SQL for Data Scientists: 1

Rahul Dave, Univ.Ai

Univ.AI

What is a relational database?

- a relation (table) is a collection of tuples. Each tuple is called a *row*
- a database is a collection of tables related to each other through common data values.
- Everything in a column is values of one attribute
- A cell is expected to be atomic, no lists, dictionaries, etc
- Tables are related to each other if they have columns called keys which represent the same values
- SQL a declarative model: a query optimizer decides how to execute the query (if a field range covers 80% of values, should we use the index or the table?). Also parallelizable

Table: contributors

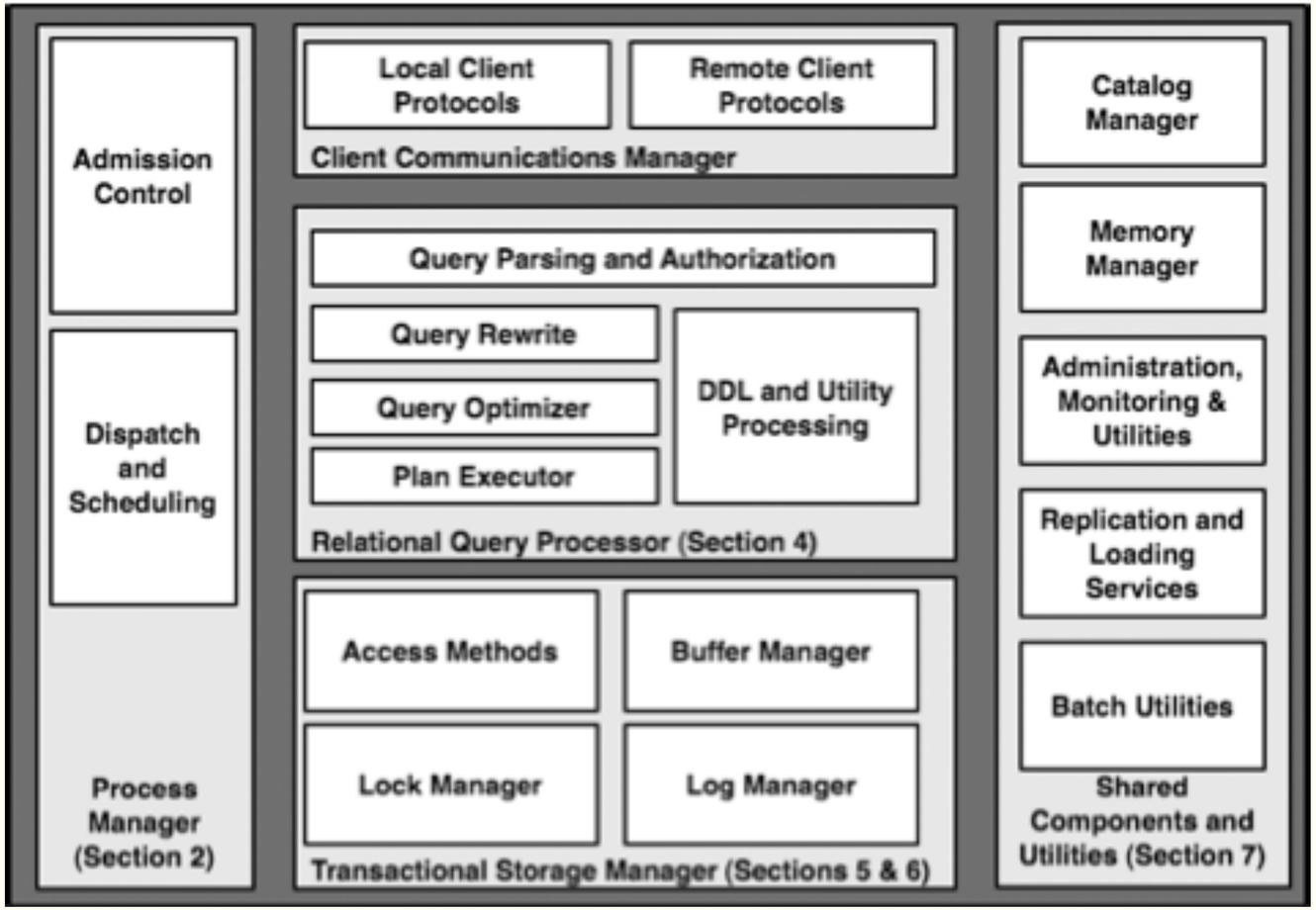
New Record | Delete Record

	id	last_name	first_name	middle_name	street_1	street_2	city	state	zip	amount	date	candidate_id
	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter
1	1	Agee	Steven	NULL	549 Laurel ...	NULL	Floyd	VA	24091	500	2007-06-30	16
2	5	Akin	Charles	NULL	10187 Suga...	NULL	Bentonville	AR	72712	100	2007-06-16	16
3	6	Akin	Mike	NULL	181 Baywo...	NULL	Monticello	AR	71655	1500	2007-05-18	16
4	7	Akin	Rebecca	NULL	181 Baywo...	NULL	Monticello	AR	71655	500	2007-05-18	16
5	8	Aldridge	Brittni	NULL	808 Capitol...	NULL	Washington	DC	20024	250	2007-06-06	16
6	9	Allen	John D.	NULL	1052 Cann...	NULL						
7	10	Allen	John D.	NULL	1052 Cann...	NULL						
8	11	Allison	John W.	NULL	P.O. Box 10...	NULL						
9	12	Allison	Rebecca	NULL	3206 Sum...	NULL						

	id	first_name	last_name	middle_name	party
	Filter	Filter	Filter	Filter	Filter
1	16	Mike	Huckabee		R
2	20	Barack	Obama		D
3	22	Rudolph	Giuliani		R
4	24	Mike	Gravel		D
5	26	John	Edwards		D
6	29	Bill	Richardson		D
7	30	Duncan	Hunter		R
8	31	Dennis	Kucinich		D
9	32	Ron	Paul		R

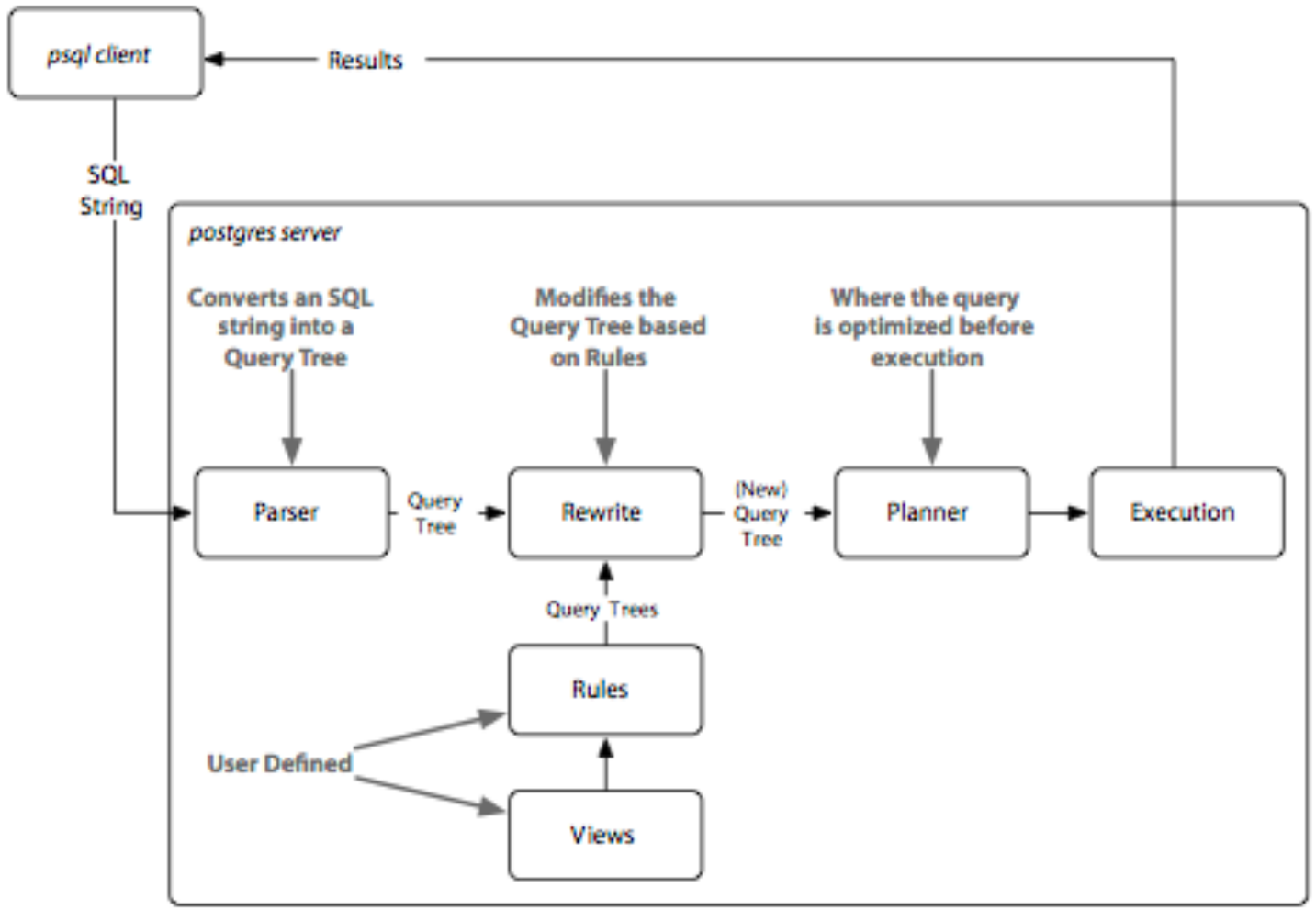
How do databases work?

- client connection manager: what to do with incomings
- transactional storage: storage data structures and the log
- process model: coroutines, threads, processes
- query model and language: query optimization

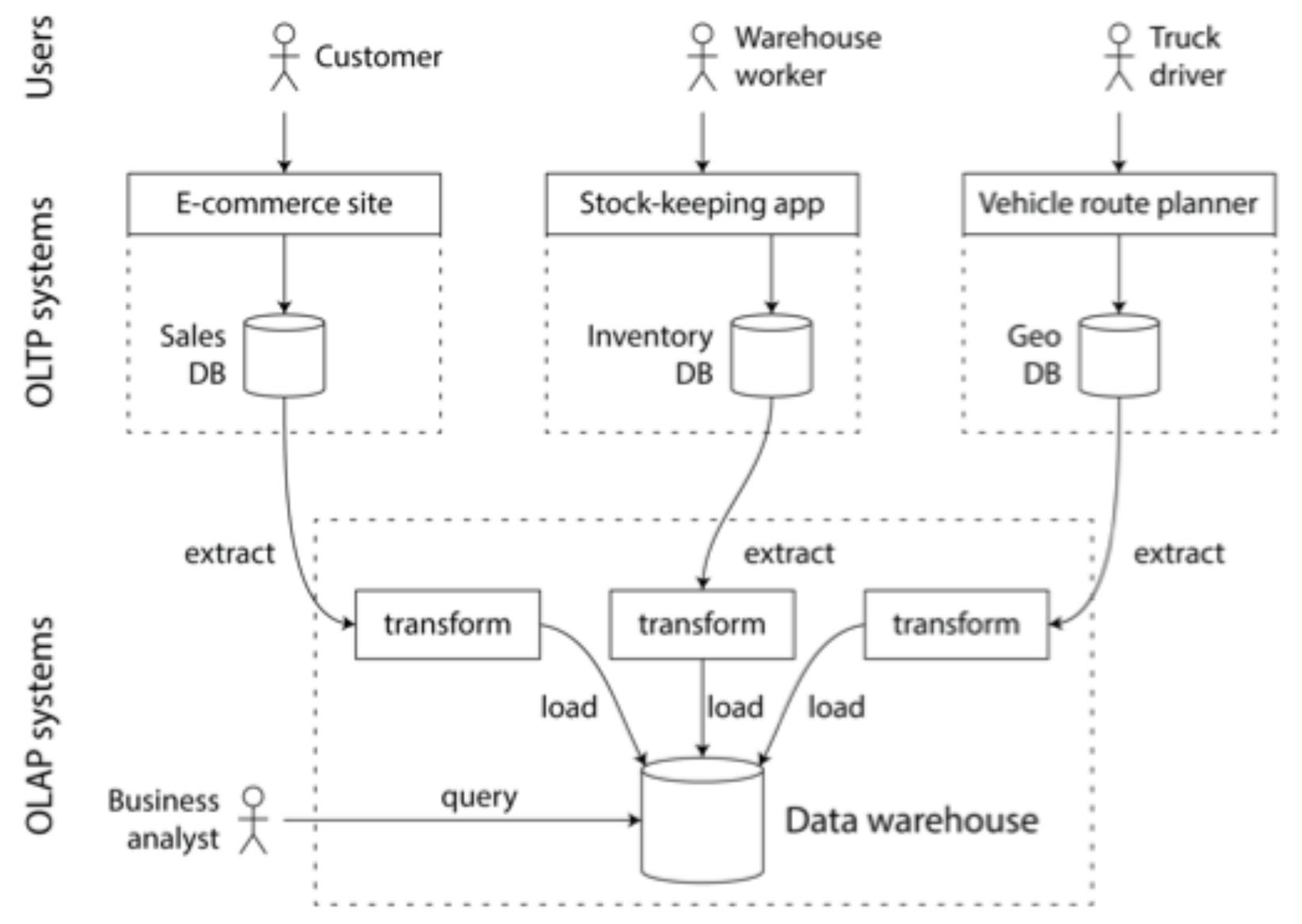


What is SQL?

SQL is a declarative model: a query optimizer decides how to execute the query (if a field range covers 80% of values, should we use the index or the table?). Also parallelize-able.



Where are the databases used?



Customer facing systems need transactions

- The general rules
  - The batch of operations is viewed as a single atomic operation, so all of the operations either succeed together or fail together.
  - The database is in a valid state before and after the transaction.
  - The batch update appears to be isolated; other queries should never see a database state in which only some of the operations have been applied.
- Databases have a mechanism for wrapping a single or multiple processes into a Transaction. This means that the batch of operations either all happen (commit) or not happen at all (abort, rollback). This is called atomicity.

How would you model data?

- The needs of OLTP databases are very different from those of OLAP databases
- OLTP databases usually need CRUD operations: CReate, Update, Delete
- OLTP tables (and incoming OLAP schemas) have a star like structure. *Fact tables* with pointers, or **keys** to *dimension tables*.
- Normalization: *The attributes of a table should be dependent on the primary key, on the whole key and nothing but the key.*

http://www.linkedin.com/in/williamhgates

Bill Gates

Greater Seattle Area | Philanthropy

Summary

Co-chair of the Bill & Melinda Gates Foundation. Chairman, Microsoft Corporation. Voracious reader. Avid traveler. Active blogger.

Experience

Co-chair • Bill & Melinda Gates Foundation
2000 – Present

Co-founder, Chairman • Microsoft
1975 – Present

Education

Harvard University
1973 – 1975

Lakeside School, Seattle

Contact Info

Blog: thegatesnotes.com
Twitter: @BillGates

users table

user_id	first_name	last_name	summary	region_id	industry_id	photo_id
251	Bill	Gates	Co-chair of … blogger.	us:91	131	57817532

regions table

id	region_name
us:7	Greater Boston Area
us:91	Greater Seattle Area

industries table

id	industry_name
43	Financial Services
48	Construction
131	Philanthropy

positions table

id	user_id	job_title	organization
458	251	Co-chair	Bill & Melinda Gates F…
457	251	Co-founder, Chairman	Microsoft

education table

id	user_id	school_name	start	end
807	251	Harvard University	1973	1975
806	251	Lakeside School, Seattle	NULL	NULL

contact_info table

id	user_id	type	url
155	251	blog	http://thegatesnotes.com
156	251	twitter	http://twitter.com/BillGates

Normalization

The process of organizing data in the database in such a way that it handles the transactions in an efficient manner. Details on Normal Forms: https://www.guru99.com/database-normalization.html

Full Names	Physical Address	Movies Rented	Salutation
Janet Jones	First Street Plot No 4	Pirates of the Caribbean, Clash of the Titans	Ms.
Robert Phil	3rd Street 34	Forgetting Sarah Marshal, Daddy's Little Girls	Mr.
Robert Phil	5th Avenue	Clash of the Titans	Mr.

Membership ID	Full Names	Physical Address	Salutation ID
1	Janet Jones	First Street Plot No 4	2
2	Robert Phil	3rd Street 34	1
3	Robert Phil	5th Avenue	1

Membership ID	Movies Rented
1	Pirates of the Caribbean
1	Clash of the Titans
2	Forgetting Sarah Marshal
2	Daddy's Little Girls
3	Clash of the Titans

Salutation ID	Salutation
1	Mr.
2	Ms.
3	Mrs.
4	Dr.

1NF: table cells/cols unique type, each row unique **2NF**: single column primary key **3NF**: no transitive functional dependencies

Normalization

Full Names	Physical Address	Movies rented	Salutation
Janet Jones	First Street Plot No 4	Pirates of the Caribbean, Clash of the Titans	Ms.
Robert Phil	3rd Street 34	Forgetting Sarah Marshal, Daddy's Little Girls	Mr.
Robert Phil	5th Avenue	Clash of the Titans	Mr.

Full Names	Physical Address	Movies rented	Salutation
Janet Jones	First Street Plot No 4	Pirates of the Caribbean	Ms.
Janet Jones	First Street Plot No 4	Clash of the Titans	Ms.
Robert Phil	3rd Street 34	Forgetting Sarah Marshal	Mr.
Robert Phil	3rd Street 34	Daddy's Little Girls	Mr.
Robert Phil	5th Avenue	Clash of the Titans	Mr.

1NF

Membership ID	Full names	Physical Address	Salutation
1	Janet Jones	First Street Plot No 4	Ms.
2	Robert Phil	3rd Street 34	Mr.
3	Robert Phil	5th Avenue	Mr.

2NF

Membership ID	Movies rented
1	Pirates of the Caribbean
1	Clash of the Titans
2	Forgetting Sarah Marshal
2	Daddy's Little Girls
3	Clash of the Titans

Salutation ID	Salutation
1	Mr.
2	Ms.
3	Mrs.
4	Dr.

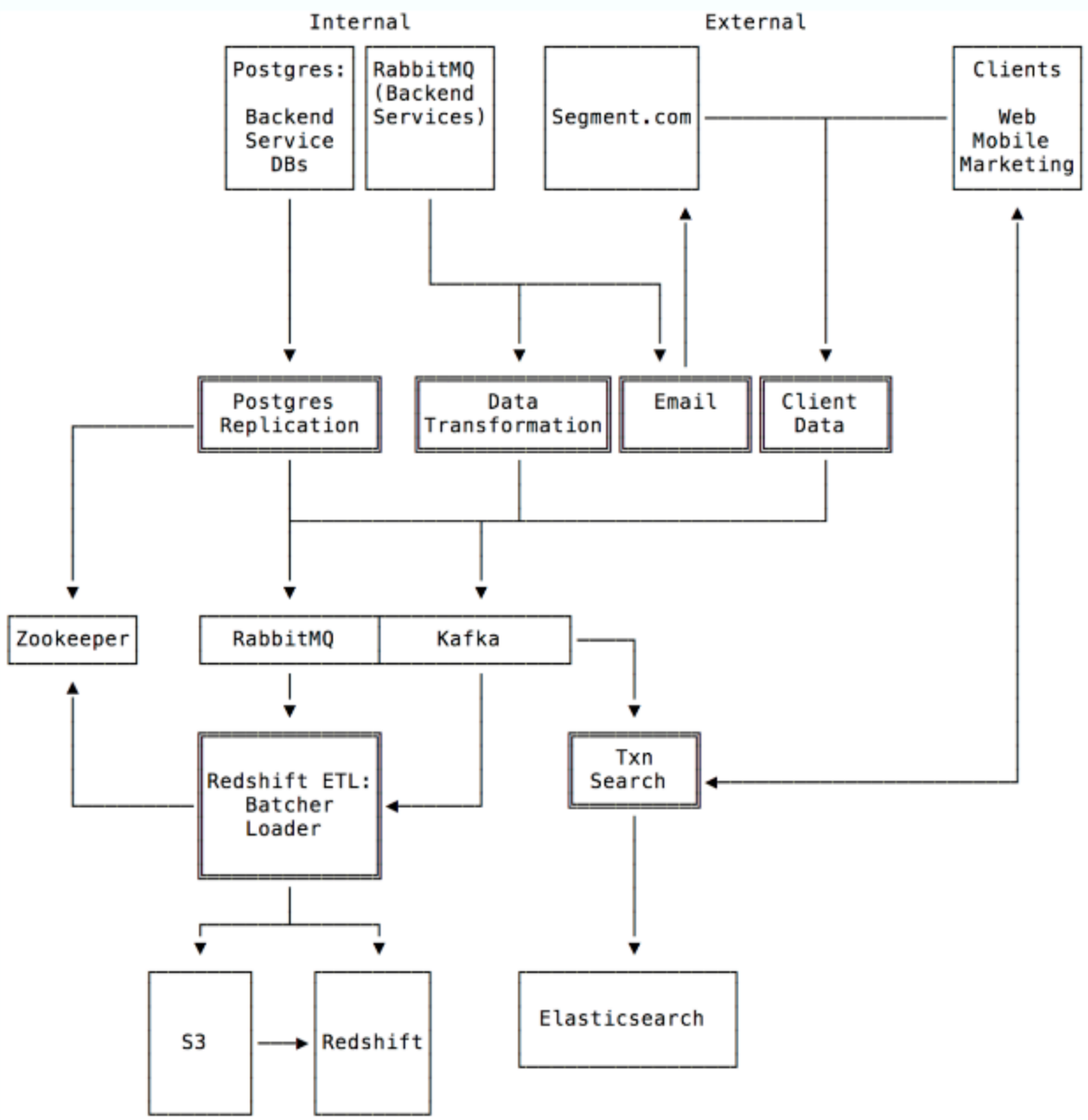
Membership ID	Full names	Physical Address	Salutation ID
1	Janet Jones	First Street Plot No 4	2
2	Robert Phil	3rd Street 34	1
3	Robert Phil	5th Avenue	1

3NF

1NF: table cells/cols unique type, each row unique **2NF**: single column primary key **3NF**: no transitive functional dependencies

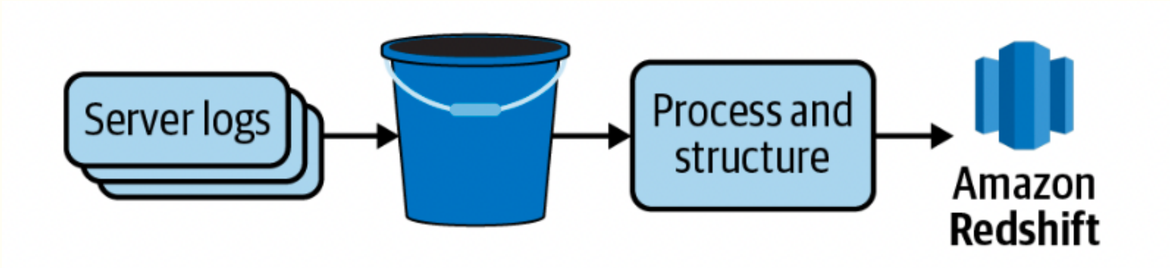
Univ.AI

Data Pipelines write data from OLTP systems and other sources into a OLAP system



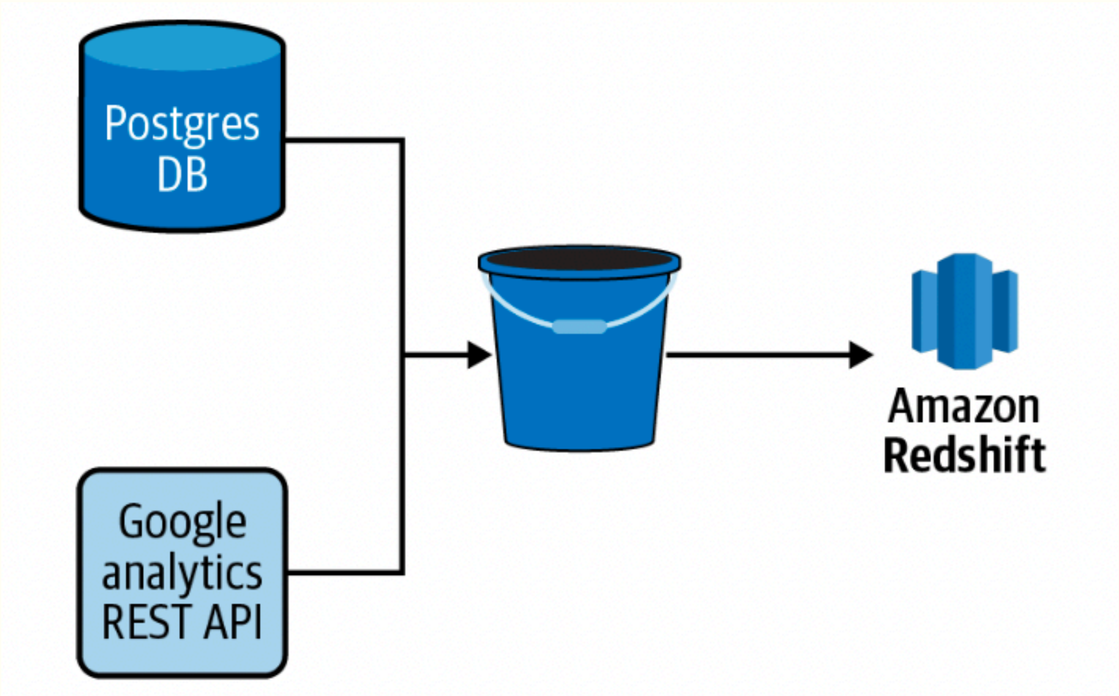
OLTP vs OLAP

- customer transaction Processing? ***vs*** analytics?
- small query size ***vs*** aggregates over large ones
- random writes from user input ***vs*** ordered stream
- end user (amazon site) ***vs*** analyst (you)
- GB to TB ***vs*** TB to PB



Where does data come from?

E-commerce databases
Logs
Customer tracking
Other sites



Where is your data?

Does not matter if you can get to it via SQL.