

# DATA CLEANING WITH SQL AND ML

LIVESPORT

DATA  
DAY

20.11.2024

# Game changing numbers

**100 M+** users worldwide

**28 M** registered users

**150 M+** apps downloaded

**50** languages

**70** websites

**16 BLN+** impressions per month

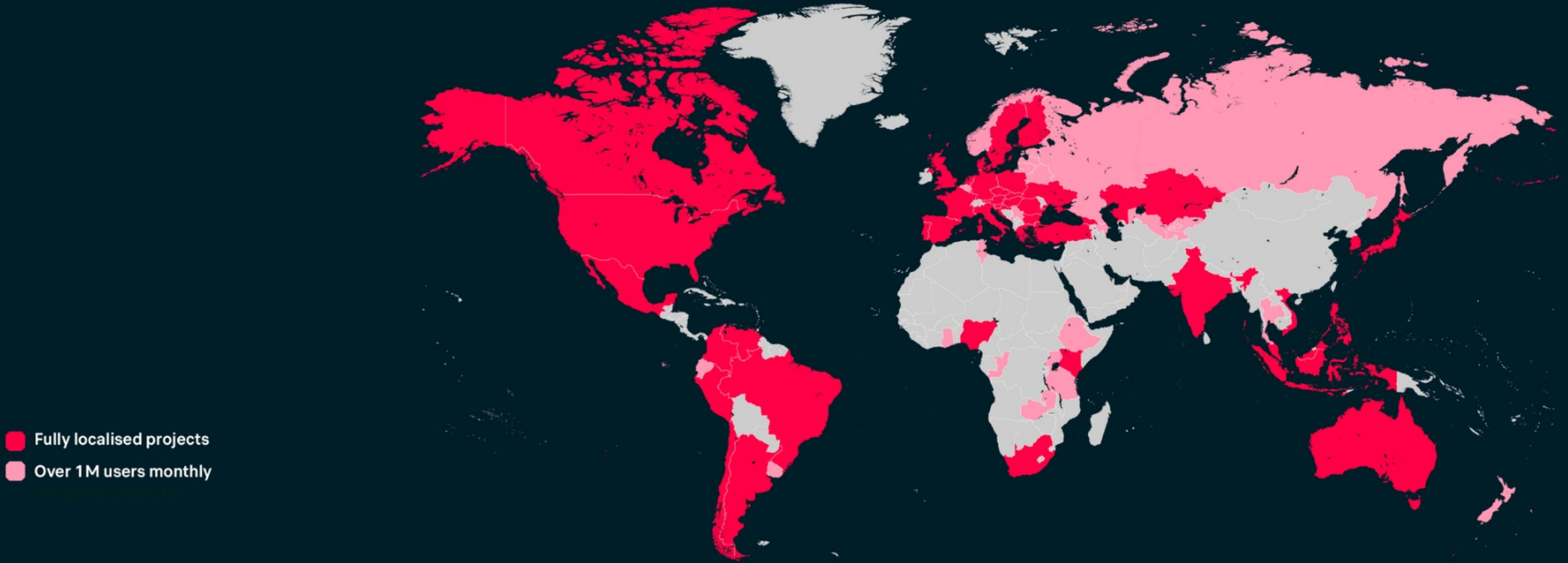
**4.8** iOS rating

**4.78** Android rating

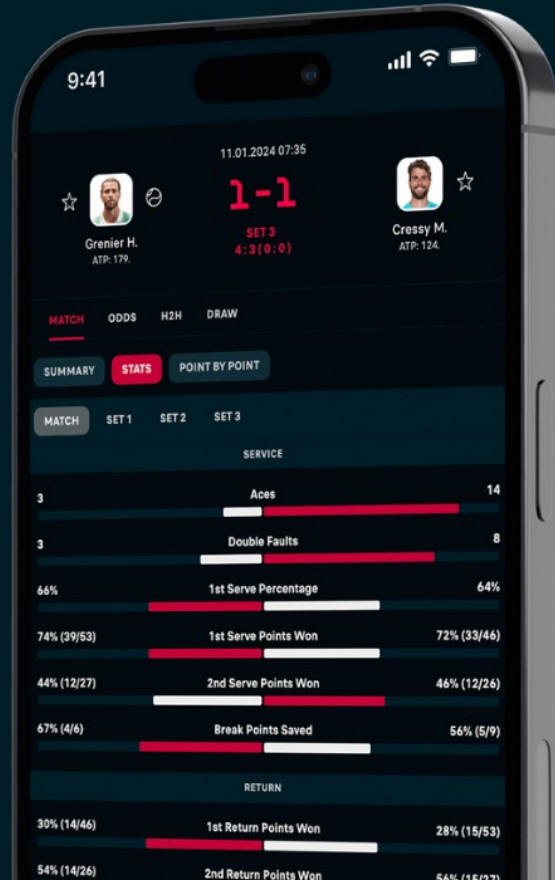
**1 M+** reviews

# Global product for local users

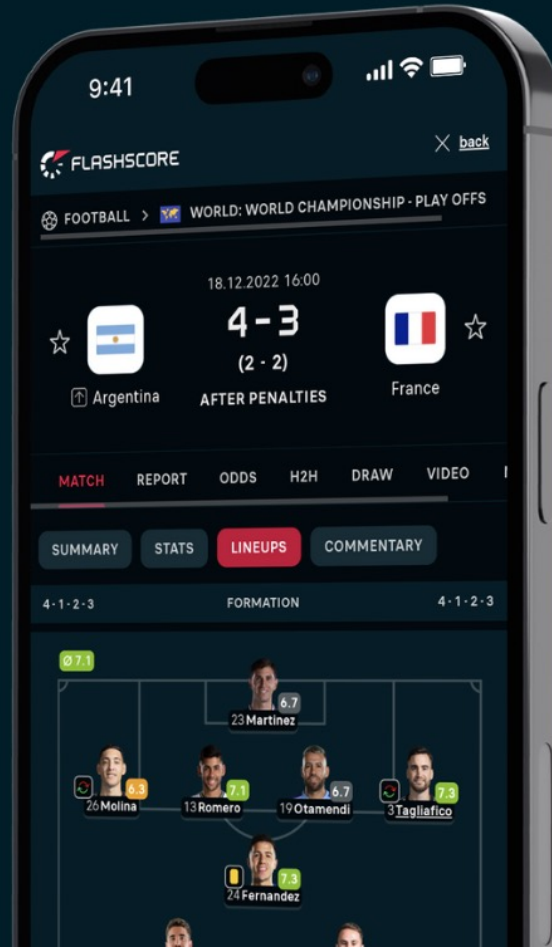
Even though we are global, we try to bring the best possible localised experience to the most important markets. Markets highlighted on the map below are those with full localisation



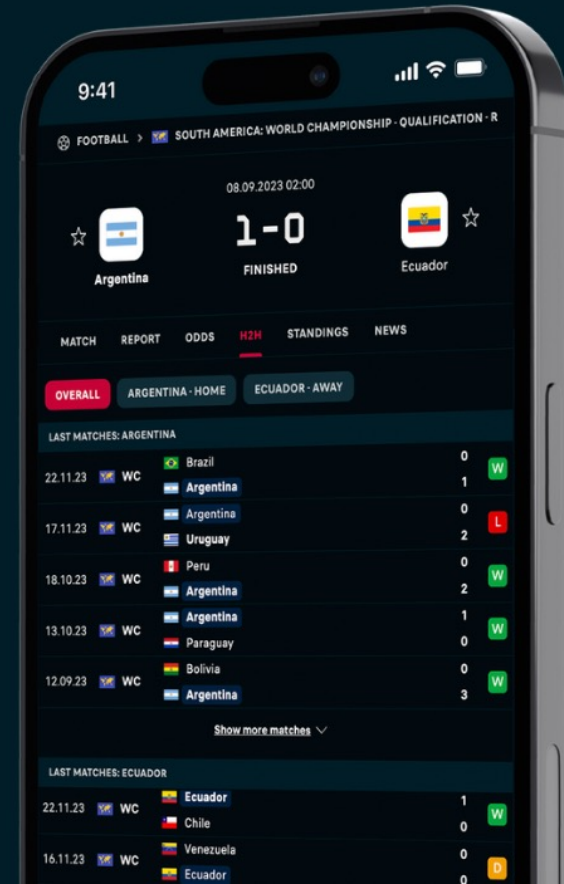
# LIVE STATS



# PLAYER RATINGS

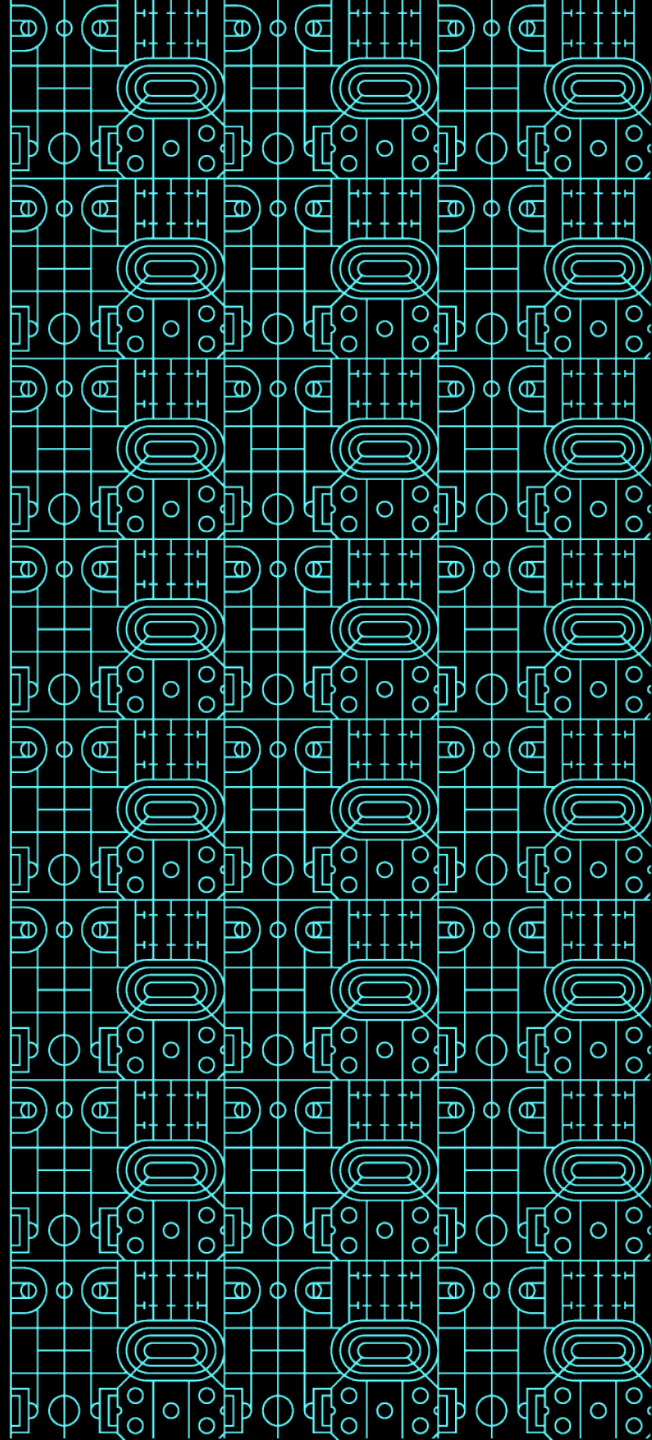


# H2H



# BUSINESS MODEL

LIVESPORT



# AFFILIATE BUSINESS MODEL

**Tipsport** 300 Kč jen za registraci CHCI BONUS

FOTBAL > EVROPA: LIGA MISTRŮ - LIGOVÁ FÁZE - 4. KOLO

05.11.2024 21:00

Real Madrid - AC Milán

PŘEHLED KURZY H2H TABULKA

PŘEDZÁPASOVÉ KURZY

Kurz	1	2
Tipsport	1.52	5.28
Betano	1.53	5.20

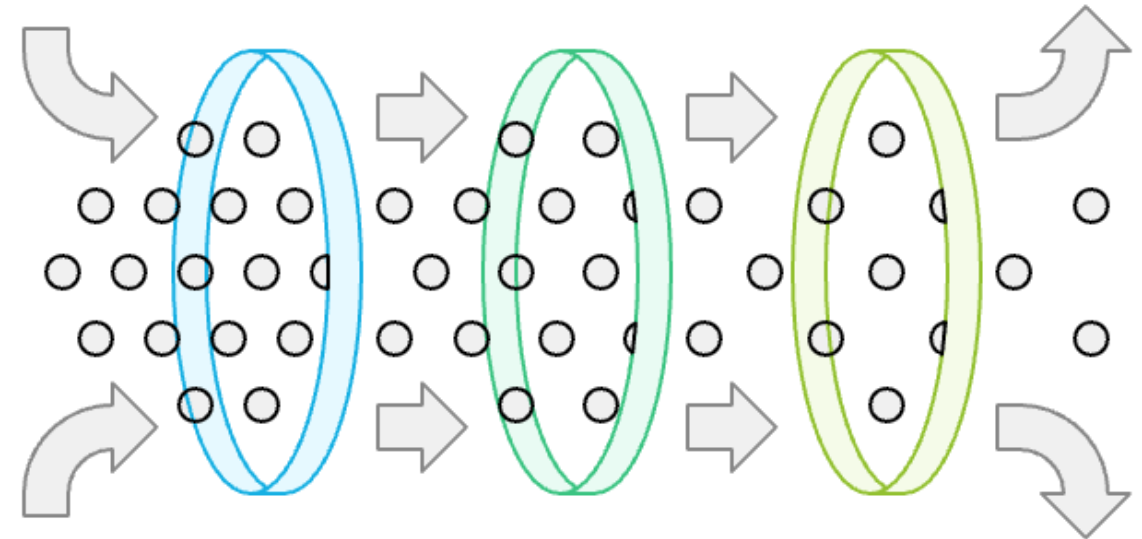
Zobrazit více

TIPSPORT - 300 Kč zdarma jen za registraci + bonus až 50 000 Kč k prvnímu vkladu! Chci bonus

BETANO - Exkluzivně všechny zápasy 1.ligy živé + 350 Kč zdarma za registraci! Registrovat

18+ Ministerstvo financí varuje: Účastí na hazardní hře může vzniknout závislost!

iGaming Affiliate Marketing Funnel



## Campaign Engagement

Users interact with the affiliate campaign

## User Conversion

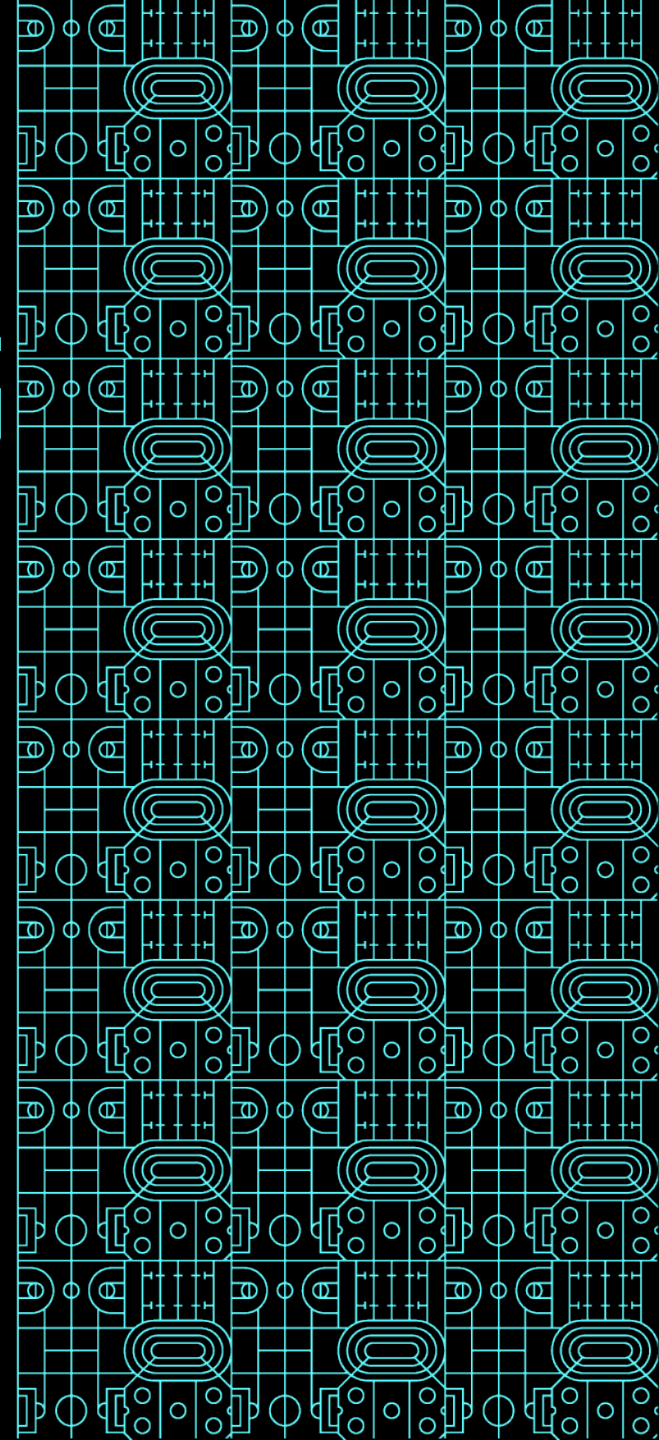
Users complete desired actions

## Revenue Sharing

Revenue is distributed based on engagement

# USE CASE: STATES REPORTING

## BRAZIL





# Brazil : Main Information

FILTERS

COUNTRY

Brazil



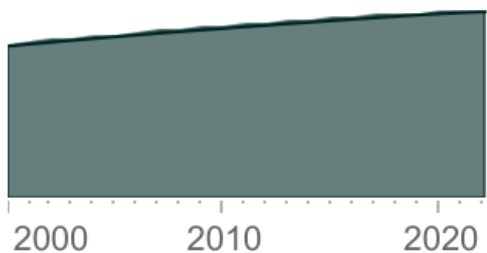
## BACKGROUND

Population

**215 313 498**

World Rank: 7

YoY: ▲ 0,5%

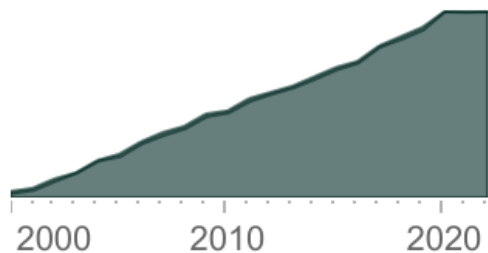


Online Population

**173 387 117**

World Rank: 5

81 % of pop.

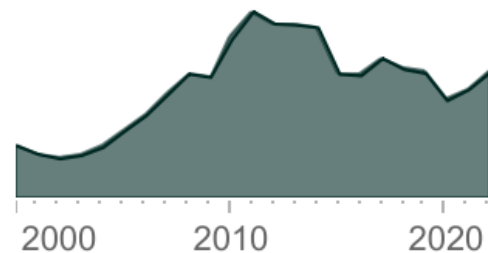


GDP per Capita

**\$8 918**

World Rank: 98

YoY: ▲ 15,9%



Regime Type

Deficient Democracy

Languages

Portuguese (official and most widely spoken language); note - less common languages include Spanish (border areas and schools), German, Italian, Japanese, English, and a large numbe..

Betting Regulation

Online

non-regulated

Retail

regulated

# REVENUE PER COUNTRY - NOT ENOUGH

## LOCAL MARKETING CAMPAIGNS

FEDERATIONS/STATES (26+1)

POPULATION: 215M / 173

## ROI/ROAS PER STATE

DETAILED STATES REPORTS

PLAYER LIFETIME VALUE, - ACTIVE PLAYERS



# CLIENT DATA: GARBAGE IN -> ?

INPUT: THOUSANDS (+12K) DISTINCT "STRING" VALUES

as_of	Affiliate Code	UserID	Signup Date	City	First Deposit Date	First Deposit Amount
2024-07-01	365_00917297	107811588	30/03/2024 14:13:39	Cuiabá	30/03/2024 14:28:05	79
2024-07-01	365_00924972	52903910	19/02/2020 12:13:05	CASTELO	20/02/2020 07:27:24	17
2024-07-01	365_00924972	53276853	04/03/2020 21:45:59	ARACAJU	06/03/2020 07:12:27	5
2024-07-01	365_00924972	59023431	08/11/2020 16:48:24	SENHOR DO BONFIM	10/11/2020 07:59:49	71
2024-07-01	365_00924972	72396605	12/10/2021 21:26:11	BARRA MANSA	11/11/2021 06:38:26	4
2024-07-01	365_00926825	56235000	17/08/2020 14:26:58	RECIFE	18/08/2020 09:54:13	4
2024-07-01	365_00926825	58695200	31/10/2020 14:12:43	Campina Grande	09/05/2022 20:27:06	4
2024-07-01	365_00926825	107078845	04/03/2024 23:31:09	Joinville	05/03/2024 14:34:50	4
2024-07-01	365_00926825	107443955	15/03/2024 16:53:55	Camargibe	15/03/2024 16:57:37	7



# TYPOS AND MISSING STATES

## EXPORTED DATA FROM THE CLIENT

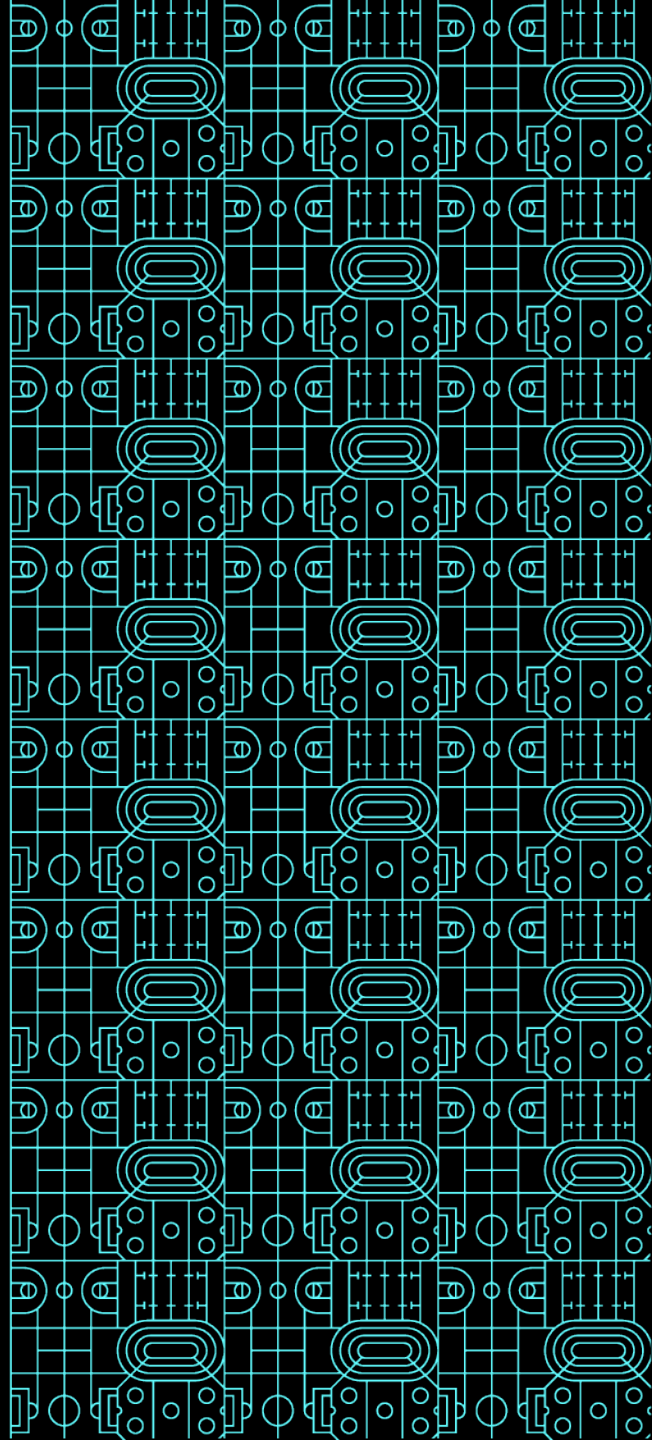
NO DATA ABOUT BR STATES

CITIES – MANUALLY ENTERED BY USERS

DISTINCT VALUES: 12 118

city ▾
Sao luis gonzaga
Sao miguel
Sao paulo
Sao pedro da aldeia
Sao vicente
SaoPauloSP
VILA SAO FRANCISCO DE ASSISTELEMACO BORBA
paco do lumiar - sao luis
sao Paulo
sao bento
sao goncalo
sao goncalo do amarante
sao gonsalo
sao gonçalo

# USE CASE: OBJECTIVE



**NO DATA DEVELOPMENT**

**CLEAN DATA**

**ASSIGN THE CITY TO THE STATE**

**SEAMLESS SOLUTION**

**WITHOUT DATA ENGINEERS**

**LOW-CODE/NO-CODE**

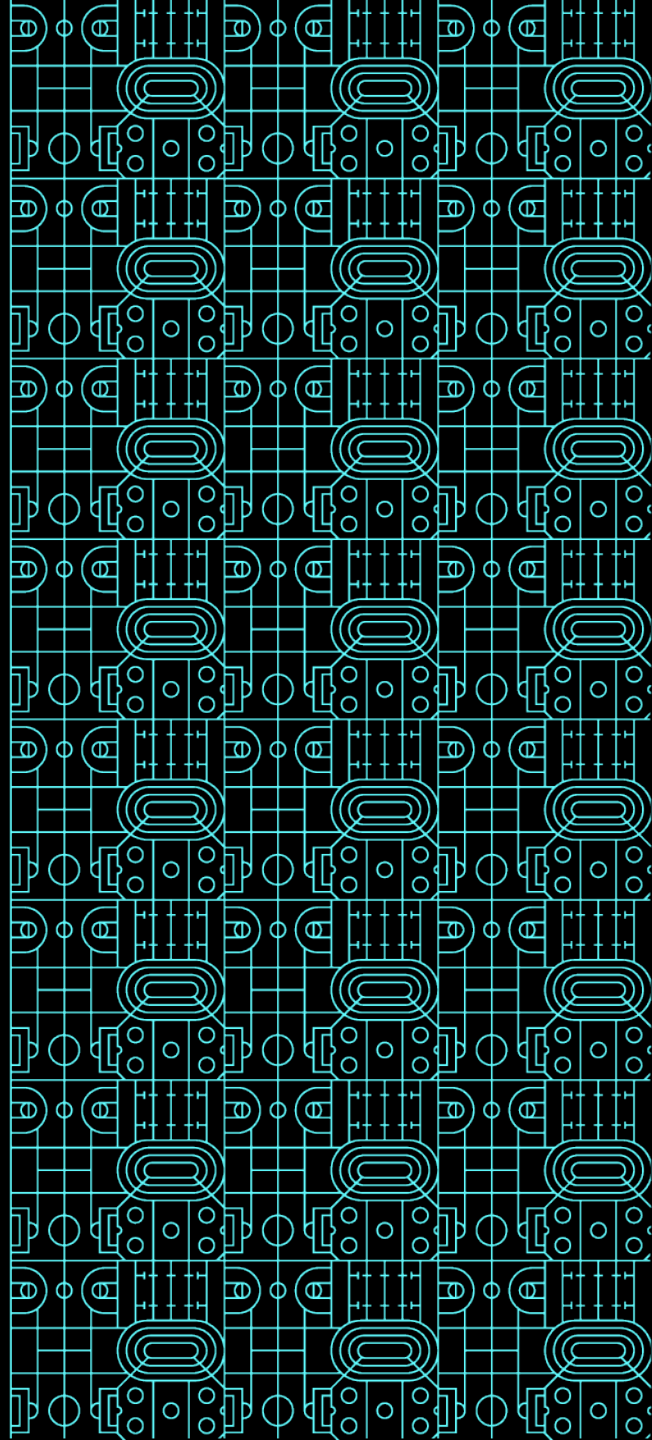
# COMMON METHODS ARE USELESS

LOWERCASE, TRIM, NORMALIZE, DISTINCT

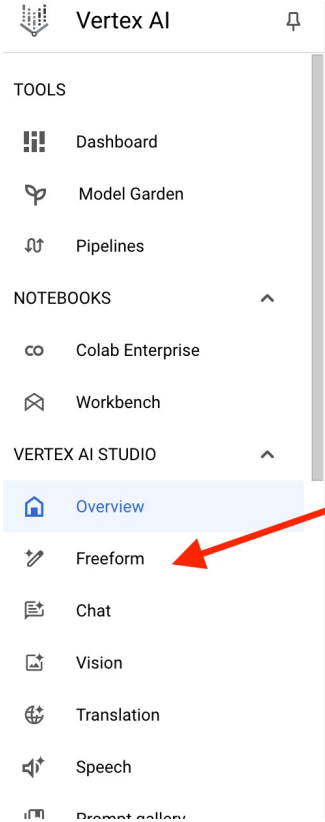
CASE – INFINITE SOLUTIONS

PROBLEM: TYPOS, DIFF VERSION OF CITY NAME

# USE CASE: ENVIRONMENT



# VERTEX AI STUDIO – FREEFORM (SANDBOX)



Vertex AI

TOOLS

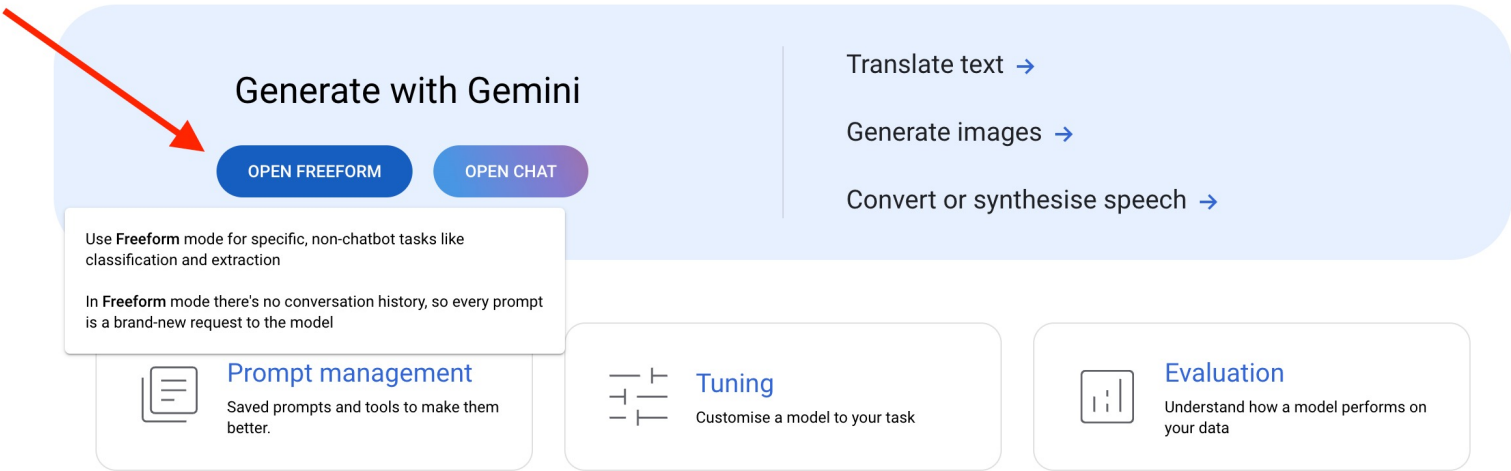
- Dashboard
- Model Garden
- Pipelines

NOTEBOOKS

- Colab Enterprise
- Workbench

VERTEX AI STUDIO

- Overview
- Freeform
- Chat
- Vision
- Translation
- Speech
- Prompt gallery



## Vertex AI Studio

Test, tune and deploy enterprise-ready generative AI

[TRY A TUTORIAL](#) [DOCUMENTATION](#) [API REFERENCE](#)

### Generate with Gemini

[OPEN FREEFORM](#) [OPEN CHAT](#)

Translate text →  
Generate images →  
Convert or synthesise speech →

**Prompt management**  
Saved prompts and tools to make them better.

**Tuning**  
Customise a model to your task

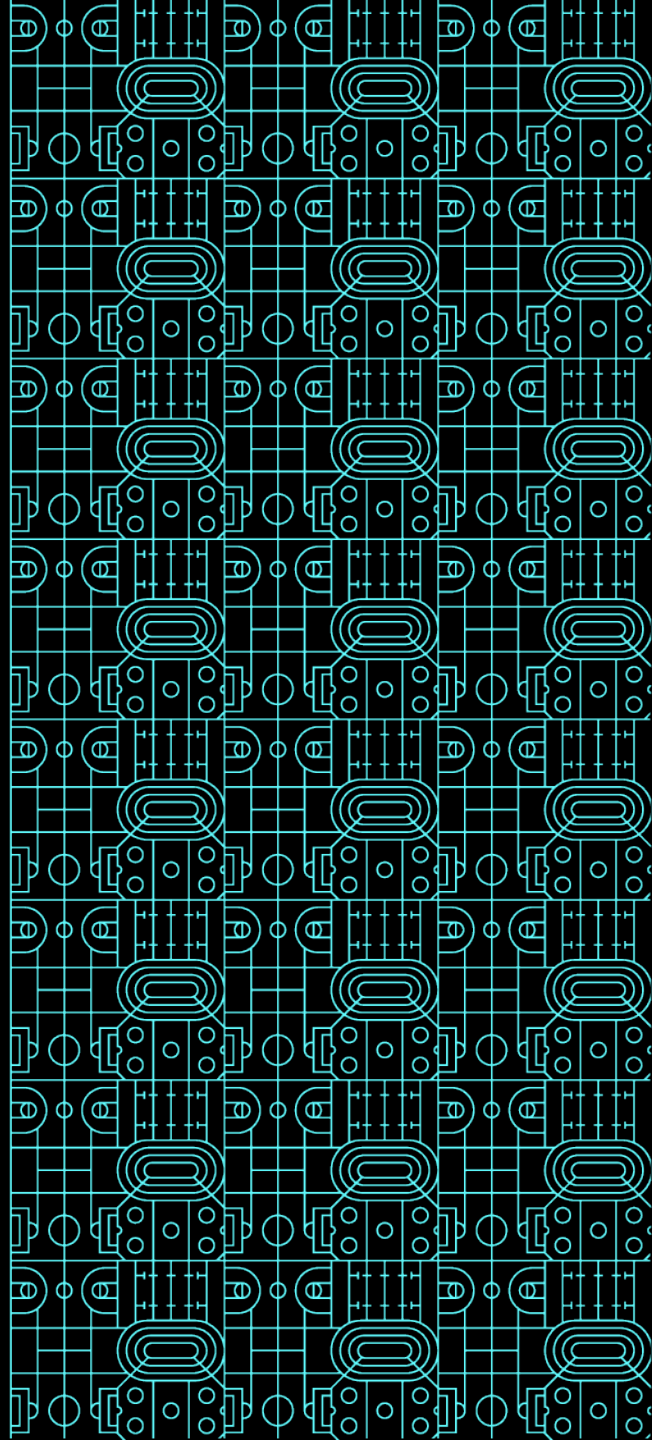
**Evaluation**  
Understand how a model performs on your data

**Freeform mode**  
Use Freeform mode for specific, non-chatbot tasks like classification and extraction  
In Freeform mode there's no conversation history, so every prompt is a brand-new request to the model

# GOOD SANDBOX FOR PLAYING

The screenshot displays the Vertex AI Freeform interface. On the left is a navigation sidebar with categories: TOOLS (Dashboard, Model Garden, Pipelines), NOTEBOOKS (Colab Enterprise, Workbench), VERTEX AI STUDIO (Overview, Freeform, Chat, Vision, Translation, Speech, Prompt gallery, Prompt management, Tuning), BUILD WITH GEN AI (Extensions), and DATA (Feature Store, Datasets, Marketplace). The main area is titled 'Untitled prompt' and contains a 'System instructions' section with the text 'Give the model context to understand the task and provide tailored responses'. Below this is a 'Prompt' input field with a placeholder 'Write a prompt, or create one with Help me write'. The 'Response' section shows three suggested actions: 'Question answering about an image', 'Extract video chapters', and 'Document classification'. On the right, a configuration panel includes settings for Model (gemini-1.5-flash-002), Region (us-central1 (Iowa)), Temperature (1), Output token limit (8192), Grounding (Source: Google Search), Add stop sequence, Output format (Plain text), Safety filter settings, and Advanced settings (Seed, Max. responses, Top-k, Top-p, Stream model responses).

# USE CASE: PROMPT DESIGN



# FIRST – FAILED BECAUSE OF BAD CHOICE

## System instructions [?](#)

Give the model context to understand the task and provide tailored responses

## Prompt

[Insert media](#) [Add examples](#) [Add variable](#) [Clear prompt](#)

You are a text standardizer. From the following text that contains the names of Brazilian municipalities, cities and towns, standardize and print one standardized, consolidated name in Portuguese language, also include the full name of the Brazilian state to which the particular city belongs tom seperated from the city by a single comma. You cannot return empty because you know how to pick up sensible data from the text in this field `{city}` CAICO,CAIRU,CAMET,CANDI,CANOO,CAPIM,C...

125 tokens

## Response

[Refine response](#)  Markdown [Copy](#)

The provided text contains abbreviations and non-standard entries that cannot be reliably mapped to Brazilian municipalities. To standardize, I need a list of Brazilian cities and their corresponding states. The abbreviations (Re, Rj, Rs, SG, SP, Sp, UB, kk) are insufficient for identification.

Please provide a complete list of the intended municipalities and their states to enable proper standardization.



Model:  [?](#)

Region:  [?](#)

Temperature:

Output token limit:

Grounding:  Source: Google Search

Customise

Add stop sequence [?](#)

Press Enter after each sequence

Output format:  [?](#) [Edit](#)

[Safety filter settings](#)

Advanced

Seed [?](#)

Max. responses:

Top-k:

Top-p:

Stream model responses  
Print responses as they're generated

BAD  
MODEL

# BETTER MODEL BUT NOT ENOUGH

## System instructions [?](#)

Give the model context to understand the task and provide tailored responses

## Prompt

[Insert media](#) [?](#) [Add examples](#) [Add variable](#) [Clear prompt](#)

You are a text standardizer. From the following text that contains the names of Brazilian municipalities, cities and towns, standardize and print one standardized, consolidated name in Portuguese language, also include the full name of the Brazilian state to which the particular city belongs tom seperated from the city by a single comma. You cannot return empty because you know how to pick up sensible data from the text in this field `{city}` `CAICO,CAIRU,CAMET,CANDI,CANOO,CAPIM,C...`

125 tokens



## Response

[Refine response](#)  Markdown [Copy](#)

- Caicó, Rio Grande do Norte
- Cairu, Bahia
- Cametá, Pará
- Candiba, Bahia
- Canoas, Rio Grande do Sul
- Capim Grosso, Bahia
- Carmo do Rio Claro, Minas Gerais
- Cascavel, Paraná (Most likely, given the other entries are Brazilian municipalities. There's also a Casca, Rio Grande do Sul, but Cascavel is much larger and better known.)
- Caçu, Goiás



Model  
gemini-1.5-pro-002 [?](#)

Region \*  
europe-west3 (Frankfurt) [?](#)

Temperature [?](#)

0  2

Output token limit [?](#)

1  8192

Grounding [?](#)

Source: Google Search

Customise

Add stop sequence [?](#)

Press Enter after each sequence

Output format [?](#)  
Plain text [?](#) [Edit](#)

[Safety filter settings](#)

Advanced

Seed [?](#)

Max. responses [?](#)

1  1

Top-k [?](#)

1  40

Top-p [?](#)

0  1

Stream model responses  
Print responses as they're generated

# PREDICTABILITY AND CREATIVITY

**System instructions** ⓘ  
Give the model context to understand the task and provide tailored responses

**Prompt**

You are a text standardizer. From the following text that contains the names of Brazilian municipalities, cities and towns, standardize and print one standardized, consolidated name in Portuguese language, also include the full name of the Brazilian state to which the particular city belongs tom seperated from the city by a single comma. You cannot return empty because you know how to pick up sensible data from the text in this field `{city} CAICO,CAIRU,CAMET,CANDI,CANOO,CAPIM,C...`

125 tokens

**Response**

- Caicó, Rio Grande do Norte
- Cairu, Bahia
- Cametá, Pará
- Candiba, Bahia
- Canoas, Rio Grande do Sul
- Capim Grosso, Bahia
- Carmo do Rio Claro, Minas Gerais
- Casca, Rio Grande do Sul
- Caçu, Goiás

Model  ⓘ

Region \*  ⓘ

Temperature ⓘ  
0 2

Output token limit ⓘ  
1 8192

Grounding ⓘ  
 Source: Google Search

Customise

Add stop sequence ⓘ  
Press Enter after each sequence

Output format  ⓘ Edit

Safety filter settings

Advanced

Seed ⓘ

Max. responses ⓘ  
1 1

Top-k ⓘ  
1 40

Top-p ⓘ  
0 1

CHANGE PARAMETERS

# USE POWER OF VERTEX AI & GEMINI

## Prompt



System Instruction

You are a world-class astronomer with an expertise on stars. You are designed to answer questions on astronomy for children. If the answer cannot be found in the context, respond with "Answer not found in context."

Below is an example of how you must provide the answer:

Question: The sun is a star. What is the sun?  
Answer: A star.

Few-Shot Examples  
(Demonstrations)

Contextual  
Information

Stars form in large clouds of gas and dust called molecular clouds. Molecular clouds range from 1,000 to 10 million times the mass of the Sun and can span as much as hundreds of light-years. Molecular clouds are cold which causes gas to clump, creating high-density pockets. Some of these clumps can collide with each other or collect more matter, strengthening their gravitational force as their mass grows. Eventually, gravity causes some of these clumps to collapse. When this happens, friction causes the material to heat up, which eventually leads to the development of a protostar – a baby star. Batches of stars that have recently formed from molecular clouds are often called stellar clusters, and molecular clouds full of stellar clusters are called stellar nurseries.

Answer the following question based on the provided context: Where are stars formed?

Task

# USE AI TO GENERATE PROMPT

 Help me write Preview

Create better prompts based on a description of your intent

Prompt intent

Write your prompt here or use an example.

Brainstorm activities Music recommendation bot Write thank you letter

Generate prompt

Suggested prompt

A new prompt will appear here after clicking **Generate prompt**

# CONTEXT + EXAMPLES: ALMOST THE BEST

**System instructions** ⓘ 🗑️

You are a data analyst tasked with cleaning a dataset of Brazilian city names. The dataset contains user-generated input, which may be messy, contain typos, abbreviations, or extra information. Your goal is to standardize the city names and identify the correct state.

**Prompt** [Insert media](#) ⓘ [Add examples](#) [Add variable](#) [Clear prompt](#)

Follow these steps:

- Analyze the input:** Carefully examine the user-provided city name.
- Identify the state:**
  - If the input includes the name of a Brazilian state (e.g., "Maranhão"), use it to determine the state. The city name should be the capital of that state (e.g., "São Luís, Maranhão").
  - If the input includes only a district, neighborhood, or administrative region (e.g., "Irajário de Janeiro"), use it as a hint to identify the city and state (e.g., "Rio de Janeiro, Rio de Janeiro").

407 tokens

**Response** [Refine response](#)  Markdown 📄

Let's apply these steps to a few examples:

**Example 1:**

- Input:** "Sao Paulo"
- Analysis:** Common misspelling of the city and state "São Paulo".
- Output:** "São Paulo, São Paulo"

**Example 2:**

- Input:** "Rio"
- Analysis:** Abbreviation for "Rio de Janeiro".
- Output:** "Rio de Janeiro, Rio de Janeiro"

**Example 3:**

- Input:** "Maranhão"
- Analysis:** This is a state name. The city should be the capital.
- Output:** "São Luís, Maranhão"

Model: gemini-1.5-pro-002 ⓘ

Region: europe-west3 (Frankfurt) ⓘ

Temperature: 0.2

Output token limit: 8192

Grounding:  Source: Google Search

Customise

Add stop sequence ⓘ

Press Enter after each sequence

Output format: Plain text ⓘ Edit

[Safety filter settings](#)

Advanced

Seed ⓘ

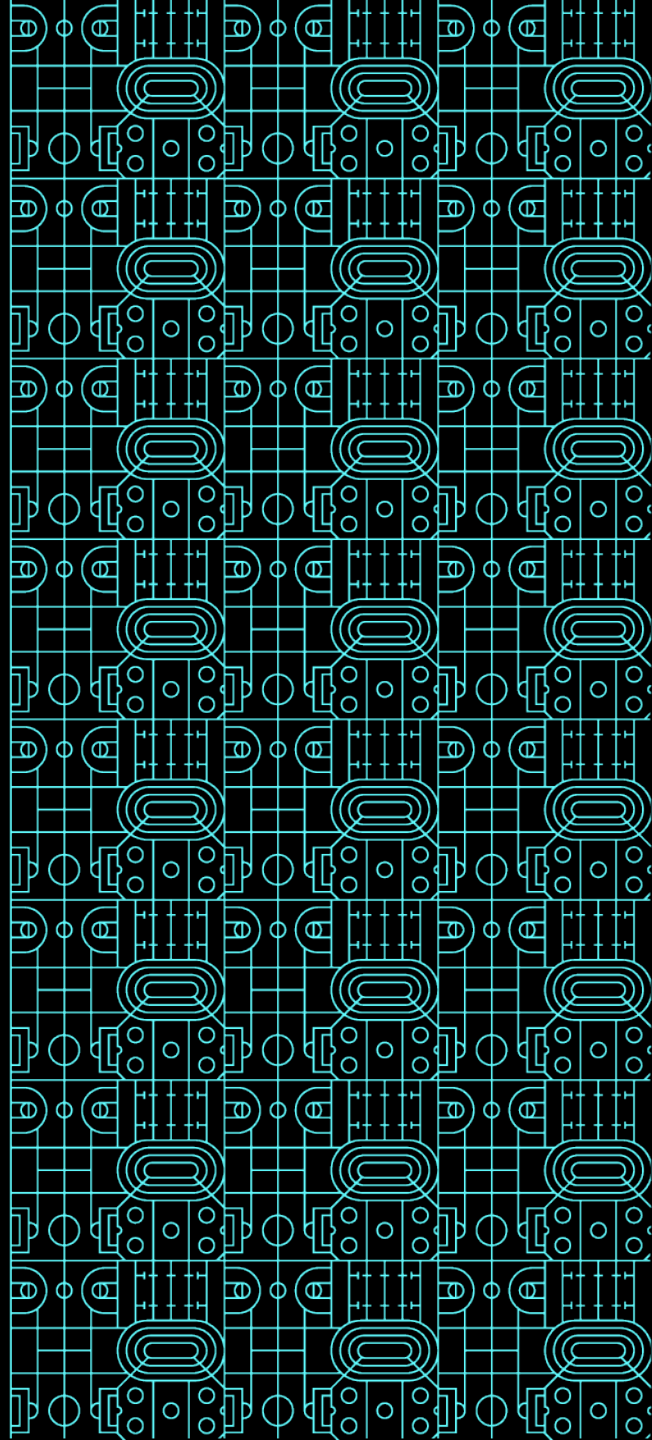
Max. responses: N/A

Top-k: N/A

Top-p: 0.1

Stream model responses  
Print responses as they're generated

# USE CASE: PROMPT ENGINEERING



# REFINE RESPONSE – GENERATE NEW PROMPT

## Refine response

Describe what you want the model to respond and get prompt suggestion

### Response feedback

Make shorter   Make longer   More professional   More casual

structure the prompt more, use context, break it into parts and give examples

Generate prompt

### Original prompt

Follow these steps:

1. **Analyze the input:** Carefully examine the user-provided city name.

### Suggested prompt

A new prompt will appear here after clicking Generate prompt

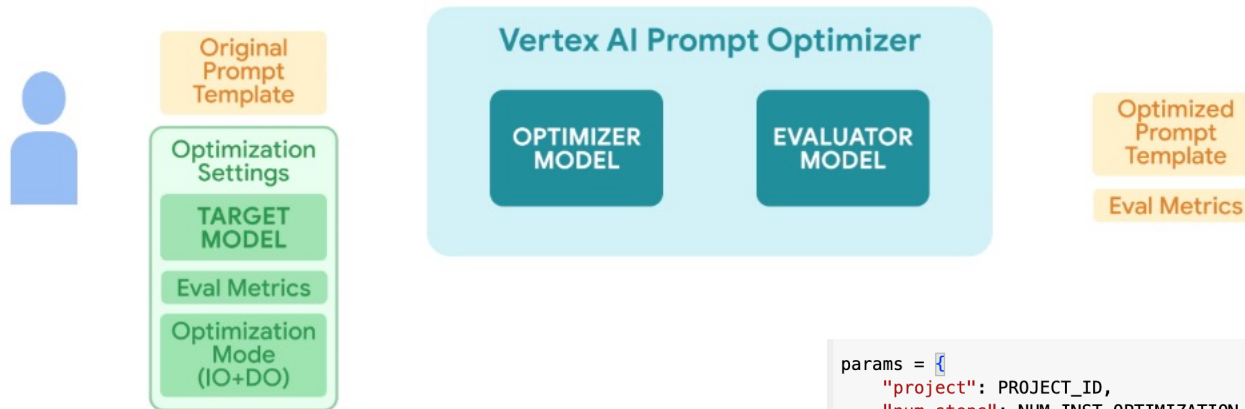


Optimise your prompt further 

Cancel

Apply and run

# VERTEX PROMPT OPTIMIZER



```
params = {  
  "project": PROJECT_ID,  
  "num_steps": NUM_INST_OPTIMIZATION_STEPS,  
  "system_instruction": SYSTEM_INSTRUCTION,  
  "prompt_template": PROMPT_TEMPLATE,  
  "target_model": TARGET_MODEL,  
  "target_model_qps": TARGET_MODEL_QPS,  
  "target_model_location": LOCATION,  
  "source_model": SOURCE_MODEL,  
  "source_model_qps": SOURCE_MODEL_QPS,  
  "source_model_location": LOCATION,  
  "eval_qps": EVAL_QPS,  
  "eval_model_location": LOCATION,  
  "optimization_mode": OPTIMIZATION_MODE,  
  "num_demo_set_candidates": NUM_DEMO_OPTIMIZATION_STEPS,  
  "demo_set_size": NUM_DEMO_PER_PROMPT,  
  "aggregation_type": METRIC_AGGREGATION_TYPE,  
  "data_limit": 50,  
  "num_template_eval_per_step": NUM_TEMPLATES_PER_STEP,  
  "input_data_path": input_data_path,  
  "output_path": output_path,  
  "response_mime_type": RESPONSE_MIME_TYPE,  
  "response_schema": RESPONSE_SCHEMA,  
  "language": TARGET_LANGUAGE,  
  "placeholder_to_content": json.loads(PLACEHOLDER_TO_VALUE),  
}
```

# THE BEST RESULTS WITH GROUNDING

## System instructions

You are a data analyst tasked with cleaning a dataset of Brazilian city names. The dataset contains user-generated input, which may be messy, contain typos, abbreviations, or extra information. Your goal is to standardize the city names and identify the correct state.

### 2. **Identify the state:**

\* If the input includes the name of a Brazilian state (e.g., "Maranhão"), use it to determine the state. The city name should be the capital of that state (e.g., "São Luís, Maranhão").

\* If the input includes only a district, neighborhood, or administrative region (e.g., "Irajário de Janeiro"), use it as a hint to identify the city and state (e.g., "Rio de Janeiro, Rio de Janeiro").

407 tokens


## Response

 Refine response

Markdown 


- **CAICO:** Caicó, Rio Grande do Norte
- **CAIRU:** Cairu, Bahia
- **CAMET:** Cametá, Pará
- **CANDI:** Candiba, Bahia
- **CANOO:** Canoas, Rio Grande do Sul
- **CAPIM:** Capim Grosso, Bahia
- **CARMO:** Carmo, Rio de Janeiro
- **CASCA:** Casca, Rio Grande do Sul
- **CAÇU:** Caçu, Goiás
- **Re:** Recife, Pernambuco
- **Rj:** Rio de Janeiro, Rio de Janeiro
- **Rs:** Porto Alegre, Rio Grande do Sul (assuming "Rs" refers to the state abbreviation)
- **SG:** São Gonçalo, Rio de Janeiro (assuming "SG" is an abbreviation, as there's no Brazilian state with this abbreviation.)
- **SP:** São Paulo, São Paulo
- **Sp:** São Paulo, São Paulo
- **UB:** Uberlândia, Minas Gerais (assuming "UB" is an abbreviation)
- **kk:** Unknown, Unknown (This input is nonsensical and cannot be mapped to a known city or state.)

Model

gemin-1.5-pro-002 

Region \*

europa-west3 (Frankfurt) 

Temperature 

0



2

0.2

Output token limit 

1



8192

8192

Grounding 


Source:  Google Search

Customise

Add stop sequence 

Press Enter after each sequence

Output format

Plain text 

Edit

Safety filter settings

Advanced

Seed 

Max. responses 

1



1

N/A

Top-k 

1



40

N/A

Top-p 

0



1

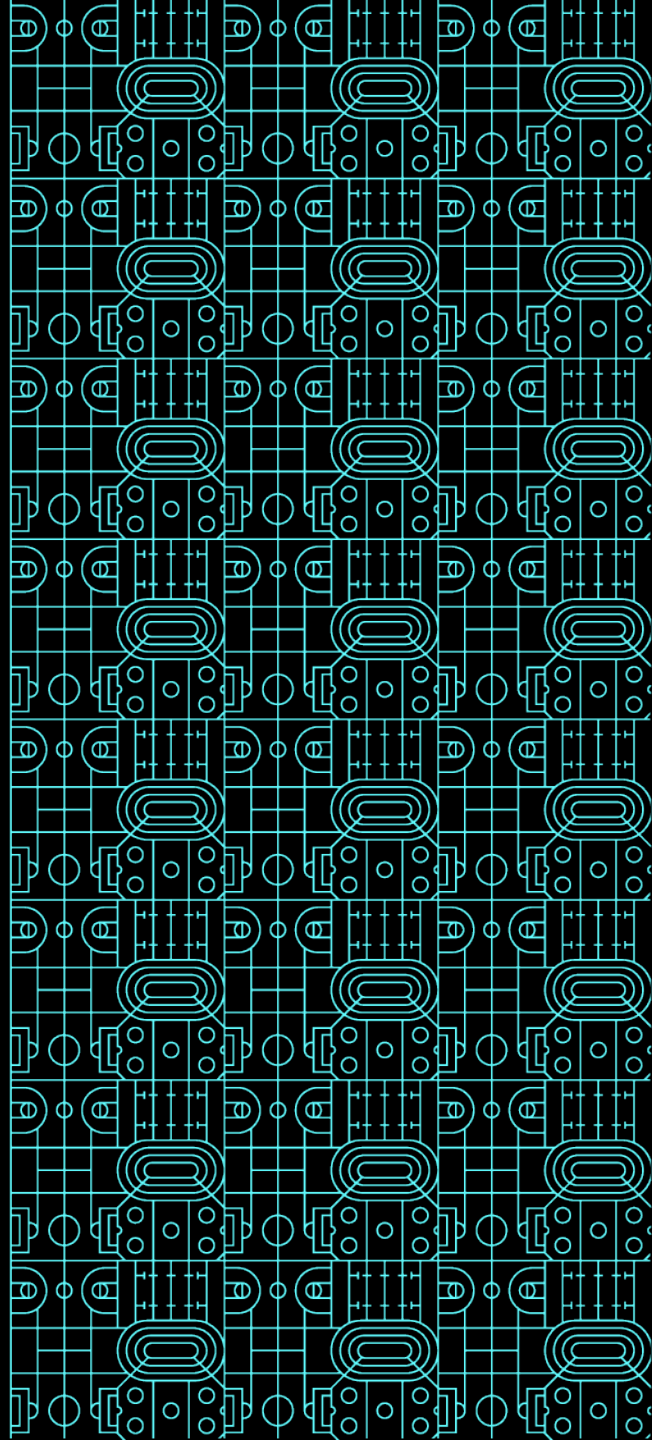
0.1

Stream model responses

Print responses as they're generated

- Up-to-date data
- Reduce hallucination

# USE CASE: FINAL PROMPT



# HYBRID PROMPT APPROACH WITH LESS CREATIVITY AND MORE PREDICTABILITY

CONTEXTUAL PROMPTING  
FEW SHOT LEARNING

MODEL: GEMINI 1.5 PRO  
TEMPERATURE: 0.1 – 0.2  
TOP\_P: 0.1

# SYSTEM INSTRUCTIONS (CONTEXT)

**You are a text standardizer.** The dataset contains user-generated input, which may be messy, contain typos, abbreviations, or extra information. Your goal is to standardize the city names and identify the correct state.

From the following input that can contain the name of Brazilian municipality, federative state, city, town, community or village, standardize and print one standardized, consolidated name in Portuguese language, also include the full name of the Brazilian state to which the particular city belongs to separated from the input by a single comma

# SYSTEM INSTRUCTIONS (CONTEXT)

## **Format Output:**

Return the cleaned city name followed by a comma and the state name (e.g., "São Luís, Maranhão").

You cannot return empty/null output because you know how to pick up the right data from the text in this field. In output include only the name of the city and region, exclude any explanation, gaps, indentation, and any other symbols

## **Unknown Input:**

If you cannot return any value, try to search for the value on the internet to get the right answer. If you cannot return any value for the name of the Brazilian state it there "Unknown State". Also if you cannot standardize the text input assign there "Unknown City". If the input does not include names of municipalities, cities and towns in Brazil assign them "Unknown City".

# FEW SHOT LEARNING - EXAMPLES

1) **Analyze the input:** Carefully examine the user-provided city name

2) **Identify the state:**

\* If the input includes the name of a Brazilian state (e.g., "Maranhão"), use it to determine the state. The city name should be the capital of that state (e.g., "São Luís, Maranhão").

\* If the input includes only a district, neighborhood, or administrative region (e.g., "Irajário de janeiro"), use it as a hint to identify the city and state (e.g., "Rio de Janeiro, Rio de Janeiro").

\* If the input is the name of a city that is not a state capital, identify the state to which the city belongs. (e.g. "Campinas" = "Campinas, São Paulo").

\* If the input contains a state abbreviation, expand it to the full state name.

# FEW SHOT LEARNING - EXAMPLES

## 3) Variations

Be aware that the selected column might already include the name of the Brazilian state, if it does, use it as a hint for you to detect which state it is and as the name for a city use the capital city of that state

(example 1: Maranhão = São Luís, Maranhão;

example 2: Segipe = Aracaju, Sergipe).

Also be aware the string might also include only the name of the city district, neighborhood, or administrative region, in that case also use it as a hint to detect the right city/town

(example: Irajário de Janeiro = Rio de Janeiro, Rio de Janeiro).

# FEW SHOT LEARNING - EXAMPLES

## 4) Other cases:

In some cases, the input contains the iso name of the country at the last two positions of the string

(example: Campo alegre al = Campo Alegre, Alagoas), use it to determine the right brazil country.

Some Brazilian names are similar to portuguese city names so before you note that the input is not Brazilian city please make sure it is not name of some brazilian municipality, community or village (example: Coimbra = Coimbra, Minas Gerais).

If you are really sure that the input name is not located in Brazil assign there "Outside of Brazil".

Do not miss names of municipalities and always try to correctly classify them

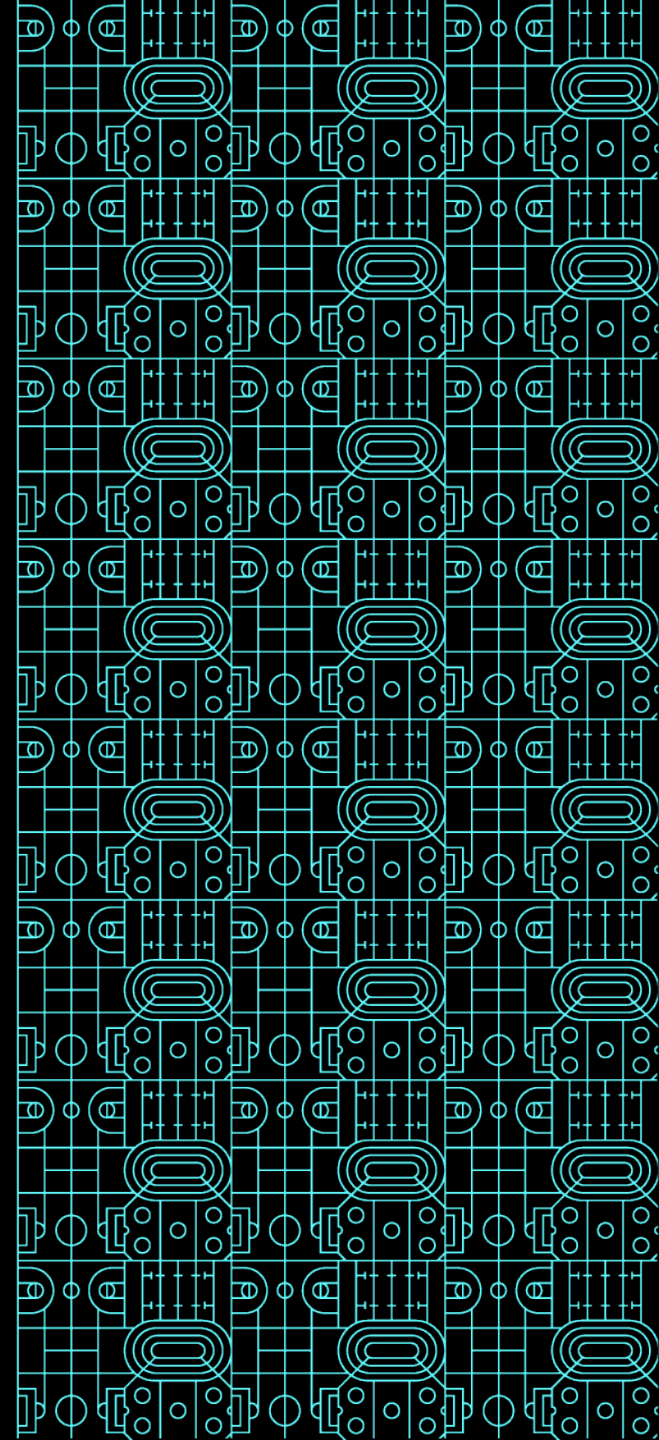
(example 1: Salete = Salete, Santa Catarina;

example 2: Pedro laurentino = Pedro Laurentino, Piauí;

example 3: Ribeira = Ribeira, São Paulo;

example 4: Pendencias = Pendências, Rio Grande do Norte).

# USE CASE: „PROMPT“ LOW-CODE DEPLOYMENT



# SQL BASED SOLUTION - SEAMLESS

~~MANUAL APPROACH~~ -> BIGQUERY AUTOMATION

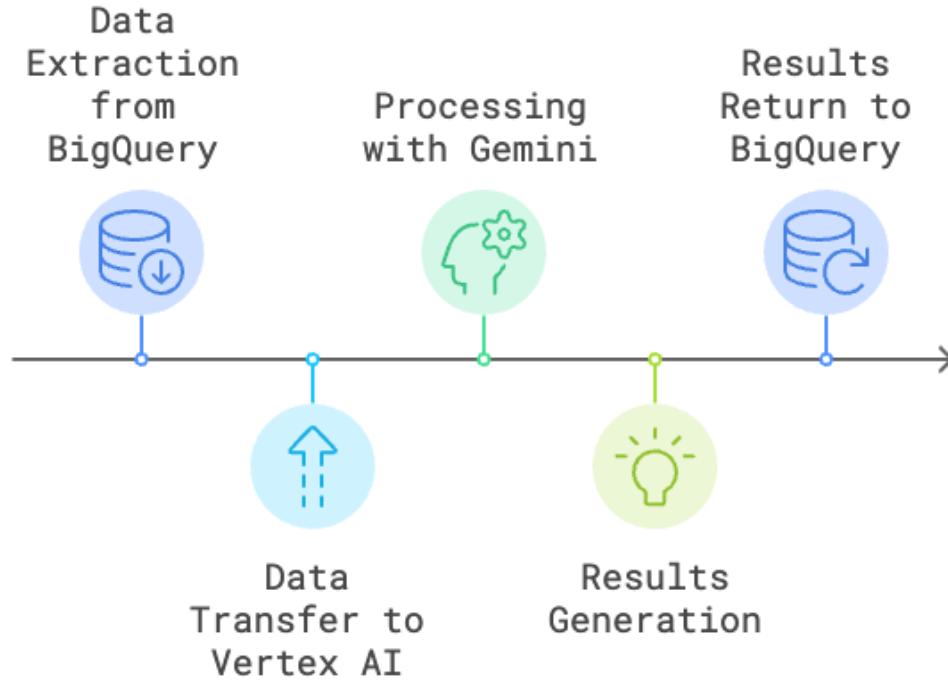
USE GEMINI (DIRECTLY FROM BIGQUERY)

ML.GENERATE\_TEXT() -> SQL

ORCHESTRATION BY DATAFORM

SQL + JS !!!

### Data Flow from BigQuery to Vertex AI



```

    ,database:"livesport"
    ,schema: "livesport_rs_data"
    ,tags: ["brazil", "players_monthly_brazil"]
    ,name: "brazil_cities_cleaned"
    ,description: "Cleaning brazilian cities and region data."
}

```

```
WITH cities_for_cleaning
```

```
AS
```

```
(SELECT
```

```
DISTINCT cleaned_city as uncleaned_city
FROM ${ref("livesport_rs_data", "players_monthly_players_cleaned")} l1
```

```
WHERE NOT EXISTS (SELECT uncleaned_city FROM ${self()} c
```

```
WHERE l1.cleaned_city = c.uncleaned_city)
```

```
),
```

```
)
```

```
gemiini_result
```

```
AS
```

```
(SELECT
```

```
ml_generate_text_llm_result as gemini_Response,
```

```
uncleaned_city
```

```
--prompt as prompt,
```

```
FROM ML.GENERATE_TEXT(MODEL 'al.l0_llm_test.model_cities_cleaning_brazil',
```

```
(select 'You are a text standardizer. From the following input that can contain the name of Brazilian municipality, federative state, city, town, commu
```

```
substring(uncleaned_city, 0, 50) as prompt,
```

```
uncleaned_city
```

```
from cities_for_cleaning),
```

```
STRUCT(
```

```
0.2 AS temperature,
```

```
TRUE AS flatten_json_output,
```

```
0.1 AS top_p,
```

```
TRUE AS ground_with_google_search
```

```
))
```

```
)
```

```
SELECT
```

```
uncleaned_city,
```

```
IF(LENGTH(gemini_response) > 60, 'Unknown City', SPLIT(gemini_response, ',')[OFFSET(0)]) AS cleaned_city,
```

```
IF(LENGTH(gemini_response) > 60, 'Unknown State', TRIM(SPLIT(gemini_response, ',')[OFFSET(1)])) AS state_full_name,
```

```
CASE
```

```
WHEN TRIM(SPLIT(gemini_response, ',')[OFFSET(1)]) = 'Acre' THEN 'AC'
```

```
WHEN TRIM(SPLIT(gemini_response, ',')[OFFSET(1)]) = 'Alagoas' THEN 'AL'
```

```
WHEN TRIM(SPLIT(gemini_response, ',')[OFFSET(1)]) = 'Amapá' THEN 'AP'
```

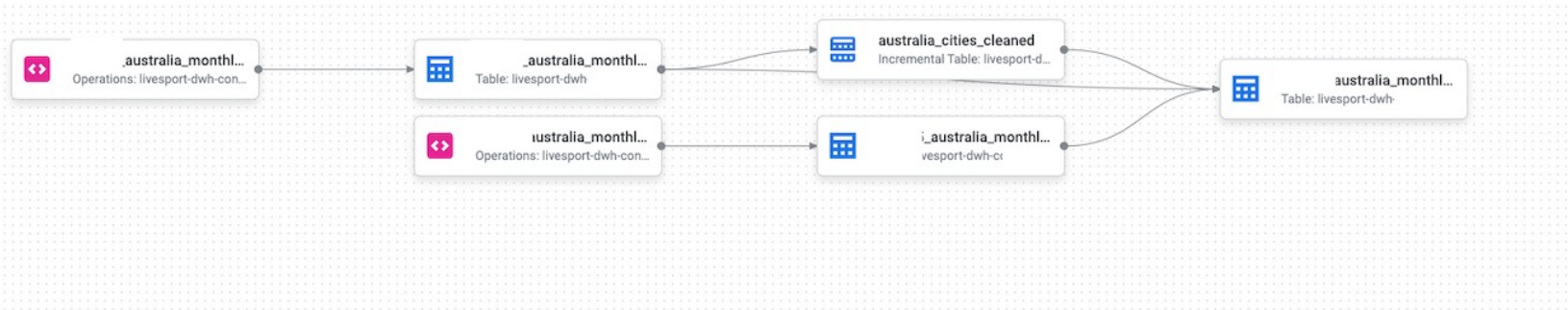
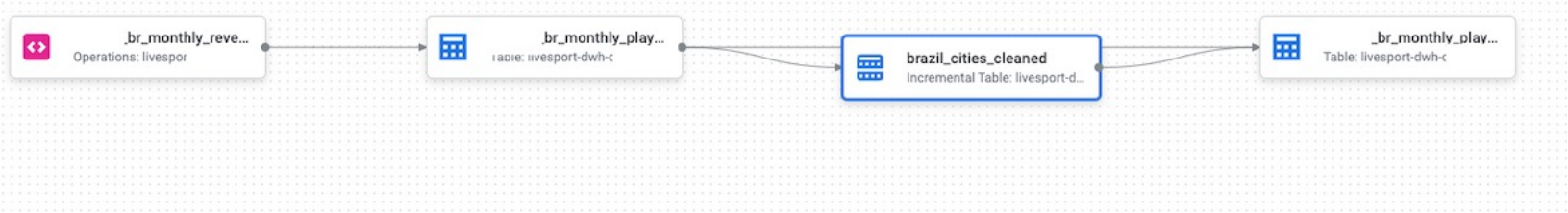
```
WHEN TRIM(SPLIT(gemini_response, ',')[OFFSET(1)]) = 'Amazonas' THEN 'AM'
```

```
WHEN TRIM(SPLIT(gemini_response, ',')[OFFSET(1)]) = 'Bahia' THEN 'BA'
```

```
WHEN TRIM(SPLIT(gemini_response, ',')[OFFSET(1)]) = 'Ceará' THEN 'CE'
```

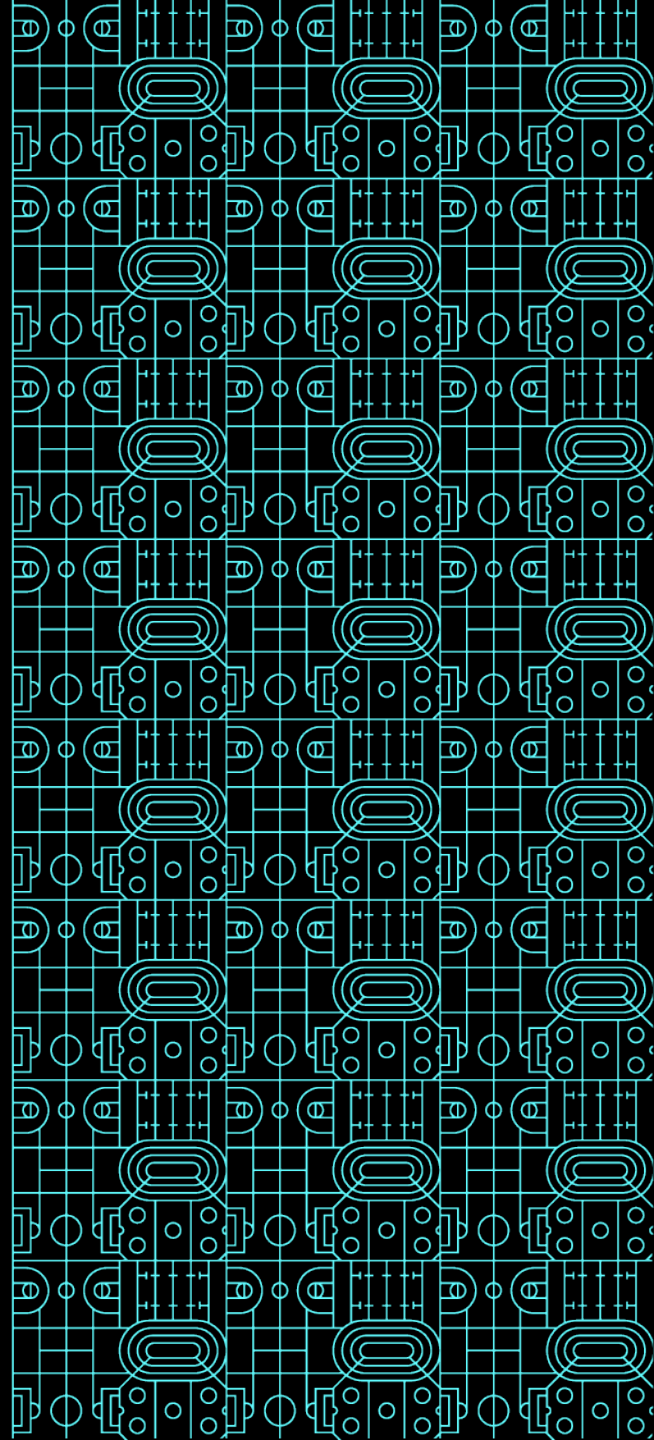
```
WHEN TRIM(SPLIT(gemini_response, ',')[OFFSET(1)]) = 'Distrito Federal' THEN 'DF'
```

# ORCHESTRATION & SCALABILITY BY DATAFORM + BIGQUERY



# USE CASE: RESULTS

LIVESPORT



# -> CLEANED DATA OUT

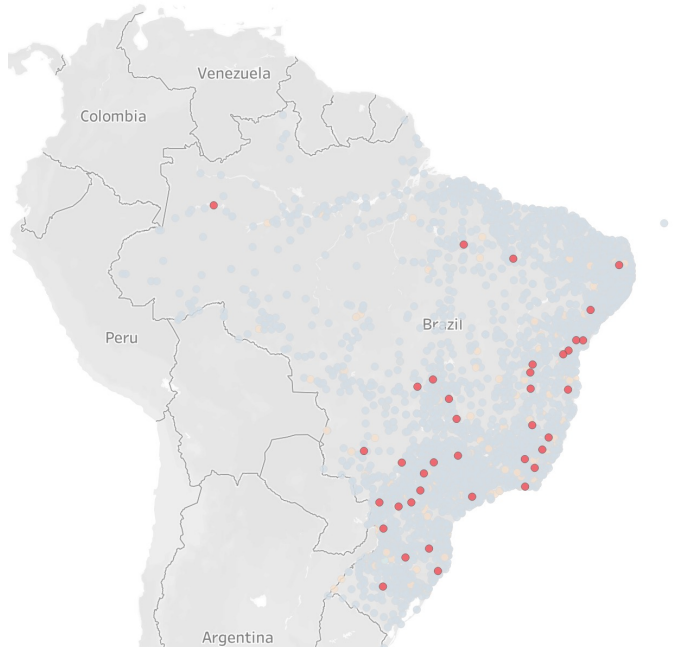
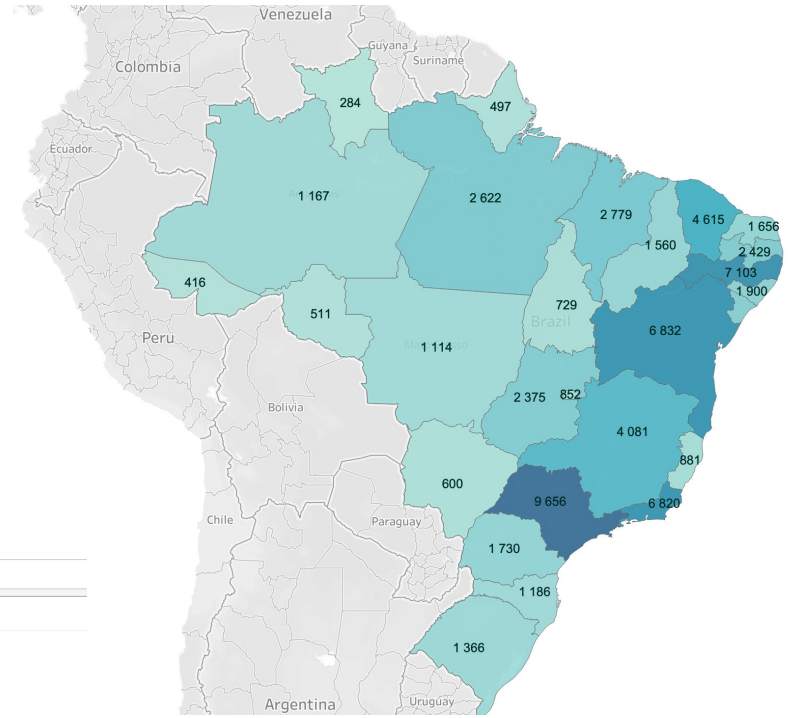
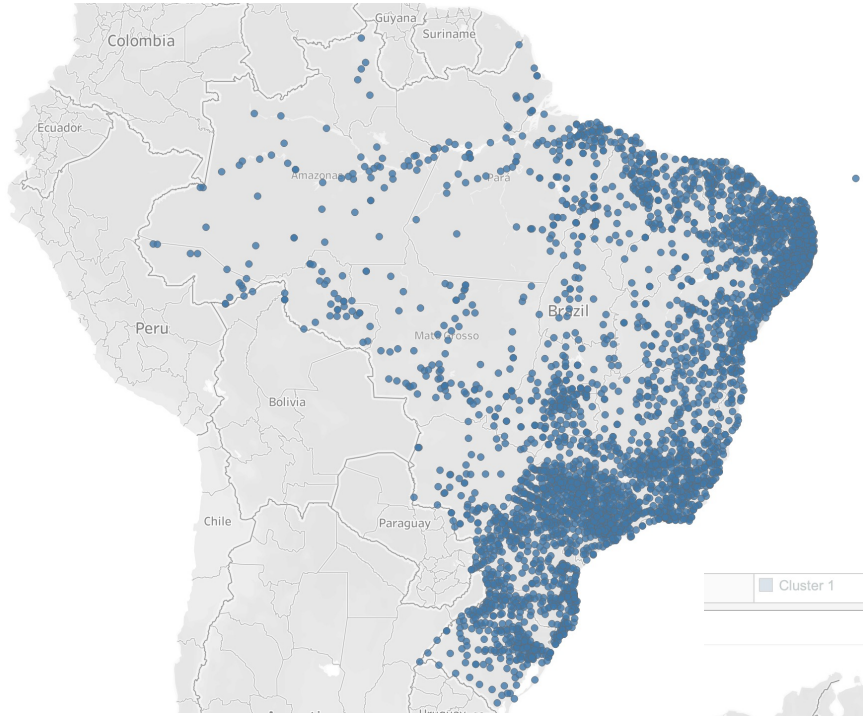
## BRAZIL

Cleaned city    Region / state

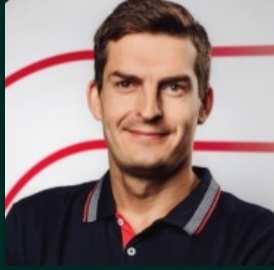
Row	uncleaned_city	cleaned_city	state_full_name	updated_at
1	Marechal thaumaturgo	Marechal Thaumaturgo	Acre	2024-09-06 07:33:40.666161 U...
2	Olho d agua das flores	Olho d'Água das Flores	Alagoas	2024-09-06 07:33:40.666161 U...
3	Maceió-al	Maceió	Alagoas	2024-09-06 07:33:40.666161 U...
4	Campo alegre al	Campo Alegre	Alagoas	2024-09-06 07:33:40.666161 U...
5	Uniao do palmares	União dos Palmares	Alagoas	2024-09-06 07:33:40.666161 U...
6	Santo antônio do içá	Santo Antônio do Içá	Amazonas	2024-09-06 07:33:40.666161 U...
7	Atalaia do norte	Atalaia do Norte	Amazonas	2024-09-06 07:33:40.666161 U...
8	Pilão arcado ba	Pilão Arcado	Bahia	2024-09-06 07:33:40.666161 U...
9	Teofilandia	Teofilândia	Bahia	2024-09-06 07:33:40.666161 U...
10	Pojuca ba	Pojuca	Bahia	2024-09-06 07:33:40.666161 U...
11	Jacobina -ba	Jacobina	Bahia	2024-09-06 07:33:40.666161 U...
12	Saúde ba	Saúde	Bahia	2024-09-06 07:33:40.666161 U...

## AUSTRALIA

	uncleaned_city	cleaned_city	region	updated_at
1	West Hoxton	Sydney	New South Wales	2024-11-04 15:40:59.926007 U...
2	Morley	Perth	Western Australia	2024-11-04 15:40:59.926007 U...
3	Birtinya	Sunshine Coast	Queensland	2024-11-04 15:40:59.926007 U...
4	Dandenong South	Melbourne	Victoria	2024-11-04 15:40:59.926007 U...
5	Dianella	Perth	Western Australia	2024-11-04 15:40:59.926007 U...
6	Melbourne	Melbourne	Victoria	2024-11-04 15:40:59.926007 U...
7	Mackenzie	Mackenzie	Unknown Region	2024-11-04 15:40:59.926007 U...
8	Coalfalls	Ipswich	Queensland	2024-11-04 15:40:59.926007 U...
9	Annandale	Sydney	New South Wales	2024-11-04 15:40:59.926007 U...
10	Haymarket	Sydney	New South Wales	2024-11-04 15:40:59.926007 U...



THAT'S ALL



DATADAY

LIVESPORT