A Distributed Framework for NLP-Based Keyword and Keyphrase Extraction From Web Pages and Documents

Paolo Nesi, Gianni Pantaleo and Gianmarco Sanesi

Department of Information Engineering, DINFO
University of Florence
Via S. Marta 3, 50139, Firenze, Florence, Italy
Tel: +39-055-2758511, fax: +39-055-2758516

DISIT Lab

http://www.disit.dinfo.unifi.it *alias* http://www.disit.org
paolo.nesi@unifi.it , gianni.pantaleo@unifi.it

21st International Conference on Distributed Multimedia Systems – DMS2015

Problem Focus & Main Issues

- The WWW continuously growing represents a massive source of knowledge, which is embedded for the most part in the textual content of web pages, documents, social media etc...

Google

100 Petabytes per day processed on 3 million servers in 2014

300 Petabytes stored in the first half of 2014

- Problem: Online web resources, pages and documents are mainly represented as unstructured natural language text data. Instead, structured data are required for the extraction and management of higher level knowledge.
- For automatic retrieval and production of structured information there is the need to ingest, process and analyze huge amounts (**Big Data** problem) of natural language data (**NLP**, **Statistical, Machine Learning & AI** problems).

Application Areas

- Many application fields and areas:
  - Keywords & keyphrase extraction for Content/Topic extraction and summarization.
  - Comprehension of natural language text documents.
  - Production of machine-readable corpora (integration with semantic resources and ontologies…).
  - Indexing for search engine functionalities: Content-based, multi-faceted search queries…
  - Design of expert systems (e.g.: Recommendation Tools, DSS, Question-Answer systems, personal assistants, etc.)
  - Social media mining and user behavior analysis for target-marketing and customized services.
- Applications and Interests in: **e-commerce, target marketing and advertising, business web services, Smart City platforms, e-Healthcare, scientific research, ICT technologies etc…**

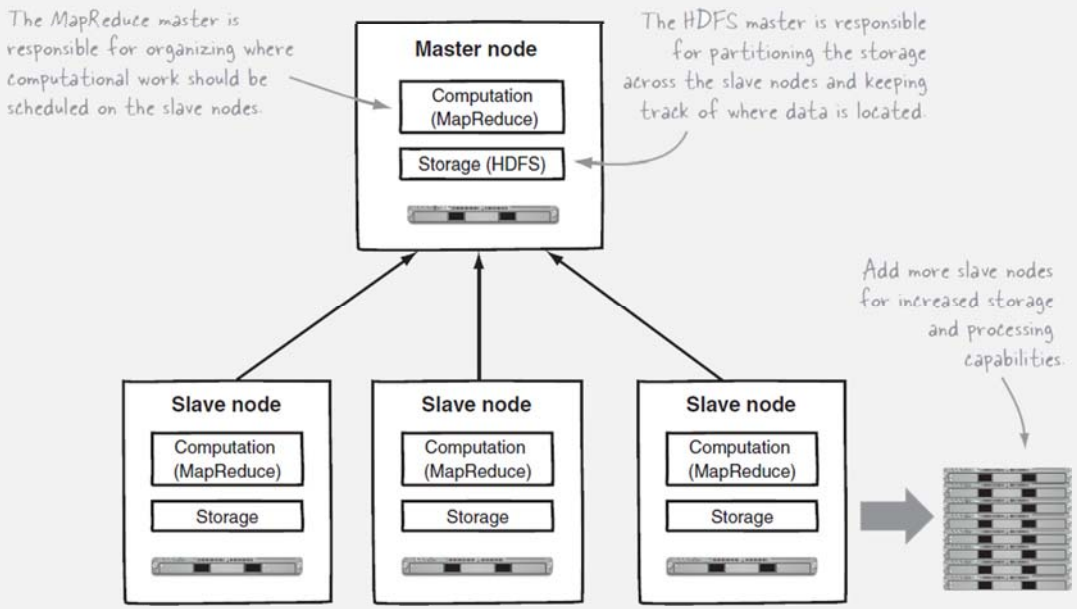
Distributed Architecture & Parallel Computing Frameworks

- Currently, single-machine, non-distributed architectures are proving to be inefficient for tasks like Big Data mining and intensive text processing and NLP analysis.
- Distributed Architectures and Parallel Computing techniques have been increasingly adopted in literature for automatic keyword extraction on Big Data sets.
  - First attempts of employing parallel computing frameworks for NLP tasks in the middle 90s: Chung and Moldovan [Chung and Moldovan, 1995] proposed a parallel memory-based parser called ***PARALLEL*** implemented on a parallel computer, the Semantic Network Array Processor (SNAP).
  - **Ogmios** [Hamon et al., 2007] is a platform for annotation enrichment and NLP analysis of specialized domain documents within a *distributed corpus*.
  - Exner and Hugues recently presented **Koshik** [Exner and Hugues, 2014], a multi-language NLP processing framework for large scale-processing and querying of unstructured natural language documents distributed upon a Hadoop-based cluster.

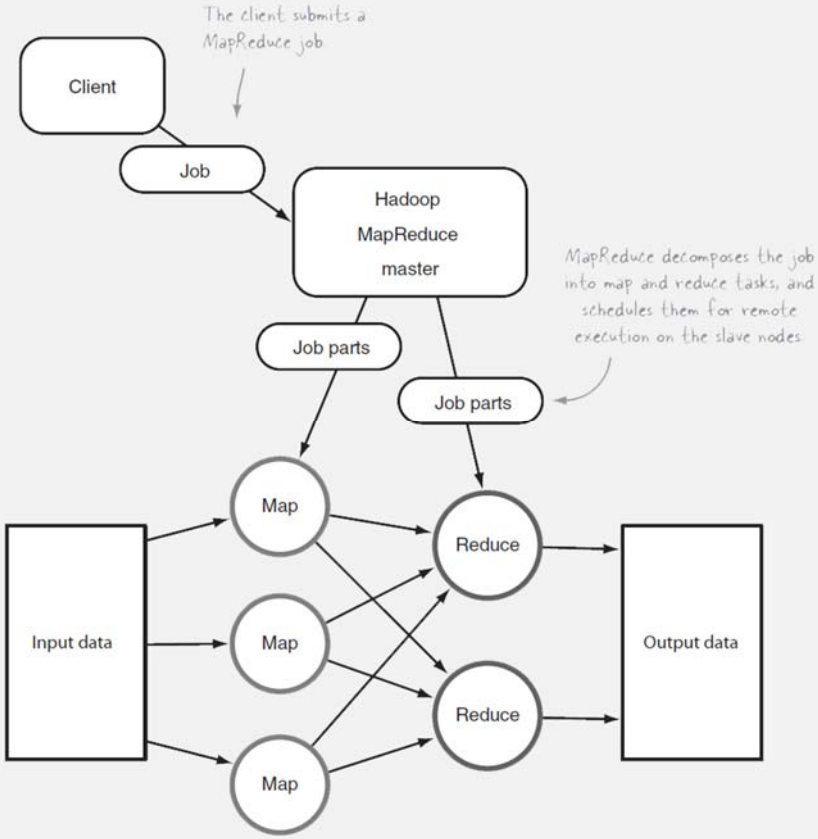
The Open Source Apache Hadoop Framework

- The proposed system has been designed to run on the **Apache Hadoop** open source framework:

Hadoop File System: HDFS



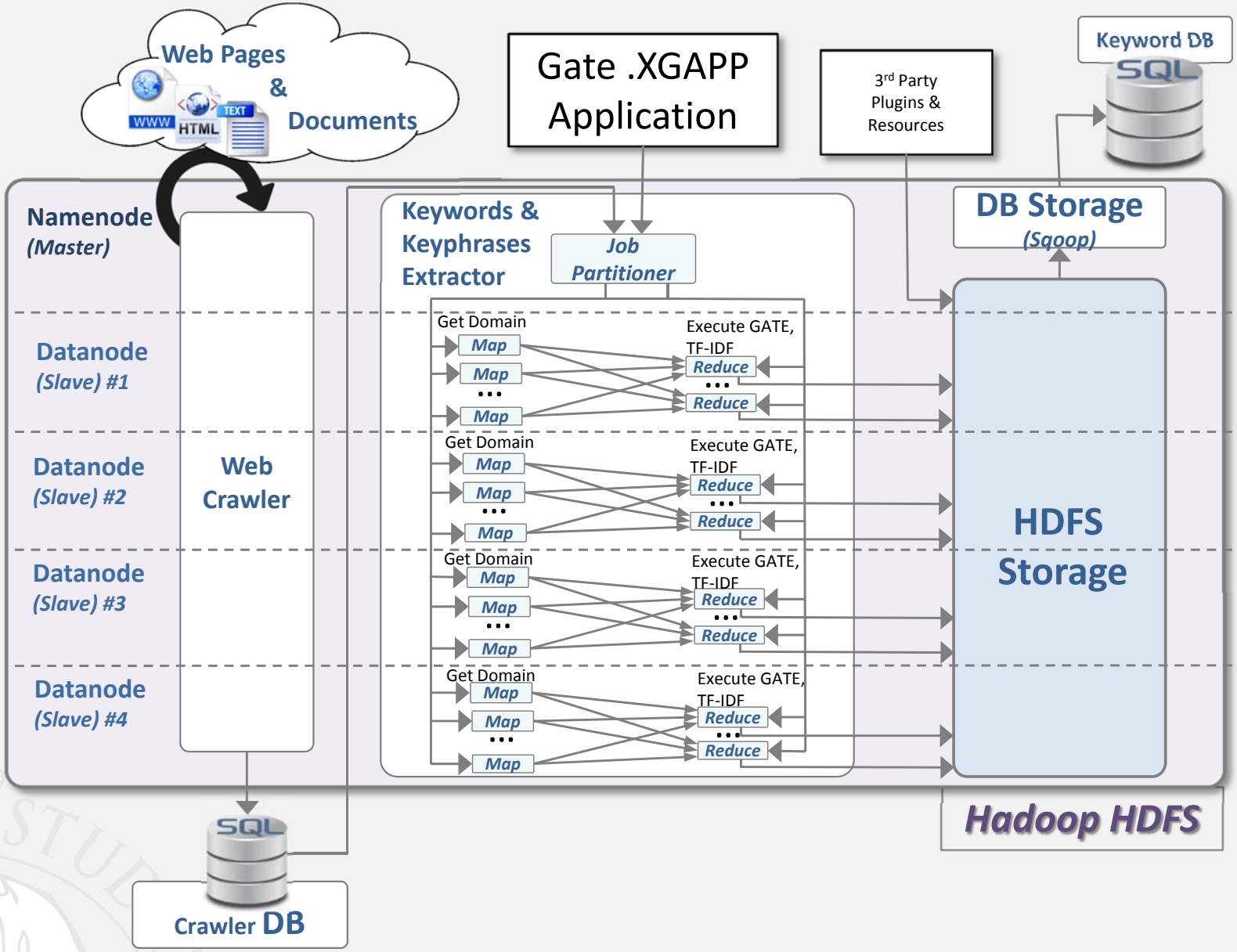
Hadoop MapReduce paradigm



- A significant advantage of using the Hadoop distributed architecture is the capability to efficiently and easily scale by adding inexpensive commodity hardware to the cluster.

[Figures reference: Holmes, A., "Hadoop in Practice", Ed. Manning (2012).]

General Architecture



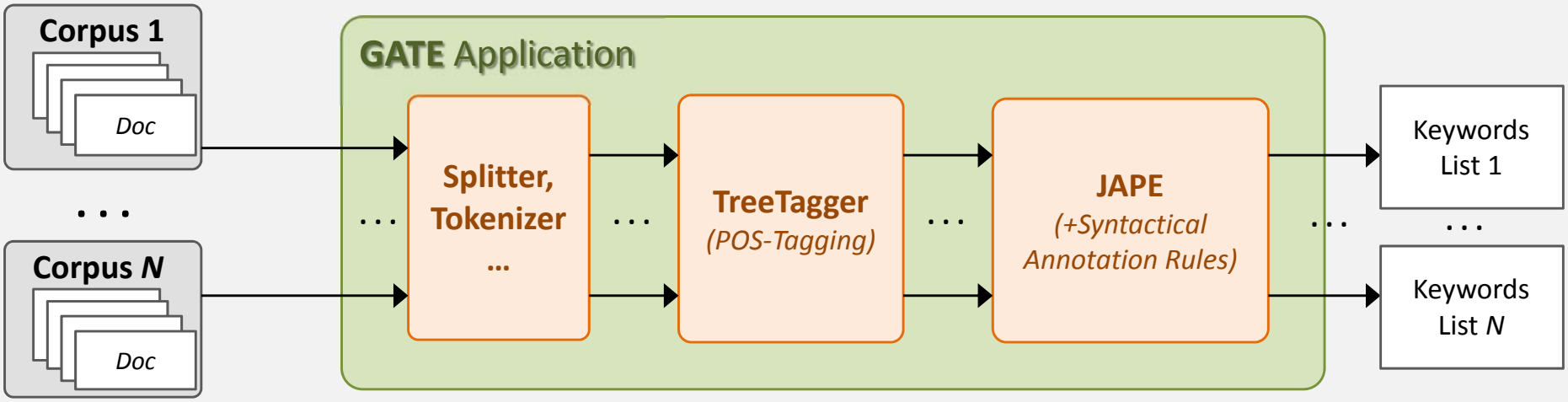
General Architecture

The system architecture is composed by 3 main modules:

- The **Web Crawler** module, based on the open source Apache Nutch tool;
  - To retrieves and parse the textual content of web pages.
- The **Keywords / Keyphrases Extractor module** is responsible for:
  - The execution of our NLP application (based on **GATE)** in a **Hadoop MapReduce** environment for text annotation and keywords / keyphrases extraction and storage in the **HDFS**.
  - The keywords and keyphrases relevance estimation, obtained by computing the TF-IDF function for each extracted keywords / keyphrases, performed to assess their relevance with respect to their whole corpus.
- The **DB Storage** module which finally stores designed keywords and keyphrases into an external SQL database, using the **Apache Sqoop** open source tool (specifically designed for data transfer between Hadoop HDFS and structured datastores).

Keywords & Keyphrases Extractor Module, .xgapp

- The **Map** function that associates key/value record pairs where the key is the URL of the single web page, and the value is the corresponding web domain (grouping web pages of the same domain).
- The **Reduce** function, in turn, fulfills the following operations:
  - Setup, launch and execution of a multi-corpora **GATE** application for keywords / keyphrases extraction:



  - Estimation of extracted keywords / keyphrases relevance at web domain level, by implementing the **TF-IDF** function:

$$\langle TF{-}IDF \rangle_k = TF_k \cdot IDF_k$$

$$TF_k = \frac{f_k}{n_d}, \qquad IDF_k = \log \frac{N_D}{N_k}.$$

Validation – Test Dataset and Configurations

- **Test Dataset**: 10000 web pages and documents ingested by the Web Crawler from a seed list of commercial companies, services and research institutes web domains.
- **Test Configurations**: Different test configurations for the Hadoop Cluster: from 2 to 5 active nodes.

Test Config. 1: *5-Nodes Cluster*

Nodes	Running Daemons
Master	*Namenode, JobTracker, TaskTracker*
Slave #1	*SecondaryNamenode Datanode, TaskTracker*
Slave #2	*Datanode, TaskTracker*
Slave #3	*Datanode, TaskTracker*
Slave #4	*Datanode, TaskTracker*

Test Config. 2: *4-Nodes Cluster*

Nodes	Running Daemons
Master	*Namenode, JobTracker, TaskTracker*
Slave #1	*SecondaryNamenode Datanode, TaskTracker*
Slave #2	*Datanode, TaskTracker*
Slave #3	*Datanode, TaskTracker*

Test Config. 3: *3-Nodes Cluster*

Nodes	Running Daemons
Master	*Namenode, JobTracker, TaskTracker*
Slave #1	*SecondaryNamenode Datanode, TaskTracker*
Slave #2	*Datanode, TaskTracker*

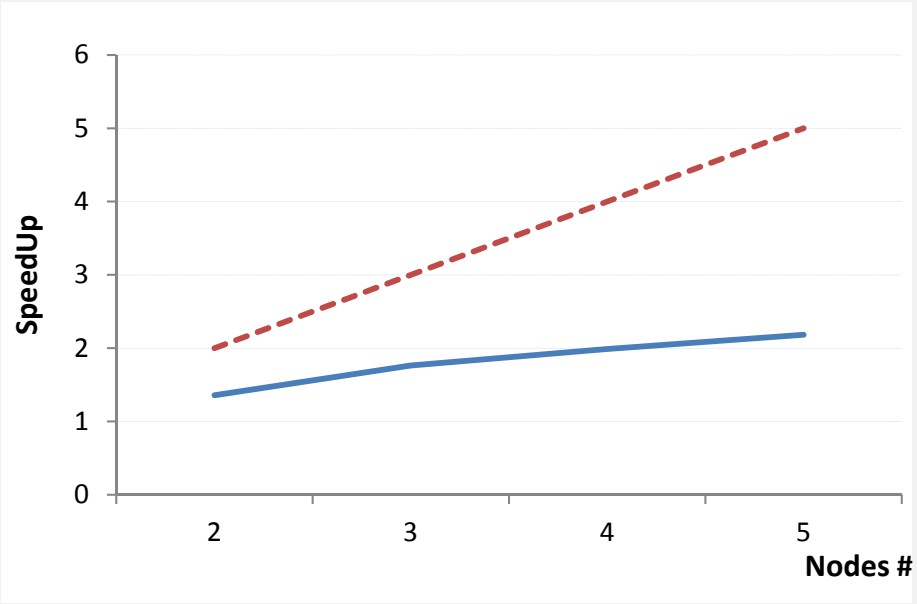
Test Config. 4: *2-Nodes Cluster*

Nodes	Running Daemons
Master	*Namenode, JobTracker, TaskTracker*
Slave #1	*SecondaryNamenode Datanode, TaskTracker*

Results and Discussion

- **Results**: For each cluster configuration, processing times for extracting a total of about 2.5 Million keywords and keyphrases have been assessed and compared.

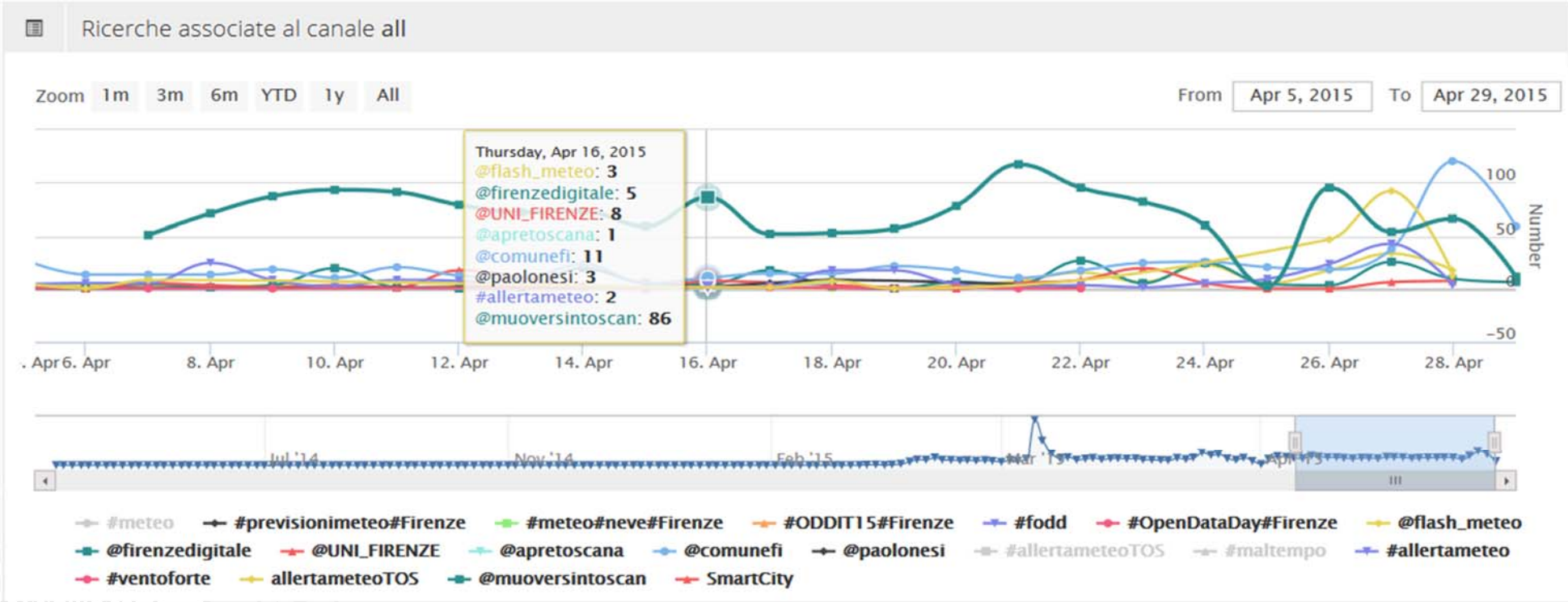
Configuration	Processing Time (hh:mm:ss)	Speed - Up
HDFS - single node	07:17:01	-
HDFS - 2 nodes	05:21:53	1.36
HDFS - 3 nodes	04:08:00	1.76
HDFS - 4 nodes	03:39:42	1,99
HDFS - 5 nodes	03:20:09	2.18



- The scaling capabilities observed confirm the nearly linear growth trend of the Hadoop architecture.
- Significant performance improvements can be achieved only for a larger number of nodes in the cluster.
- Single cpu/thread execution in 60 hours

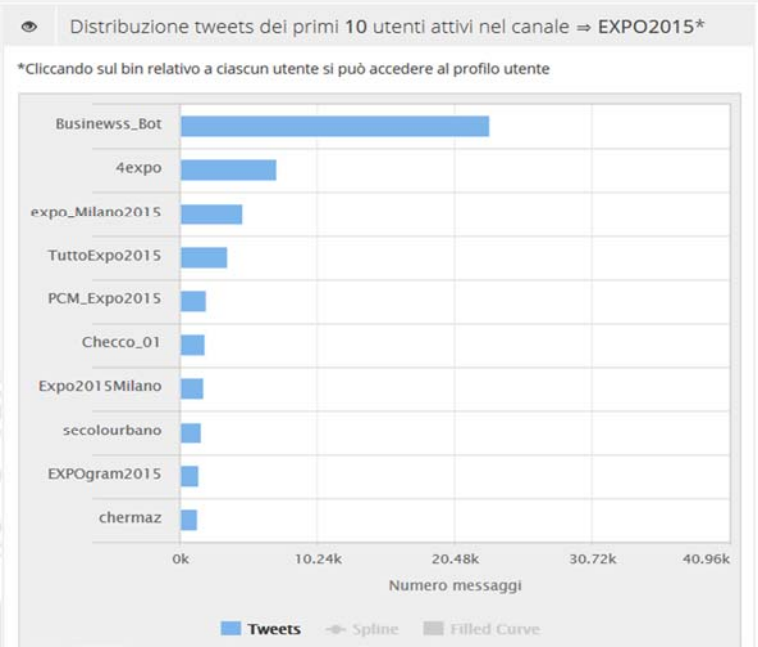
A Current Application

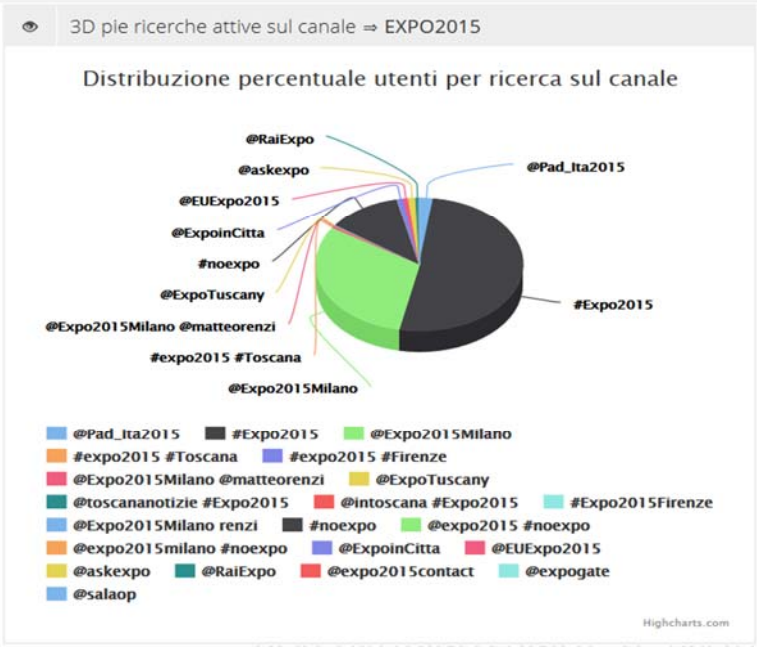
- Integration in the **Twitter Vigilance** tool (http://www.disit.org/tv/), developed at DISIT Lab, University of Florence.
- Twitter Vigilance is a multi user tool for making analysis of "Twitter channels". Each channel can be tuned to monitor one or more Search Queries on Twitter with a sophisticated and expressive syntax.
- Some public channels are publically accessible as examples.



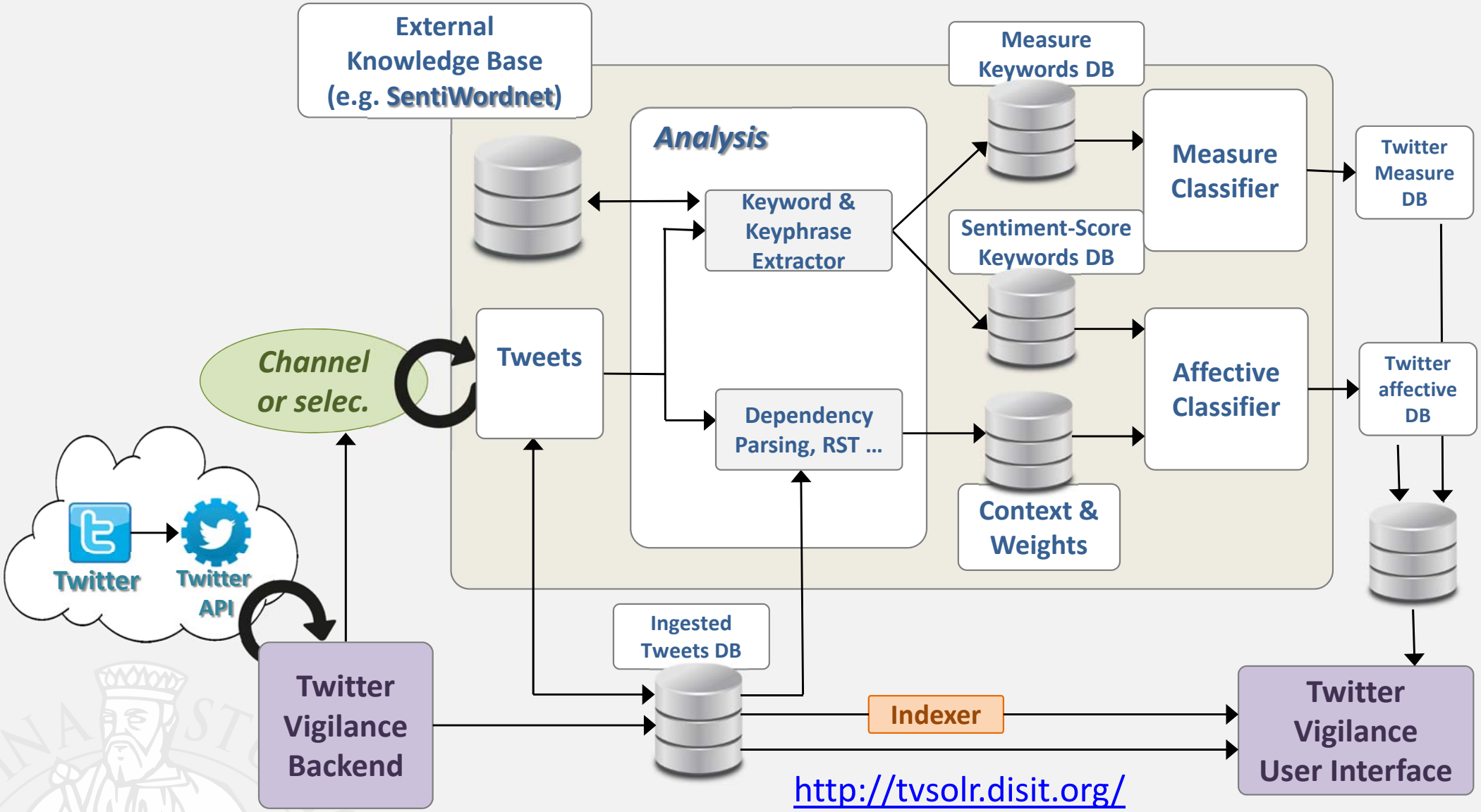
A Current Application

- Currently more than 12 Millions Twitters have been processed, at a rate of approximately 200000 New TW per day, so a distributed architecture is required.
- Main Channel focus:
  - City Service Assessment: smart city application
  - Weather conditions and measures
  - Weather communication channel assessment and qualification
  - Events appreciation assessment
  - Product market appreciation assessment
  - Drugs vs people appreciation assessment





Sentiment and Measure Analysis of Social Media (Twitter)



http://www.disit.org/tv/

Conclusions

- A NLP application to enforce the NLP capabilities into Hadoop has been designed and built
- Advantages with respect to the state of the art are:
  - usage of GATE, instead of creating a new one
  - Different kind of text sources and not only documents
  - Larger flexibility
- The resulting solution
  - has been assessed in terms of performance
  - Is currently in further development and use for Twitter Vigilance Affective and Measuring analyses in multiple projects with different institutions: CNR IBINET, LAMMA, NEUROFARBA

Sentiment and Measure Analysis of Social Media (Twitter)

- Extracted keywords could be monitored by different POS (nouns, adjectives and verbs)
  or by special characters for user mentions (@) and hashtags (#), in order to asses temporal trends and perform statistical analysis.

- By exploiting external knowledge bases and lexicons specifically designed for Sentiment Analysis (such as **SentiWordnet**), sentiment scores can be assigned to each extracted keyword and keyphrases (as compositions of keywords) in order to estimate the general sentiment polarity of collected tweets.

- An advanced Sentiment Analysis could be eventually performed by employing more sophisticated techniques, such as dependency parsing, Rhetorical Structure Theory (RST) and Discourse Following, in order to perform syntactical sentence parsing and extract relations and dependencies among keywords and syntagmas.
  - *These resources can be then used to weight sentiment coefficients previously assigned to extracted keywords, in order to improve the estimation of sentiment.*

References

[Chung and Moldovan, 1995] M. Chung and D. I. Moldovan, "Parallel Natural Language Processing on a Semantic Network Array Processor", IEEE Transactions on Knowledge and Data Engineering, Vol. 7(3), pp. 391-404, 1995.

[Exner and Nugues, 2014] Exner, P. and Nugues, P., "KOSHIK - A Large-scale Distributed Computing Framework for NLP", in Proc. of the International Conference on Pattern Recognition Applications and Methods (ICPRAM 2014), pp. 463-470, 2014.

[Hamon et al., 2007] T. Hamon, J. Deriviere and Nazarenko, "Ogmios: a scalable NLP platform for annotating large web document collections", in Proc. of Corpus Linguistics, Birmingham, United Kingdom, 2007.

[Holmes, 2012] Holmes, A., "Hadoop in Practice", Ed. Manning (2012).

Thanks for your attention !