I can't pass an extremely competitive test to become a surgeon. But you give me any operation on a heart. I can perhaps do much better than most people. I am like an artist. Don't expect me to compete in an exam. Give me the job and I will show you how good I am.

NoSQL DB

Venkatesh Vinayakarao
venkateshv@cmi.ac.in
http://vvtesh.co.in

Chennai Mathematical Institute

The cost of managing traditional databases is high. Mistakes made during routine maintenance are responsible for 80 percent of application downtime. **– Dev Ittycheria, MongoDB.**

A Relation as a Data Model

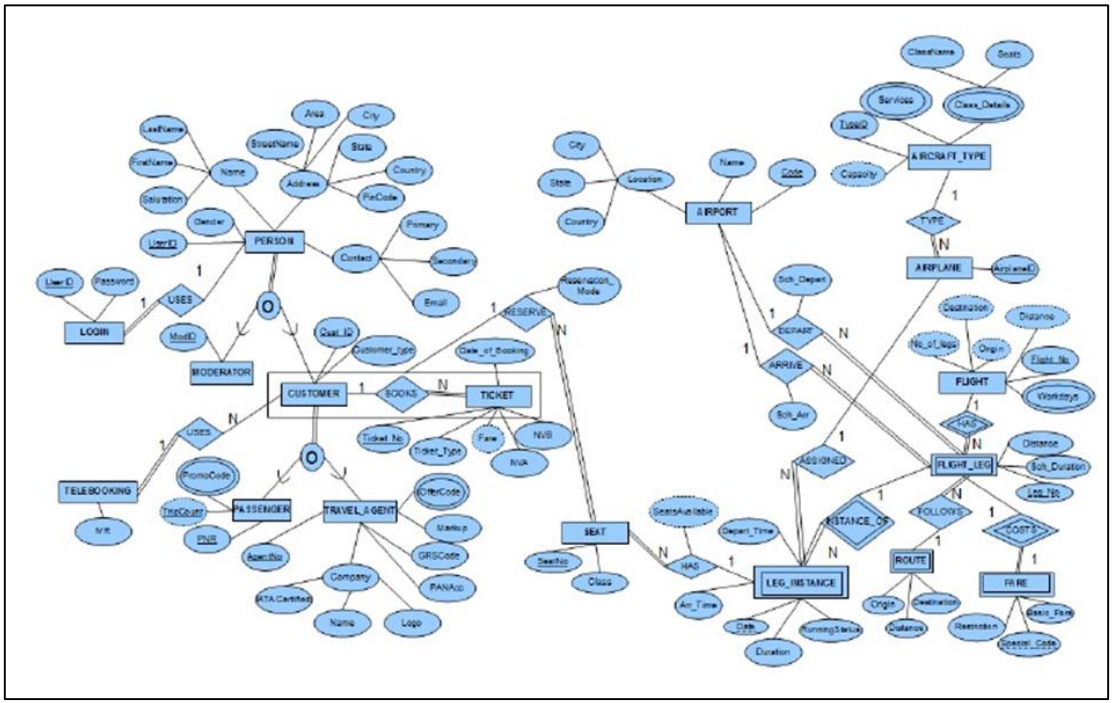
- Let the set, id = {1,2,3}
- Let the set, names = {vv, sd}
- What is id x names?
- We have a relation if we assign a sequential id to each name.

id	name
1	sd
2	vv

id	name
1	sd
1	vv
2	sd
2	vv
3	sd
3	vv

… and thus we had the relational database.

An Entity-Relationship Design



DB Designs:

- **Can get too complex!**
- **May become too hard to maintain!!**

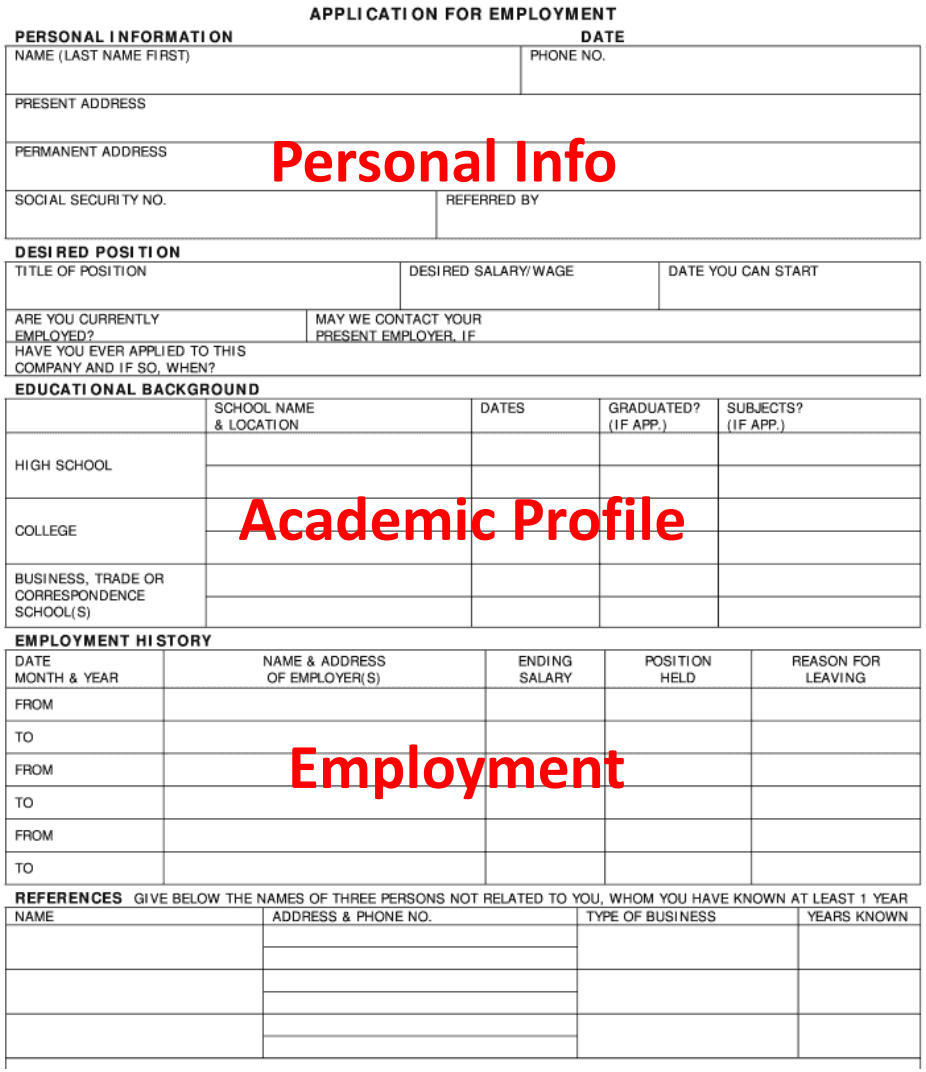
Source: https://www.slideshare.net/NarendranThangarajan/airline-reservation-system-db-design

Key Challenges of Relational DB

- Schema needs to be defined.
- Maintenance becomes harder over time.
- Impedance mismatch problem.
- Does not scale out by design.
- ACID Transactions – Consistency Vs. Availability Trade-off.

Impedance Mismatch

DB Design



APPLICATION FOR EMPLOYMENT

PERSONAL INFORMATION **DATE**

NAME (LAST NAME FIRST)	PHONE NO.
PRESENT ADDRESS	
PERMANENT ADDRESS	
SOCIAL SECURITY NO.	REFERRED BY

Personal Info

DESIRED POSITION

TITLE OF POSITION	DESIRED SALARY/WAGE	DATE YOU CAN START
ARE YOU CURRENTLY EMPLOYED?	MAY WE CONTACT YOUR PRESENT EMPLOYER, IF	
HAVE YOU EVER APPLIED TO THIS COMPANY AND IF SO, WHEN?		

EDUCATIONAL BACKGROUND

	SCHOOL NAME & LOCATION	DATES	GRADUATED? (IF APP.)	SUBJECTS? (IF APP.)
HIGH SCHOOL				
COLLEGE				
BUSINESS, TRADE OR CORRESPONDENCE SCHOOL(S)				

Academic Profile

EMPLOYMENT HISTORY

DATE MONTH & YEAR	NAME & ADDRESS OF EMPLOYER(S)	ENDING SALARY	POSITION HELD	REASON FOR LEAVING
FROM				
TO				
FROM				
TO				
FROM				
TO				

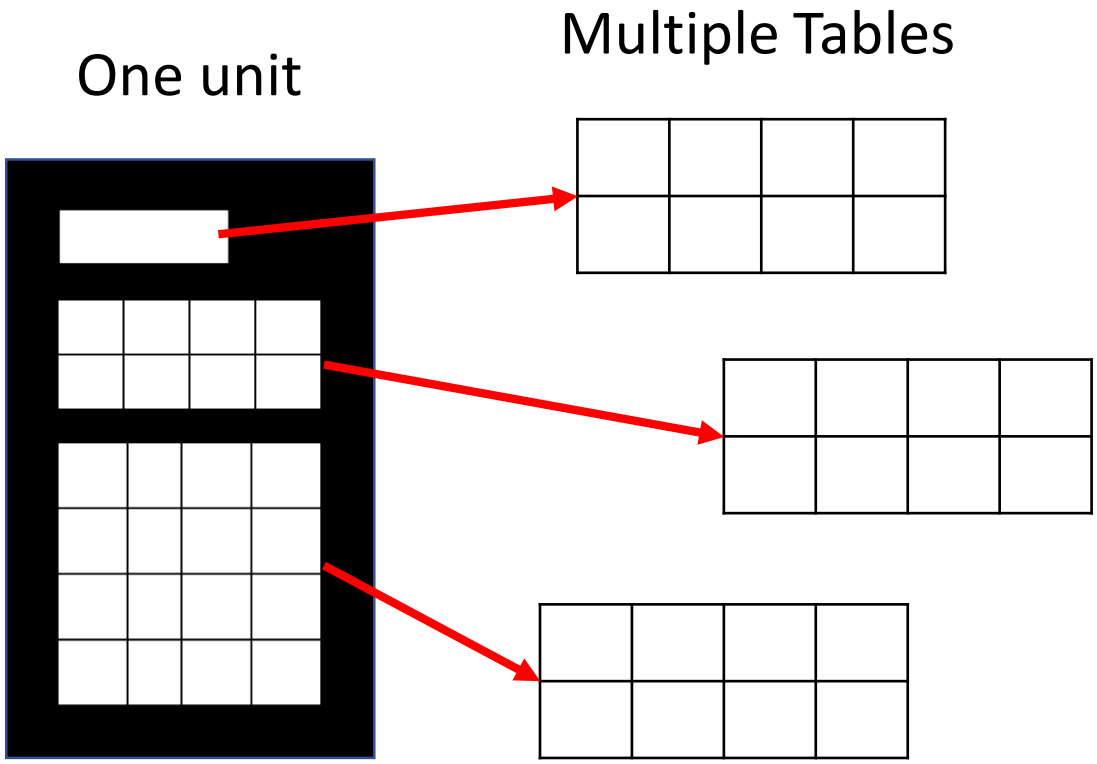
Employment

REFERENCES GIVE BELOW THE NAMES OF THREE PERSONS NOT RELATED TO YOU, WHOM YOU HAVE KNOWN AT LEAST 1 YEAR

NAME	ADDRESS & PHONE NO.	TYPE OF BUSINESS	YEARS KNOWN

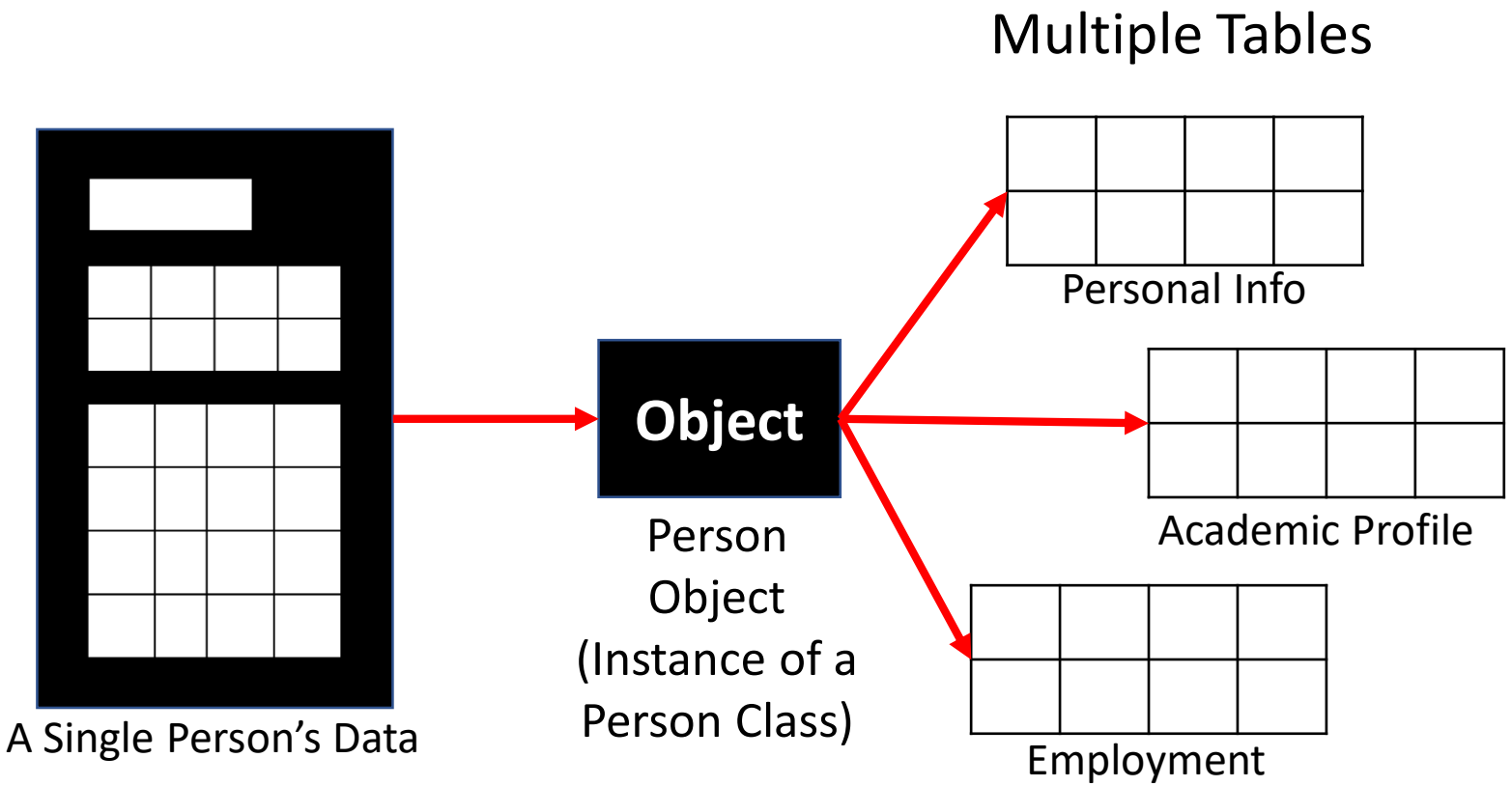
How will you design the DB for this content?

Impedance Mismatch Problem



Intermediate Solution: Object Relational Mapping (ORM)

Object Relational Mapping



Hibernate Framework, Java Data Objects, … and many other ORM frameworks emerged.

Scaling Out

Table Joins Using MapReduce

- How would you do it?

Map-side Join

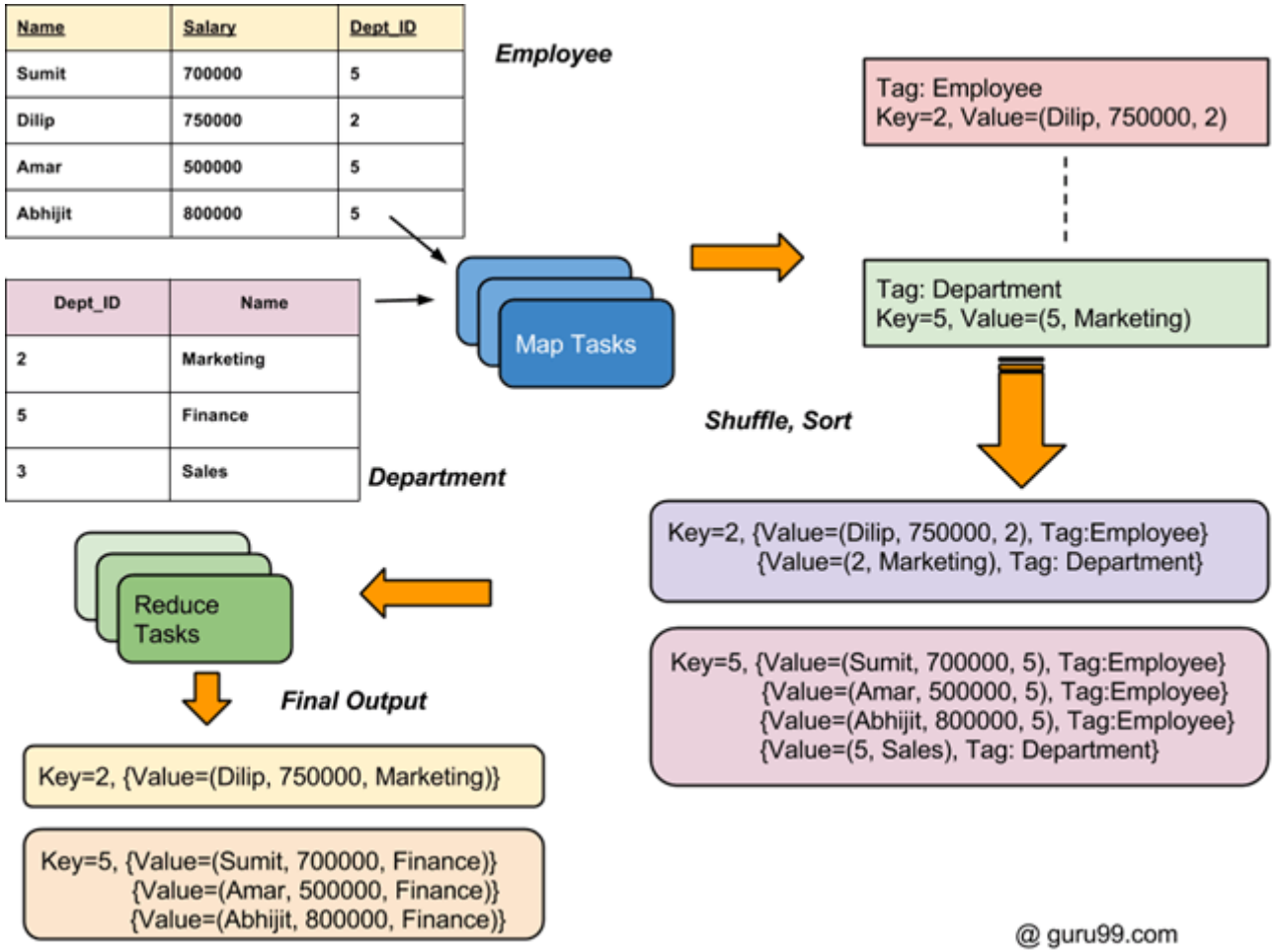
Join is performed by the mapper.

Reduce-side Join

Join is performed by the reducer.

Table joins are expensive. So, new solutions emerged. Google BigTable, Amazon Dynamo…

Join Pattern



See https://www.guru99.com/introduction-to-counters-joins-in-map-reduce.html

A New Movement was Born

- We needed a
  - Not only relational
  - Cluster friendly
  - Schemaless

  way to store and retrieve data.
- Johan Oskarsson proposed a meetup. He needed a twitter hashtag. He used, “**nosql**”.

Transactions, Consistency and CAP Theorem

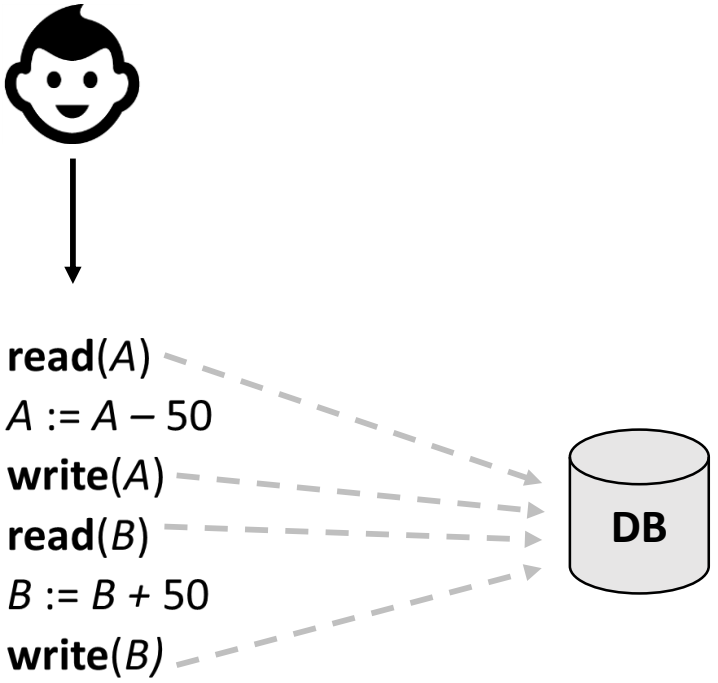
Transaction

1. **read**(*A*)
2. *A* := *A* – 50
3. **write**(*A*)
4. **read**(*B*)
5. *B* := *B* + 50
6. **write**(*B*)

transfer $50 from account A to account B

A transaction is a *unit* of program execution that accesses and possibly updates various data items.

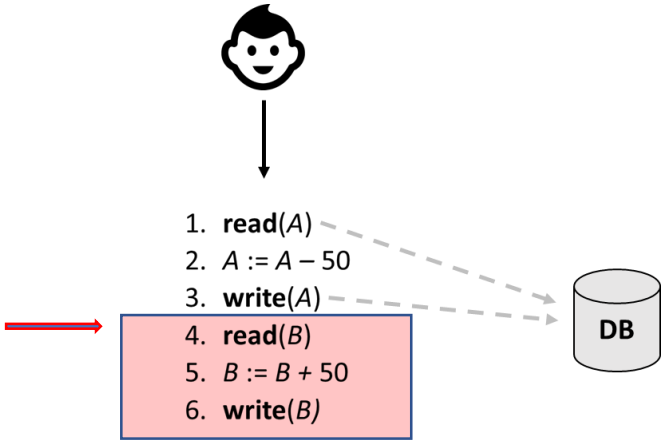
Do You See Any Issues Here?



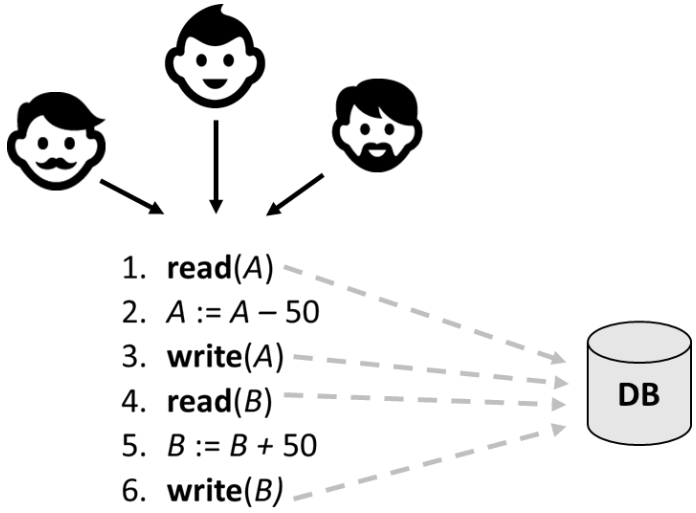
A transaction that reads
and writes to disk.

Issues

- Two main issues to deal with:



Failure (hardware failure, system crash, software defect…)



concurrent execution

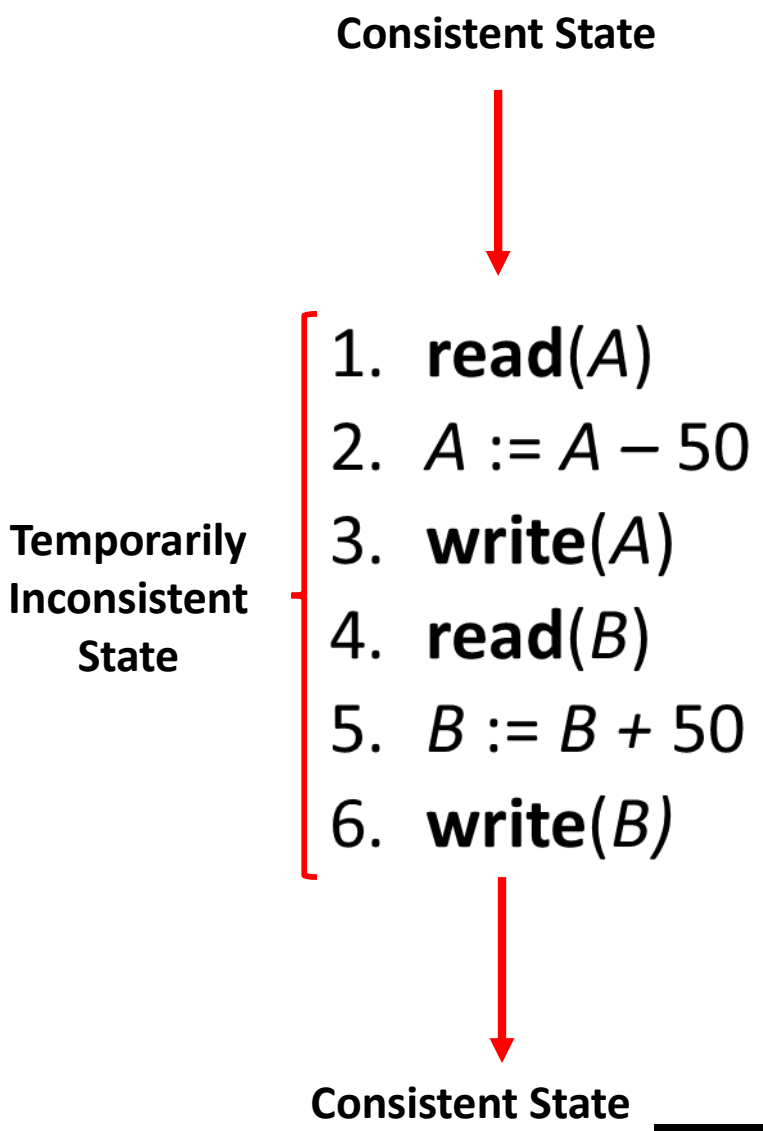
Atomicity

- **What happens if step 3 is executed but not step 6?**
  - Failure could be due to software or hardware
- The system should ensure that updates of a partially executed transaction are not reflected in the database.

1. **read**(A)
2. $A := A - 50$
3. **write**(A)
4. **read**(B)
5. $B := B + 50$
6. **write**(B)

Consistency

- Respect
  - Explicitly specified integrity constraints
  - Implicit integrity constraints
    - e.g., sum of balances of all accounts stays constant



Isolation

- T2 sees an inconsistent database if T1 and T2 are concurrent.

T1	T2
1. **read**(*A*)	
2. *A* := *A* – 50	
3. **write**(*A*)	
	read(A), read(B), print(A+B)
4. **read**(*B*)	
5. *B* := *B* + 50	
6. **write**(*B*)	

- Isolation can be ensured trivially by running transactions **serially**
  - That is, one after the other.

Durability

- After step 6, the updates to the database by the transaction must
  - persist even if there are software or hardware failures.

1. **read**(A)
2. $A := A - 50$
3. **write**(A)
4. **read**(B)
5. $B := B + 50$
6. **write**(B)

ACID Properties

- **Atomicity.** Either all operations of the transaction are properly reflected in the database or none are.
- **Consistency.** Execution of a transaction in isolation preserves the consistency of the database.
- **Isolation.** Although multiple transactions may execute concurrently, each transaction must be unaware of other concurrently executing transactions. Intermediate transaction results must be hidden from other concurrently executed transactions.
  - That is, for every pair of transactions T_i and T_j, it appears to T_i that either T_j finished execution before T_i started, or T_j started execution after T_i finished.
- **Durability.** After a transaction completes successfully, the changes it has made to the database persist, even if there are system failures.

But, as a facebook user, I had a different observation…

Eventual Consistency

- I updated my facebook status and asked my friend to check it out.
- But she found nothing there!!!
- Asked her to wait a bit and check again.
- Now, she finds it!

Eventual Consistency

- Facebook is eventually consistent.
- Why not use a strongly consistent model?
  - Stores Petabytes of data.
  - We have Availability vs. Consistency tradeoff.

CAP Theorem

- Concerns while designing distributed systems:
  - **Consistency** –all clients of a data store get responses to requests that ‘make sense’. For example, if Client A writes 1 and later 2 to location X, Client B cannot read 2 followed by 1.
  - **Availability** – all operations on a data store eventually return successfully. We say that a data store is ‘available’ for, e.g. write operations.
  - **Partition tolerance** – if the network stops delivering messages between two sets of servers, will the system continue to work correctly?

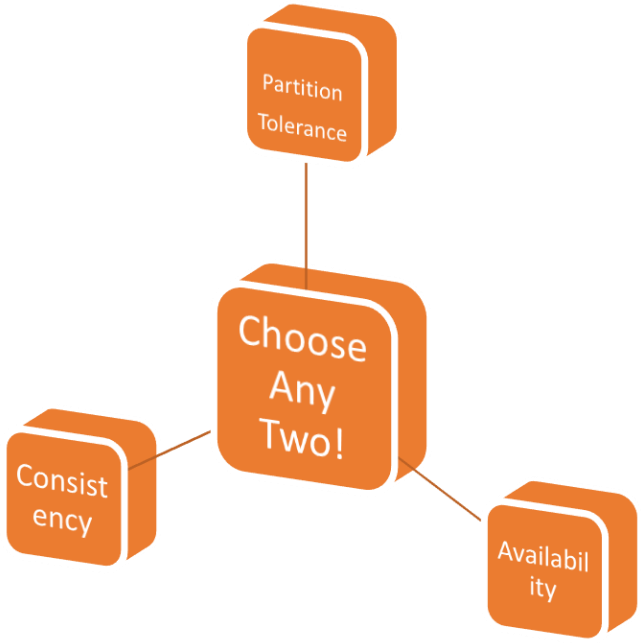
The CAP Message

If you:

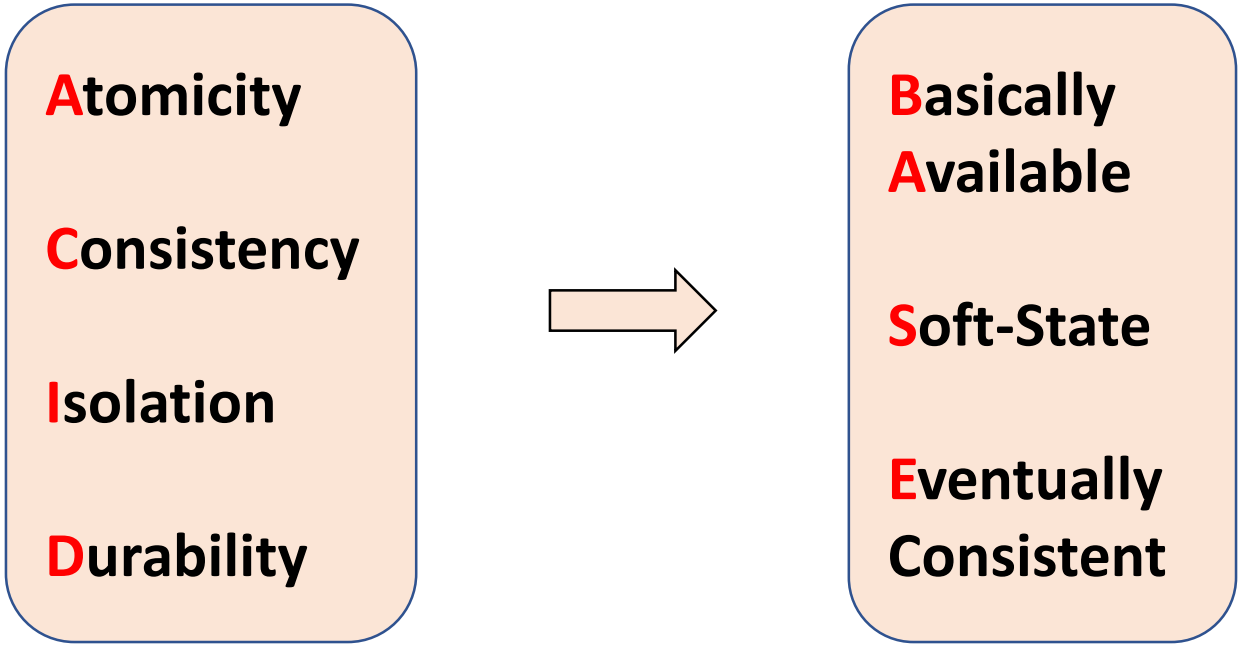
- cannot limit the number of faults,
- requests can be directed to any server, and
- insist on serving every request you receive,

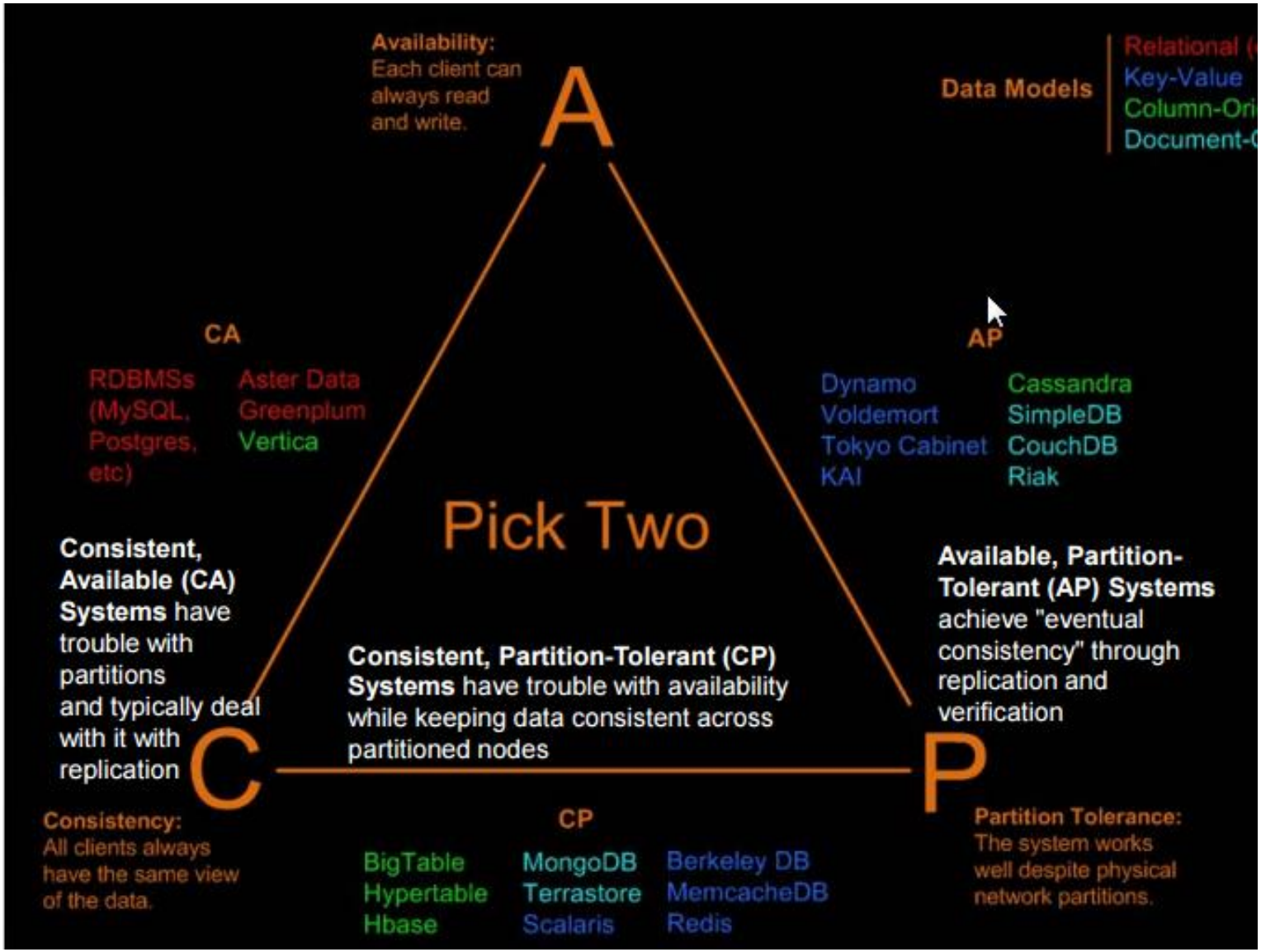
Then:

- you cannot possibly be consistent.



The Transaction Properties




Availability:
Each client can always read and write.
A
Data Models
Relational (
Key-Value
Column-Ori
Document-
CA
RDBMSs (MySQL, Postgres, etc)
Aster Data
Greenplum
Vertica
AP
Dynamo
Voldemort
Tokyo Cabinet
KAI
Cassandra
SimpleDB
CouchDB
Riak
Pick Two
Consistent, Available (CA) Systems have trouble with partitions and typically deal with it with replication
Consistent, Partition-Tolerant (CP) Systems have trouble with availability while keeping data consistent across partitioned nodes
Available, Partition-Tolerant (AP) Systems achieve "eventual consistency" through replication and verification
C
P
Consistency:
All clients always have the same view of the data.
CP
BigTable
Hypertable
Hbase
MongoDB
Terrastore
Scalaris
Berkeley DB
MemcacheDB
Redis
Partition Tolerance:
The system works well despite physical network partitions.

NoSQL DB Types

Types of NoSQL DB

- Key-Value Stores
  - Simplest. Every item is a key-value pair.
  - Examples: Riak, Voldemort, and Redis
- Document DB
  - Complex data structures are represented as documents.
  - Examples: MongoDB
- Wide-Column Stores
  - Data stored as columns.
  - Examples: Cassandra and Hbase
- Graph DB
  - Examples: Neo4J and HyperGraphDB

Redis DB – Key Value Store

```
redis> GET nonexisting
(nil)
redis> SET mykey "Hello"
"OK"
redis> GET mykey
"Hello"
redis>
```

Read https://redis.io/commands/get

Voldemort DB

```
> bin/voldemort-shell.sh test tcp://localhost:6666
Established connection to test via tcp://localhost:6666
> put "hello" "world"
> get "hello"
version(0:1): "world"
> delete "hello"
> get "hello"
null
> help
...
> exit
k k thx bye.
```

mongoDB – Document Database

- mongoDB = “Hu**mongo**us DB”
  - Open-source
  - Document-based data model
  - “High performance, high availability”
  - Automatic scaling
  - C-P on CAP

MongoDB vs. RDBMS

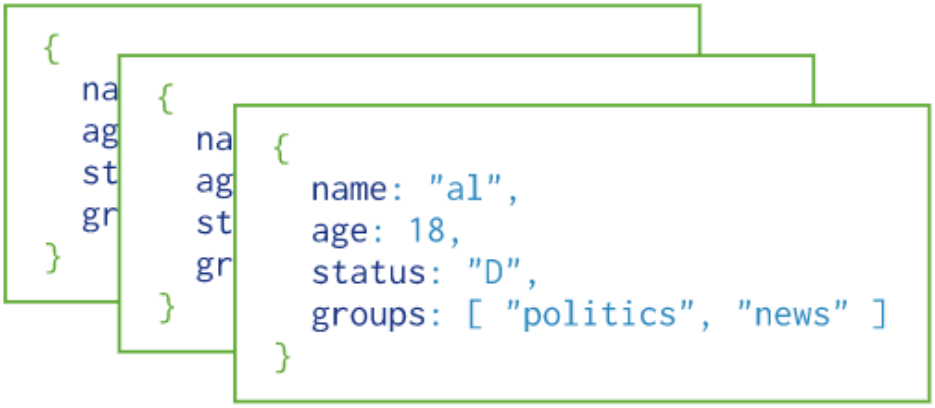
- Collection vs. table
- Document vs. row
- Field vs. column
- Schema-less

Document Data Model

- Documents are a natural way to represent data.
- Here is a “Person” object represented as a JSON document.
- MongoDB stores this as a BSON document (Binary representation of JSON).

```
{
  name: "sue",                    <-- field: value
  age: 26,                        <-- field: value
  status: "A",                    <-- field: value
  groups: [ "news", "sports" ]    <-- field: value
}
```

A record in MongoDB is a document

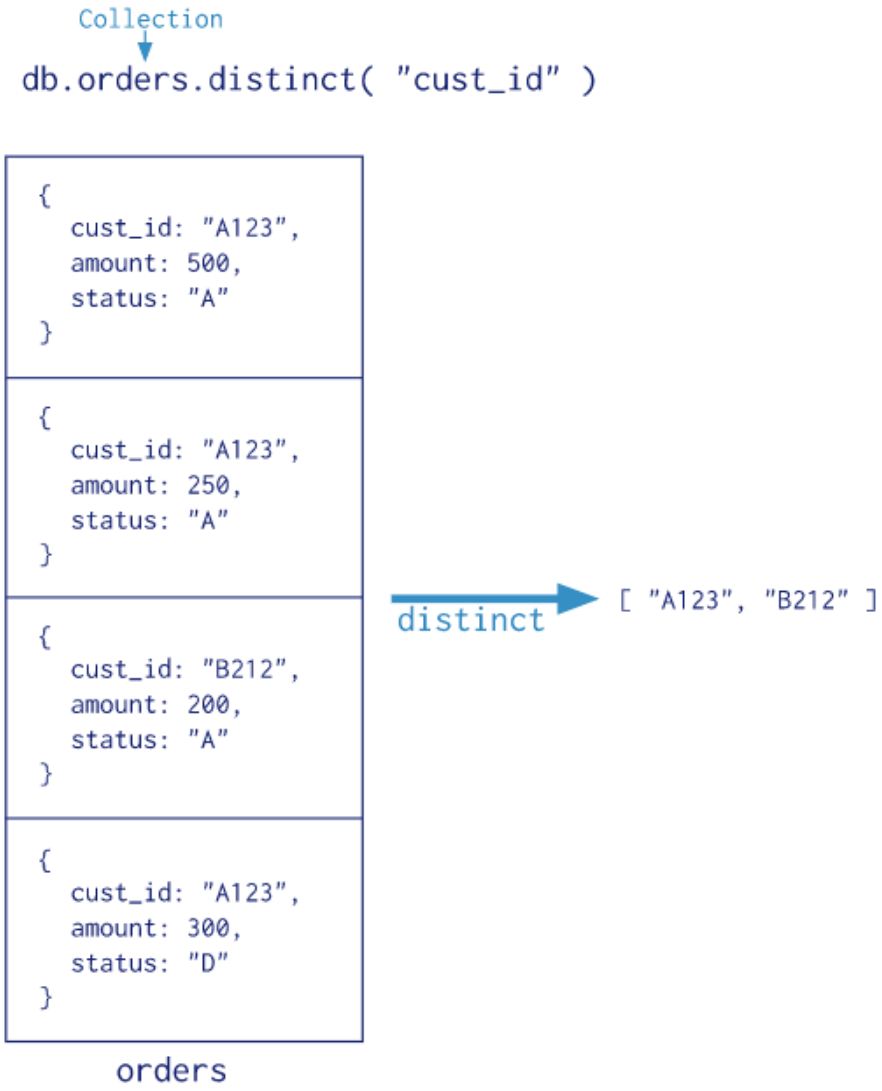



```
db.myNewCollection2.insertOne( { x: 1 } )
db.orders.deleteOne({ “name" : “al" } );
```

Commands

Read https://docs.mongodb.com/manual/core/databases-and-collections/

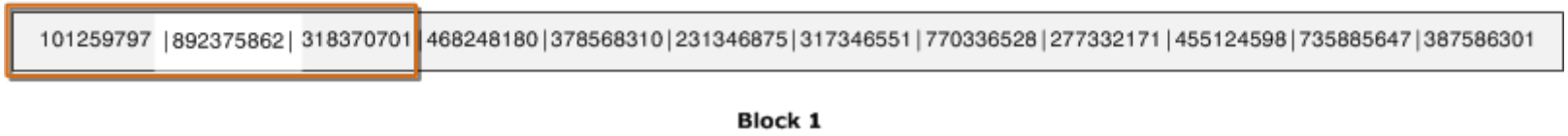
Operations on MongoDB Data



Columnar Storage

SSN	Name	Age	Addr	City	St
101259797	SMITH	88	899 FIRST ST	JUNO	AL
892375862	CHIN	37	16137 MAIN ST	POMONA	CA
318370701	HANDU	12	42 JUNE ST	CHICAGO	IL

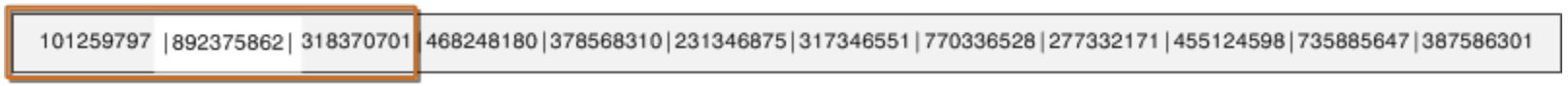
SSN	Name	Age	Addr	City	St
101259797	SMITH	88	899 FIRST ST	JUNO	AL
892375862	CHIN	37	16137 MAIN ST	POMONA	CA
318370701	HANDU	12	42 JUNE ST	CHICAGO	IL



Read https://docs.aws.amazon.com/redshift/latest/dg/c_columnar_storage_disk_mem_mgmnt.html

Columnar Storage

SSN	Name	Age	Addr	City	St
101259797	SMITH	88	899 FIRST ST	JUNO	AL
892375862	CHIN	37	16137 MAIN ST	POMONA	CA
318370701	HANDU	12	42 JUNE ST	CHICAGO	IL



Block 1

Same datatype in a block helps in devising efficient compression schemes. Therefore, improve storage efficiency.

Assumption: "*OLTP transactions typically involve most or all of the columns in a row for a small number of records, data warehouse queries commonly read only a few columns for a very large number of rows*"

Cassandra - Wide-Column Store

- A **column** is the basic data structure of Cassandra.
- A Column has three values, namely key or column name, value, and a time stamp.

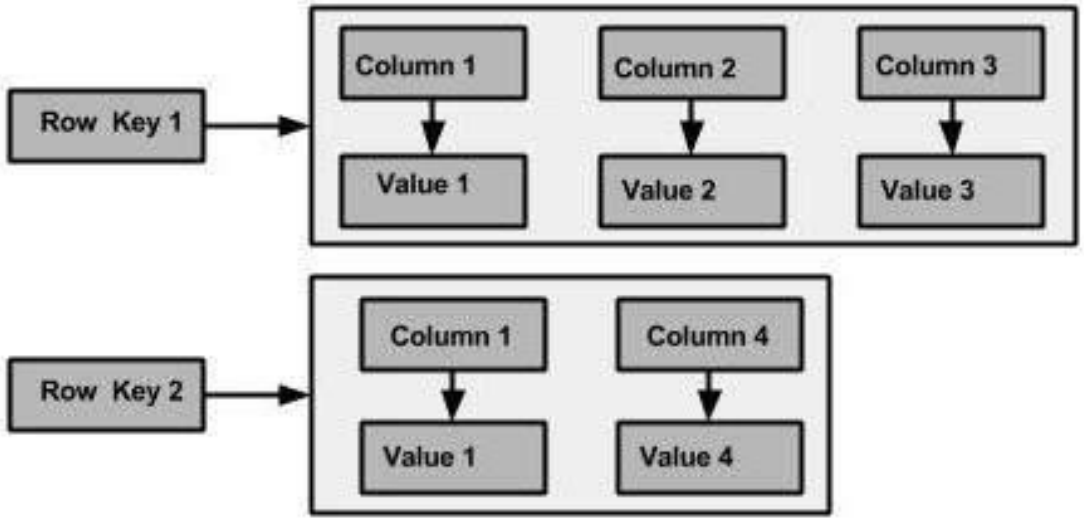
Column		
name : byte[]	value : byte[]	clock : clock[]

- A **super column** is a special column. stores a map of sub-columns.

Super Column	
name : byte[]	cols : map<byte[], column>

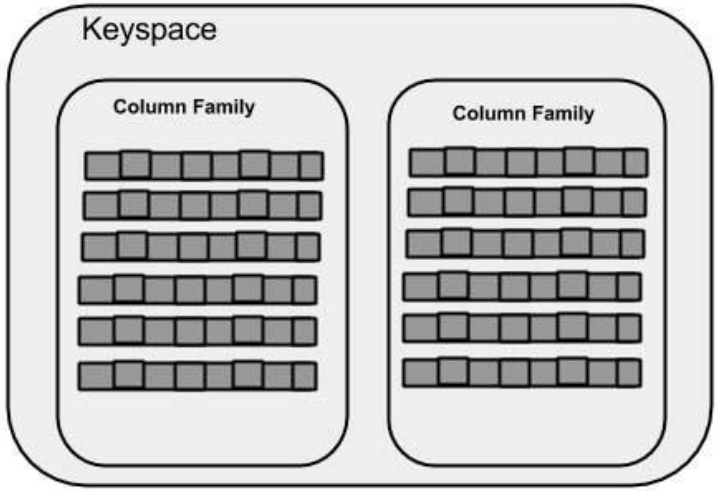
Column-Family DB

- Cassandra does not force individual rows to have all the columns.
- An example of a Cassandra column family:



Cassandra Keyspace

- Keyspace is a container for a list of one or more column families.
- A column family, in turn, is a container of a collection of rows.
- Each row contains ordered columns.



cqlsh

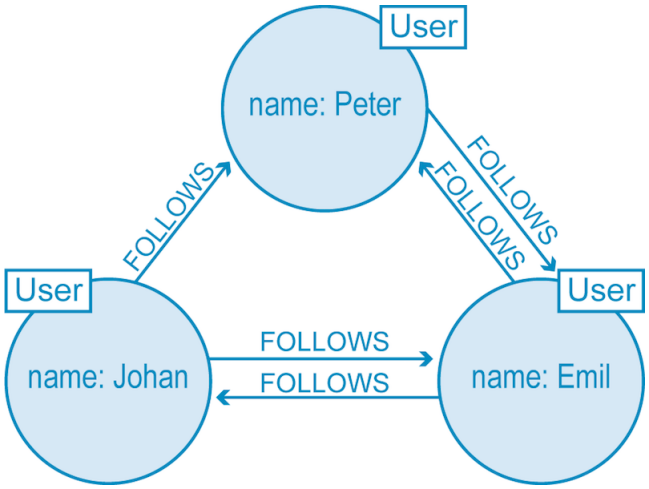
- Cassandra Query Language Shell

```
[hadoop@linux bin]$ cqlsh
Connected to … Cluster at ….
cqlsh>  select * from emp;
```

- Note: Cassandra does not join!
- If you need to lookup several tables, create another column-family.

Graph DB

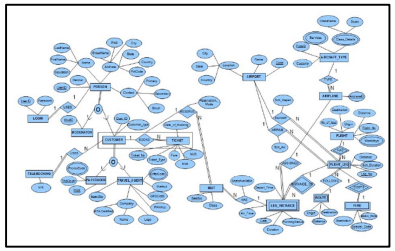
- Facebook, LinkedIn, Google …have connected data.
- It is natural to store and retrieve data as graphs.



Twitter users represented in a graph database model.

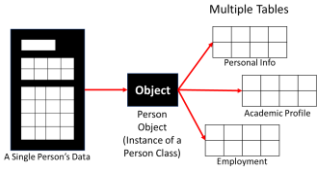
Read https://neo4j.com/blog/why-graph-databases-are-the-future/

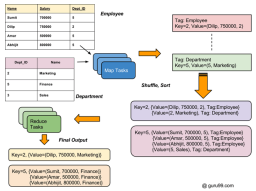
Summary



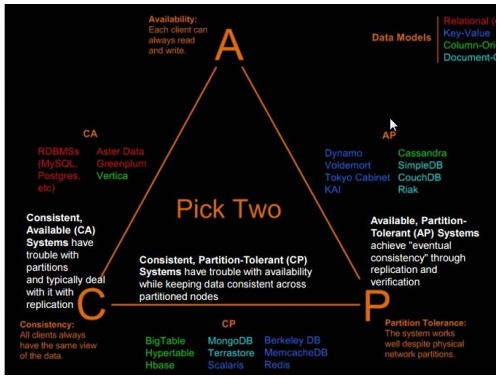
Schema-based Relational Model - maintenance problems

Impedance Mismatch





Scale-up Challenges



CAP Theorem

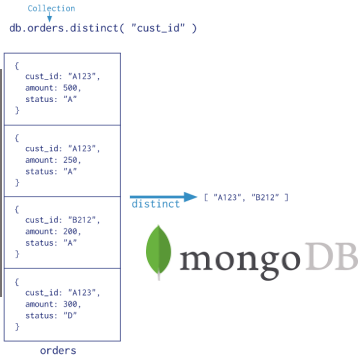
Types of NoSQL datastores

Key-Valuecv

```
redis> GET nonexisting
(nil)
redis> SET mykey "Hello"
"OK"
redis> GET mykey
"Hello"
redis>
```

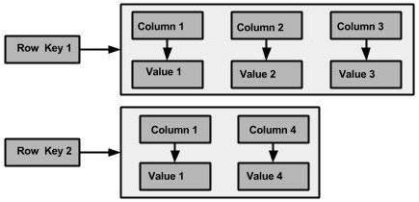
redis

Doc-based



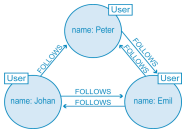
mongoDB

Columnar DB



Apache CASSANDRA

Graph DB



neo4j

Thank You