

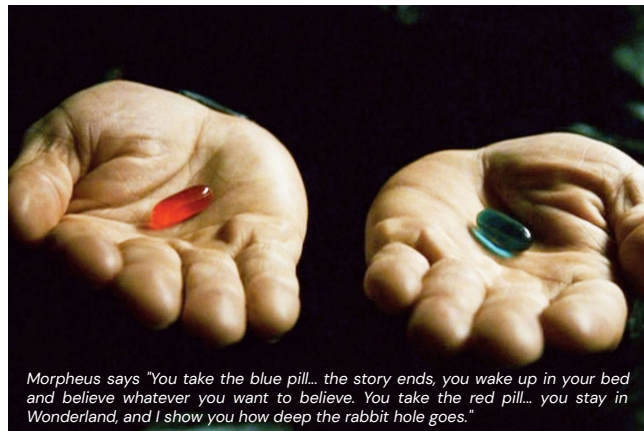
**DATA+AI**  
SUMMIT 2022

# Building an Analytics Lakehouse at Grab

ORGANIZED BY  databricks



# Data Lake



# Data Warehouse

- Freedom & flexibility ✓
- Agility & velocity ✓
- Unlimited scalability ✓

**but..**

- Messy ⚠
- Difficult to standardize ⚠
- Data management becomes impossible ⚠

- ✓ Standardization
- ✓ Accuracy & reliability
- ✓ Coherence/Single-source-of-truth

**but..**

- ⚠ Heavy design investment upfront
- ⚠ Tedious to implement changes
- ⚠ Scalability issues

**DATA+AI**  
SUMMIT 2022

# Building an Analytics Lakehouse at Grab

**Zulfikar Lazuardi Maulana**

Lead Data Scientist (Analytics) at Grab

[linkedin.com/in/zulfikar-lazuardi](https://www.linkedin.com/in/zulfikar-lazuardi)

28 June 2022

ORGANIZED BY  databricks



# Grab in a Nutshell

## MOBILITY



Safe & efficient 4W transport



Affordable 2W option transport



Affordable 4W carpooling option



3W culturally popular Localized transport

## DELIVERIES



On demand food delivery

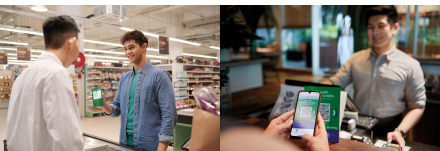


On demand package delivery



On demand grocery delivery

## DIGITAL FINANCIAL SERVICES



Digital payment solution



Digital offline lending, insurance, and many more

many more..



**25 million** monthly transacting users\*

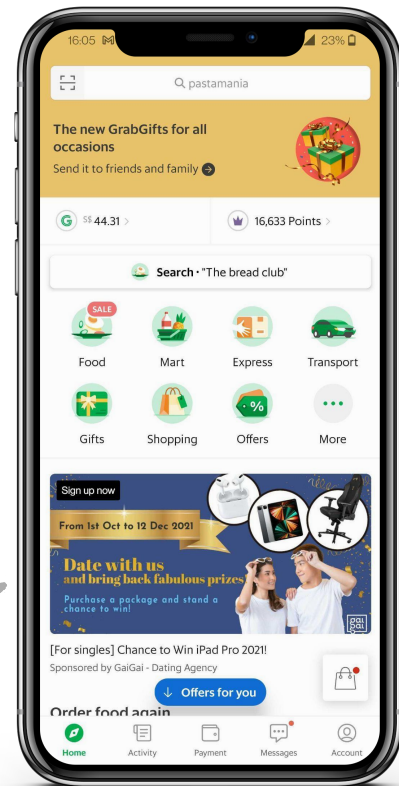


Over **9 million** registered driver-partners, merchant-partners and GrabKios agents across our network\*



Over **30 petabytes** of data

**480 Cities 8 Countries**



\*as of December 2020

# The Dawn of Analytics Lakehouse at Grab

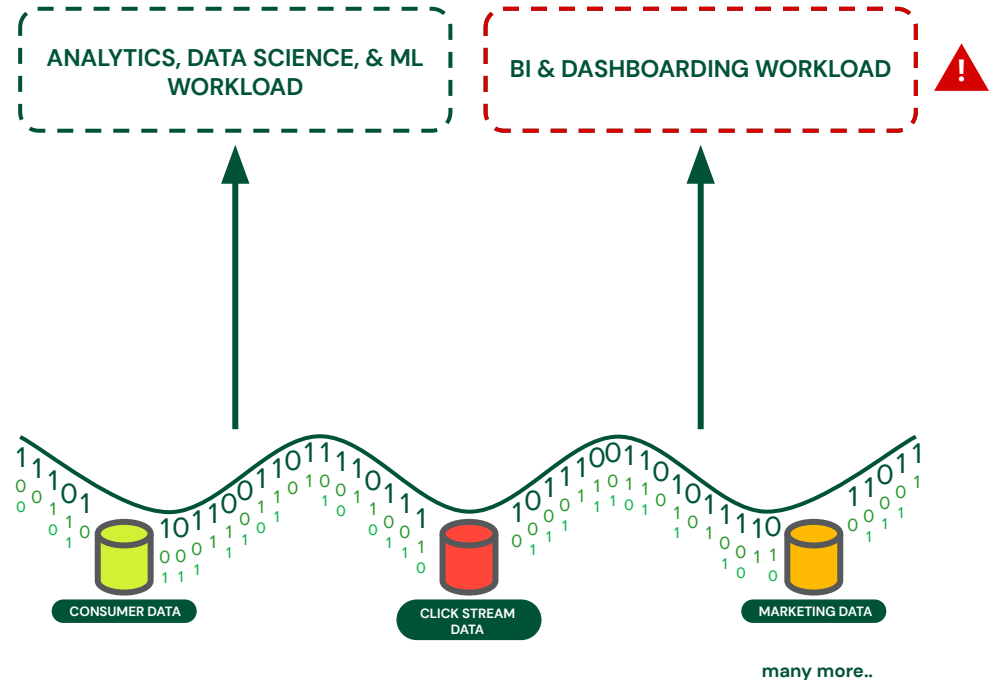
# Data lake: choosing freedom, accept messiness

## Key advantages:

- ✓ Flexibility
- ✓ Agility
- ✓ Scalability
- ✓ Supports DS & ML use-cases

## Key problems:

- ⚠ Difficult to standardize
- ⚠ Poor for BI use-cases



  
DATA LAKE

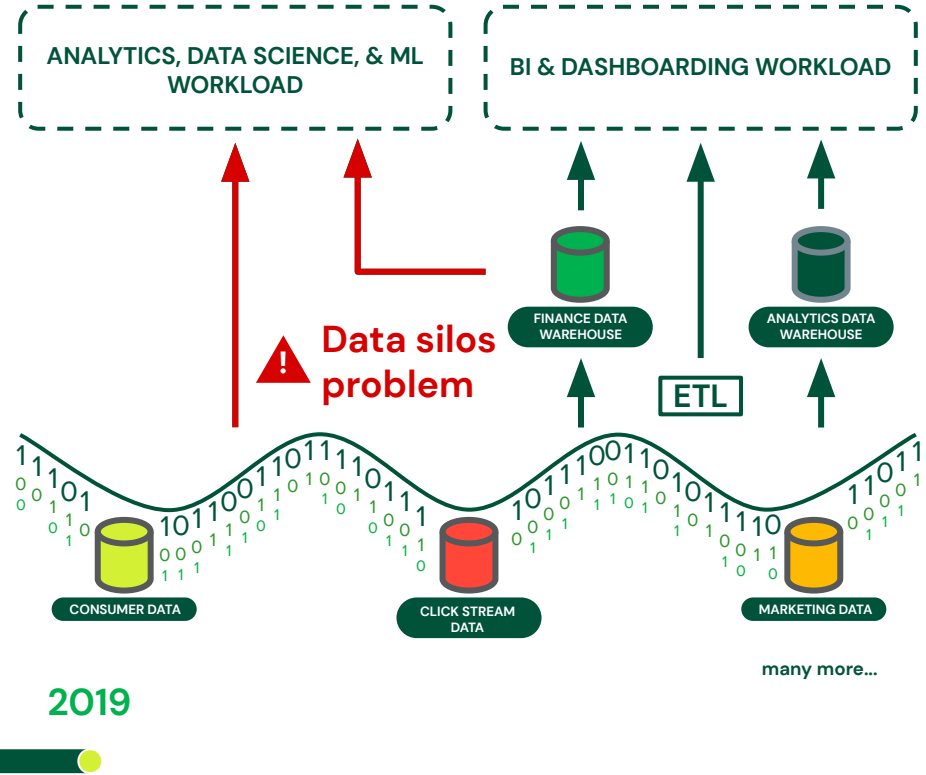
# Data warehouse: choosing standardization, sacrificing agility

## Key advantages:

- ✓ Great for BI use-cases
- ✓ Enhanced data quality & consistency

## Key problems:

- ⚠ Heavy design investment upfront
- ⚠ Tedious to implement changes



2019

DATA LAKE

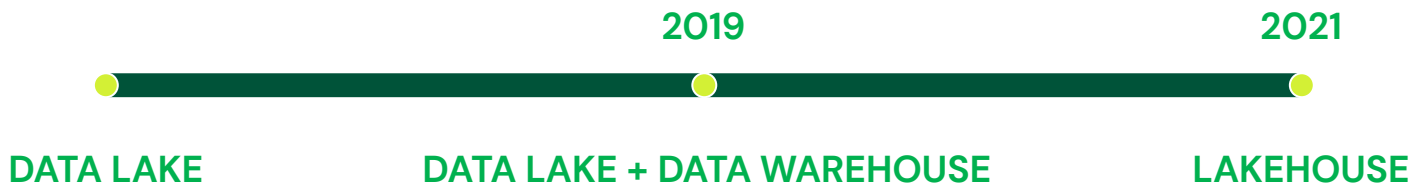
DATA LAKE + DATA WAREHOUSE

# We can have both at once

- Freedom & flexibility ✓
- Agility & velocity ✓
- Unlimited scalability ✓



- ✓ Standardization
- ✓ Accuracy & reliability
- ✓ Coherence/  
Single-source-of-truth





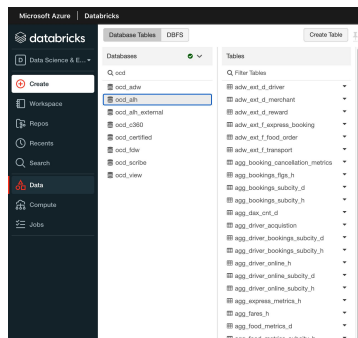
# Introducing One Central Data (OCD)

# One Central Data as Analytics Lakehouse

CENTRAL,  
CERTIFIED,  
CURATED DATA



ONE CENTRAL DATA



ANALYTICS, DATA SCIENCE & ML WORKLOAD



IN-HOUSE DS  
TOOLS

+ Spark + Delta Engine to accelerate our DS workloads

BI & AD-HOC SQL WORKLOAD



+ Familiar SQL syntax and integrated with BI workloads

Schemas

Show 25 rows ▾ Filter

★	✓	Schema	Title	Tables	Popularity
★	✓	data_analytics	Data Analytics	13568	
★		temptables	Temptables	13419	
★		product_analytics	Product Analytics	10476	
★	+5	slide	Slide	8954	
★		stg_user_trust	Staging User Trust	5787	
★		grab_marketing	Grab Marketing	4250	
★					
★					
★	+1	crm	Customer Relationship Management	2155	
★					
★		data_science	Data Science	1459	
★	+1				
★					
★		user_trust	User Trust	732	
★	+1	user_trust_datamart	User Trust Datamart	695	
★					
★		econs_id	Economics Indonesia	433	
★		ds_presto	Data Science Presto	407	
★		chimera_gaia		293	
★		ds_econs_public	Data Science Economics Public	263	
★					
★					
★		geo	Geo	204	

Showing 1 to 25 of 474

Page 1 Prev Next



CONSUMER DATA



CLICK STREAM DATA



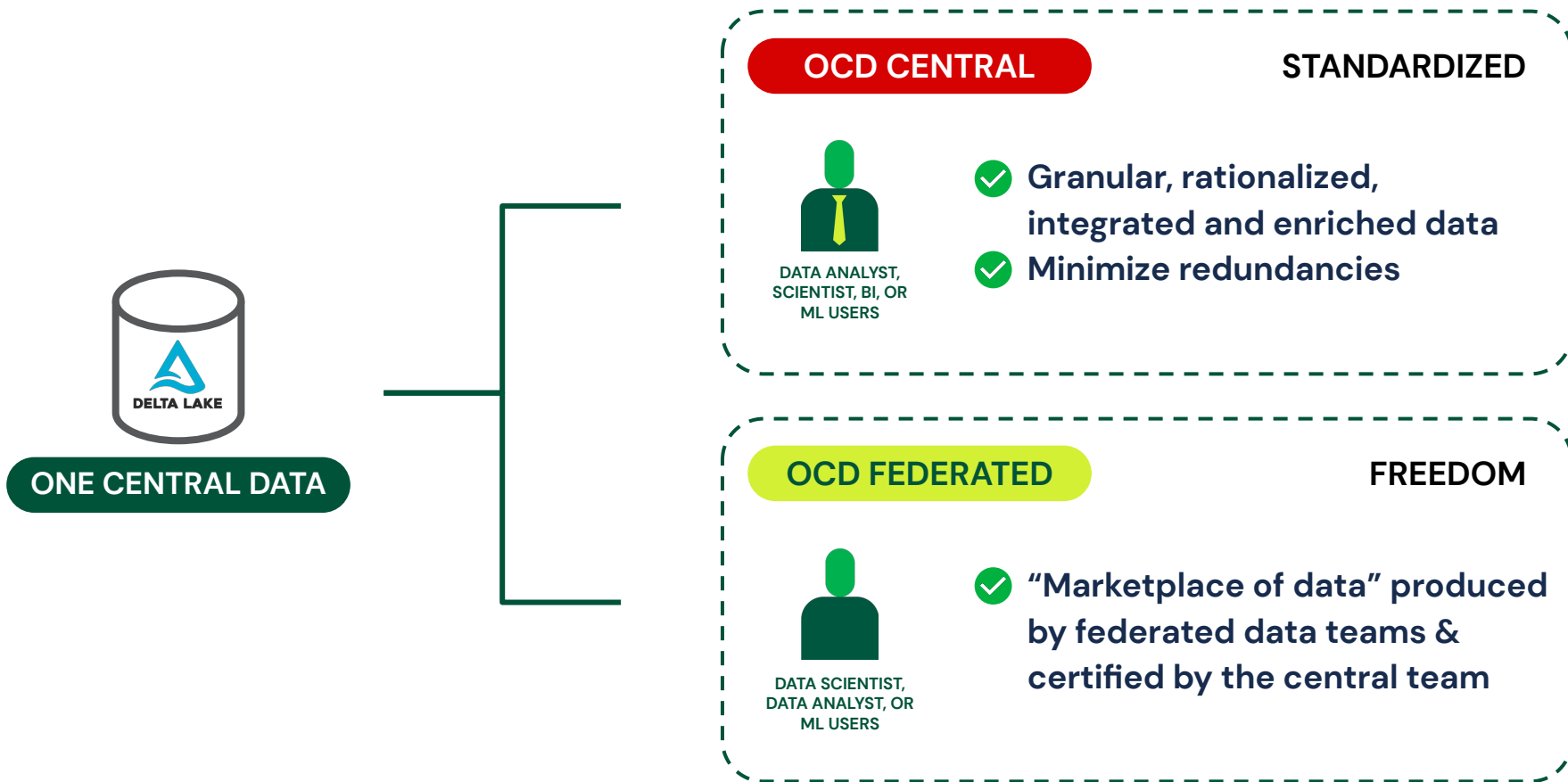
FINANCE DATA



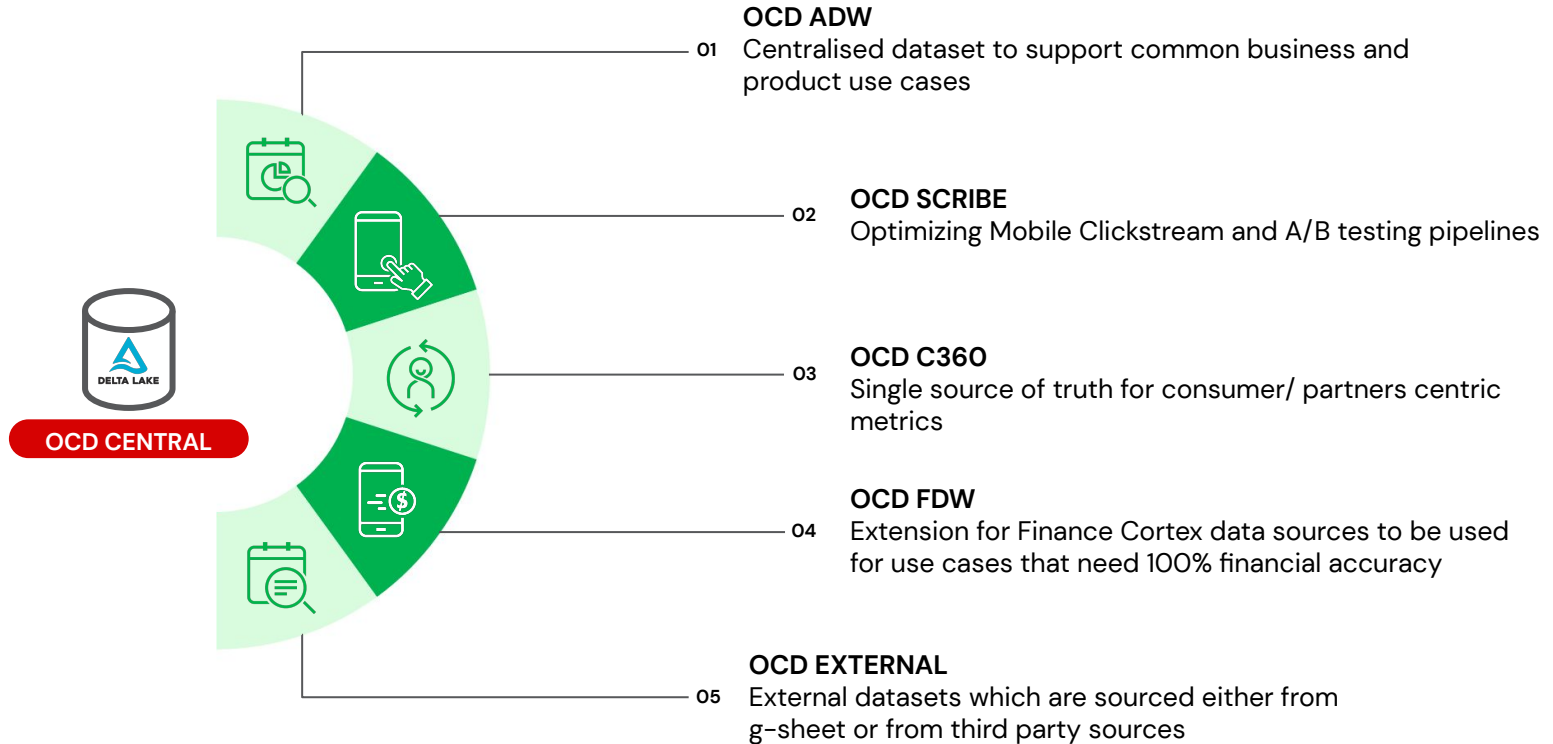
DATA WAREHOUSES

>100K tables

# Balancing freedom and standardization



# OCD Central: single source of truth for all personas



# OCD Central: standardization that matters

**BRONZE DATA**

Schemas Show 25 rows ▾ Filter

★	✓	Schema	Title	📄	Tables	Popularity
☆	+1	data_analytics	Data Analytics	📄	13568	📊
☆		temptables	Temptables	📄	13419	📊
☆		product_analytics	Product Analytics	📄	10476	📊
☆	+5	slide	Slide	📄	8854	📊
☆		stg_user_trust	Staging User Trust	📄	5787	📊
☆		grab_marketing	Grab Marketing	📄	4250	📊
☆						
☆						
☆	+1	crm	Customer Relationship Management	📄	2155	📊
☆						
☆		data_science	Data Science	📄	1459	📊
☆	+1					
☆						
☆		user_trust	User Trust	📄	722	📊
☆	+1	user_trust_datamart	User Trust Datamart	📄	695	📊
☆						
☆		econs_id	Economics Indonesia	📄	433	📊
☆		ds_presto	Data Science Presto	📄	407	📊
☆		chimera_gaia		📄	293	📊
☆		ds_econs_public	Data Science Economics Public	📄	263	📊
☆						
☆						
☆		geo	Geo	📄	204	📊

Showing 1 to 25 of 474 Page 1 Prev Next

➤ 100K tables in Bronze data

**SILVER DATA & GOLD DATA**

Tables Show 25 rows ▾ Filter

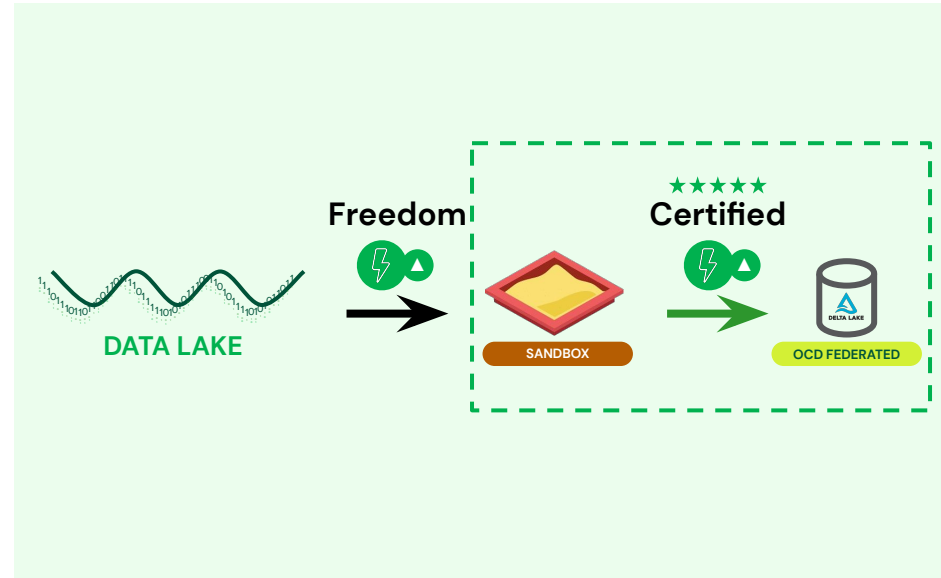
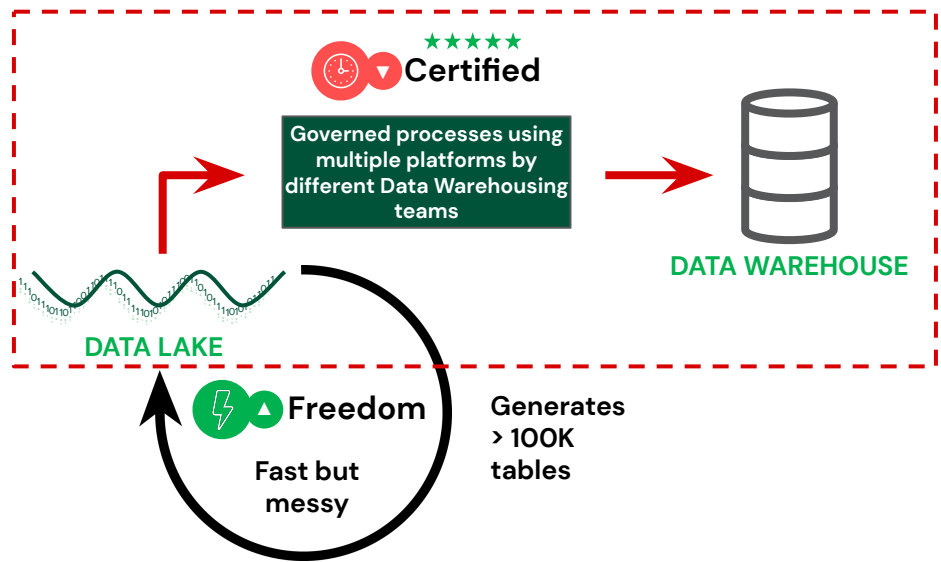
★	✓	Table ID	Title	📄	Popularity	Columns	Rows
☆		agg_food_metrics_d	Aggregate Food Metrics Daily	📄	141	0	0
☆	+1	agg_express_metrics_h	Aggregate Express Metrics Hourly	📄	125	0	0
☆	+2	agg_driver_bookings_subcity_h	Aggregate Driver Bookings Subcity Hourly	📄	122	0	0
☆		agg_driver_bookings_subcity_d	Aggregate Driver Bookings Subcity Daily	📄	121	0	0
☆							
☆		agg_bookings_subcity_h	Aggregate Bookings Subcity Hourly	📄	109	0	0
☆		agg_bookings_subcity_d	Aggregate Bookings Subcity Daily	📄	108	0	0
☆							
☆		f_elo		📄	95	0	0
☆		agg_bookings_flag_h	Aggregate Bookings Hourly with Flags	📄	91	0	0
☆		agg_driver_acquisition	Aggregate Driver Acquisition	📄	48	0	0
☆		f_driver_online_detail_subcity_h		📄	45	0	0
☆		f_driver_online_detail_subcity_d		📄	44	0	0
☆		f_hc_user_event	Help Center Portal User Clickstream Events	📄	41	0	0
☆		f_hc_user_clickstream_event	Help Center Portal User Clickstream Events	📄	41	0	0
☆		agg_food_metrics_subcity_h		📄	39	0	0
☆		dim_merchant_classification	Aggregate	📄	37	0	0

Standardize & streamline into 500+ tables

# OCD Federated: allowing freedom to create with agility & velocity

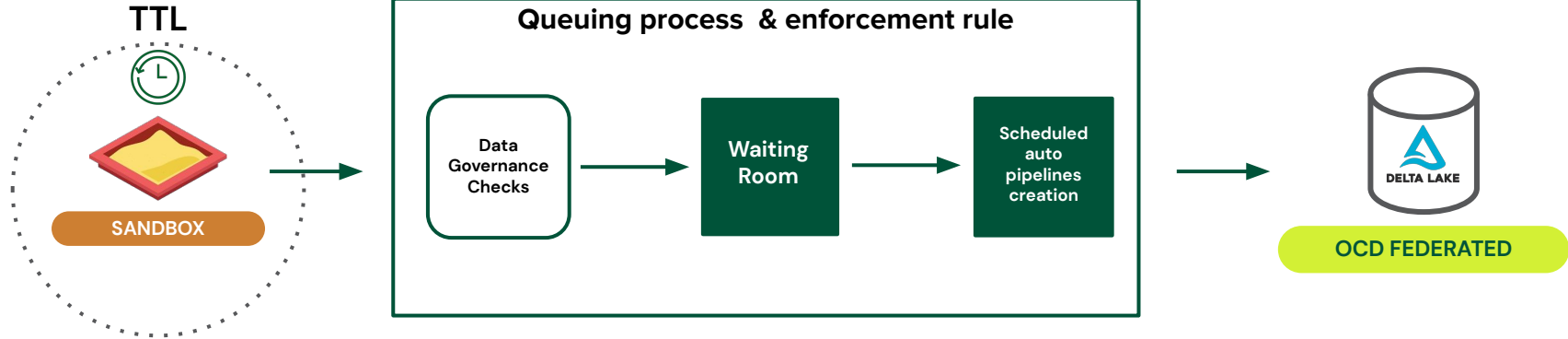
BEFORE

AFTER

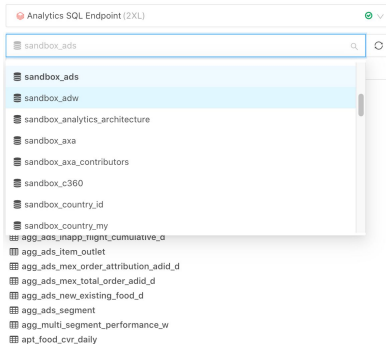


User can complete the certified process in a day

# Certifying data from laissez-faire sandbox to standardized OCD federated



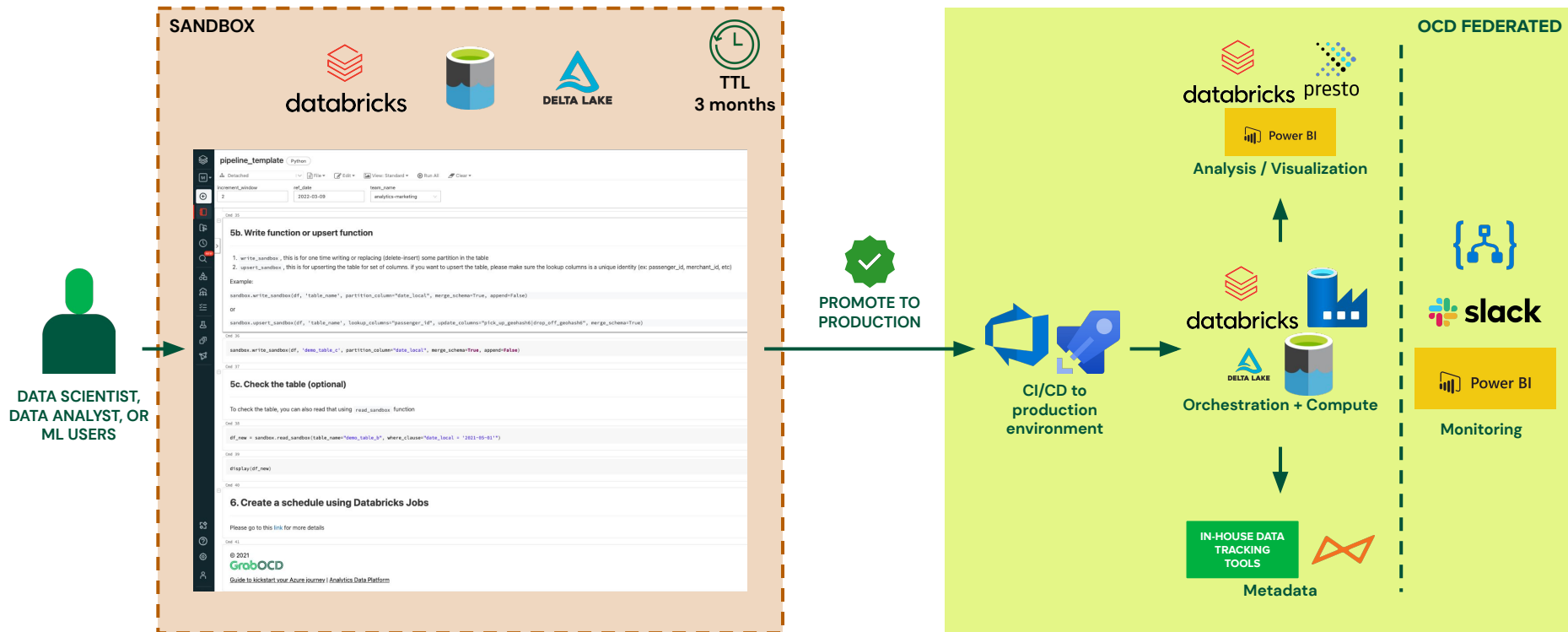
Temporary isolated area per team



```
def promote_to_prd(self, table_name, job_id, description,
                  backfill_start_date=None, backfill_end_date=None,
                  backfill_interval_hours=None, backfill_timezone_id=None):
    """
    promote sandbox data to production environment
    """
    Args:
        table_name (str): sandbox table name
        job_id (int): sandbox job id in databricks
        description (str): can be string or html string for description
            in Alation
        backfill_start_date (str): optional, backfill start date
        backfill_end_date (str): optional, backfill end date
        backfill_interval_hours (int): optional, interval for backfill
            dates using hours
        backfill_timezone_id (str): optional, timezone id for backfill.
            can be UTC or SGT
    Returns:
        None
```

- ✓ Reduce time to production data/ ML models
- ✓ Powered by Databricks (Spark + Delta Engine)
- ✓ Sandbox help to reduce unwanted tables in production

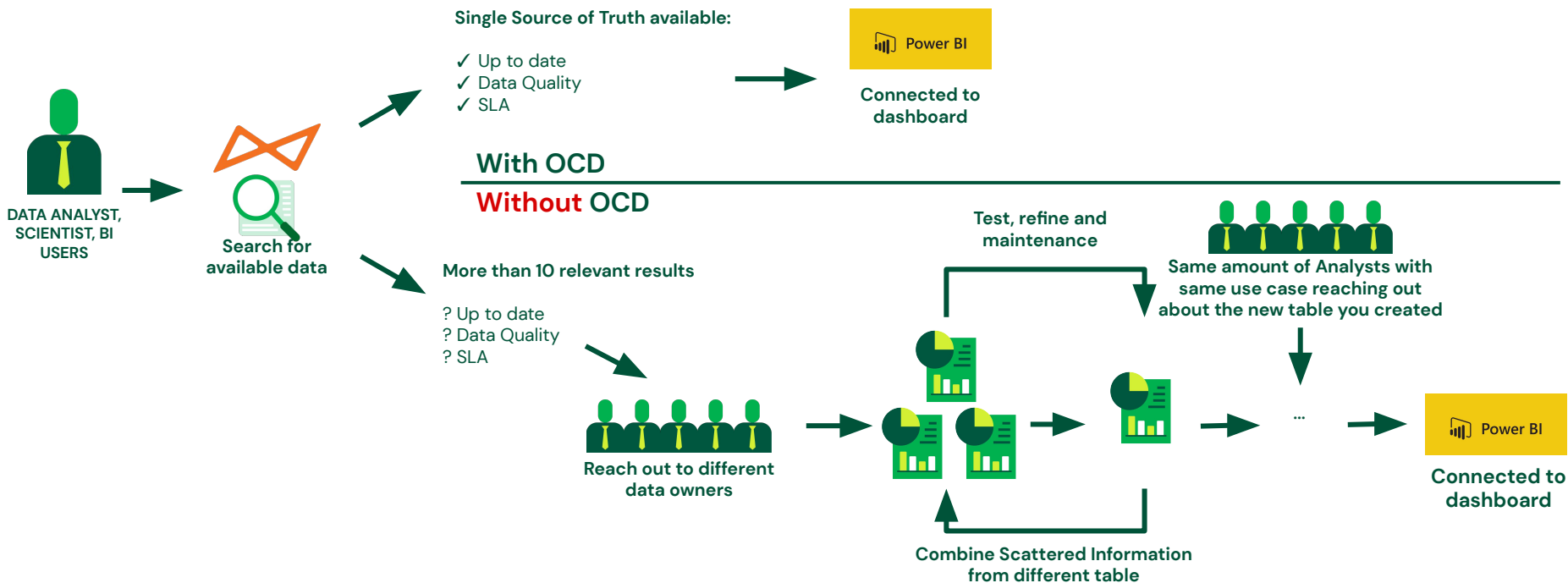
# OCD Federated in details





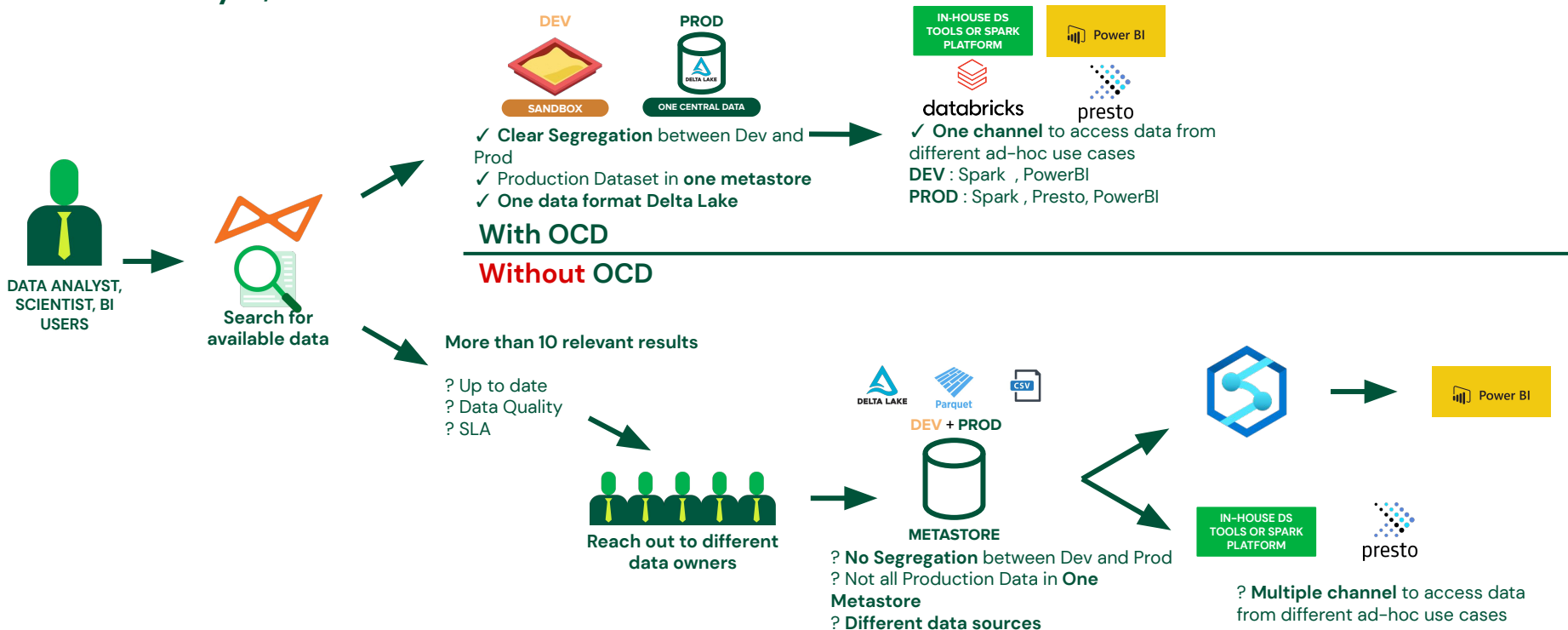
# User journey: example to query the data from different use cases – dashboard

As an Analyst, I want to build a dashboard...



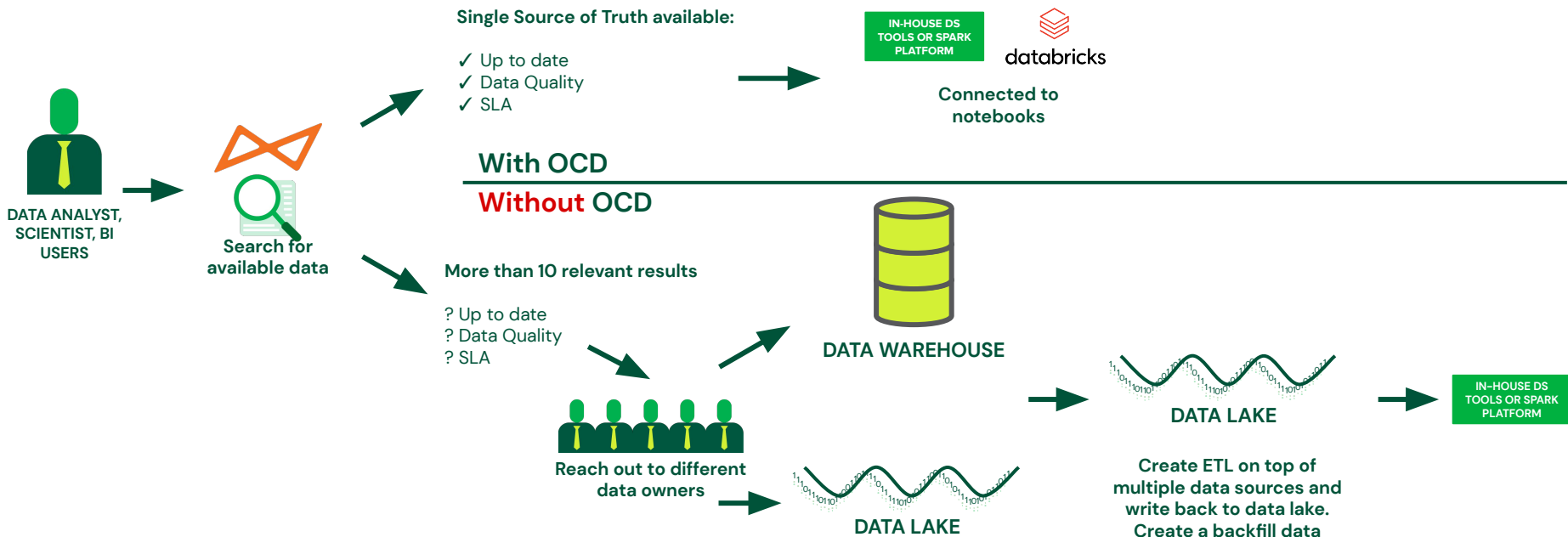
# User journey: example to query the data from different use cases - ad-hoc

As an Analyst, I want to build an ad-hoc...

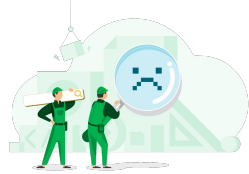


# User journey: example to query the data from different use cases – data science

As an Analyst, I want to build a data science model...



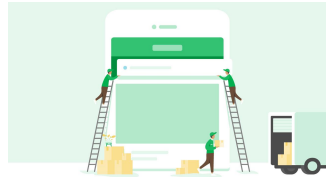
# “Data Lake”-ish use-cases on Analytics Lakehouse, some examples:



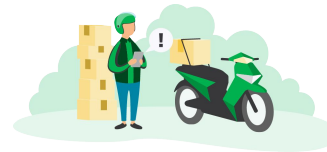
**Merchant Thumbnail Image  
Recognition**



**Customer Lifetime Value**

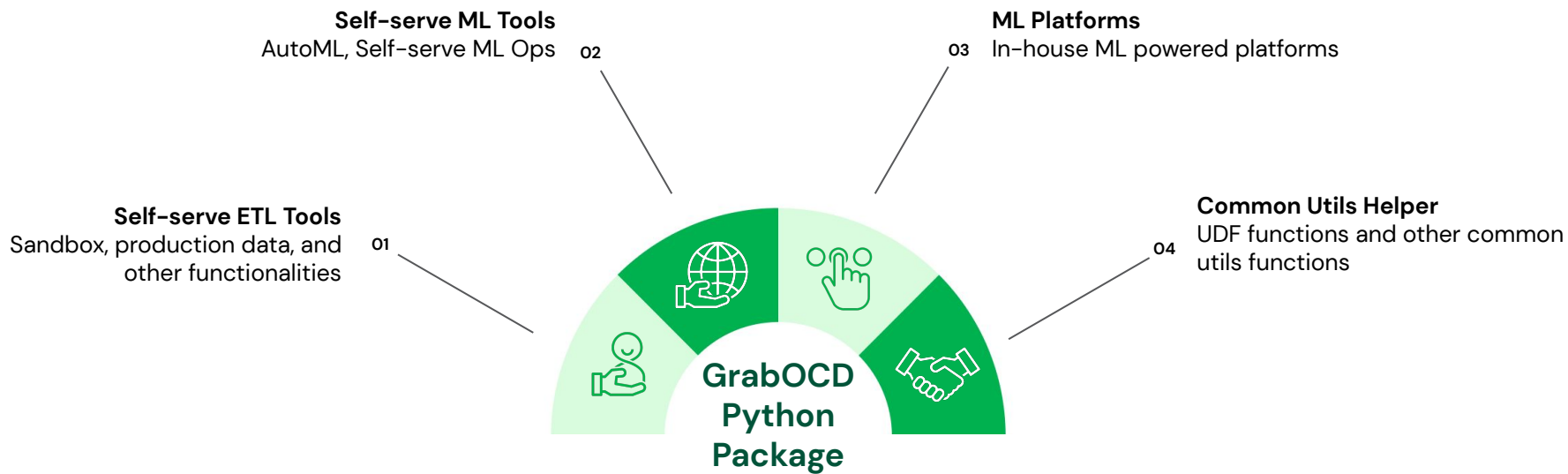


**Click Stream Data**

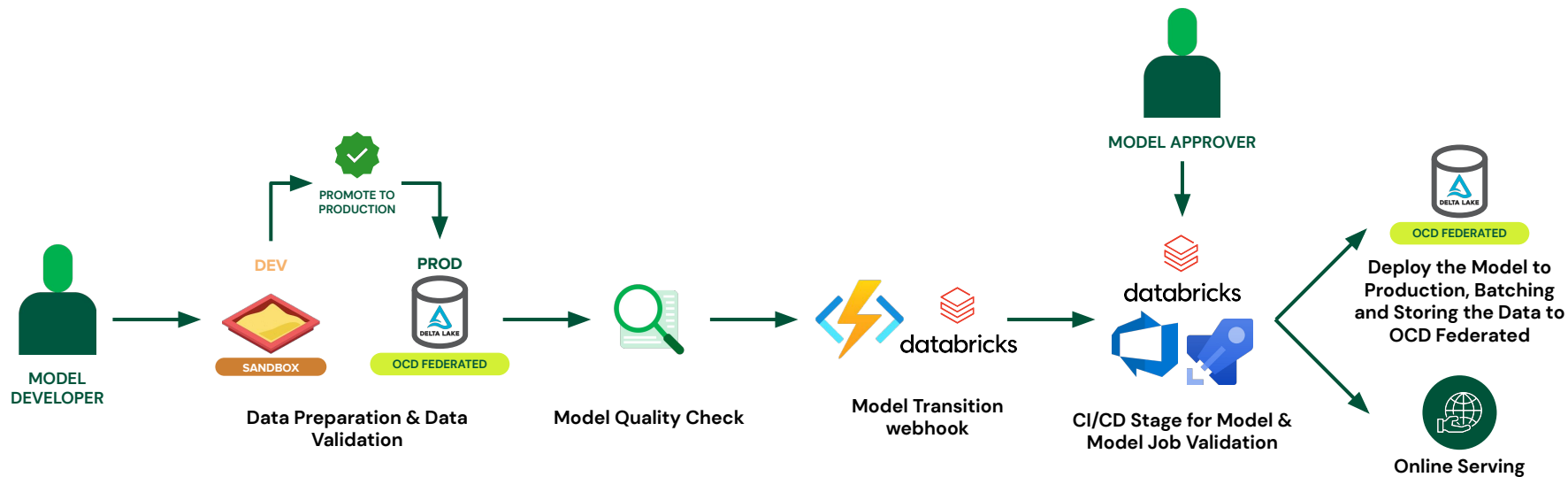


**Demand Shaping Model**

# “Data Lake”-ish use-cases on Analytics Lakehouse, python package:

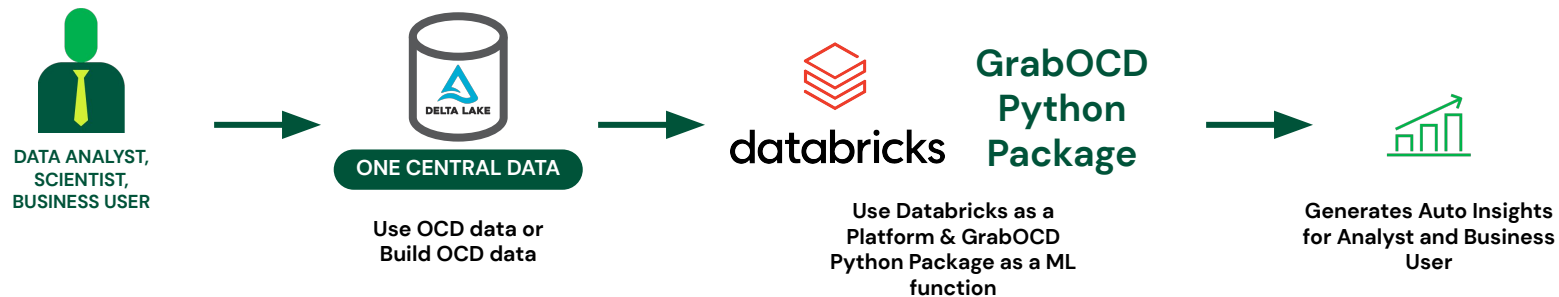


# Also helps to standardize self-serve ML Ops at scale in Analytics



Made easy relying on certified federated datasets

# Also helps to automate the insights using ML Platforms



```
buddy.predict(df=df, target_column="insured", delta_path="/mnt/analytics_storage/zulfikar/buddy/insured_usecase_v5/", problem_type="classification")
```

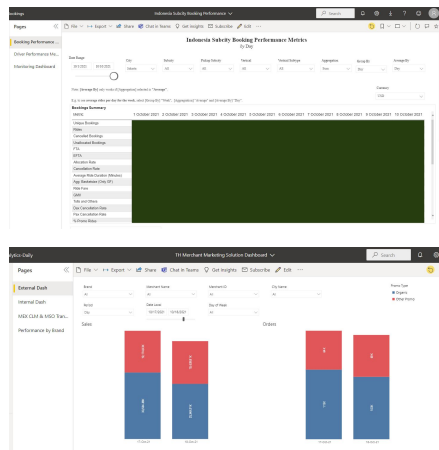
» (11) Spark Jobs

**NOTE:** The dataset loaded below is a sample of the original dataset. Stratified sampling using pyspark's sampleBy method is used to ensure that the distribution of the target column is retained. Rows were sampled with a sampling fraction of 0.1795236456727593

# “Data Warehouse”-ish use-cases on Analytics Lakehouse, some examples:

## Country Analytics Dashboards

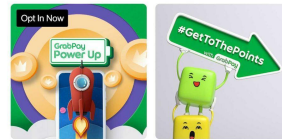
- Indonesia Sub-city Booking Performance
- Thailand Merchant Marketing Solution Dashboard



## Marketing Projects

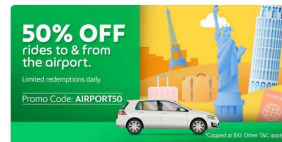
- Insight the in-app channels for attribution model and demand shaping

Get more points & GrabPay deals →



Tap for thousands of bonus points

This way to up to 2% back points



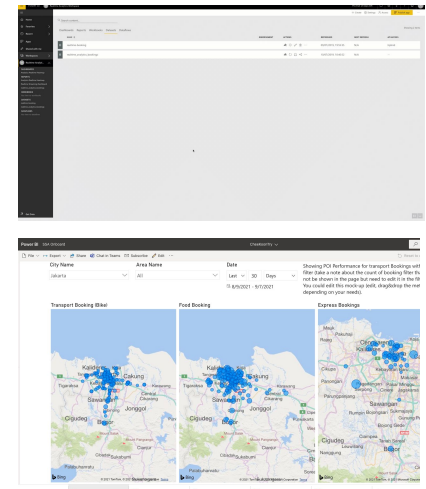
Begin your journey with a treat, from us.

Delightful Deals this Deepavali →



## Self-Service Analytics

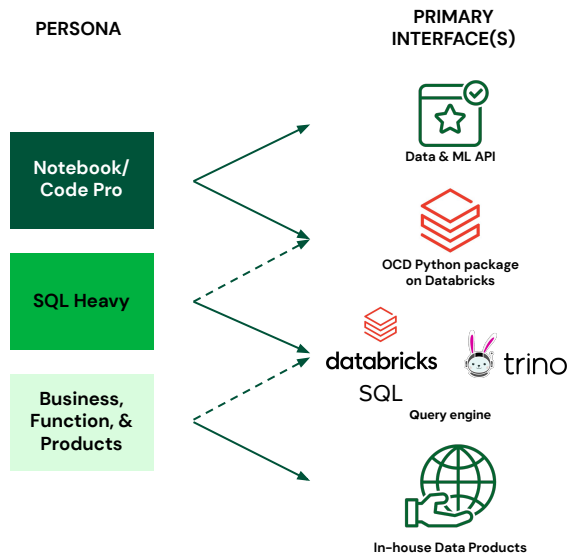
- No-code, self-service analytics for business users and citizen data scientists to make data-driven decisions





# Our journey continues..

- ✓ Single abstraction layer of truth
- ✓ Addressing SQL persona use-cases:



- ✓ No-Code for Citizen Data Scientists & Self-service analytics

**DATA+AI**  
SUMMIT 2022

ORGANIZED BY  databricks

# Thank you



Zulfikar Lazuardi Maulana  
[linkedin.com/in/zulfikar-lazuardi](https://www.linkedin.com/in/zulfikar-lazuardi)  
[lazuardi32@gmail.com](mailto:lazuardi32@gmail.com)