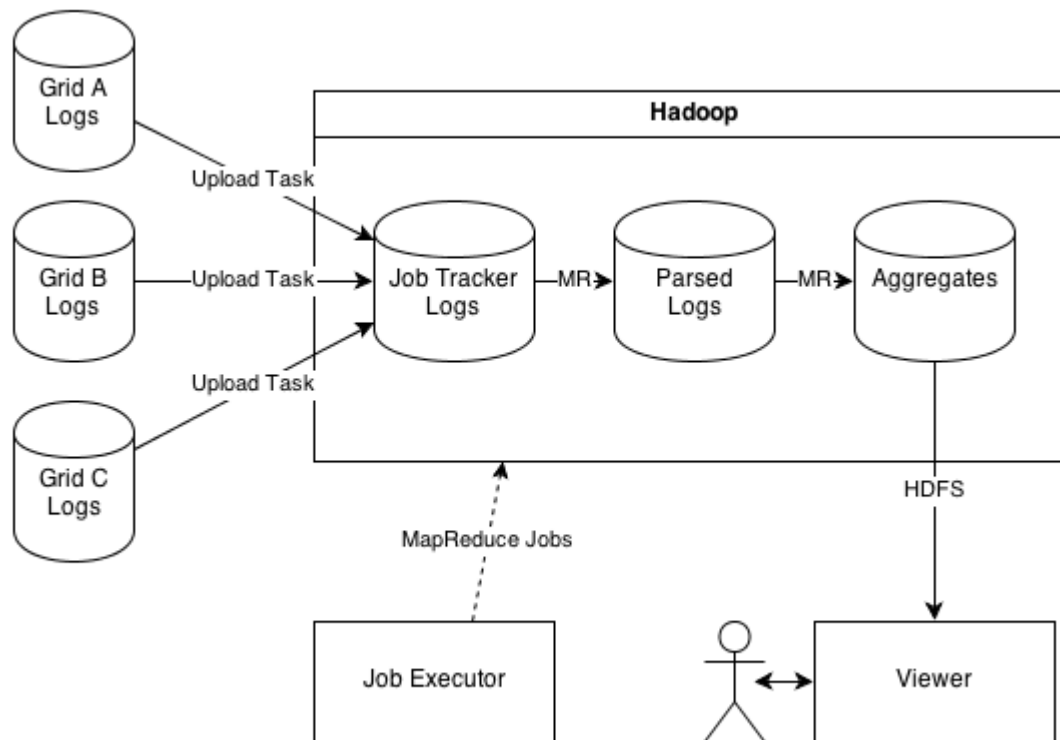
- › Co to znamená?
- › Všechno na Mesos
- › Dynamická alokace zdrojů
- › Omezení přesunů dat
- › Zatím spíš okrajová, resp. nejsou články ani zmínky

Několik ukázek ze života

- › Sběr logů
- › Reklamní platforma
- › DWH Offloading
- › Analytické pískoviště

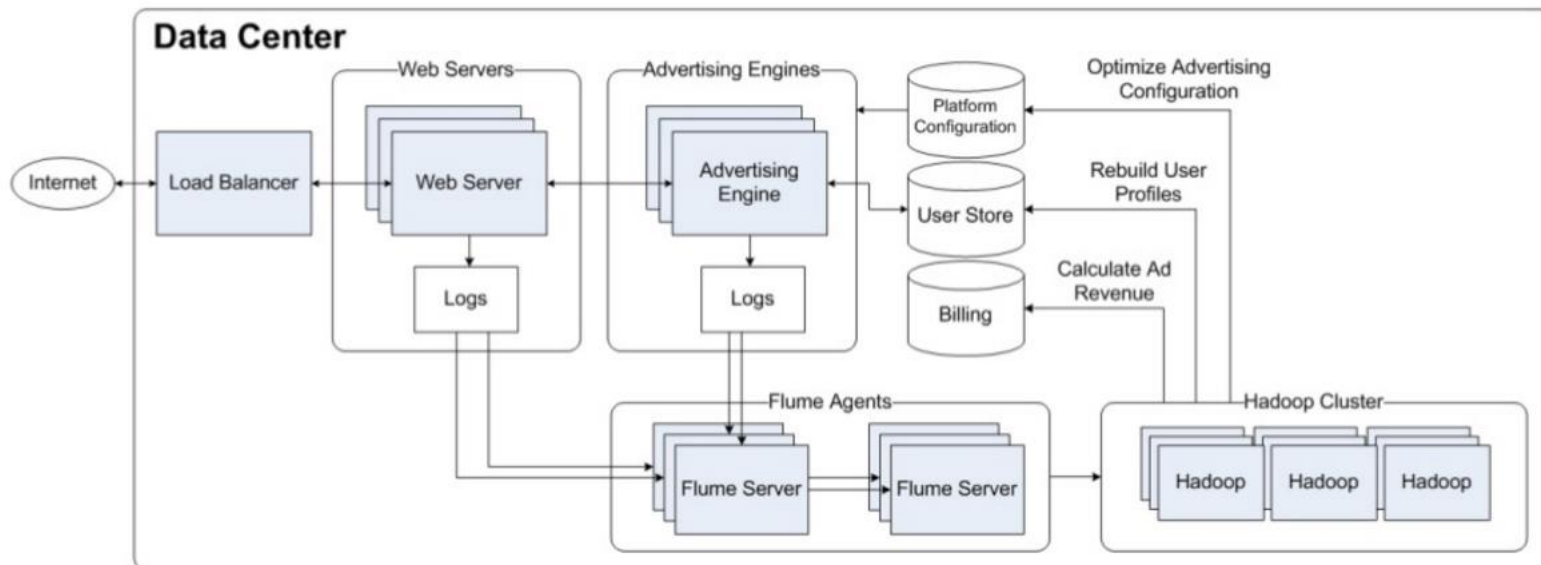
Sběr logů

- › Web server tvoří logy
- › Ty se ukládají na disk – rotace
- › Pak se posílají na jiné servery
- › Logy se zpracovávají



Reklamní platforma

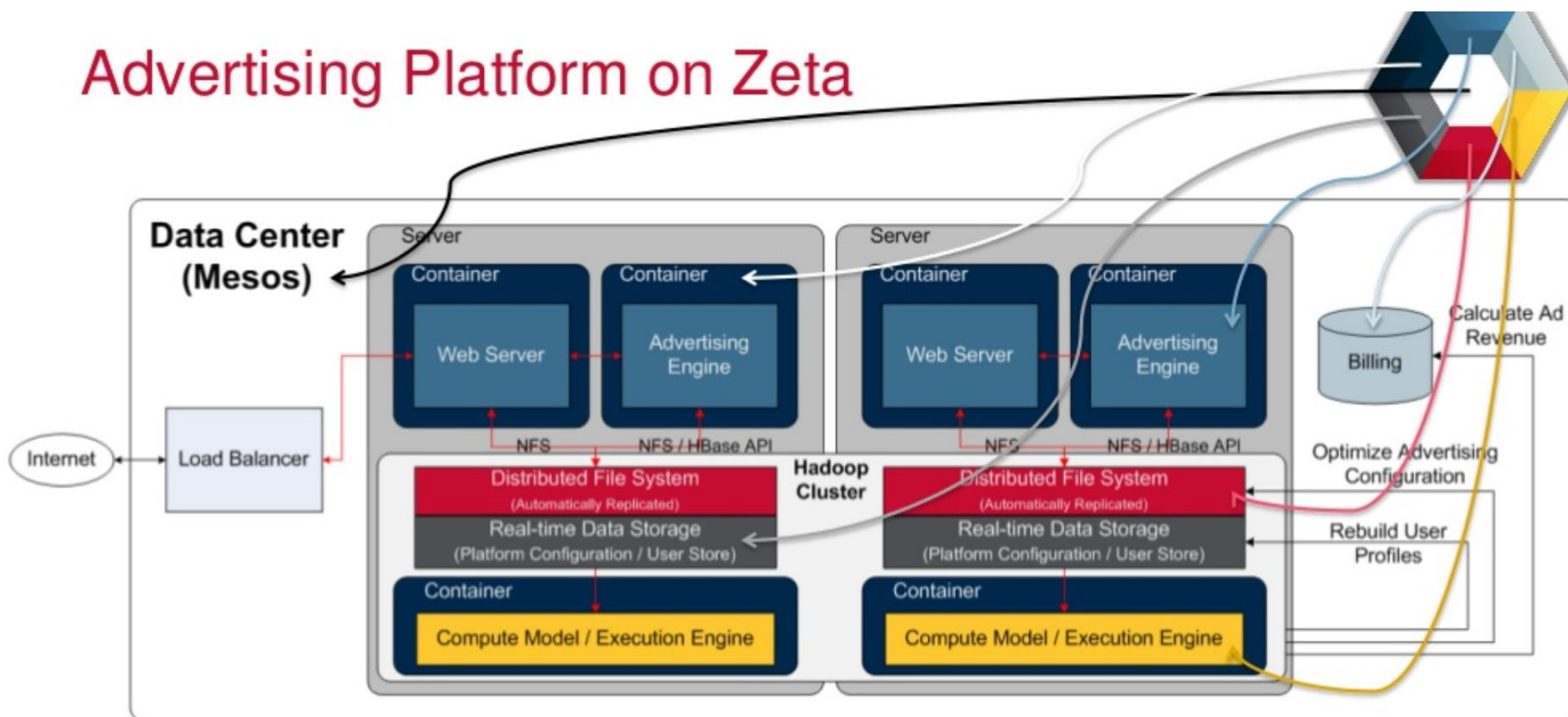
- › Web logy a informace o zobrazování reklamy
- › Logy Flumem do HDFS
- › Pak počítáme a vracíme zpět na znovuzpracování



Reklamní platforma

- › Web logy a informace o zobrazování reklamy
- › Logy Flumem do HDFS
- › Pak počítáme a vracíme zpět na znovuzpracování

Advertising Platform on Zeta



DWH offloading

- › Aktivní archiv
- › Počítání náročných agregací
- › Levnější úložiště dat

- › Typicky Sqoop
- › Flume
- › ETL řešené v Hive, Sparku, nebo přes nástroje třetích stran

DWH offloading

- › Je možné dělat vrstvy jako v normálním DWH
- › Tzn. L0, L1, L2
- › ETL řízené např. pomocí Oozie
- › Často ale komerční nástroje – Talend ETL, Informatica BDM, Oracle ODI
- › Ne vždy to je ale výhra

Analytické pískoviště

- › Data nahrávána většinou ad-hoc
- › Standardní přísun dat přes Flume/Sqoop/scp
- › Velkou roli má R, python a Spark (pySpark)
- › Využití toolů jako Zeppelin, Jupyter, Hue Notebook, či Cloudera Workbench
- › Většinou se moc neřeší bezpečnost

The background of the slide is a dark gray field filled with numerous overlapping, translucent, light gray geometric shapes. These shapes, which include rectangles, parallelograms, and polygons of various sizes and orientations, create a complex, layered effect that suggests architectural structures or a digital network. The text is centered over this background.

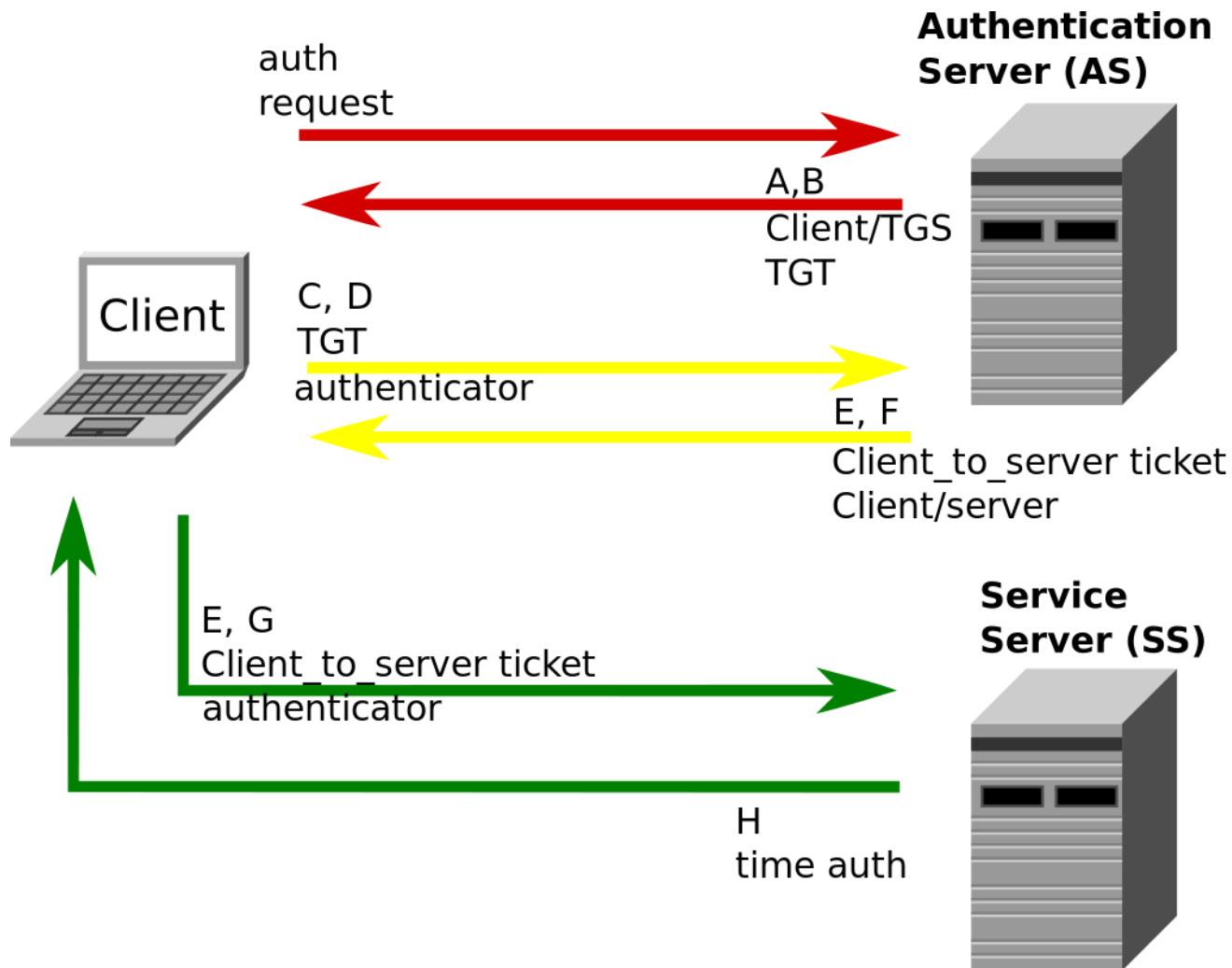
Security a vliv na architekturu

Kerberos

- › Lze rozdělit na tři části
 - KDC – Kerberos Distribution Center
 - Server – poskytuje služby
 - Klienti – uživatelé, počítače, služby
- › KDC nabízí
 - AS – Autentizační server
 - TGS - Ticket Granting Service – řídící server
- › Pojmy
 - TGT – Ticket Granting Ticket - – tiket opravňující uživatele ke komunikaci s řídícím serverem
 - Service Server - Servisní středisko

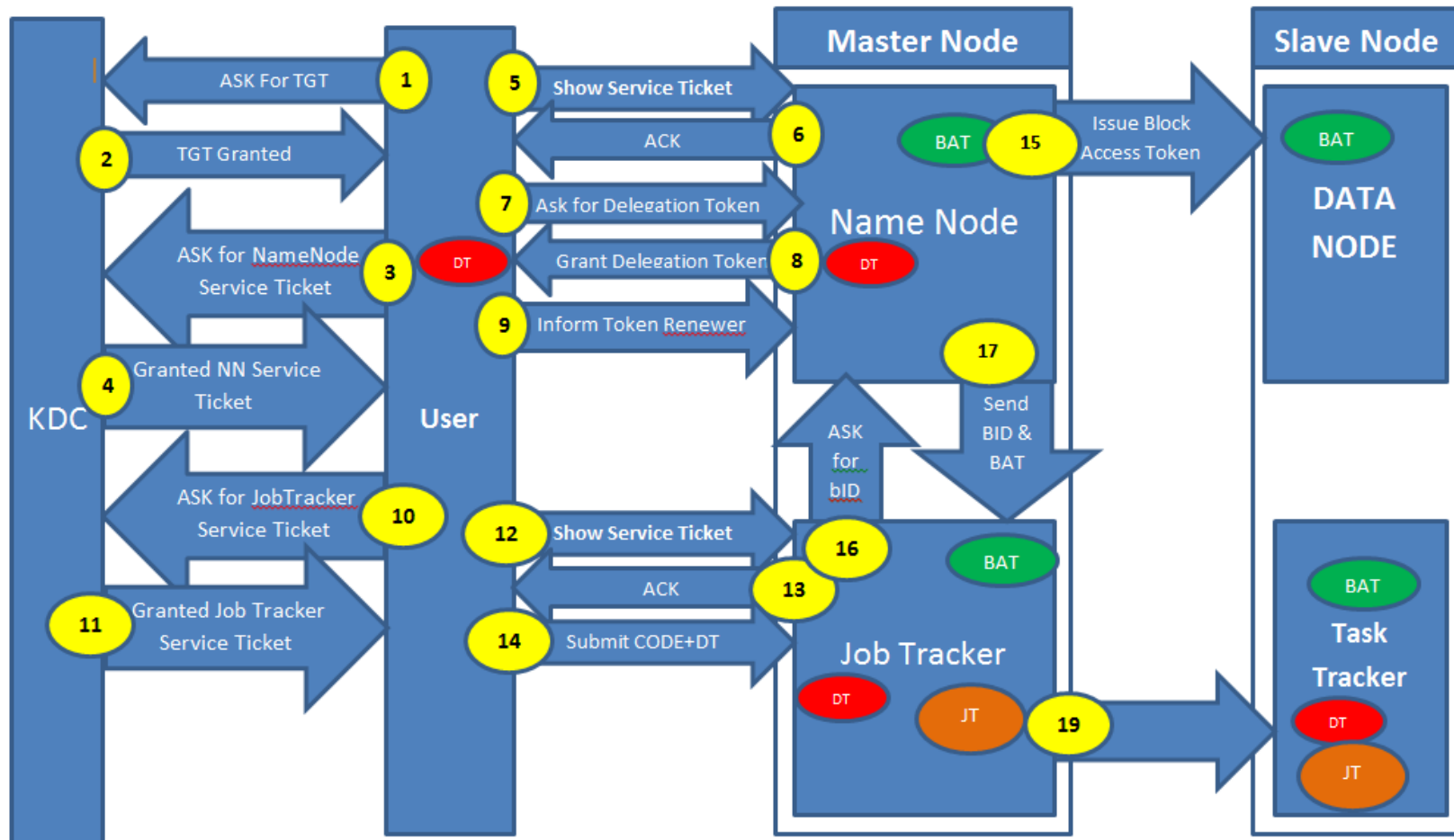
Kerberos

- [https://en.wikipedia.org/wiki/Kerberos_\(protocol\)](https://en.wikipedia.org/wiki/Kerberos_(protocol))



Kerberos

› Jak to funguje



Security

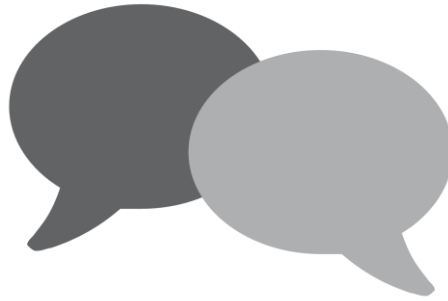
- › HDFS encryption
- › End to end encryption
- › Security komponenty
 - Sentry
 - Ranger

Security

- › Data locality
 - › Data privacy
 - › Data labeling
 - › Expirace dat
-
- › GDPR
 - Anonymizace
 - Pseudonymizace

Data masking

- › Jak získat data pro testovací prostředí?
- › Syntetická data?
- › Jak zajistit byznys relevantnost těchto dat?
- › Jak to udělat výkonné a škálovatelné?



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