DĚLAT
DOBRÝ SOFTWARE
NÁS BAVÍ

PROFINIT

B0M33BDT – 7. přednáška Architektury a bezpečnost

Marek Sušický
Milan Kratochvíl

5. prosinec 2018

Osnova

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

PROFINIT

Jak vypadá Hadoop?

- Yahoo

PROFINIT

Jak vypadá Hadoop?

- Facebook

PROFINIT

Jak vypadá Hadoop?

- Google

PROFINIT

Jak vypadá Hadoop?

Několik otázek

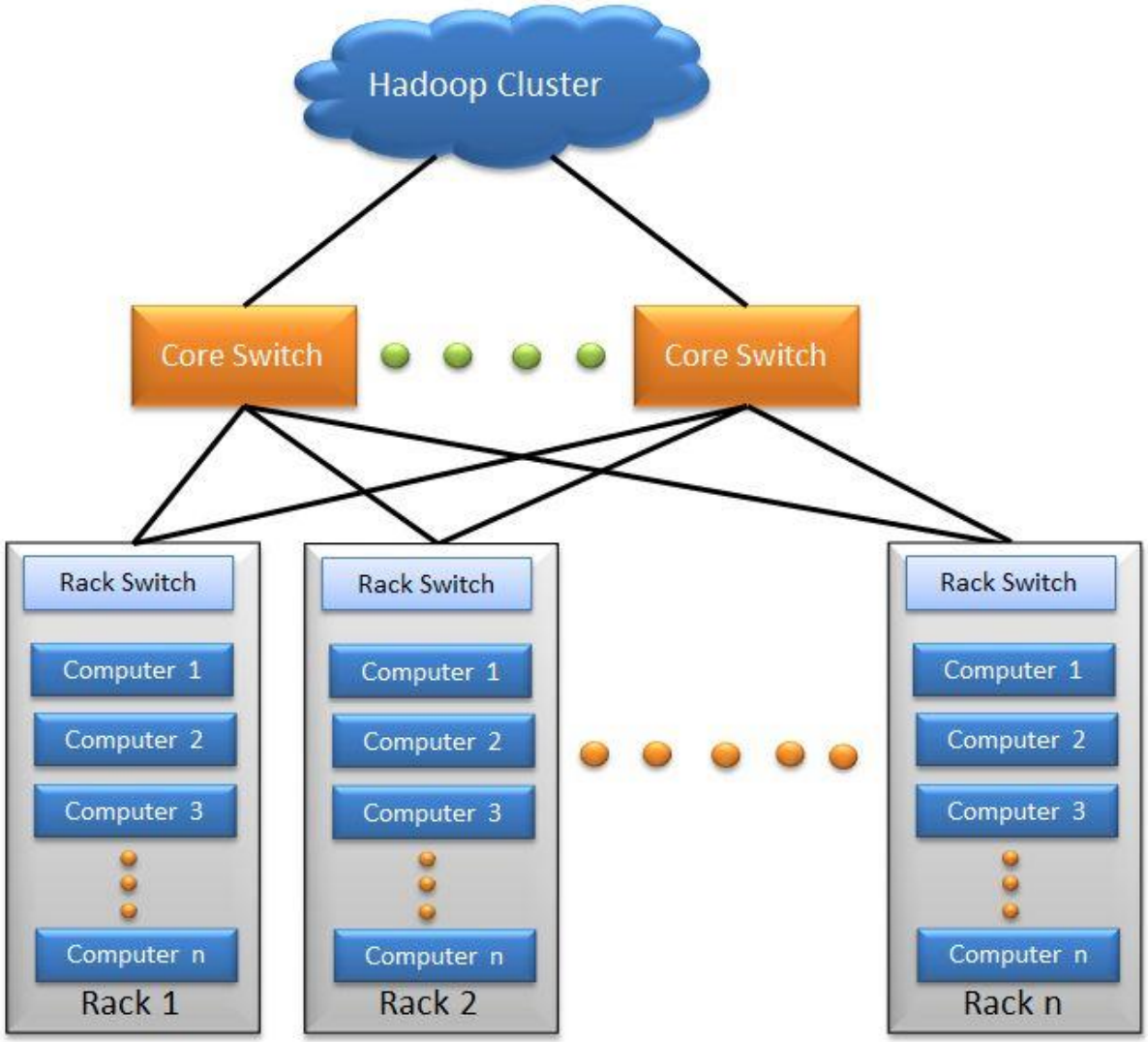
- Jaká je rychlost světla v optickém kabelu?
- Jaká je akceptovatelná latence pro telefonní hovor?
- Kolik událostí za sekundu zvládnou konvenční velké databáze?
- Kolik stojí malý clusteřík? (5x 2x 10core, 256GB RAM, 10x2TB HDD)

Několik otázek

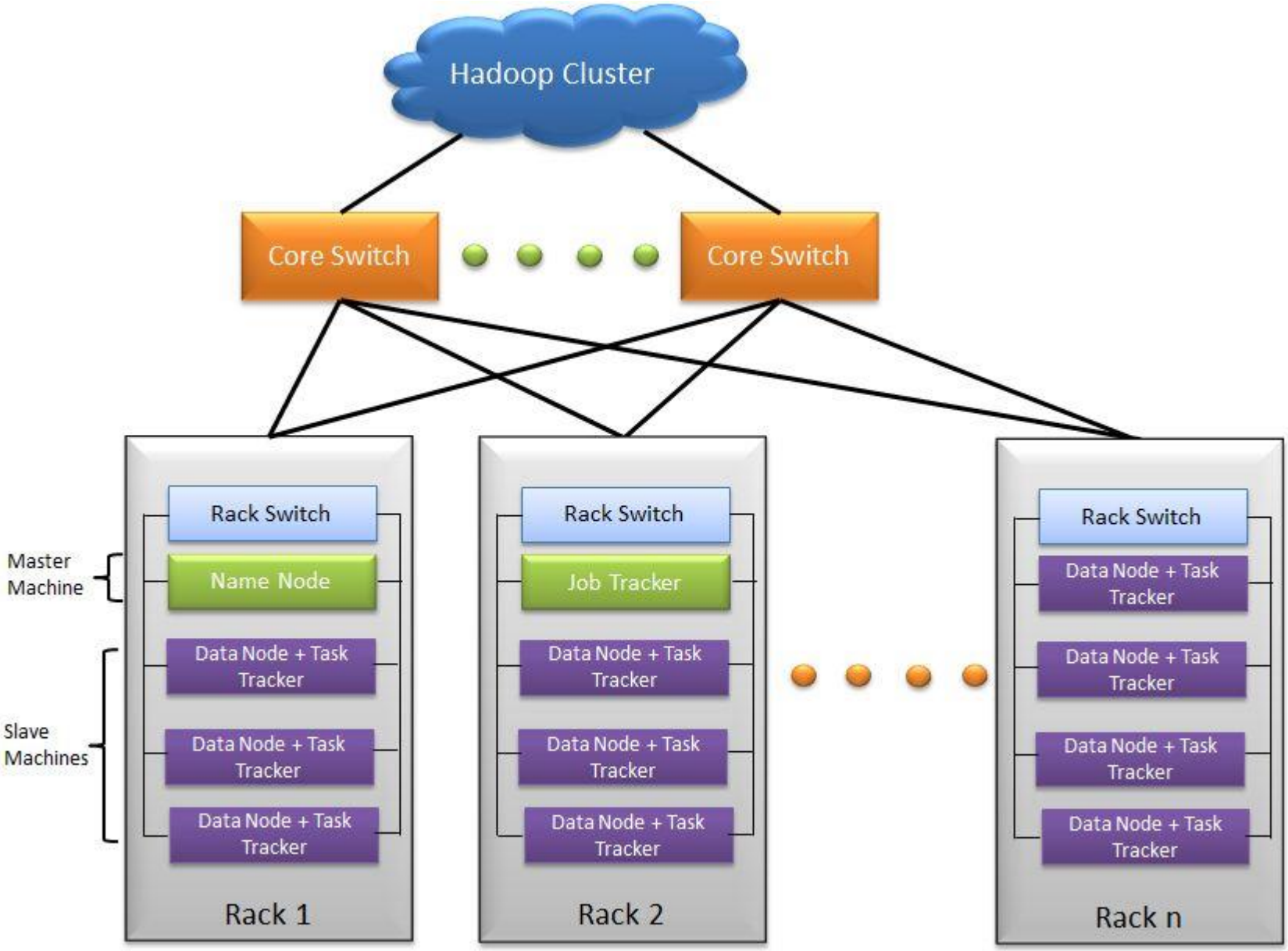
- Jaká je rychlost světla v optickém kabelu?
  - 200 000km/s
- Jaká je akceptovatelná latence pro telefonní hovor?
  - 50ms
- Kolik událostí za sekundu zvládnou konvenční velké databáze?
  - Cca 10 000
- Kolik stojí malý clusteřík? (5x 2x 10core, 256GB RAM, 10x2TB HDD)
  - Pod 5 M

Architektury

Hadoop

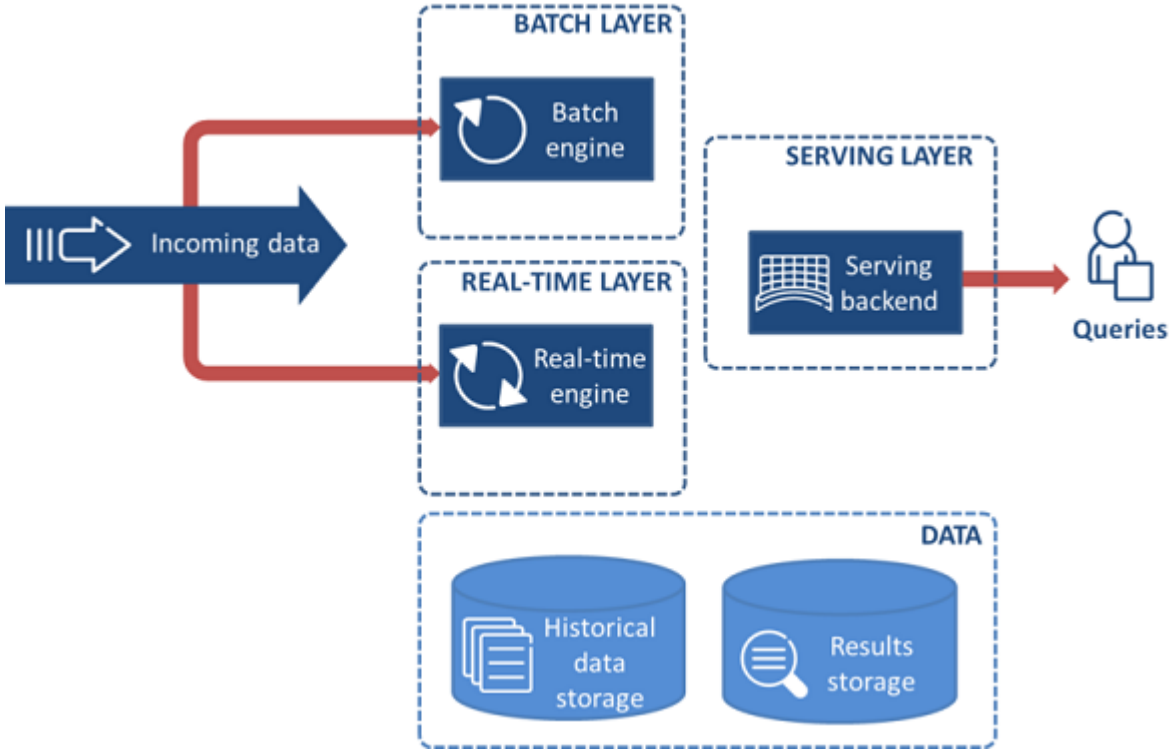

Hadoop Cluster
Core Switch
Core Switch
Rack Switch
Computer 1
Computer 2
Computer 3
Computer n
Rack 1
Rack Switch
Computer 1
Computer 2
Computer 3
Computer n
Rack 2
Rack Switch
Computer 1
Computer 2
Computer 3
Computer n
Rack n

Hadoop


Hadoop Cluster
Core Switch
Core Switch
Rack Switch
Name Node
Data Node + Task Tracker
Data Node + Task Tracker
Data Node + Task Tracker
Master Machine
Slave Machines
Rack 1
Rack Switch
Job Tracker
Data Node + Task Tracker
Data Node + Task Tracker
Data Node + Task Tracker
Rack 2
Rack Switch
Data Node + Task Tracker
Data Node + Task Tracker
Data Node + Task Tracker
Data Node + Task Tracker
Rack n

Lambda

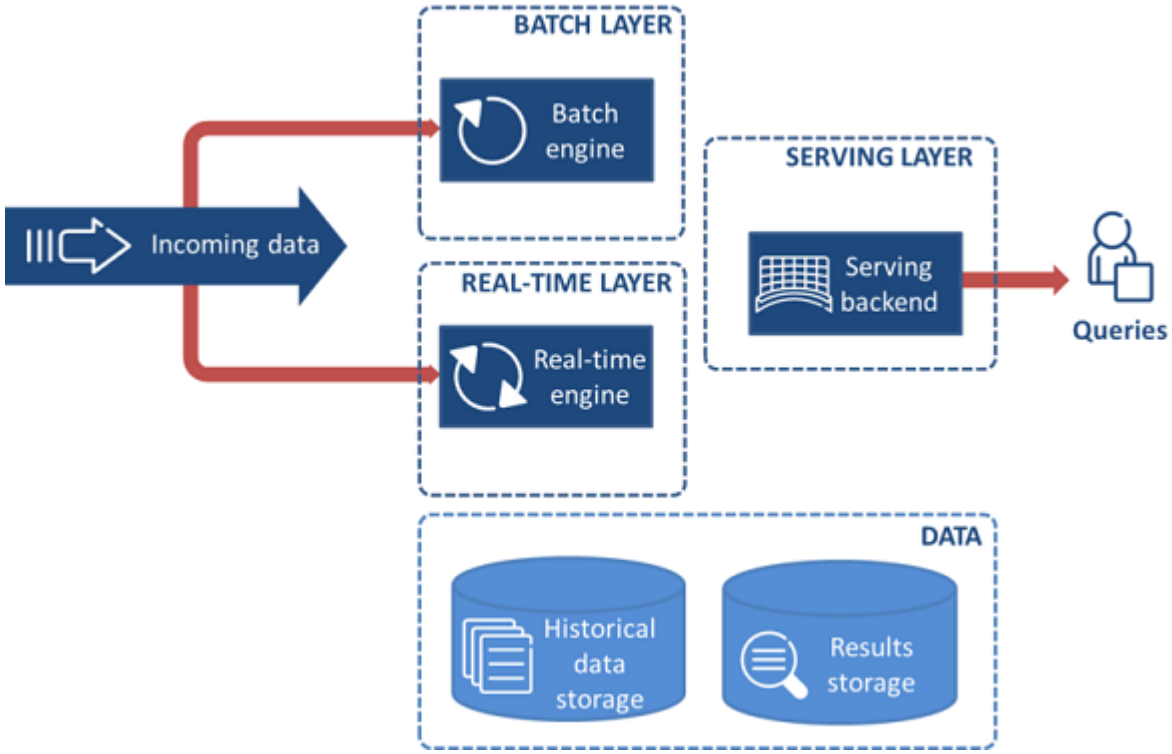
- Z Apache Storm
- Nathan Marz, 2011
- http://nathanmarz.com/blog/how-to-beat-the-cap-theorem.html
- Yahoo, Netflix



PROFINIT

Lambda

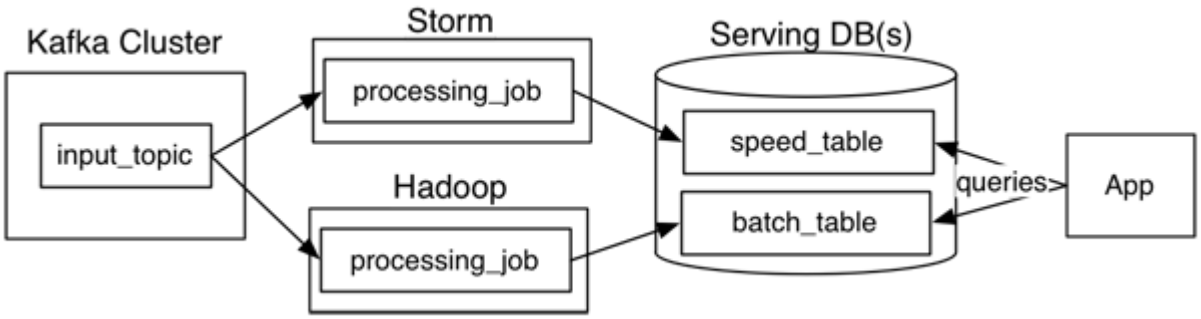
- 4 vrstvy



PROFINIT

Lambda

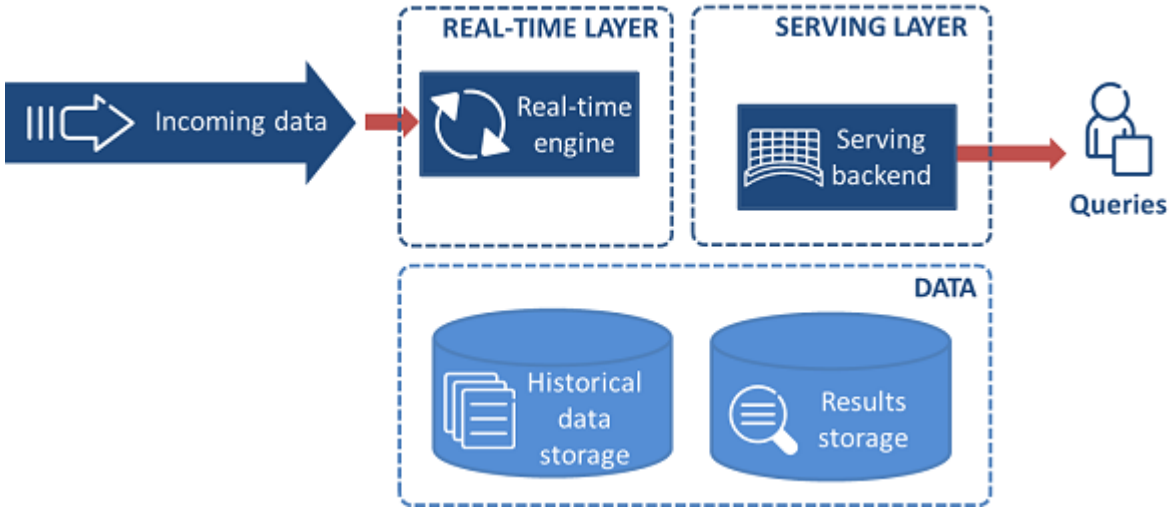
- Konkrétní technologie



PROFINIT

Kappa

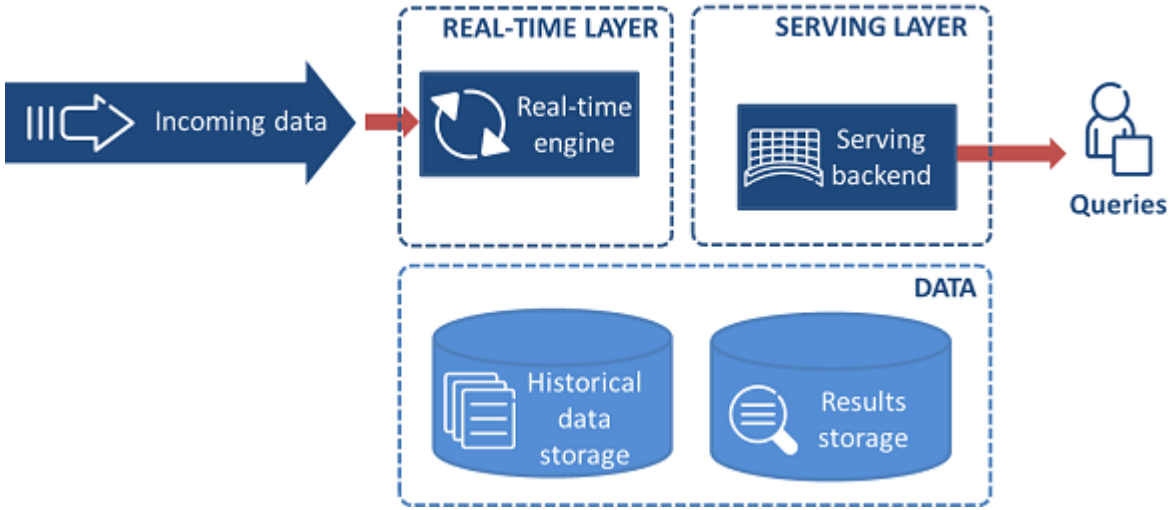
- 2014 Jay Kreps – Linkedin
- https://www.oreilly.com/ideas/questioning-the-lambda-architecture



PROFINIT

Kappa

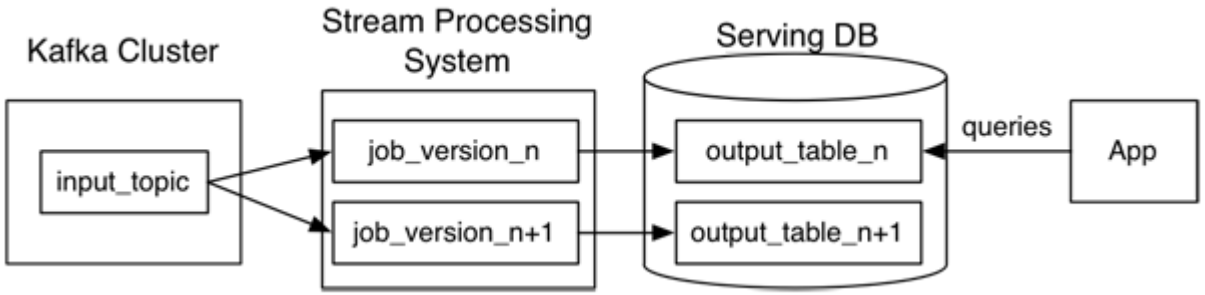
- 3 vrstvy – odstranění batch vrstvy
- Lze použít dlouhou retenci
- Problém se stavem – microbatche?



PROFINIT

Kappa

- Konkrétní technologie



PROFINIT

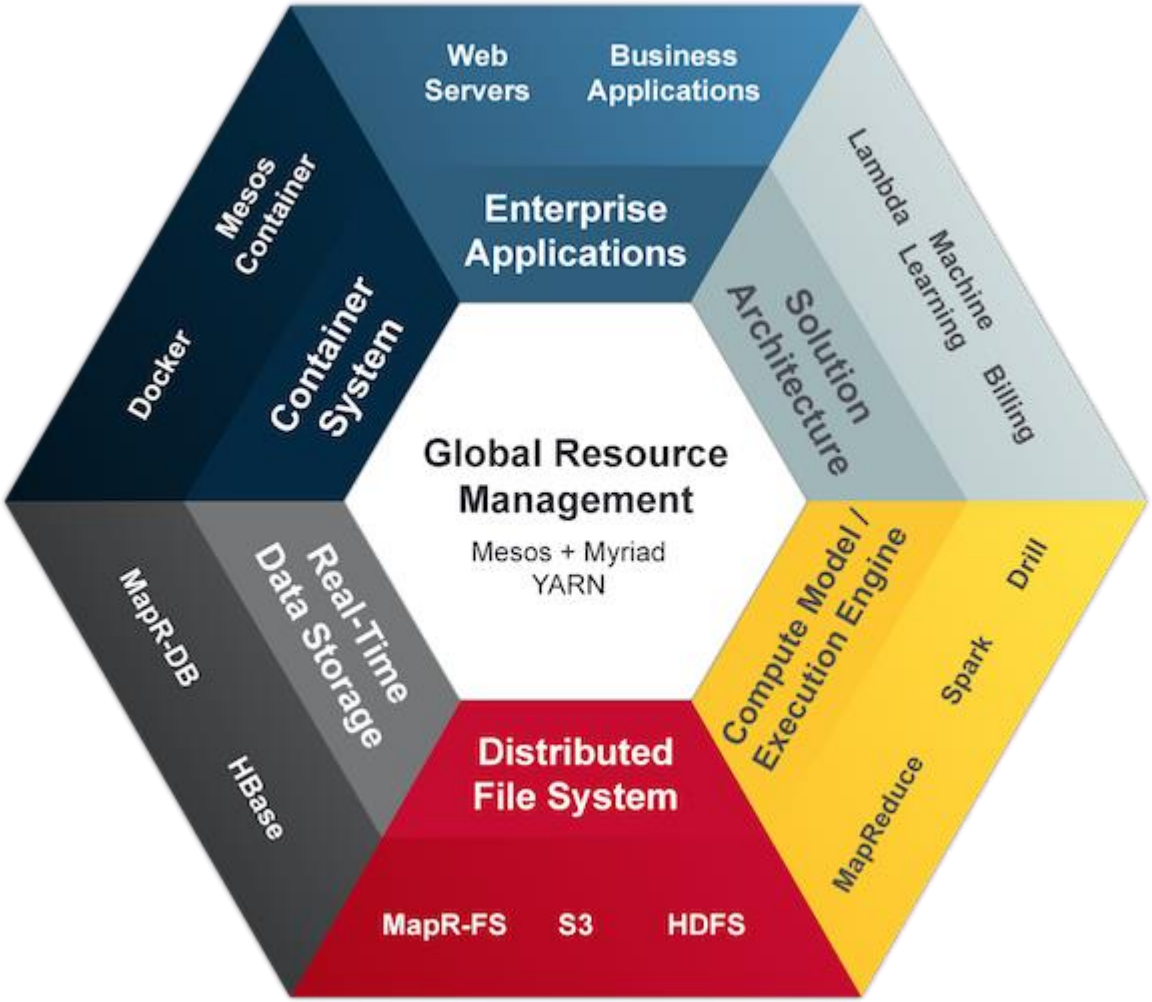
Kappa

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

PROFINIT

Zetta

- Jim Scott - MapR
- (Zetta je 6 číslo řecké abecedy), data-centric



Zetta

- Google Zetta

PROFINIT

Zetta

- Co to znamená?
- Všechno na Mesos
- Dynamická alokace zdrojů
- Omezení přesunů dat
- Zatím spíš okrajová

PROFINIT

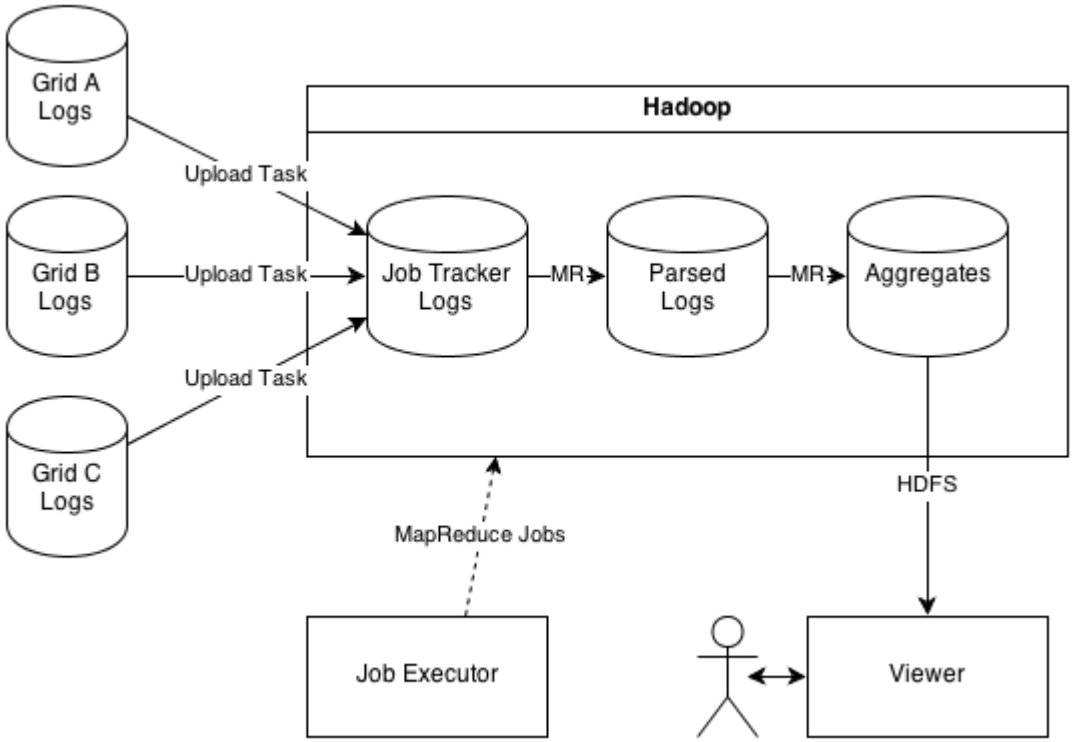
Několik ukázek ze života

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

PROFINIT

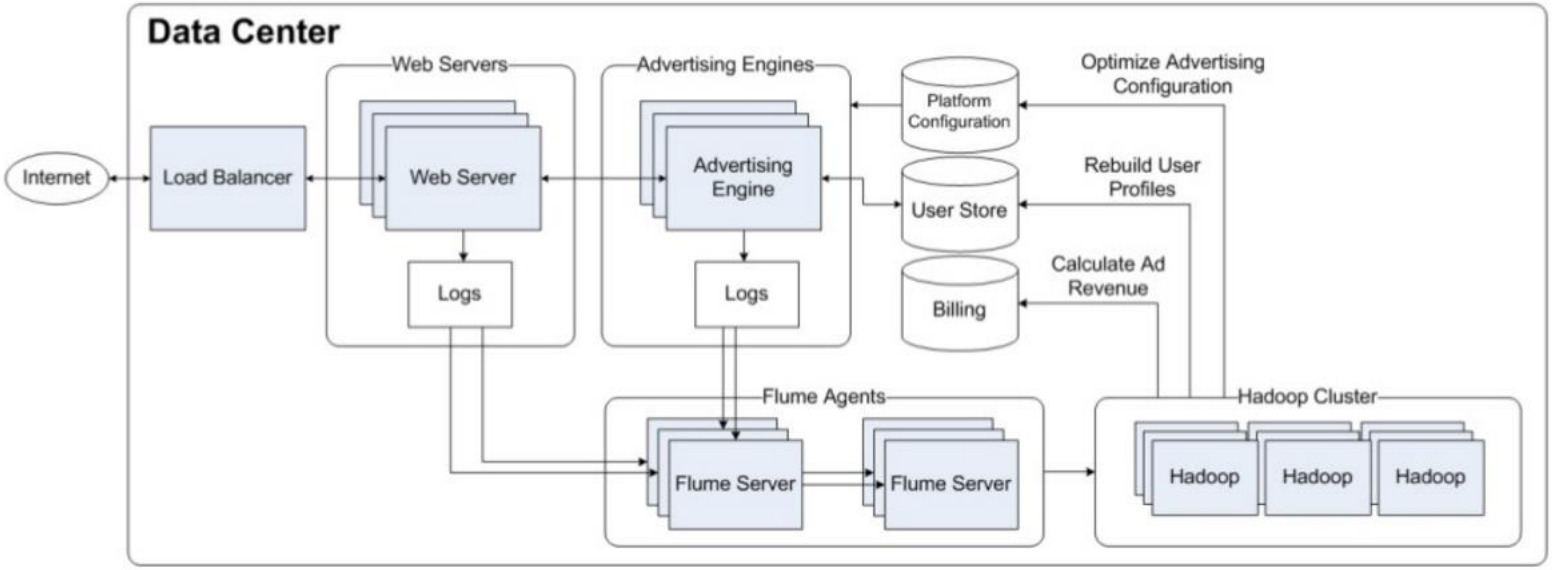
Sběr logů

- 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



DWH offloading

- Aktivní archiv
- Levnější úložiště dat

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

PROFINIT

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

PROFINIT

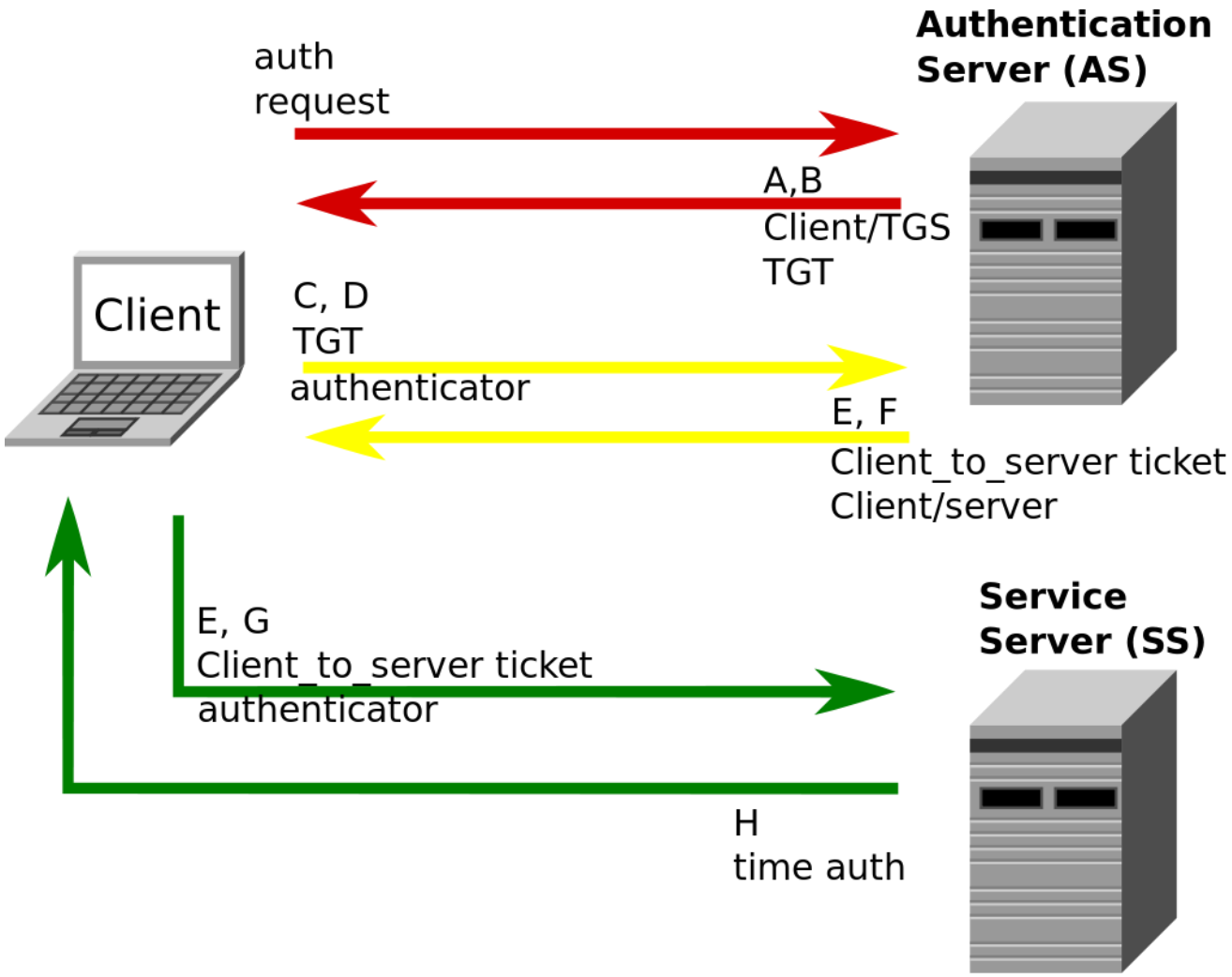
Security a vliv na architekturu

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
- Pojmy
  - TGT – Ticket Granting Ticket
  - Service Server

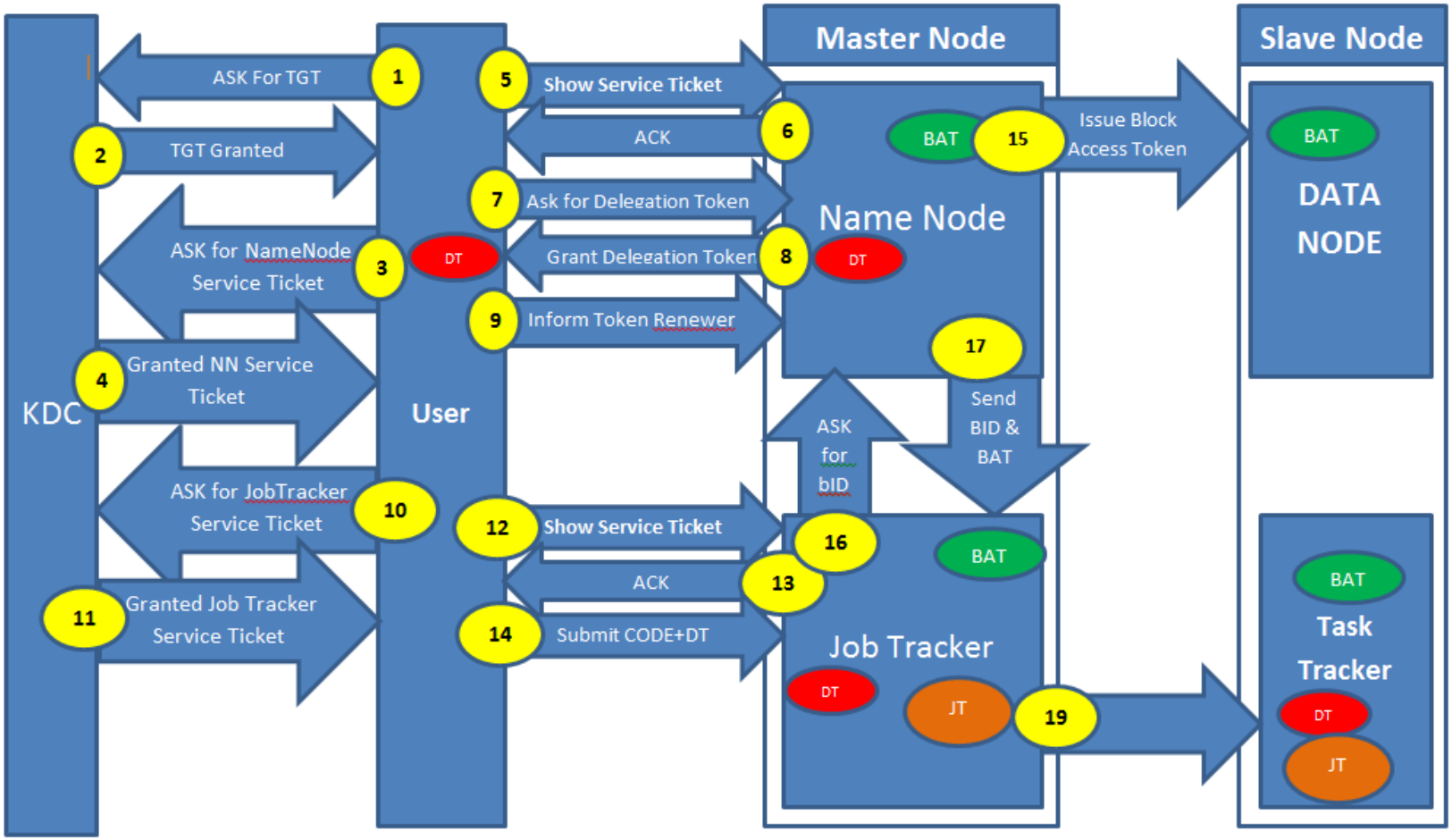
Kerberos

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



Kerberos

- Jak to funguje



Security

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

PROFINIT

Security

- Data locality
- Data privacy
- Data labeling
- Expirace dat

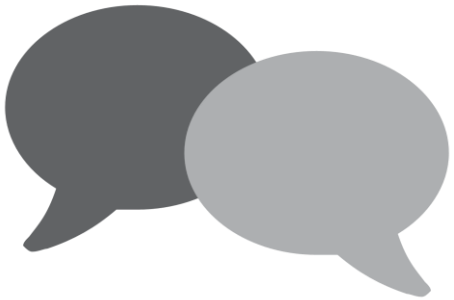
- GDPR
  - Anonymizace
  - Pseudonymizace

PROFINIT

Data masking

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

PROFINIT



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