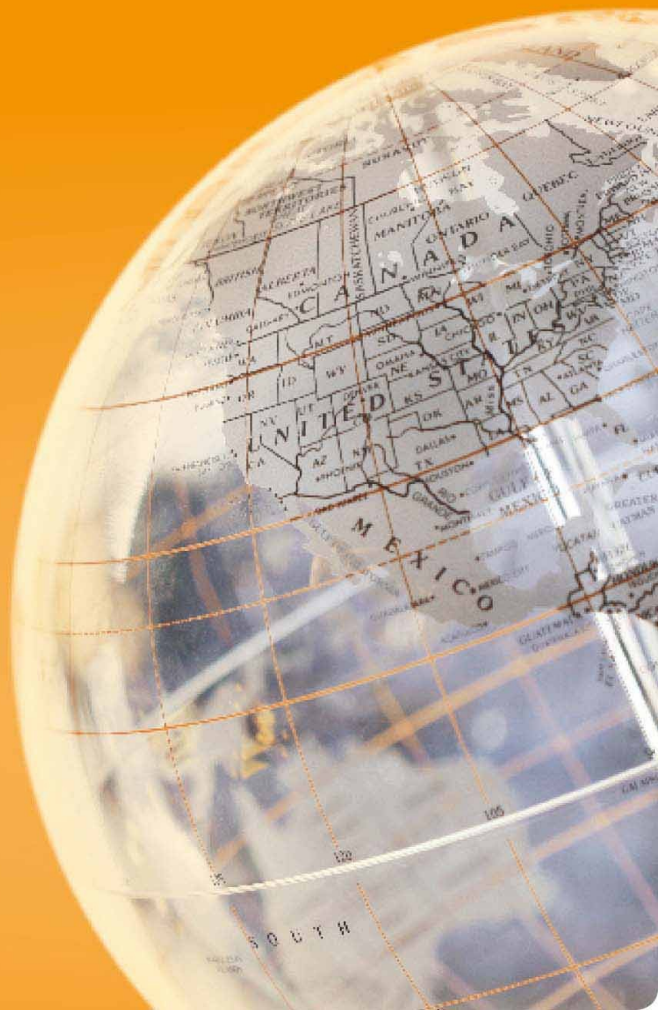
zanox
we create partners

Product Search and Reporting powered by Hadoop

10. March 2010 | Dr. Dragan Milosevic



Who am I?

- Senior Architect at zanox AG
  - Over the last two years I have been writing map-reduce jobs which help applications cope with millions of products and billions of clicks

- I have applied different machine-learning techniques mainly to optimise resource usage while performing distributed search during my PhD
  - See my book: “Beyond Centralised Search Engines An Agent-Based Filtering Framework”

What is it about?

zanox
we create partners

- Part I: Processing product and tracking data by Map-Reduce
  - Normalising and categorising product data
  - Joining and aggregating tracking data
- Part II: Lucene-powered distributed search and aggregation
  - Merger-based coordination of multiple searchers
  - Observer infrastructure to ensure robust and reliable services
- Part III: Technical details
  - Hardware, how much data, number of jobs, how many requests

Part I: Product Normalisation

- Problem: Manufacturer names are not normalised in imported data
  - Single manufacturer has sometimes more than 50 different names
  - There are more than 1 million different names, which are too much for exhaustive comparison

- Solution: Divide-and-Conquer to make it suitable for Map-Reduce
  - Use fast clustering that puts together potentially identical names
  - Each Map task applies on cluster-level several distance computation algorithms:
    - Coding-based (Soundex) – *code(“samsung”) = s525*
    - Edit-distance (Levenshtein) – *d(“gumbo”,”gambol”) = 2*
    - N-gram-based – *code(“samsung”)={‘sa’,’am’,’ms’,’su’,’un’,’ng’}*
    - Suffix-Tree-based (Longest-Common-Substring)
      d(“megaphon importservice”, “import megaphon”) = 8 + 6 = 14

Part I: Product Categorisation

- Every category (out of 600) has been assigned language specific-model to be used in categorisation process
  - Models are compact and suitable to be loaded in memory
  - They can be seen as collection of words and phrases together with heuristic-rules helping to correctly categorise
  - Models are semi-automatically updated to improve categorisation

- Compact models are loaded by Map tasks
  - Markov-Chain-based language detection of a product to select model
  - Appling rules to reduce the set of possible categories
  - Computing scores based on word and phrase belongingness

Part I: Joining and Aggregating Tracking Data

zanox we create partners

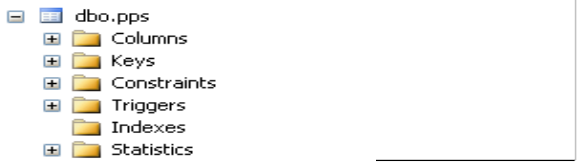
Map-Reduce Inputs

Custom Report Definition

```
<report-configuration xmlns="http://www.zanox.com/xml
  ...
  <report>
    <name>Conversion per Campagne and Chanel</name>
    <filter-key>PARTNER_VALUE_CLIENT_ID</filter-key>
    <filter-key>PARTNER_VALUE_UNIT_ID</filter-key>
    <group-key>PARTNER_VALUE_CAMPAGNE</group-key>
    <group-key>PARTNER_VALUE_CHANEL</group-key>
    <value-key>PPC_VISITS</value-key>
    <value-key>IMMEDIATE_SALES</value-key>
    <value-key>DELAYED_SALES</value-key>
    <value-key>IMMEDIATE_SALE_AMOUNT</value-key>
    <value-key>DELAYED_SALE_AMOUNT</value-key>
    <value-key>GOOGLE_COSTS</value-key>
    <value-key>GOOGLE_AVG_POSSITION</value-key>
  </report>
  ...
</report-configuration>
```

+

zanox Tracking Data



+

Search Engine Data

```
"impressions";"clicks";"ctr";"cpc";"cost";"pos"
27;1;0.0370370370370372;0.05;0.05;9.29629629631
```

+

Custom Tracking Data

partner_value
146;HUK24+DE+Kfz;78;Kfz-Tarifrechner;;;929653201

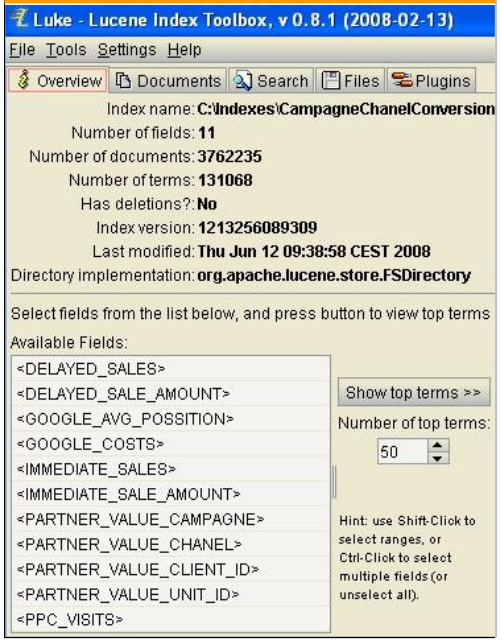
Custom Tracking Data Definition

```
<reader-configuration xmlns="http://www.zanox.com/
  ...
  <key>
    <name>PARTNER_VALUE_CLIENT_ID</name>
    <possition>10</possition>
    <sub-possition>0</sub-possition>
    <operation>EMPTY_TO_ZERO</operation>
    <type>SHORT</type>
    <rs-column>partner_value</rs-column>
  </key>
  <key>
    <name>PARTNER_VALUE_UNIT_ID</name>
    <possition>10</possition>
    <sub-possition>0</sub-possition>
    <operation>EMPTY_TO_ZERO</operation>
    <operation>CLIENT_TO_UNIT</operation>
    <type>SHORT</type>
    <rs-column>partner_value</rs-column>
  </key>
  <key>
    <name>PARTNER_VALUE_CAMPAGNE</name>
    <possition>10</possition>
    <sub-possition>1</sub-possition>
    <operation>ISO_URL_DECODE</operation>
    <operation>TO_LOWER_CASE</operation>
    <rs-column>partner_value</rs-column>
  </key>
  <key>
    <name>PARTNER_VALUE_CHANEL</name>
    <possition>10</possition>
    <sub-possition>2</sub-possition>
    <operation>EMPTY_TO_ZERO</operation>
    <type>SHORT</type>
    <rs-column>partner_value</rs-column>
  </key>
  ...
</reader-configuration>
```

=

Map-Reduce Outputs

Lucene Indexes

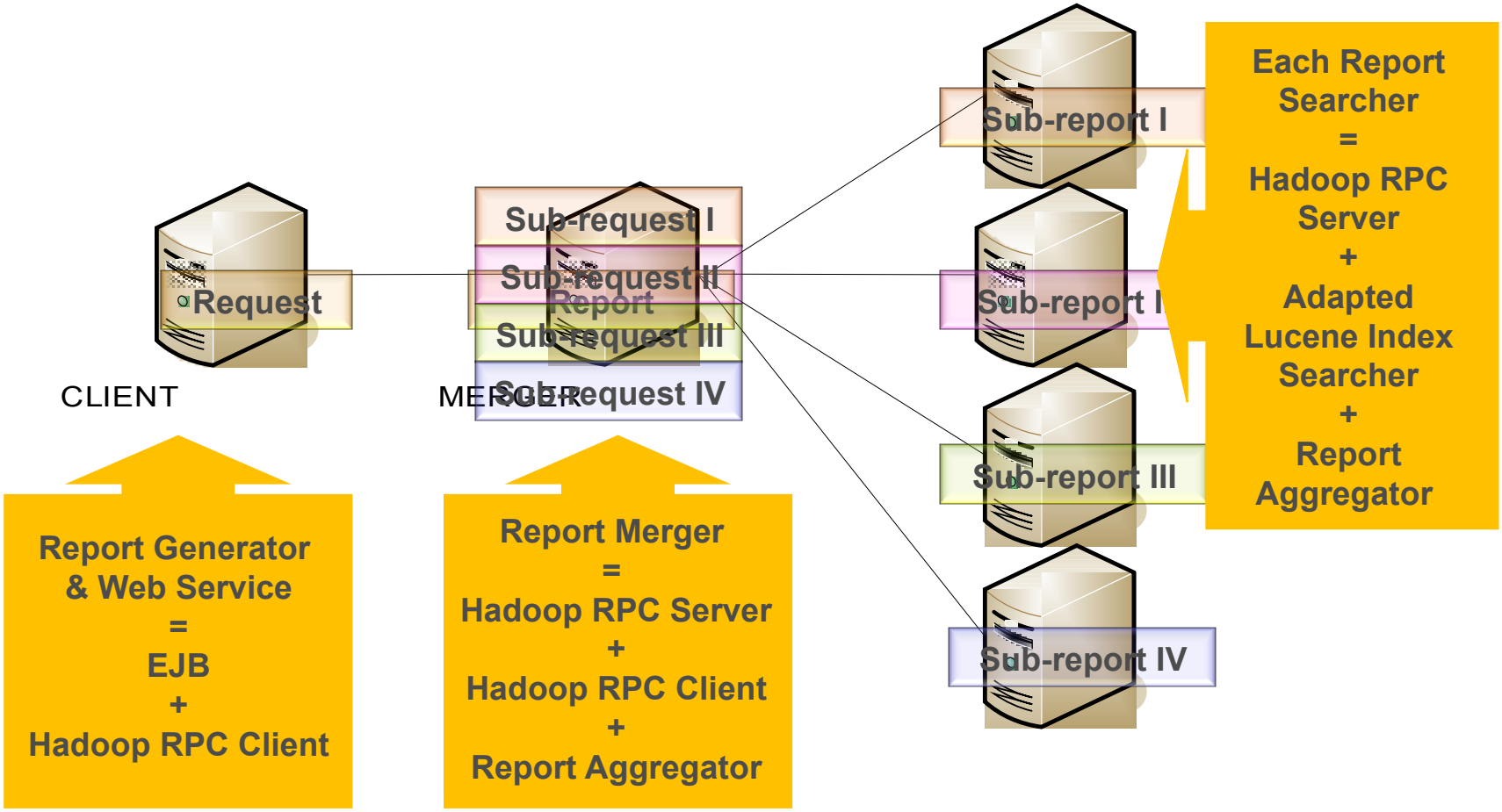


Part II: Distributed Search and Aggregation

- Problem: Indexes are so large that they cannot be handled by a single machine
  - Combined size of daily produced indexes is over 500 GB
  - Neither searching nor aggregation can be done by one machine

- Solution: Distributed search
  - Indexes are loaded by several Lucene searchers
  - Searchers are capable of finding matching documents, building facets, aggregating (reducing) selected data
  - Mergers select searchers to be used, adapt query to be sent to every searcher and aggregate results received from searchers
  - Observers control how searchers and mergers are performing

Part II: Merger-Based Coordination of Searchers



Part III: Technical Details - Hardware

10.12.2007

3 machines
=
Single Core
+
1GB RAM
+
40GB HD

10.03.2010

42 machines
(18 + 24)
=
8 Core
+
16GB RAM
+
2 x 1TB HD

01.01.2013

?

Part III: Technical Details – Data, Jobs, Queries

- Data in HDFS
  - Data volume growing by 50 GB/day
    (30 million clicks, 500 million views and 2 million product updates)
  - 500 GB Lucene indexes built on daily basis
  - Total data volume of 14 TB for 11 billion clicks, 90 billion views and 85 million products
- Jobs
  - More than 800 scheduled jobs per day
- Queries
  - 5 queries per second and more than 20 million queries in the last 2 months

Thank you for your attention

21 “Different” Samsung Manufacturers

“Samsung", "SAMSUNG - MONITORS", "SAMSUNG - PLASMA",
"SAMSUNG - PRESENTATION", "Samsung (Electronics)", "SAMSUNG (SA)",
"Samsung Books", "Samsung BW", "SAMSUNG by NORTEK",
"SAMSUNG Compatible", "SAMSUNG COMPUTER", "SAMSUNG DEUTSCHLAND",

"Samsung Music", "Samsung Notebook", "SAMSUNG, TELECOM“
"Samsung Opto-Electronics UK Ltd.", "SAMSUNG ORIGINAL",
"SAMSUNG PLEOMAX", "SAMSUNG SEMICONDUCTOR",
"SAMSUNG SGH-E390", "Samsung UK Ltd“ and "Samsung WW“.