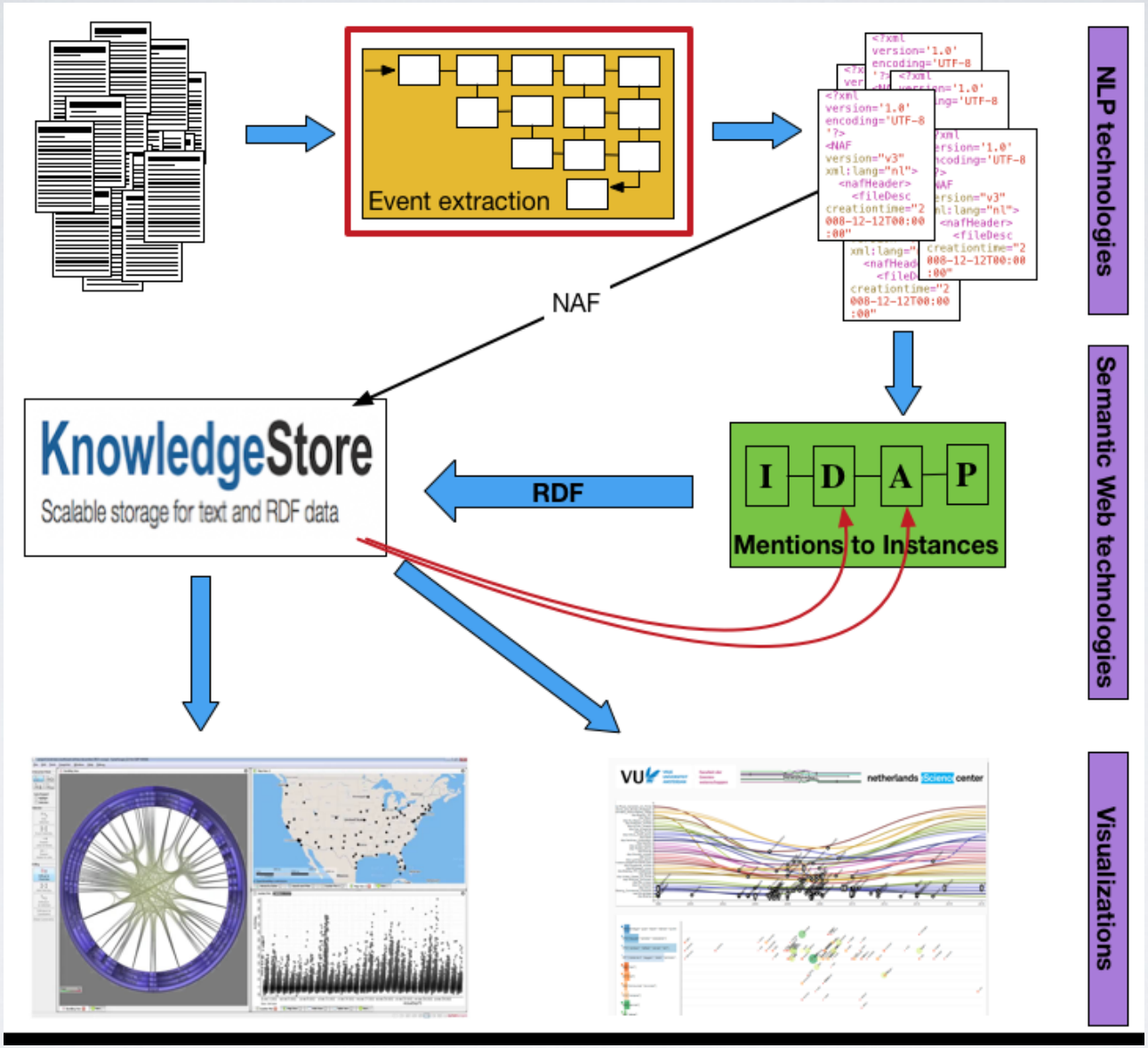
Newsreader distributed cluster

The Newsreader project

- Event detection in four languages (English, Dutch, Spanish and Italian)
  - what, who, where and when and relations between events (time, cause)
  - factual/non-factual or speculative
  - provenance: who tells what and when
- Large-scale processing of events
  - millions of articles spanning ten years or more
  - process new information as fast as possible

The Newsreader project



NLP modules for event mining

- English pipeline: 19 modules
  - Tokenization
  - POS tagger
  - Word Sense Disambiguation
  - Named Entity Recognition and Disambiguation
  - Semantic Role Labeling
  - Time expressions
  - ...

Distributed processing: Main goals

- Analyze and implement alternatives for dealing with large volumes of textual data.
- Having a pipeline comprised of third-party NLP modules, execute them in a distributed environment as efficiently as possible.
  - Avoid rewriting existing modules.
  - Requisite for modules: consume NAF (STDIN), produce NAF (STDOUT)

Distributed processing: Main goals

- Two processing paradigms:
  - *Batch*: a large set of documents available before any computation starts. Computations start and end within a given time frame.
  - *Streaming*: documents may arrive at any moment (typically, very often). Processing never ends.

Batch architecture. Characteristics

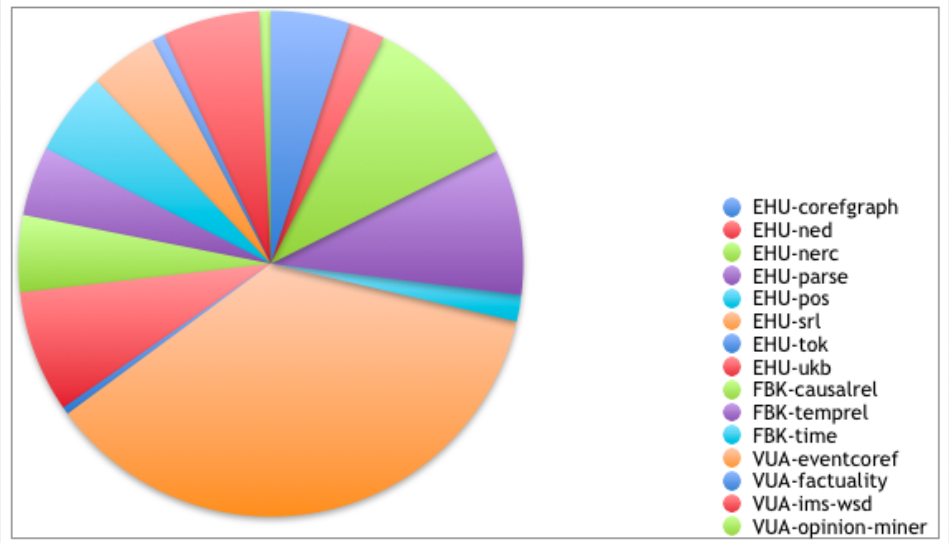
- Hadoop architecture
  - processing large data sets distributed across clusters.
- Cascading
  - supports complex workflows.
  - no need to adapt modules to *MapReduce*.

Processing dataset

- Batch architecture used to process 2.5 million articles.
- 86 nodes, 8 cores, 64GB RAM and 16TB disk.
  - 4 worker per node
  - 10 documents per worker

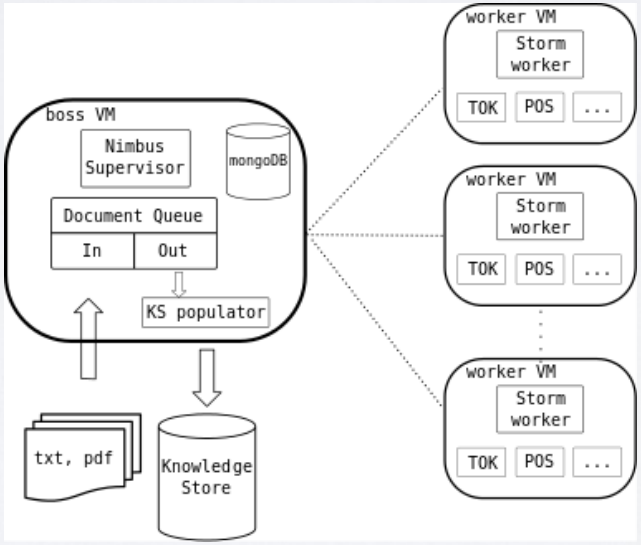
Processing dataset

- Based on the processing:
  - 264 seconds per document on average.
  - ~327 documents processed per core per day.
  - SURFsara has 1,400 cores on Hadoop
    - 458K documents per day at SURFsara.
  - 80,000 core hours per million documents:
    - We need 3,350 cores for a day.
- SURFsara can handle a day of news within 2.5 days
  - A single computer will need 9 years.



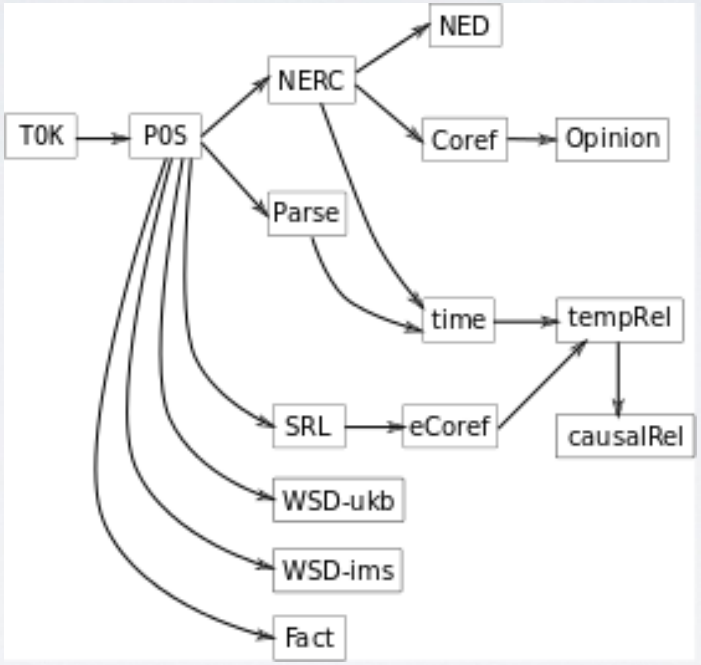
Streaming architecture. Characteristics.

- A system composed of many virtual machines (VMs) distributed across one or more physical machines.
- Two types of VMs:
  - *boss VM*: The main/manager node. It contains the in/out queues, database, Zookeeper, Storm, ...
  - *worker VM*: The VMs that performs the processing. They contain all the NLP modules.



Non-linear topologies

- Instead of running the modules serially, run those which do not depend on each other in parallel.



Non-linear topologies. Experiment.

- Send 1000 documents randomly (*Poisson* process)
- Using 1 boss and 7 worker nodes

Proc. time	Elapsed time	Idle	Latency (doc/sent/token)
290,276	148,156	7,558	148.60/4.18/0.17

- Parallel processing decreases latency by 50%

Knowing more

- NewsReader website

```
http://www.newsreader-project.eu/
```

- Deliverable D2.3

```
http://kyoto.let.vu.nl/newsreader_deliverables/NWR-D2-3.pdf
```