Petr Nevrlý <petr.nevrly@firma.seznam.cz>

Obsah přednášky

- Vyhledávání
  - Cíl vyhledávání
  - Architektura
  - Vyhledávání
  - Robot
  - Údaje z provozu

- Novinky ve fulltextu za rok 2010
  - Robot v3.0
  - Termové hledání
  - Rozšířené hledaní

Cíl fulltextového vyhledávání

- Poskytnout odpověď na dotaz uživatele

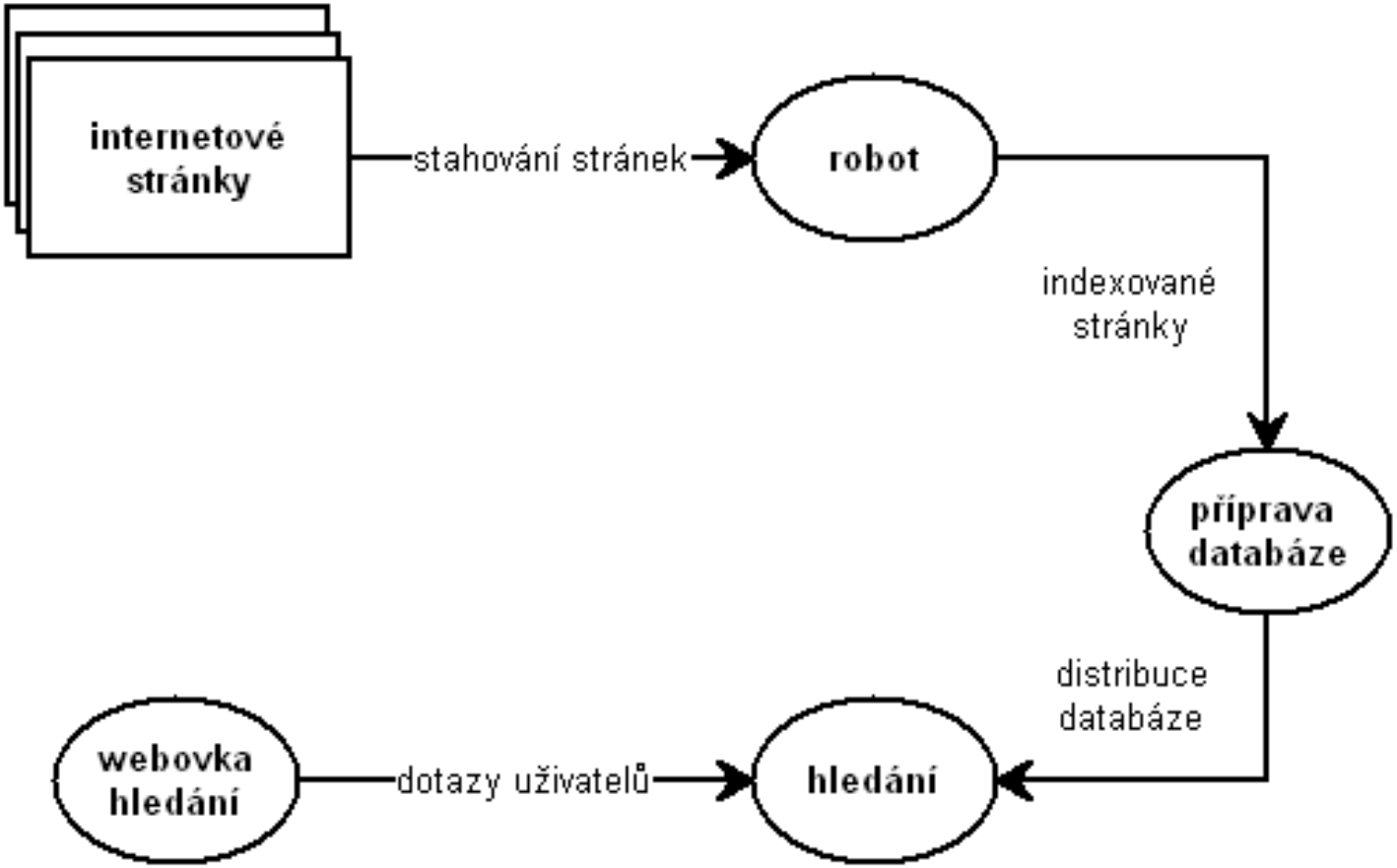
Cíl fulltextového vyhledávání

- Poskytnout odpověď na dotaz uživatele
  - Shromažďování dat
    - Rychlý robot
    - Spolehlivá indexace
    - Zakládání „správných“ dokumentů
  - Zpracování
    - Vhodná architektura (rozšiřitelnost)
  - Vydání, řazení a prezentace
    - Výkon (rychlost)
    - Dostupnost
    - Konzistence
    - Kvalita a UX

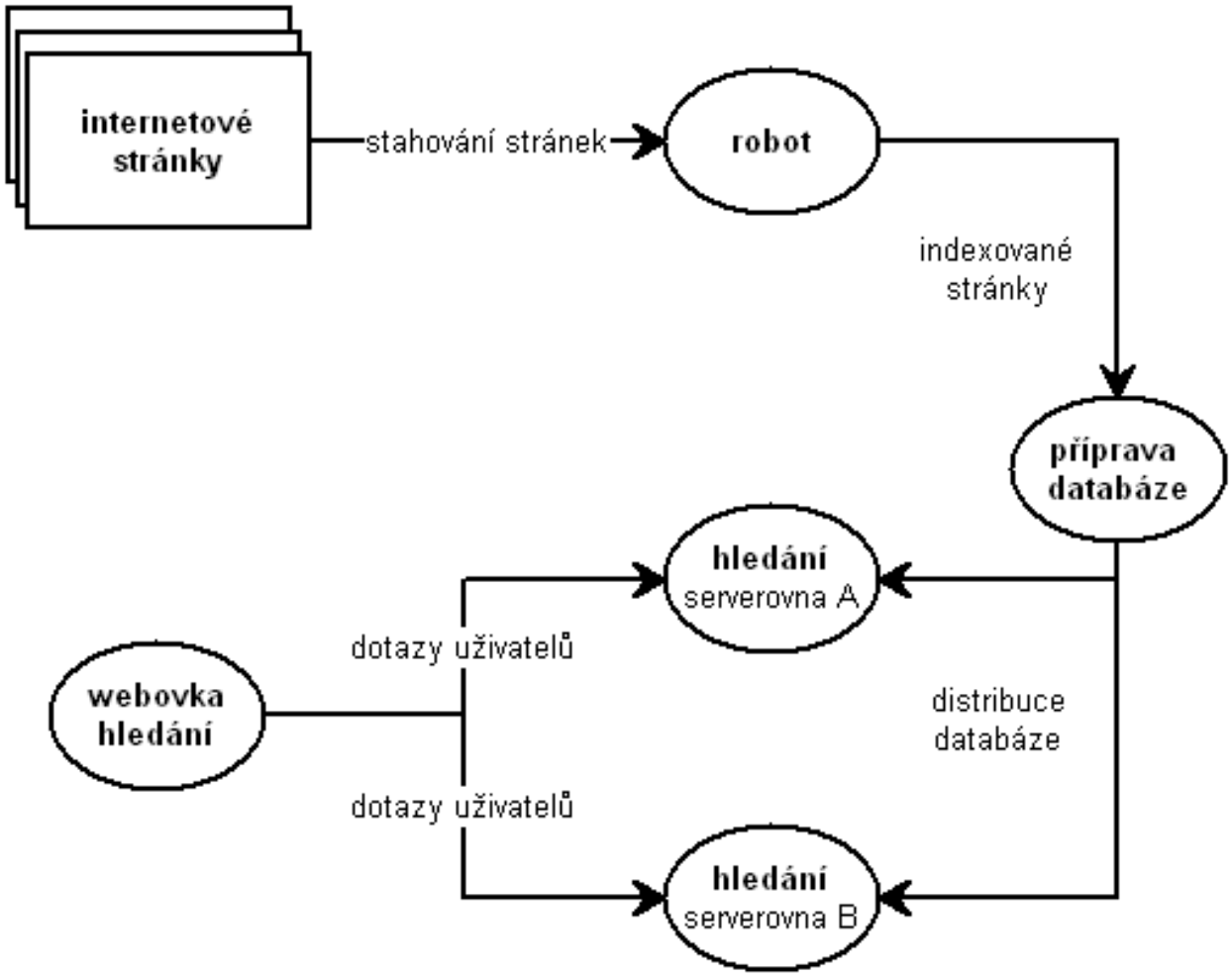
Část 1 – Architektura ve zkratce

1. Hlavní části
2. Redundance v provozu
3. Blokové schéma

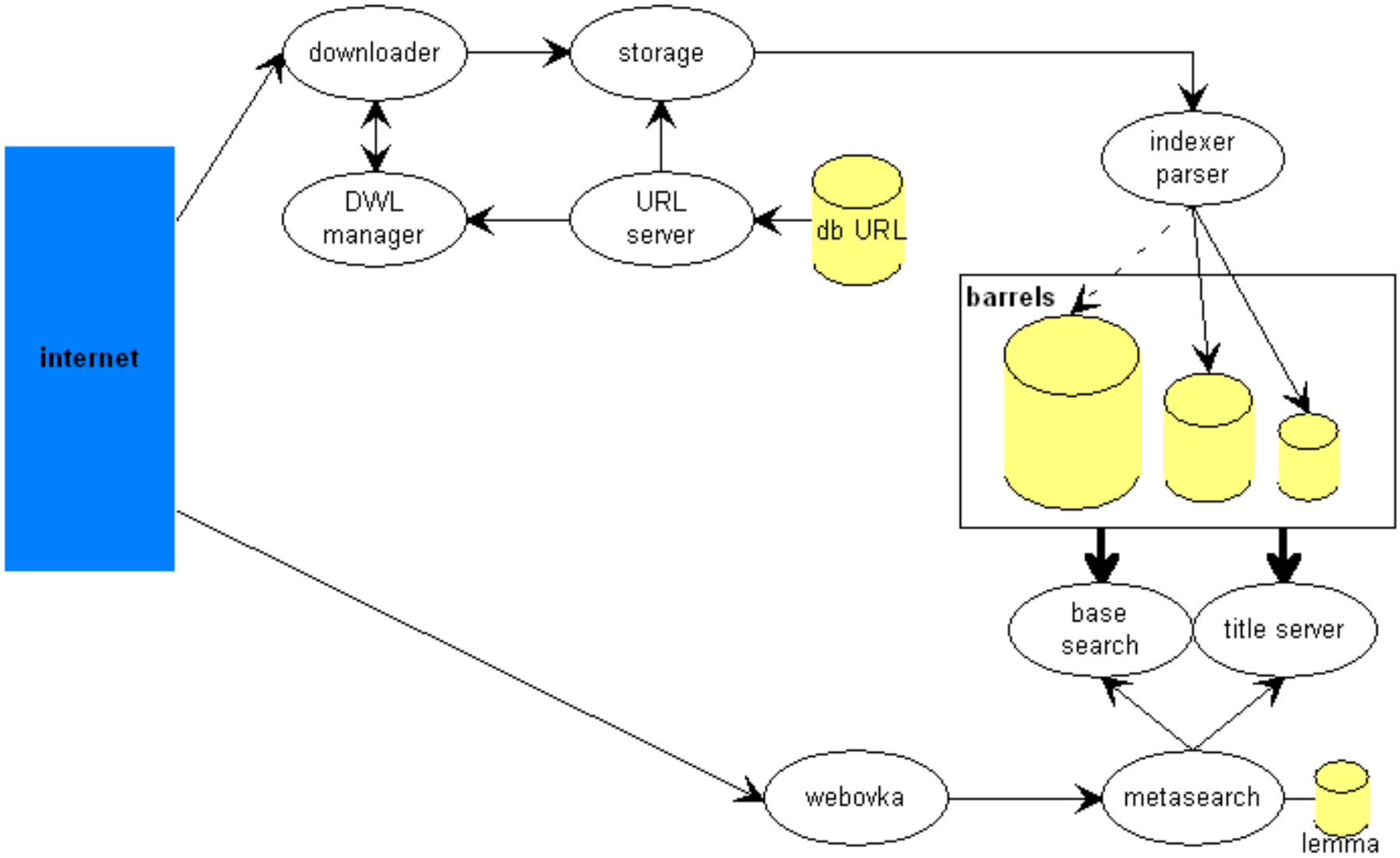
Hlavní části FTX



Hlavní části FTX – Redundance v provozu



Blokové schéma



Část 2 – Vyhledávání

1. Zadávané dotazy
2. Expanze dotazu
3. Hodnocení stránek

Zadávané dotazy (1)

- 10 náhodných dotazů
  - kurzy boxu
  - brzdový válec java 125
  - nokia c 5 recenze
  - blond sestrihy
  - seznam her na ps3
  - poslat sms
  - wikipeda
  - okresni soud chrudim
  - mafia 2 auta
  - linkedin.com

Zadávané dotazy (2)

- Forma dotazů:

  Nejedná se přímo o otázky

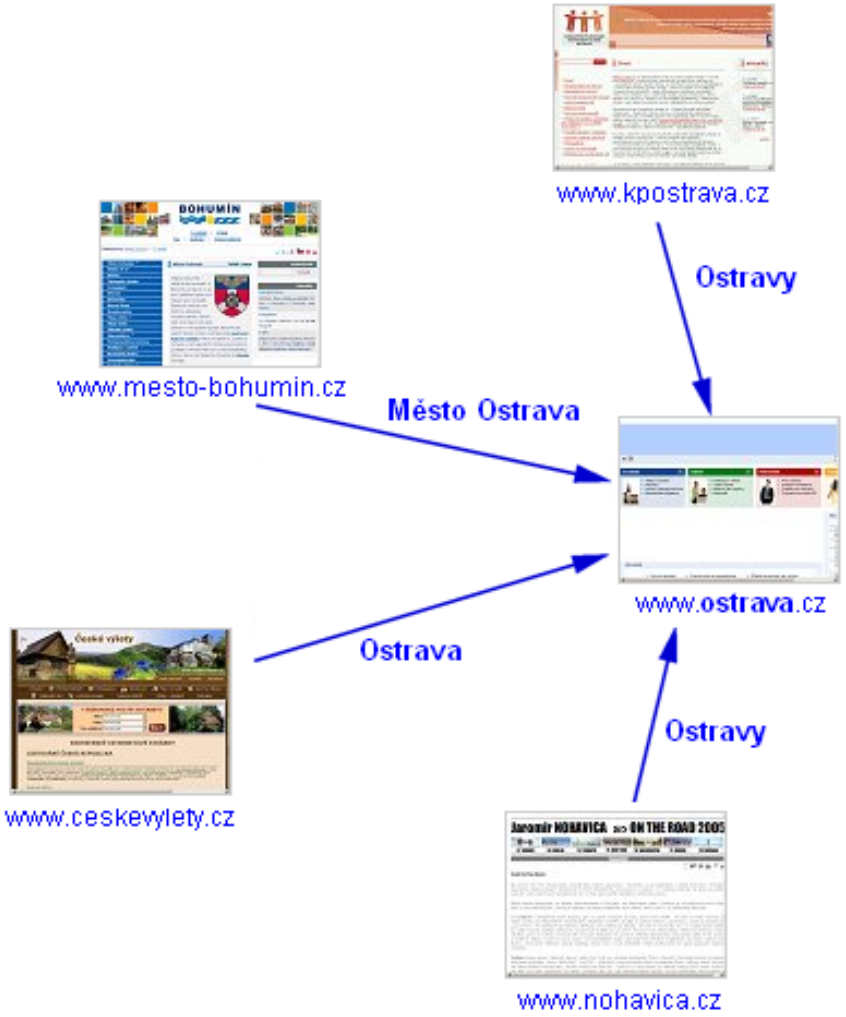
  – heslovité
  – překlepy
  – často bez diakritiky
  – 1. pád a podstatná jména

Hodnocení stránek (1) – onPage

Hlavní město **Praha** - Informační server pražské radnice
Praha udělá maximum, aby dopravní investice město bolely co nejméně. Informační server Hlavního Města **Prahy**
magistrat.**praha**-mesto.cz/ - Hlavní město Praha

- Titulek
- Obsah stránky
- URL
- Meta dascription

Hodnocení stránek (2) - offPage



Citační analýza pro dotaz „Ostrava“

Hodnocení stránek (3)

- Pagerank = statická „důležitost“ stránky založená na citační analýze
- Předpoklad: statisticky náhodné chování
- SPAM, Gray&Black Hat SEO
- celkově desítky signálů

Část 3 – Robot

1. Hledání nových stránek
2. Reindexace stránek
3. Ne-HTML formáty
4. offline výpočty

Hledání nových stránek (1)

- Před 6 lety start
- Procházení nalezených odkazů
  - Domény .cz, .sk, .com, .org, .net, .info, …
- Hledá stránky v českém jazyce
- Alternativní zdroje: RSS, sitemap, AddForm

Hledání nových stránek (2)

- Robots.txt – standardní protokol pro zakázání přístupu robotů (www.robotstxt.org)

- Textový soubor http://example.com/robots.txt

```
# comment
User-Agent: *
Disallow: /statistiky

User-Agent: Bot
Disallow: /
```

Reindexace stránek (1)

- Každý den se vybere množina stránek pro reindexaci

- Při výběru se hodnotí
  - Datum poslední návštěvy
  - Rank (Srank)
  - Frekvence změn

Reindexace stránek (2)

- Přetěžování webserverů
  - Shapování podle IP adresy
  - Omezení max počet URL / sec
  - Limity na straně serveru (v robots.txt)

Ne-HTML formáty

- konverze do html

- PDF
- DOC (MS Word)
- RTF
- PPT

- Operátor filetype:

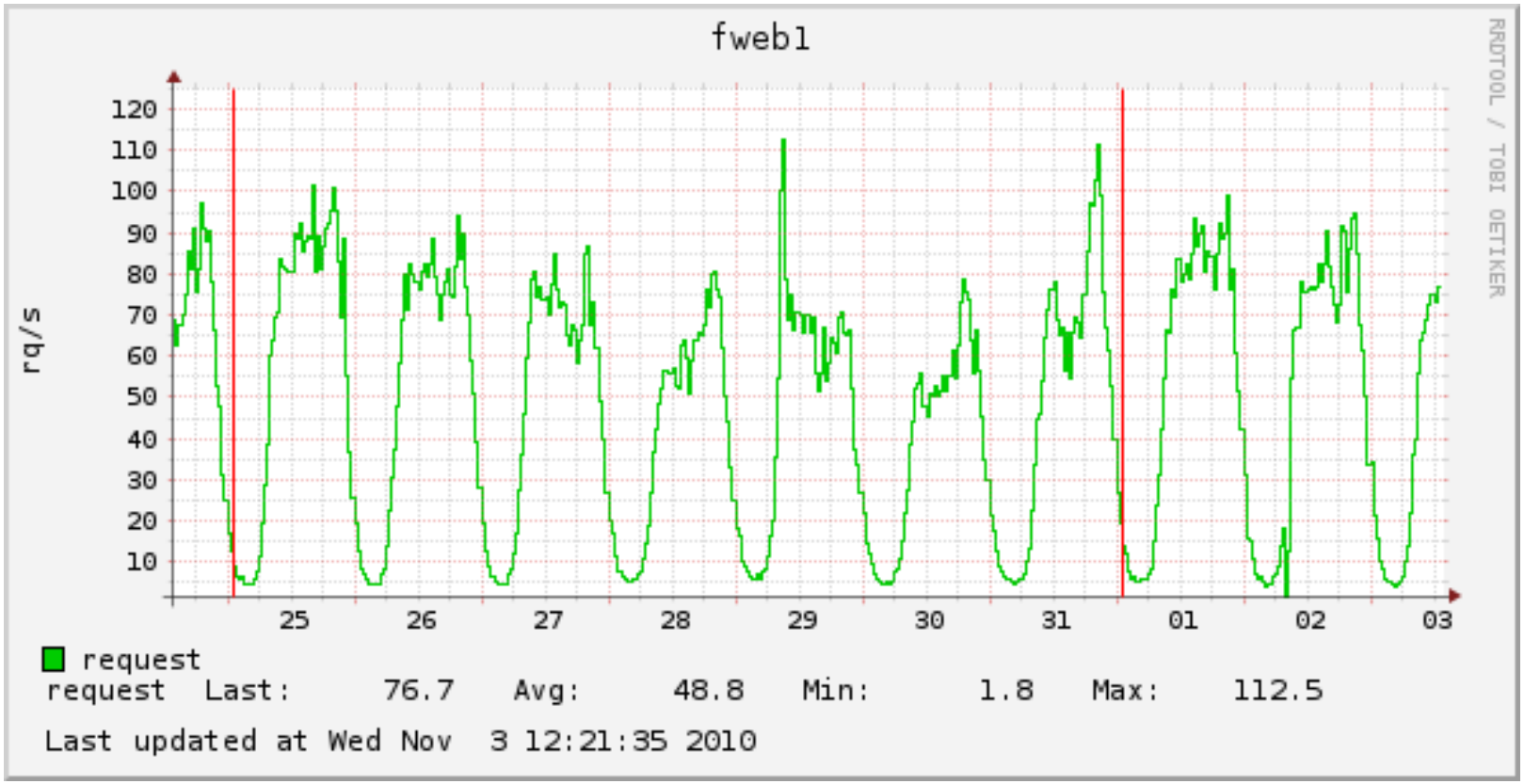
query filetype:pdf,html,ppt | Vyhledat Seznamem

Část 4 – Aktuální údaje z provozu

Velikost databáze (1)

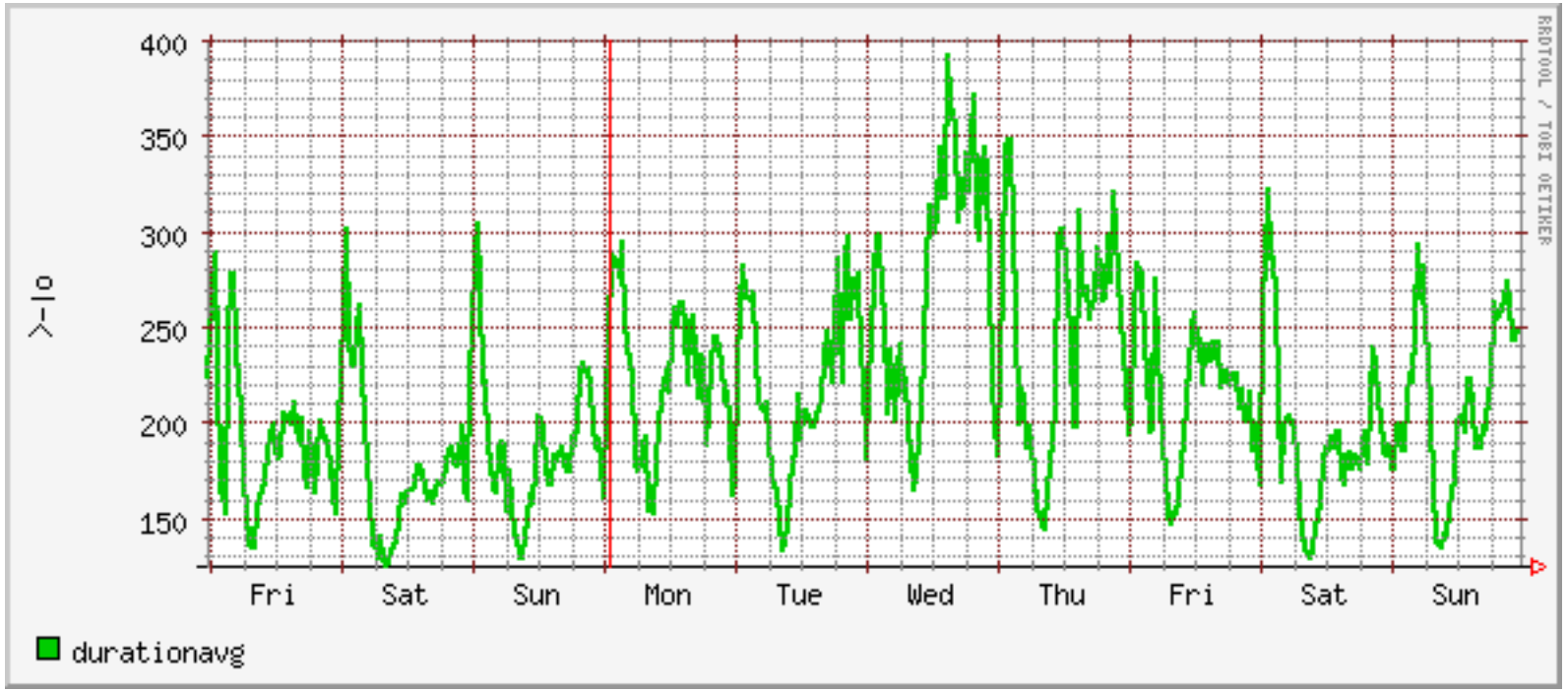
Počet dokumentů	340 miliónů
Indexy	1,7 TB
Obsah dokumentů (texty)	1,4 TB
Průměrný text	6 kB / dokument

Zátěž během týdne



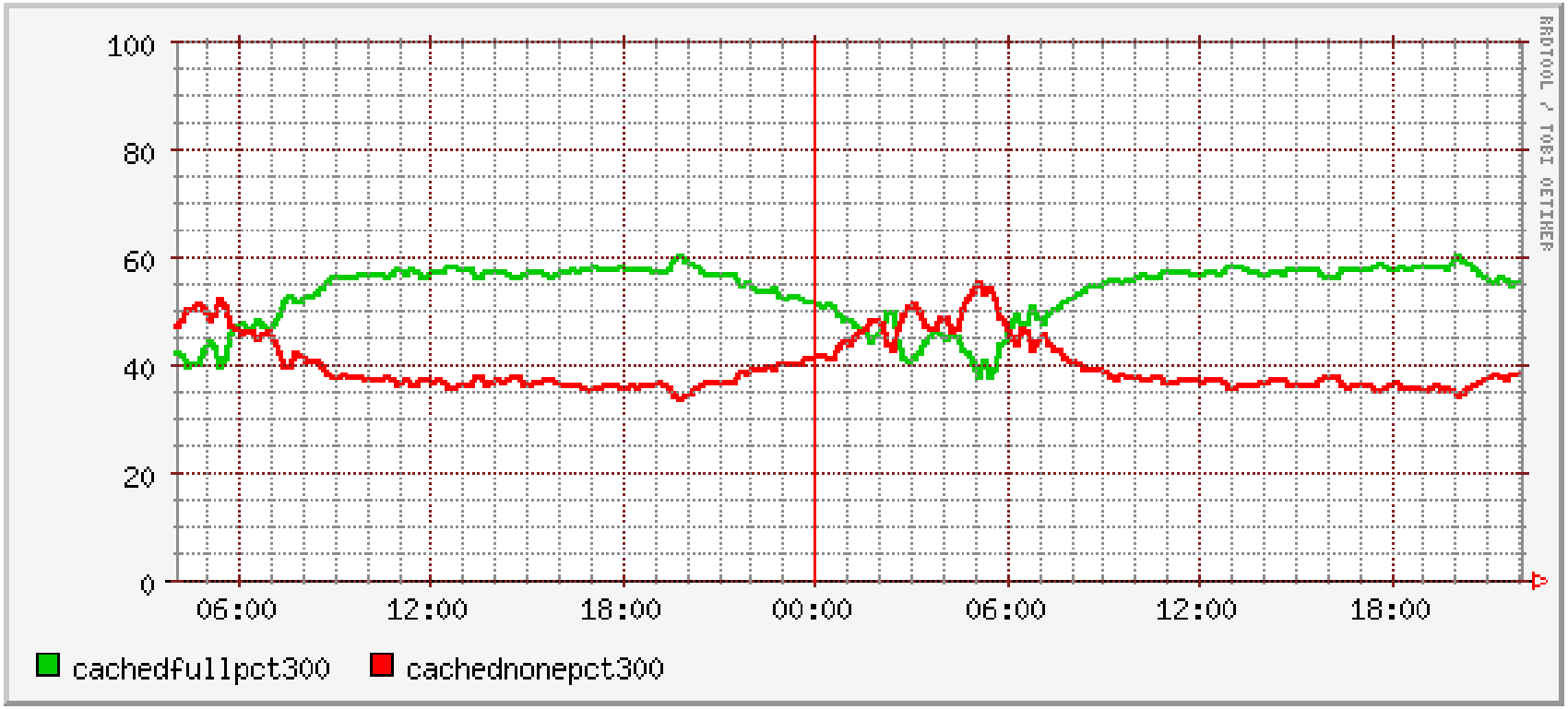
- 1/4 zátěže
- max ~480 dotazů/s

Doba odezvy během týdne



- Doba odezvy v msec

Úspěšnost query cache



- Úspěšnost cache v %

Výkon robota

Rychlost stahování	> 450 stránek / sec
Průměrná stránka	~11 kB (zdrojový kód)
Denní objem	~40 miliónů dokumentů cca 410 GB dat

Stáří dokumentů ve dnech

Minimální	<1
Maximální	135
Průměr	5,7
Nejčastěji	1,2 – 9,5

Novinky v roce 2010

- Termové hledání
- Rozšířené hledání
- Robot v3.0

Termová verze vyhledávání

- Návaznost na OR + expanze dotazu (2009)

- Hlavní změnou - *rušení* lemmatizace
  - indexujeme neupravená slova
  - rozhodnutí o slovu a dotazu až při hledání

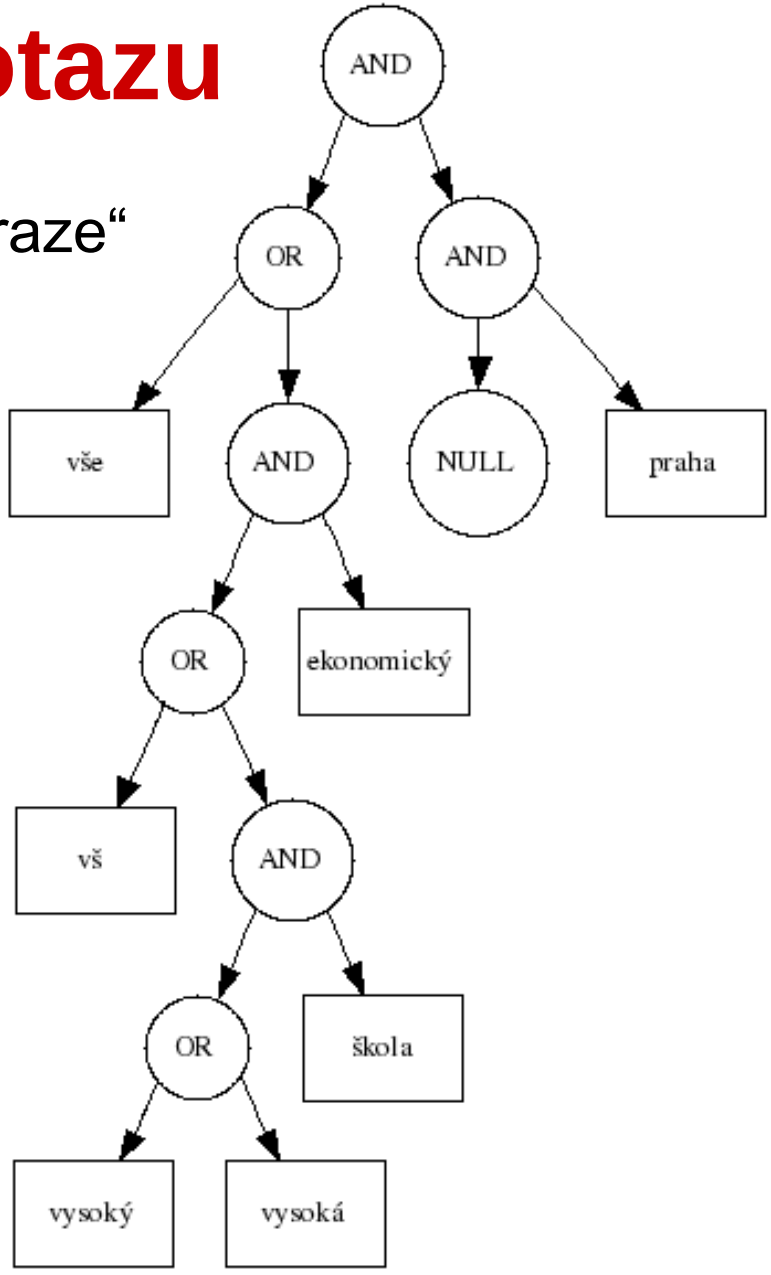
Expanze dotazu

- Lemma = základní tvar slova
- Věta:
  „Jeden z nejlepších zdrojů o německých tancích.“
- Lemmatizováno:
  Jedna/Jíst z dobrý zdroj o německý tank/tanec.
- Disambiguace = vyloučení nejednoznačnosti

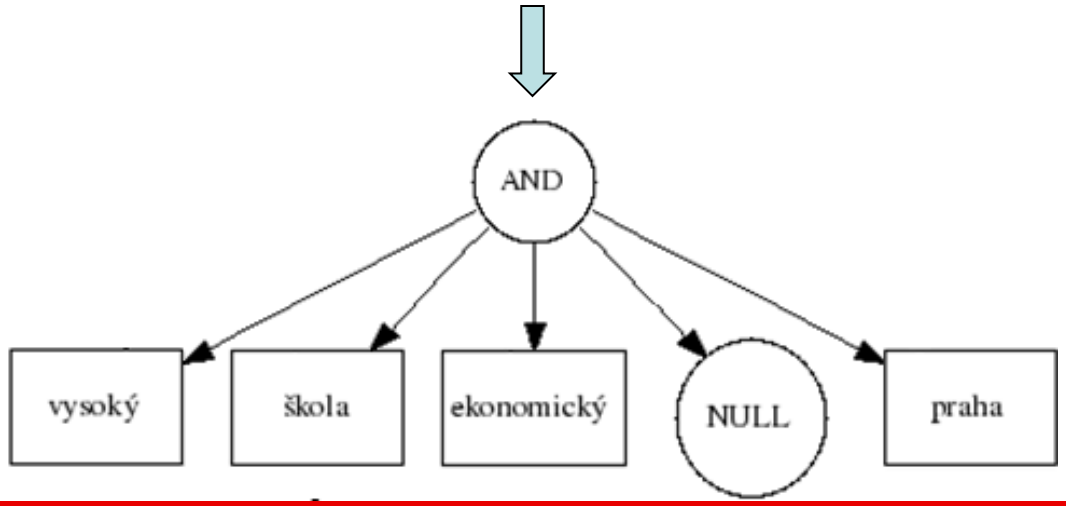
OR, expanze dotazu

Query: „Vysoká škola ekonomická v Praze“

Nové hledání



Staré hledání

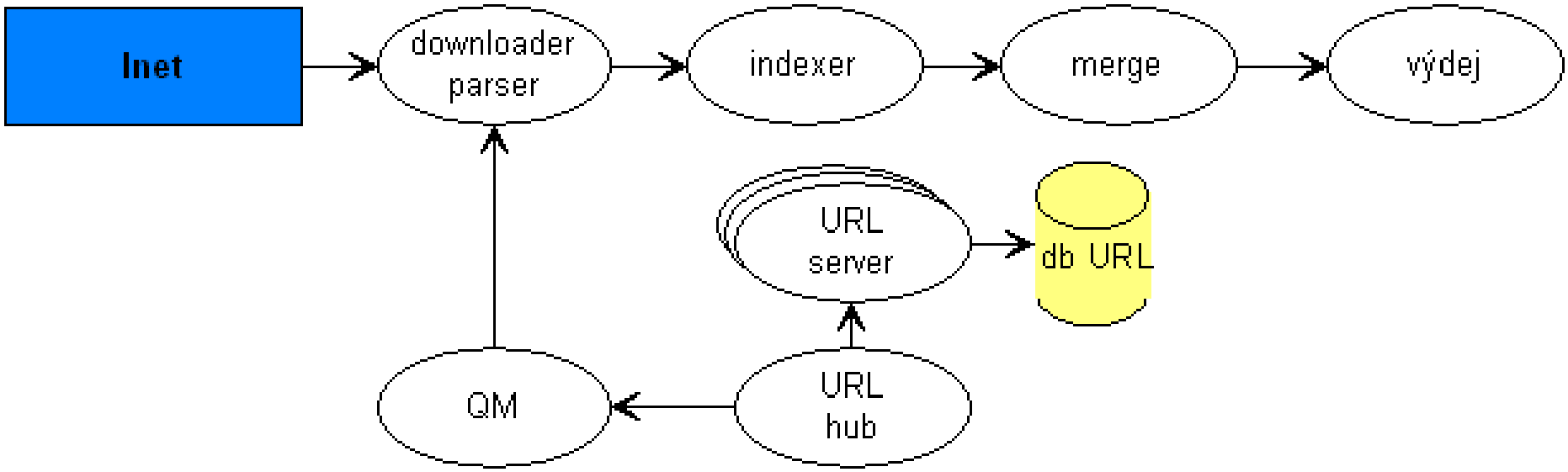


Rozšířené hledání

- zprostředkování interní funkčnosti

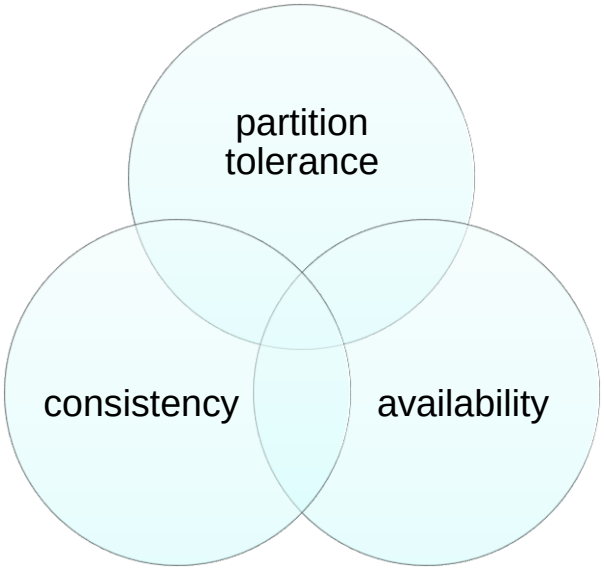
Robot v3.0 (2010)

Robot v2.x


Inet
downloader
parser
indexer
merge
výdej
URL
server
db URL
QM
URL
hub

Robot (2.x vs. 3.0)

- konec inkrementální indexace
- CAP theorem
  - r. 2000 (Brewer)



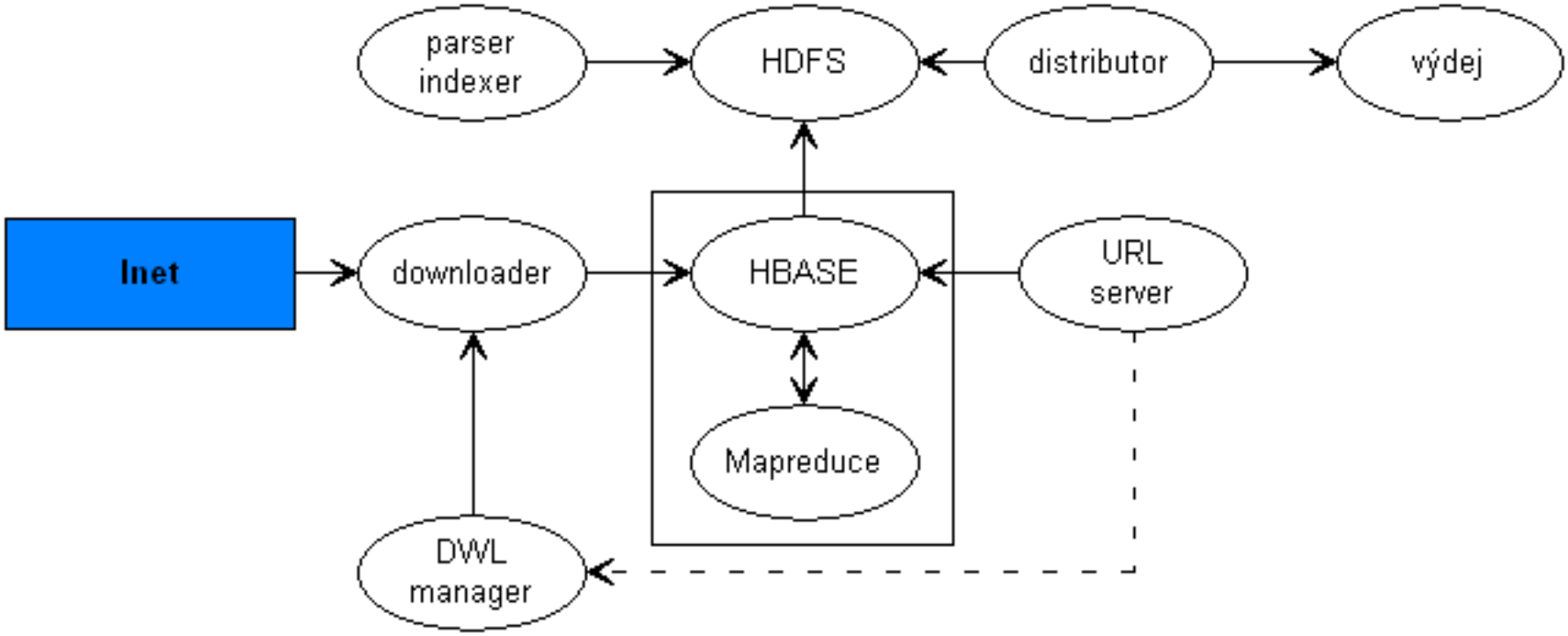
- storage 28 serverů
  - 3x replikace

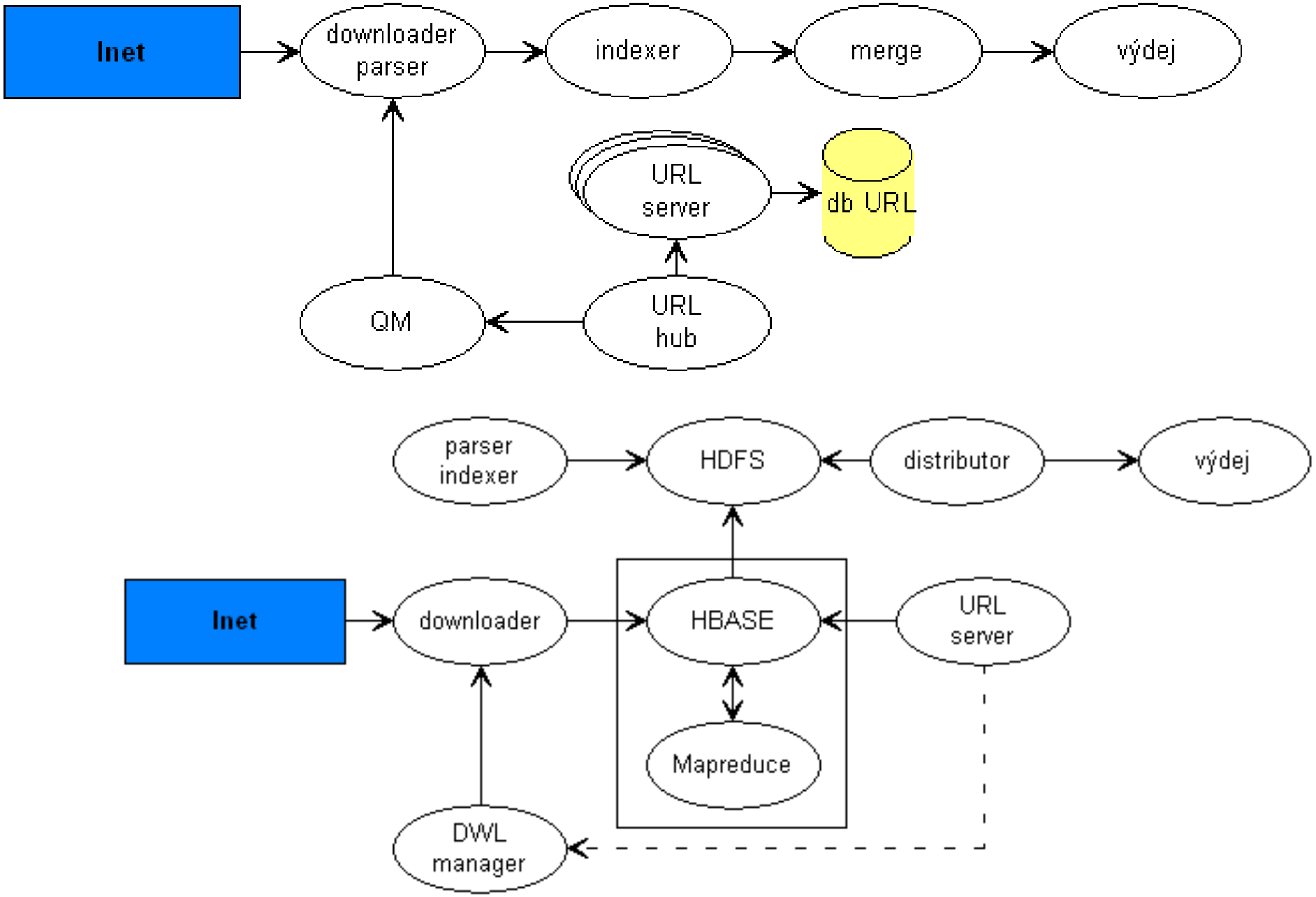
Robot v3.0

- v testu (od ledna 2011 venku)

- storage **Hbase**
  - Bigtable-like
  - MapReduce
- **Hadoop** (HDFS)
  - Open Souce
  - reimplementace Google FS
  - velké soubory (512MB)
  - místo RAID je replikace

Robot v3.0


parser
indexer
HDFS
distributor
výdej
Inet
downloader
HBASE
URL
server
Mapreduce
DWL
manager


Inet
downloader
parser
indexer
merge
výdej
URL
server
db URL
QM
URL
hub
parser
indexer
HDFS
distributor
výdej
Inet
downloader
HBASE
URL
server
Mapreduce
DWL
manager

Konec

Děkuji za pozornost

http://fulltext.sblog.cz

„Bonusy“

1. TOP 10 dotazů
2. o optimalizaci

Top 10 dotazů

r. 2008	r. 2009	r. 2010
1. ""	1. ""	1. ""
2. youtube	2. youtube.com	2. facebook.com
3. libimseti.cz	3. libimseti.cz	3. youtube
4. superhry	4. superhry	4. super hry
5. freefoto	5. o2	5. freevideo
6. freevideo	6. freevideo	6. idos
7. redtube.com	7. facebook	7. aukro
8. sms zdarma	8. aukro.cz	8. google
9. google	9. google	9. super hry
10. porno	10. porno	10. sms zdarma

SEO

(search engine optimalization)

1. URL
2. Obsah stránky
3. JavaScript a Flash

URL

- Vhodně zvolená doména
  - www.csas.cz
  - www.ceskasporitelna.cz
- Optimalizované URL a rewrite
  - super.cz/index.php?clid=18656
  - novinky.cz/vladni-spis-jak-zabranit-uniku-informaci-na-internet-unikl-na-internet
- Minimalizovat duplicity!!

Obsah stránky

- Titulek
  - Důležitá součást stránky
  - Unikátní na každé stránce

- Text
  - Správně používat sémantické značky
  - Nepoužívat text jen na obrázku

JavaScript a Flash

- Robot neumí procházet přes:
  - formuláře
  - JavaScript navigaci
  - Flash presentace
  - JavaScript přesměrování

- Textová alternativa k dynamické navigaci

Konec (2)