

Installation de l'environnement pour le TP

Configuration d'un reseau host-only via Virtualbox

1. Démarrez l'application **Virtualbox**
2. Verifiez/creez un réseau host-only ***vboxnet0*** ("File/Host network Manager" ...)

Network



Create



Remove



Properties

Name	IPv4 Address/Mask	IPv6 Address/Mask	DHCP Server
vboxnet0	192.168.56.1/24		<input checked="" type="checkbox"/> Enable

Close

Configuration d'un reseau host-only via VBoxManage

- Si vous avez rencontré des difficultes a l'etape precedente vous pouvez creer l'interface en ligne de commande:

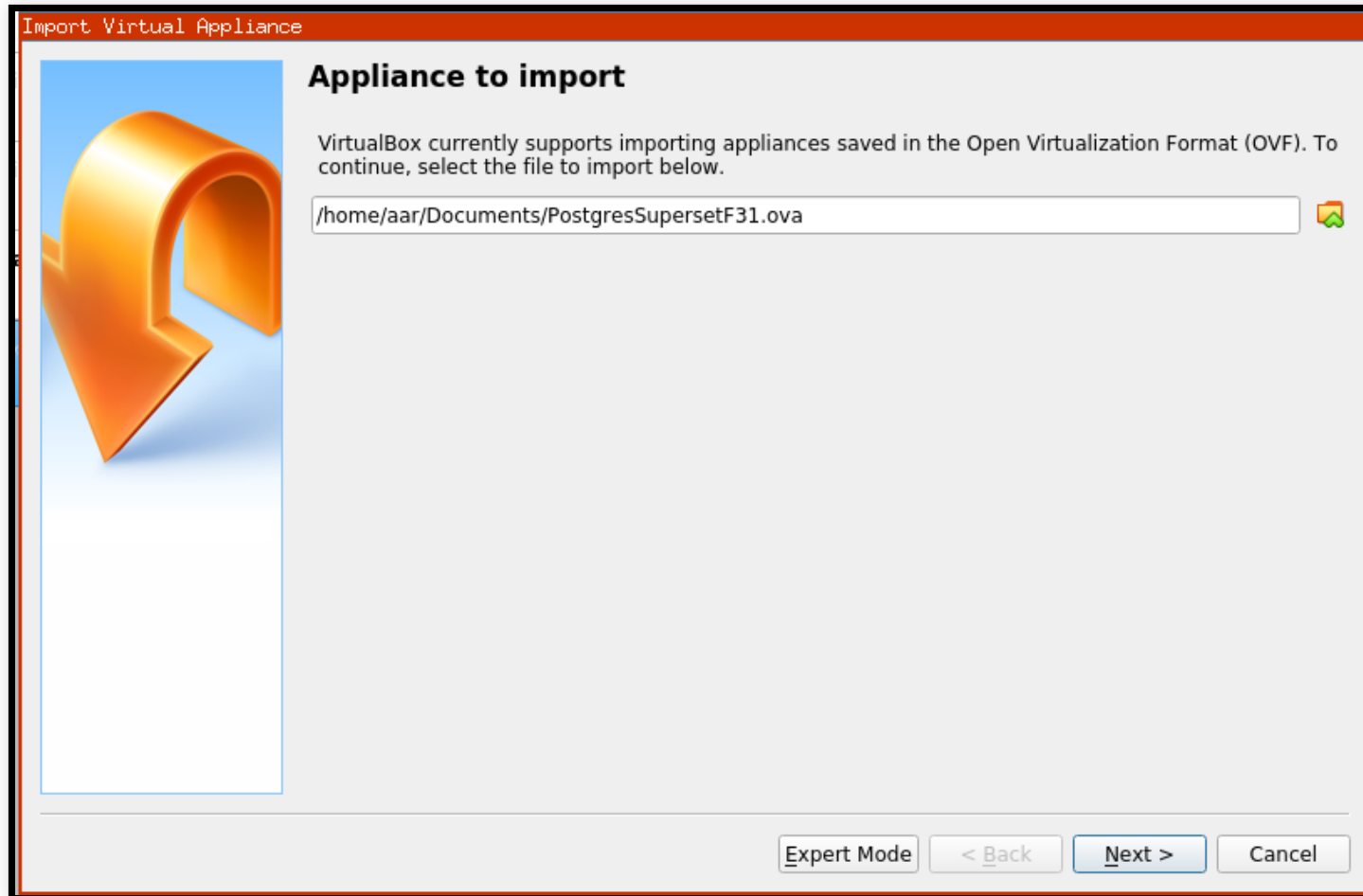
```
VBoxManage hostonlyif create
VBoxManage hostonlyif ipconfig vboxnet0 --ip 192.168.56.1
VBoxManage dhcpserver add --ifname vboxnet0 --ip 192.168.56.1\
    -netmask 255.255.255.0 --lowerip 192.168.56.100\
    --upperip 192.168.56.200
VBoxManage dhcpserver modify --ifname vboxnet0 --enable
```

Telecharger la VM pour le tp

Telecharger l'image Virtualbox: <http://bit.ly/telecom-postgresql-superset>

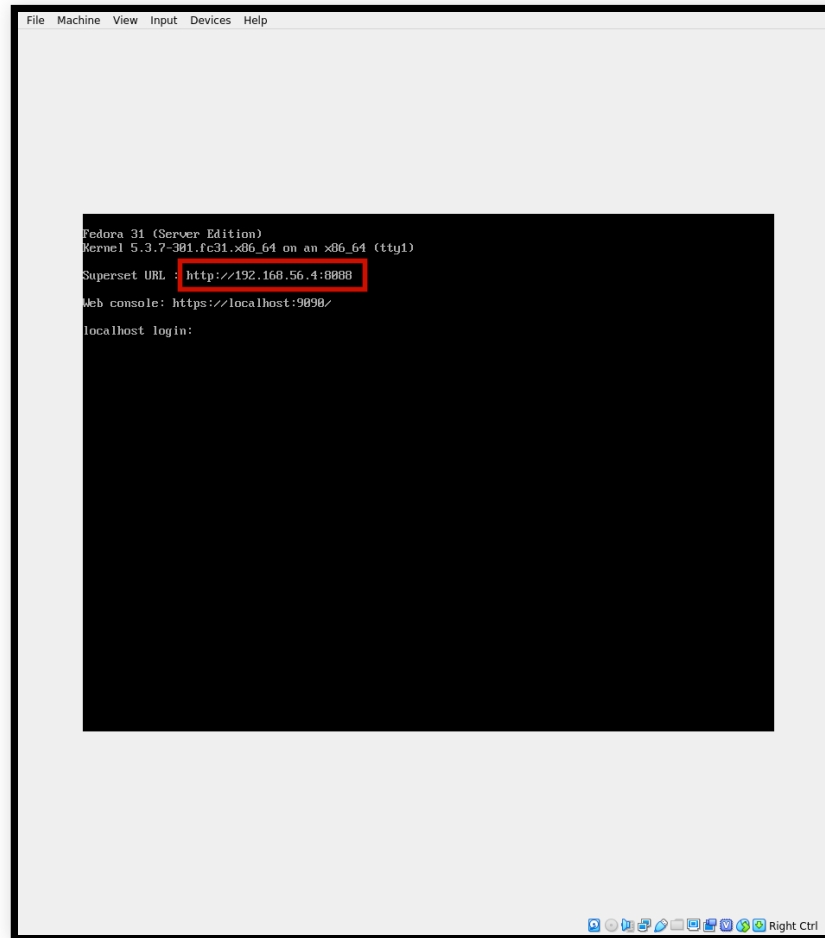
Importer la VM

File/Import appliance ...



Demarrer la VM

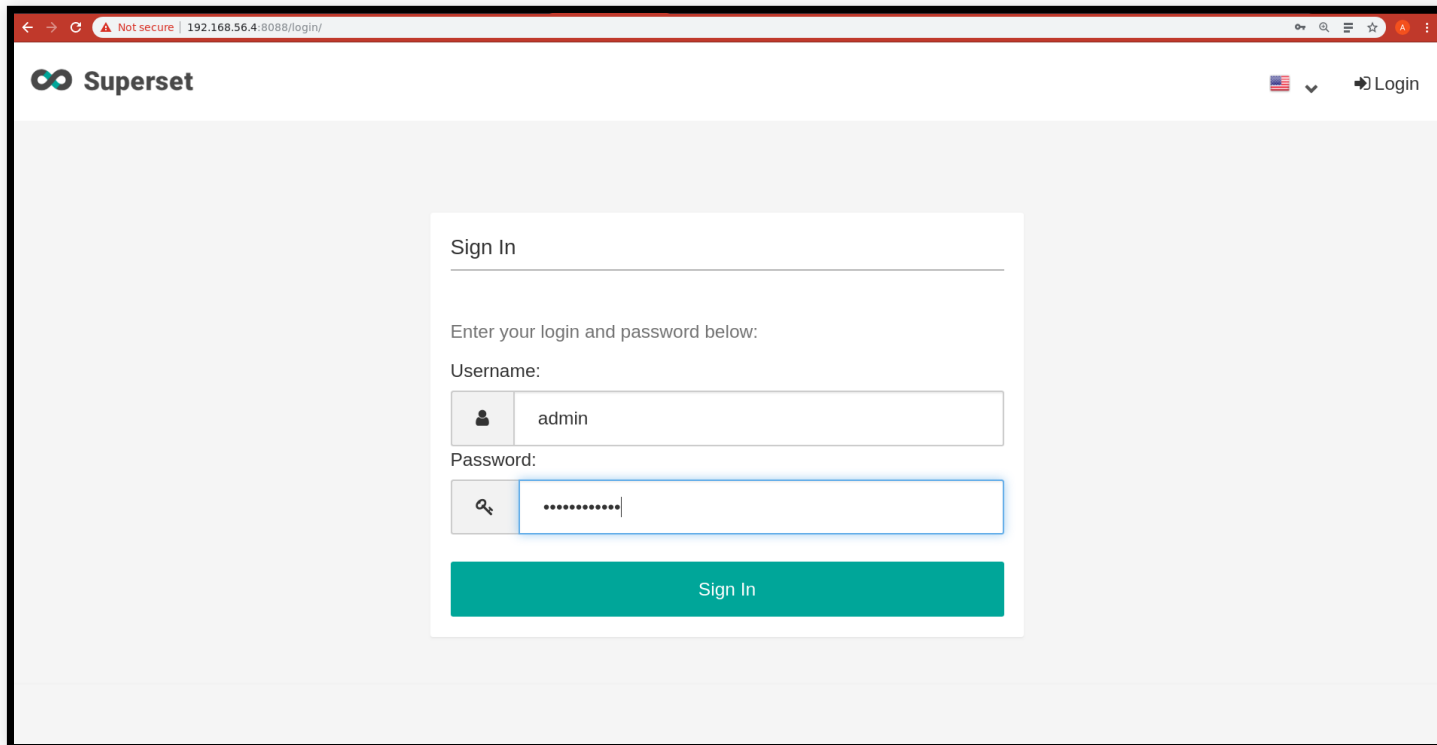
- via le button Run/Start et noter l'URL pour acceder a Superset



Si votre vm ne démarre pas ou vous n'avez pas de IP, vous pouvez essayer quelques [workarounds ici](#)

Connectez vous a Superset via votre navigateur

Username: admin Password: bigdatafuret



The screenshot shows a web browser window with the Superset login page. The browser's address bar shows the URL `192.168.56.4:8088/login/` and a "Not secure" warning. The Superset logo is in the top left, and a "Login" link with a flag icon is in the top right. The main content area features a "Sign In" form. The form has a title "Sign In", a subtitle "Enter your login and password below:", and two input fields. The "Username:" field contains the text "admin". The "Password:" field contains a series of dots, indicating a masked password. A teal "Sign In" button is at the bottom of the form.

Sign In

Enter your login and password below:

Username:

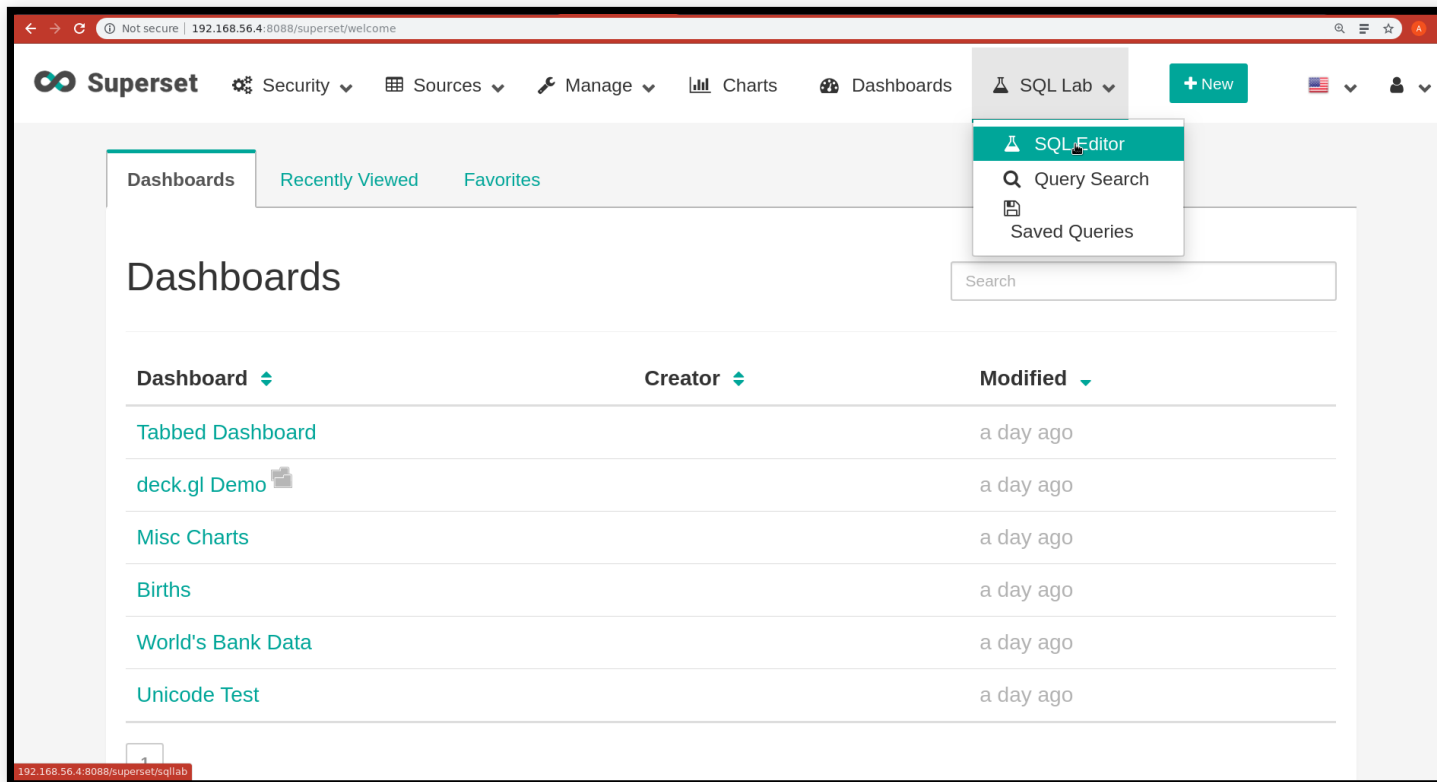
admin

Password:

.....

Sign In

Ouvrez le SQL Editor



Dans le SQLEditor lancez une requete

Tester la requete `_SELECT * FROM MOVIES_` (dans la Database Movies/ Schema: public.

Si vous avez eu des résultats, l'installation s'est bien passe, felicitations !

The screenshot shows the Superset SQL Editor interface. The top navigation bar includes the Superset logo, a 'Security' dropdown, and links for 'Sources', 'Manage', 'Charts', 'Dashboards', and 'SQL Lab'. A 'New' button is on the right. The main area is titled 'Untitled Query 2'. On the left, the 'Database' is set to 'postgresql' and the 'Schema' is 'public'. Below this, a section for 'See table schema (4 in public)' shows a dropdown for 'Select table or type table name'. The 'movies' table schema is displayed: 'movie_id' (INTEGER), 'title' (TEXT), and 'genre' (NullType). The SQL query editor contains the following code:

```
1 -- Note: Unless you save your query, these tabs will NOT persist if you clear your cookies or c
2
3 SELECT * FROM movies ;
```

Below the query editor are buttons for 'Run Query', 'Save Query', 'Share Query', 'new table name', 'CTAS', and 'parameters'. The 'Run Query' button is highlighted. The 'Results' tab is active, showing a preview of the 'movies' table data:

movie_id	title	genre
1	Star Wars	(0, 7, 0, 0, 0, 0, 0, 0, 0, 7, 0, 0, 0, 0, 10, 0, 0, ...)
2	Forrest Gump	(0, 0, 0, 5, 0, 0, 0, 7, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)

TP1: PostgreSQL Recherche et recommandation (1h)

- **Objectifs:**
 - prise en main de l'environnement de TP basé sur VirtualBOX
 - écrire des requêtes SQL
 - recherche: textuelle, approximative, phonétique
 - moteur de recommandation très basique

Recherche et recommandation

Moteur de recherche et recommandation des films:

- recherche: *textuelle, approximative, phonétique*
- recherche de type *graph*
- moteur de recommandation très basique

Schéma (déjà créé)

```
CREATE TABLE genres (  
    name text UNIQUE,  
    position integer  
);  
CREATE TABLE movies (  
    movie_id SERIAL PRIMARY KEY,  
    title text,  
    genre cube  
);  
CREATE TABLE actors (  
    actor_id SERIAL PRIMARY KEY,  
    name text  
);  
  
CREATE TABLE movies_actors (  
    movie_id integer REFERENCES movies NOT NULL,  
    actor_id integer REFERENCES actors NOT NULL,  
    UNIQUE (movie_id, actor_id)
```

```
CREATE INDEX movies_actors_movie_id ON movies_actors (movie_id);  
CREATE INDEX movies_actors_actor_id ON movies_actors (actor_id);  
CREATE INDEX movies_genres_cube ON movies USING gist (genre);
```

Create schema script Import data script

Recherche

- *Recherche exacte / pattern matching*
- *Distance de Levenshtein* → typos simples
- *N-gram/similarité* → trouver les erreurs modérées
- *Full text match @@* → similarité grammaticale
- *Métaphone* → similarité phonétique

Recherche textuelle/patterns

Utilisez les opérateurs **LIKE** ou **RegEX** pour les requêtes suivantes:

1. Tous les films qui ont le mot ***stardust*** dans leur nom.
2. Compter tous les films dont le titre ne commence pas par le mot ***the***
3. Tous les films qui ont le mot ***war*** dans le titre mais pas en dernière position

Distance Levenshtein

I	N	T	E	*	N	T	I	O	N
*	E	X	E	C	U	T	I	O	N
d	s	s		i	s				

- Opérations: **S**ubstitute, **I**nsert, **D**elete
- Distance *Levenshtein* : nb minimal d'opérations

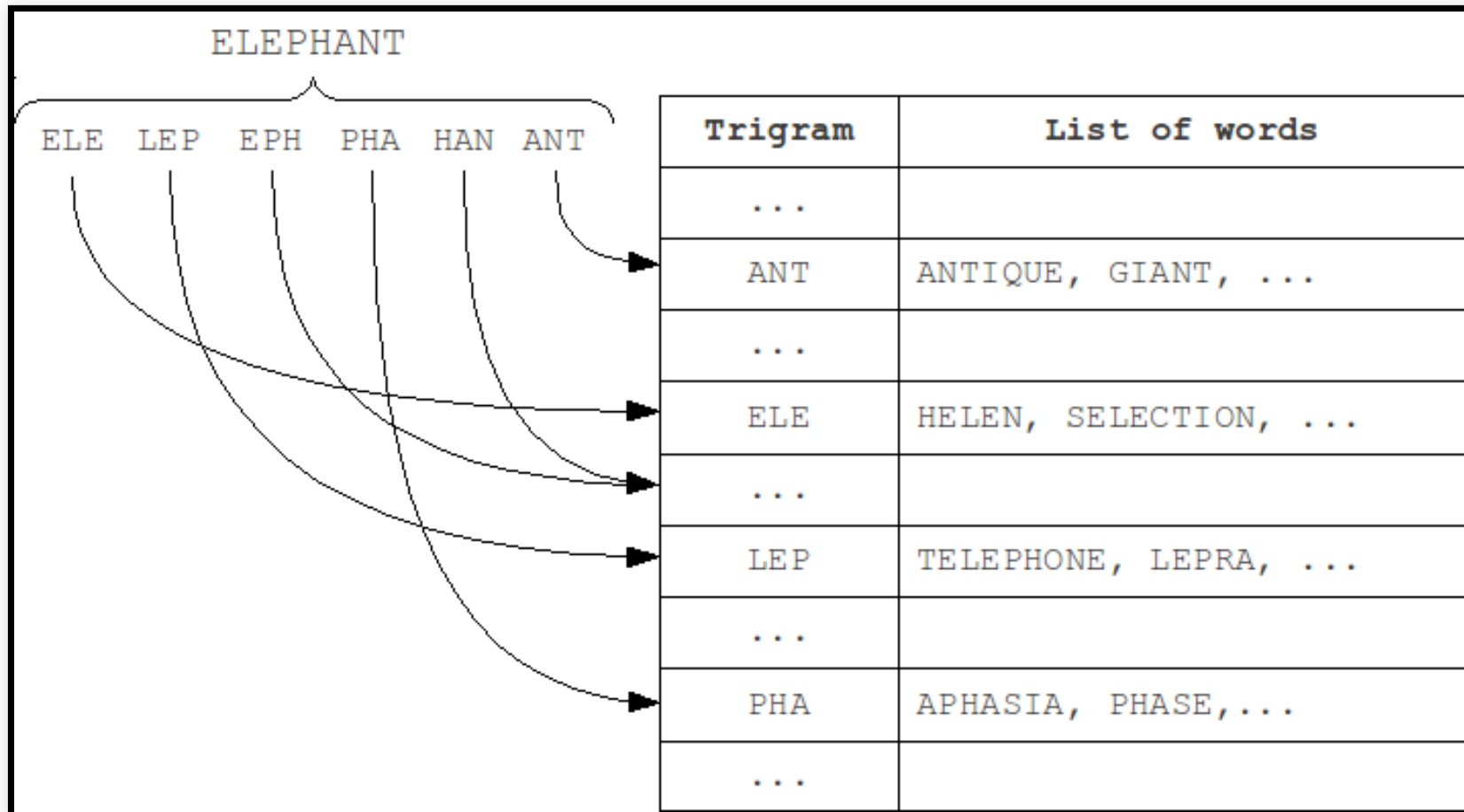
Distance Levenshtein

Utilisez les fonctions du package [fuzzystrgmatch](#) pour trouver :

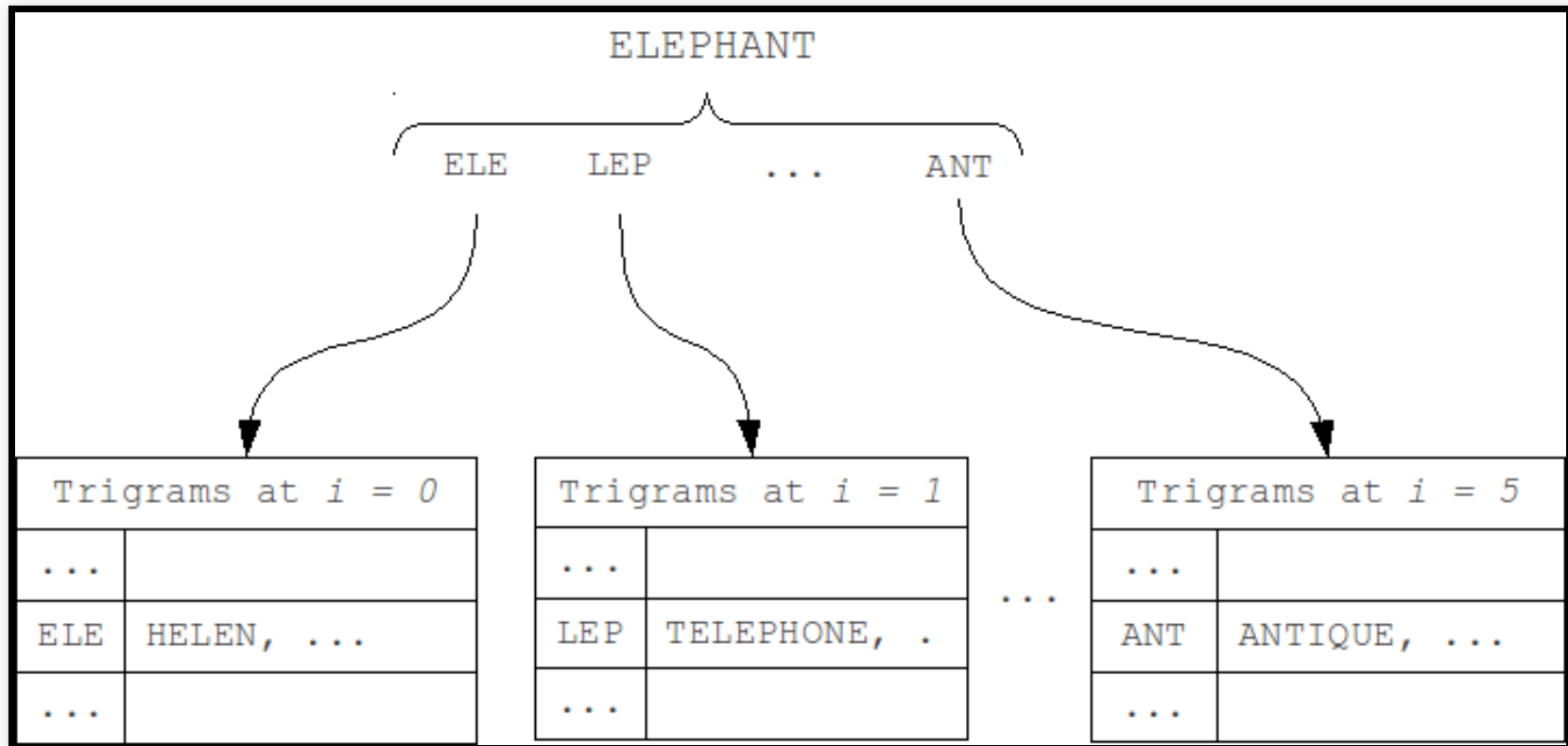
1. La distance levenshtein entre les mots ***execution*** et ***intention***
2. Tous les films qui sont a une distance *levenshtein* inférieure a 9 de la chaine suivante: ***a hard day nght***

[Documentation extension fuzzystrgmatch](#)

N-gram



N-gram, similarity search



N-gram, similarity search (%)

Écrivez les requêtes pour trouver :

1. Tous les tri-grammes du mot *Avatar*
2. La similarité entre **VOTKA** et **VODKA**
3. Tous les films dont le titre est similaire a plus de 0.1% du titre *Avatar* .

Documentation extension trgm

Full text search

Trouver les films qui contiennent les formes grammaticales des mots 'night' et 'day':

(ignorer les mots de liaison/ pluriel/etc..)

Algorithme:

1. ***extraire les racine des mots (lexèmes) → spécifiques au langage !***
2. comparer les vecteurs des lexèmes

Full text search

```
SELECT to_tsvector('A Hard Day's Night'),  
       to_tsquery('english', 'night & day');  
  
      to_tsvector      |      to_tsquery  
-----+-----  
'day':3 'hard':2 'night':5 | 'night' & 'day'
```

- ***tsvector*** : lexèmes :position
- ***tsquery*** : lexèmes séparées par &
- ***spécifique au langage !***

Documentation [recherche plein text ...](#)

Full text search

```
SELECT title
FROM movies
WHERE to_tsvector(title) @@ to_tsquery('english', 'night & day');
```

```
SELECT title
FROM movies
WHERE title @@ 'night & day';
```

```
A Hard Day's Night
Six Days Seven Nights
Long Day's Journey Into Night
```

Recherche phonétique

- plusieurs fonctions pour la codification phonétique des mots

```
SELECT name, dmetaphone(name), dmetaphone_alt(name),  
       metaphone(name, 8), soundex(name)  
FROM actors;
```

name	dmetaphone	dmetaphone_alt	metaphone	soundex
50 Cent	SNT	SNT	SNT	C530
Aaron Eckhart	ARNK	ARNK	ARNKHRT	A652
Agatha Hurlé	AK0R	AKTR	AK0HRL	A236

Documentation fuzzystmatch...

Recherche phonétique

1. Trouver les films qui ont des acteurs dont les noms se prononcent pareil.
2. Trouver les acteurs avec un nom similaire a ***Robin Williams***, triés par similarité (combiner %, metaphone et levenshtein):

actor_id		name
4093		Robin Williams
2442		John Williams
4479		Steven Williams
4090		Robin Shou

Search

- *Recherche exacte / pattern matching*
- *Distance de Levenstein* → typos simples
- *N-gram/similarite* → trouver les erreurs modérées
- *Full text match @@* → similarité grammaticale
- *Métaphone* → similarité phonétique

Recherche "graph"

- Trouvez le graph des acteurs connectees a Tom Hanks (ont deja joue dans un film avec l'acteur ou bien il y a un chemin films/acteurs qui mene a l'acteur)

Hint: vous pouvez utiliser les **Common Table Expressions**

title	name
Forrest Gump	Tom Hanks
The Green Mile	Tom Hanks
Apollo 13	Tom Hanks
Saving Private Ryan	Tom Hanks
Sleepless in Seattle	Tom Hanks
Toy Story	Tom Hanks
Toy Story 2	Tom Hanks
Big	Tom Hanks
Splash	Tom Hanks
Cast Away	Tom Hanks
You've Got Mail	Tom Hanks
The Bonfire of the Vanities	Tom Hanks
Philadelphia	Tom Hanks
Dragnet	Tom Hanks
The Money Pit	Tom Hanks
The Man with One Red Shoe	Tom Hanks
A League of Their Own	Tom Hanks
The 'Burbs	Tom Hanks
Bachelor Party	Tom Hanks
Sleeping Dogs	Tom Hanks
Forrest Gump	Robin Wright Penn
Forrest Gump	Gary Sinise
Forrest Gump	Mykelti Williamson
The Green Mile	Michael Clarke Duncan
The Green Mile	Bonnie Hunt
The Green Mile	David Morse
Apollo 13	Bill Paxton
Apollo 13	Kevin Bacon

Recherche multi-dimensionnelle

```
CREATE TABLE movies (  
    movie_id SERIAL PRIMARY KEY,  
    title text,  
    genre cube 1  
);  
  
INSERT INTO movies (movie_id,title,genre) VALUES  
(1, 'Star Wars',  
'(0,7,0,0,0,0,0,0,0,7,0,0,0,0,10,0,0,0)') 2  
),
```

- on utilise le type cube <1> pour mapper les notes sur un vecteur n-dimensionnel de valeurs (= score du film <2>)

Recherche multi-dimensionnelle

- les noms pour les dimensions sont définis dans la table genres

```
CREATE TABLE genres (  
    name text UNIQUE,  
    position integer  
);  
  
INSERT INTO genres (name,position) VALUES  
( 'Action',1),  
( 'Adventure',2),  
( 'Animation',3),  
...  
( 'Sport',16),  
( 'Thriller',17),  
( 'Western',18);
```


Recherche multi-dimensionnelle

Utiliser le module **cube** pour recommander des films *similaires*
(du même genre)

- Afficher les notes du film ***Star Wars***
- Quelle est la note du film ***Star Wars*** dans la catégorie 'Animation'
- Afficher les films avec les meilleurs notes dans la catégorie SciFi

Recherche multi-dimensionnelle

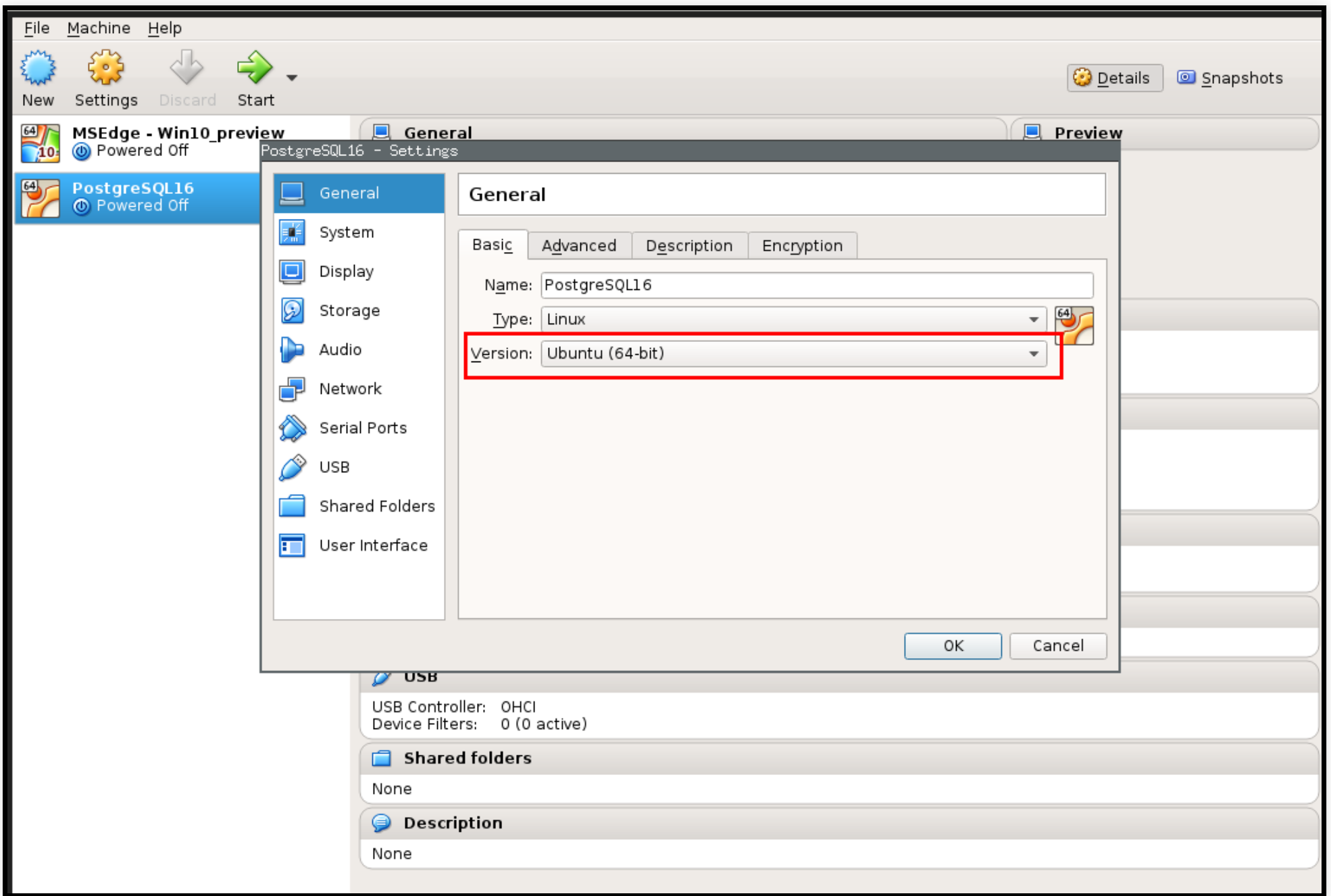
- Afficher les films similaires (**cube_distance**) a **Star Wars** (vecteur = (0, 7, 0, 0, 0, 0, 0, 0, 0, 7, 0, 0, 0, 0, 10, 0, 0, 0)) du plus similaire au moins similaire

title	dist
Star Wars	0
Star Wars: Episode V - The Empire Strikes Back	2
Avatar	5
Explorers	5.74456264653803
Krull	6.48074069840786
E.T. The Extra-Terrestrial	7.61577310586391

- Écrivez une requête pour trouver les films qui sont a moins de 5 points de différence sur chaque dimension (utiliser *cube_enlarge* et *@>*).

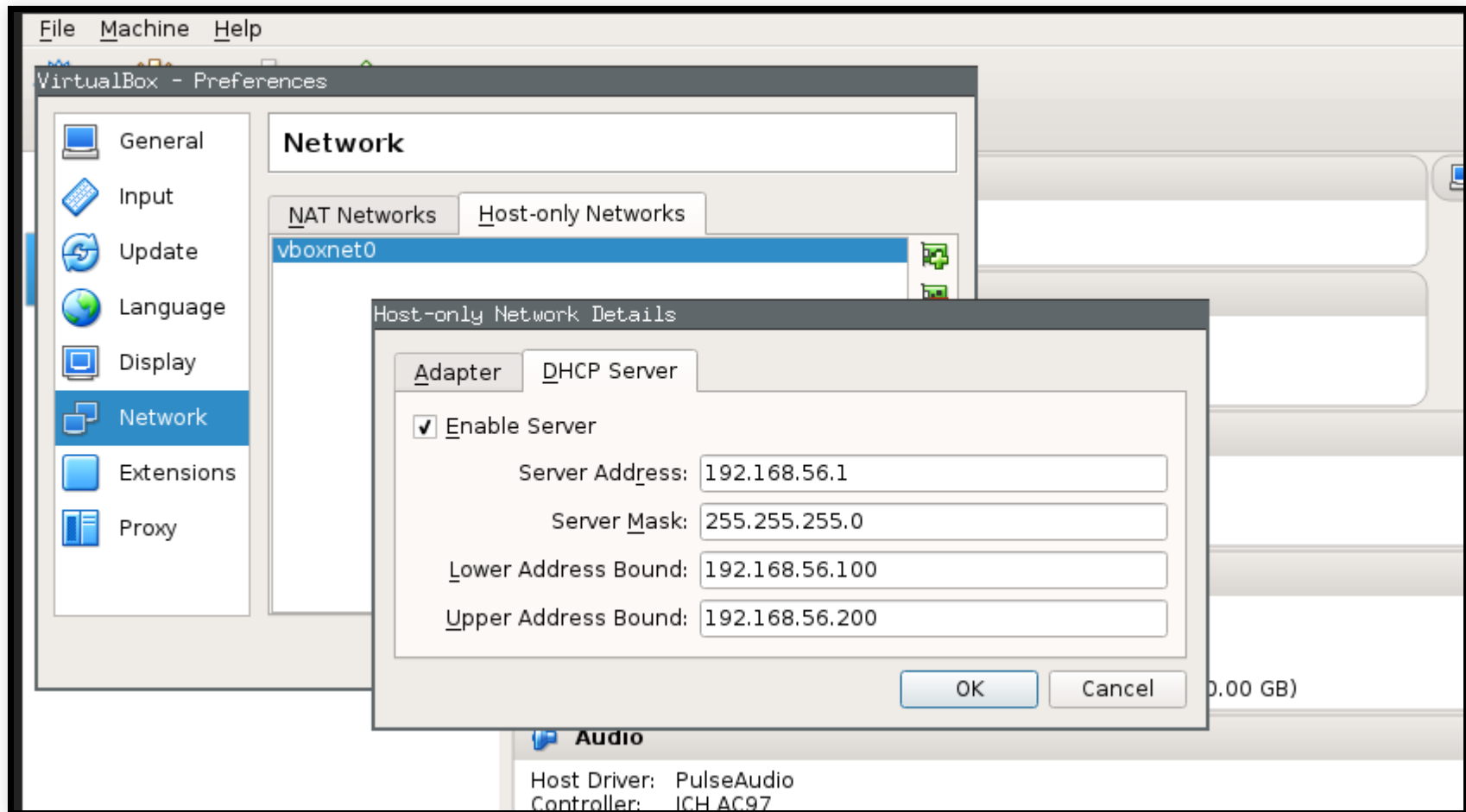
Troubleshoot VM freeze

- VM hangs at `*Loading initial ramdisk ... *`
- check VM type \Rightarrow Version Ubuntu(64 bit)



Troubleshoot no IP

- check DHCP settings



Troubleshoot no IP

- try to re-create the vboxnet0 interface via cmdline
 - poweroff VM
 - re-create the network interface:

```
VBoxManage hostonlyif create remove vboxnet0
VBoxManage hostonlyif create
VBoxManage hostonlyif ipconfig vboxnet0 --ip 192.168.56.1
VBoxManage dhcpserver add --ifname vboxnet0 --ip 192.168.56.1\
-netmask 255.255.255.0 --lowerip 192.168.56.100\
--upperip 192.168.56.200
VBoxManage dhcpserver modify --ifname vboxnet0 --enable
```

- restart VM

Installation en détail

Fedora Ubuntu

```
dnf install postgresql postgresql-server postgresql-contrib 1  
postgresql-setup initdb 2  
  
systemctl start postgresql.service 3  
  
yum install pgadmin3 4  
  
CREATE EXTENSION tablefunc; 5  
CREATE EXTENSION dict_xsyn;  
CREATE EXTENSION fuzzystrmatch;  
CREATE EXTENSION pg_trgm;  
CREATE EXTENSION cube;
```

1 Installation du client/serveur/extensions supplémentaires

2 Initialisation de la base

3 Démarrage du serveur

4 Front-end requetage

5

Installation des extensions **Verifier les extensions
installees**

Create index

```
CREATE INDEX [ nom ] ON table [ USING method ]  
  ( { colonne | ( expression ) } [ classeop ] ... )
```

- ***method***: btree/hash/gin/gist
- ***classeop***: operator class that can use the index

[Documentation](#)

Ressources:

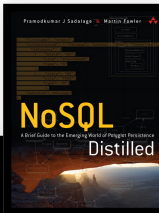
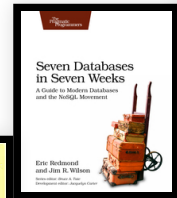
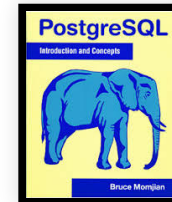
Bigdata - book by Nathan Marz book

NoSQL Distilled - book by Martin Fowler

7 databases in 7 days book

BigTable paper

MovieLens dataset



Ressources:

Why SQL is beating NoSQL, and what this means for the future of data

MapReduce: A major step backwards

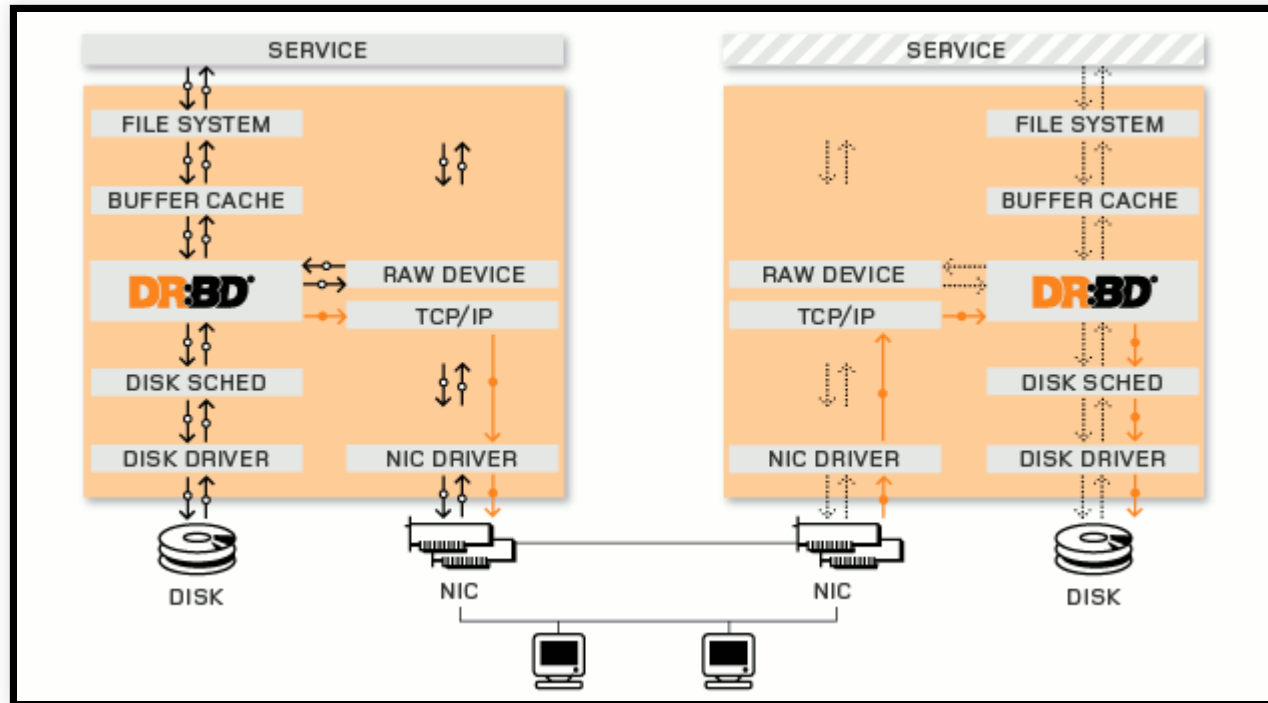
Why PostgreSQL

<http://momjian.us/main/presentations/Postgres> - Books and
ressources by Bruce Momjian

Other

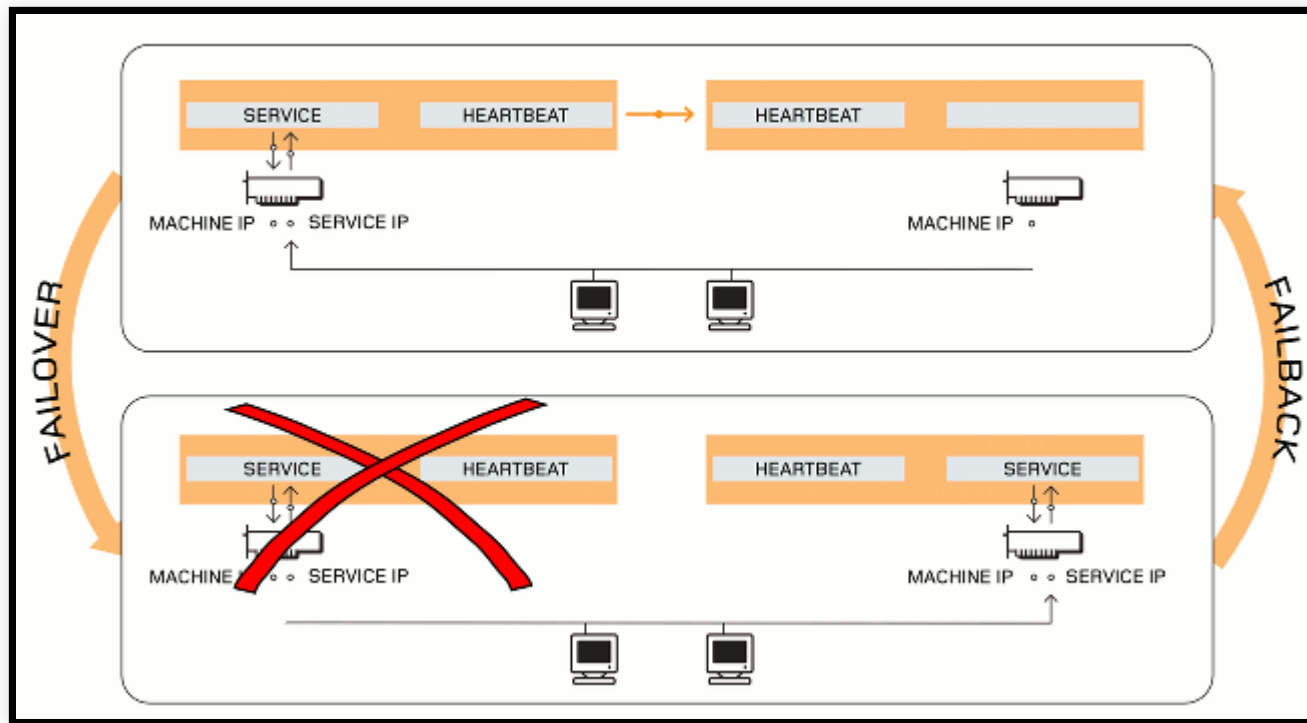
DRBD mirroring

- mirroring a linux partition over IP (sync/async)



DRBD HA

- heartbeat protocol monitors failures
- triggers service switch via IPFOs



more...

DRBD recovery

- node(s) outage
 - background sync (most up-to date node if both were down)
- replication network outage
 - automatic recovery
- storage subsystem
 - mostly transparent
- network partition
 - ***split brain!*** both nodes switched to the primary role while disconnected
 - ***Manual intervention needed***