Conception Avancée de Bases de Données

Tree Node

Selectivity

Attribute Cardinality

Traduction en cours

From Ullman

Parse Query

Query expression tree

Query Optimization

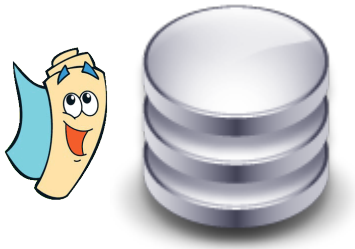
Select Logical Query plan

Logical Query Plan tree

Select physical Query plan

Physical Query Plan tree

From Ullman



Parse Query

Query expression tree

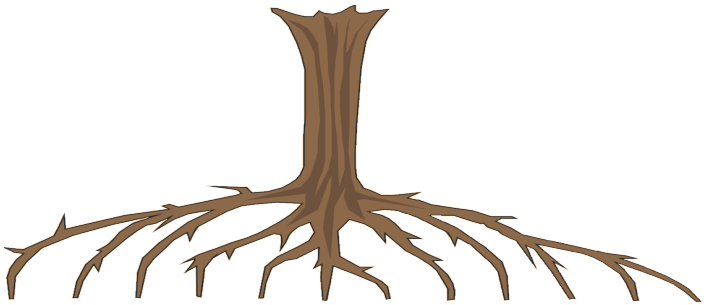
Select Logical Query plan

Logical Query Plan tree

Select physical Query plan

Physical Query Plan tree

Arbre Logique



Arbre Physique

Niveaux d'abstraction

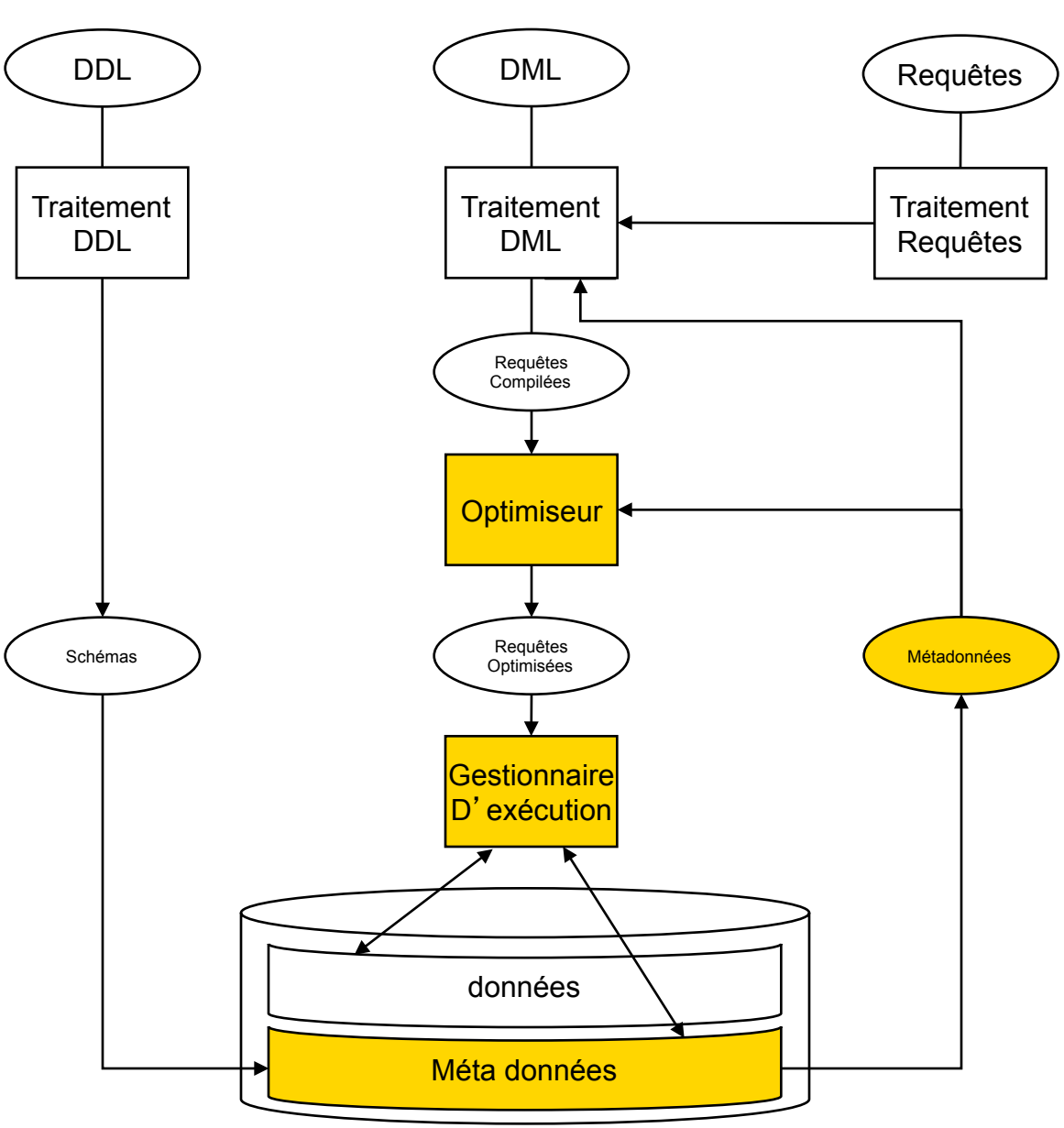
Modèle

Algèbre

σ **owner1=owner2 (Cats** $\otimes$ **Dogs) = Cat** $\bowtie$ **Dogs**

Logiciel

Java, C++,..



D' après C.J DATE

DDL : langage de définition des données; DML : langage de manipulation des données

planner/optimizer

- The task of the planner/optimizer is to create an optimal execution plan

- The planner/optimizer starts by generating plans for scanning each individual relation (table) used in the query.

http://www.postgresql.org/docs/9.1/static/planner-optimizer.html

Cost Based Optimization

- Optimiser adapts request plans as data characteristics change :
  - Selectivity
  - Cardinality
  - Frequencies
  - Max
- The cost of a request plan varies according to :
  - Cardinalities of intermediate joins and selections.
  - Selectivity of join predicates.

Optimisations

- Tous les arbres ont des avantages et des inconvénients.

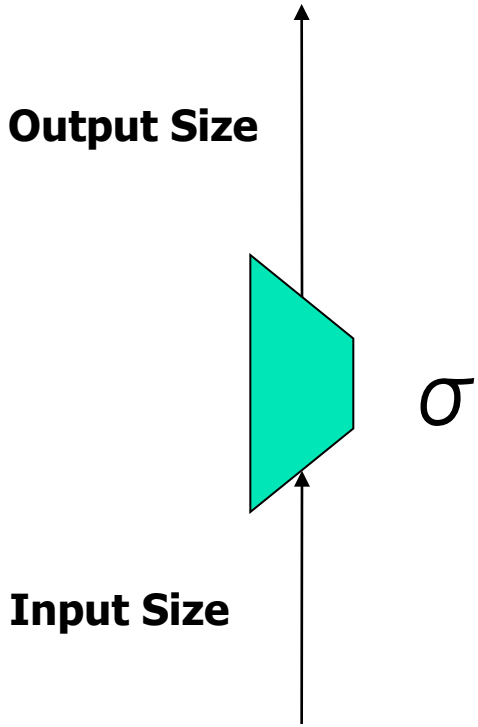
- Il faut choisir l'arbre en fonction des critères statistiques.

Selectivities

- Attribut Selectivity
- Predicate Selectivity
- Column Selectivity
- Index Selectivity

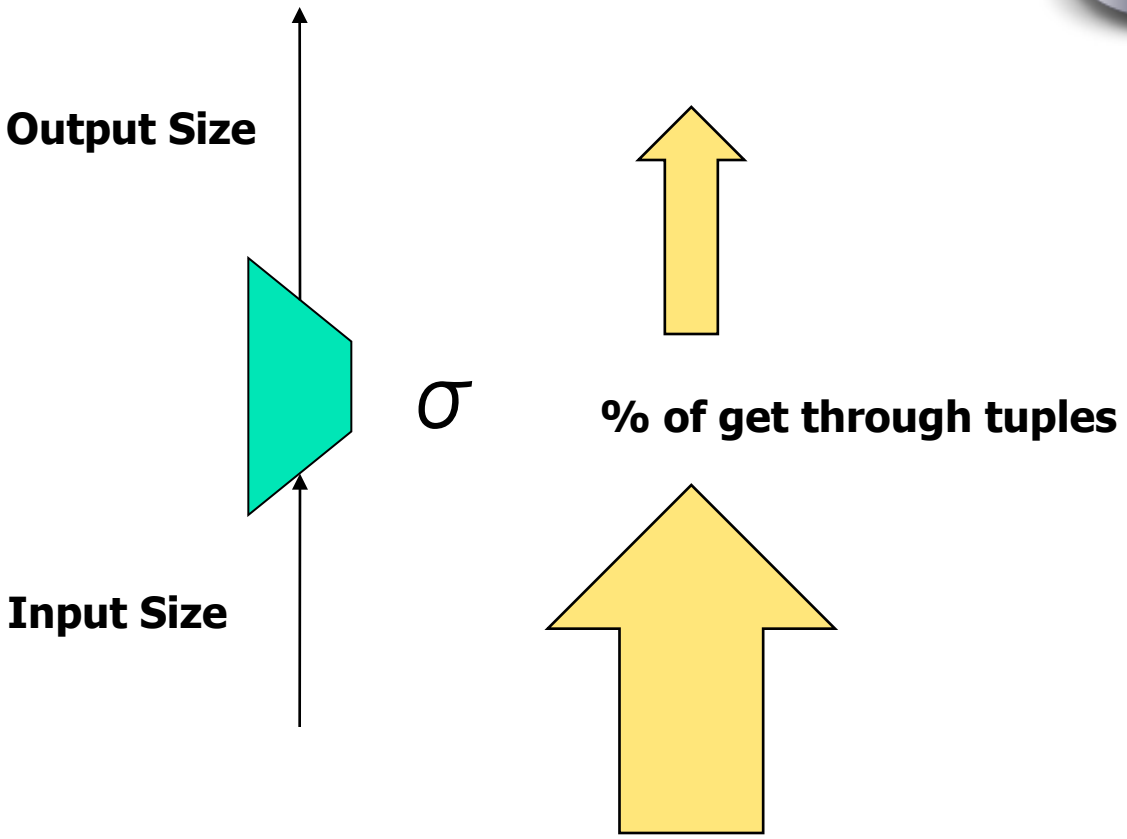
- Planed Selectivity
- Runtime Selectivity

selection



Output Size / Input Size === > 0

selection



Output Size / Input Size === > 0

Sélectivité d'un attribut A d'une table R

- Rapport entre le nombre de Tuples pour A = c et le nombre total de Tuples (T).

- Hypothèse (h1) : Dans le cas d'une répartition uniforme des valeurs de A.

- Ex1 : 1 seul tuples pour lequel A=C : 1/T
  - Attribut très sélectif
- Ex2 : A « true », « false » (h1) : ½
  - Attribut peu sélectif

Traduction en cours

Sélectivité d'un attribut A d'une table R

$$S = \frac{\text{Nombre de Tuples ayant un attribut de même valeur dans la table}}{\text{Nombre de Tuples dans la table}}$$

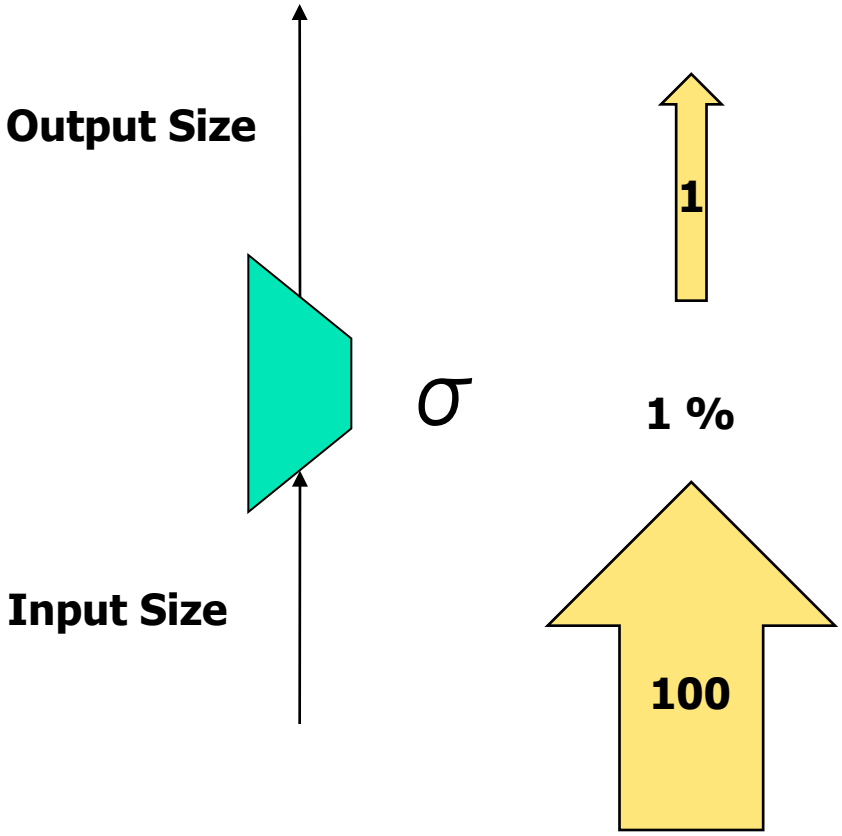
Traduction en cours

Sélectivité d'un attribut A d'une table R

$$S = \frac{\text{1 seul attribut de même valeur}}{\text{100 Tuples dans la table}} = \frac{1}{100}$$

Traduction en cours

selection



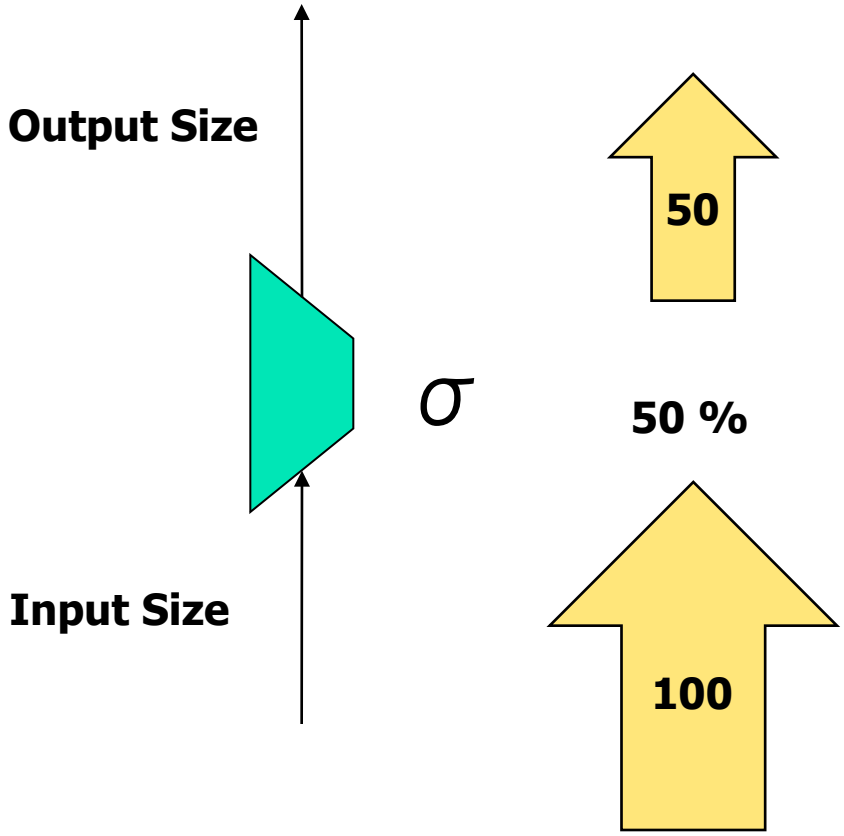
Sélectivité d'un attribut A d'une table R

$$S = \frac{\text{50 valeurs « True »}}{\text{100 Tuples dans la table}} = \frac{\text{50 valeurs « False »}}{\text{100 Tuples dans la table}} = \frac{1}{2}$$

Hypothèse (h1) : Dans le cas d'une répartition uniforme des valeurs de A = 50, !

Traduction en cours

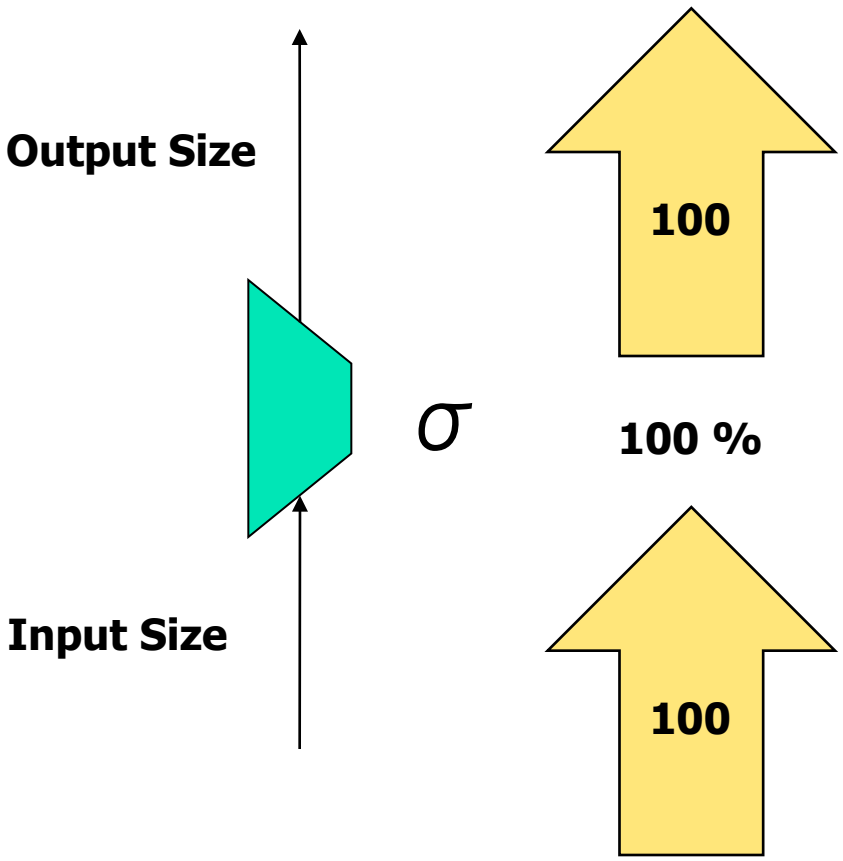
Selection



Sélectivité d'un attribut A d'une table R

$$S = \frac{\text{100 valeurs de A}}{\text{100 Tuples dans la table}} = 1$$

Selection



Column Selectivity

- A column that has a selectivity of 100%, then all the values in that column are unique.
- Column selectivity reveals how many different values are available in a given column.
- Low selectivity means there is no variation in the values contained in the column.

Index Selectivity

- Index with low selectivity mean that the index index is not efficient for the current request.
- Low selectivity index means no variation in data set.
- If index has a low selectivity then seq scan is more effcient than index.

Cardinality

- Cardinality is used to calculate selectivity

- The cardinality is the number of rows returned by each operation in an execution plan.

Data distribution uniformity

- Main statistics for selectivity estimation:
  - The number of rows contained in a table
  - The number of distinct values contained in a column

- But the selectivity computation is biased by data distribution uniformity.

Data Skewing

Big Data

- When to do data set are to big ?
- How to count attributs value ?