Temporal characteristics of dynamic bibliographic networks

Michaël Waumans

Université Libre de Bruxelles - ULB

mwaumans@ulb.ac.be

Overview

Context

Context

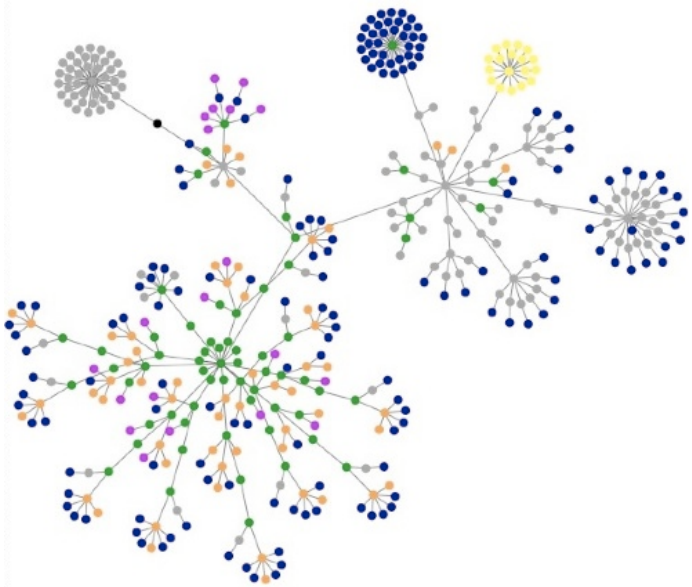
Graph Theory

Study of graphs in mathematics and computer sciences

Graphs

Mathematical structures used to model pairwise relations between objects. It is made up of vertices (nodes) and edges that connect them.

- **Directed** :
  From one vertice to another
- **Undirected** :
  No distincton made for the direction



Context

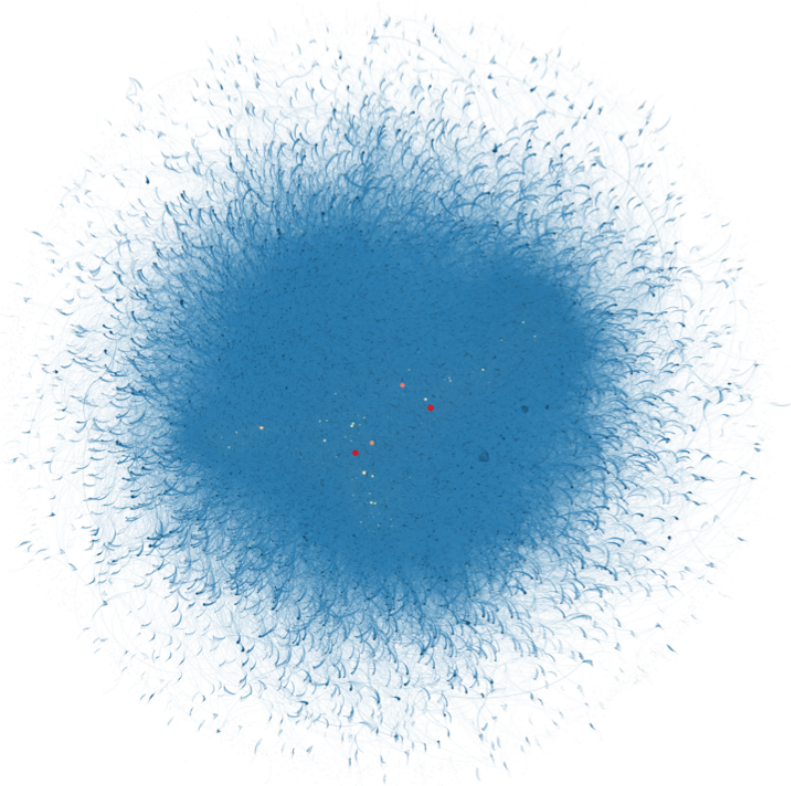
Network Analysis

Part of graph theory that concerns itself with the study of graphs as a representation of relation between discrete objects.

1. Temporality
   - Static
   - Dynamic / Temporal
2. Edge weight
   - Unweighted
   - Weighted networks

Examples of networks are numerous

- Social, Friendship, Communication,...
- Biological, Gene Regulatory,...
- Text, Speech,...
- Reference / Bibliographic, Co-author, Affiliations,...



Context

Metrics and algorithms

- Degree : In degree, Out degree, degree distribution
- Centrality : Betweeness, Closeness, Distance,...
- Clustering, Hubness,...
- Prestige : PageRank[S. Brin L.Page, 1998], Hits[Jon M. Kleinberg, 1999],...
- Prediction : EventRank[O'Madadhain Smyth, 2005], FutureRank[H.Sayyadi L.Getoor, 2008], T-Rank[Berberich Vazirgiannis Weikum, 2004]...
- ...

Models

- Random
- Erdos-Renyi
- Barabasi-Albert[Barabasi Albert-Laszlo Reka Albert, 1999], Preferential Attachment
- ...

Context

Subjects covered

1. **Bibliometric / Scientometric** : Development of methods to analyze networks of articles from the scientific litterature
2. **Model** : Verify the validity of preferential attachment in reference networks and propose and alternative model if necessary
3. **Classification** : Ranking nodes among such networks
4. **Prediction** : Predict the ranking of a node inside the network

Bibliographic Networks

Bibliometrics : Datasets used

1. **ArXiV HEP-Th**
   - More than 29.000 nodes
   - From 1992 to 2002
2. **ArXiV HEP-Ph**
   - More than 35.000 nodes
   - From 1992 to 2002
3. **APS**
   - More than 450.000 nodes
   - From 1892 to 2012

[ArXiV, 2013][American Physical Society, 2013]

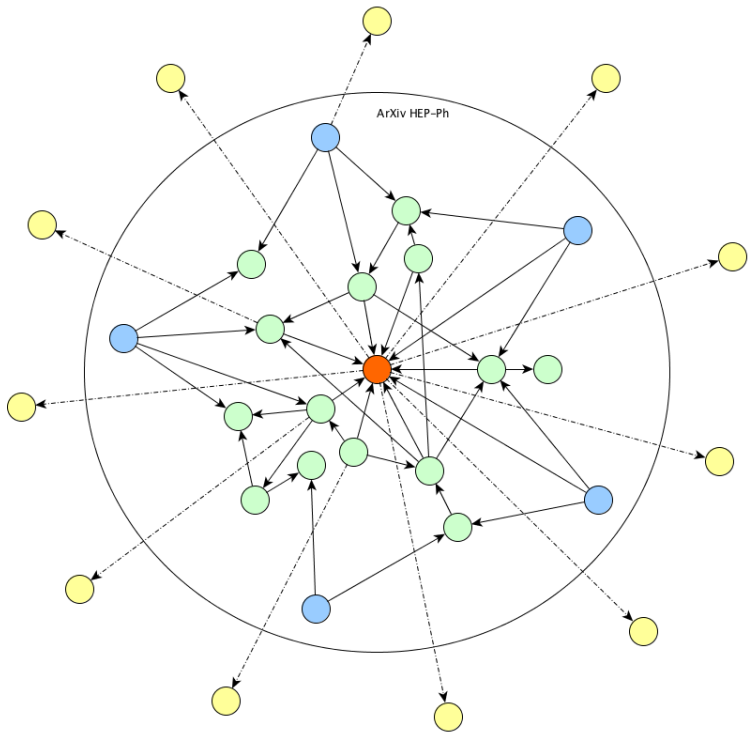
Bibliometrics : Datasets Pitfalls

Internal references only !

1. Articles with high in-degree value may have a very low or null out-degree value
2. Lots of articles have no incoming references (Roughly 20%)
3. Lots of articles do present a very low variation of growth
   (Roughly 50 to 70% depending on the characterization of "low variation")

But also...

1. Short time period
   (10 years)
2. Specific field
   (Physics)



Bibliometrics : Summary

Questions to answer

1. Is it possible to study the actual properties of reference networks using those datasets ?
2. Are the properties observable in such datasets the same as in other fields of research than physics ?
3. Are reference networks really evolving following the model of preferential attachment or are we missing something ?

Solution

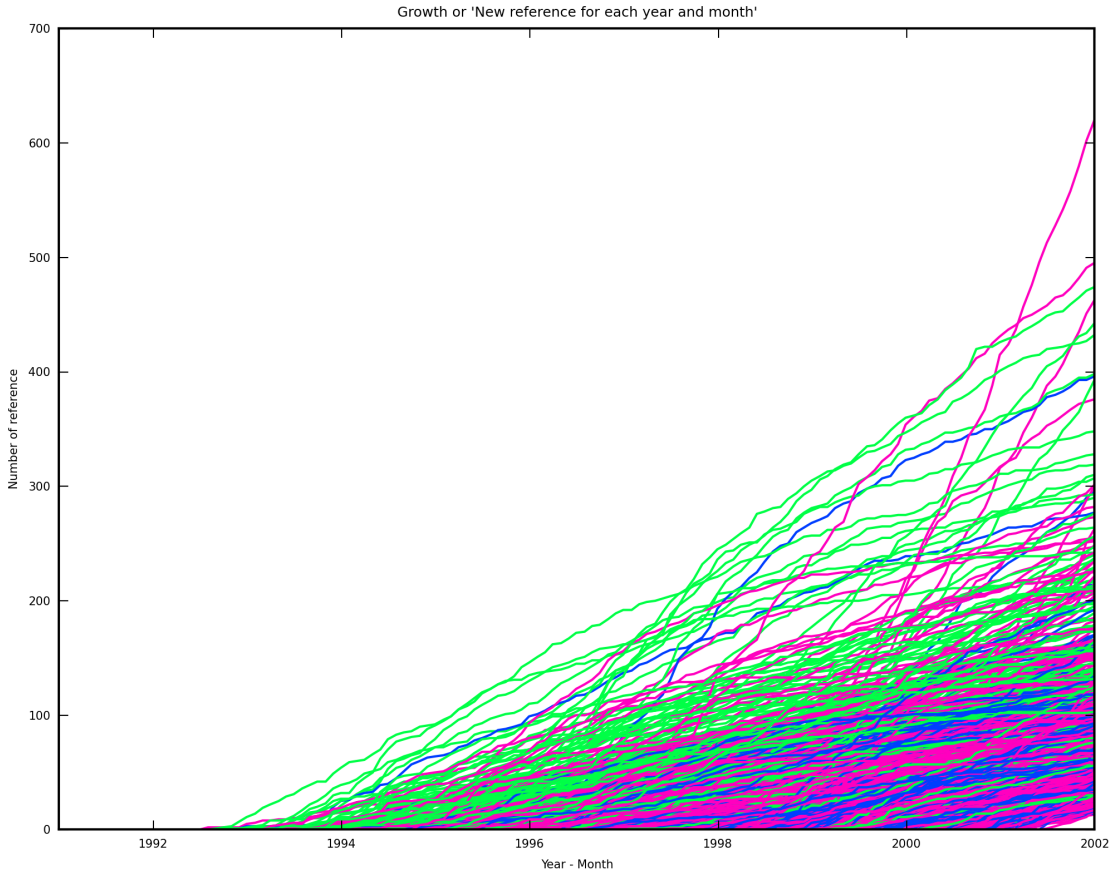
Build a widder network including reference sub-networks from different publishers (ArXiV[ArXiV, 2013], APS[American Physical Society, 2013], SpringerLink[SpringerLink, 2013], PubMed[?],...) and also use full-text reference matching to limit the loss of information to a minimum.

BUT

Considering a denser network will obviously increase the size of the graph which won't be treatable by an any algorithm on a personal machine in a reasonable time. (Cluster ?)

Experimentations

Experimentation : Evolution of the growth

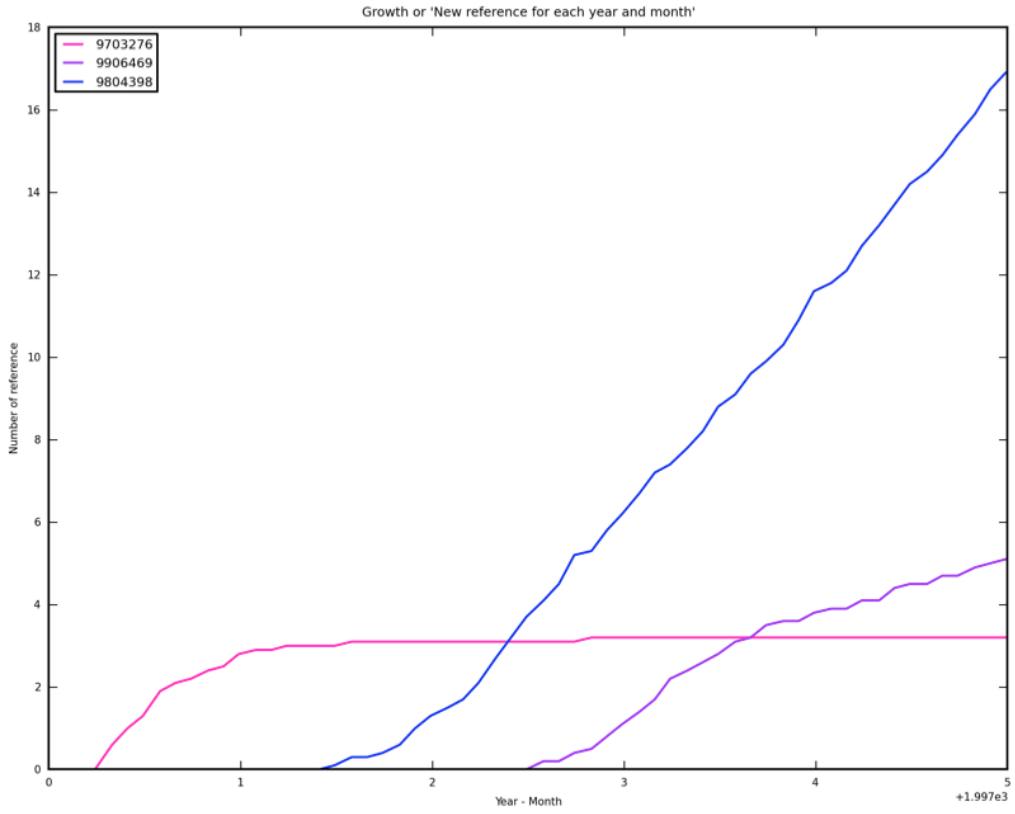


1. Are there any recognizable patterns among the individual progression of the articles in terms of growth ?
2. Is it possible to classify articles among those classes of "growth type" ?
3. Is it possible to predict the popularity of a node according to this classification ?

Experimentation : Observations

Characteristics

1. Specific growth behaviors
2. Life of an article
   - First years of life are determinant
   - Following years usually display linear progression
3. "Death" of an article
   - Sudden death of an article (Log-growths)
   - Natural death of an article usually implies a growth lowering down after some time



Experimentation : Summary

Questions

1. Is it possible to rank nodes according to their growth behavior ?
2. Is it possible to predict the class of a node in terms of popularity ? (unpopular, popular, dying, normal,...)

Still...

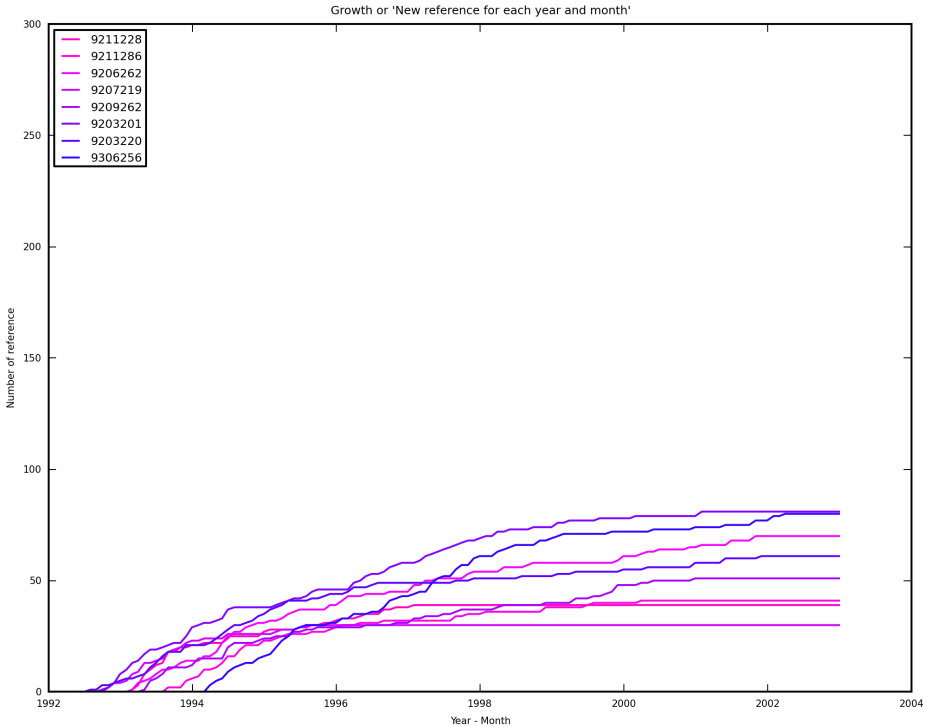
Importance of more sub-networks and bigger networks since those observation may differ fron one network to another. Depending on the field of research covered for example, or depending on the age of the articles (more difficult access to scientific ressources)
$\Rightarrow$ DOI widely used only since 2005-2007
$\Rightarrow$ Impact of systems like SpringerLink, ArXiV or even Google Scholar.

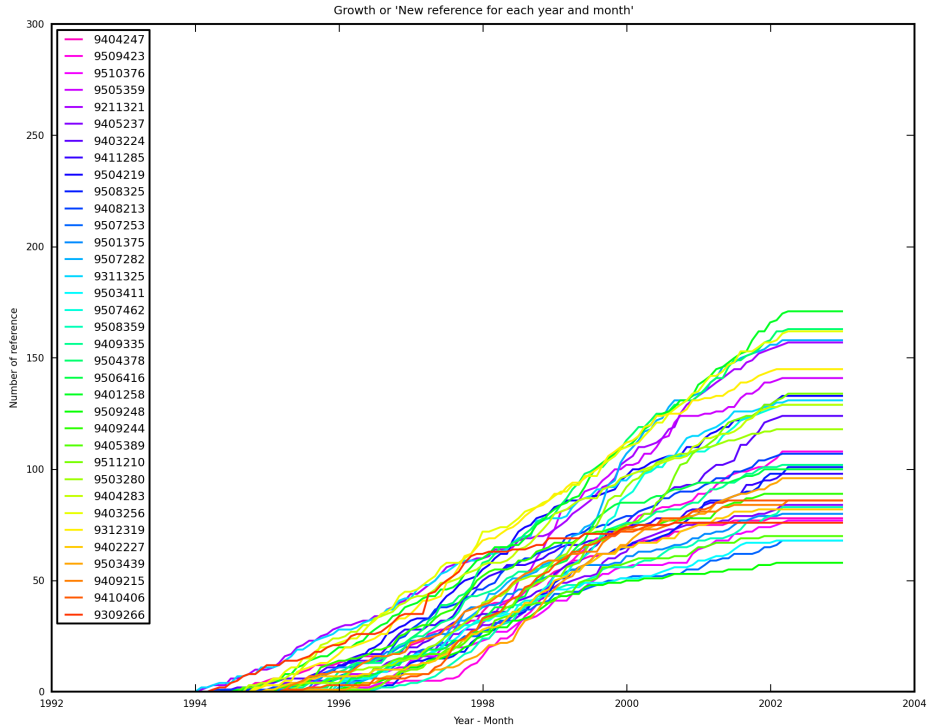
Classification

Classification : Proposal

Ranking nodes according to their growth rate over time

Compare the growths of each node in the network to separate them in four classes at least. The impopular nodes, the nodes with an early high growth variation ("buzz" in social networks), the nodes with a high and long term impact and the nodes with a "normal" popularity.





Classification : Algorithm

1. Building the network (Unweighted acyclic network with temporal metadata for each node)
2. Building the growth data from the network.
3. Pre-processing
   - Identifying nodes with no growth (Null final in-degree)
   - Identifying nodes with very low variation of growth (...)
4. Normalization (Examining fixed time-frames)
   - X : Simple normalization $x \exists [0, 1]$
   - Y : Normalization following a uniformization of the growth curves $y \exists [0, 1]$ where max in-degree always equals 1.
5. Polynomial fitting of each growth curve (degree 2 or more)
6. Execution of the KMeans clustering algorithm on the features extracted (polynomial coefficients)

Classification : Algorithm illustration

Intermediate results

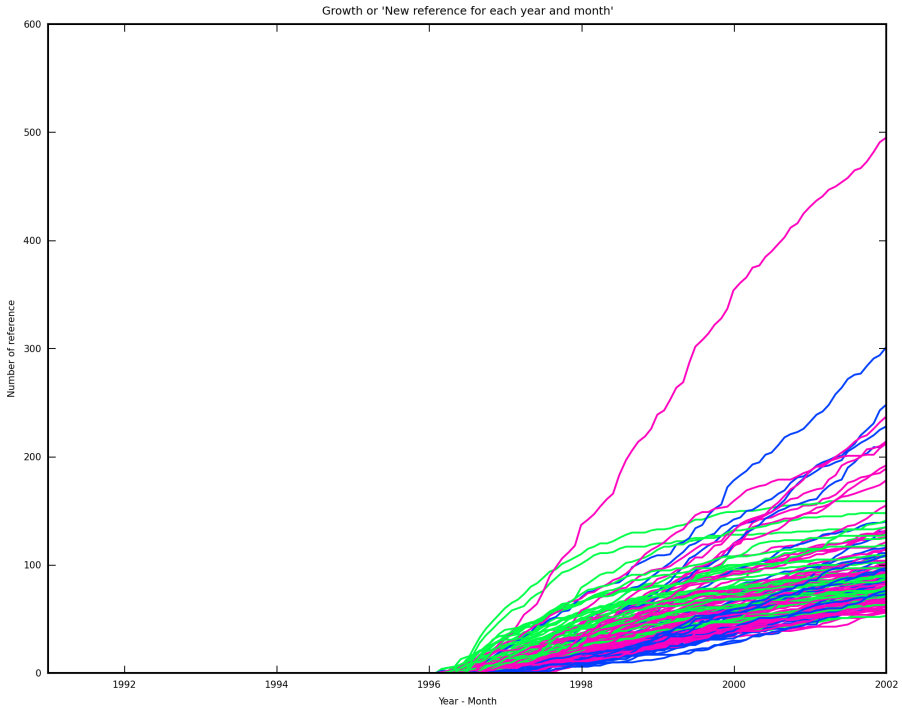


Figure : Original growth

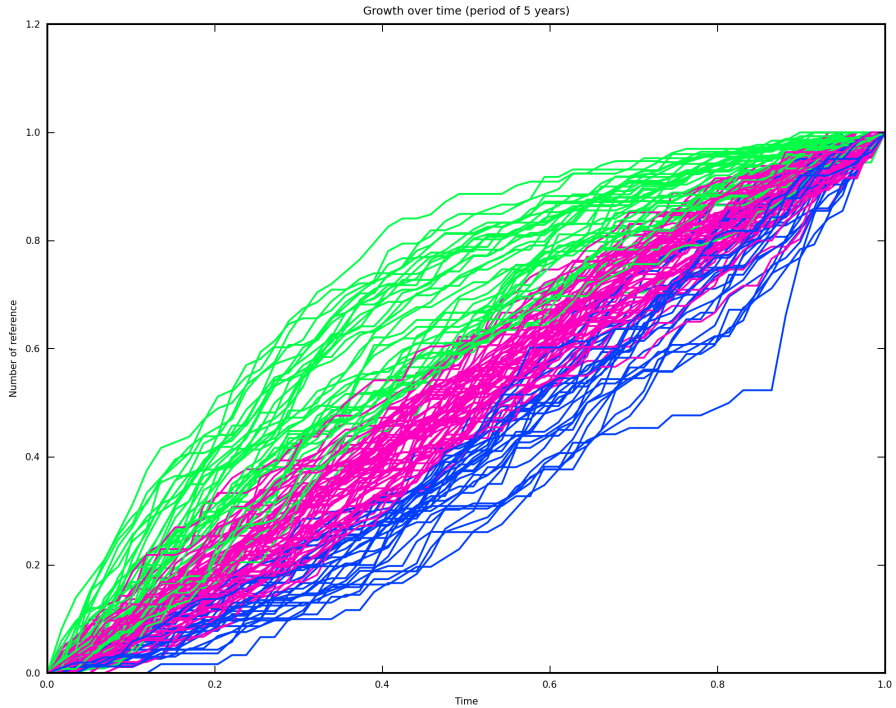


Figure : Normalized growths

Classification : Summary

Current results

Classification using three classes plus the two others, the articles with an in-degree value of zero and the ones with a very low growth variations.

Work in progress

1. Use total in-degree for classification
2. Split into more classes
3. Use of weighted edges to reflect the importance of each reference
4. Increase the importance of high and quick variations of growth
5. Possible smoothing pre-processing on the lower growths
6. If metadata available, remove suspect edges due to collaboration
7. Split growth curves after a few years and establish the ranking according first, to the first years of life, then the end. The first variations being apparently the most determinant in an article's life.

Prediction

Prediction : How ?

Fitting in each classes, the articles with an appropriate growth

1. No influence
2. Low influence (very low variation of growth)
3. "Normal" influence
4. High influence
5. High impact but short life
6. ...

Important

The first years after the publication of an article are determinant for its future. A period of 3 to 5 years (depending on the field) is enough to predict with a good precision the category in which it fits.

Prediction

Problem

Working with temporal networks implies that the articles are not published at the same date and thus may need to be compared even tough we do not possess the same amount of information about their growth (the periods covered are not the same)

Proposal

1. **Static time window :** Consider the same period of time but not the same starting point. Implies a loss of information.
2. **Opening time window :** Consider all the data available for each article. A normalization step is then required but the comparison may be very wrong if the time frames are too different.
3. **Sliding time window :** Consider the same period of time and slide the frame considered with time. Making this approach very pertinent in the context of real time social network analysis (buzz detection,...) but not relevant here.

Future work

Future Work

Bibliometric

Particular growth behaviors observed as well as potential characteristics specific to reference networks.
$\Longrightarrow$ Use other datasets to verify those preliminary observations and build a wider network to eliminate the bias induced by the consideration of internal networks only.

Classification / Ranking

Method proposed and currently tested with different heuristics to rank nodes according to their growth rate.
$\Longrightarrow$ Apply other clustering algorithms, Use different sets of features, test a two step classification method,...

Prediction

The mechanism for predicting the popularity of nodes depends heavily on the classification algorithm and is a bit unappropriate in reference networks.
$\Longrightarrow$ Apply this method to real-time Twitter networks to target potential buzz or popular persons inside a networks of Tweets at some point in time

Other work

Other work

Network Analysis

1. Extraction of friendship and dialog networks from litterature
2. Building SpringerLink-based reference networks
3. Social network analysis (Tweets networks)
4. Text classification using text network topology comparison
5. Development of a database to store one large reference network using DOI matching as well as full text reference matching.

Hadoop Cluster

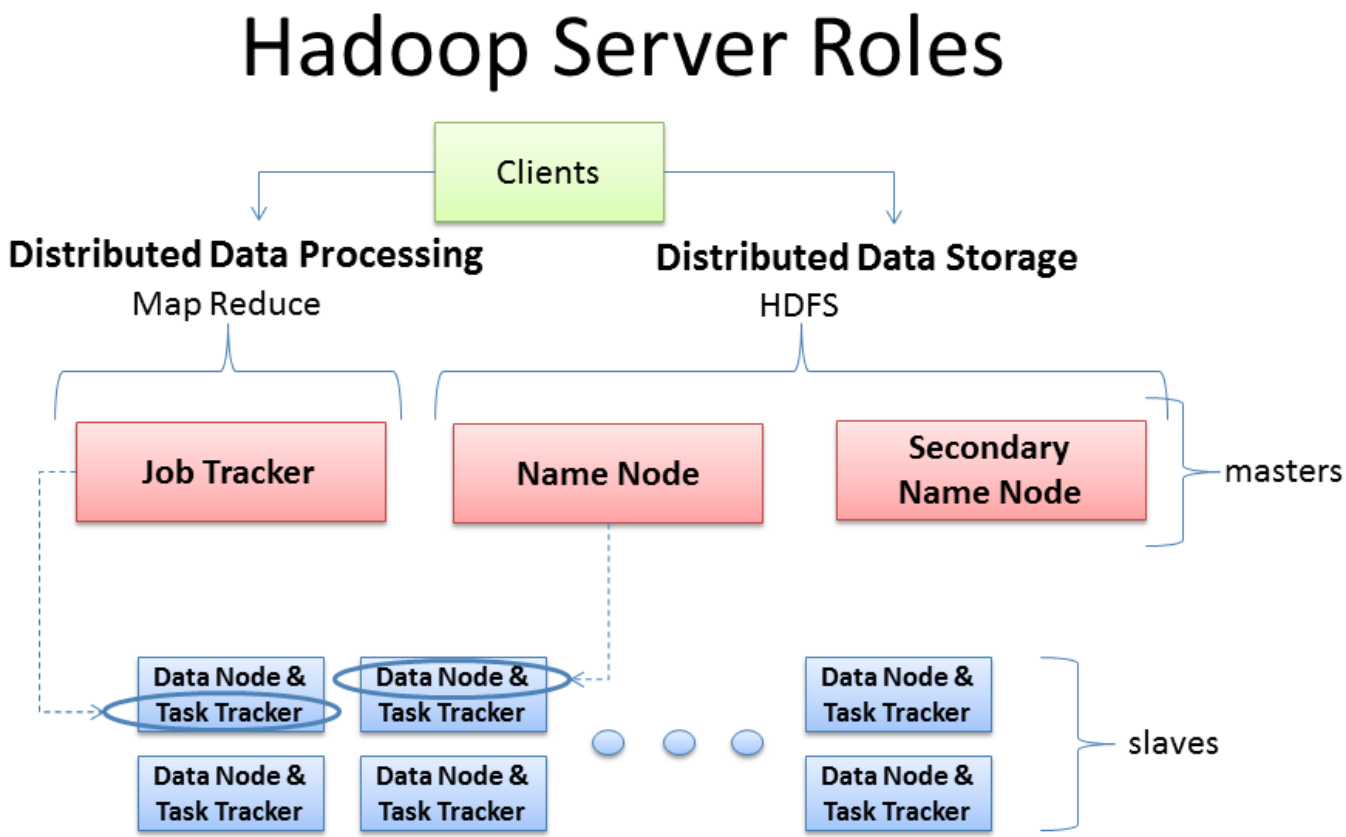
Deployment, configuration and testing of Cloudera on a cluster composed of 20 machines.

Hadoop Cluster

Hadoop Cluster : Architecture

Principle

Distributed computing and storage without single point of failure using simple server configuration (no raid).

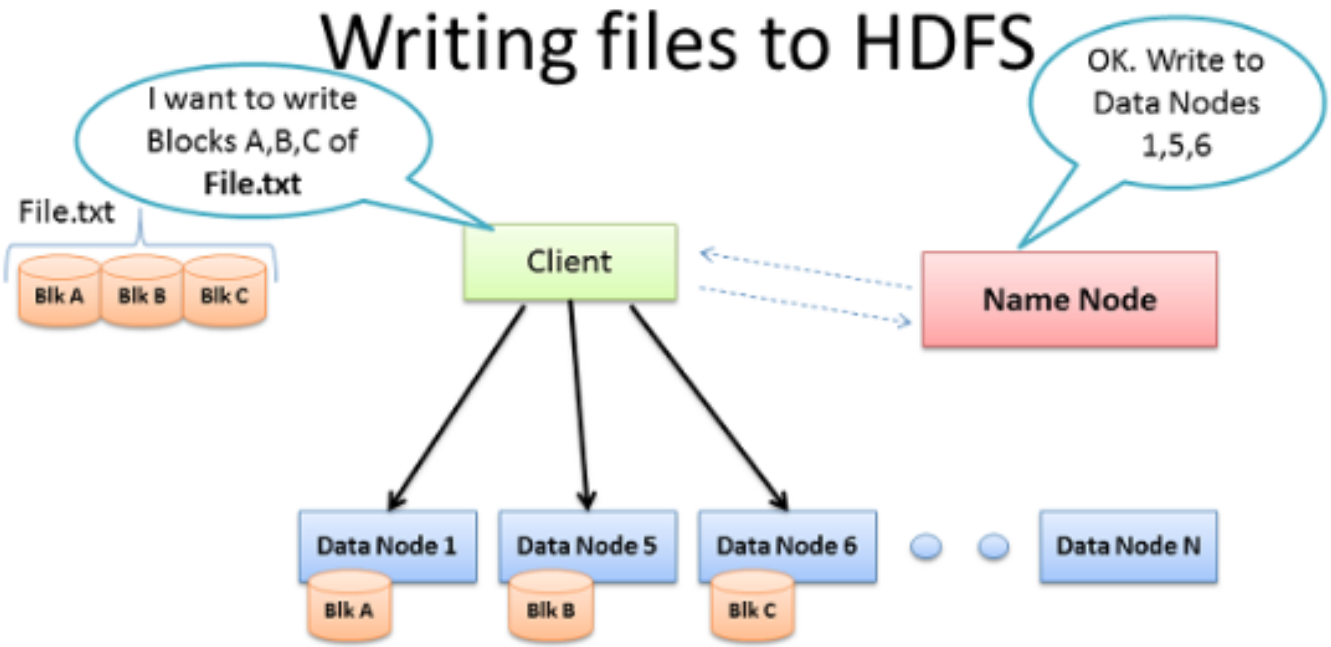


Typical Workflow

- Load data into the cluster (HDFS writes)
- Analyze the data (Map Reduce)
- Store results in the cluster (HDFS writes)
- Read the results from the cluster (HDFS reads)

Figure : Hadoop Workflow

Hadoop Cluster : Write Hdfs



- Client consults Name Node
- Client writes block directly to one Data Node
- Data Nodes replicates block
- Cycle repeats for next block

Figure : Hadoop Write Hdfs

Hadoop Cluster : Block replication

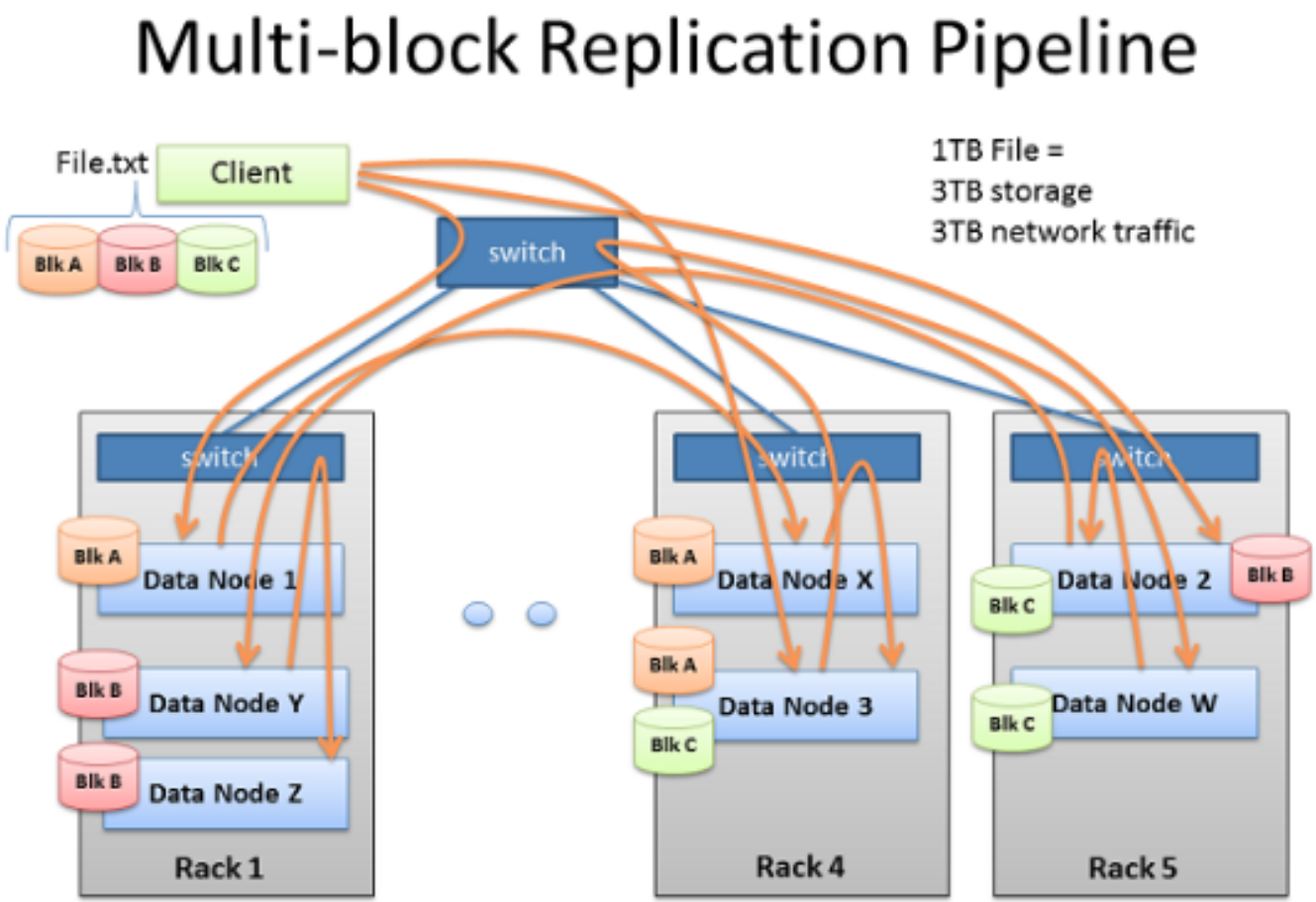
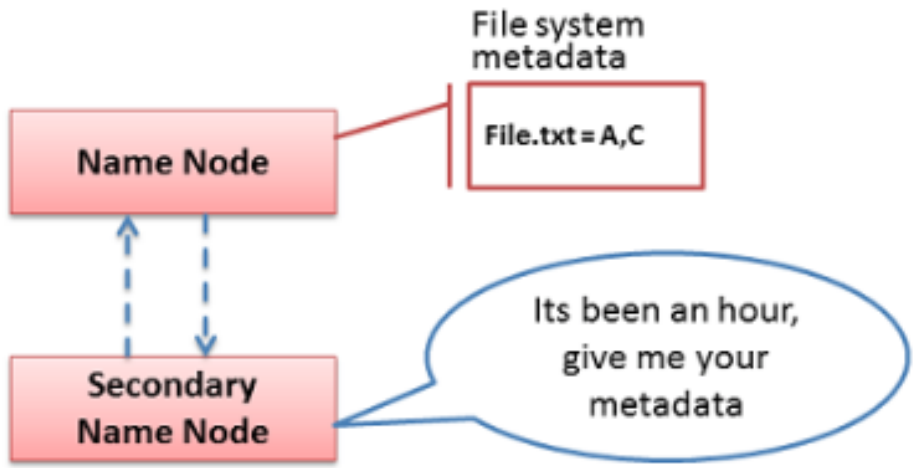


Figure : Hadoop Block Replication



- Not a hot standby for the Name Node
- Connects to Name Node every hour*
- Housekeeping, backup of Name Node metadata
- Saved metadata can rebuild a failed Name Node

Figure : Hadoop Secondary NameNode

Hadoop Cluster : Software Solution

CDH : Cloudera's Open Source Distribution of Hadoop

Hadoop solution that offers batch processing, interaction SQL and interactive search as well as enterprise-grade continuous availability.

Services

- Storage : HDFS, HBASE
- Batch processing : MAPREDUCE, HIVE, PIG
- Interactive SQL : IMPALA
- Machine Learning : MAHOUT
- Graphs : GIRAPH

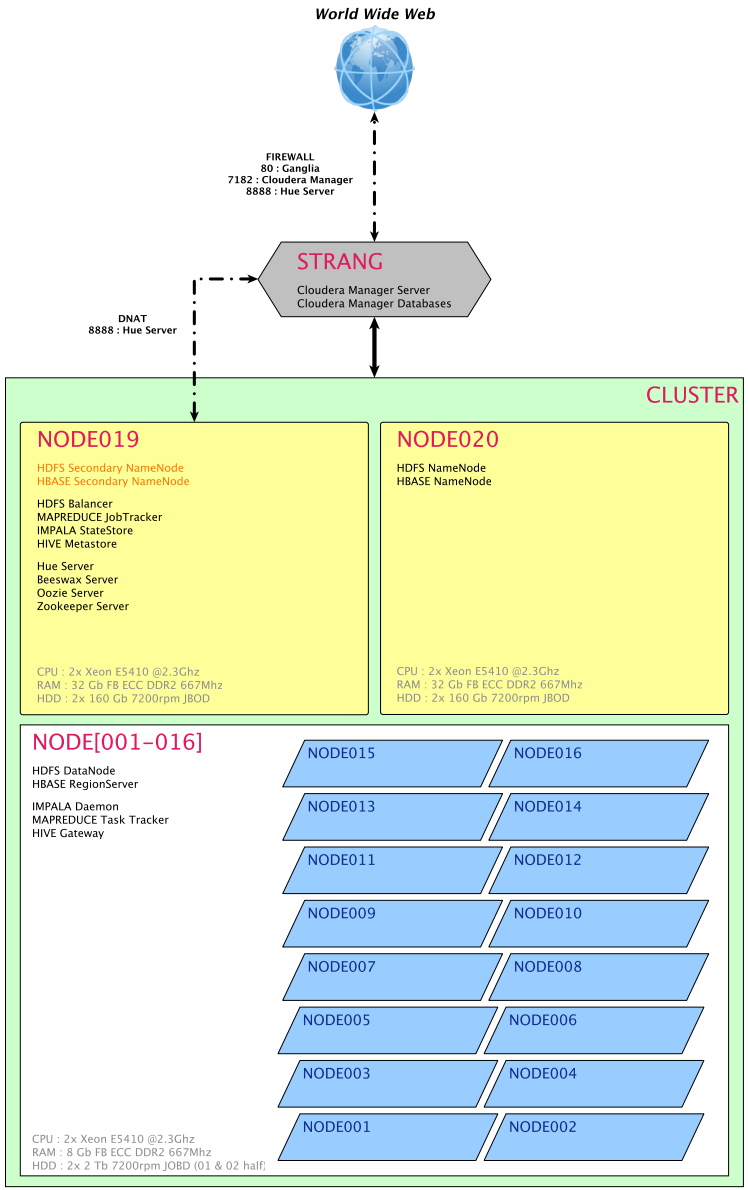
Hadoop Cluster : Nodes configuration

1. 20 nodes
   - 1 Head (STRANG) : DHCP, Firewall, Cloudera-Manager, Cloudera-Databases
   - 1 Inactive (ROENTGEN)
   - 2 Masters (019-020) : NameNode, Secondary NameNode,...
   - 16 Slaves (001-016) : DataNode, TaskTracker,...
2. CPU : 2x Intel XEON E5410 Quad 2.33Ghz
3. RAM : From 8 (Slaves) to 32 (Masters) Gb
4. HDD : 4 Tb JBOD (Slaves), Varying (Masters)

Suboptimal configuration

16 to 32 Gb of RAM required for each slave node, as well as RAID mounted HDD 1 Tb for the NameNode and Secondary NameNodes.
No heap space left for backup of lack of space to store the metadatas for HDFS.

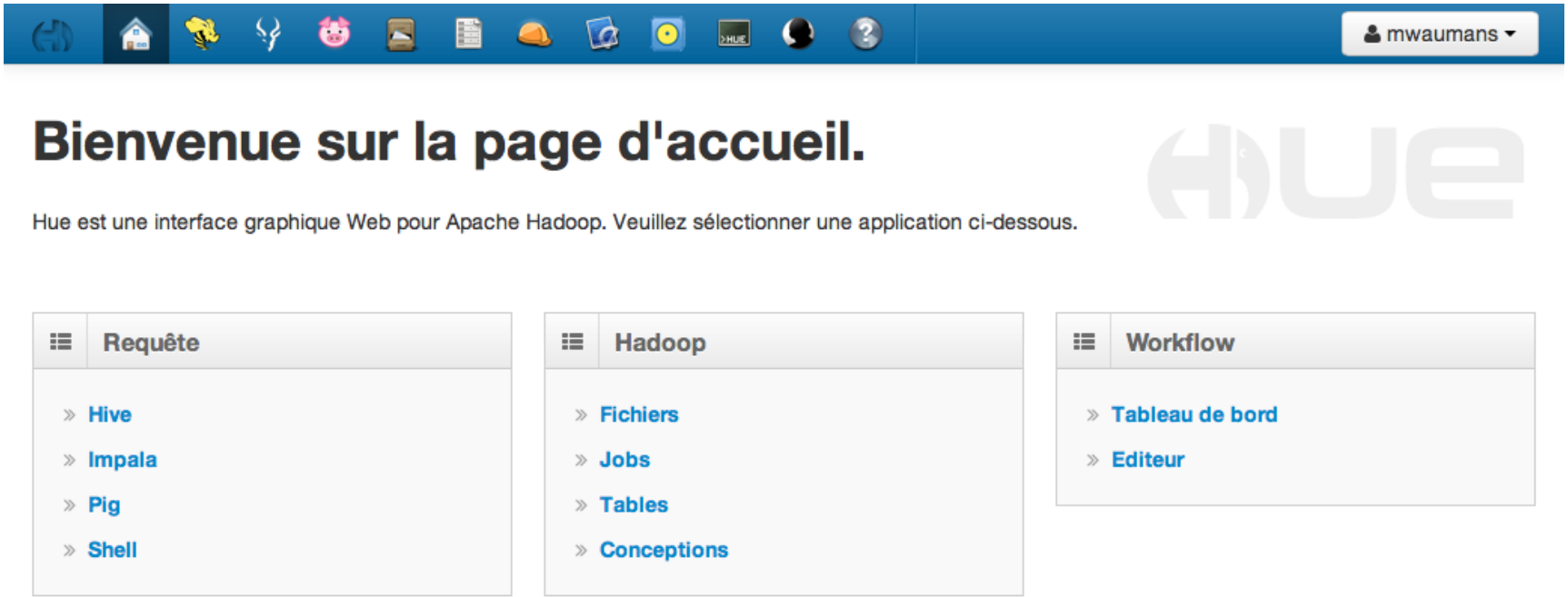
Hadoop Cluster : Cluster Architecture



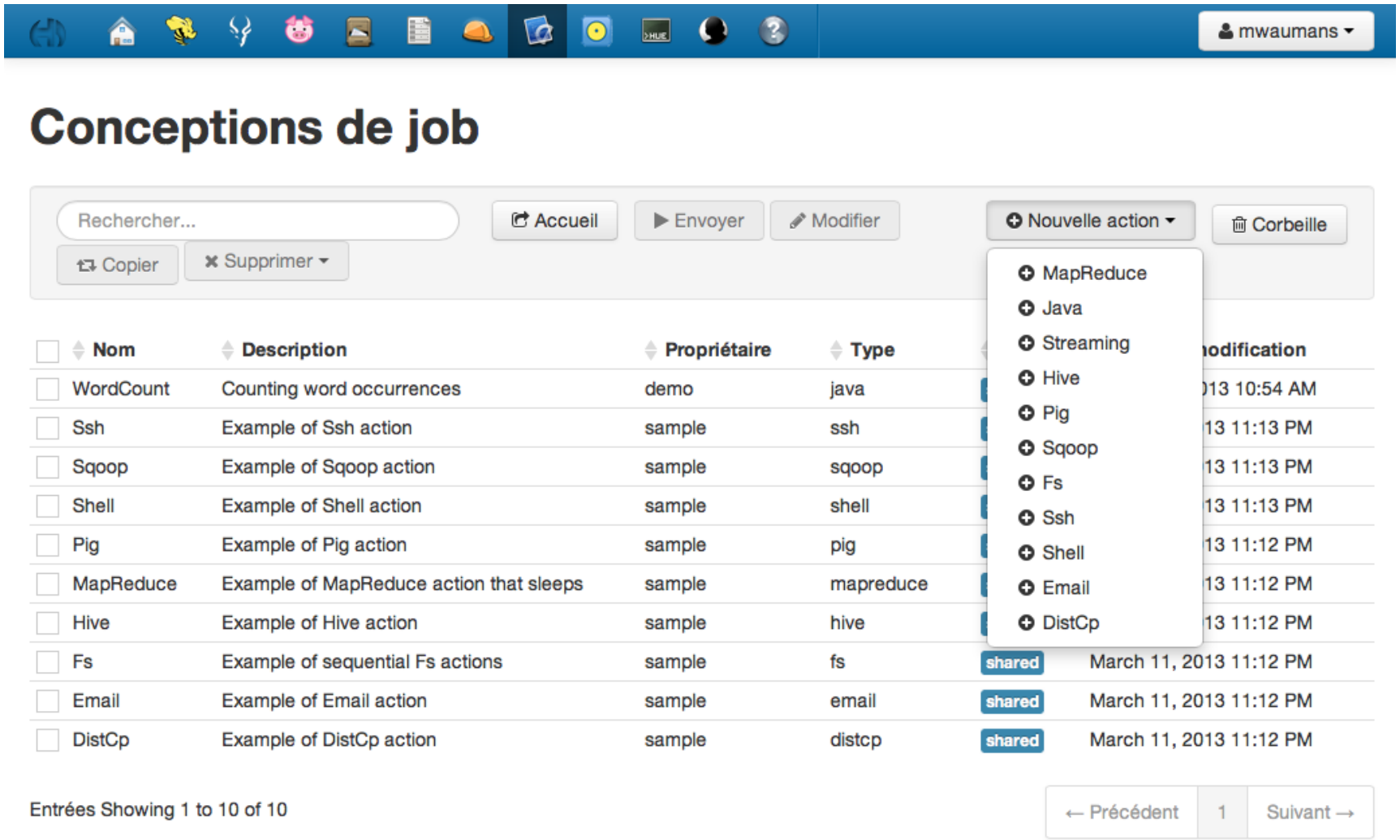
Hadoop Cluster : Access Through HUE

HUE

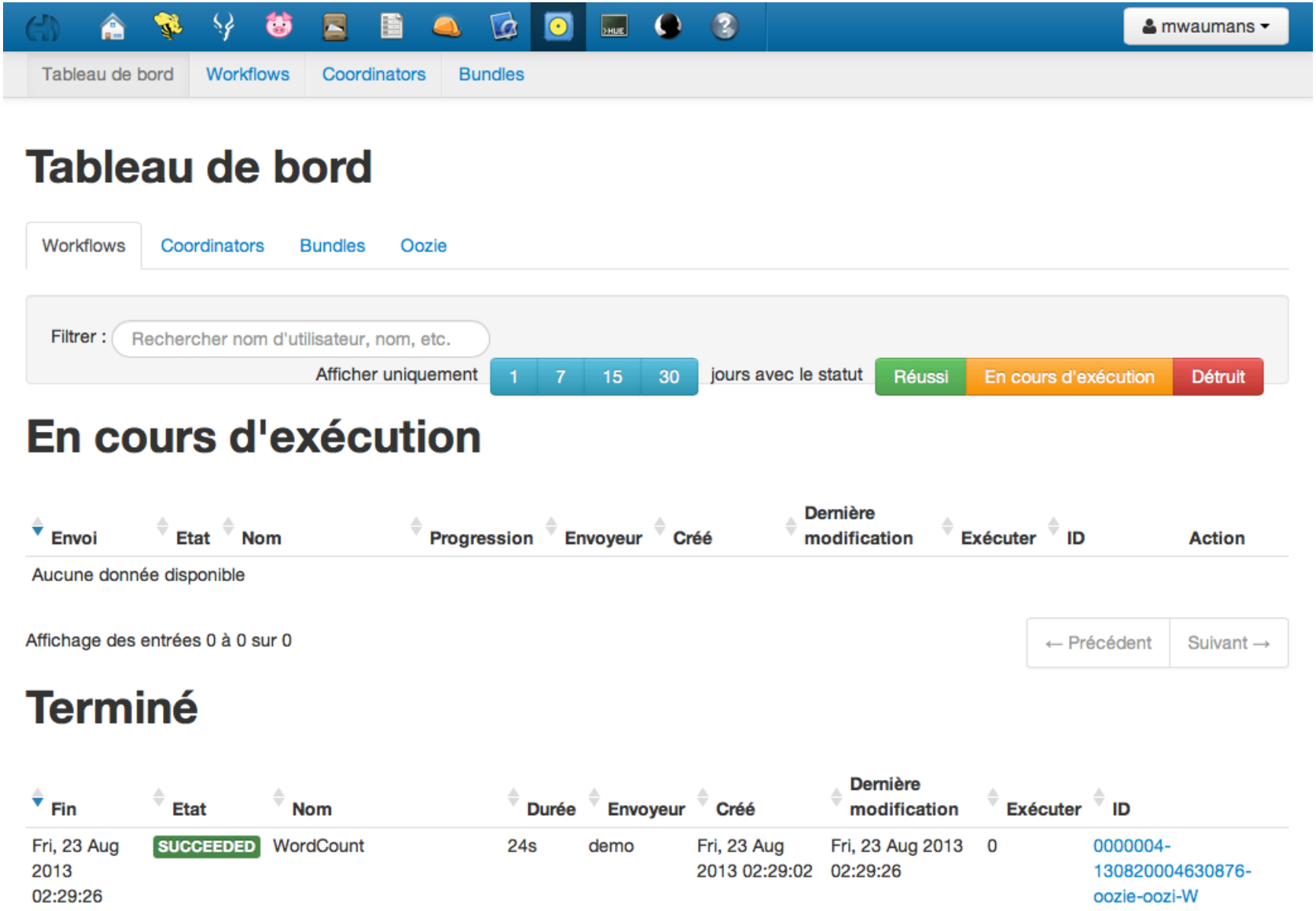
Open Source Apache Hadoop UI.
Features a file browser for HDFS, a Job Browser for Map Reduce, an HBase browser, query editors for Hive, Pig, Cloudera Impala and Sqoop,...



HUE Jobs


mwaumans
Conceptions de job
Rechercher...
Accueil
Envoyer
Modifier
Nouvelle action
Corbeille
Copier
Supprimer
MapReduce
Java
Streaming
Hive
Pig
Sqoop
Fs
Ssh
Shell
Email
DistCp
Nom
Description
Propriétaire
Type
WordCount Counting word occurrences demo java
Ssh Example of Ssh action sample ssh
Sqoop Example of Sqoop action sample sqoop
Shell Example of Shell action sample shell
Pig Example of Pig action sample pig
MapReduce Example of MapReduce action that sleeps sample mapreduce
Hive Example of Hive action sample hive
Fs Example of sequential Fs actions sample fs shared March 11, 2013 11:12 PM
Email Example of Email action sample email shared March 11, 2013 11:12 PM
DistCp Example of DistCp action sample distcp shared March 11, 2013 11:12 PM
Entrées Showing 1 to 10 of 10
← Précédent
1
Suivant →

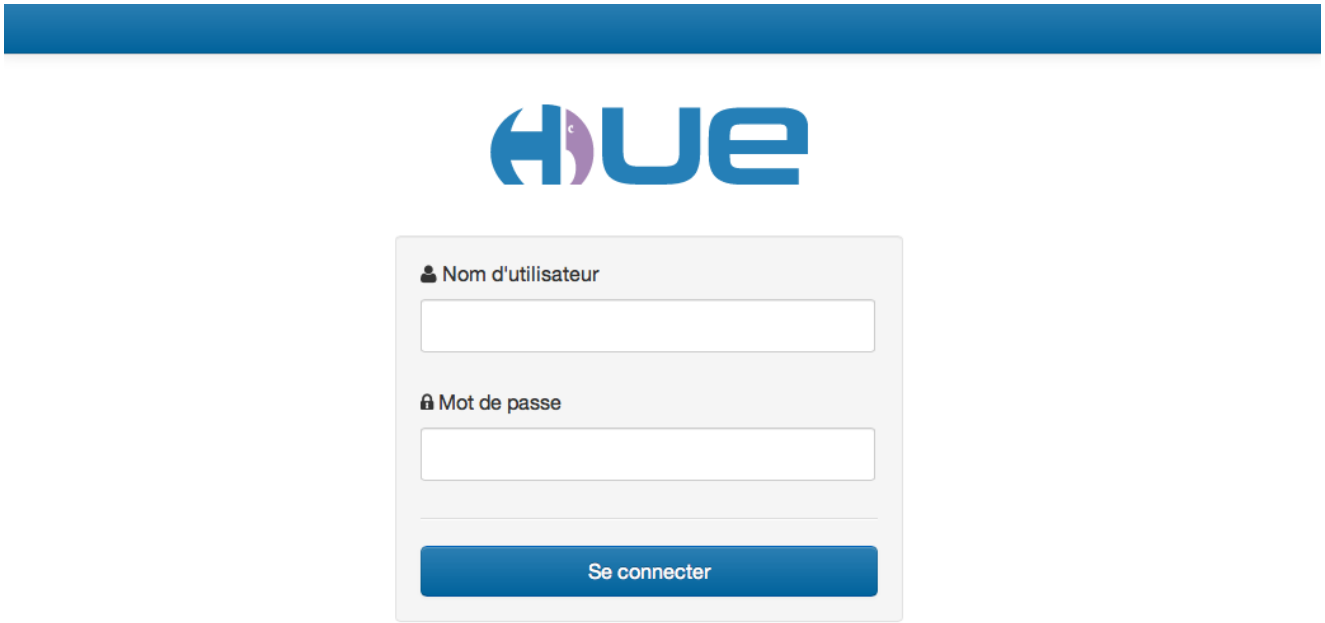
HUE Workflow



Try it !

URL

http://insilicodb.ulb.ac.be:8888



Access

Login : demo

Password : demo

Questions ?

Hassan Sayyadi, Lise Getoor (2008)

Future Rank : Ranking scientific articles by predicting their future pagerank
SIAM.

Berberish, Vazirgiannis, Weikum (2004)

T-Rank : time-aware authority ranking
Springer.

O'Madadhain, Smyth (2005)

EventRank : A framework for ranking time-varying networks
ACM.

S. Brin L.Page (1998)

The Anatomy of a Large-Scale Hypertextual Web Search Engine
Stanford.

Jon M. Kleinberg (1999)

Authoritative Sources in a Hyperlinked Environment
Journal of the ACM.

Barabasi Albert-Laszlo Reka Albert (1999)

Emergence of Scaling in random networks
Science.

ArXiV (2013)

ArXiV Cornell University Library
arxiv.org.

American Physical Society (2013)

APS Physics
aps.org.

SpringerLink (2013)

SpringerLink
link.springer.com.

PubMed (2013)

PubMed.gov US National Library of Medicine National Institutes of Health
http://www.ncbi.nlm.nih.gov/pubmed.