POLITECNICO
MILANO 1863

SYSTEMS AND METHODS FOR BIG AND UNSTRUCTURED DATA

Data Architecture Concepts

Marco Brambilla

marco.brambilla@polimi.it

@marcobrambi

Schema

The data schema

- Typing
- Coherence
- Uniformity

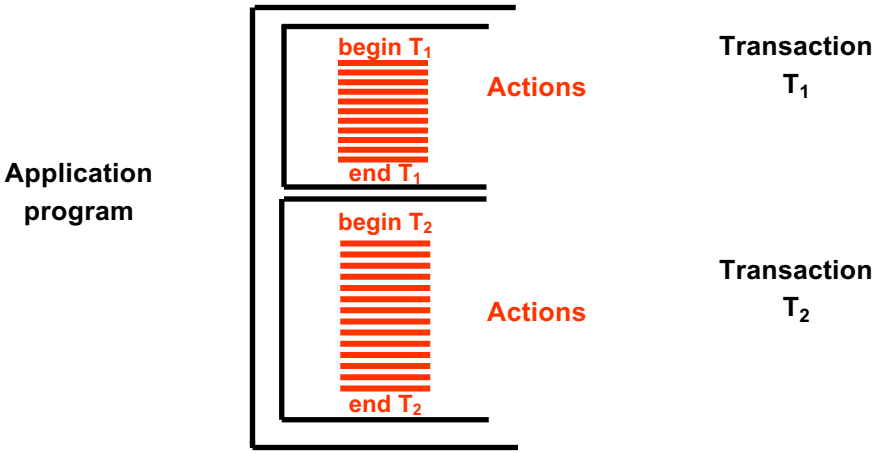
Transactions

- The relational world
- Multi-user
- Distributed systems

Definition of Transaction

- An elementary unit of work performed by an application
- Each transaction is encapsulated within two commands:
  - **`begin transaction`** (bot) and **`end transaction`** (eot)
- Within a transaction one of the commands below is executed (exactly once):
  - **`commit work`** (commit) and **`rollback work`** (abort)
- Transactional System (OLTP): a system capable of providing the definition and execution of transactions on behalf of multiple, concurrent applications
  - As opposed to OLAP

Application and Transaction



Transaction: Example

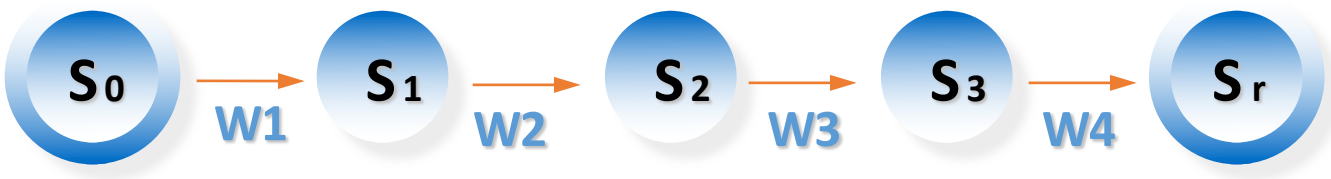
```
begin transaction;

update Account
  set Balance = Balance + 10 where AccNum
  = 12202;

update Account
  set Balance = Balance - 10 where AccNum
  = 42177;
commit work;

end transaction;
```

Transaction: Example with Alternative

```
begin transaction;
update Account
  set Balance = Balance + 10 where AccNum =
  12202;
update Account
  set Balance = Balance - 10 where AccNum =
  42177;
select Balance into A from Account
  where AccNum = 42177;
if (A>=0) then commit work
          else rollback work;
end transaction;
```

Well-formed Transactions

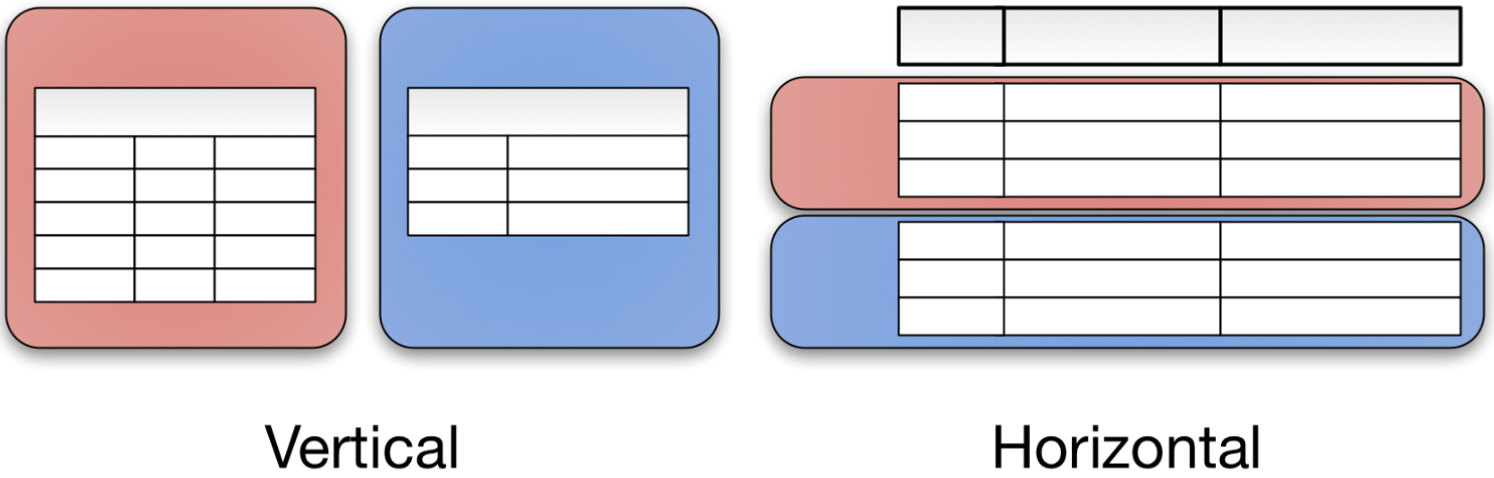
- **`begin transaction`**
- code for data manipulation (reads and writes)
- **`commit work`** – **`rollback work`**
- no data manipulation
- **`end transaction`**



Partitioning and Replication

Horiz. Vs. Vert. Partitioning

- The distinction of **horizontal** vs **vertical** comes from the traditional tabular view of a database.



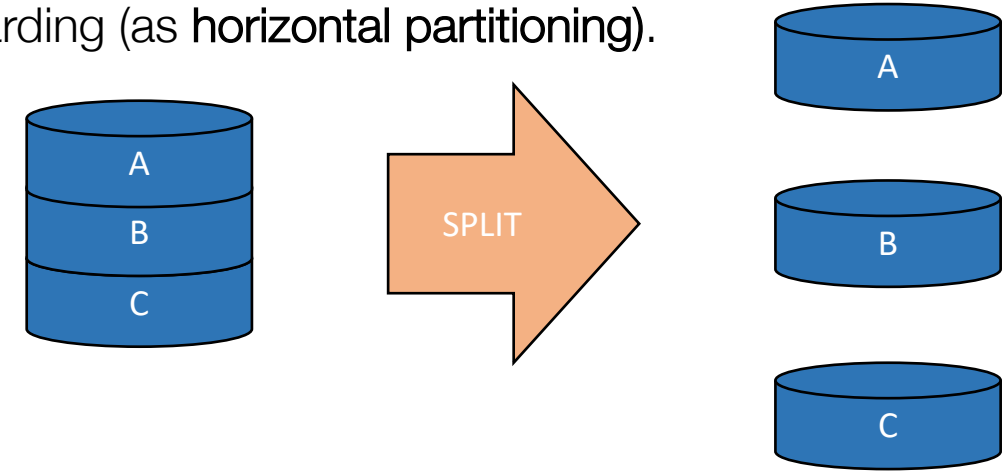
Data Partitioning

Aim: scalability, distribution.

Partitioning splits the data in the database and partitions pieces of it to different storage nodes.

Databases can be sharded horizontally (by rows) or vertically (by columns).

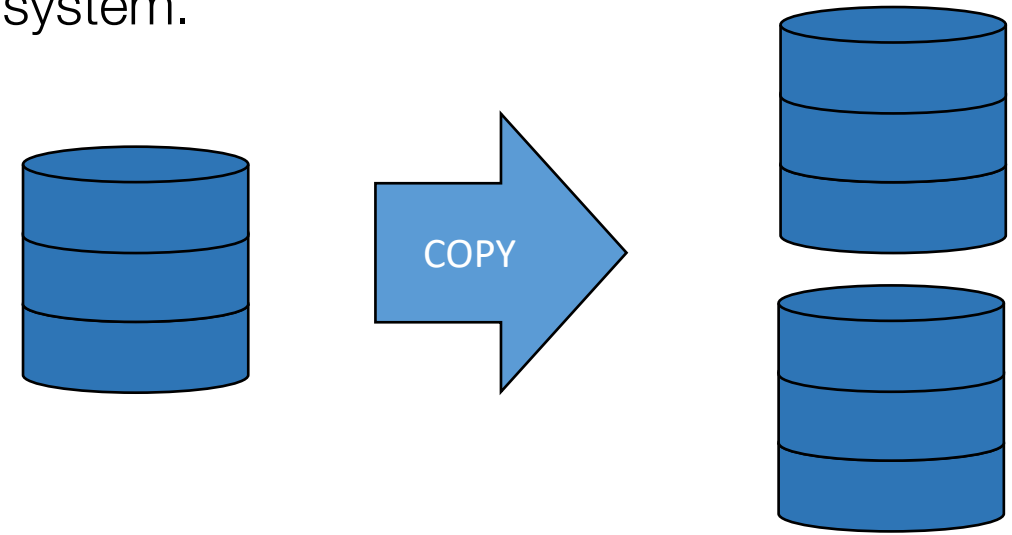
See also: Sharding (as **horizontal partitioning**).



Replication

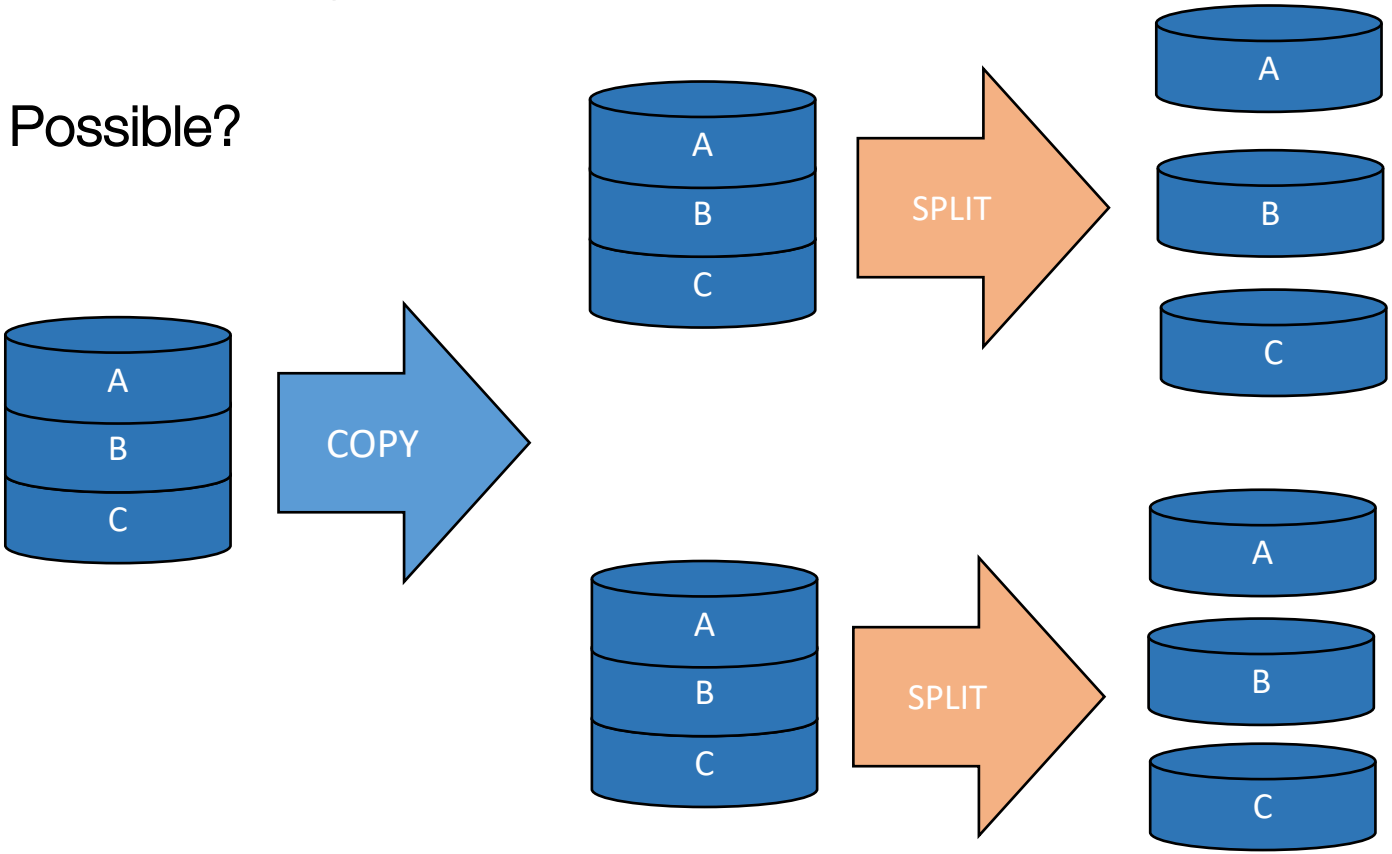
Aim: fault-tolerance, backup

Replication copies the entire database across all nodes in the distributed system.



Partitioning + Replication

Possible?



Pros / Cons of Each

	Partitioning	Replication
Pros	Fast data writing / reading. Low memory overhead.	Fast data reading. High data reliability.
Cons	Potential data loss	High network overhead. High memory overhead.

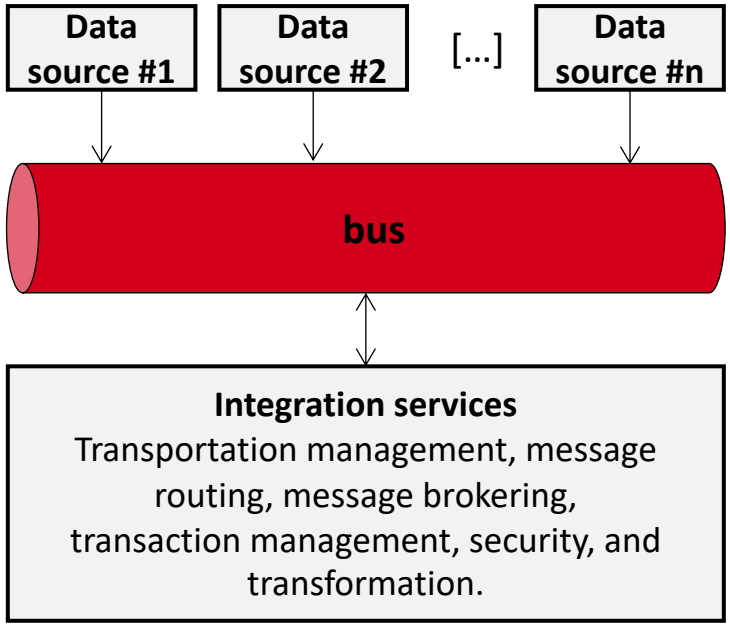
Scale and Ingestion

Scalability

- How big is big?
- Not only scaling up
  - Elasticity

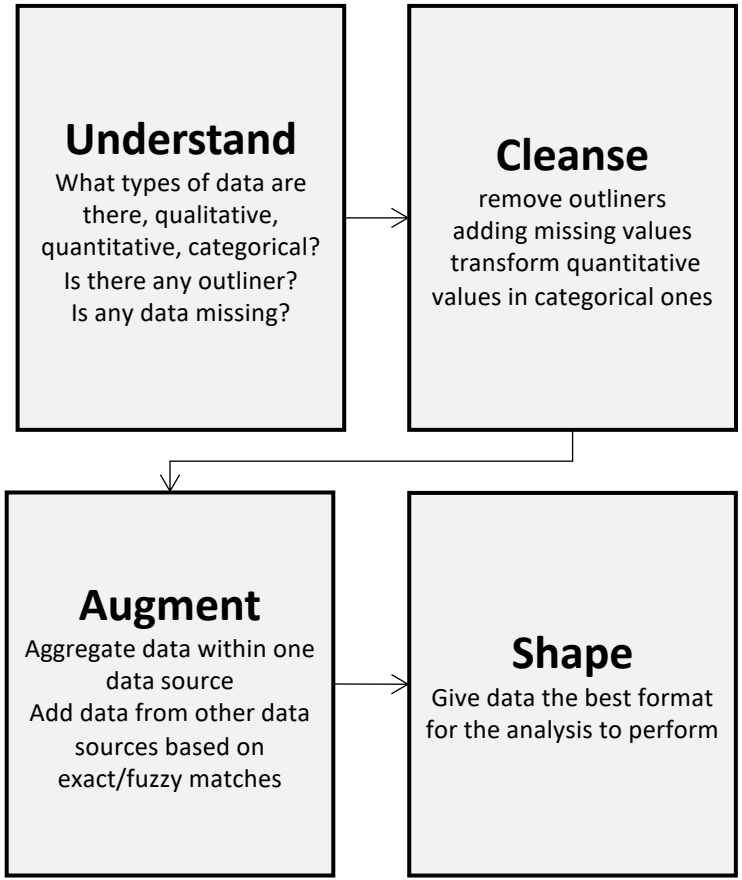
Data Ingestion

- The process of importing, transferring and loading data for storage and later use
- It involves loading data from a variety of sources
- It can involve altering and modification of individual files to fit into a format that optimizes the storage
- For instance, in Big Data small files are concatenated to form files of 100s of MBs and large files are broken down in files of 100s of MB.



Data Wrangling

The process of cleansing "raw" data and transforming raw it into data that can be analysed to generate valid actionable insights.
It includes understanding, cleansing, augmenting and shaping data.
The results is data in the best format (e.g., columnar) for the analysis to perform.



POLITECNICO
MILANO 1863

SYSTEMS AND METHODS FOR BIG AND UNSTRUCTURED DATA

Data Architecture Concepts

Marco Brambilla

marco.brambilla@polimi.it

@marcobrambi