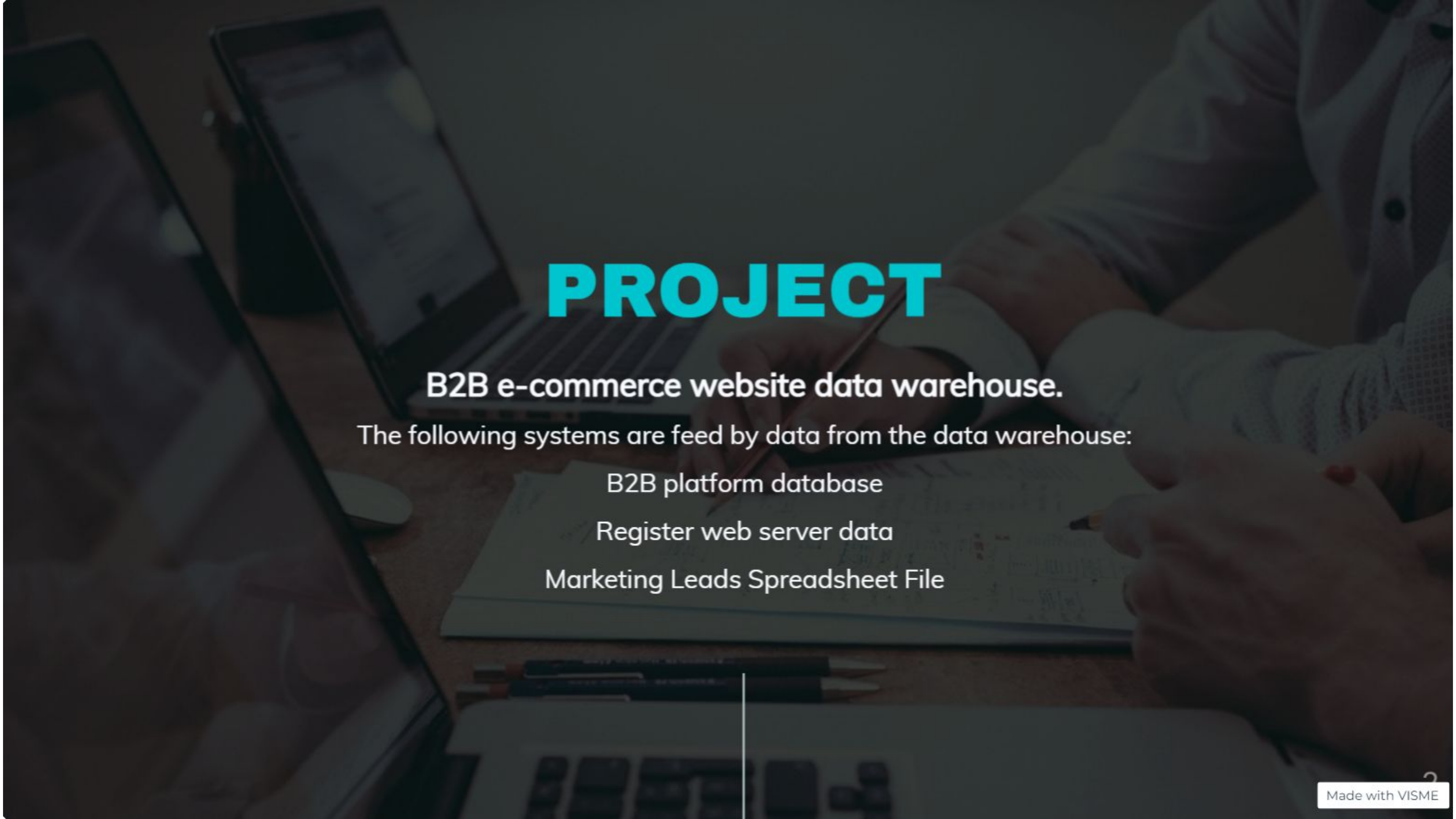
PRESENTED BY: BRENDA ALEXSANDRA

B2B TOPTAL

DATA WAREHOUSE



PROJECT

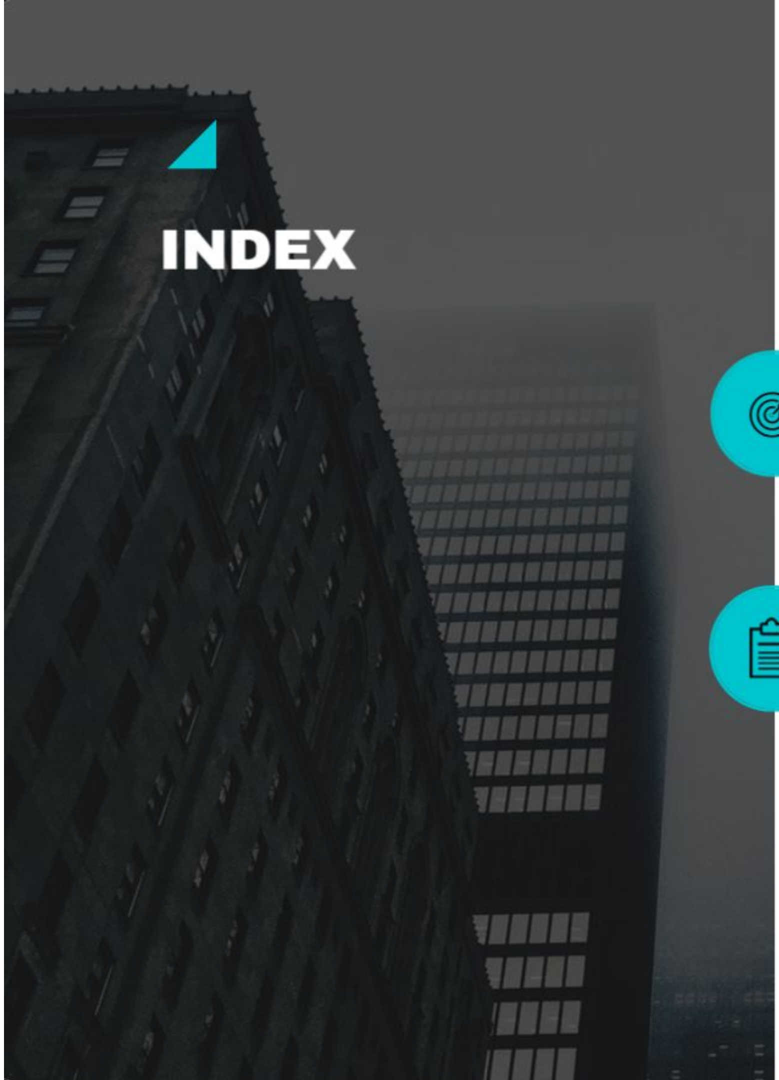
B2B e-commerce website data warehouse.

The following systems are feed by data from the data warehouse:

B2B platform database

Register web server data

Marketing Leads Spreadsheet File



INDEX

Objective

All project requirements and objectives

Methodology

- Architecture
- Data Modeling
- Data Pipeline
- Data Tracking
- Tests
- Marketing Lead Sheet
- Infrastructure as Code

OBJECTIVE

PROJECT REQUIREMENTS

DATABASE IMPLEMENTATION

Database implementation with generated data for the B2B.
Weblog generated via script in a language of your choosing

DATA MODELING

A target database which represents the data warehouse

ETL PIPELINE

ETL process with transformations that will fill the initial load of the target datastore and can be restarted if tasks or subtasks fail.

HANDLE ERRONEOUS DATA

Test, identify and handle with erroneous data.

DATA TRACKING

Track data and ETL/ELT metadata

DATA PROCESSING

Transform the data into a readable data format for reporting
Transform large data size.

METHODOLOGY

- ARCHITECTURE
- DATA MODELING
- DATA PIPELINE
- DATA TRACKING
- TESTS
- MARKETING LEAD SHEET
- INFRASTRUCTURE AS CODE

Architecture Project

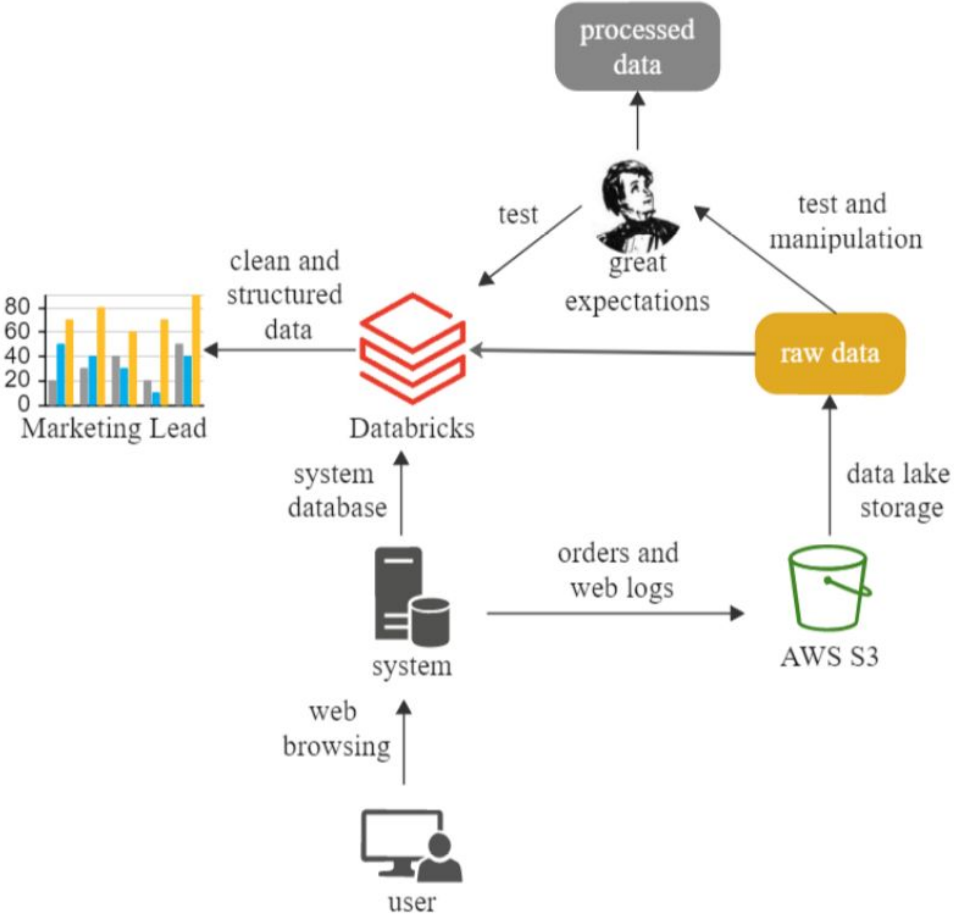
E-COMMERCE B2B DATA SYSTEM TOOLS

DATA STACK:

- AWS S3
- AWS EC2
- AWS Cloud Formation
- Databricks Data Science & Engineering
- Databricks SQL Query Editor
- Great Expectations
- Terraform

Architecture Project

E-COMMERCE B2B DATA SYSTEM TOOLS



Made with VISME

METHODOLOGY

- ARCHITECTURE ✓
- DATA MODELING
- DATA PIPELINE
- DATA TRACKING
- TESTS
- MARKETING LEAD SHEET
- INFRASTRUCTURE AS CODE

DATA MODELING

RELATIONAL DATA WAREHOUSE

TRANSACTIONAL DATA BASE

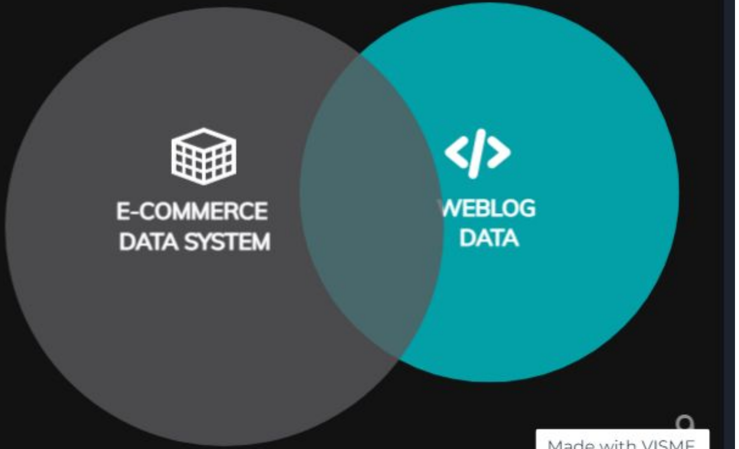
- COMPANIES
- CUSTOMERS
- PRODUCTS
- ORDERS (INCLUDE ITEMS)

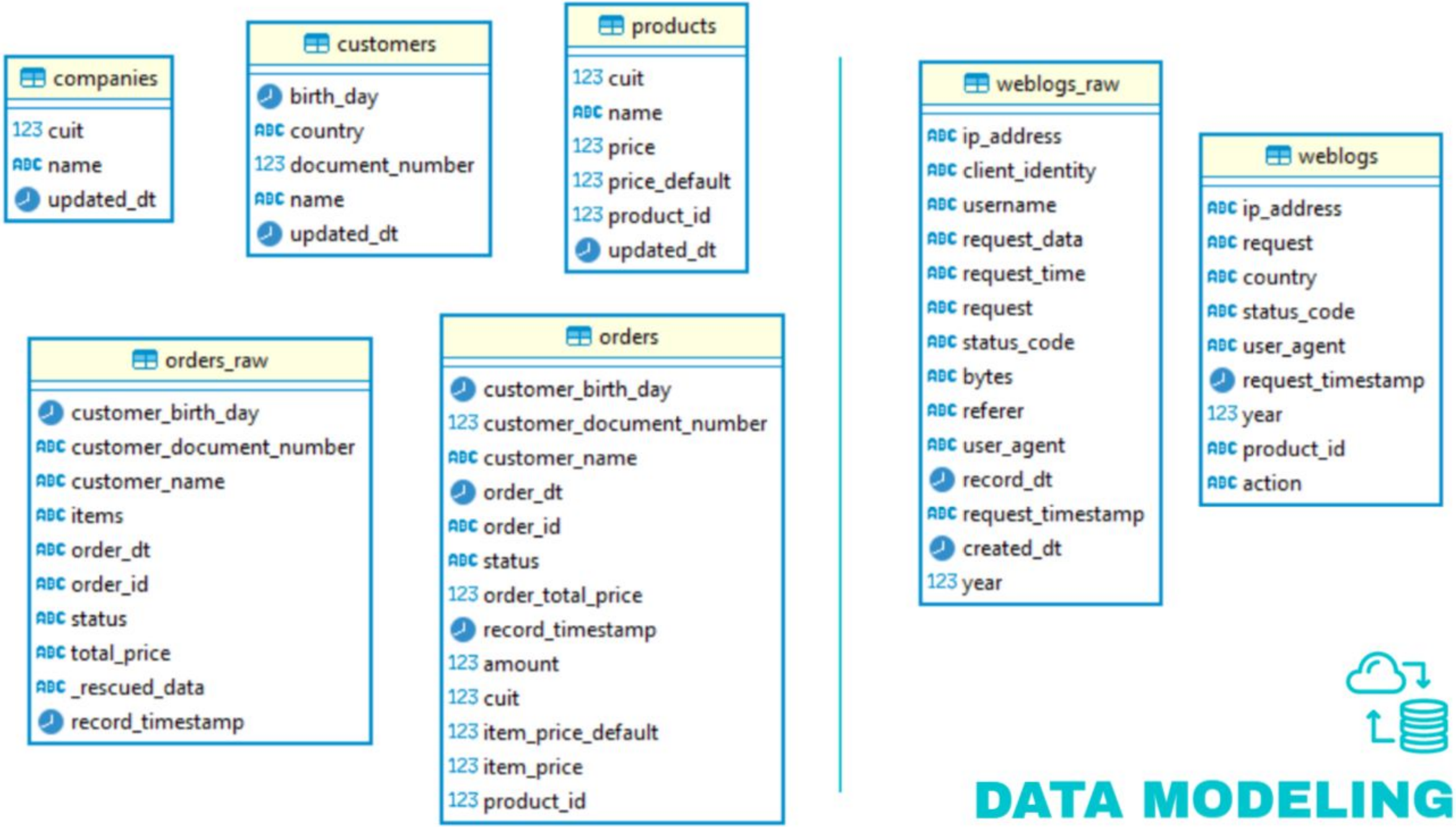
ANALYTICAL DATA BASE

- WEBLOG

TECNICAL FEATURES

- SQL
- DENORMALIZED
- DELTA TABLES
- PARTITIONED TABLE
  - WEBLOG PARTITIONED BY COUNTRY





companies

- 123 cuit
- ABC name
- 🕓 updated_dt

customers

- 🕓 birth_day
- ABC country
- 123 document_number
- ABC name
- 🕓 updated_dt

products

- 123 cuit
- ABC name
- 123 price
- 123 price_default
- 123 product_id
- 🕓 updated_dt

orders_raw

- 🕓 customer_birth_day
- ABC customer_document_number
- ABC customer_name
- ABC items
- ABC order_dt
- ABC order_id
- ABC status
- ABC total_price
- ABC _rescued_data
- 🕓 record_timestamp

orders

- 🕓 customer_birth_day
- 123 customer_document_number
- ABC customer_name
- 🕓 order_dt
- ABC order_id
- ABC status
- 123 order_total_price
- 🕓 record_timestamp
- 123 amount
- 123 cuit
- 123 item_price_default
- 123 item_price
- 123 product_id

weblogs_raw

- ABC ip_address
- ABC client_identity
- ABC username
- ABC request_data
- ABC request_time
- ABC request
- ABC status_code
- ABC bytes
- ABC referer
- ABC user_agent
- 🕓 record_dt
- ABC request_timestamp
- 🕓 created_dt
- 123 year

weblogs

- ABC ip_address
- ABC request
- ABC country
- ABC status_code
- ABC user_agent
- 🕓 request_timestamp
- 123 year
- ABC product_id
- ABC action

DATA MODELING

RELATIONAL DATA WAREHOUSE

METHODOLOGY

- ARCHITECTURE ✓
- DATA MODELING ✓
- DATA PIPELINE
- DATA TRACKING
- TESTS
- MARKETING LEAD SHEET
- INFRASTRUCTURE AS CODE

Data Pipeline

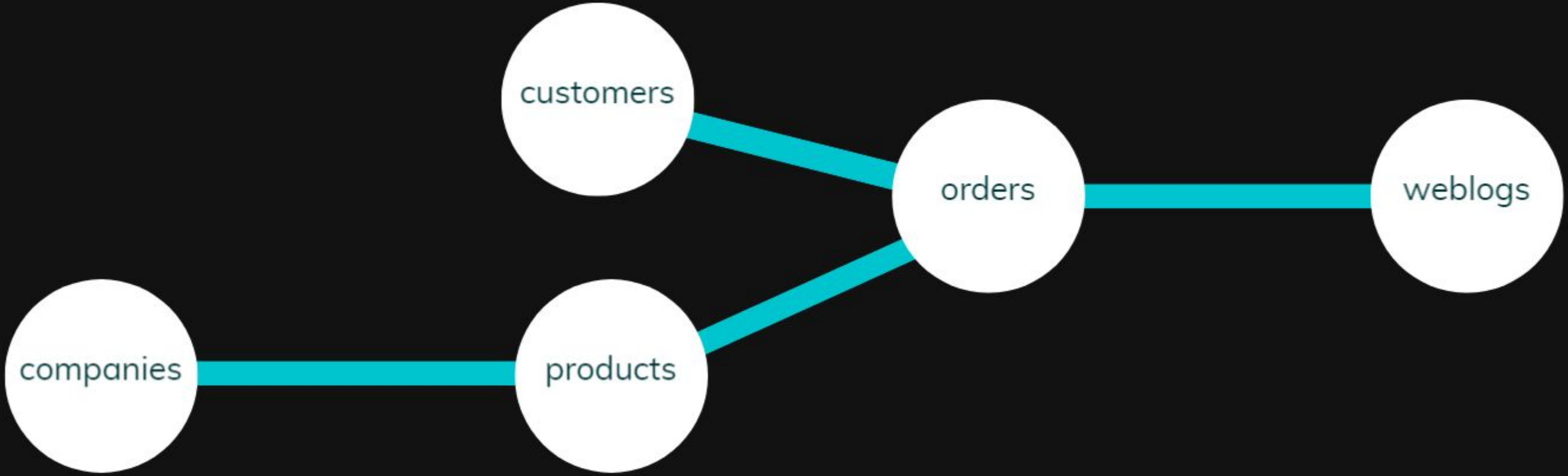
E-COMMERCE B2B ETL PIPELINE

- Ingestion and creation of random data used in this project.
- Daily ingestion of customer, companies and products data.
- Daily ingestion of order data.
- Ingestion of streaming weblog data.

DATA PIPELINE

DATA GENERATION

Pipeline that generates all the data used in the project.
The data was generated using Faker python package.



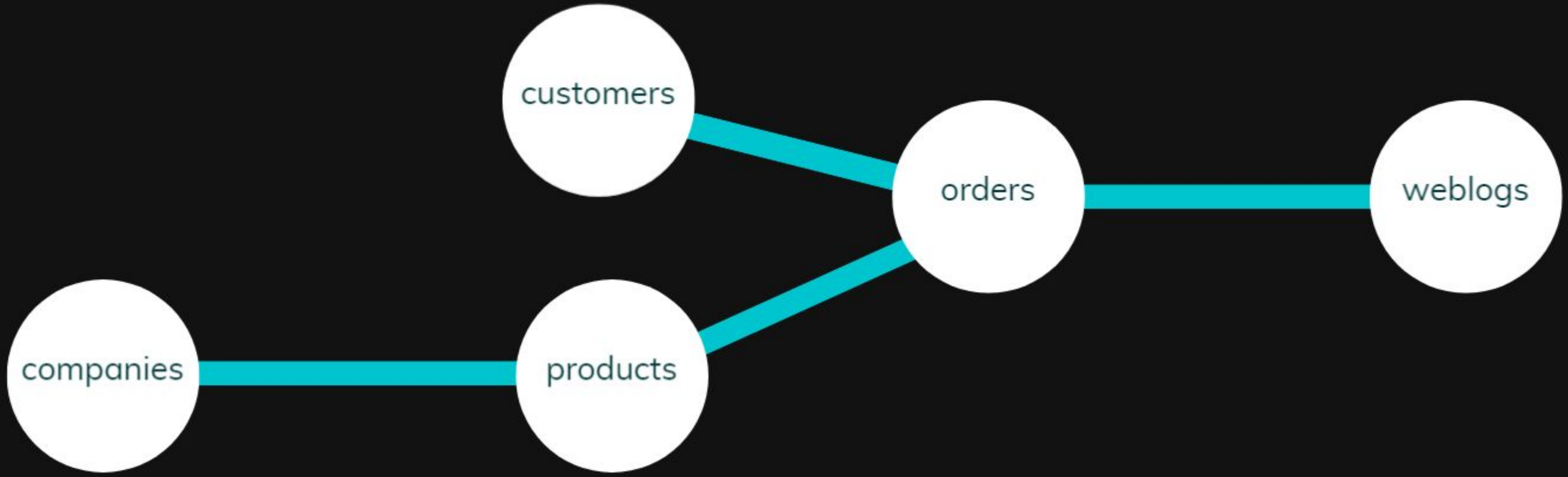
DATA PIPELINE

DAILY CUSTOMER, PRODUCTS AND COMPANIES INGESTION

keep customer, company and product data updated.
This data is written directly to the data warehouse (system -> data warehouse)

FEATURES

Updates backup data whenever a record is updated.
If the pipeline tries to ingest data into a table that has been deleted, it will recreate it with data from the backup in the datalake.
Retry and e-mail notification on failure.



DATA PIPELINE

STREAMING DAILY ORDERS INGESTION

Keep orders data updated.

Pipeline steps:

1. Generate random order data in json format and store in **datalake**.
2. Ingest the data into `orders_raw` into the **data warehouse**.
3. Test the data ingested in `orders_raw`.
4. Make the necessary manipulations in `orders_raw` and store in `orders` table.
5. Test the data ingested in `orders`.

FEATURES

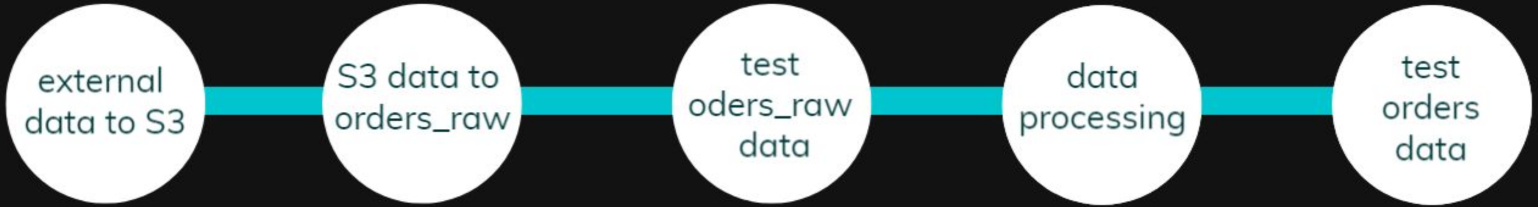
Data avalaible on data lake.

Autoloader was used for ingestion in these two tables.

The autoloader is a Databricks tool that needs 3 paths to start the ingestion:

i. schema,
ii. folder for badly formatted/erroneous records and
iii. checkpoint.

Retry and e-mail notification on failure.



external data to S3

S3 data to orders_raw

test oders_raw data

data processing

test orders data

DATA PIPELINE

STREAMING WEB LOG INGESTION

Collect all logs.

Pipeline steps:

1. Generate random logs in text format and store in **datalake**.
2. Ingest the data in web_logs_raw into the data **warehouse**.
3. Make the necessary manipulations in `web_logs_raw` and storage in `web_logs` table.

FEATURES

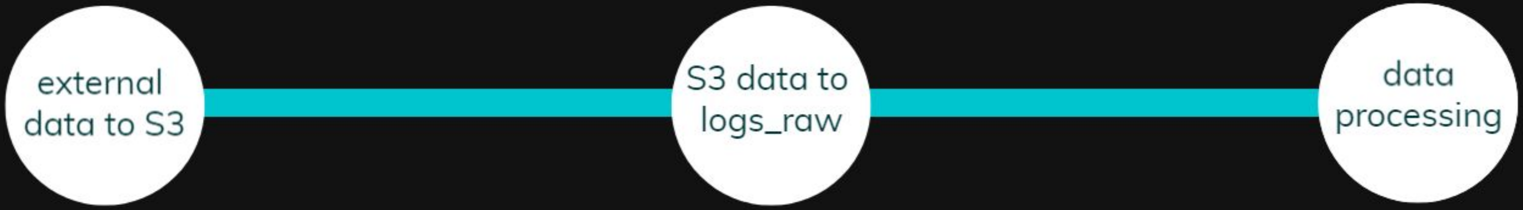
Data avalaible on data lake.

Uses Autoloader.

Uses an mmdb file to get the country from the IP.

Retry and e-mail notification on failure.

Big Data manipulation: +700M rows.



METHODOLOGY

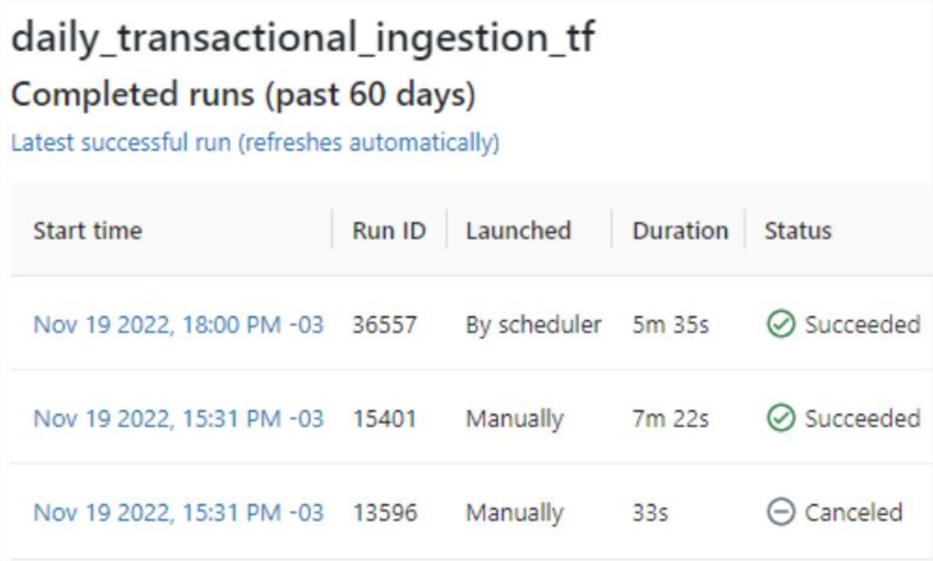
- ARCHITECTURE ✓
- DATA MODELING ✓
- DATA PIPELINE ✓
- DATA TRACKING
- TESTS
- MARKETING LEAD SHEET
- INFRASTRUCTURE AS CODE

DATA TRACKING

PIPELINES AND DATA TRACKING INFORMATION

Databricks UI

- each job generates a history with the information of each execution



daily_transactional_ingestion_tf

Completed runs (past 60 days)

Latest successful run (refreshes automatically)

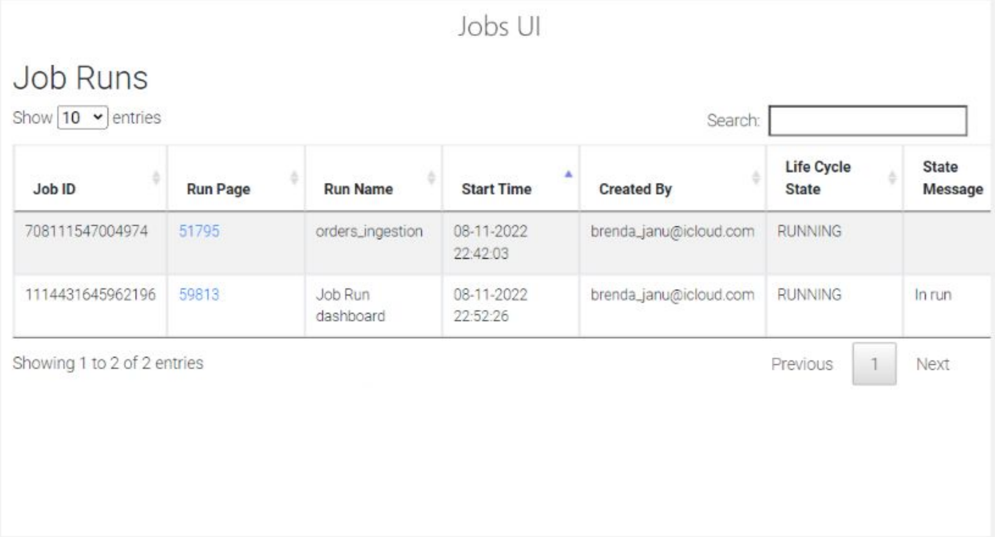
Start time	Run ID	Launched	Duration	Status
Nov 19 2022, 18:00 PM -03	36557	By scheduler	5m 35s	✓ Succeeded
Nov 19 2022, 15:31 PM -03	15401	Manually	7m 22s	✓ Succeeded
Nov 19 2022, 15:31 PM -03	13596	Manually	33s	⊖ Canceled

DATA TRACKING

PIPELINES AND DATA TRACKING INFORMATION

Job Run dashboard

- panel with the information about the execution of the jobs.



Jobs UI

Job Runs

Show 10 entries

Search:

Job ID	Run Page	Run Name	Start Time	Created By	Life Cycle State	State Message
708111547004974	51795	orders_ingestion	08-11-2022 22:42:03	brenda_janu@icloud.com	RUNNING	
1114431645962196	59813	Job Run dashboard	08-11-2022 22:52:26	brenda_janu@icloud.com	RUNNING	In run

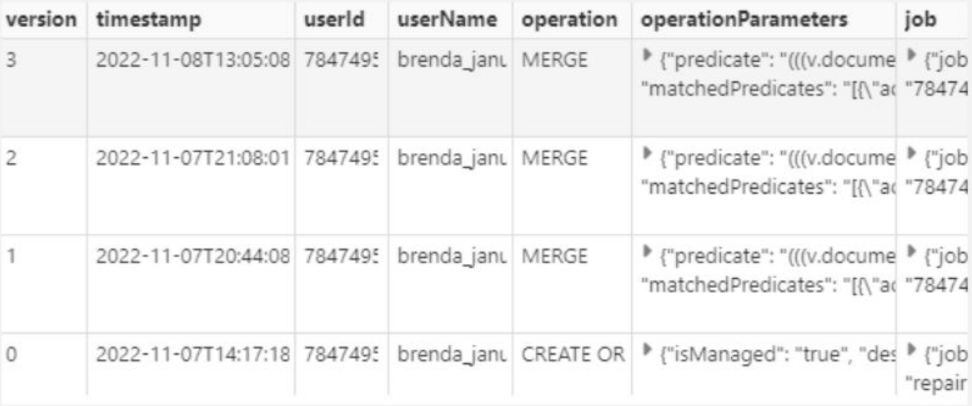
Showing 1 to 2 of 2 entries

DATA TRACKING

PIPELINES AND DATA TRACKING INFORMATION

Change Data Feed (CDC)

- tables with CDC enable to consult the entire history of changes in a table, informing the operation, the time and the user/job that made the change.



version	timestamp	userId	userName	operation	operationParameters	job
3	2022-11-08T13:05:08	784749!	brenda_janu	MERGE	▸ {"predicate": "(((v.docume "matchedPredicates": "[{\"ac	▸ {"job "78474
2	2022-11-07T21:08:01	784749!	brenda_janu	MERGE	▸ {"predicate": "(((v.docume "matchedPredicates": "[{\"ac	▸ {"job "78474
1	2022-11-07T20:44:08	784749!	brenda_janu	MERGE	▸ {"predicate": "(((v.docume "matchedPredicates": "[{\"ac	▸ {"job "78474
0	2022-11-07T14:17:18	784749!	brenda_janu	CREATE OR	▸ {"isManaged": "true", "des	▸ {"job "repair

METHODOLOGY

- ARCHITECTURE ✓
- DATA MODELING ✓
- DATA PIPELINE ✓
- DATA TRACKING ✓
- TESTS
- MARKETING LEAD SHEET
- INFRASTRUCTURE AS CODE

Data Tests

DATA QUALITY WITH GREAT EXPECTATIONS

Great Expectations for tests and data quality.
It helps data teams eliminate pipeline debt, through data testing, documentation, and profiling.
The execution results are saved in a history and can be displayed in html.

- For this project the tests results are in great_expectations/index.html

great_expectations

EXPECTATIONS

Tests included in the Orders data pipeline

orders data

TEST

EXPECTED COLUMN VALUES TO NOT BE NULL:

- customer_birth_day
- customer_document_number
- customer_name
- product_id
- order_dt
- order_id
- status
- order_total_price
- record_timestamp

EXPECTED COLUMN PAIR VALUES A TO BE GREATER THAN B:

- order_total_price > item_price_default
- item_price > Item_price_default

EXPECTED COLUMN MAX TO BE BETWEEN:

- 0 < order_total_price

METHODOLOGY

- ARCHITECTURE ✓
- DATA MODELING ✓
- DATA PIPELINE ✓
- DATA TRACKING ✓
- TESTS ✓
- MARKETING LEAD SHEET
- INFRASTRUCTURE AS CODE

Marketing Lead Sheet

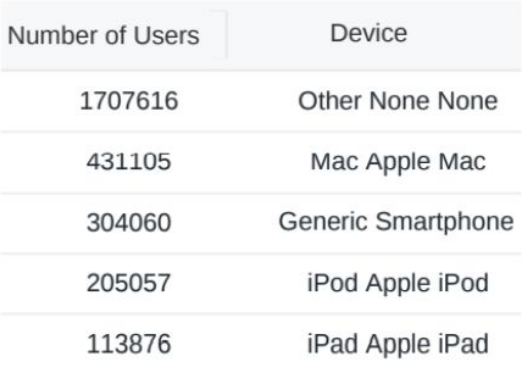
DASHBOARD ANALYSIS

Panel with dashboards and tabular information to data analysis using SQL Query Editor on Databricks

MARKET LEAD VIEWS

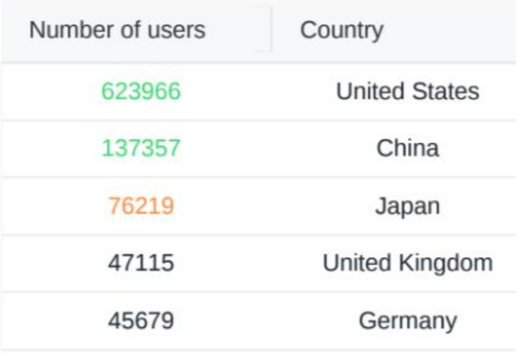
MAIN DEVICES USED

Tóp 5 most popular used devices for B2B clients



Number of Users	Device
1707616	Other None None
431105	Mac Apple Mac
304060	Generic Smartphone
205057	iPod Apple iPod
113876	iPad Apple iPad

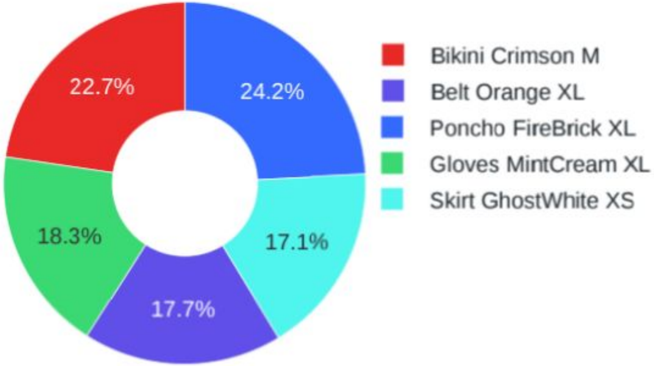
COUNTRIES WITH MORE LOGGED USERS



Number of users	Country
623966	United States
137357	China
76219	Japan
47115	United Kingdom
45679	Germany

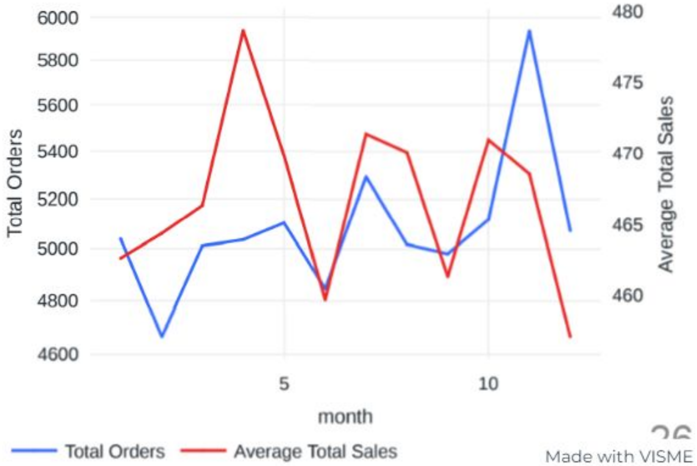
MOST VIEWED PRODUCTS

Top 5 most popular products in the country from which most users log into



MONTHLY SALES LAST YEAR

All sales of B2B platform displayed monthly for the last year



METHODOLOGY

- ARCHITECTURE ✓
- DATA MODELING ✓
- DATA PIPELINE ✓
- DATA TRACKING ✓
- TESTS ✓
- MARKETING LEAD SHEET ✓
- INFRASTRUCTURE AS CODE

Infrastructure As Code

PIPELINE AND CLUSTER SETTINGS

The Infrastructure as code in **terraform** to generate automatically the structure of the pipelines and cluster settings.

The state file (.tfstate) configured to store inside the datalake to safely find the resources created previously and update them properly.

METHODOLOGY

- ARCHITECTURE ✓
- DATA MODELING ✓
- DATA PIPELINE ✓
- DATA TRACKING ✓
- TESTS ✓
- MARKETING LEAD SHEET ✓
- INFRASTRUCTURE AS CODE ✓

FINISH

THANK YOU!

DO YOU HAVE ANY QUESTIONS?