



La qualité des données

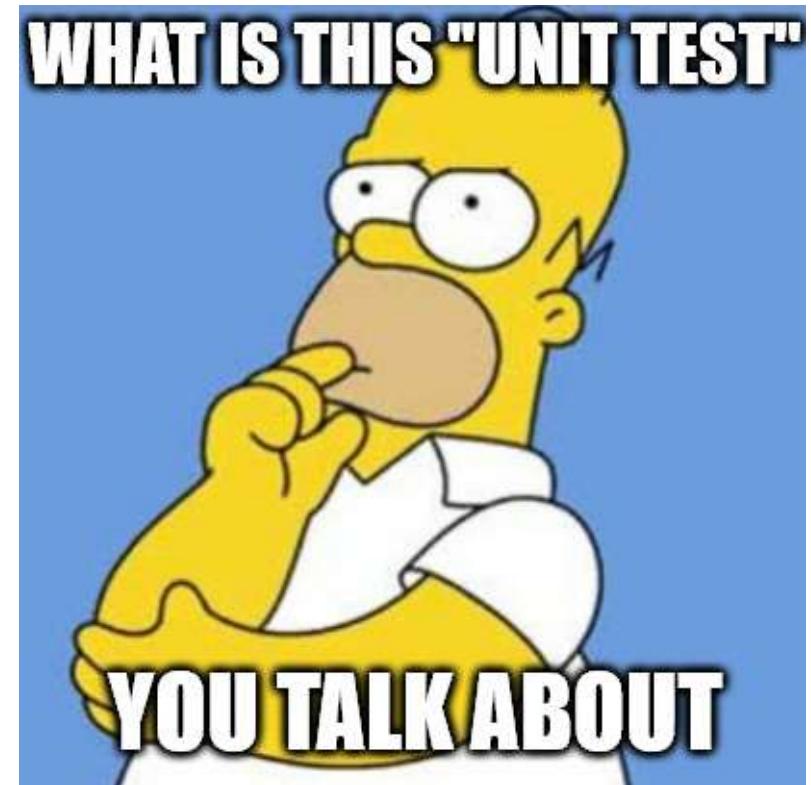
Un objectif atteignable ?

Sherif Behna
Architecte Logiciel et de Données
Hikari Data inc.

Les deux solitudes



Développement logiciel



BI / Analytics / Data Science

Qualité des données

Une (douce) introduction...

Les données : La ressource stratégique du 21e siècle ?

CIO

Data: The oil of the 21st Century

BrandPost • By Phil Dawson
Mar 25, 2022

Cloud Security IT Leadership

in X

Credit: metamorworks/tstock

If there was any doubt that the trend towards globalisation had decelerated, the emergence of COVID-19 has put that to rest. The shift towards more inward-looking policy settings around the globe, has been emboldened by the COVID-19 pandemic.

towards data science

ARTIFICIAL INTELLIGENCE

Is Data Really the New Oil in the 21st Century?

Exploring the strengths and limitations of this metaphor in the information age.

Amol Mavuduru
Dec 12, 2020 • 11 min read

Photo by Robin Sommer on Unsplash

Forbes

INNOVATION

The Strategic Value Of Data Infrastructure In R&D

By Akshay Talekar, Forbes Councils Member.
for [Forbes Technology Council](#), COUNCIL POST | Membership (fee-based)

Published Aug 11, 2025, 10:15am EDT

Share Save

Akshay Talekar, Lead Data Scientist at UL Research Institutes.

GETTY

Pourquoi on a besoin de données ?

- Pour prendre des décisions éclairées
- Pour trouver des solutions à un problème
- Pour mieux comprendre un phénomène
- Pour optimiser un processus

Utilisation des données



RAPPORTS



ANALYTICS

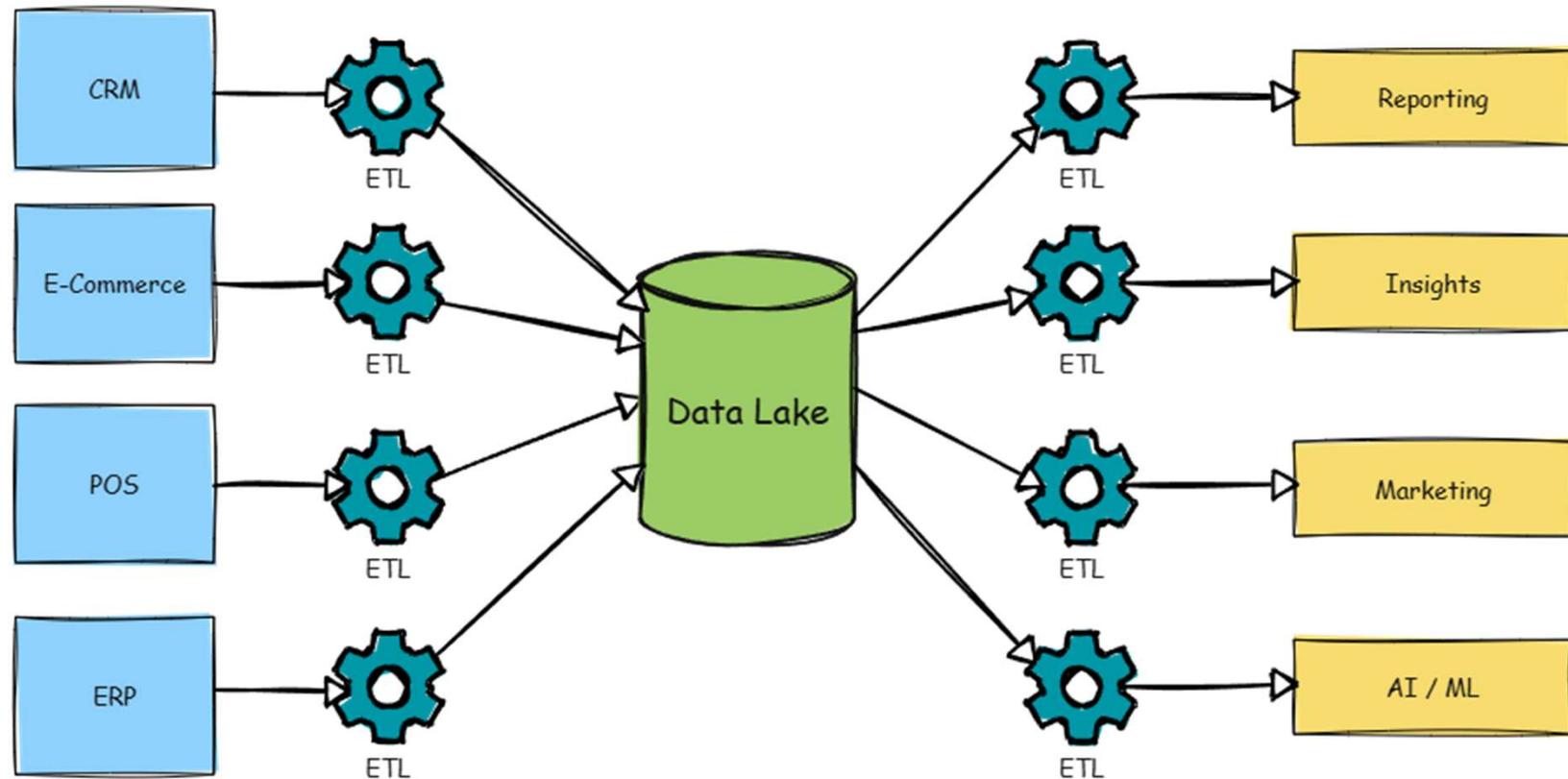


ALIMENTER UNE
AUTRE APPLICATION

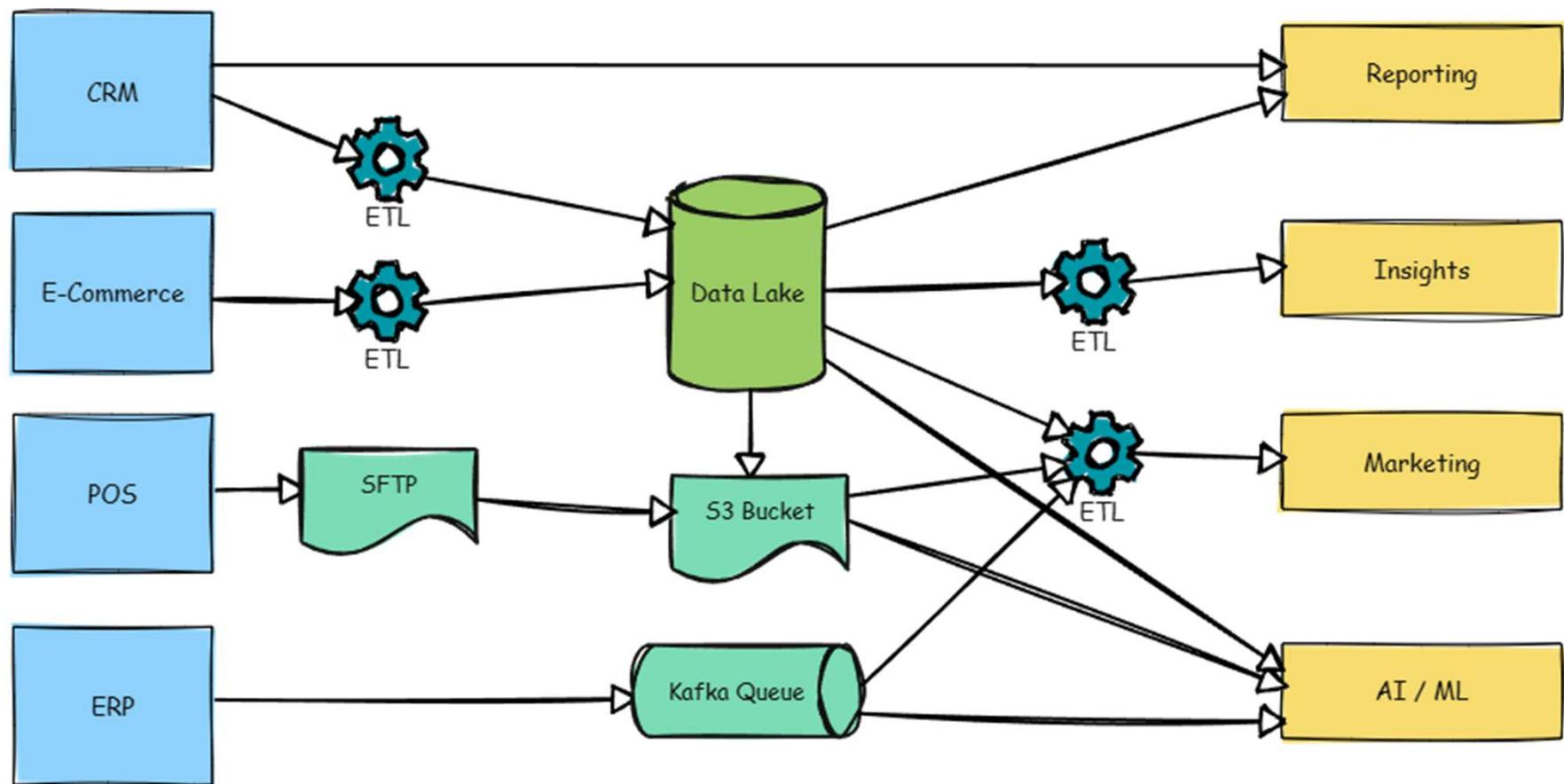


ML / AI

L'architecture de données idéalisée



La vraie vie...



Les problèmes de qualité de données

Jeu de données

pickup_date	pickup_time	dropoff_date	dropoff_time	psg_count	distance	rate_id	pu_loc_id	do_loc_id	fare
2025-01-30	18:55	2025-01-30	19:00	1	1,06	1	2	2	7,9
2025-01-30	18:55	2025-01-30	19:00	1	1,06	3	2	2	11,4
2025-03-23	21:32	2025-03-27	21:27	1	3,36	1	1	2	18,4
2025-04-07	13:14	2025-04-07	13:25	1	0,42	1	3	5	10
2025-04-10	10:28	2025-04-10	10:51	1	4,71		4	1	34,99
2025-04-12	18:40	2025-04-12	19:05	1	7,02	1	1	1	32,4
2025-05-06	16:11	2025-05-06	16:27	1	0		2	1	
2025-05-09	22:16	2025-05-09	15:58	1	2,07	2	6	2	14,2
05/17/2025	0h15	2025-05-17	00:20	1	140,32		1	1	13,79
2025-06-18	07:46	2025-06-18	07:58	1	1,77	1	2	3	13,5

Les problèmes de qualité de données

Format de données invalide

pickup_date	pickup_time	dropoff_date	dropoff_time	psg_count	distance	rate_id	pu_loc_id	do_loc_id	fare
2025-01-30	18:55	2025-01-30	19:00	1	1,06	1	2	2	7,9
2025-01-30	18:55	2025-01-30	19:00	1	1,06	3	2	2	11,4
2025-03-23	21:32	2025-03-27	21:27	1	3,36	1	1	2	18,4
2025-04-07	13:14	2025-04-07	13:25	1	0,42	1	3	5	10
2025-04-10	10:28	2025-04-10	10:51	1	4,71		4	1	34,99
2025-04-12	18:40	2025-04-12	19:05	1	7,02	1	1	1	32,4
2025-05-06	16:11	2025-05-06	16:27	1	0		2	1	
2025-05-09	22:16	2025-05-09	15:58	1	2,07	2	6	2	14,2
05/17/2025	0h15	2025-05-17	00:20	1	140,32		1	1	13,79
2025-06-18	07:46	2025-06-18	07:58	1	1,77	1	2	3	13,5

Les problèmes de qualité de données

Données incomplètes (valeurs manquantes)

pickup_date	pickup_time	dropoff_date	dropoff_time	psg_count	distance	rate_id	pu_loc_id	do_loc_id	fare
2025-01-30	18:55	2025-01-30	19:00	1	1,06	1	2	2	7,9
2025-01-30	18:55	2025-01-30	19:00	1	1,06	3	2	2	11,4
2025-03-23	21:32	2025-03-27	21:27	1	3,36	1	1	2	18,4
2025-04-07	13:14	2025-04-07	13:25	1	0,42	1	3	5	10
2025-04-10	10:28	2025-04-10	10:51	1	4,71		4	1	34,99
2025-04-12	18:40	2025-04-12	19:05	1	7,02	1	1	1	32,4
2025-05-06	16:11	2025-05-06	16:27	1	0		2	1	
2025-05-09	22:16	2025-05-09	15:58	1	2,07	2	6	2	14,2
05/17/2025	0h15	2025-05-17	00:20	1	140,32		1	1	13,79
2025-06-18	07:46	2025-06-18	07:58	1	1,77	1	2	3	13,5

Les problèmes de qualité de données

Données irréalistes

pickup_date	pickup_time	dropoff_date	dropoff_time	psg_count	distance	rate_id	pu_loc_id	do_loc_id	fare
2025-01-30	18:55	2025-01-30	19:00	1	1,06	1	2	2	7,9
2025-01-30	18:55	2025-01-30	19:00	1	1,06	3	2	2	11,4
2025-03-23	21:32	2025-03-27	21:27	1	3,36	1	1	2	18,4
2025-04-07	13:14	2025-04-07	13:25	1	0,42	1	3	5	10
2025-04-10	10:28	2025-04-10	10:51	1	4,71		4	1	34,99
2025-04-12	18:40	2025-04-12	19:05	1	7,02	1	1	1	32,4
2025-05-06	16:11	2025-05-06	16:27	1	0		2	1	
2025-05-09	22:16	2025-05-09	15:58	1	2,07	2	6	2	14,2
05/17/2025	0h15	2025-05-17	00:20	1	140,32		1	1	13,79
2025-06-18	07:46	2025-06-18	07:58	1	1,77	1	2	3	13,5

Les problèmes de qualité de données

Données incohérentes

pickup_date	pickup_time	dropoff_date	dropoff_time	psg_count	distance	rate_id	pu_loc_id	do_loc_id	fare
2025-01-30	18:55	2025-01-30	19:00	1	1,06	1	2	2	7,9
2025-01-30	18:55	2025-01-30	19:00	1	1,06	3	2	2	11,4
2025-03-23	21:32	2025-03-27	21:27	1	3,36	1	1	2	18,4
2025-04-07	13:14	2025-04-07	13:25	1	0,42	1	3	5	10
2025-04-10	10:28	2025-04-10	10:51	1	4,71		4	1	34,99
2025-04-12	18:40	2025-04-12	19:05	1	7,02	1	1	1	32,4
2025-05-06	16:11	2025-05-06	16:27	1	0		2	1	
2025-05-09	22:16	2025-05-09	15:58	1	2,07	2	6	2	14,2
05/17/2025	0h15	2025-05-17	00:20	1	140,32		1	1	13,79
2025-06-18	07:46	2025-06-18	07:58	1	1,77	1	2	3	13,5

Les problèmes de qualité de données

Données en double

pickup_date	pickup_time	dropoff_date	dropoff_time	psg_count	distance	rate_id	pu_loc_id	do_loc_id	fare
2025-01-30	18:55	2025-01-30	19:00	1	1,06	1	2	2	7,9
2025-01-30	18:55	2025-01-30	19:00	1	1,06	3	2	2	11,4
2025-03-23	21:32	2025-03-27	21:27	1	3,36	1	1	2	18,4
2025-04-07	13:14	2025-04-07	13:25	1	0,42	1	3	5	10
2025-04-10	10:28	2025-04-10	10:51	1	4,71		4	1	34,99
2025-04-12	18:40	2025-04-12	19:05	1	7,02	1	1	1	32,4
2025-05-06	16:11	2025-05-06	16:27	1	0		2	1	
2025-05-09	22:16	2025-05-09	15:58	1	2,07	2	6	2	14,2
05/17/2025	0h15	2025-05-17	00:20	1	140,32		1	1	13,79
2025-06-18	07:46	2025-06-18	07:58	1	1,77	1	2	3	13,5

Les problèmes de qualité de données

Données non crédibles

pickup_date	pickup_time	dropoff_date	dropoff_time	psg_count	distance	rate_id	pu_loc_id	do_loc_id	fare
2025-01-30	18:55	2025-01-30	19:00	1	1,06	1	2	2	7,9
2025-01-30	18:55	2025-01-30	19:00	1	1,06	3	2	2	11,4
2025-03-23	21:32	2025-03-27	21:27	1	3,36	1	1	2	18,4
2025-04-07	13:14	2025-04-07	13:25	1	0,42	1	3	5	10
2025-04-10	10:28	2025-04-10	10:51	1	4,71		4	1	34,99
2025-04-12	18:40	2025-04-12	19:05	1	7,02	1	1	1	32,4
2025-05-06	16:11	2025-05-06	16:27	1	0		2	1	
2025-05-09	22:16	2025-05-09	15:58	1	2,07	2	6	2	14,2
05/17/2025	0h15	2025-05-17	00:20	1	140,32		1	1	13,79
2025-06-18	07:46	2025-06-18	07:58	1	1,77	1	2	3	13,5

Les problèmes de qualité de données

Données non intégrées

pickup_date	pickup_time	dropoff_time	psg_count	distance	pu_loc_id	do_loc_id	fare
2025-01-30	18:55	19:00	1	1,06	2	2	7,9
2025-01-30	18:55	19:00	1	1,06	2	2	11,4
2025-03-23	21:32	21:27	1	3,36	1	2	18,4
2025-04-07	13:14	13:25	1	0,42	3	5	10
2025-04-10	10:28	10:51	1	4,71	4	1	34,99
2025-04-12	18:40	19:05	1	7,02	1	1	32,4
2025-05-06	16:11	16:27	1	0	2	1	
2025-05-09	22:16	15:58	1	2,07	6	2	14,2
05/17/2025	0h15	00:20	1	140,32	1	1	13,79
2025-06-18	07:46	07:58	1	1,77	2	3	13,5

location_id	name
1	Manhattan
2	Queens
3	Brooklyn
4	Bronx
5	Staten Island

Les problèmes de qualité de données

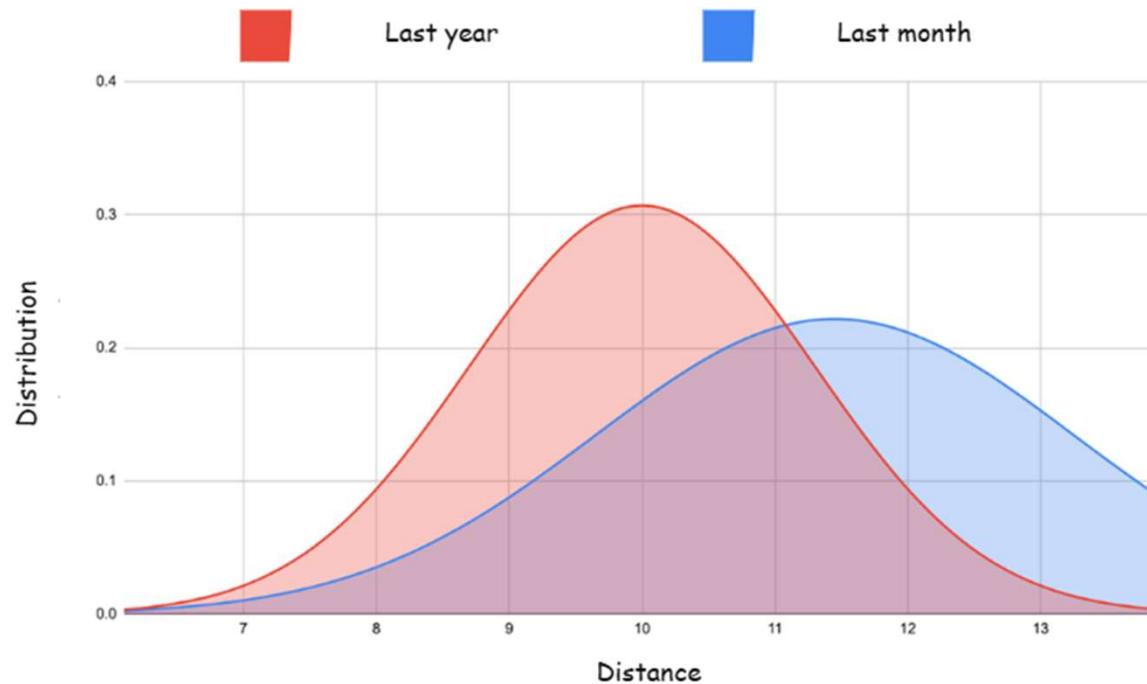
Données qui ne sont pas à jour (manque de fraîcheur)

pickup_date	pickup_time	dropoff_date	dropoff_time	psg_count	distance	rate_id	pu_loc_id	do_loc_id	pmt_type	fare
2025-01-30	18:55	2025-01-30	19:00	1	1,06	1	2	2	1	7,9
2025-01-30	18:55	2025-01-30	19:00	1	1,06	3	2	2	1	11,4
2025-03-23	21:32	2025-03-27	21:27	1	3,36	1	1	2	2	18,4
2025-04-07	13:14	2025-04-07	13:25	1	0,42	1	3	5	1	10
2025-04-10	10:28	2025-04-10	10:51	1	4,71		4	1	0	34,99
2025-04-12	18:40	2025-04-12	19:05	1	7,02	1	1	1	1	32,4
2025-05-06	16:11	2025-05-06	16:27	1	0		2	1	0	
2025-05-09	22:16	2025-05-09	15:58	1	2,07	2	6	2	1	14,2
05/17/2025	0h15	2025-05-17	00:20	1	140,32		1	1	0	13,79
2025-06-18	07:46	2025-06-18	07:58	1	1,77	1	2	3	1	13,5

Date actuelle : 2025-09-18

Les problèmes de qualité de données

Données biaisés (distribution)



Les problèmes de qualité de données

Données incompatibles

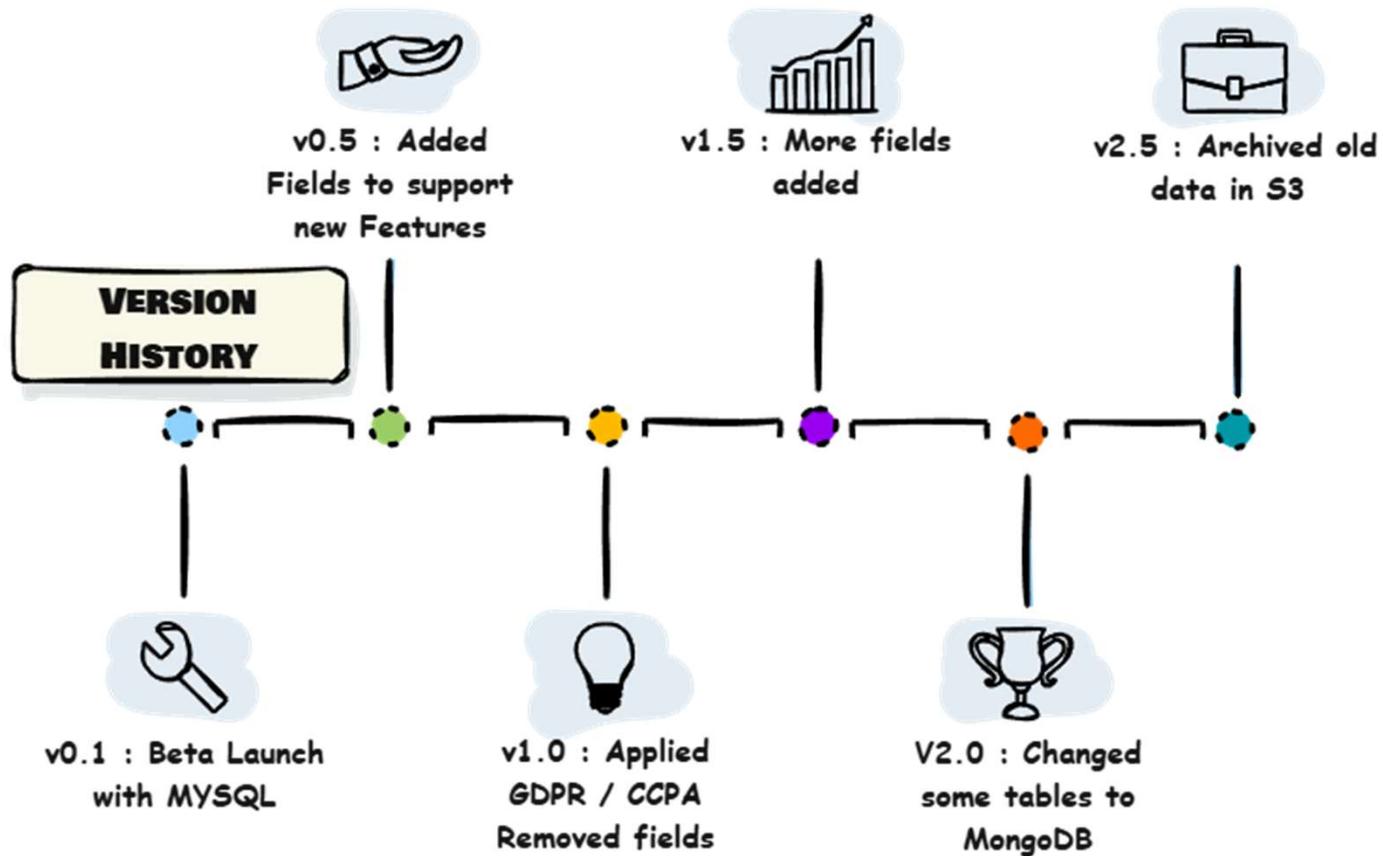
pickup_date	pickup_time	dropoff_date	dropoff_time	psg_count	distance	rate_id	pu_loc_id	do_loc_id	fare	toll
2025-01-30	18:55	2025-01-30	19:00	1	1,06		2	2	7,9	2,1
2025-01-30	18:55	2025-01-30	19:00	1	1,06		2	2	11,4	0
2025-03-23	21:32	2025-03-27	21:27	1	3,36		1	2	18,4	1,5
2025-04-07	13:14	2025-04-07	13:25	1	0,42		3	5	10	2,25
2025-04-10	10:28	2025-04-10	10:51	1	4,71		4	1	34,99	0
2025-04-12	18:40	2025-04-12	19:05	1	7,02		1	1	32,4	1,98
2025-05-06	16:11	2025-05-06	16:27	1	0		2	1		
2025-05-09	22:16	2025-05-09	15:58	1	2,07		6	2	14,2	2,25
05/17/2025	0h15	2025-05-17	00:20	1	140,32		1	1	13,79	0
2025-06-18	07:46	2025-06-18	07:58	1	1,77		2	3	13,5	3,5

Les problèmes de qualité de données

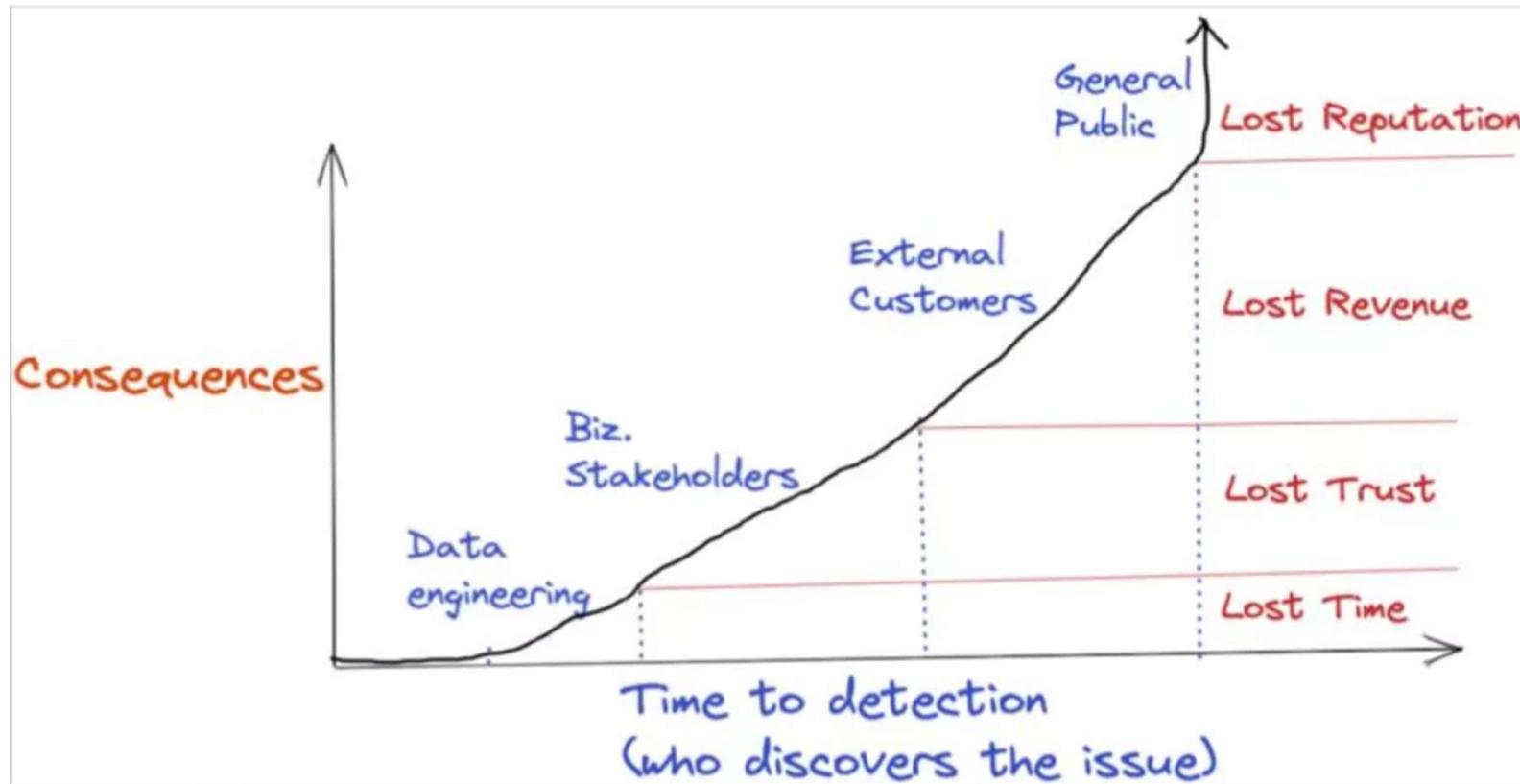
Fuite de renseignements personnels

pickup_date	pickup_time	dropoff_date	dropoff_time	psg_count	distance	pu_loc_id	do_loc_id	fare	comment
2025-01-30	18:55	2025-01-30	19:00	1	1,06	2	2	7,9	Premium customer
2025-01-30	18:55	2025-01-30	19:00	1	1,06	2	2	11,4	
2025-03-23	21:32	2025-03-27	21:27	1	3,36	1	2	18,4	
2025-04-07	13:14	2025-04-07	13:25	1	0,42	3	5	10	Drop off at 20 W. 34 th st.
2025-04-10	10:28	2025-04-10	10:51	1	4,71	4	1	34,99	
2025-04-12	18:40	2025-04-12	19:05	1	7,02	1	1	32,4	
2025-05-06	16:11	2025-05-06	16:27	1	0	2	1		
2025-05-09	22:16	2025-05-09	15:58	1	2,07	6	2	14,2	Zip code : 10001
05/17/2025	0h15	2025-05-17	00:20	1	140,32	1	1	13,79	
2025-06-18	07:46	2025-06-18	07:58	1	1,77	2	3	13,5	

Qualité variable selon le temps



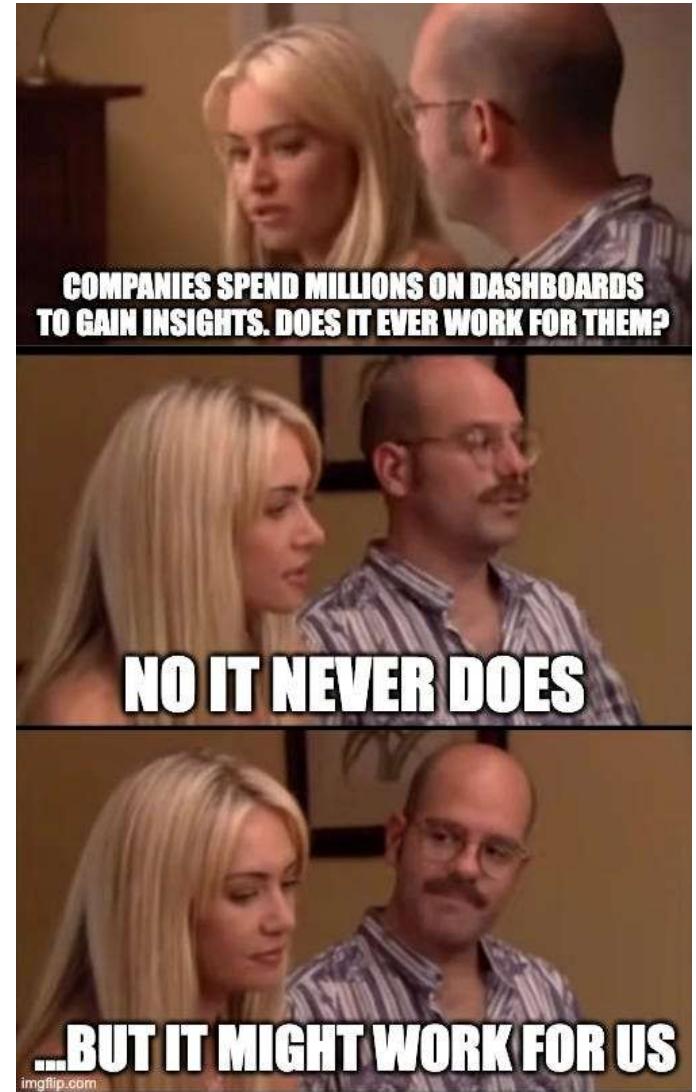
Pourquoi c'est important



Les projets d'exploitation de données sont voués à l'échec ?

Taux élevé d'échec dans les projets de données

60% des projets seraient voués à l'échec (Gartner - 2015)



Les conséquences...

Zillow's home-buying debacle shows how hard it is to use AI to value real estate



By [Rachel Metz](#), CNN Business

⌚ 7 minute read · Published 7:32 AM EST, Tue November 9, 2021

The decision, [announced last week](#), marks a stunning defeat for Zillow. The real estate listing company took a \$304 million inventory write-down in the third quarter, which it blamed on having recently purchased homes for prices that are higher than it thinks it can sell them. The company saw its stock plunge and it now plans to cut 2,000 jobs, or 25% of its staff.

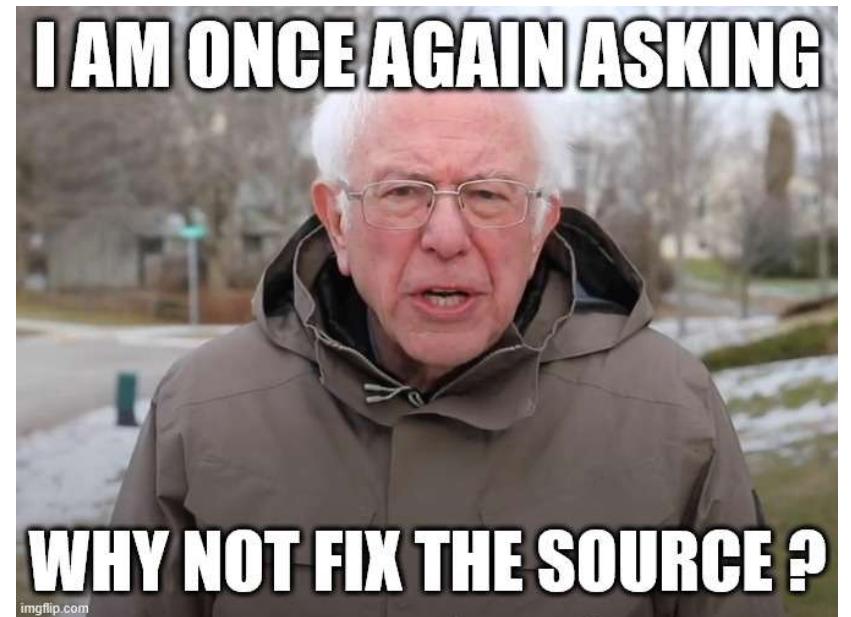
<https://edition.cnn.com/2021/11/09/tech/zillow-ibuying-home-zestimate/index.html>

Causes profondes

- Les données en tant qu'externalité (« byproduct »)
- Les producteurs et consommateurs travaillent en silo
- Les attentes envers les données ne sont pas explicites
- Manque d'approches pro-actives pour assurer la qualité des données
- Pas de responsabilité sur les données produites (manque de « ownership »)

Pourquoi ne pas tout régler à la source ?

- Oui, idéalement, c'est ce qu'on veut... mais comment faire ?
- La qualité est une question d'attentes...
 - Les attentes **varient** en fonction du **use case**



Data burn out ?

in Search

All / Business Administration / Data Analysis

What do you do if your team members are experiencing burnout?

Powered by AI and the LinkedIn community

1 [Recognize Signs](#)

SeattleDataGuy's Newsletter

Beyond The Hype: Data Teams Have A Burnout Problem

And Five Tactics On How You Can Keep It At Bay

SEATTLEDATAGUY

FEB 21, 2025



DATA ENGINEERING

How to Avoid Burnout as a Data Engineer

By: Chris Garzon | March 6, 2025 | 13 mins read

Pistes de solution

Profilage des données
Exploration des données (EDA)

Exploration / Profilage des données

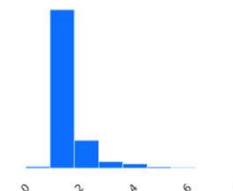
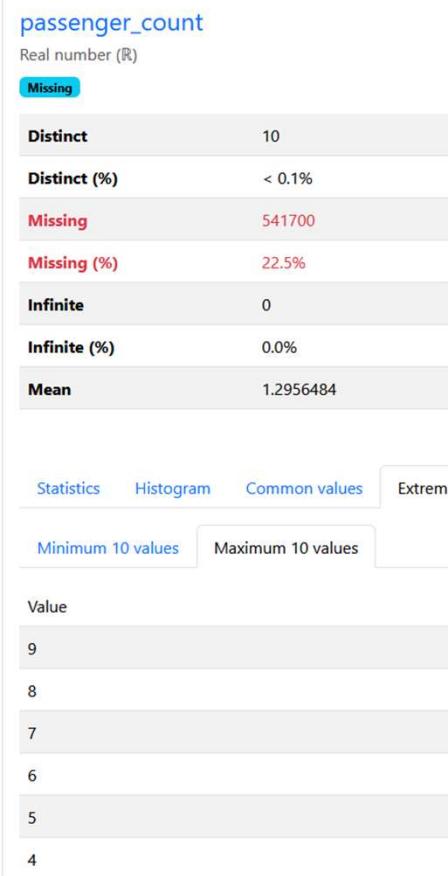
Overview	Alerts 17	Reproduction
Dataset statistics		
Number of variables	20	Variable types
Number of observations	2409114	Categorical 5
Missing cells	2708500	DateTime 2
Missing cells (%)	5.6%	Numeric 12
Duplicate rows	0	Boolean 1
Duplicate rows (%)	0.0%	
Total size in memory	358.4 MiB	
Average record size in memory	156.0 B	



Exploration / Profilage des données



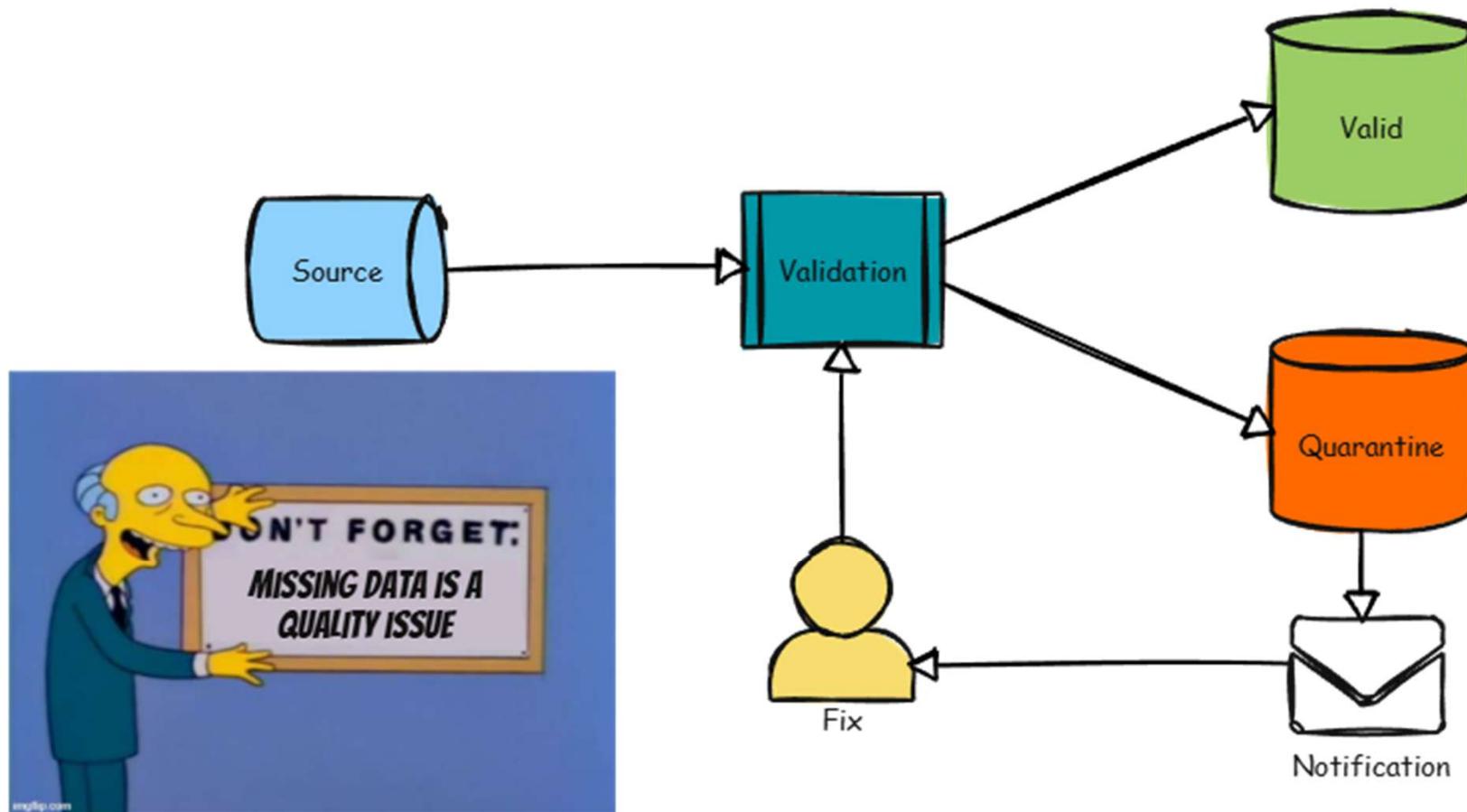
Exploration / Profilage des données



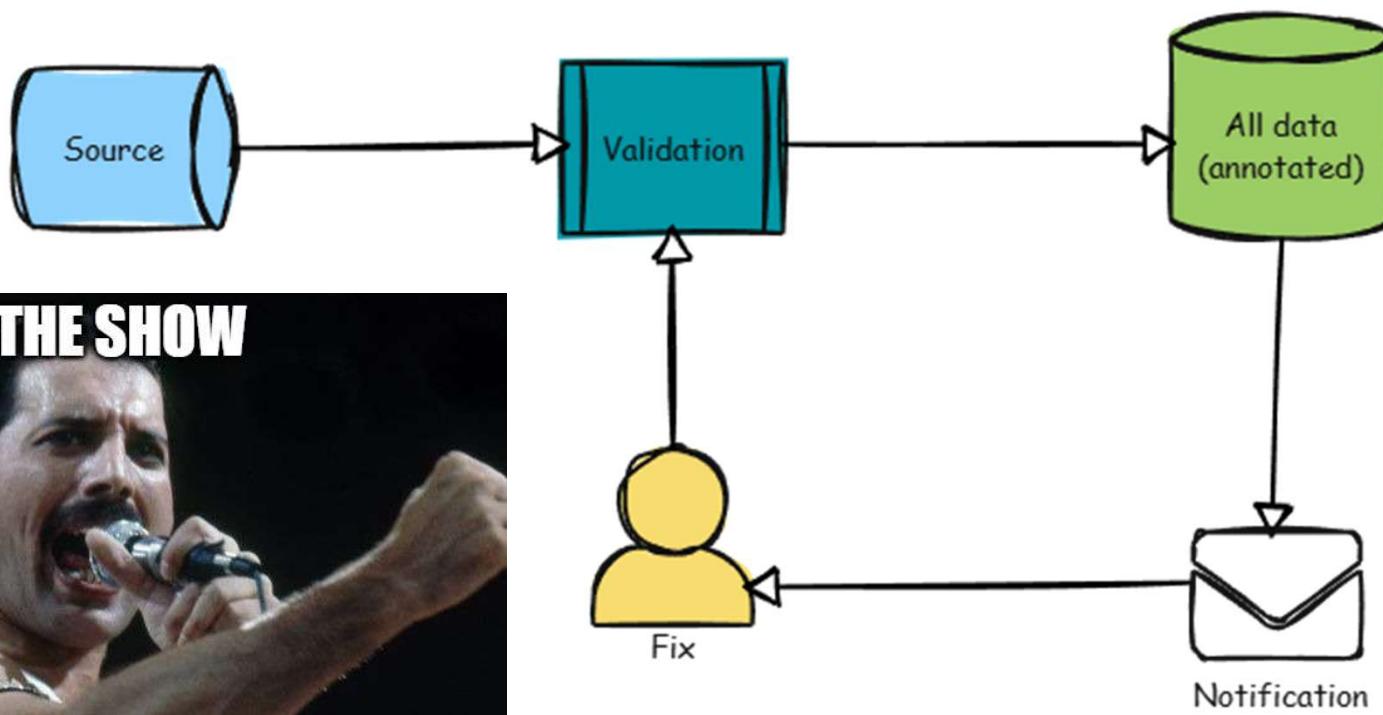
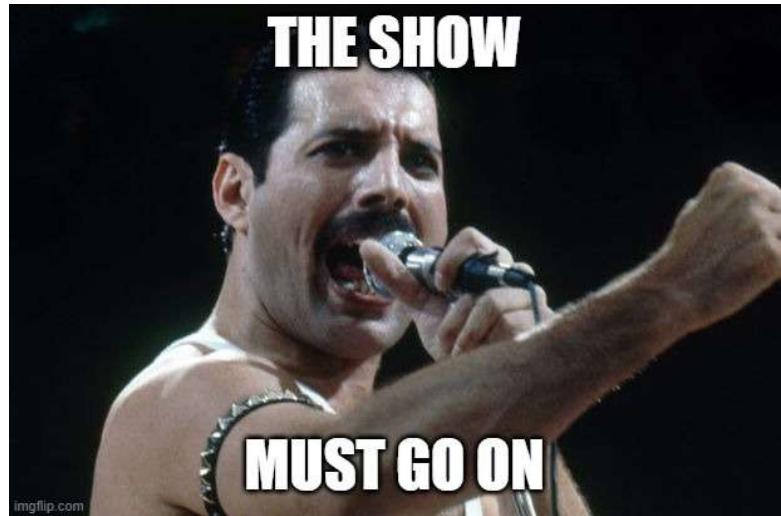
Pistes de solution

Data Quality Checks

Data Quality Checks



Data Quality Checks – Annotator



Data Quality Checks - DQX

```
# passenger_count checks
DQRowRule(
    criticality="error",
    check_func=is_not_null,
    column="passenger_count",
    name="passenger_count_is_null"
),
DQRowRule(
    criticality="warn",
    check_func=is_in_range,
    column="passenger_count",
    check_func_kwargs={"min_limit": 1, "max_limit": 4},
    name="passenger_count_is_in_range"
),
```



DQX - Data Quality Framework

Data Quality Checks - DQX

```
# dataset rules
DQDatasetRule(
    criticality="warn",
    check_func=is_unique,
    columns=[ "tpep_pickup_datetime", "PULocationID"],
    name="uniqueness"
),
DQDatasetRule(
    criticality="error",
    check_func=foreign_key,
    columns=[ "PULocationID"],
    check_func_kwargs={
        "ref_columns": [ "LocationID"],
        "ref_df_name": "zone",
    },
    name="PULocationID_foreign_key"
),
```



DQX - Data Quality Framework

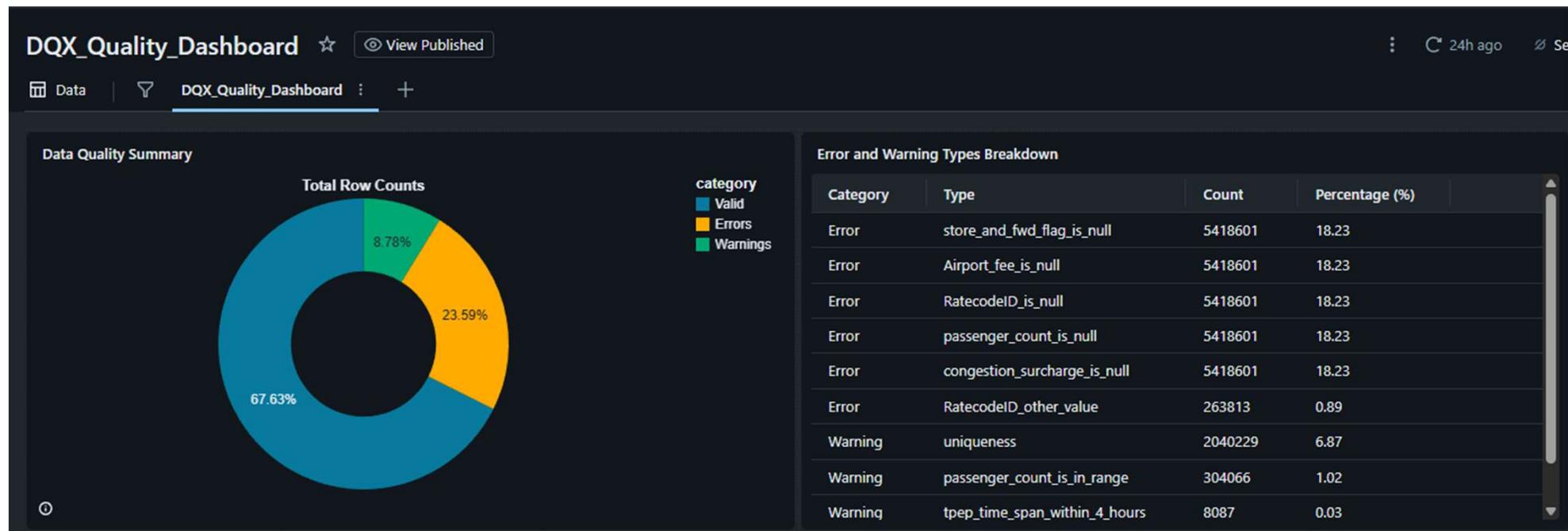
Data Quality Checks - DQX

```
└── _errors
    └── array
        ├── 0: {"name": "passenger_count_is_null", "message": "Column 'passenger_count' value is null", "columns": ["passenger_count"], "filter": null, "function": "is_not_null", "run_time": "2025-09-01T14:40:27.512Z", "user_metadata": {}}
        ├── 1: {"name": "RatecodeID_is_null", "message": "Column 'RatecodeID' value is null", "columns": ["RatecodeID"], "filter": null, "function": "is_not_null", "run_time": "2025-09-01T14:40:27.512Z", "user_metadata": {}}
        ├── 2: {"name": "store_and_fwd_flag_is_null", "message": "Column 'store_and_fwd_flag' value is null", "columns": ["store_and_fwd_flag"], "filter": null, "function": "is_not_null", "run_time": "2025-09-01T14:40:27.512Z", "user_metadata": {}}
        ├── 3: {"name": "congestion_surcharge_is_null", "message": "Column 'congestion_surcharge' value is null", "columns": ["congestion_surcharge"], "filter": null, "function": "is_not_null", "run_time": "2025-09-01T14:40:27.512Z", "user_metadata": {}}
        └── 4: {"name": "Airport_fee_is_null", "message": "Column 'Airport_fee' value is null", "columns": ["Airport_fee"], "filter": null, "function": "is_not_null", "run_time": "2025-09-01T14:40:27.512Z", "user_metadata": {}}
```



DQX – Data Quality Framework

Data Quality Checks - DQX



DQX - Data Quality Framework

Data Quality Checks - Soda

yellow_trip

tic

Add Check Last scan - Yesterday, 8:05 PM ::

Checks Metric Monitors Agreements Columns Sample Data Incidents

Last updated - Today, 2:00 PM

Check Coverage ①

14

Health ① 71% 0%

Incidents ①

Sun Mon Tue Wed Thu Fri Sat

Jul 25 Aug 25 Sep 25

Search RESULTS: 14

CHECK	VALUE	LAST RESULTS	ORIGIN	LAST EVALUATED	INCIDENT	AGREEMENT
! passenger_count_is_null	6,457,356	□ ✗ ✗ ✗ ✗ ✗	cloud	about 18 hours ago ⑦	-	-
! passenger_count_is_in_range	344,630	□ ✗ ✅ ✅ ✅ ✅	cloud	about 18 hours ago ⑦	-	-
! Duplicates	1,157,938	███████████	cloud	about 18 hours ago ⑦	-	-
! trip_distance_is_in_range	9,231	□ ██████████	cloud	about 18 hours ago ⑦	-	-
✓ tpep_dropoff_datetime_not_in_future	0	□ ✅ ✅ ✅ ✅ ✅	cloud	about 18 hours ago ⑦	-	-

SODA 

Data Quality Checks - Soda

CHECK HISTORY

passenger_count_is_in_range

Last scan - abc

Last 30 days

Result Failed Rows Analysis

Failed Rows (Sample of 100 rows)

VENDORID	TPEP_PICKUP_DATETIME	TPEP_DROPOFF_DATETIME	PASSENGER_COUNT	TRIP_DISTANCE	RATECODEID	STORE_AND_FWD_FLAG	PULOCATIONID	DLOCATIONID
1	2025-05-01 00:18:14	2025-05-01 00:27:38	0	1.5	1	N	140	263
...

DATASET **tic / yellow_trip**

CHECK NAME **passenger_count_is_in_range** ADD TO SCAN DEFINITION **tic Default Scan** Sunday (9/7/25) at 20:00

Filters

Define Valid Values

COLUMN **passenger_count**

VALIDITY RULE VALUE

max 4

VALIDITY RULE VALUE

min 1

+ Add Validity Rule

Alerts

ALERT LEVEL Warn when the number of invalid values

WARN CONDITION VALUE VALUE TYPE

greater than 0 Absolute

Attributes

[View SodaCL](#) [Test Check](#) [Save](#)

SODA

Data Quality Checks - Soda

INCIDENT

INC-1: passenger_count is null for many new rows

Save ::

Reported Investigating Fixing Resolved

TITLE: passenger_count is null for many new rows

SEVERITY: Major

STATUS: Reported LEAD: Sheriff Behna REPORTED BY: Sheriff Behna

DESCRIPTION: Scans since 09/02 are failing because of this check.

CHECK: passenger_count_is_in_range

CHECK RESULTS	VALUE	TIME	INCIDENT
<input type="checkbox"/> !	344,630	Yesterday, 8:05 PM	
<input type="checkbox"/> !	304,066	Sep 5, 8:05 PM	
<input type="checkbox"/> !	304,066	Sep 4, 8:07 PM	

Duration: 4 days and 19 hours
Reported: Sep 2, 7:05 PM
Last update: Sep 2, 7:05 PM

Integrations: NO LINKED INTEGRATIONS.

Relates to:

- Check: [passenger_count_is_in_range](#)
- Agreement: -
- Dataset: [yellow_trip](#)
- Data Source: [tlc](#)



Data Quality Checks - Soda

DISCUSSION

#1 - Validate payment_type

SB Sherif Behna 5 days ago (edited)

The payment_type field should be validated to make sure it is within valid values. I believe the valid values are 0 to 6 inclusively, but this needs deeper analysis.

SB Sherif Behna PROPOSAL 5 days ago (edited)

CHECK	VALUE	RELATES TO
payment_type_other_value	0	yellow_trip tic / yellow_trip

Review

Details

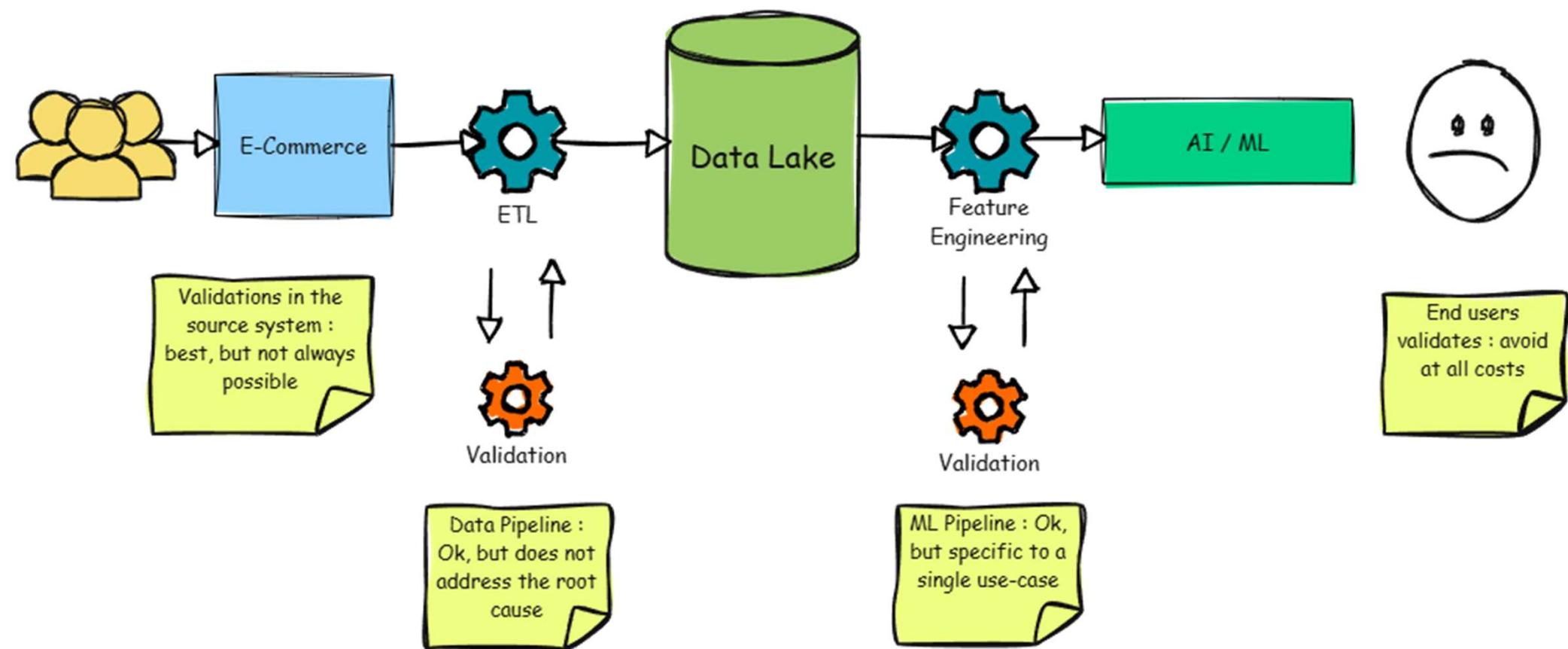
Dataset	yellow_trip tic / yellow_trip
Created By	Sherif Behna
Created	5 days ago
Last update	5 days ago

Add a comment...

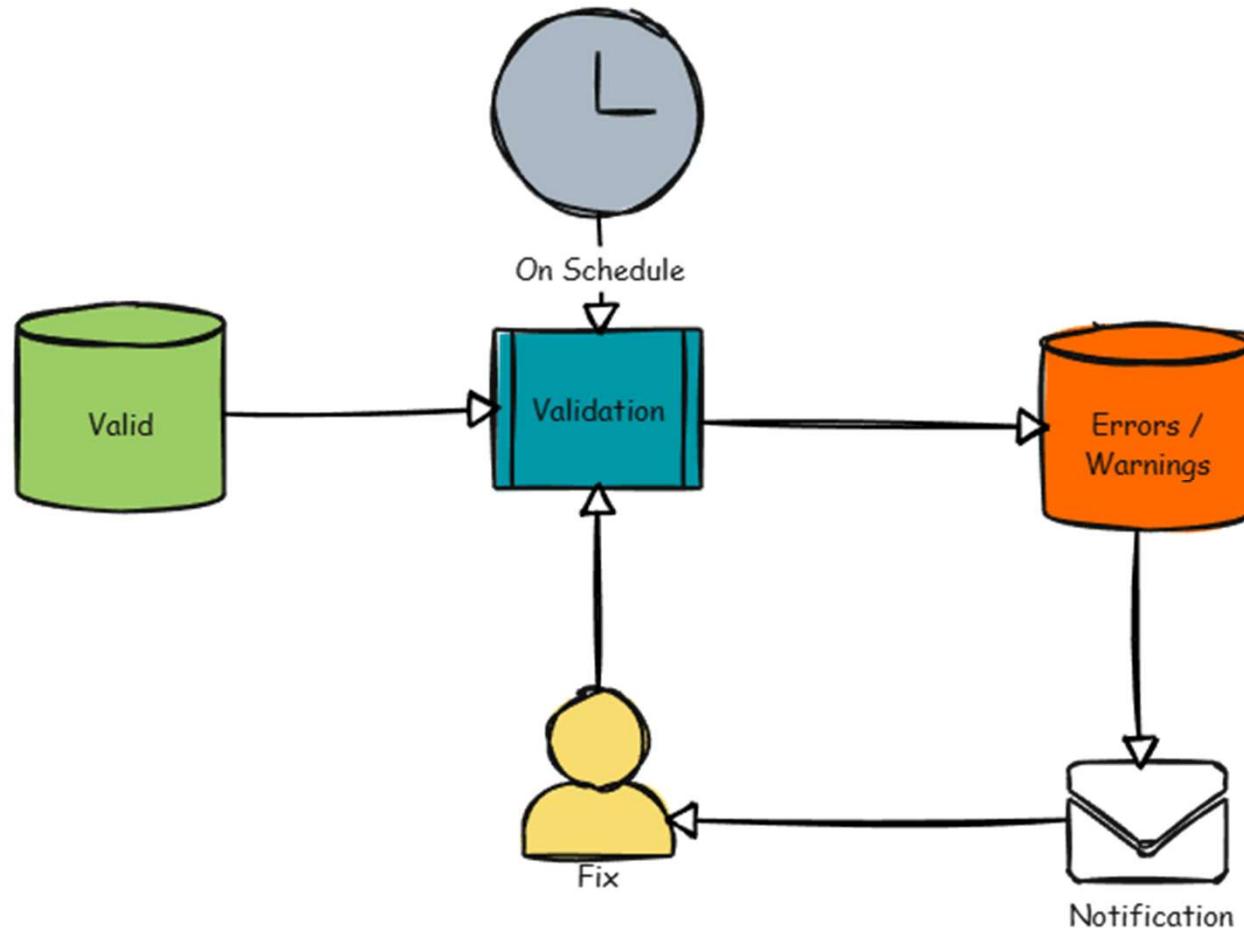
Propose Check Post



Data Quality Checks – Quand les appliquer ?



Data Quality Checks – Cédule



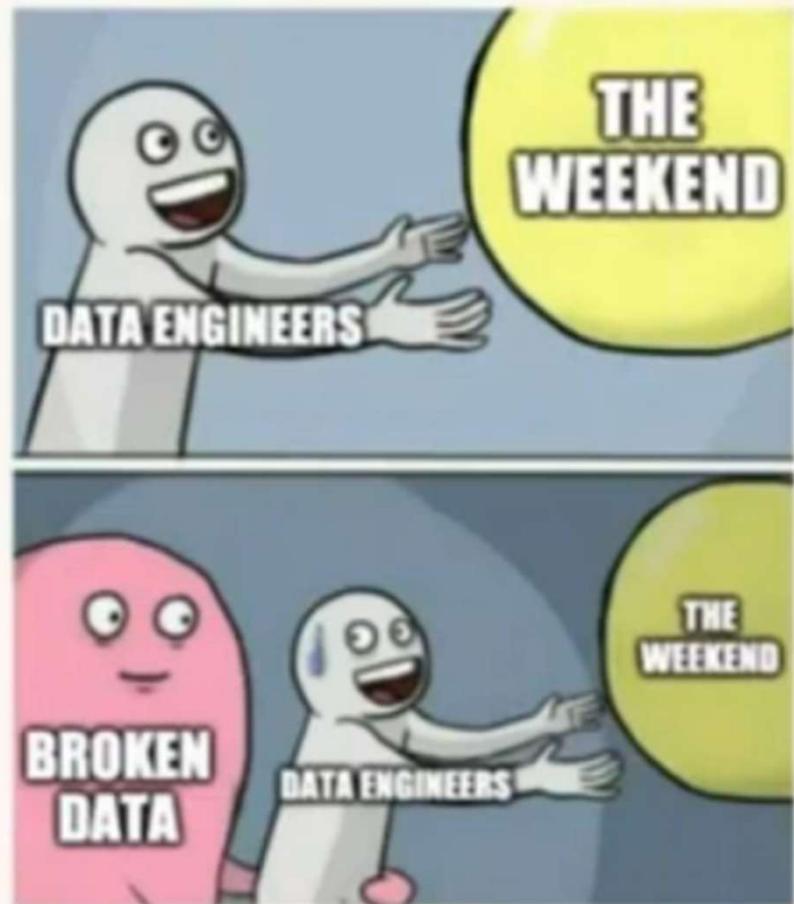
Pistes de solution

Data observability

Data observability

Data downtime

Périodes où les données sont partielles, erronées, manquantes, ou inexactes

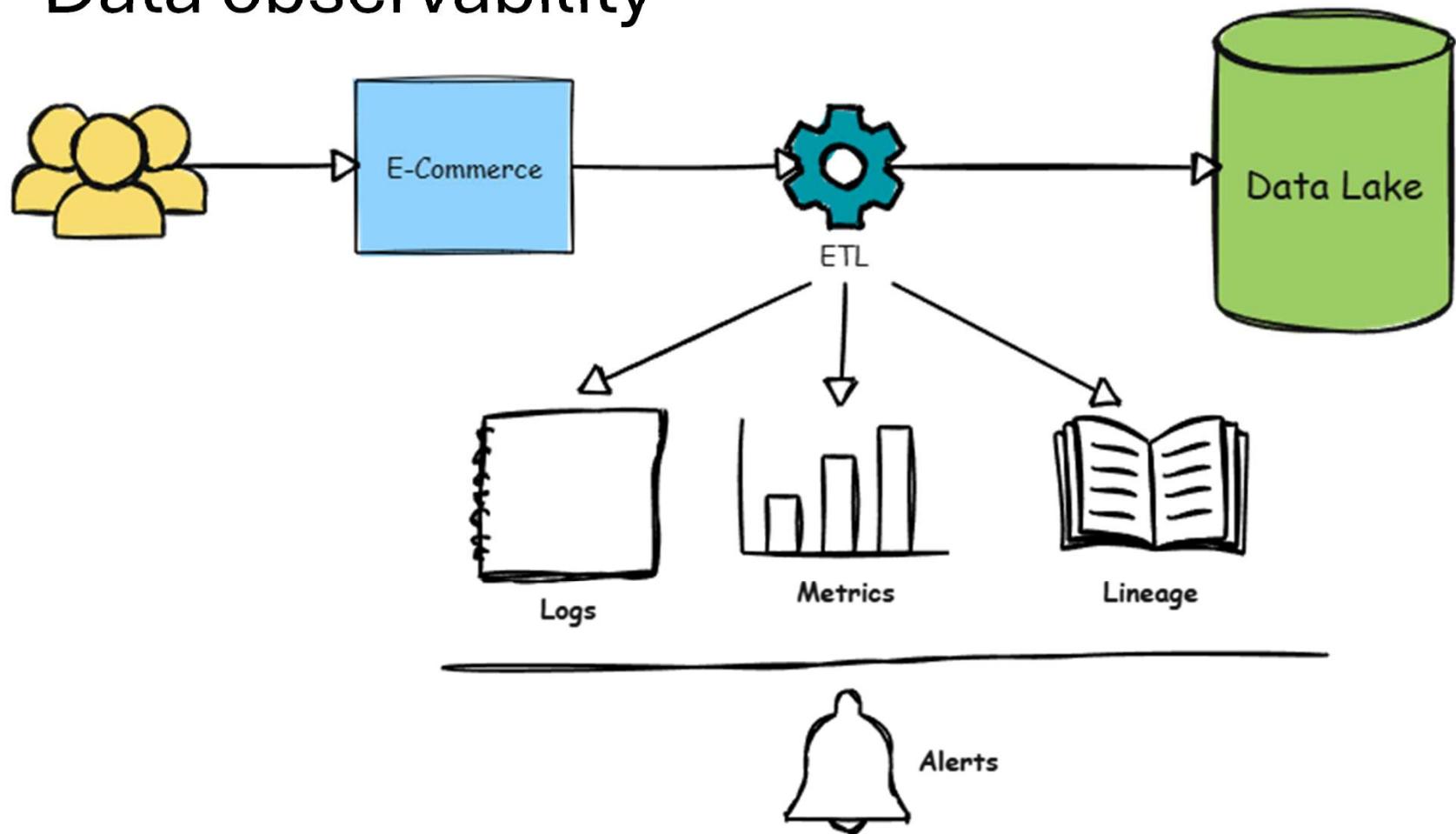


Cinq piliers de l'observabilité

- Freshness
- Quality / Distribution
- Volume
- Schema
- Lineage



Data observability



Data observability - Logs

```
print(f"transforming bronze -> silver. Got {num_rows} rows.") # ✘
```

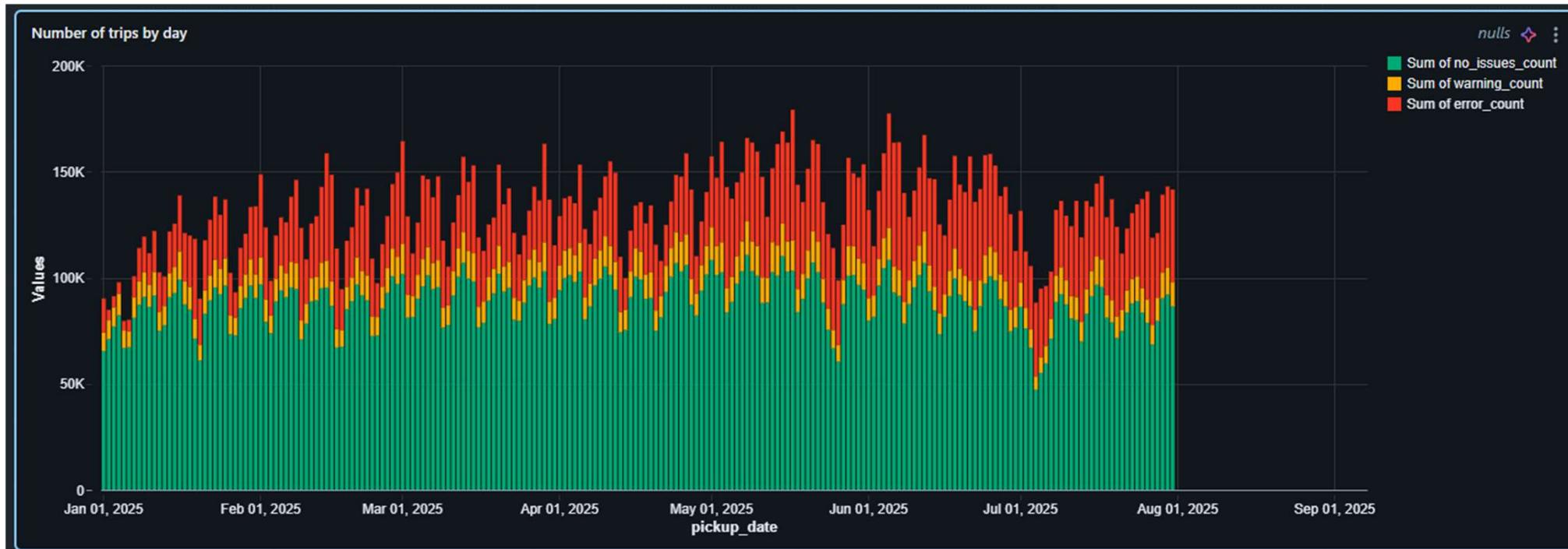
```
logger.debug("This is a debug message")
logger.info(f"This is an info message. x = {x}")
logger.warning("This is a warning message")

# Handle exceptions with logging.
try:
    raise RuntimeError("This is a runtime error")
except RuntimeError:
    logger.error("This is an error message", exc_info=True)
```

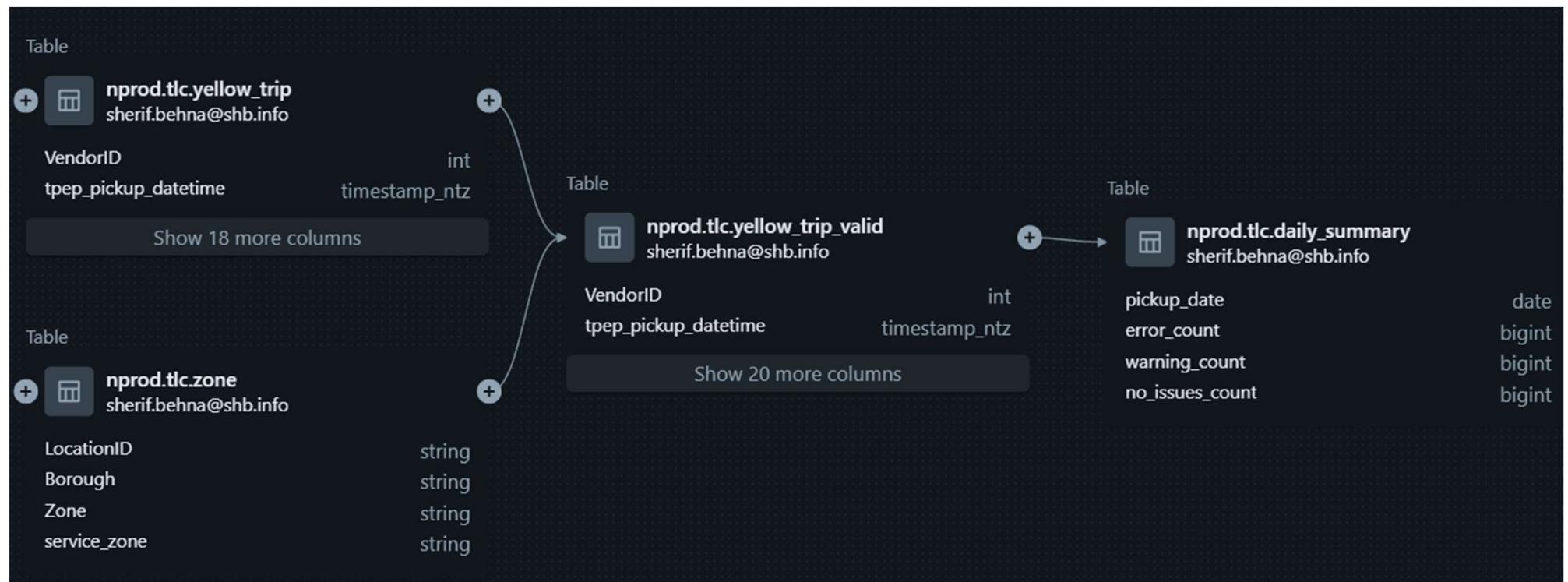
Data observability - Logs

Search Message and Job ID...				Source	Cluster
Timestamp	Level	Message			
2025-06-12 18:15:22.052	INFO	▶ Cleaning the wrapper ReplInfo(driverReplId=ReplId-19766-1bd8e-a, chauffeurReplId=ReplId-19766-1bd8e-a, executionCont...	log4j	JOB	
2025-06-12 18:15:21.430	INFO	Current cluster load: 0, Old Ema: 0.85, New Ema: 0.0	log4j	JOB	
2025-06-12 18:15:18.830	INFO	Received SAFEr configs with version 1749764676494	log4j	JOB	
2025-06-12 18:15:18.430	INFO	Current cluster load: 0, Old Ema: 1.0, New Ema: 0.85	log4j	JOB	
2025-06-12 18:15:18.002	INFO	Removed result fetcher for 1749764790506_7260688975682456956_f062332a-4f1e-43a7-be4c-7acab60d9f3f	log4j	JOB	
2025-06-12 18:15:17.941	WARN	The context seems empty, likely executing non-notebook command, not setting sys.path	log4j	JOB	
2025-06-12 18:15:17.929	INFO	Added result fetcher for 1749764790506_7260688975682456956_f062332a-4f1e-43a7-be4c-7acab60d9f3f	log4j	JOB	
2025-06-12 18:15:17.902	INFO	▶ Removed result fetcher for 1749764790506_8070518749520008224_job-65084945506620-run-127266188482036-acti...	log4j	JOB	
2025-06-12 18:15:17.781	INFO	▶ Added result fetcher for 1749764790506_8070518749520008224_job-65084945506620-run-127266188482036-action...	log4j	JOB	
2025-06-12 18:15:17.498	INFO	▶ Removed result fetcher for 1749764790506_5458236457776631310_job-65084945506620-run-127266188482036-acti...	log4j	JOB	
2025-06-12 18:15:17.172	INFO	Code generated in 36.096927 ms	log4j	JOB	
<	1	2	3	4	5
	...		4000	>	

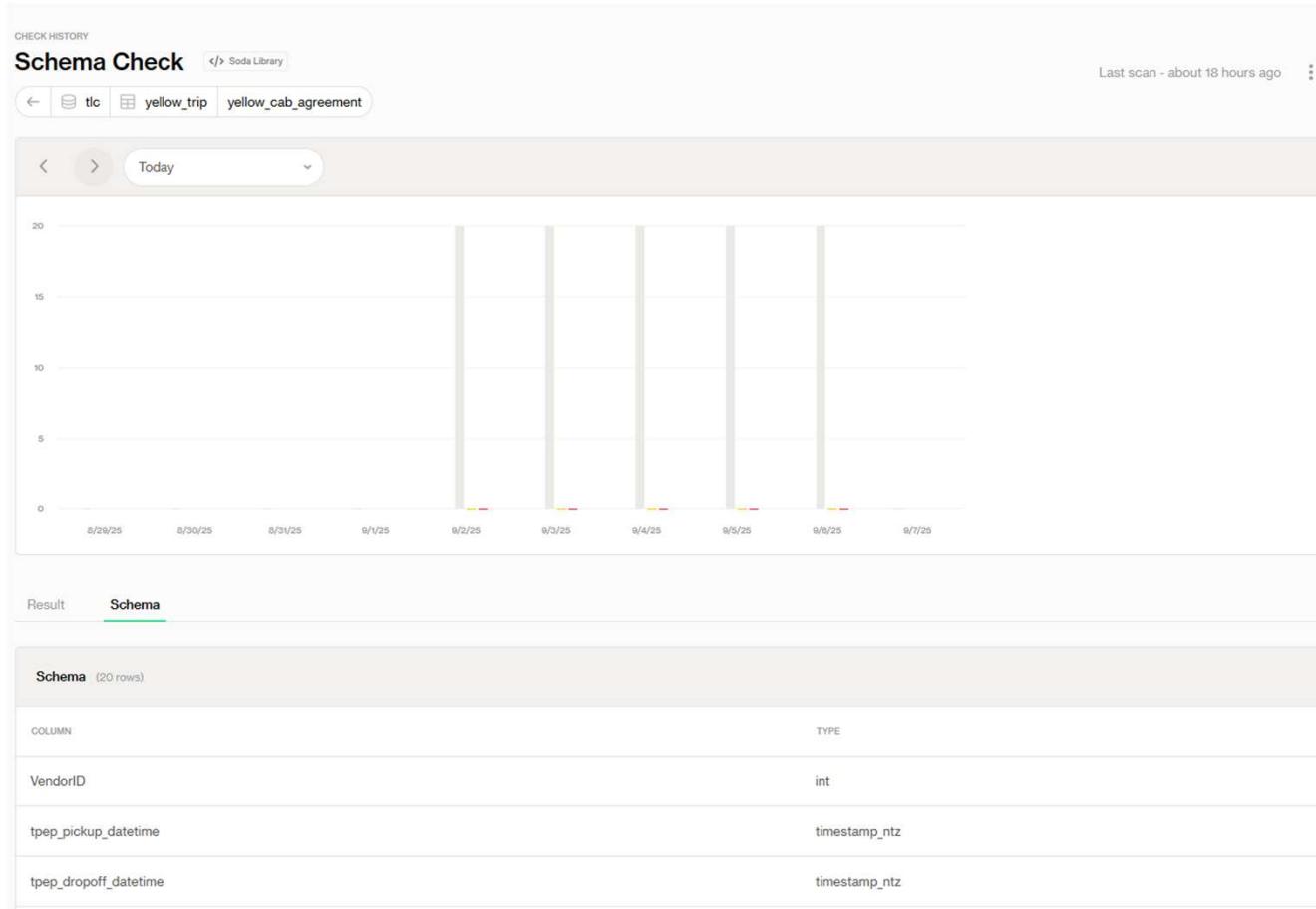
Data observability - Metrics



Data observability - Lineage



Data Observability – Schema check



SODA

Pistes de solution

Data contracts / Data products

Il y a encore un problème...

Comment codifier /
communiquer les attentes
d'une équipe à l'autre ?



Data Contract / Data Product

Data Product : NYC Trip Data v1.1.0

Schema

- pickup_dt: datetime
- dropoff_dt: datetime
- psg_count: int
- distance: float
- rate_id: int (references rates)
- pu_loc_id: int (references locations)
- do_loc_id: int (references locations)
- pmt_type: int (references payments)

Quality

- pickup_dt < dropoff_dt
- psg_count > 0
- etc...

SLA

- Updated : every day at 3 AM
- Availability : 99,5%
- Retention : 7 years

Ports

- Databricks SQL : jdbc:databricks://dbc-....
- S3 : [https://s3.amazonaws.com/...](https://s3.amazonaws.com/)
- Web API : <https://acme.com/api/tlc/v1/trips>

Team

- DPO : jane.doe@acme.com
- Support : tlc_support@acme.com
- Slack : #tlc_support

Data Contracts - ODCS

```
schema:
  - name: nyc_yellow_taxi_trip
    logicalType: object
    physicalType: table
    description: NYC Yellow Taxi Trip Records
    authoritativeDefinitions:
      - url: https://www.nyc.gov/assets/tlc/downloads/pdf/data\_dictionary\_trip\_records\_yellow.pdf
        type: businessDefinition
    properties:
      - name: tpep_pickup_datetime
        logicalType: date
        description: Date and time when the meter was engaged.
      - name: tpep_dropoff_datetime
        logicalType: date
        description: Date and time when the meter was disengaged.
      - name: passenger_count
        logicalType: integer
        description: Number of passengers in the vehicle.
      - name: trip_distance
        logicalType: number
        description: Trip distance in miles.
      - name: RatecodeID
        logicalType: integer
        description: Rate code (e.g., Standard, JFK, Newark, etc.)
```

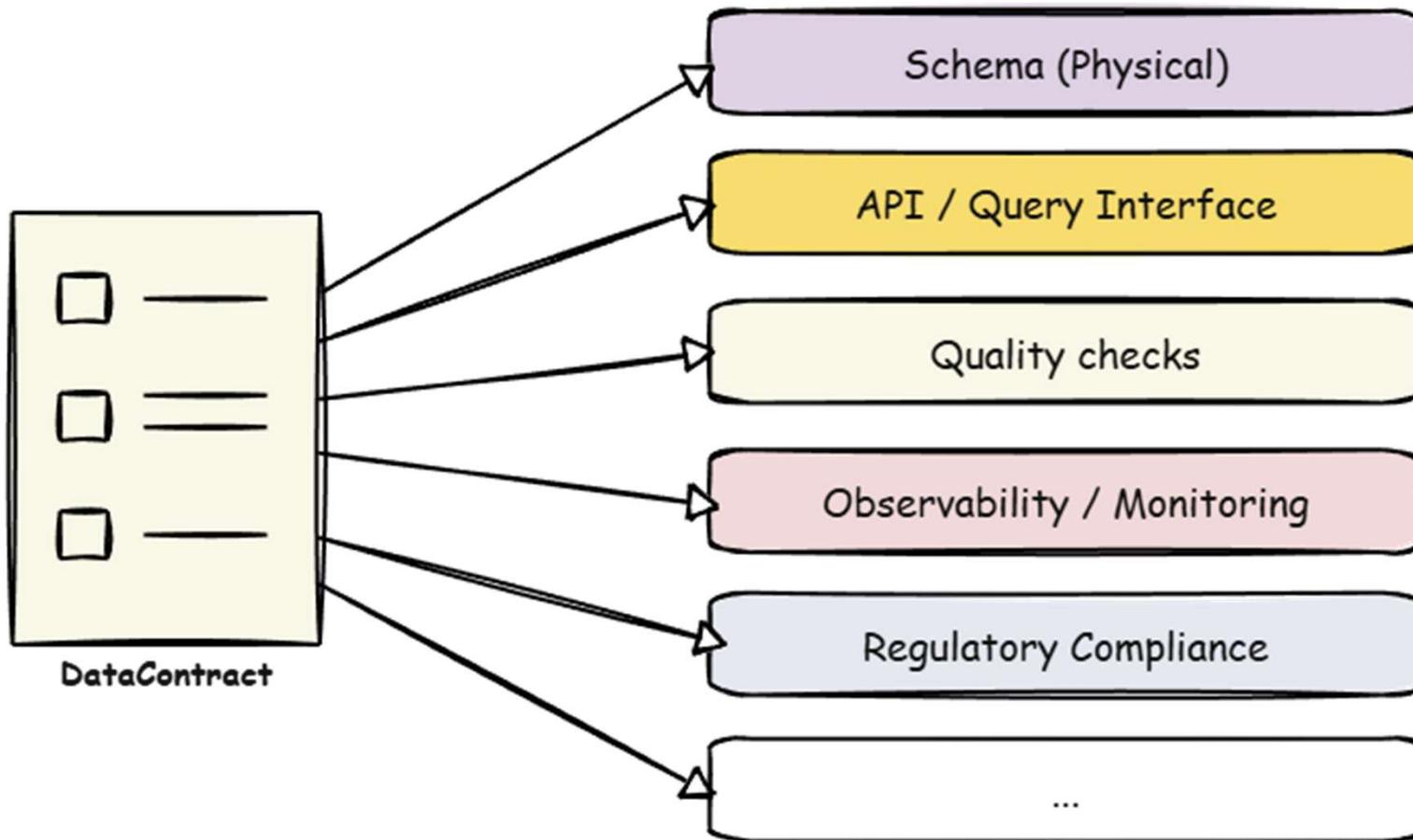
Data Contracts - ODCS

```
- name: VendorID
  logicalType: string
  description: TPEP provider
  quality:
    - rule: validValues
      validValues: [
        '1', # Creative Mobile Technologies, LLC
        '2', # Curb Mobility, LLC
        '6', # Myle Technologies Inc
        '7', # Helix
      ]
```

Data Contracts - ODCS

```
slaProperties:
  - property: latency
    value: 1
    unit: h
    element: nyc_yellow_taxi_trip.tpep_pickup_datetime
  - property: generalAvailability
    value: 2025-05-02T09:30:10-08:00
  - property: retention
    value: 7
    unit: y
    element: nyc_yellow_taxi_trip.tpep_pickup_datetime
```

Data contracts



Spécifications exécutables

Result	Check	Field	Details
failed	Freshness		
failed	Check RatecodeID is in a list of known values.	RatecodeID	Value: 319120.0 Fail: {'greaterThan': 0.0}
failed	Check outliers in passenger_count	passenger_count	Value: 86.0 Fail: {'greaterThan': 0.0}
failed	Check outliers in trip_distance	trip_distance	Value: 3206.0 Fail: {'greaterThan': 0.0}
passed	Check that field 'Airport_fee' is present	Airport_fee	
passed	Check that field Airport_fee has type DOUBLE	Airport_fee	
passed	Check that field 'DOLocationID' is present	DOLocationID	
passed	Check that field DOLocationID has type INT	DOLocationID	
passed	Check that field DOLocationID has no missing values	DOLocationID	
passed	Check that field 'PUlocationID' is present	PUlocationID	
passed	Check that field PUlocationID has type INT	PUlocationID	
passed	Check that field PUlocationID has no missing values	PUlocationID	
passed	Check that field 'RatecodeID' is present	RatecodeID	
passed	Check that field RatecodeID has type BIGINT	RatecodeID	
passed	Check that field 'VendorID' is present	VendorID	
passed	Check that field VendorID has type INT	VendorID	
passed	Check that field VendorID has no missing values	VendorID	
passed	Check VendorID is in a list of known values.	VendorID	
passed	Check that field 'cbd_congestion_fee' is present	cbd_congestion_fee	
passed	Check that field cbd_congestion_fee has type DOUBLE	cbd_congestion_fee	

Data Contracts – Document Excel ??

Schema

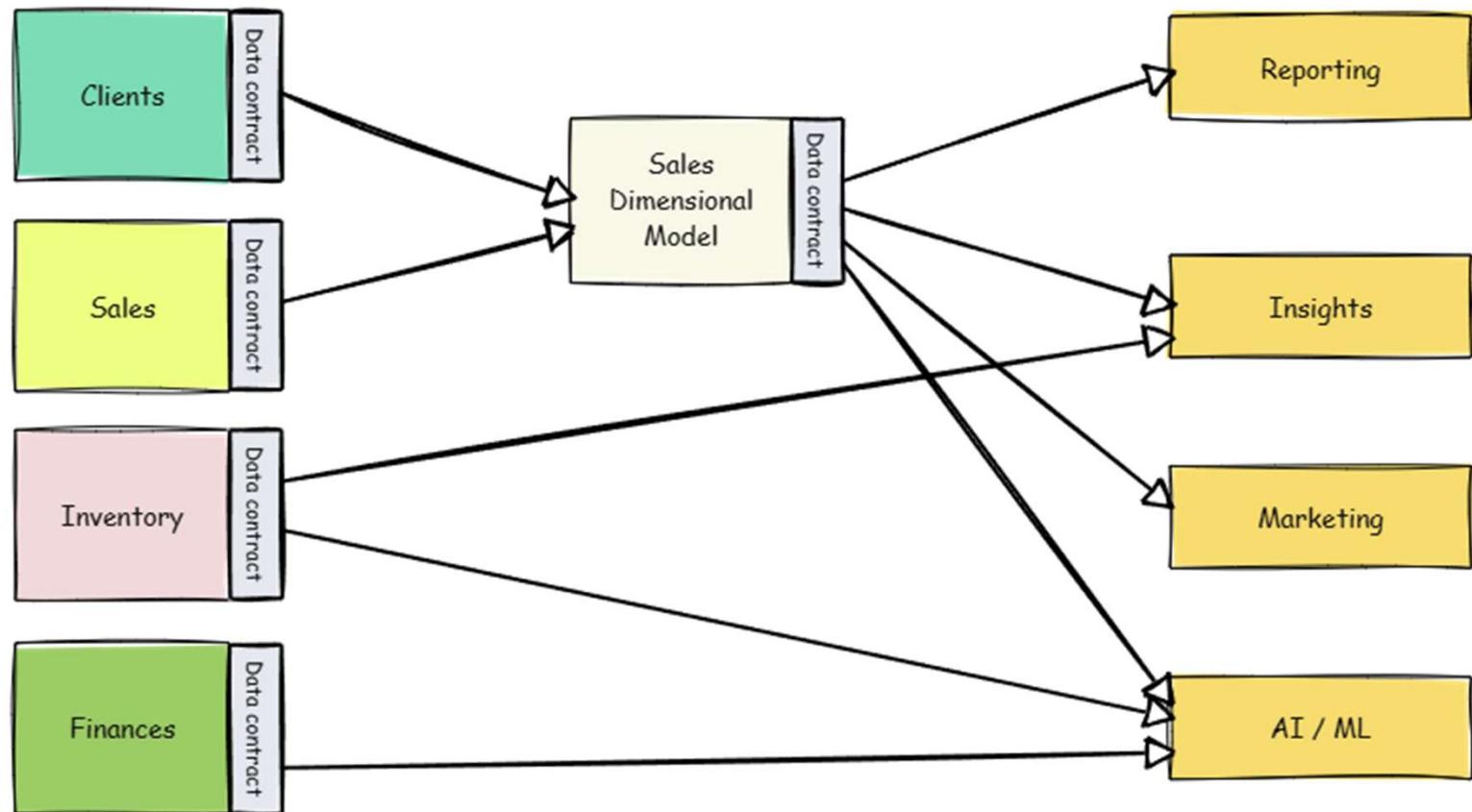
This section describes the data model for one table, message, or object with their properties.

Copy this sheet for every model in your data contract.

Name	yellow_trip
Type	table
Description	Provide data about the NYC Taxicab and Ride History
Business Name	Yellow cab trips
Physical Name	nprod.tlc.yellow_trip
Data Granularity	transactional
Tags	

Property	Logical Type	Physical Type	Description	Required	Unique	Classification
vendor_id	integer	bigint	Vendor id for the TPEP provider	true	false	public
tpep_pickup_datetime	timestamp	timestamp	Pickup date and time	true	false	public
tpep_dropoff_datetime	timestamp	timestamp	Dropoff date and time	true	false	public
passenger_count	integer	int	Number of passengers	true	false	public
trip_distance	number	decimal	Trip distance in miles	true	false	public
ratecode_id	integer	int	Rate code (e.g., Standard, JFK, Newark, etc.)	false	false	public
pu_location_id	integer	int	Pickup location id	true	false	public
do_location_id	integer	int	Dropoff location id	true	false	public
fare	number	decimal	Total fare including taxes and tip	true	false	public

Réseau de data products – Data Mesh



La culture organisationnelle

- Il y a une limite à ce que les moyens techniques peuvent accomplir
- La culture, les rôles et responsabilités sont des aspects critiques
- Les producteurs doivent être imputables de leurs données
- Les consommateurs doivent définir leurs attentes



Conclusion

La qualité des données est une question d'attentes...

- Explorer les données pour mieux connaître leur valeur (EDA)
- Mettre en place des vérifications pour assurer la qualité (Quality Checks)
- Surveiller proactivement le flow de données (Data Observability)
- Mieux définir les attentes, les rôles et les responsabilités (Data contracts / Data products)

Librairies et frameworks

- YData Profiling : <https://github.com/ydataai/ydata-profiling>
- Databricks Labs DQX : <https://databrickslabs.github.io/dqx/>
- Soda : <https://www.soda.io/>
 - Soda core : <https://github.com/sodadata/soda-core>

Ressources

- Data Quality
 - Intro : <https://www.soda.io/resources/no-bs-guide-to-data-quality-dimensions>
- Data Observability
 - Intro : <https://www.ibm.com/think/topics/data-observability>
- Data Contracts & Data Products
 - ODCS : <https://bitol-io.github.io/open-data-contract-standard/latest/>
 - Tutorial : <https://medium.com/data-mesh-learning/so-you-want-to-work-with-data-contracts-and-data-products-03e86f099710>
 - Talk (en français !) : <https://www.linkedin.com/events/odps-l-opendataproductstandard-7353360934775529472/theater/>
 - Gabarit Excel : <https://github.com/datacontract/open-data-contract-standard-excel-template>
 - Data contract CLI : <https://cli.datacontract.com/>
- Data Mesh
 - <https://www.linkedin.com/company/data-mesh-learning/>

Projet GitHub

- <https://github.com/shbehna/data-contract-demo>

Autres références

- Gartner
 - <https://www.gartner.com/en/newsroom/press-releases/2015-09-15-gartner-says-business-intelligence-and-analytics-leaders-must-focus-on-mindsets-and-culture-to-kick-start-advanced-analytics>
- NYC Taxi Data
 - <https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page>

Questions

