Use case: Merging heterogeneous network measurement data

Jorge E. López de Vergara and Javier Aracil

Jorge.lopez_vergara@uam.es

Credits to Rubén García-Valcárcel, Iván González, Rafael Leira, Víctor Moreno, David Muelas, Javier Ramos, Paula Roquero, Carlos Vega and the rest of the HPCN-UAM team

SMART Internet Monitoring Study 3rd Workshop, Barcelona, Spain, 22nd April 2015

Contents

- Introduction
- Technologies
  - High-speed traffic measurements
  - Data integration alternatives
  - Hadoop for network measurements
  - Log processing
- Conclusions

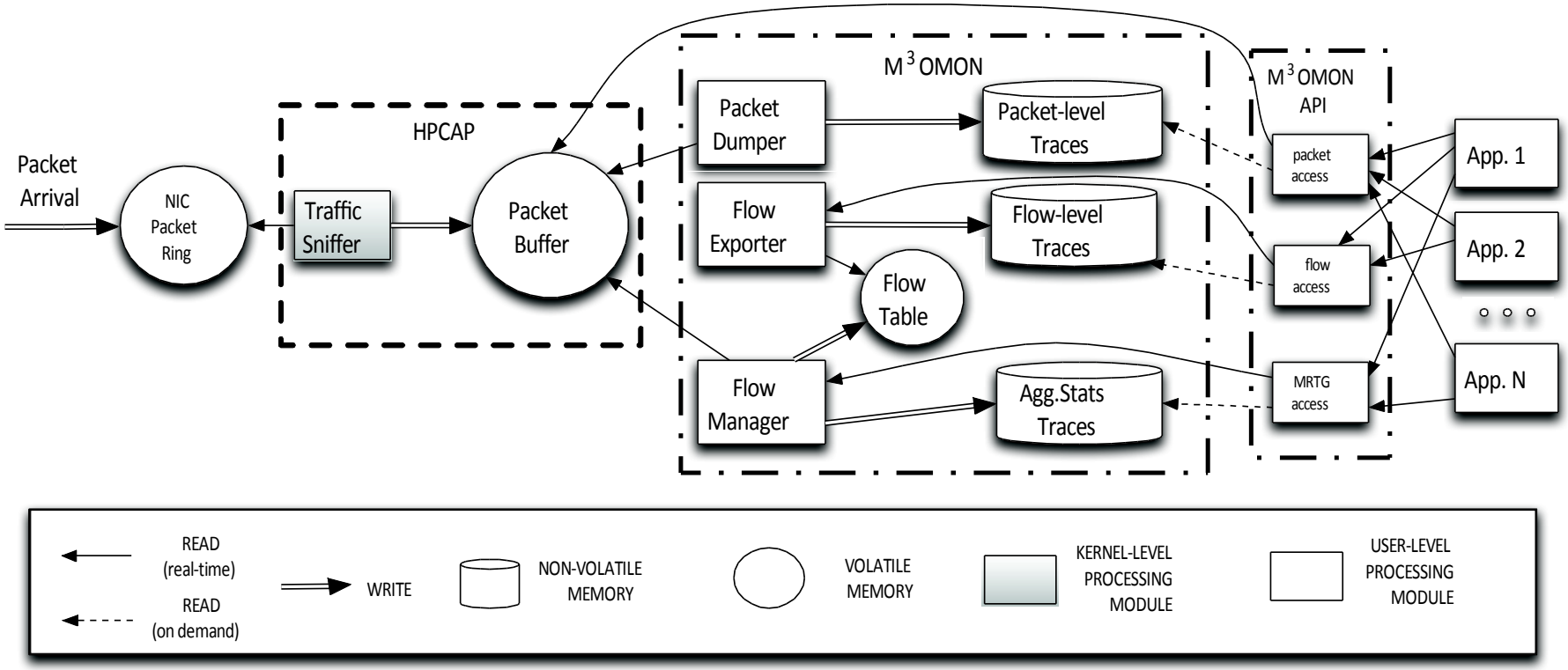
Introduction

- Network measurement data can come from different sources
  - Network-oriented sources:
    - SNMP MIB instances
    - Netflow records
    - Pcap files
    - …
  - Application-oriented sources:
    - Logs
      - Some standard (Apache web log)
      - Some proprietary (application specific)
    - Important to deal with encrypted traffic
  - It is necessary to provide ways to merge them

High-speed traffic measurements

- Requirements
  - Capture at core networks
  - +10 Gbps links, sometimes virtualized
  - No packet drops
- Available off-the-shelf resources that help on this tough task
  - Intel +10G network cards
    - Intel DPDK
    - Other developments: HPCAP at UAM
  - Mellanox +10G network cards
    - Mellanox Messaging Accelerator (VMA)
  - Multicore processors
    - CPU affinity and isolation for key tasks
  - Lots of RAM memory
    - Use of hugepages and mmap

High-speed traffic measurements



Data integration alternatives

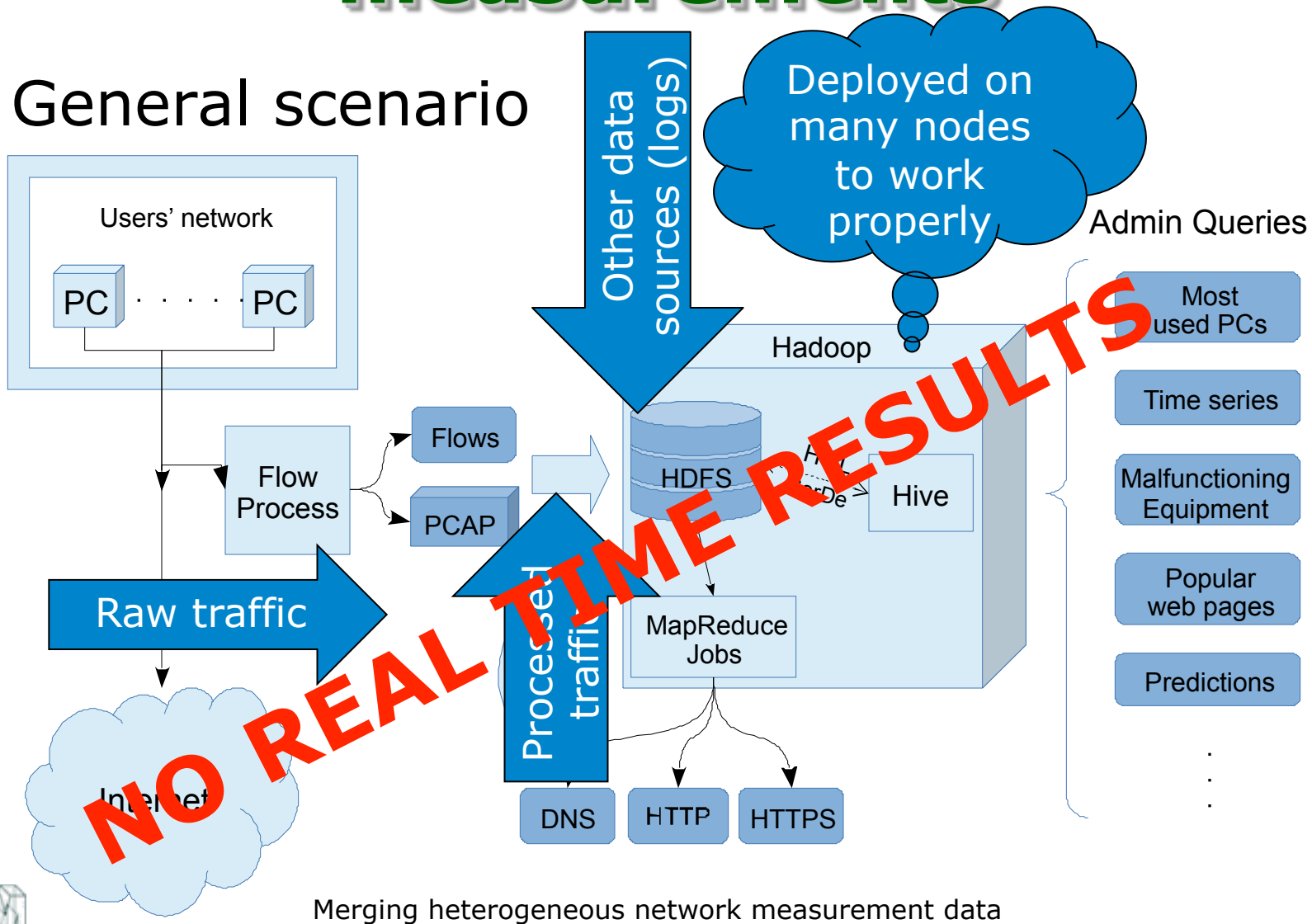
- SQL databases
  - Pros: reliable, normalized schemas, consistent with defined constraints
  - Cons: slow, need to use materialized views to go faster, creating materialized views is costly and sometimes it can't be done concurrently

    → Not valid to deal with high-speed network measurements
- Plain files, no SQL
  - Pros: much faster
  - Cons: inconsistencies, lack of normalization

    → Necessary to deal with high-speed network measurements, but keeping in mind its limitations

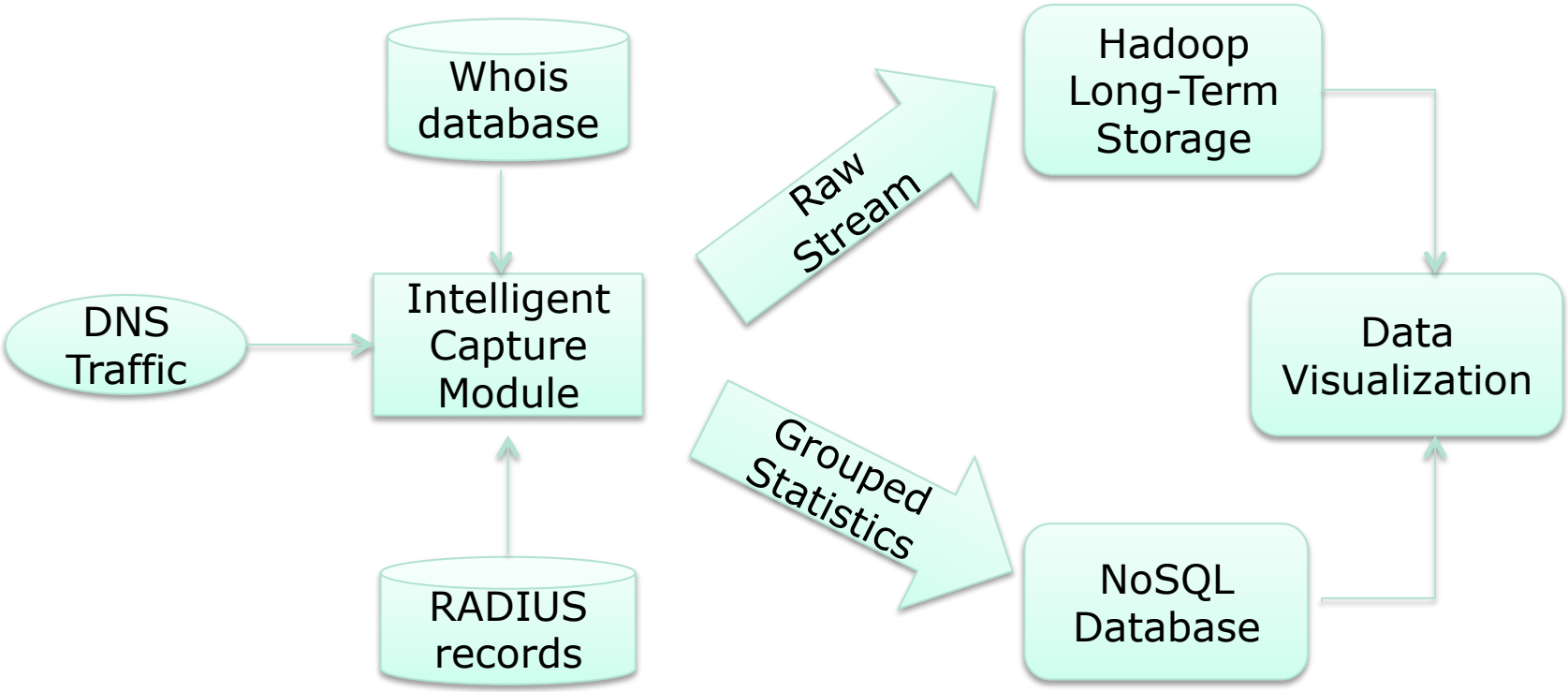
Hadoop for network measurements

- General scenario



Hadoop for network measurements

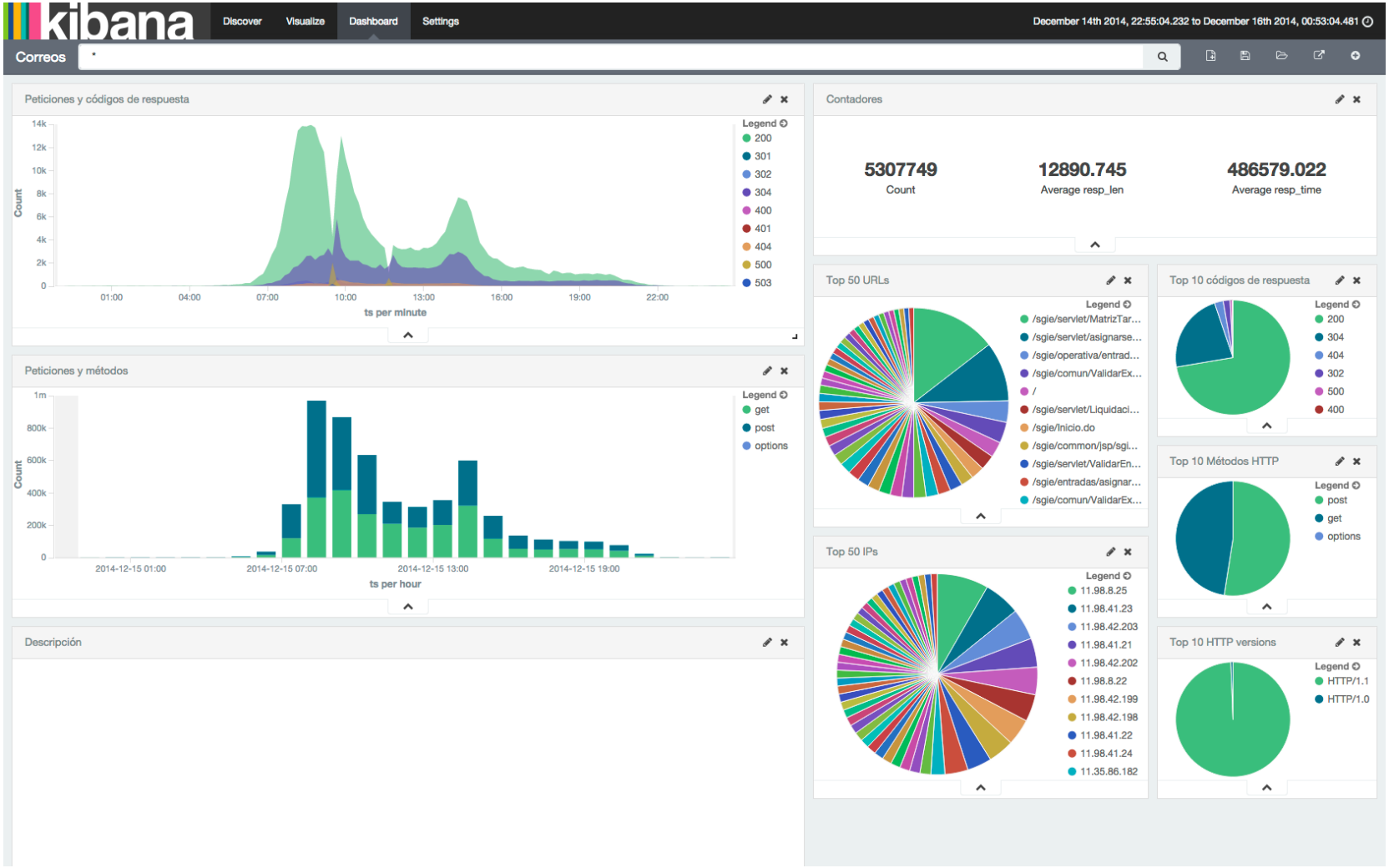
- DNS analysis



Log processing

- Requirements (real scenario):
  - Process 3M events per second (about 5 Gbps)
    - Put together application logs and network flows
  - Several disks in parallel are necessary to store the events at appropriate rate
  - Fast access to time series and aggregated statistics
    → What operators demand
  - Slower access to raw data
    → What IT analysts demand
- Elasticsearch and Kibana tools provide some support, but it is necessary to tune them

Kibana UI


kibana
Discover
Visualize
Dashboard
Settings
December 14th 2014, 22:55:04.232 to December 16th 2014, 00:53:04.481
Correos
Peticiones y códigos de respuesta
Count
ts per minute
Legend
200
301
302
304
400
401
404
500
503
Contadores
5307749
Count
12890.745
Average resp_len
486579.022
Average resp_time
Top 50 URLs
Top 10 códigos de respuesta
Peticiones y métodos
ts per hour
get
post
options
Top 10 Métodos HTTP
Top 50 IPs
11.98.8.25
11.98.41.23
11.98.42.203
11.98.41.21
11.98.42.202
11.98.8.22
11.98.42.199
11.98.42.198
11.98.41.22
11.98.41.24
11.35.86.182
Top 10 HTTP versions
HTTP/1.1
HTTP/1.0
Descripción

Conclusions

- Need for different network measurement data sources
  - Combine network and application data
  - Necessary to find sources of problems
    - Is the slowdown caused by the network, the server or the application?
    - This question can only be answered if all information from different layers is provided and analyzed
- Huge amount of data → it is necessary to work with fast processing systems
- Availability of processing tools from the Cloud Computing community
  - It is necessary to adapt them to the network measurment processes