Report 1: Understanding the topic and orientation

7.

Název tématu	***Měření výkonu DB enginů***
Určení	*PVS*
Anotace	*Realizace sady experimentů pro porovnání výkonu DB s implementovaným rozhraním typu Postgres (Postgres, Yagabyte, Cocroach).*
Výstupy práce	*Výzkumná zpráva + software*
Požadavky na studenta	*Student alespoň 3. ročníku KB, případně student s nadprůměrnými znalostmi jazyku Python*
Navrhovatel	*Prof. Dr. Ing. Alexandr ŠTEFEK*

How I understand what I have to do:

Topic Title:	Measuring the **Performance** of **DB Engines**
Application:	PVS
Annotation	Implementation of a **set of experiments to compare** the **performance of DB** with the implemented interface of the Postgres type (Postgres, Yagabyte, Cocroach).
Outputs	Research Report + Software
Student Requirements	Student of at least the 3rd year of KB, or a student with above-average knowledge of Python
Applicant:	Prof. Dr. Ing. Alexandr ŠTEFEK

Key word:

- Database
- Measuring, Performance of database,**DB Engines**
- Postgres type (Postgres, Yagabyte, Cocroach).

○ First Questions:

1. What is a database? Their characteristics?

2. What is Performance of database?

1. Database

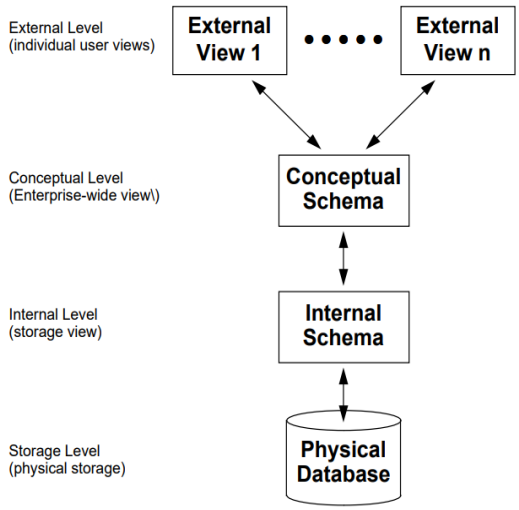
1.1: definition

General:
• A database is any **collection of related data.**
Restrictive:
• A database is a persistent, logically coherent collection of inherently meaningful data, relevant to some aspects of the real world.

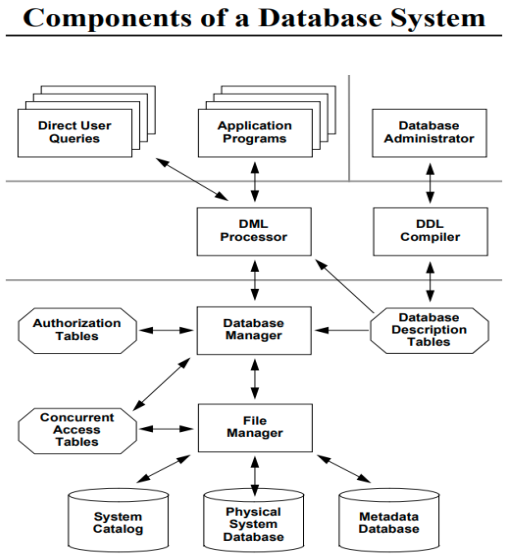
What is a Database Management System?

A database management system (**DBMS**) is a c**ollection of programs that enables users to create and maintain a database**. According to the ANSI/SPARC DBMS Report (1977), a DBMS should be envisioned as a multi-layered system:

Database management systems provide several functions in addition to simple file management:
• allow concurrency
• control security
• maintain data integrity
• provide for backup and recovery
• control redundancy
• allow data independence
• provide non-procedural query language
• perform automatic query optimization



Robert J. Robbins Johns Hopkins University



RJ-ROBBINS database Fundamentals

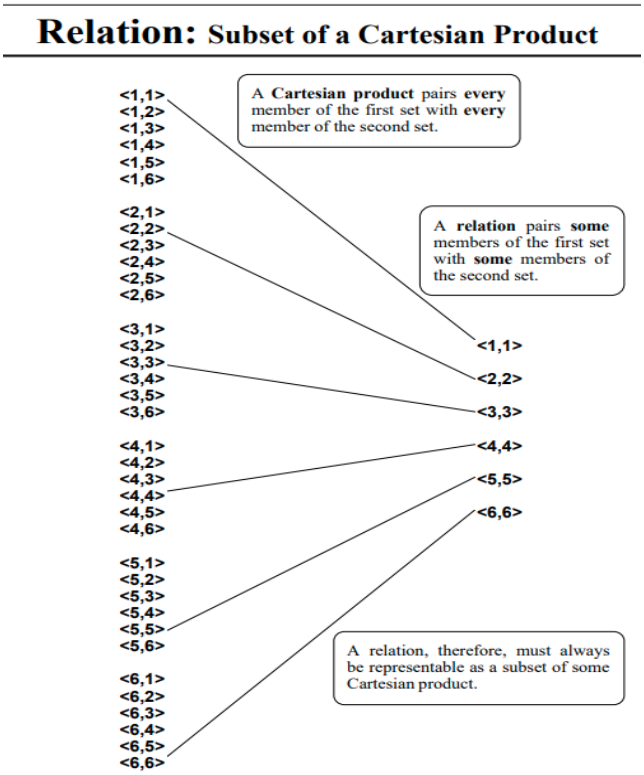
1.2: **Relations as a Database**

*"PostgreSQL is a powerful, o**pen source object-relational database system** that uses and extends the S**QL language** combined with many features that safely **store and scale the most complicated data workloads.**"*

From <https://www.postgresql.org/about/>

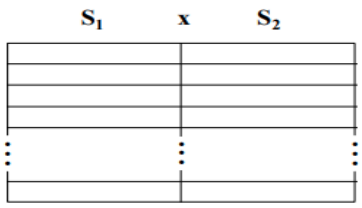
What is a relation?

- a kind of set
- a subset of a Cartesian product
- an unordered set of ordered tuples

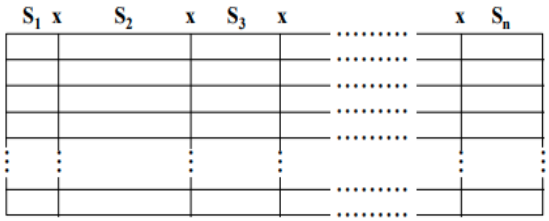


Relations as a Database?

A binary relation is a set of ordered doubles, with one element a member of the first set and one element a member of the second set. Generally, we could represent a set of ordered doubles as below. S1 is the first set and S2 the second.



By adding sets, relations can be extended to include ordered triples, ordered quadruples or, in general, any ordered n-tuple, as below. A relation with n participating sets is said to be of degree n or to possess arity n.



An n-ary relation (i.e., a subset of a Cartesian product of n sets) could be be represented in a computer system as an n-column tabular file, with one member from the first set named in the first column of each record and one member of the second set in the second column, etc

Codd recognized that many of the files used in computerized information systems were very similar in structured to tabularized relations.

Smith	Robert	L.	1154 Elm Street	Glendale	MD	21200
Smith	Judy	F.	1154 Elm Street	Glendale	MD	21200
Jones	Greg	G.	765 Cedar Lane	Towson	MD	21232
Harris	Lloyd	K.	2323 Maple Dr	Towson	MD	21232
⋮	⋮	⋮	⋮	⋮	⋮	⋮
Ziegler	Fred	K.	7272 Cherry Ln.	Baltimore	MD	21208

The business data file resembles a relation in a number of ways. The tabular file itself corresponds to a relation. Each column, or attribute, in the file corresponds to a particular set and all of the values from a particular column come from the same domain, or set. Each row, or record, in the file corresponds to a tuple

If such a file is to be genuinely interchangeable with a relation, certain contraints must be met:
• every tuple must be unique
• every attribute within a tuple must be single-valued
• in in all tuples, the values for the same attribute must come from the same domain or set
• no attributes should be null

```
13
14  select e.first_name, e.last_name, d.department_name
15  from employees e, departments d
16  where e.department_id = d.department_id;
17
```

Data Output | Messages | Notifications

	first_name character varying (20)	last_name character varying (25)	department_name character varying (30)
1	Steven	King	Executive
2	Neena	Kochhar	Executive
3	Lex	De Haan	Executive
4	Alexander	Hunold	IT
5	Bruce	Ernst	IT
6	David	Austin	IT
7	Valli	Pataballa	IT
8	Diana	Lorentz	IT
9	Nancy	Greenberg	Finance
10	Daniel	Faviet	Finance
11	John	Chen	Finance

- We will clarify the characteristics of Relational Database, E-R Data Model with specific examples in the second report.
    - Key
    - foreign key.
    - definition of normal forms.
    - What is the E-R Data Model?
    - OLTP (Online Transaction Processing, ACID....
- Examples can be taken from the databases that will build to serve in the process "Measuring Performance of database"

.....to be continued.....

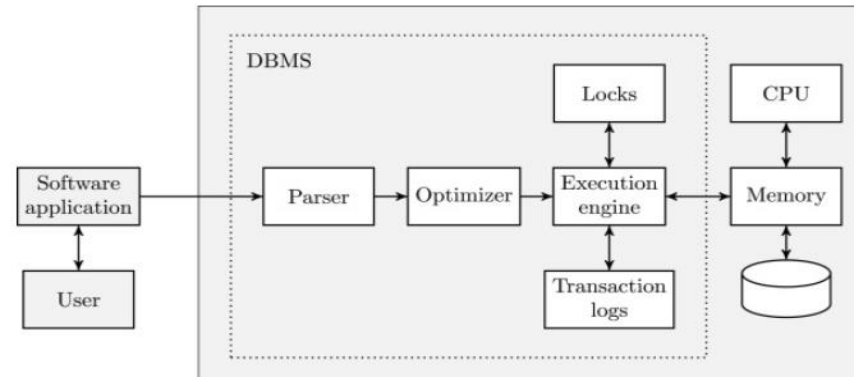
2. Measuring, Performance of database

2.1: Performance measurement

In general, performance is a measurement of how efficiently a software system completes its tasks. Performance is typically measured in: *response time, throughput, or in some cases, utilization of computing resources....*

Response time is the time taken for a call in the system to traverse to some other part of the system and back. This is also sometimes called **latency**, and in the context of database systems, the response time may be measured as the response time to the first or the last result item.

The response time might be the time taken after the end-user sends a request to the software application (e.g., an online store), which passes the request to a DBMS, which returns a set of data to the software application, which finally presents the data to the end-user's device.



In database **benchmarking**, however, response time might be measured by running the benchmark on the same device the DBMS and the database reside, effectively eliminating inter device-induced performance drawbacks such as network latency +and firewalls, and mitigating the effects of other software running on the devices.

*Comment: "Transaction processing environments often process a large number of concurrent transactions, response time alone might not reliably account for the effects of concurrent transactions, unless **response time is measured as an average of multiple concurrent transactions**". Question: Within the scope of the research article: how is the processing of large numbers of concurrent transactions limited? How will the average response time be calculated?*

Performance can also be measured by **throughput, i.e., how many transactions the DBMS can execute in a given time frame**. T**hroughput is often expressed as transactions per second and requires a more sophisticated approach, e.g., benchmarking software**. Again, throughput may be measured either locally (i.e., using only the hardware the DBMS and the database reside on), or over a network in case the database is distributed).

Finally, performance may be measured by **resource utilization**, either CPU time, I/O, memory allocation, or energy consumption in systems striving for energy-efficiency due to, e.g., limited battery power, or due to environmental concerns.

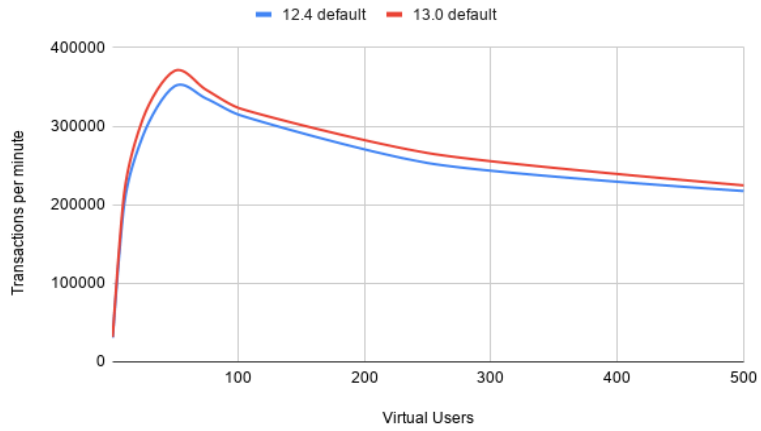
Question:

- *How to measure Response time and throughput? (local? benchmarking software?)*
- *Within the scope of the research, what other parameters need to be paid attention to (CPU time, I/O, memory allocation, energy consumption....)? How to measure those parameters?*

2.2 Database performance benchmarks

There are several database performance benchmarks available, each typically consisting of a sample database and a workload that **simulates how the database could be used**.
"The benchmarks usually measure the efficiency of querying while taking into account factors such as concurrency but disregarding other DBMS tasks such as efficiency in data structure definition or bulk loading."

In the domain of relational databases, the Transaction Processing Council (TPC) benchmarks are perhaps the most utilized, and **test the throughput of the DBMS with various parameters**.
For example, the TPC-A benchmark simulates a database of a bank with four tables and with one transaction, the TPC-B benchmark a database of a wholesale supplier with nine tables and with five transactions, and the TPC-E benchmark a brokerage database with 33 tables and 12 transactions. All these benchmarks have the option for simulating strong consistency, and while TPC-A and TPC-B have transactions typical for transaction processing, TPC-E includes also decision support transactions



eg: [PostgreSQL TPC-C Benchmarks: PostgreSQL 12 vs. PostgreSQL 13 Performance (enterprisedb.com)]

In the more general DBMS domain, the Yahoo! Cloud Serving Benchmark (YCSB) is a framework for benchmarking transaction processing in systems with different data models and architectures (Cooper et al., 2010). . Other benchmarks : LUBM, OLTP-Bench , JOB...

Regardless of the data model and DBMS, transaction processing benchmarks have typically been the de facto method of comparing different DBMSs and hardware

Question:

- *Within the scope of the research: should we use TPC benchmarks? And how to apply it?*

3. IDEA about TESTING AND EVALUATION the RESULTS

These are first-ideas of the experiment. These ideas may change when i have more ideas become about performance measurement methods or about the characteristics of specific DBSMs (Postgres, Yagabyte, Cocroach).

The benchmark itself consists of following steps and procedures:
● generates a set of alerts; (Alert sets were generated using vulerability scanners and tools).
● connects to databases; (modul?? How to create? How to use?)
● deal with data: creates tables; inserts data; perform SELECT operations; deletes all data from the table;

3.1 Building a database: need to learn more about the other 3 database management systems to build a suitable database

- 1
- 2
- 3

3.2 Configure the testing environment: Hardware and software characteristics
Sets specific requirements for hardware configuration, software installation and work characteristics to ensure fairness and comparability between different database systems.
tpc-c_v5.11.0.pdf

For each DBSM, the environment installation is carried out according to the process:
format the hard drive, install the operating system, install the necessary systems/modules, restore the database including the generated data, perform the necessary configurations. Thus, we have eliminated as many other factors as possible that could affect the measurement results
.....
Question:

- *How to set up the best environment to be able to compare the performance of DBSMs?? What does that environment include?? (new computers to use as servers and clients, their configuration parameters...)*

4. My ideas and Questions:

Within the scope of the research:

- *Should we use TPC benchmarks? And how to apply it?*
- *How to measure Response time and throughput? (local? benchmarking software?)*
- *Within the scope of the research article: how is the processing of large numbers of concurrent transactions limited? How will the average response time be calculated?*
- *What other parameters need to be paid attention to (CPU time, I/O, memory allocation, energy consumption....)? How to measure those parameters?*
- *How to set up the best environment to be able to compare the performance of DBSMs?? What does that environment include?? (new computers to use as servers and clients, their configuration parameters...)*

Report1 ' s questions:

1. What do you think about my idea? (Going in the right direction, not enough, too big for me....)

2. Can you guide me furthe (What should I do to improve the idea? Or I need to choose the right idea?)

3. ***My difficulty:***
   My research method is to always ask a question: "How to .." and learn how to solve it on the Internet. So I think sometimes I still don't have enough basic knowledge background,
   Can you recommand me some references, documents, ebook., your github learning repo....To improve my basic background?

Thank you very much in advance!
Předem děkuju Vám moc !
Spozdravem,
Dang Quy Tai

Report 2:

- Solve some problems of report 1
- Learn about Postgres, Yagabyte, Cocroach