

Airbyte.

2 года в продакшене

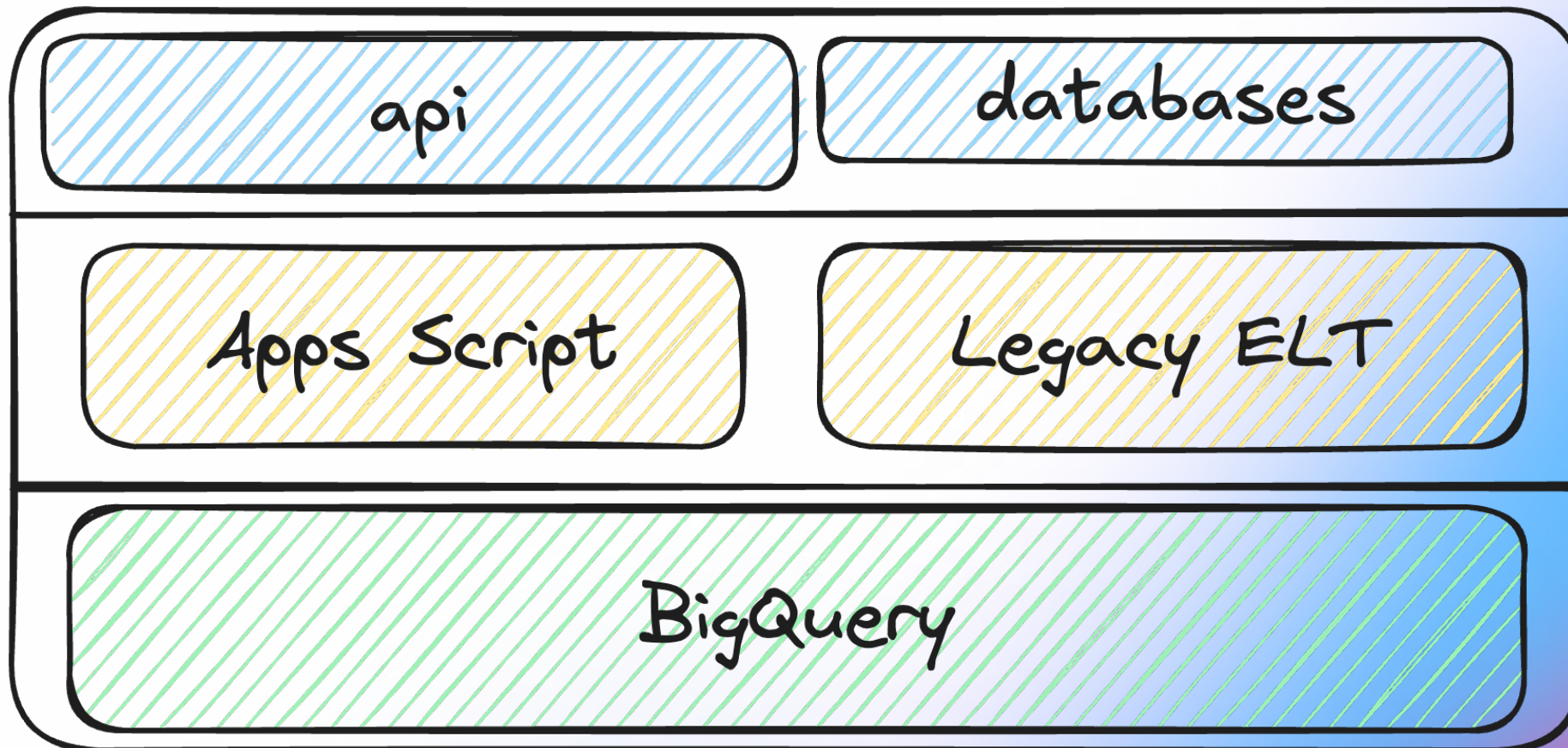
Спикер:

Попова Александра, Lead Data Engineer

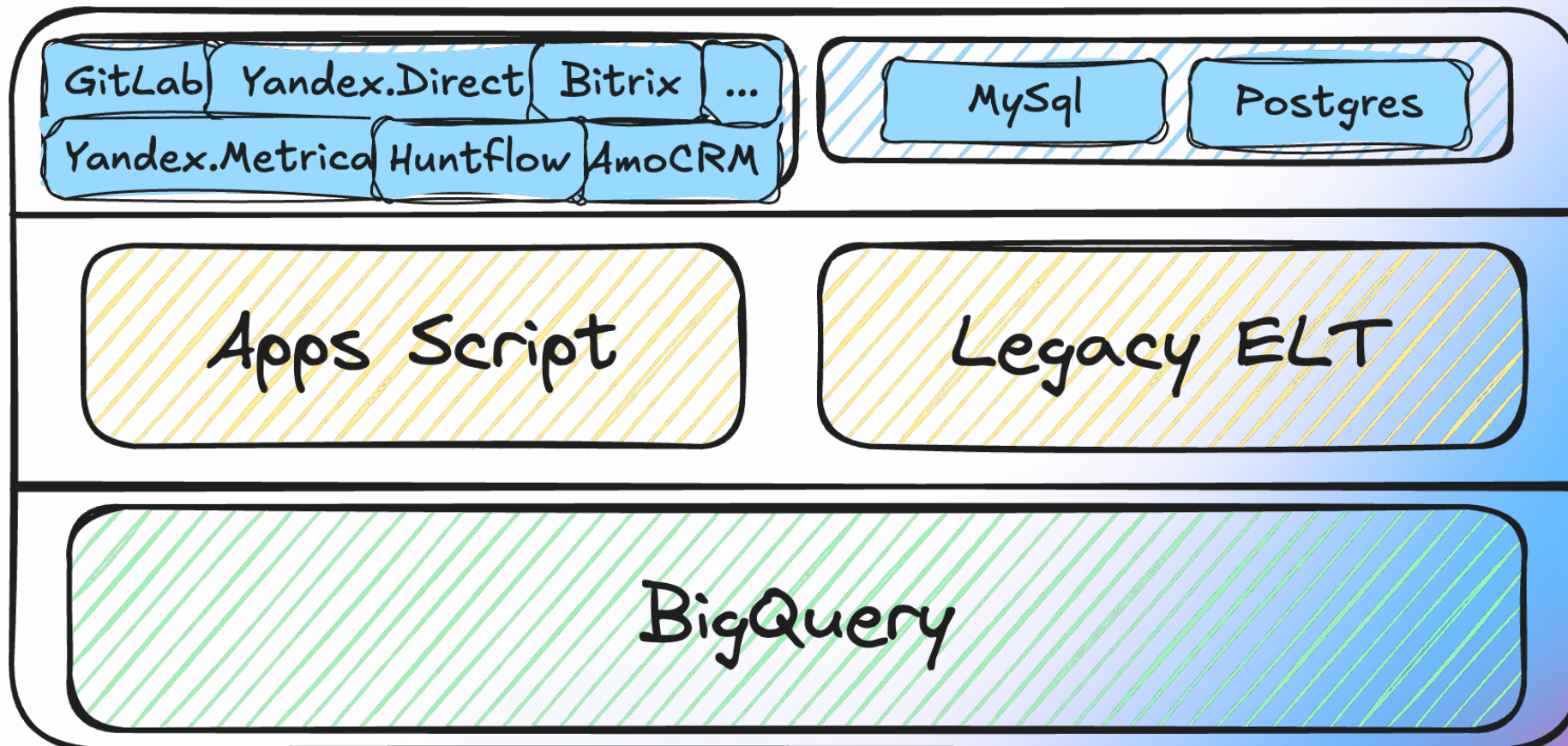
План



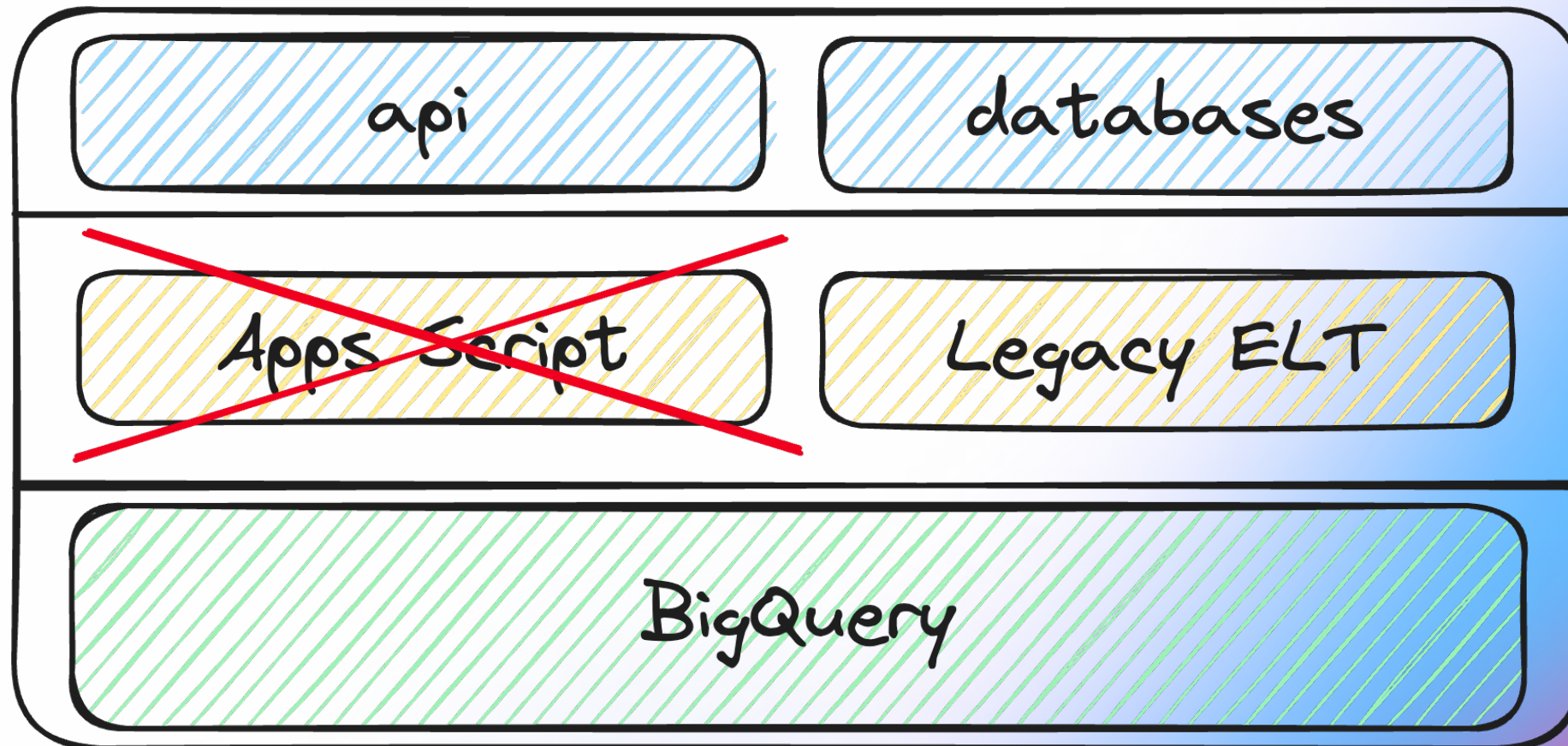
BACK TO end-2021



BACK TO end-2021



BACK TO end-2021



ВАСК ТО end-2021. Что хотели?

- Общий подход
- Шаблонизировать и стандартизировать
- Знать кто и что меняет
- Open source



Airbyte

15,000+

community members

10,000+

Github stars

1PB+

synced/month

800+

contributors

100k+

deployments

Airbyte. Что было?

Airbyte. Что было? Connector Development Kit

- Python CDK
- Java CDK

Airbyte. Что было? Connector Development Kit

- Python CDK:

```
class Source(BaseConnector):
    def spec() -> ConnectorSpecification:
        ...

class AbstractSource(Source):
    def check() -> AirbyteConnectionStatus:
        ...
    def streams() -> List[Stream]:
        ...
    def discover() -> AirbyteCatalog:
        ...
    def read() -> Iterator[AirbyteMessage]:
        ...
```

Airbyte. Что было? Connector Development Kit

- Python CDK:

```
class Stream():
    def read() -> Iterable[StreamData]:
        ...
class HttpStream(Stream):
    def path() -> str:
        ...
    def request_params() -> MutableMapping[str, Any]:
        ...
    def request_headers() -> Mapping[str, Any]:
        ...
    def next_page_token() -> Optional[Mapping[str, Any]]:
        ...
    def get_backoff_strategy() ->
        Optional[Union[BackoffStrategy, List[BackoffStrategy]]]:
```

Airbyte. Что было? UI

Connections + New connection

NAME ▲	SOURCE NAME ▲	DESTINATION NAME ▲
✓ cabinet_legal_eam.message...	Postgres - cabinet_legal_eam	Greenplum - g
✓ treatment_plan.treatment_pl...	Postgres - treatment_plan	Greenplum - g
✓ treatment_plan.treatment_pl...	Postgres - treatment_plan	Greenplum - g

Airbyte. Что было? API

The screenshot shows the Airbyte API documentation for the 'Create a connection' endpoint. The left sidebar contains a navigation menu with categories like CONNECTIONS, JOBS, and STREAMS. The main content area is titled 'Create a connection' and includes the endpoint URL, a table for request history, a description of connections, required fields, and default connection settings. The right sidebar shows the 'LANGUAGE' dropdown (Shell, Node, Ruby, PHP, Python), 'AUTHORIZATION' (Bearer token), and a 'CURL REQUEST' section with a 'Try It!' button. Below the curl request is a 'RESPONSE' section with an example response.

Create a connection
POST https://api.airbyte.com/v1/connections

TIME	STATUS	USER AGENT
Make a request to see history.		
0 Requests This Month		

Connections

A connection is a configuration for syncing data between a source and a destination.

Required fields

Required fields include `sourceId` and `destinationId`. You can discover these by using the endpoints for [List Sources](#) and [List Destinations](#). Alternatively, these can also be found in the address bar of the Airbyte UI.

Default Connection Settings

If no other configuration settings set, then default configuration settings will be used:

- All streams will be set to the sync mode: `full_refresh_overwrite`
- Replication Frequency will be set to `Manual`
- Destination Namespace will be set to `Destination Default`
- Data Residency will be set to `AUTO` (most efficient data plane to sync data selected)
- Connection name defaults are formatted as: `source-name <-> destination-name`
- Non-Breaking Schema Updates Behavior: `ignore`

LANGUAGE
Shell Node Ruby PHP Python

AUTHORIZATION
Bearer token

CURL REQUEST

```
1 curl --request POST \  
2 --url https://api.airbyte.com/v1/connections \  
3 --header 'accept: application/json' \  
4 --header 'content-type: application/json' \  
5 --data '  
6 {  
7   "schedule": {  
8     "scheduleType": "manual"  
9   },  
10  "dataResidency": "auto",  
11  "namespaceDefinition": "destination",  
12  "namespaceFormat": null,  
13  "nonBreakingSchemaUpdatesBehavior": "ignore"  
14 }  
15 '
```

RESPONSE

Click **Try It!** to start a request and see the response here!
Or choose an example:
application/json

Airbyte. Что было? ГОТОВЫЕ КОННЕКТОРЫ

Hundreds of connectors out-of-the-box

We've got all your needs covered.
350+ connectors and counting. [Check our docs!](#)

Search

Build your own connector in only 10 min with our [no-code Connector Builder](#)

Types

- Sources
- Destinations

Availability

- Cloud
- Open-source

Categories

- All
- New & Noteworthy

Azure Blob Storage
Warehouses and Lakes

BigQuery
Warehouses and Lakes

DuckDB
Databases

Airbyte. Что есть сейчас?

Airbyte. Что есть сейчас? No-code Connector builder

The screenshot displays the Airbyte No-code Connector builder interface for the 'ya-tracker' connector. The interface is divided into three main sections:

- Left Panel:** Contains a sidebar with a 'UI' / 'YAML' toggle, a 'ya-tracker' connector icon, a 'Draft saved' status, and a list of configuration sections: 'Global Configuration', 'User Inputs (1)', 'STREAMS (1)', and 'projects'.
- Central Panel:** Titled 'Global Configuration', it shows the 'Connector Name' as 'ya-tracker'. Below this, the 'API Base URL' is set to 'https://api.tracker.yandex.net/v2/'. The 'Authentication' section is configured with the 'Method' set to 'Bearer' and a 'Bearer Token' field.
- Right Panel:** Features a 'Testing values' section with a red notification badge '1', a 'Projects' dropdown menu, and a prominent blue 'Test' button.

Airbyte. Что есть сейчас? Low code CDK

```
spec:
  type: Spec
  connection_specification:
    type: object
    $schema: http://json-schema.org/draft-07/schema#
    required:
      - api_key
    properties:
      api_key:
        type: string
        order: 0
        title: API Key
        airbyte_secret: true
      additionalProperties: true
  type: DeclarativeSource
  check:
    type: CheckStream
    stream_names:
      - projects
  streams:
    - name: projects
      type: DeclarativeStream
      retriever:
        type: SimpleRetriever
      paginator:
        type: NoPagination
      requester:
        path: projects
        type: HttpRequester
        url_base: https://api.tracker.yandex.net/v2/
        http_method: GET
        authenticator:
          type: BearerAuthenticator
          api_token: '{{ config['api_key'] }}'
        request_headers: {}
        request_body_json: {}
        request_parameters: {}
```

```
streams:
  - name: projects
    type: DeclarativeStream
    retriever:
      type: SimpleRetriever
    paginator:
      type: NoPagination
    requester:
      path: projects
      type: HttpRequester
      url_base: https://api.tracker.yandex.net/v2/
      http_method: GET
      authenticator:
        type: BearerAuthenticator
        api_token: '{{ config['api_key'] }}'
      request_headers: {}
      request_body_json: {}
      request_parameters: {}
    record_selector:
      type: RecordSelector
    extractor:
      type: DpathExtractor
      field_path: []
    primary_key: []
    schema_loader:
      type: InlineSchemaLoader
    schema:
      type: object
      $schema: http://json-schema.org/draft-07/schema#
      properties: {}
      additionalProperties: true
version: 0.51.41
metadata:
  autoImportSchema:
    projects: true
```

Airbyte. Что есть сейчас? PyAirbyte

- `pip install airbyte`

Airbyte. Что есть сейчас? PyAirbyte

Для чего?

- Для локального тестирования без сборки/пересборки контейнера

Airbyte. Что есть сейчас? PyAirbyte

```
import airbyte as ab

source = ab.get_source(
    "source-faker",
    config={"count": 5_000},
    install_if_missing=True,
)
source.check()
result = source.read()

for name, records in result.streams.items():
    print(f"Stream {name}: {len(list(records))} records")
```

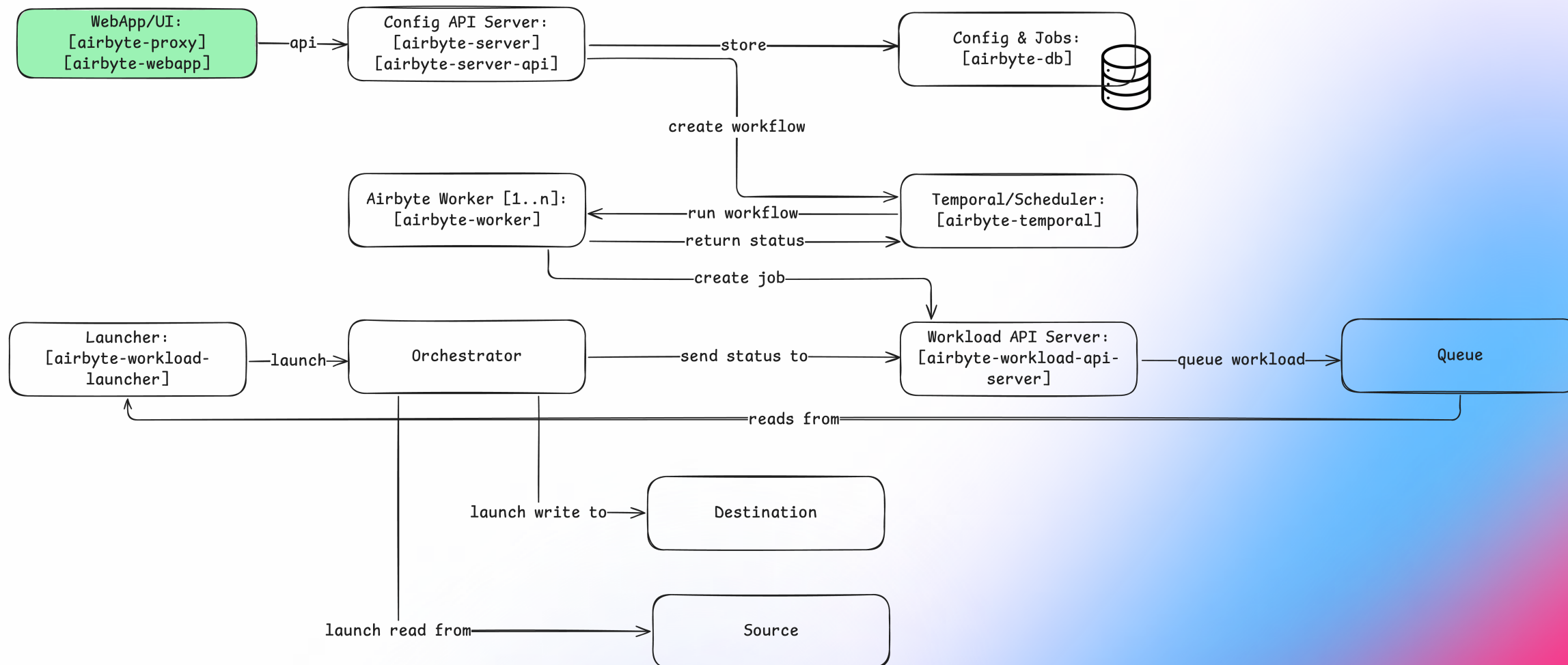
Airbyte. Что есть сейчас? Terraform Provider

```
resource "airbyte_source_stripe" "my_source_stripe" {
  configuration = {
    sourceType = "stripe"
    account_id = "acct_123"
    client_secret = "sklive_abc"
    start_date = "2023-07-01T00:00:00Z"
    lookback_window_days = 0
    slice_range = 365
  }
  name = "Stripe"
  workspace_id = var.workspace_id
}
```

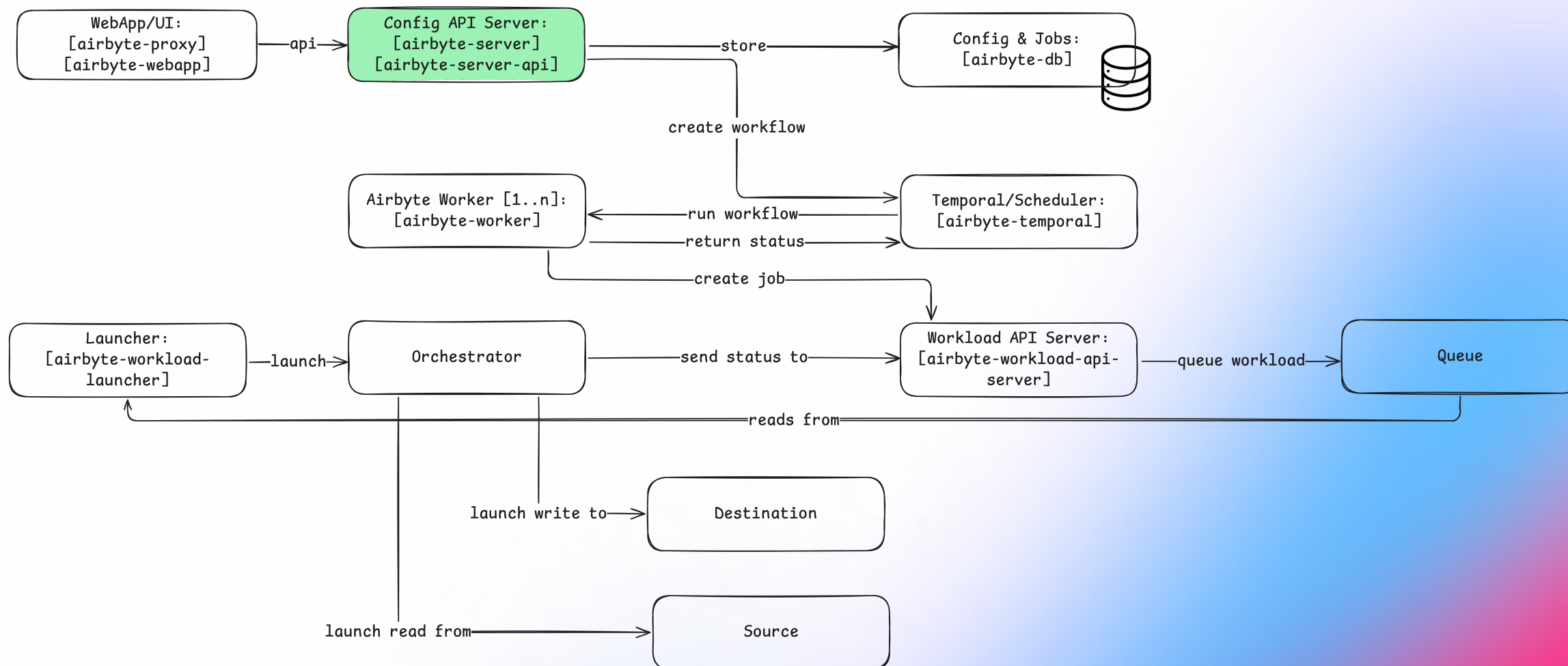
Airbyte. Что есть сейчас? Terraform Provider

```
terraform init  
terraform plan  
terraform apply
```

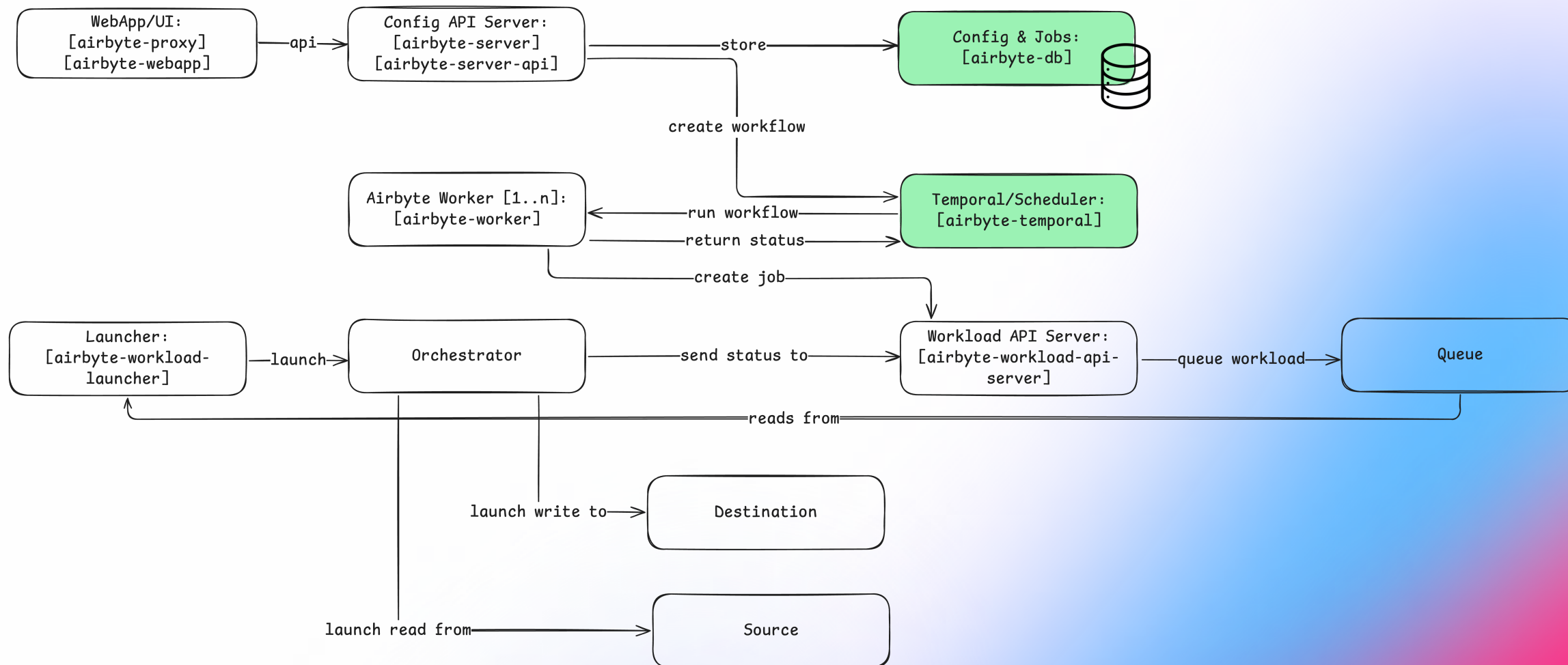
Airbyte. Архитектура сегодня



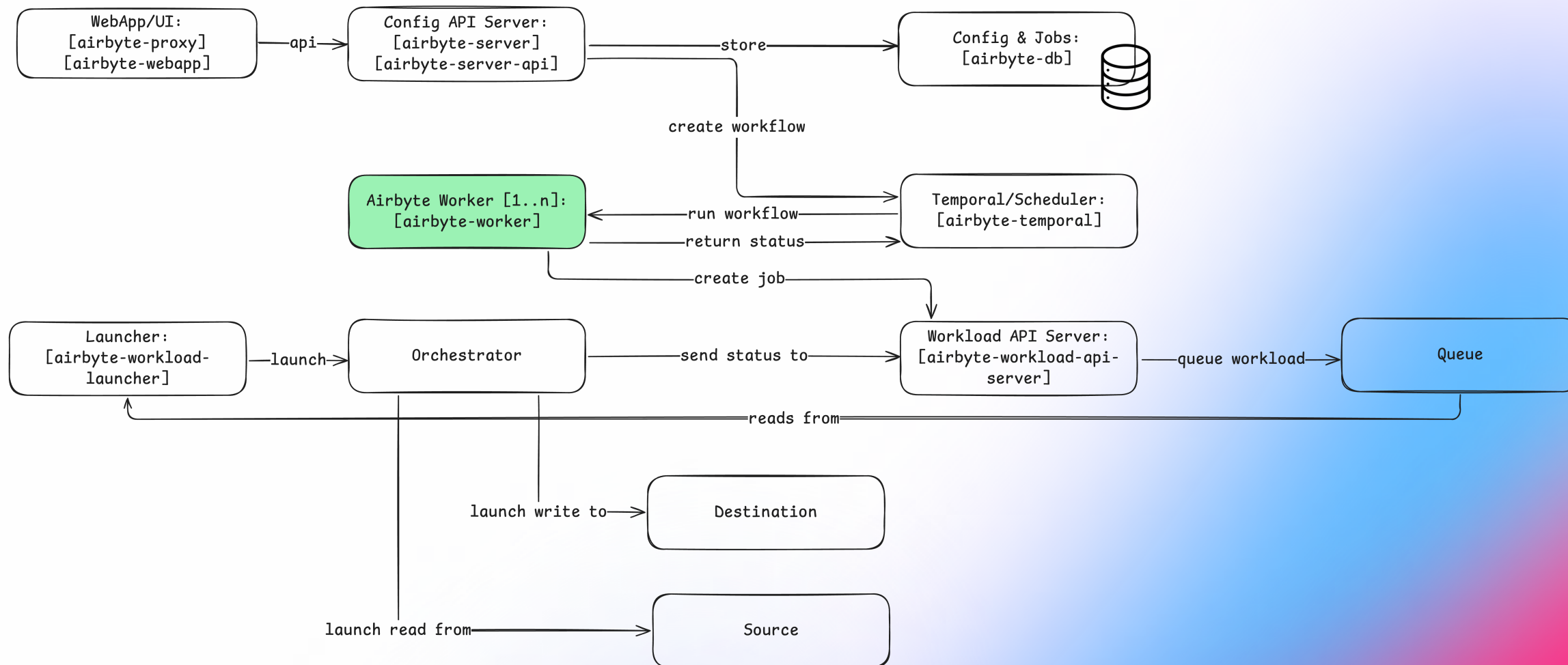
Airbyte. Архитектура сегодня



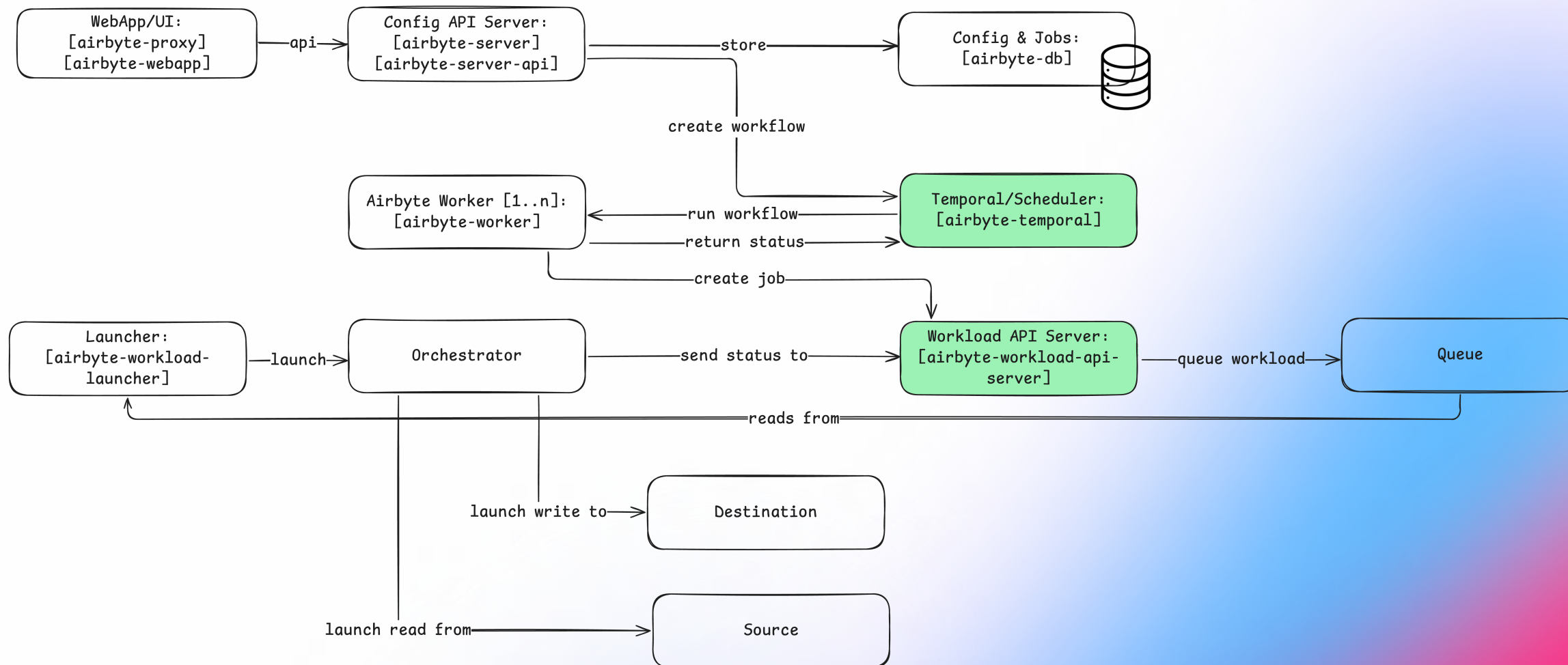
Airbyte. Архитектура сегодня



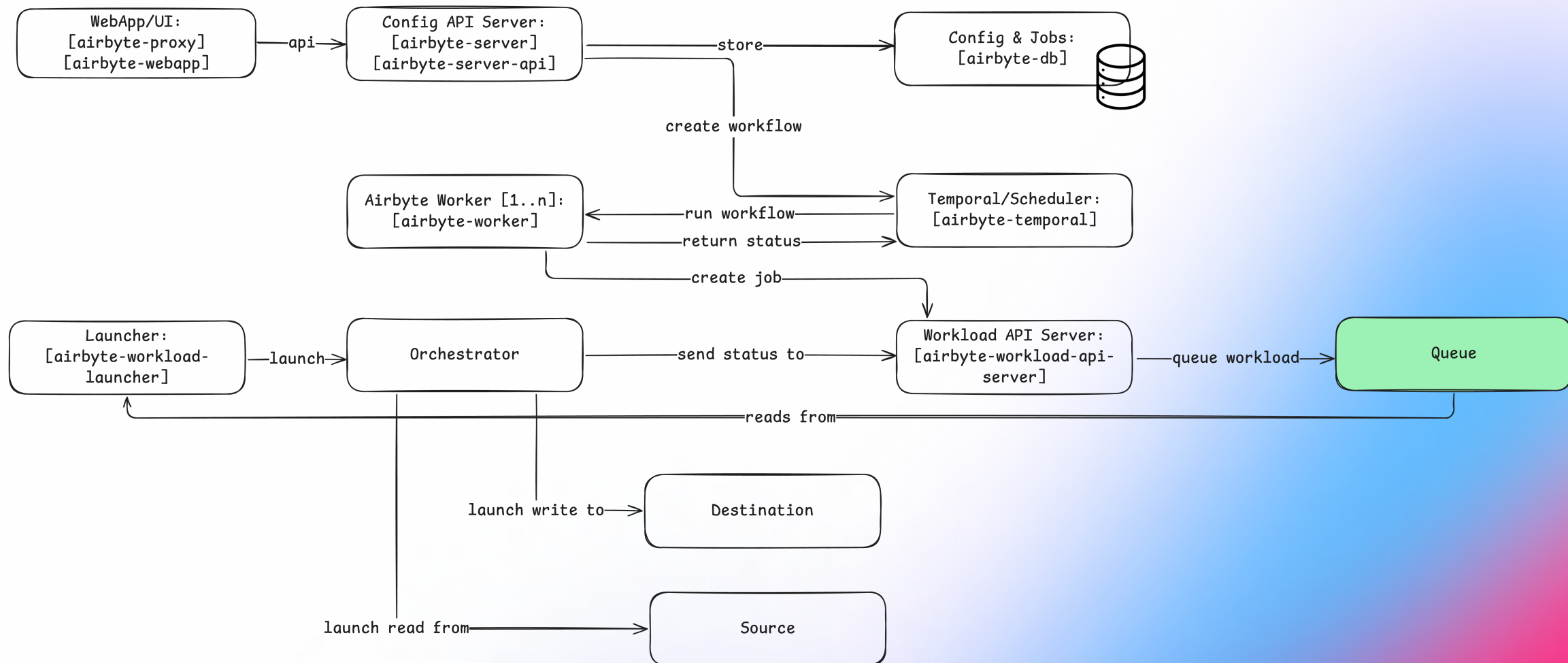
Airbyte. Архитектура сегодня



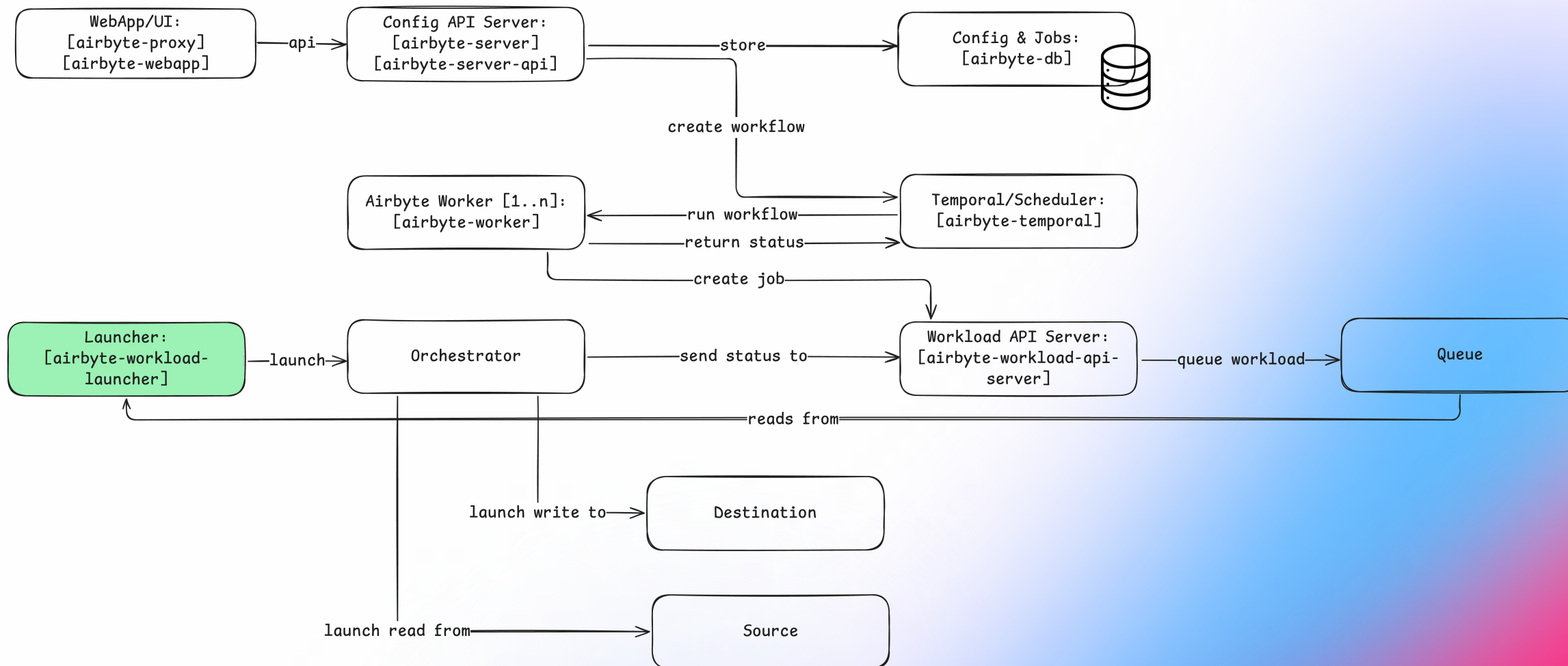
Airbyte. Архитектура сегодня



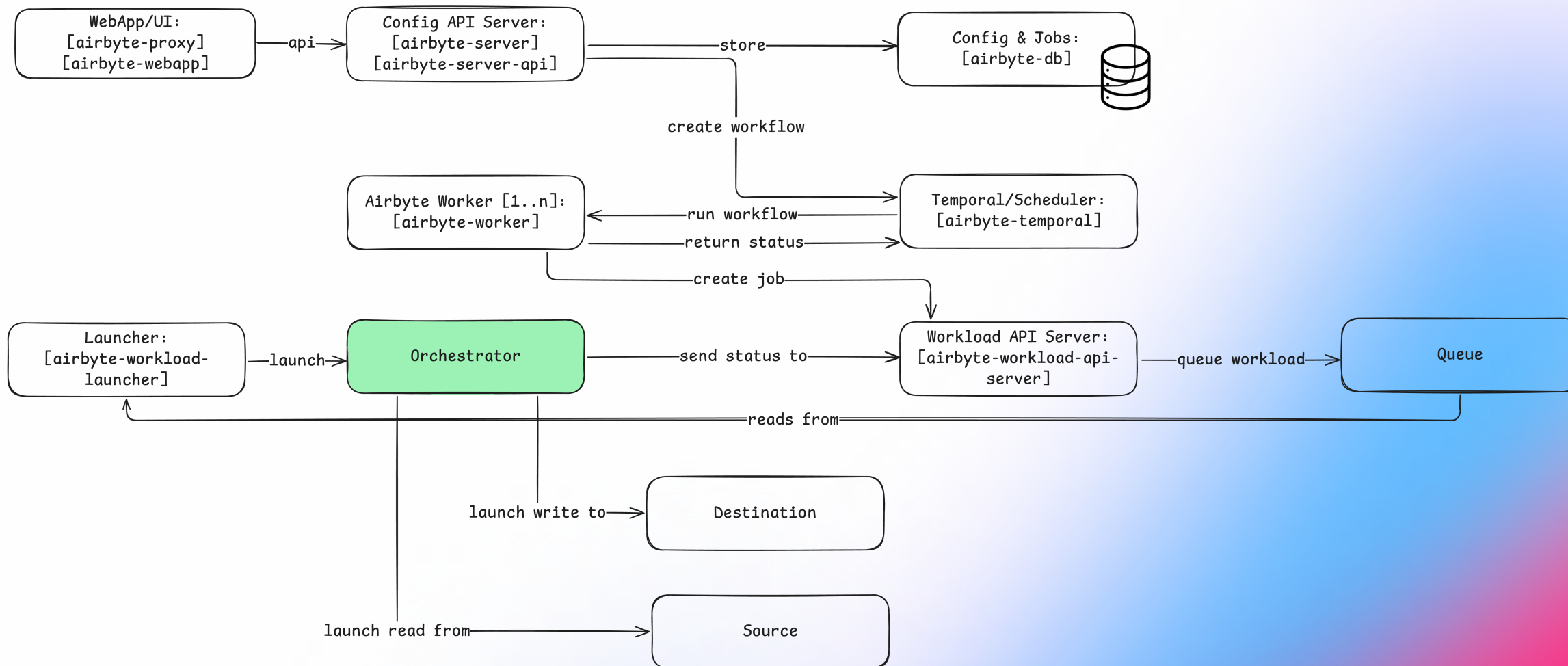
Airbyte. Архитектура сегодня



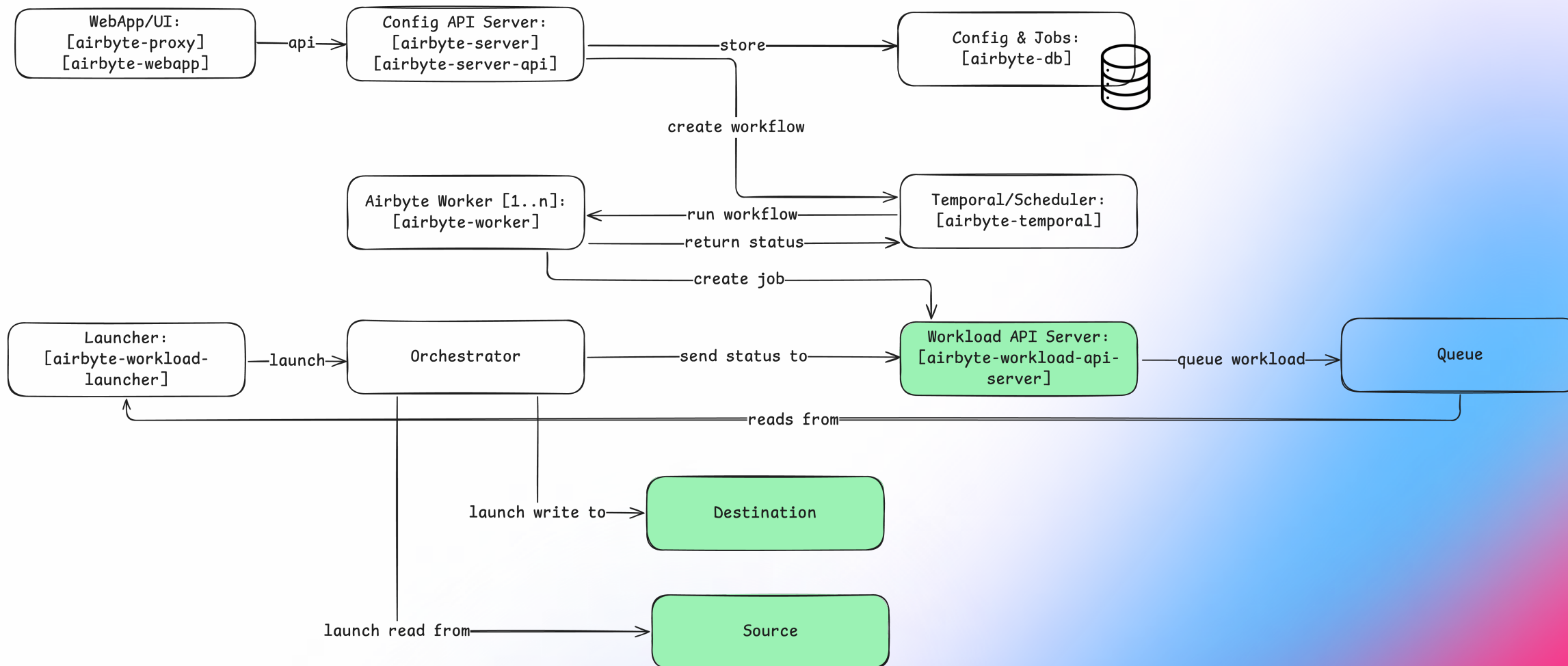
Airbyte. Архитектура сегодня



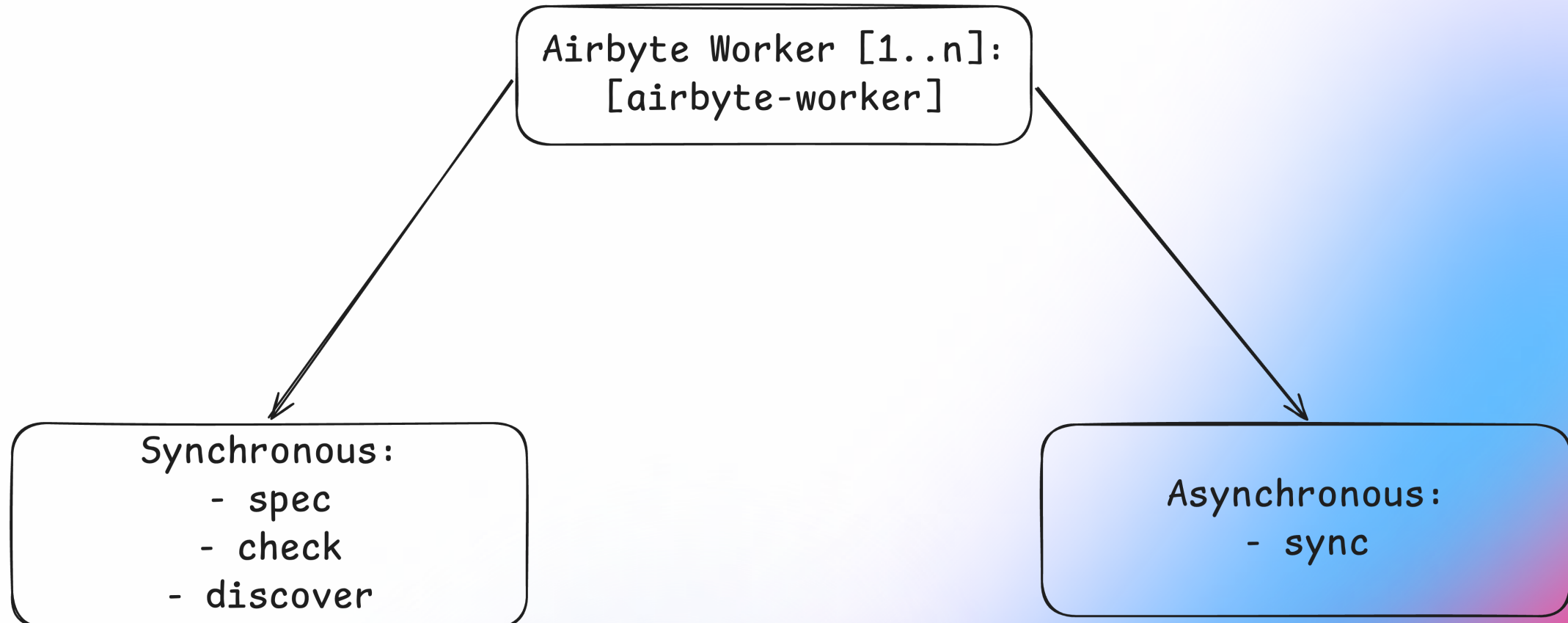
Airbyte. Архитектура сегодня



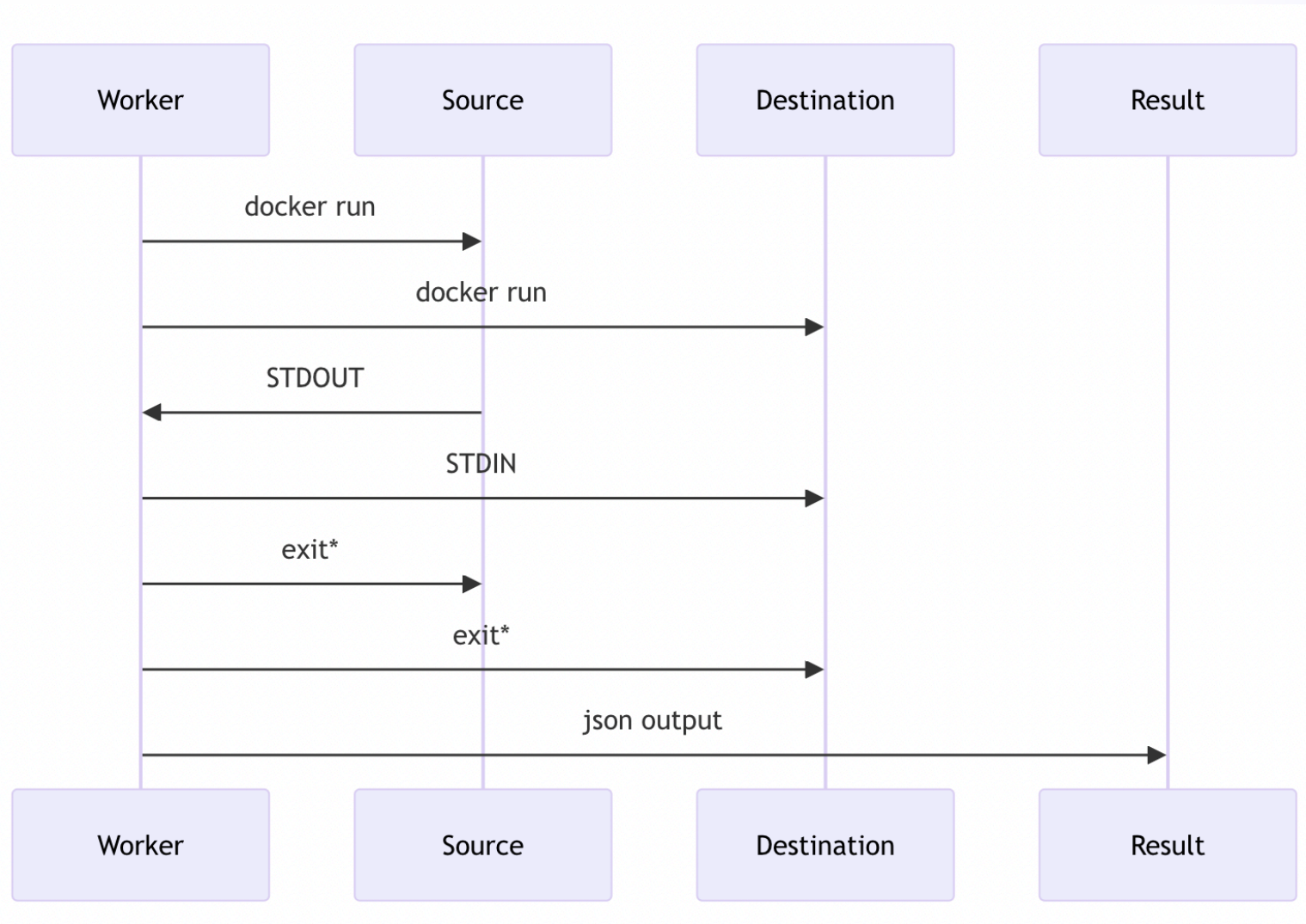
Airbyte. Архитектура сегодня



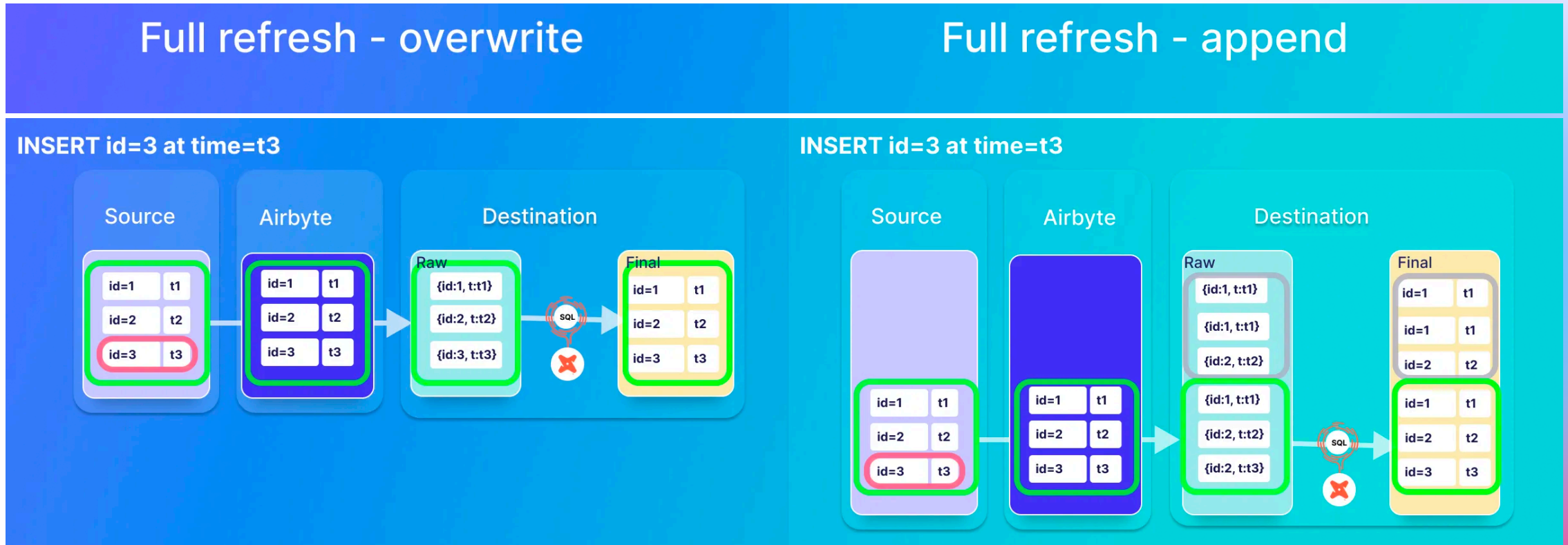
Airbyte. Workers



Airbyte. Workers



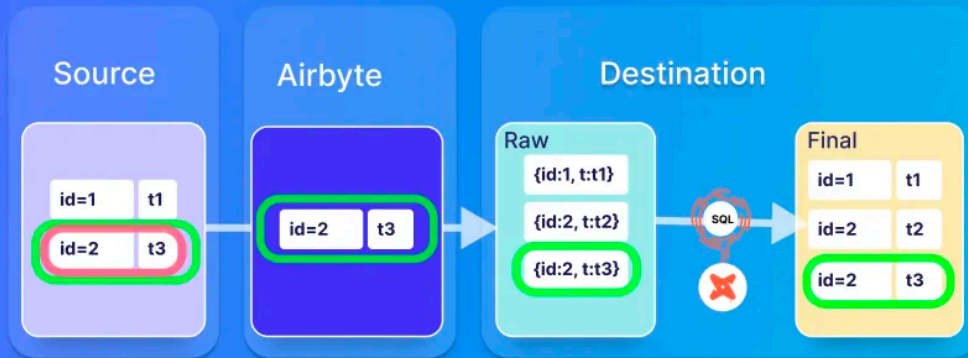
Airbyte. Sync Modes: Full Refresh



Airbyte. Sync Modes: Incremental

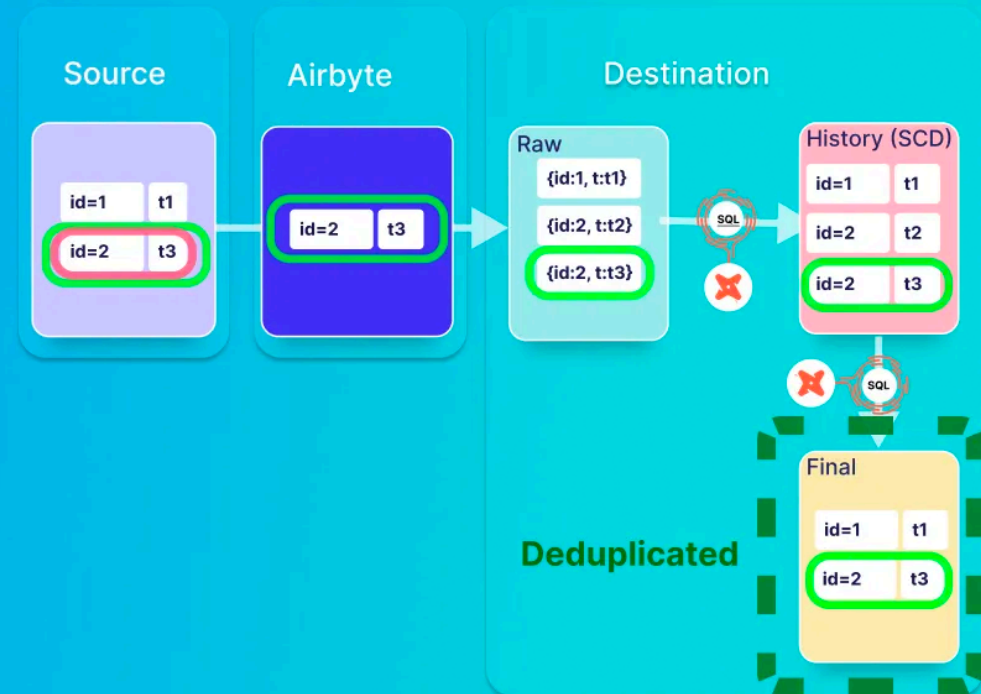
Incremental sync - append

UPDATE id=2 at time=t3

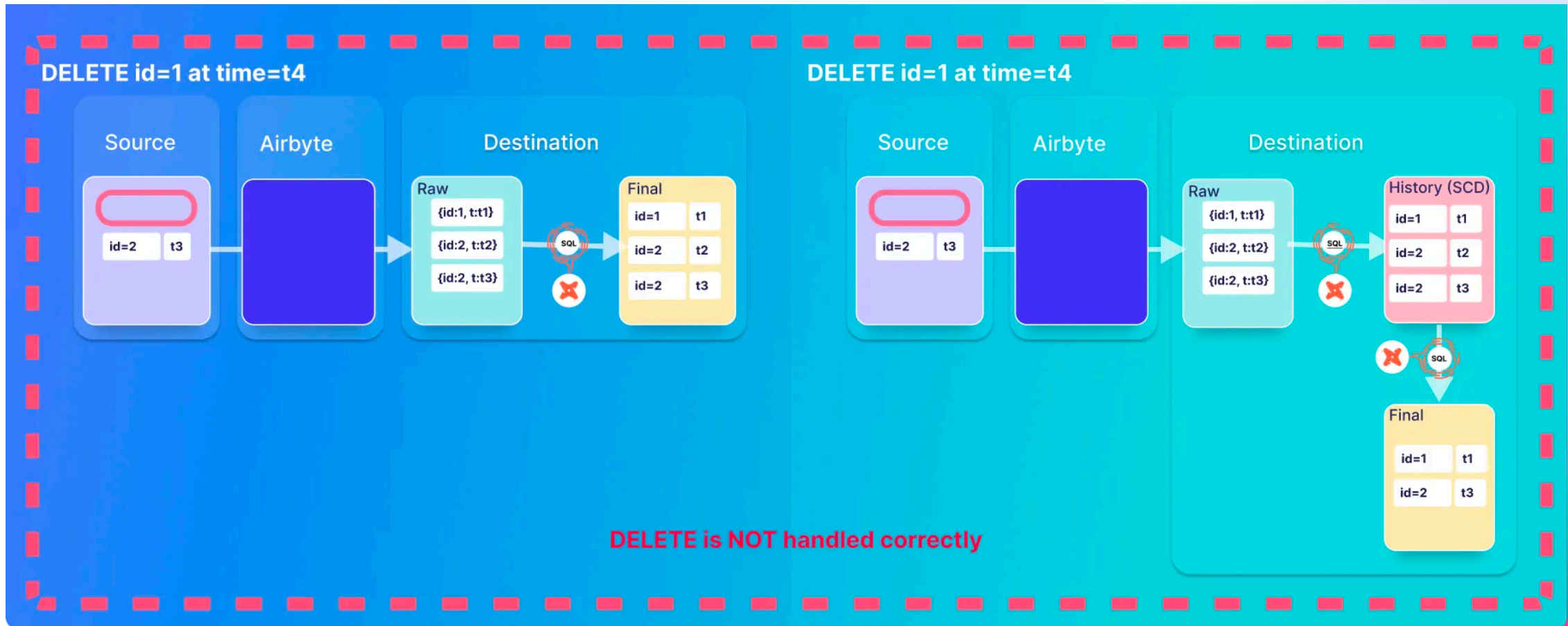


Incremental sync - deduped history

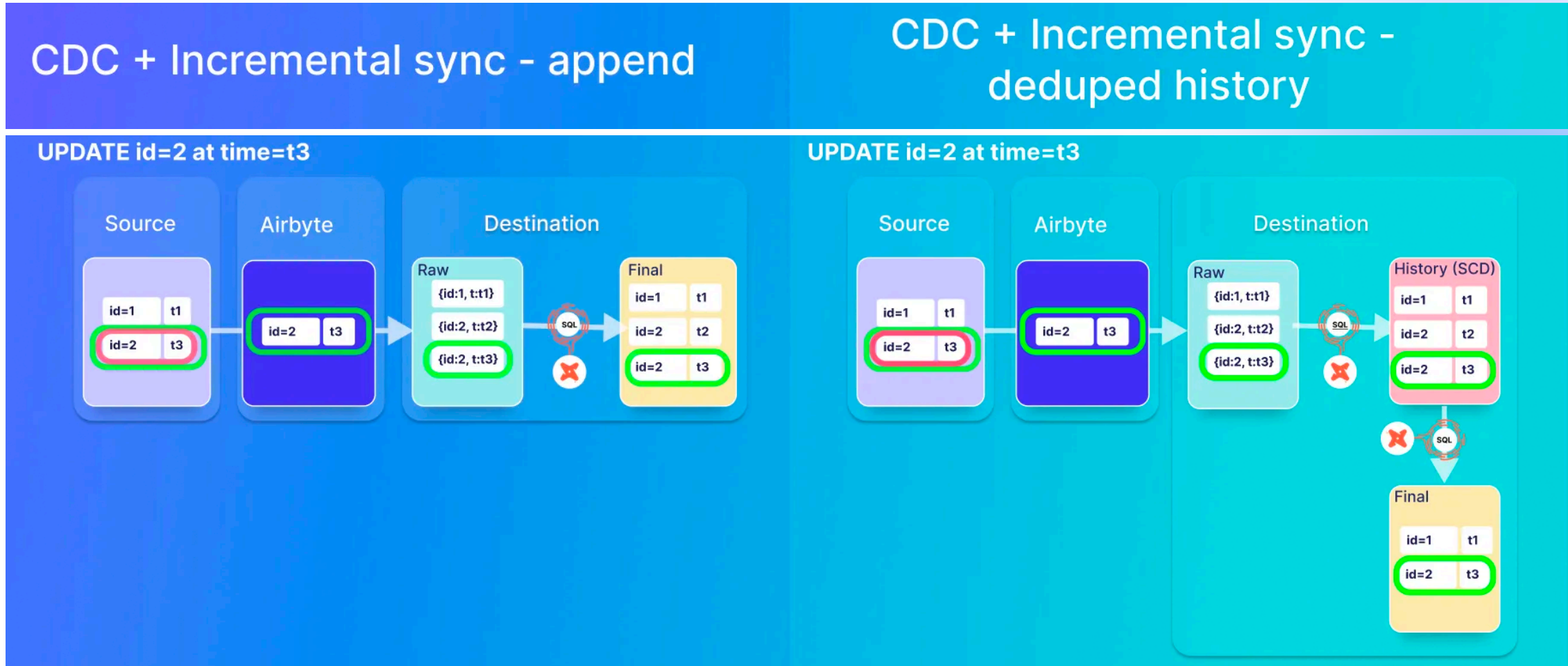
UPDATE id=2 at time=t3



Airbyte. Sync Modes: Incremental

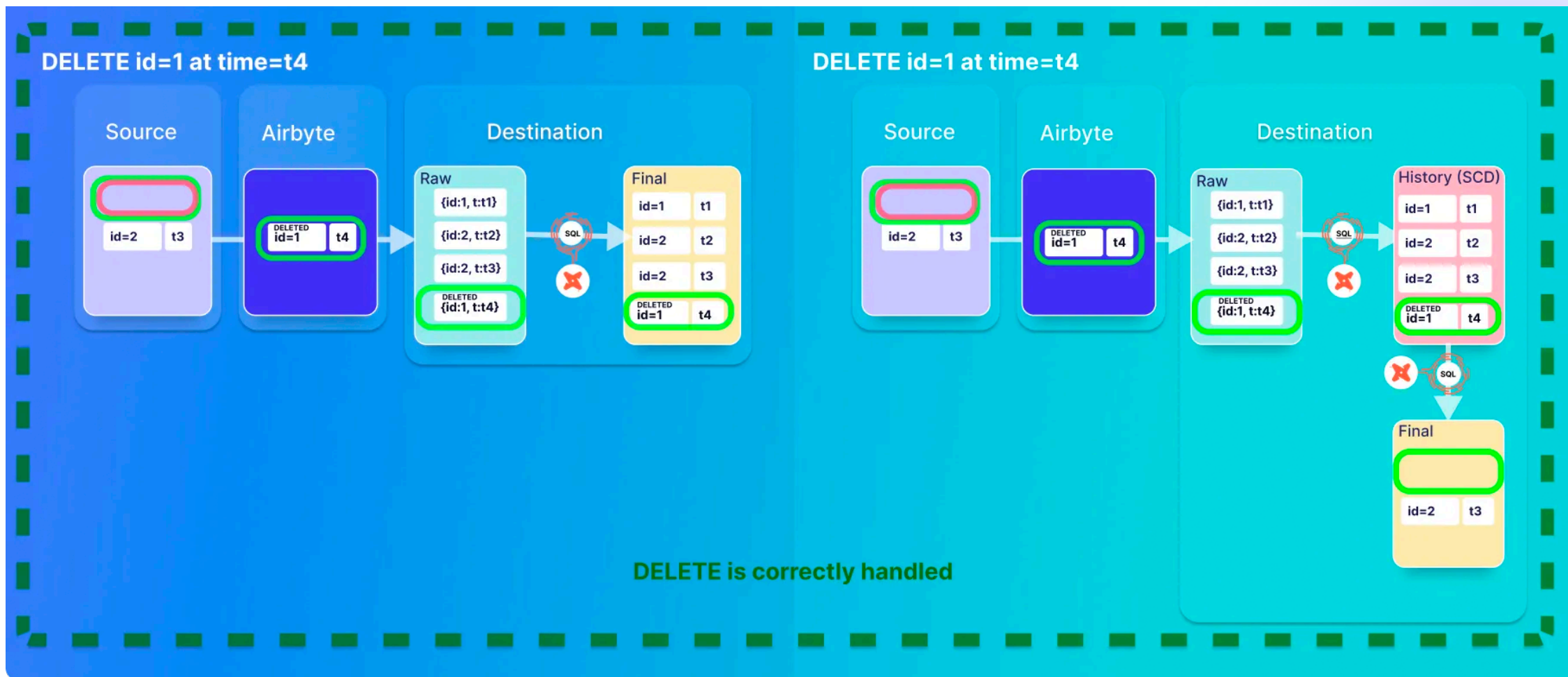


Airbyte. Sync Modes: CDC*



*Postgres, MySQL, MSSQL и MongoDB

Airbyte. Sync Modes: CDC*



*Postgres, MySQL, MSSQL и MongoDB

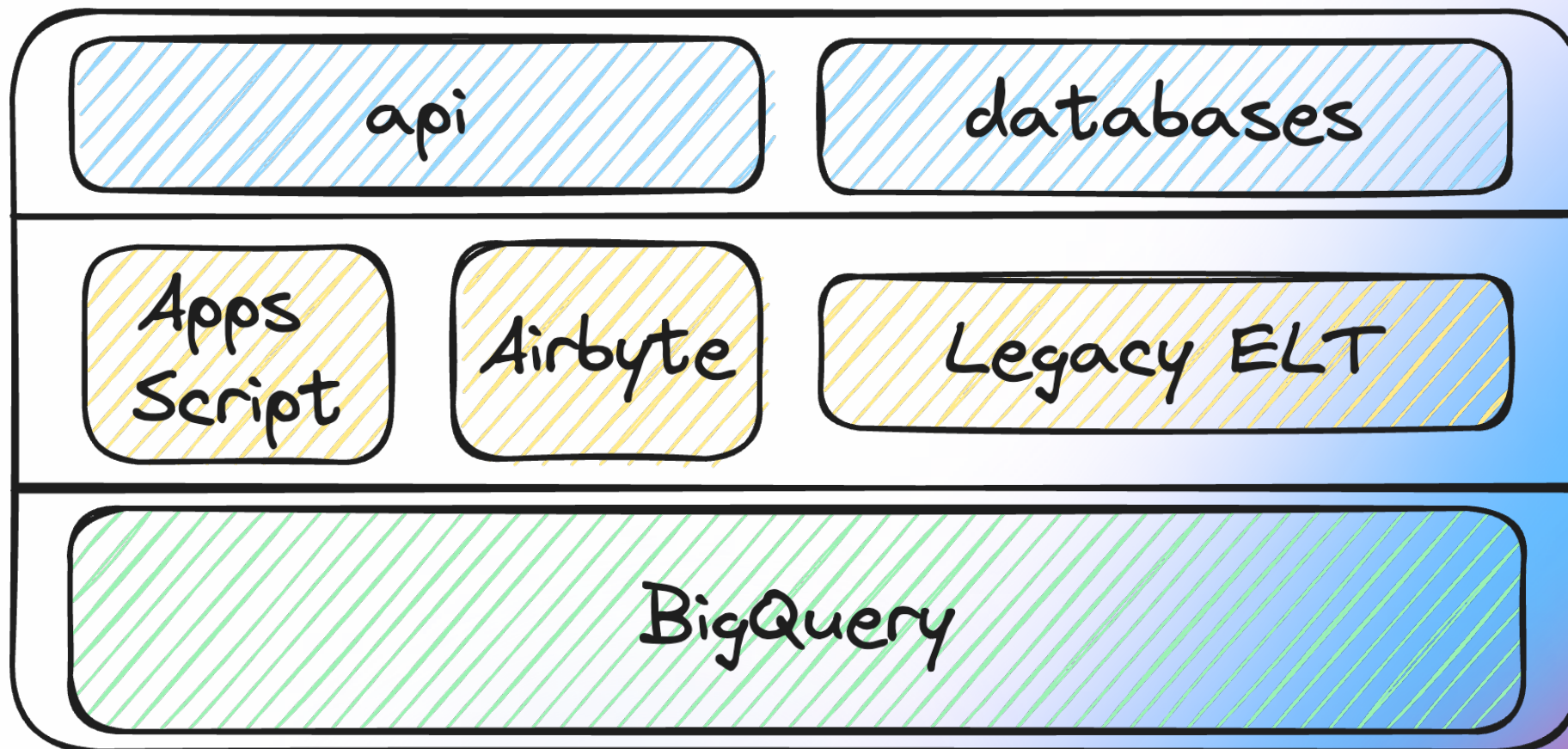
Airbyte. Sync Schedules

- Scheduled (e.g. every 24 hours, every 2 hours)
- Cron*
- Manual

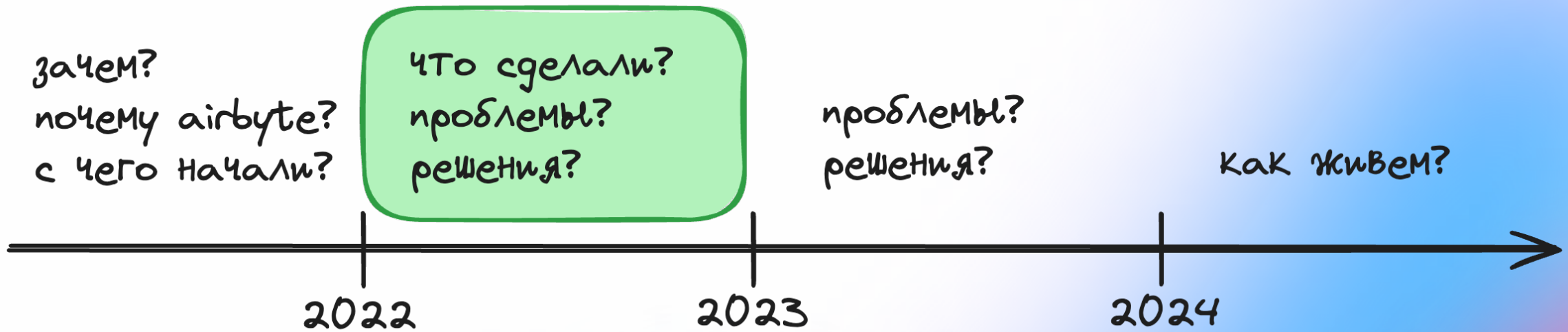
BACK TO end-2021. Что решили?

- Airbyte предоставляет CDK для простого создания коннекторов до API
- У Airbyte есть коннектор до BigQuery

BACK TO end-2021. Что получили?



План



BACK TO begin-2022

A Airbyte connectors		0	14	2
S	source-amocrm	★ 0		2 weeks ago
S	source-bitrix	★ 0		1 month ago
S	source-calltouch	★ 0		1 month ago
S	source-erp-1c-hr	★ 0		2 weeks ago
S	source-file	★ 0		1 month ago
S	source-huntflow	★ 0		1 month ago
S	source-omnidesk	★ 0		1 month ago
S	source-smsc	★ 0		1 month ago
S	source-topvisor	★ 0		1 month ago
S	source-usedesk	★ 0		1 week ago
S	source-vk-ads	★ 0		1 month ago
S	source-voximplant	★ 0		1 month ago
S	source-yandex-direct	★ 0		1 month ago

BACK TO begin-2022

	yandex-direct - yandex-direct-docdoc-clinics	BigQuery - BQ-yandex-direct	24 hours	1 hour ago	<input checked="" type="checkbox"/>
	yandex-direct - yandex-direct-docdoc-diagnostica	BigQuery - BQ-yandex-direct	24 hours	1 hour ago	<input checked="" type="checkbox"/>
	yandex-direct - yandex-direct-docdoc-general	BigQuery - BQ-yandex-direct	24 hours	59 minutes ago	<input checked="" type="checkbox"/>
	yandex-direct - yandex-direct-docdoc-network	BigQuery - BQ-yandex-direct	manual		<input type="button" value="Launch"/>
	yandex-direct - yandex-direct-docdoc-services	BigQuery - BQ-yandex-direct	manual		<input type="button" value="Launch"/>
	yandex-direct - yandex-direct-docdoc-spb	BigQuery - BQ-yandex-direct	24 hours	57 minutes ago	<input checked="" type="checkbox"/>
	yandex-direct - yandex-direct-docdoc-vrachi	BigQuery - BQ-yandex-direct	24 hours	29 minutes ago	<input checked="" type="checkbox"/>
	yandex-direct - yandex-direct-sberhealth-diagnostica	BigQuery - BQ-yandex-direct	24 hours	29 minutes ago	<input checked="" type="checkbox"/>
	yandex-direct - yandex-direct-sberhealth-msk	BigQuery - BQ-yandex-direct	24 hours	28 minutes ago	<input checked="" type="checkbox"/>
	yandex-direct - yandex-direct-sberhealth-spb	BigQuery - BQ-yandex-direct	manual		<input type="button" value="Launch"/>

BACK TO 2022



Google Cloud

Миграция
→



Yandex Cloud

BACK TO 2022. Миграция

- Сжатые сроки ~ 2 месяца
- Перенести как было
- Более 600 таблиц на расписании
- Расписание от раз в сутки до раз в час

BACK TO 2022



Airbyte

BACK TO 2022

Проблема: Нет коннектора для Greenplum
Но есть коннектор для Postgres
Решение: написать свой, переписав
постгресовый :)

BACK TO 2022

Проблема:

marketplace → greenplum ← **нейминг :(**

MySQL marketplace COMMUNITY → greenplum COMMUNITY

Status Job History **Replication** Transformation Settings

Activate the streams you want to sync Refresh source schema

Search stream name

Hide disabled streams

<input checked="" type="checkbox"/> Sync	Data destination ⚙	Stream ⚙	Sync mode ⓘ	Fields
<input checked="" type="checkbox"/> +	marketplace	agent_order_item	Full refresh Overwrite	All
<input checked="" type="checkbox"/> +	marketplace	amo_crm_config	Full refresh Overwrite	All
<input checked="" type="checkbox"/> +	marketplace	analysis	Full refresh Overwrite	All
<input checked="" type="checkbox"/> +	marketplace	analysis_biomaterial_instruction	Full refresh Overwrite	All
<input checked="" type="checkbox"/> +	marketplace	analysis_biomaterial_obtaining_method	Full refresh Overwrite	All
<input checked="" type="checkbox"/> +	marketplace	analysis_biomaterial_obtaining_method_price	Full refresh Overwrite	All

← все таблицы в одном коннекшене :(

BACK TO 2022

Решение:

- 1 таблица = 1 коннекшен
- , но создавать руками = БОЛЬ 🌀
- + хотели управлять конфигурациями как
КОДОМ

BACK TO 2022

Что делать?

- Использовать API для создания коннекшенов
- Написать CLI
- Сгенерировать YAML для всех коннекшенов

BACK TO 2022

marketplace:

type: mysql

schema: marketplace

streams:

- **timezone_list:**

source_sync_mode: full_refresh

destination_sync_mode: overwrite

schedule: 24

- **telecheckup_settings:**

source_sync_mode: incremental

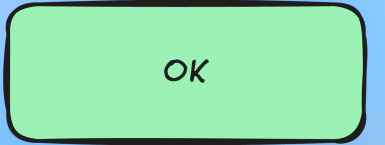
destination_sync_mode: append

cursorField: id

schedule: 24

...

== airbyte-cli validate ==>



BACK TO 2022

marketplace:

type: mysql

schema: marketplace

streams:

- **timezone_list:**

source_sync_mode: *overwrite*

destination_sync_mode: *full_refresh*

schedule: 24

- **telecheckup_settings:**

source_sync_mode: incremental

destination_sync_mode: append

cursorField: id

schedule: 24

...

== airbyte-cli validate ==>

не OK

BACK TO 2022

marketplace:

type: mysql

schema: marketplace

streams:

- **timezone_list:**

source_sync_mode: full_refresh

destination_sync_mode: overwrite

schedule: 24

- **telecheckup_settings:**

source_sync_mode: incremental

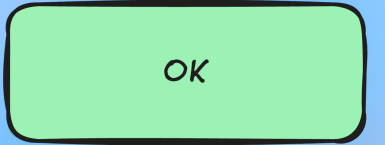
destination_sync_mode: append

cursorField: id









schedule: 24

...

————airbyte-cli apply————→



BACK TO 2022

NAME	DESTINATION NAME	CONNECTOR	FREQUENCY	LAST SYNC	ENABLED
✓ marketplace.timezone_list → greenplu...	greenplum	 Greenplum	24 hours	2 hours ago	<input checked="" type="checkbox"/>
✓ marketplace.telecheckup_settings → ...	greenplum	 Greenplum	24 hours	1 hour ago	<input checked="" type="checkbox"/>
✓ marketplace.subscription_pay_rule → ...	greenplum	 Greenplum	24 hours	2 hours ago	<input checked="" type="checkbox"/>
✓ marketplace.specialization → greenpl...	greenplum	 Greenplum	7 hours	5 hours ago	<input checked="" type="checkbox"/>
✓ marketplace.snils → greenplum	greenplum	 Greenplum	24 hours	1 hour ago	<input checked="" type="checkbox"/>
✓ marketplace.sms_log → greenplum	greenplum	 Greenplum	7 hours	4 hours ago	<input checked="" type="checkbox"/>
✓ marketplace.smartblock_smartcard_s...	greenplum	 Greenplum	7 hours	4 hours ago	<input checked="" type="checkbox"/>
✓ marketplace.shop → greenplum	greenplum	 Greenplum	24 hours	3 hours ago	<input checked="" type="checkbox"/>

BACK TO 2022

Проблема: Tinyint as bool

```
select distinct payment_type from partner;
```

```
+-----+
| payment_type |
+-----+
|      true    |
+-----+
|      null    |
+-----+
```


BACK TO 2022

Проблема: Tinyint as bool

```
select distinct payment_type from partner;
```

```
+-----+
| payment_type |
+-----+
|      true    |
+-----+
|      null    |
+-----+
```



BACK TO 2022

Проблема: Tinyint as bool

```
case BOOLEAN -> putBoolean(json, columnName, resultSet, colIndex);
case TINYINT, TINYINT_UNSIGNED -> {
    if (field.getLength() == 1L) {
        // TINYINT(1) is boolean
        putBoolean(json, columnName, resultSet, colIndex);
    } else {
        putShortInt(json, columnName, resultSet, colIndex);
    }
}
```

Решение:

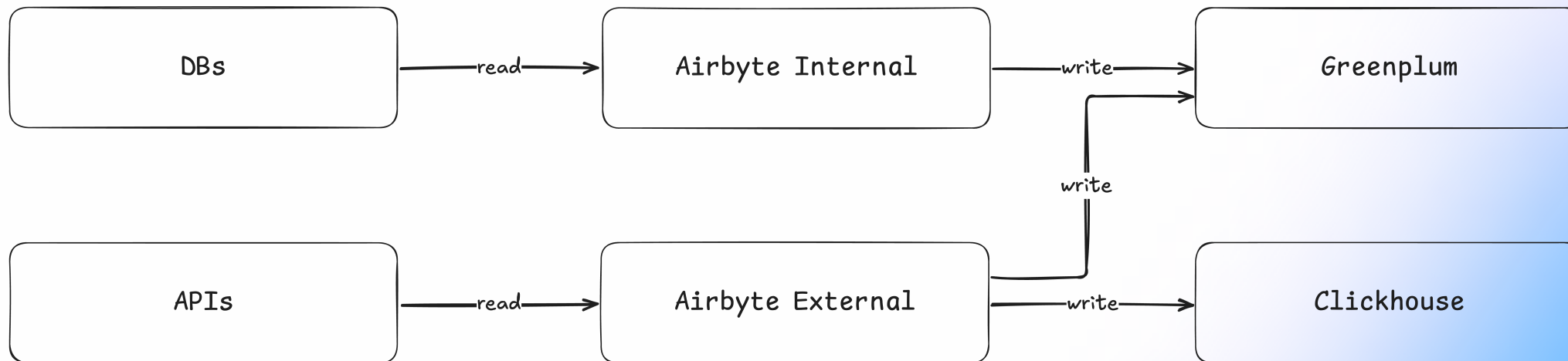
JDBC URL Params: &tinyInt1isBit=false

BACK TO 2022

payment_type
1
2
3
4
5
null



BACK TO end-2022. Что получили?

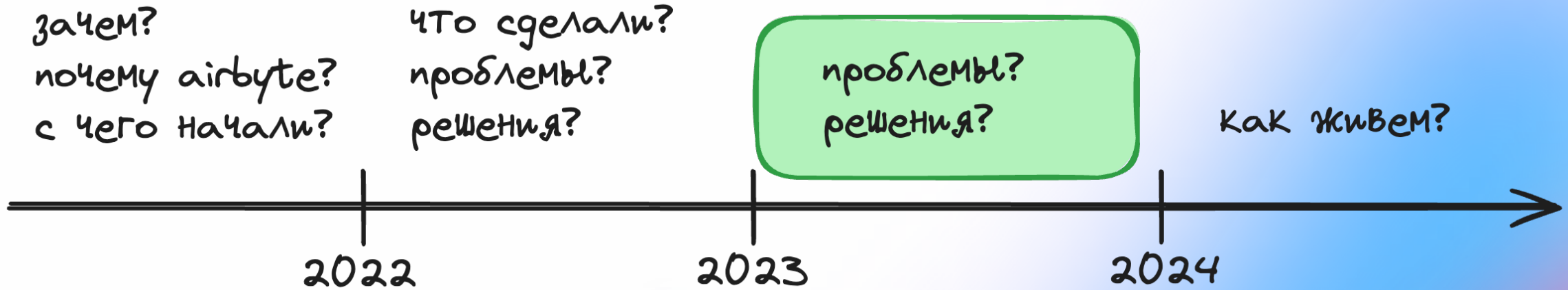


- Забирали данные по jdbc из баз данных
- Забирали данные из API
- Писали в Greenplum
- Писали в Clickhouse

BACK TO 2022. Какие проблемы оставались?

- Не было расписания по крону
 - Не было возможности исключать колонки из синхронизации
- > приходилось обновлять мету в базе :(

План



Проблема: аудитное поле на источнике с точностью до секунд, забираем по строгому неравенству

id	rating	created_at	updated_at
1	0	2024-07-17 11:38:23	2024-07-17 11:38:23
2	3	2024-08-02 00:44:19	2024-08-02 00:44:19
3	1	2024-08-02 14:12:43	2024-08-03 14:00:43

cursor_field: updated_at

cursor: 2024-08-03 14:00:43

Проблема: аудитное поле на источнике с точностью до секунд, забираем по строгому неравенству

id	rating	created_at	updated_at
1	5	2024-07-17 11:38:23	2024-08-03 14:00:43
2	3	2024-08-02 00:44:19	2024-08-02 00:44:19
3	1	2024-08-02 14:12:43	2024-08-03 14:00:43

```
cursor_field: updated_at
```

```
cursor: 2024-08-03 14:00:43
```

```
select * from rating where updated_at > '2024-08-03 14:00:43';
```


BACK TO 2023

Проблема: аудитное поле на источнике с точностью до секунд, забираем по строгому неравенству -> теряем данные

Решение: использовать CDC

BACK TO 2023

marketplace:

type: mysql

host: \$marketplace_host

database: \$marketplace_database

port: \$marketplace_port

username: \$marketplace_username

password: \$marketplace_password

marketplace:

type: mysql

host: \$marketplace_host

database: \$marketplace_database

port: \$marketplace_port

username: \$marketplace_username

password: \$marketplace_password

replication_method: cdc

BACK TO 2023

Support cron strings

Проблема: При ручных запусках расписание ломалось

Решение: перенос запусков на расписание по крону

BACK TO 2023

Support cron strings

marketplace:

type: mysql

schema: marketplace

streams:

- **timezone_list:**

source_sync_mode: full_refresh

destination_sync_mode: overwrite

schedule: 24

...

marketplace:

type: mysql

schema: marketplace

streams:

- **timezone_list:**

source_sync_mode: full_refresh

destination_sync_mode: overwrite

schedule:

cron: "0 0 4 * * ?"

...

BACK TO 2023

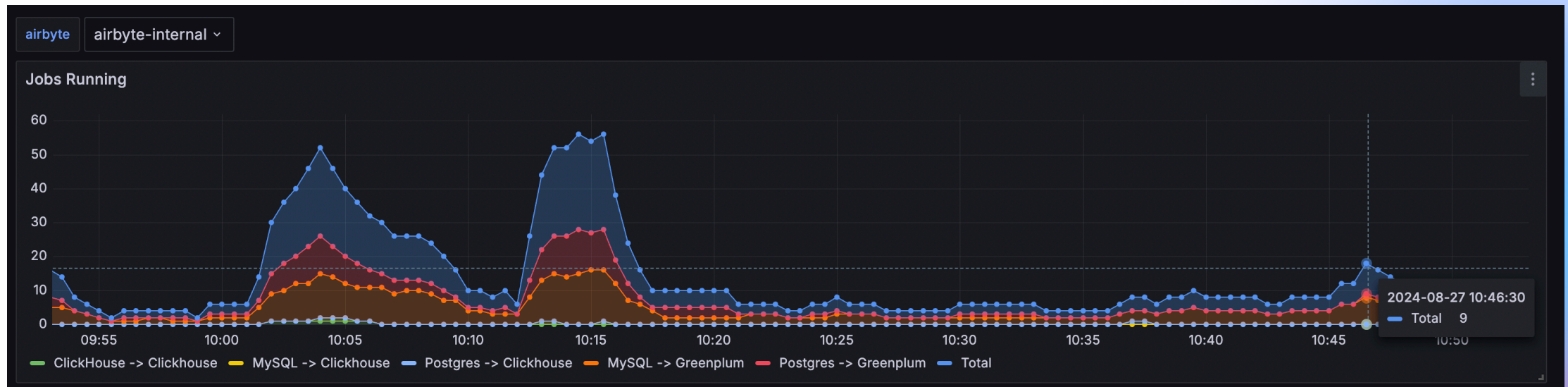
Проблема: красивый мониторинг для
DataDog из коробки, но у нас его нет

Решение: собираем метрики через sql-
exporter

BACK TO 2023



BACK TO 2023



BACK TO 2023

Проблема: каждый запуск требует минимум 3 пода для операций:

- orchestrator
- read
- write

-> большое количество «мертвых» подов от запусков.

Решение: cleanup-policy

BACK TO 2023

```
apiVersion: batch/v1
```

```
kind: CronJob
```

```
metadata:
```

```
  name: airbyte-completed-pods-cleanup
```

```
  namespace: airbyte
```

```
spec:
```

```
  ...
```

```
  command: ["sh", "-c", "kubectl -n airbyte delete pods --  
field-selector status.phase!=Running,status.phase!  
=Pending"]
```

BACK TO 2023

Проблема: большое количество запусков
может влиять на работу ui и api

Решение: разнести статичные сервисы и
синки на разные ноды

BACK TO 2023

```
# static services
```

```
webapp:
```

```
  nodeSelector:
```

```
    type: static
```

```
✓airbyte-bootloader
```

```
✓server
```

```
✓temporal
```

```
# dynamic services
```

```
worker:
```

```
  nodeSelector:
```

```
    type: dynamic
```

```
✓orchestrator
```

```
✓workload-api-server
```

```
✓workload-launcher
```

ВАСК ТО end-2023. Что получили?

- Забирали данные по jdbc из баз данных
- Забирали данные по cdc из баз данных
- Забирали данные из API
- Писали в Greenplum
- Писали в Clickhouse
- Писали в s3

BACK TO end-2023

Какие проблемы оставались?

- Ролевая модель (только в Cloud)
- Бэкфилл данных

План



NOW (2024)

- Проблем пока не было 😎

Как живем?

- Синки из баз переехали в Debezium

Почему?

- Целевое решение
- Гибкость в управлении коннекторами

NOW (2024)

- Проблем пока не было 😎

Как живем?

- Продолжаем писать коннекторы до API

NOW (2024)

Что еще сделали?

- Перенесли запуск коннекшенов в Dagster

Почему?

- Оркестрация в одном месте
- Можем строить единый пайплайн загрузки данных

Выводы

🔥 Python CDK для API

👍 Для баз, если:

- синхронизации as-is
- не нужен запуск с окном
- достаточно стандартной конфигурации для cdc

🤔 Если синхронизаций >1000

Вопросы?