СБЕР ЗДОРОВЬЕ

Airbyte.

2 года в продакшене

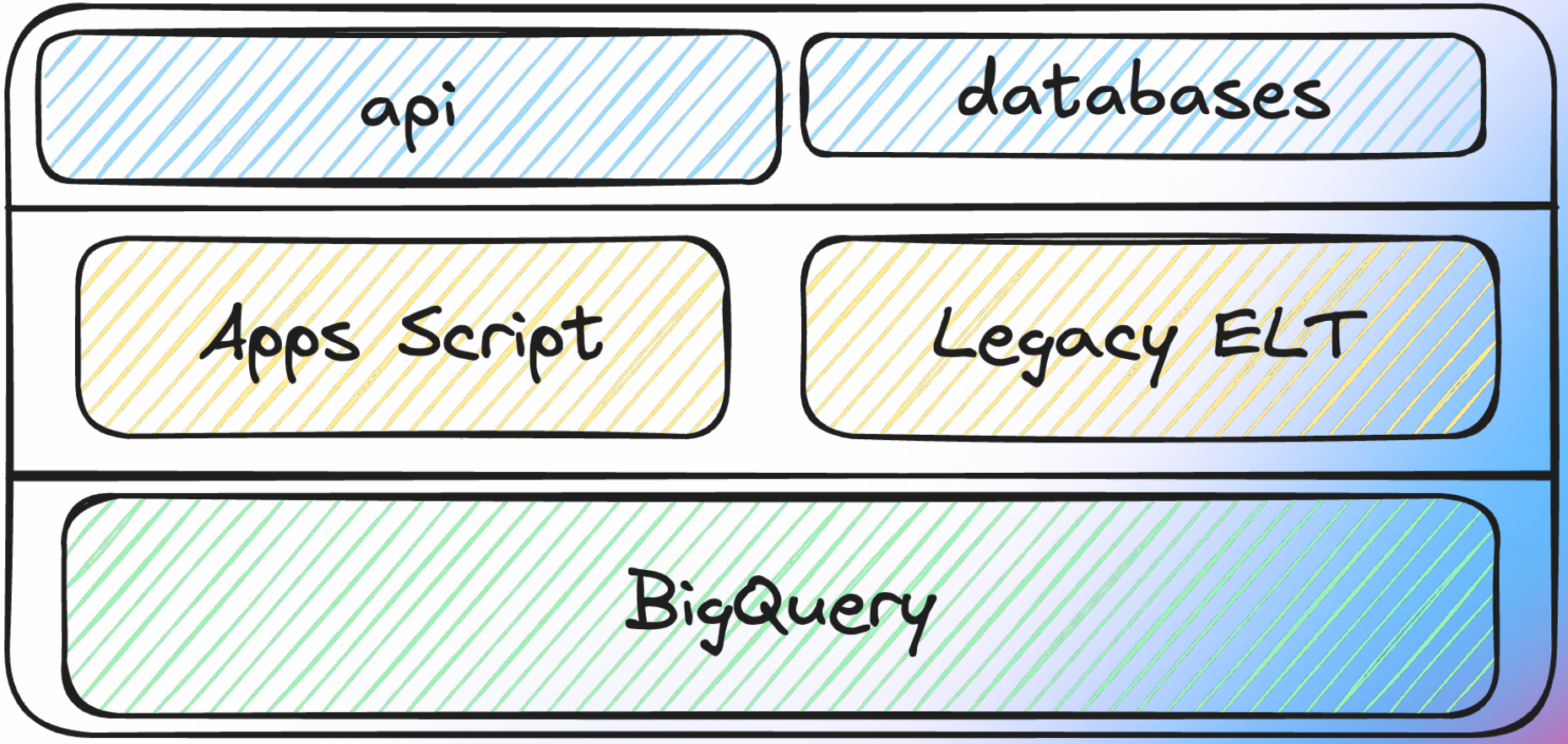
Спикер:
Попова Александра, Lead Data Engineer

План

СБЕР ЗДОРОВЬЕ

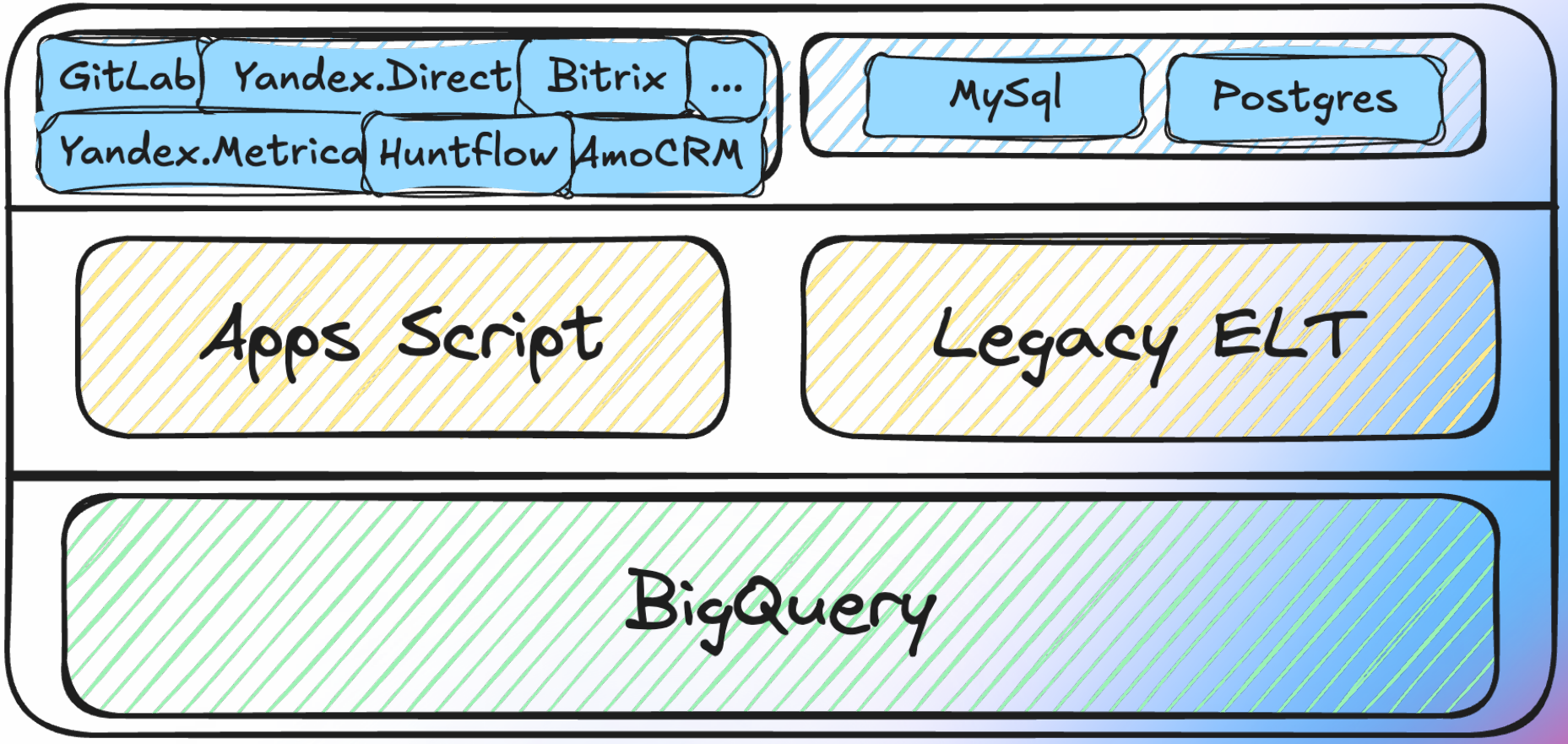
BACK TO end-2021

СБЕР ЗДОРОВЬЕ



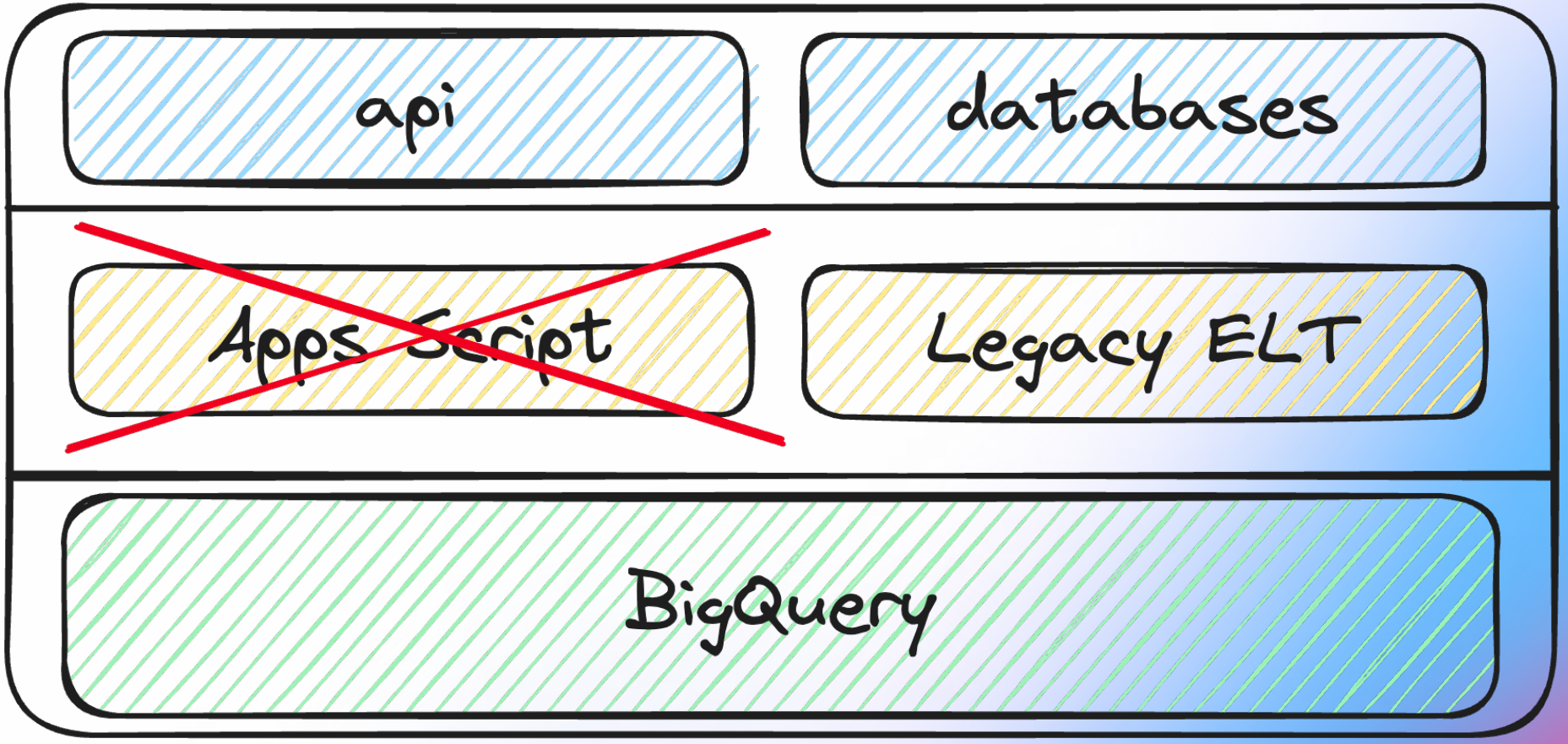
BACK TO end-2021

СБЕР ЗДОРОВЬЕ



BACK TO end-2021

СБЕР ЗДОРОВЬЕ



BACK TO end-2021. Что хотели?

СБЕР ЗДОРОВЬЕ

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

СБЕР ЗДОРОВЬЕ

Airbyte

15,000+
community members

10,000+
Github stars

1PB+
synced/month

800+
contributors

100k+
deployments

*на момент подготовки презентации

*про выбор было тут https://smartdataconf.ru/

СБЕР ЗДОРОВЬЕ

Airbyte. Что было?

Airbyte. Что было? Connector Development Kit

СБЕР ЗДОРОВЬЕ

- Python CDK
- Java CDK

Airbyte. Что было? Connector Development Kit

СБЕР ЗДОРОВЬЕ

- 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

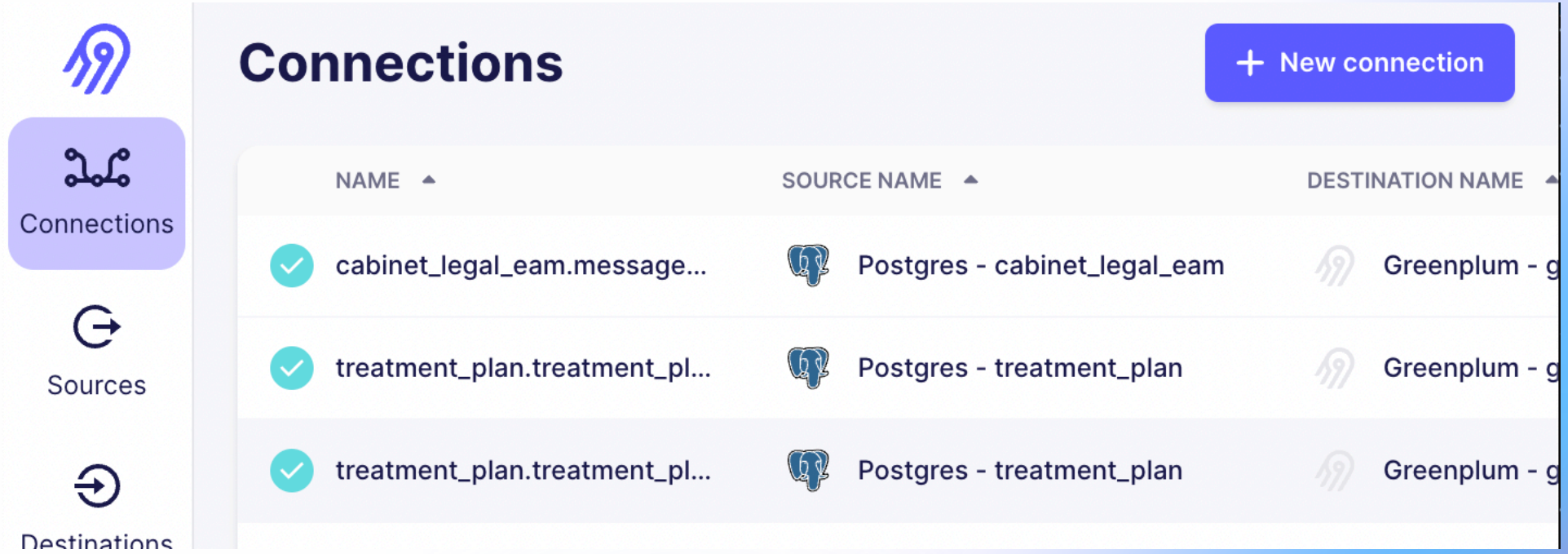
СБЕР ЗДОРОВЬЕ

- 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]]]:
```

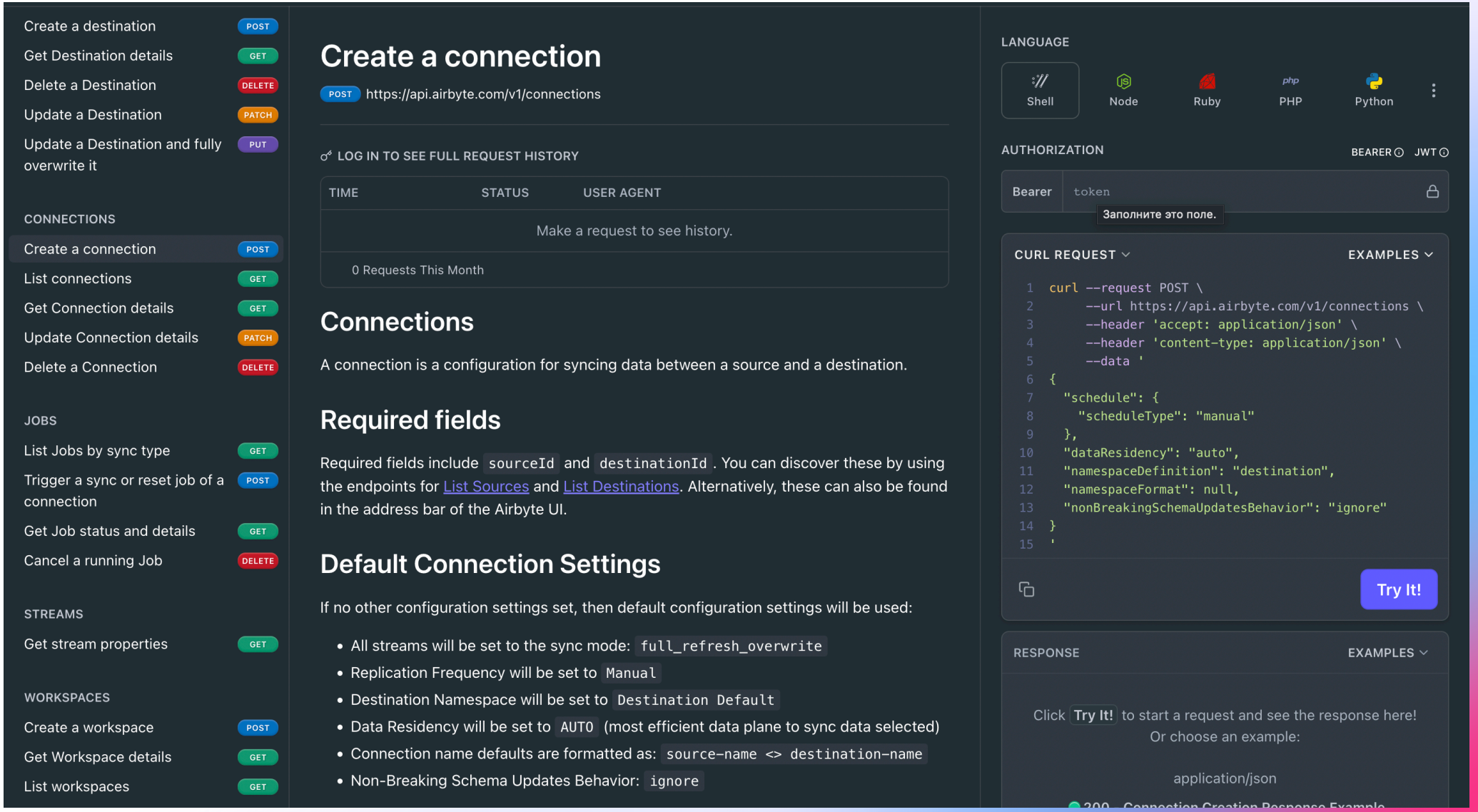
Airbyte. Что было? UI

СБЕР ЗДОРОВЬЕ



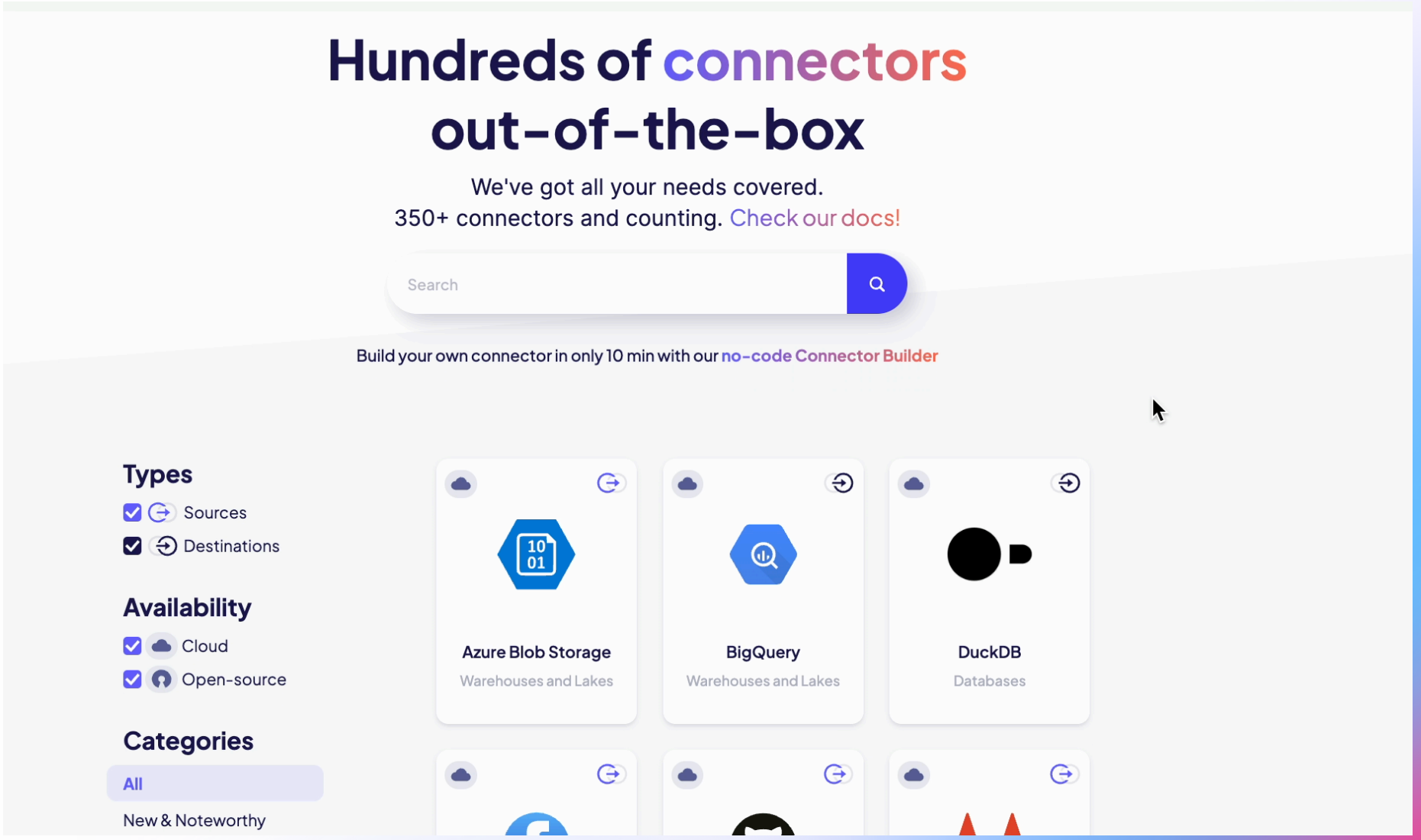
Airbyte. Что было? API

СБЕР ЗДОРОВЬЕ



Airbyte. Что было? Готовые коннекторы

СБЕР ЗДОРОВЬЕ



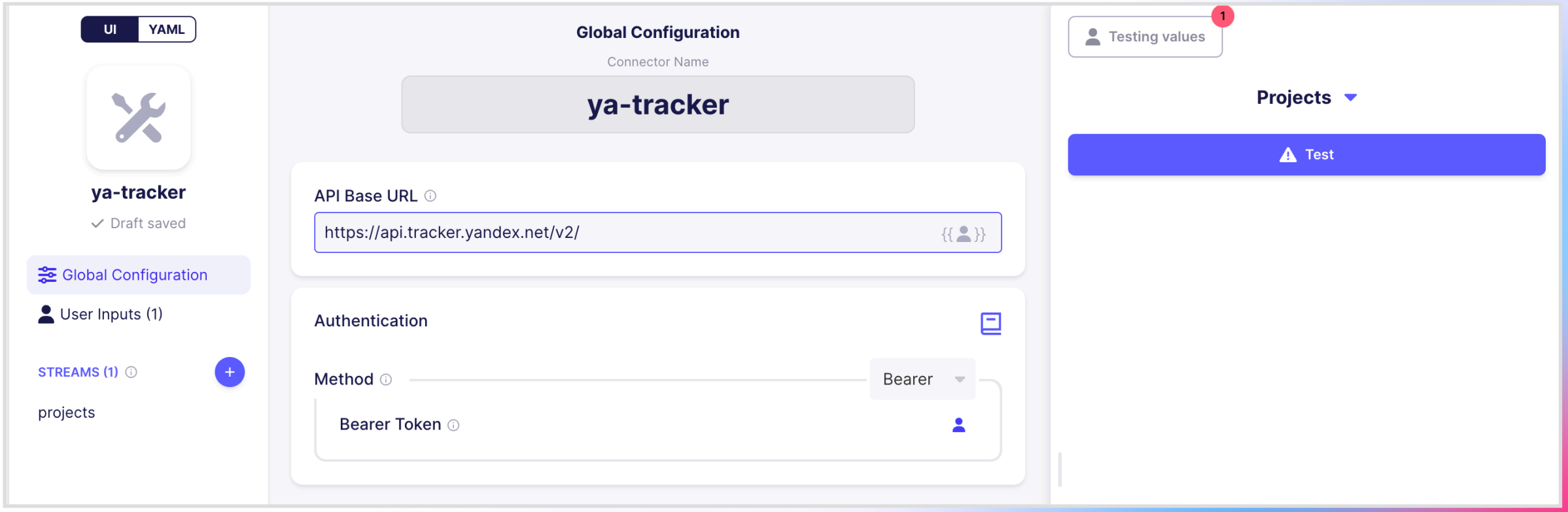
https://airbyte.com/connectors

СБЕР ЗДОРОВЬЕ

Airbyte. Что есть сейчас?

Airbyte. Что есть сейчас? No-code Connector builder

СБЕР ЗДОРОВЬЕ



Airbyte. Что есть сейчас? Low code CDK

СБЕР ЗДОРОВЬЕ

```yaml
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: {}
```

```yaml
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

СБЕР ЗДОРОВЬЕ

- pip install airbyte

Airbyte. Что есть сейчас? PyAirbyte

СБЕР ЗДОРОВЬЕ

Для чего?

- Для локального тестирования без сборки/ пересборки контейнера

Airbyte. Что есть сейчас? PyAirbyte

СБЕР ЗДОРОВЬЕ

```python
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

СБЕР ЗДОРОВЬЕ

```
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
}
```

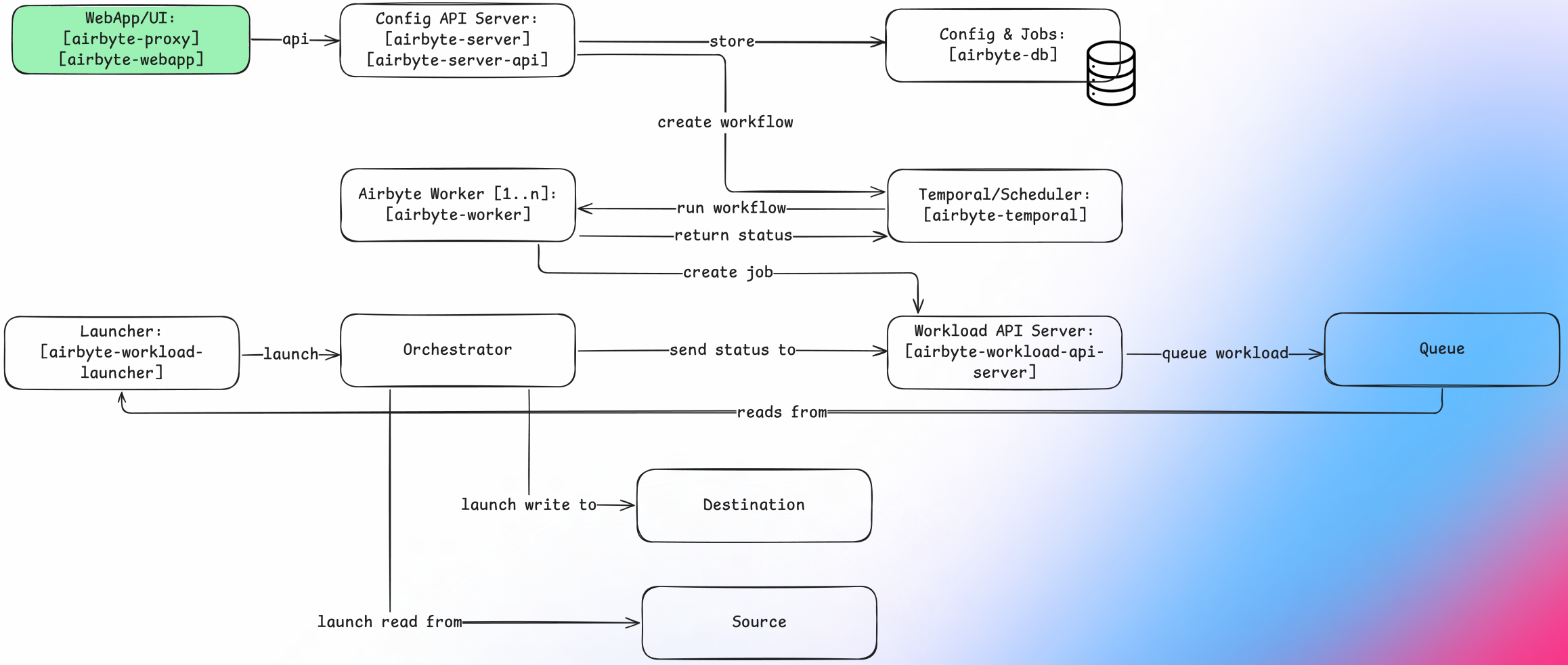
Airbyte. Что есть сейчас? Terraform Provider

СБЕР ЗДОРОВЬЕ

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

Airbyte. Архитектура сегодня

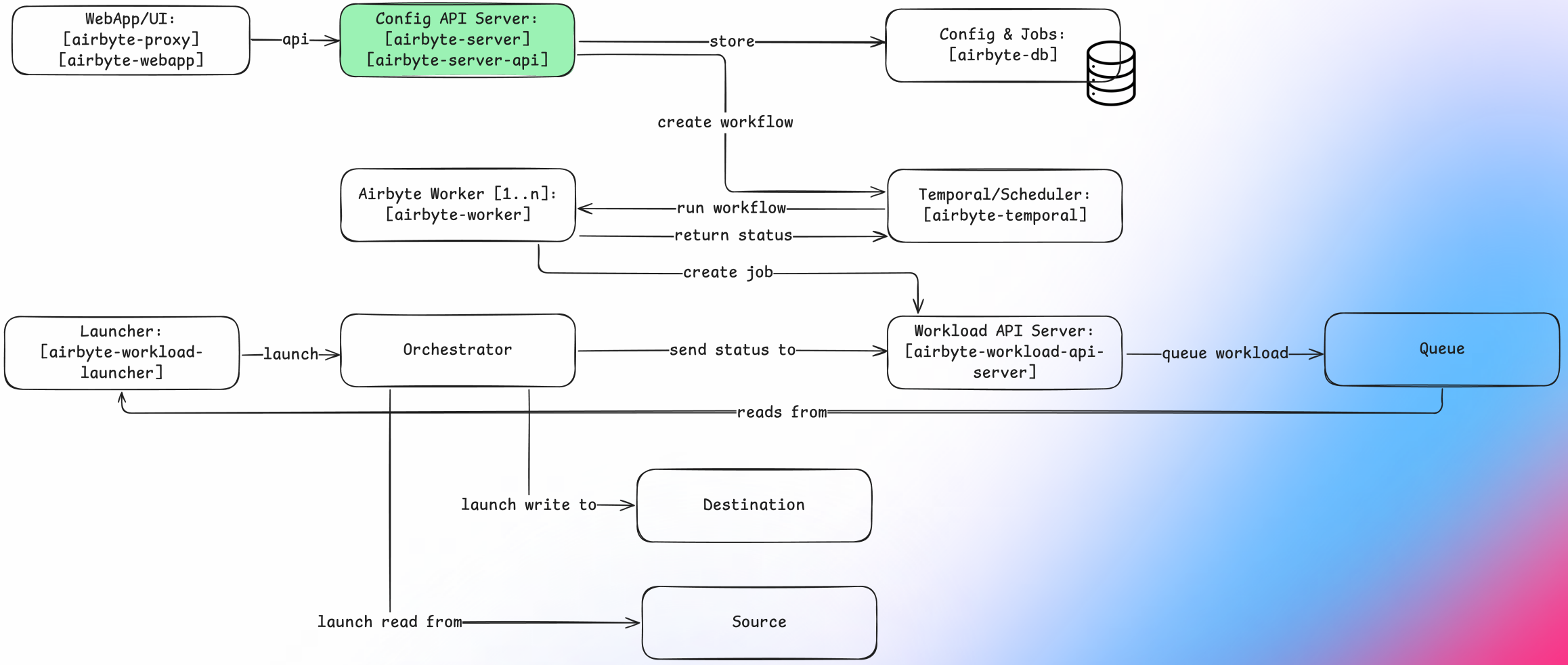
СБЕР ЗДОРОВЬЕ



https://docs.airbyte.com/understanding-airbyte/high-level-view

Airbyte. Архитектура сегодня

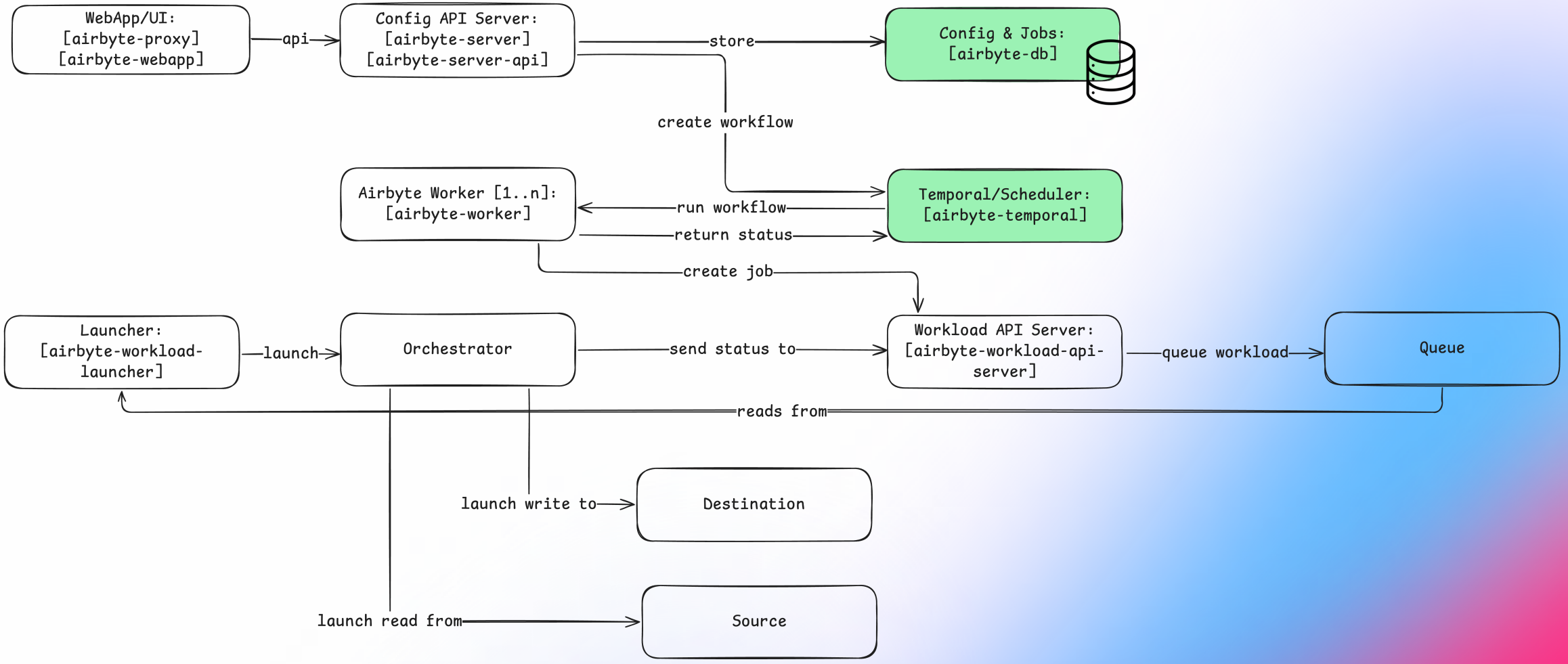
СБЕР ЗДОРОВЬЕ



https://docs.airbyte.com/understanding-airbyte/high-level-view

Airbyte. Архитектура сегодня

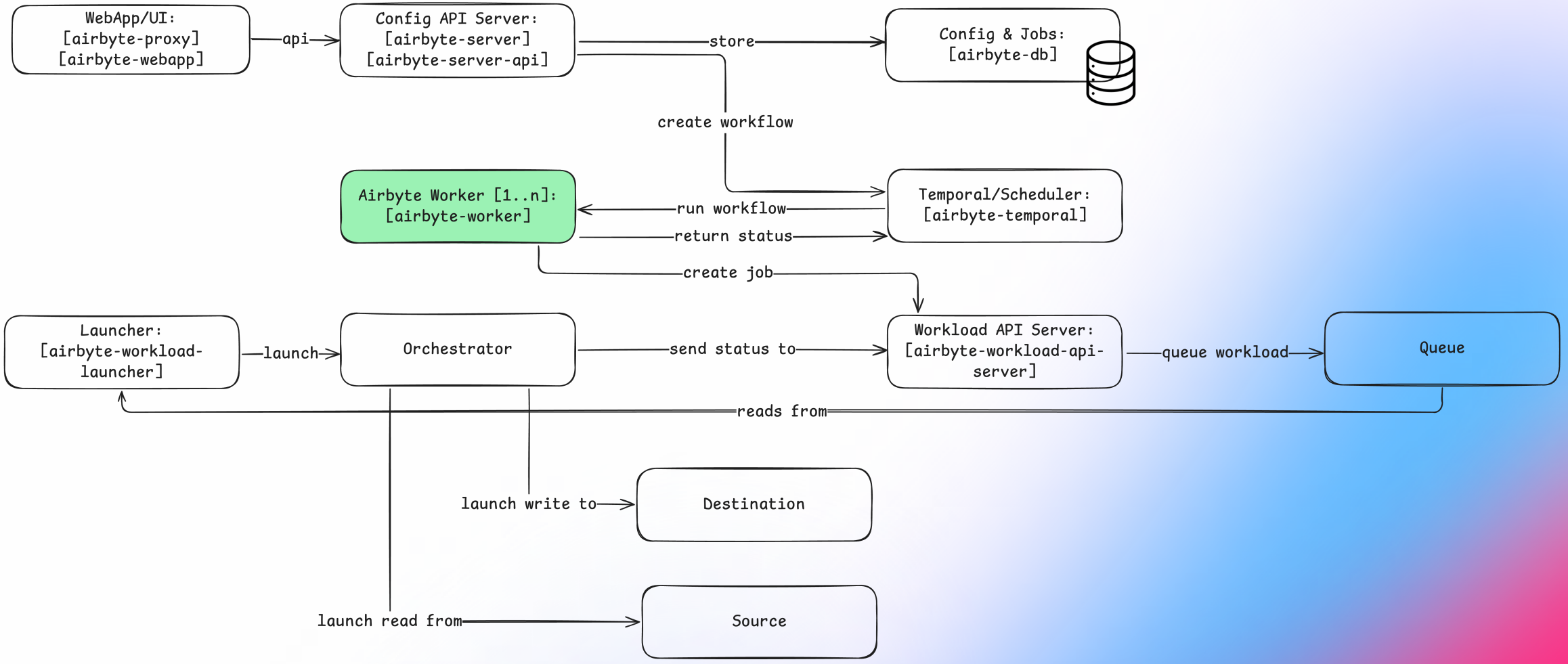
СБЕР ЗДОРОВЬЕ



https://docs.airbyte.com/understanding-airbyte/high-level-view

Airbyte. Архитектура сегодня

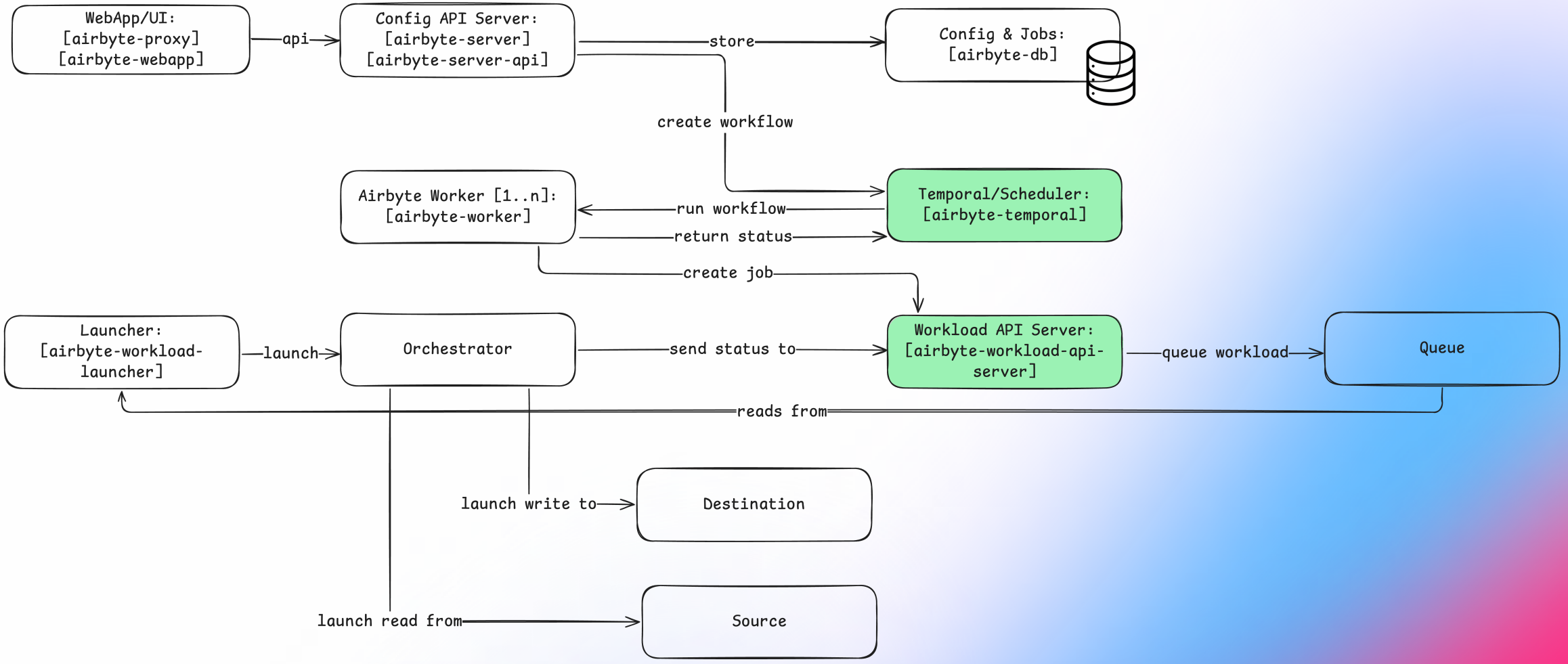
СБЕР ЗДОРОВЬЕ



https://docs.airbyte.com/understanding-airbyte/high-level-view

Airbyte. Архитектура сегодня

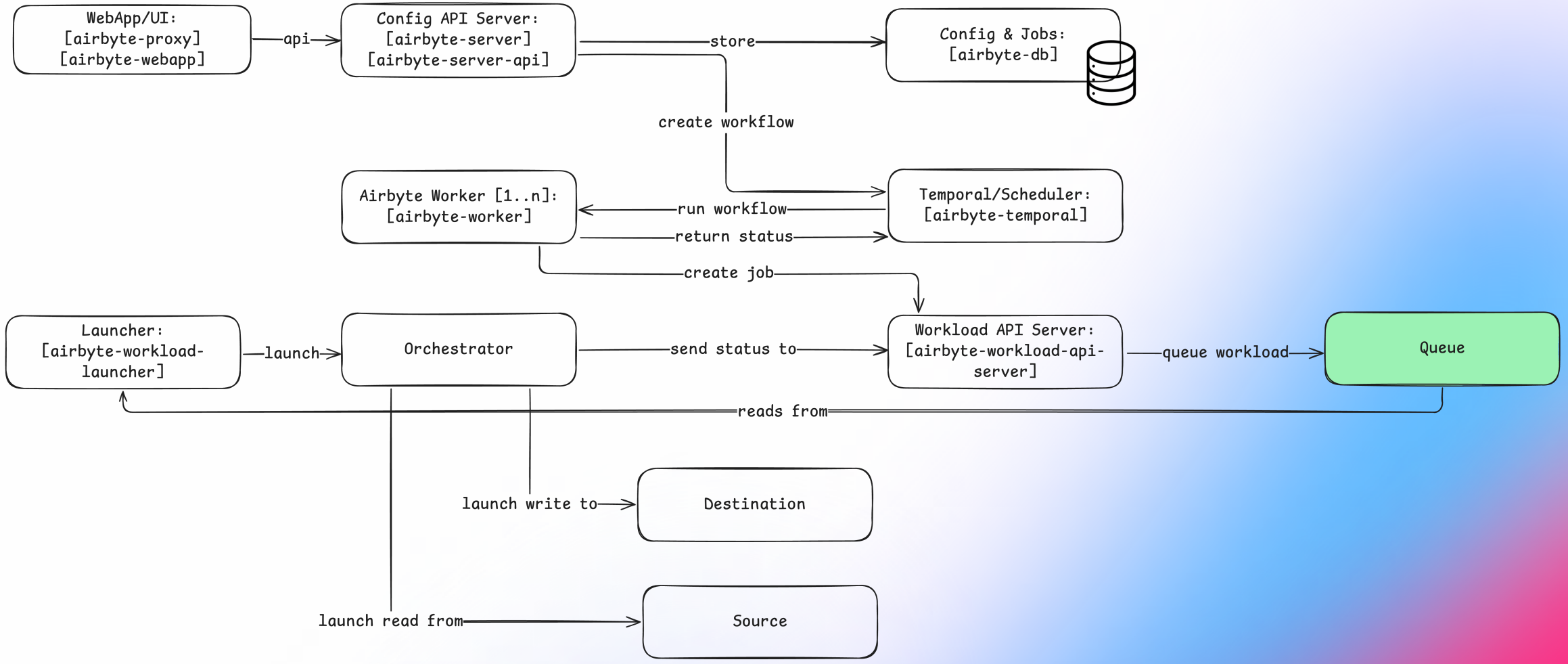
СБЕР ЗДОРОВЬЕ



https://docs.airbyte.com/understanding-airbyte/high-level-view

Airbyte. Архитектура сегодня

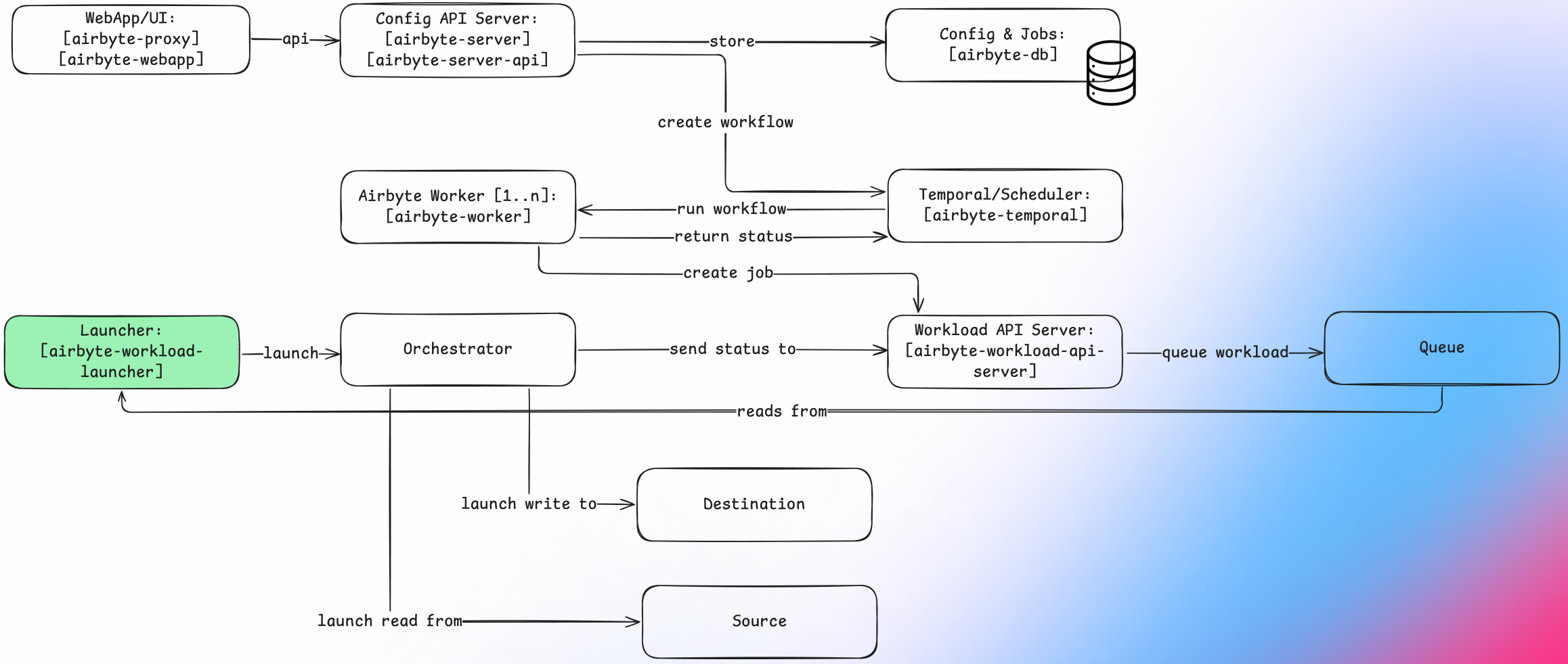
СБЕР ЗДОРОВЬЕ



https://docs.airbyte.com/understanding-airbyte/high-level-view

Airbyte. Архитектура сегодня

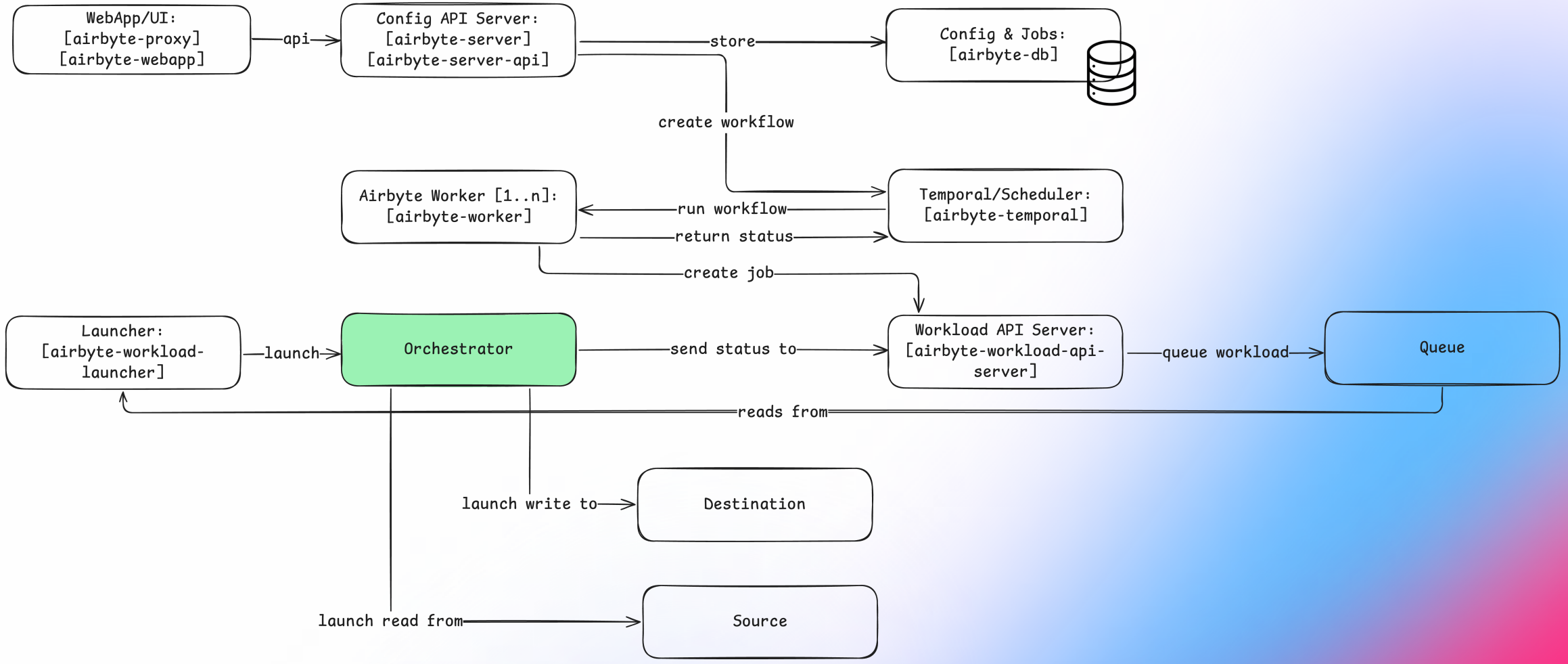
СБЕР ЗДОРОВЬЕ



https://docs.airbyte.com/understanding-airbyte/high-level-view

Airbyte. Архитектура сегодня

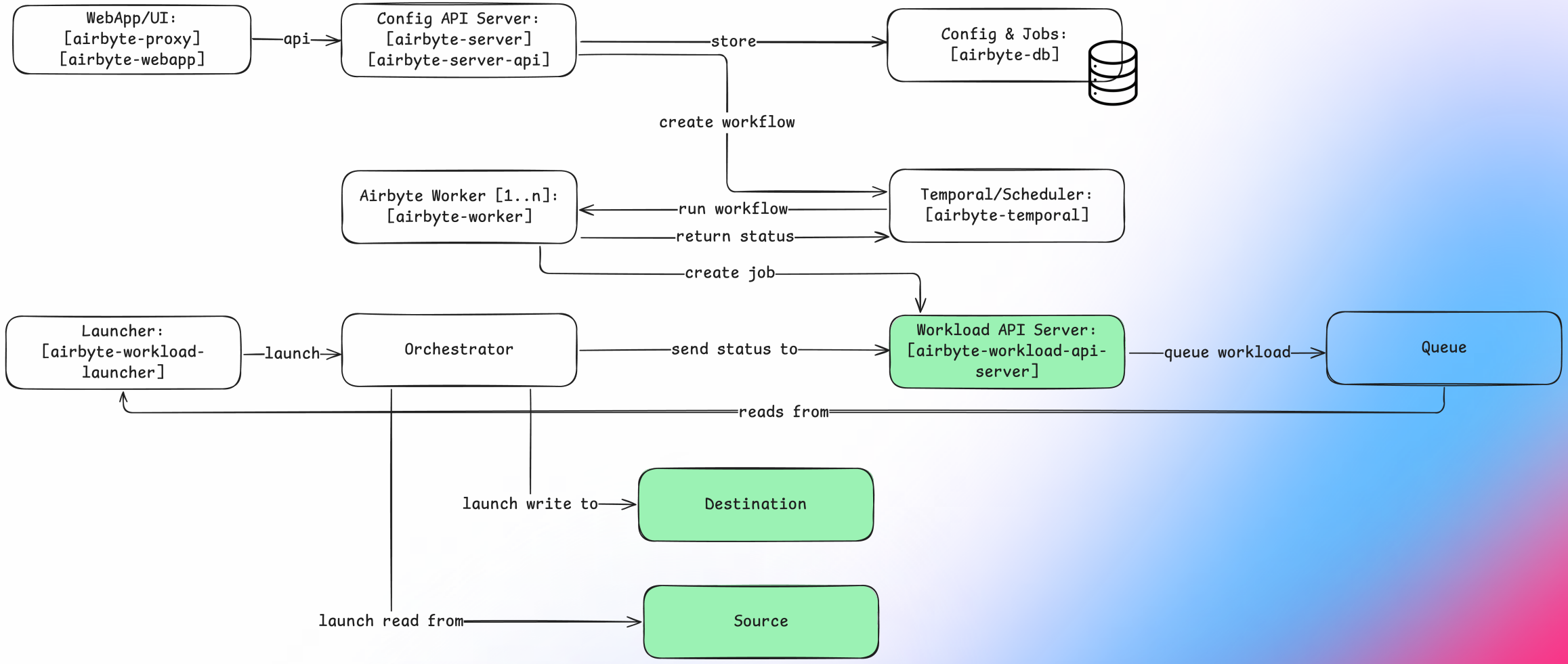
СБЕР ЗДОРОВЬЕ



https://docs.airbyte.com/understanding-airbyte/high-level-view

Airbyte. Архитектура сегодня

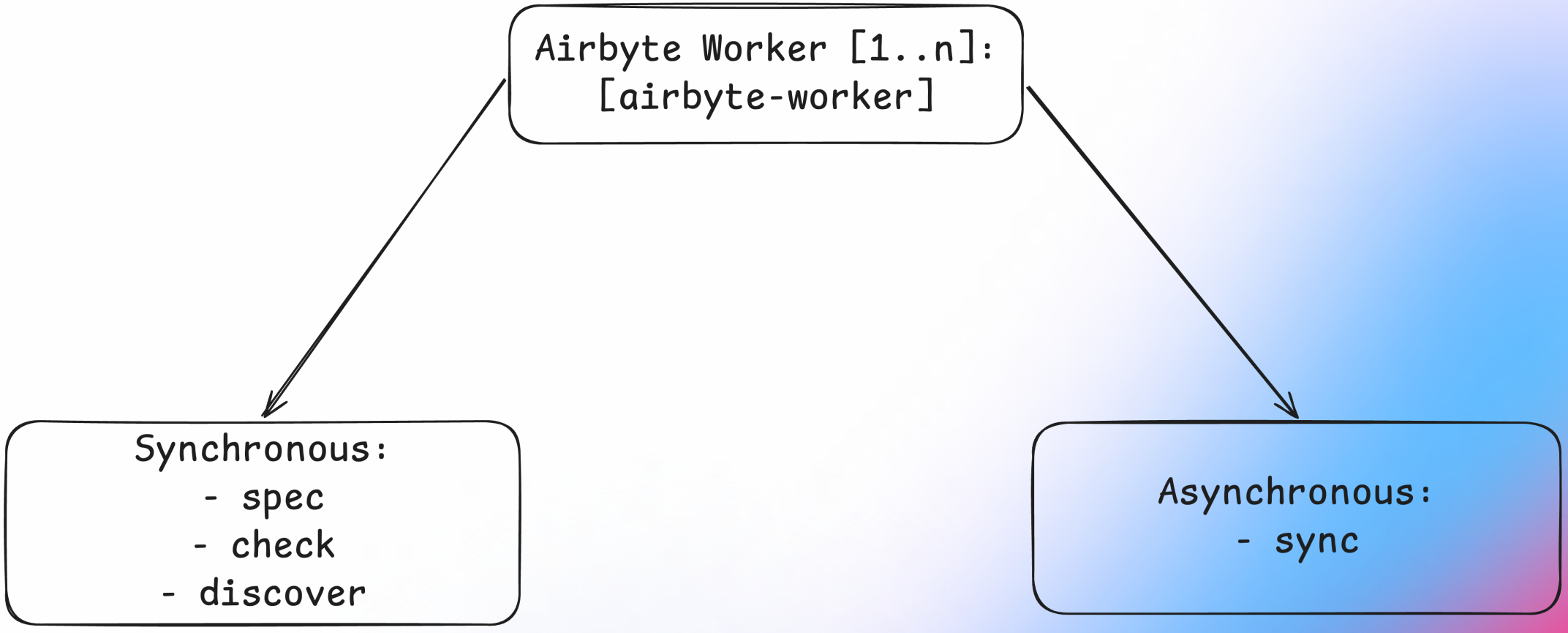
СБЕР ЗДОРОВЬЕ



https://docs.airbyte.com/understanding-airbyte/high-level-view

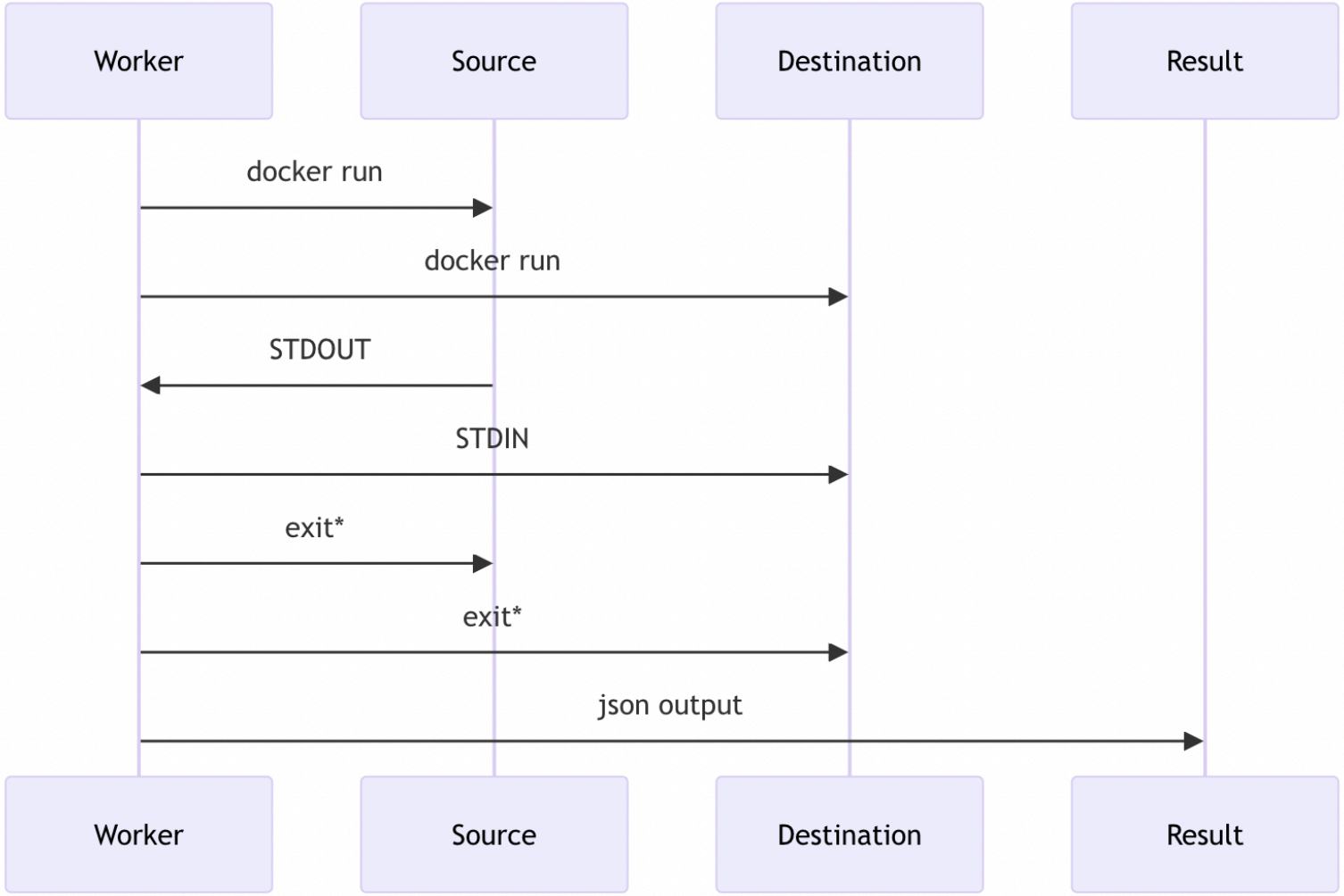
Airbyte. Workers

СБЕР ЗДОРОВЬЕ



Airbyte. Workers

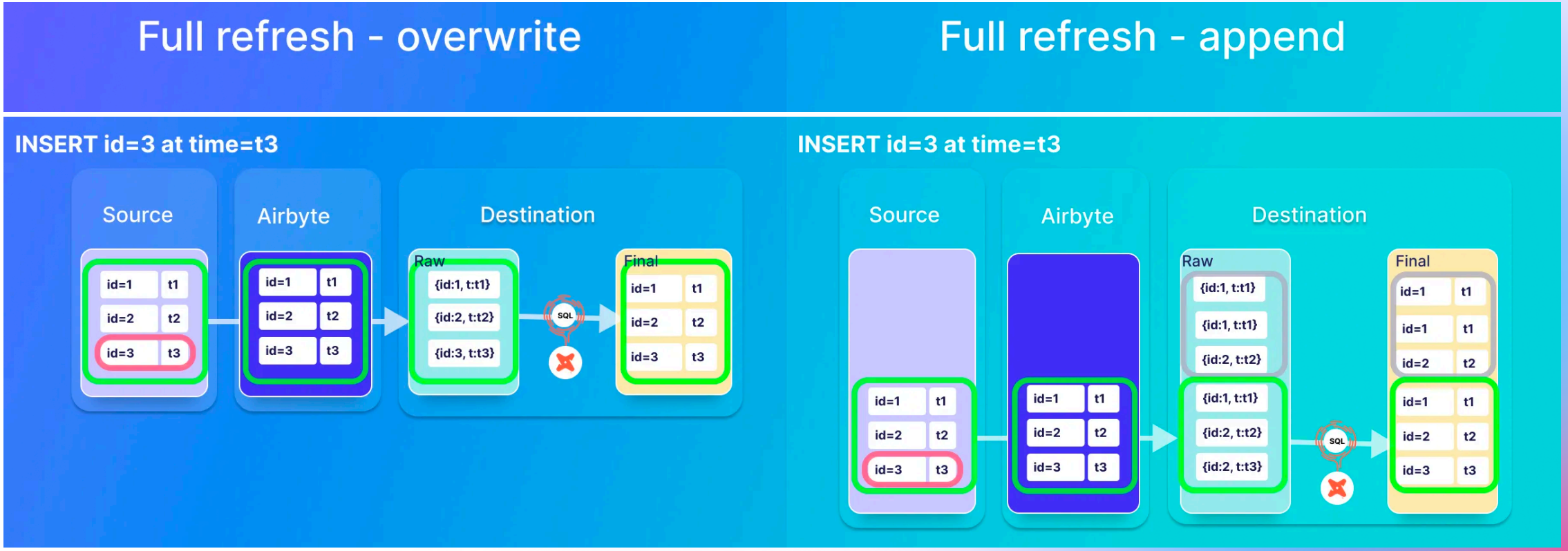
СБЕР ЗДОРОВЬЕ



https://docs.airbyte.com/understanding-airbyte/jobs

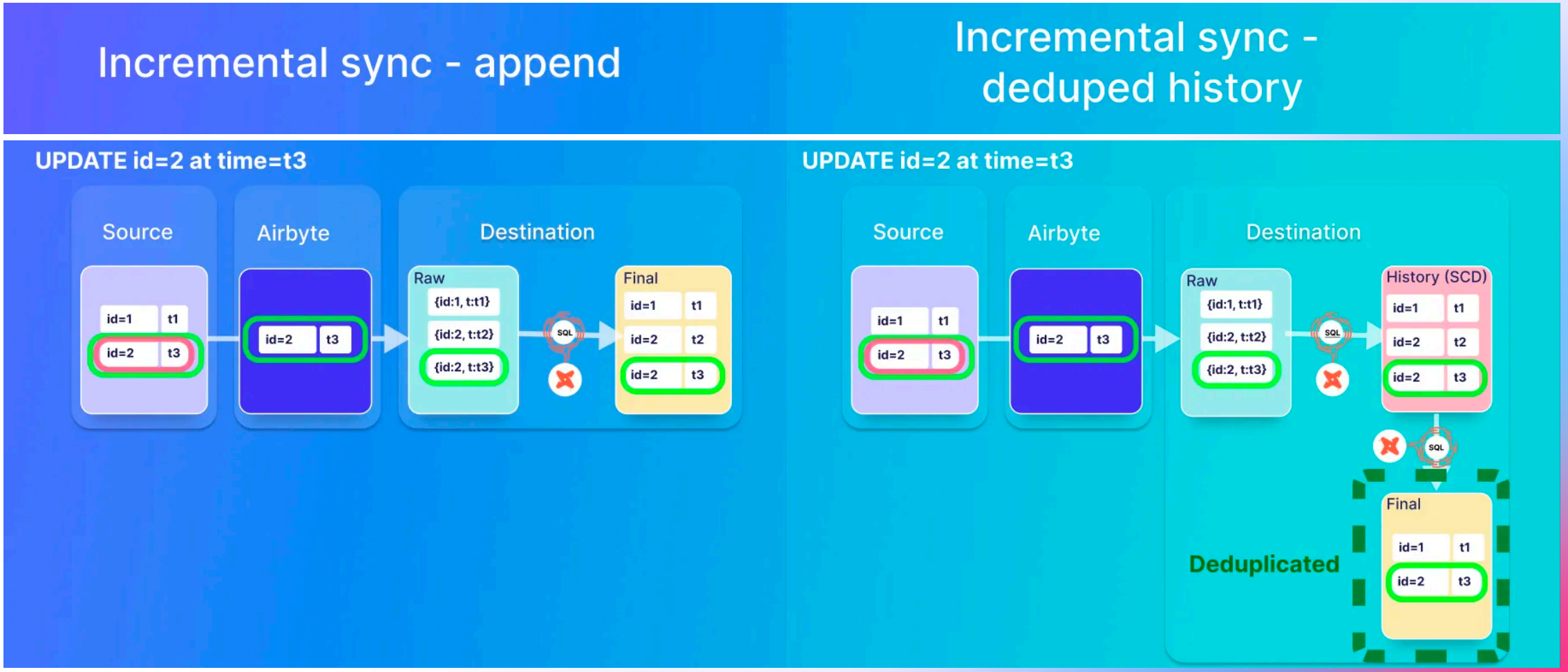
Airbyte. Sync Modes: Full Refresh

СБЕР ЗДОРОВЬЕ



Airbyte. Sync Modes: Incremental

СБЕР ЗДОРОВЬЕ

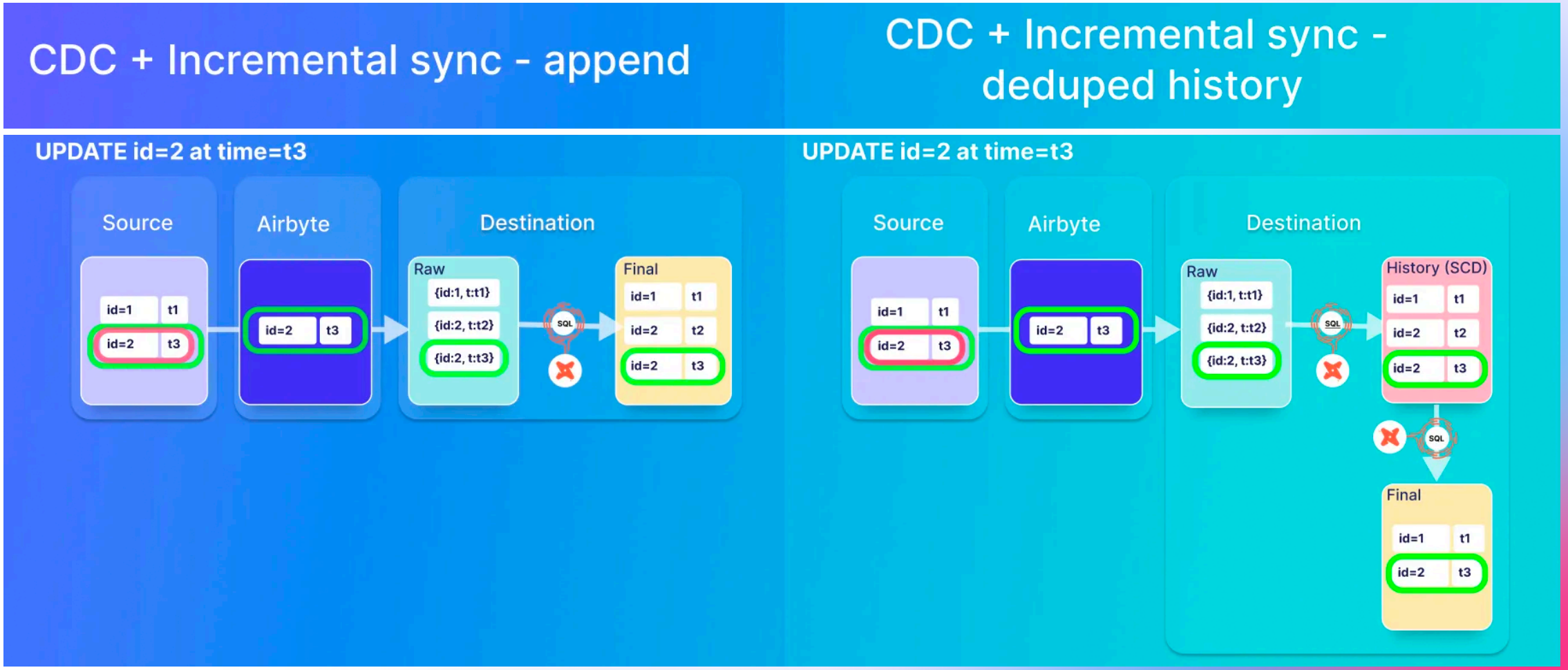


Airbyte. Sync Modes: Incremental

СБЕР ЗДОРОВЬЕ

Airbyte. Sync Modes: CDC*

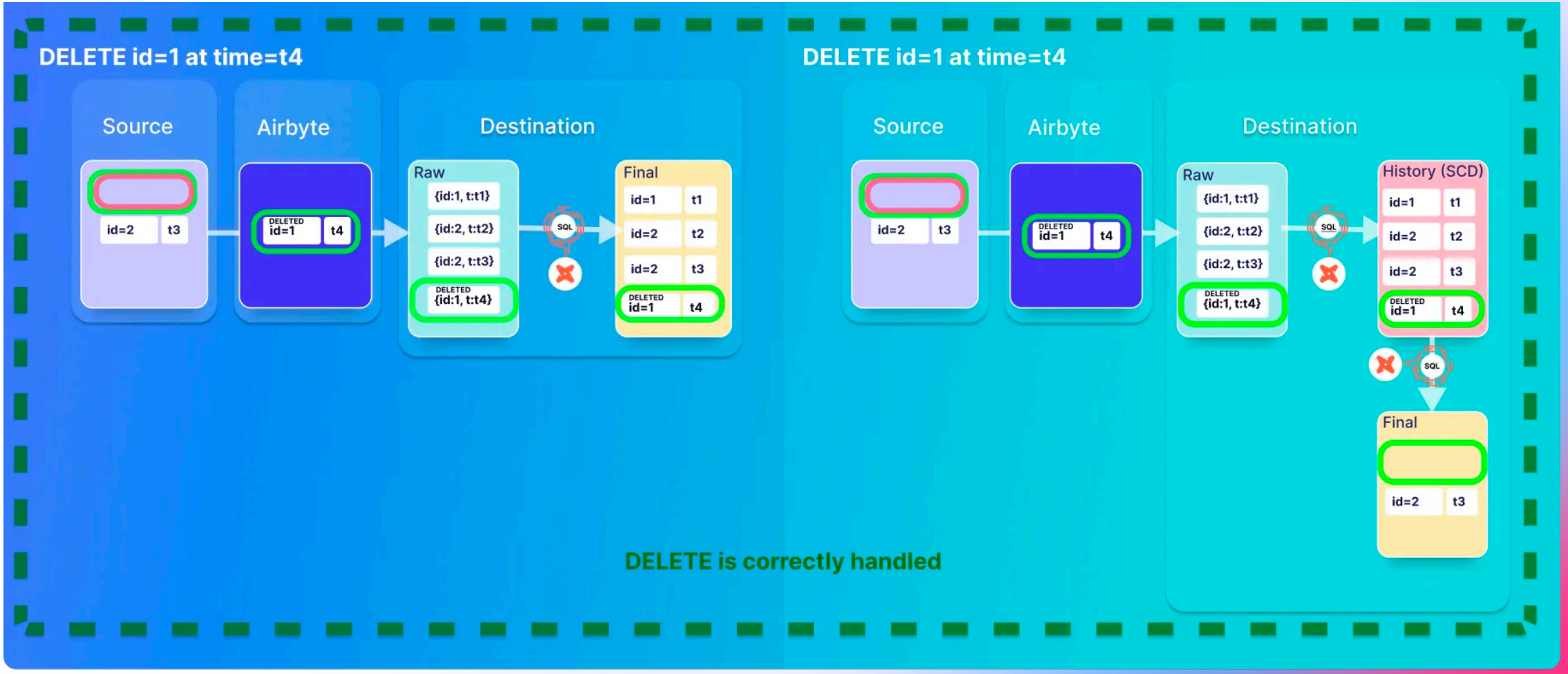
СБЕР ЗДОРОВЬЕ



*Postgres, MySQL, MSSQL и MongoDB

Airbyte. Sync Modes: CDC*

СБЕР ЗДОРОВЬЕ



*Postgres, MySQL, MSSQL и MongoDB

Airbyte. Sync Schedules

СБЕР ЗДОРОВЬЕ

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

*

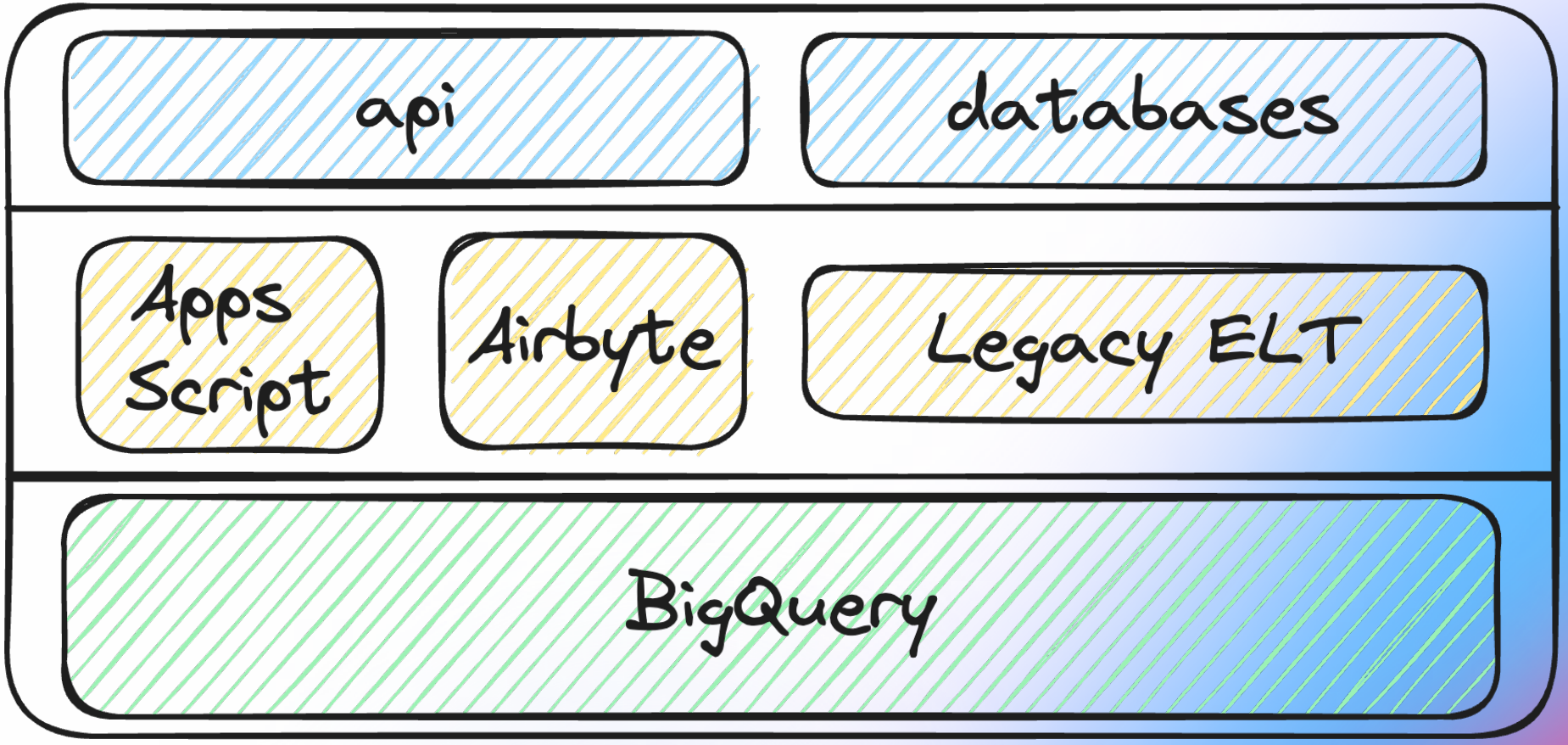
BACK TO end-2021. Что решили?

СБЕР ЗДОРОВЬЕ

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

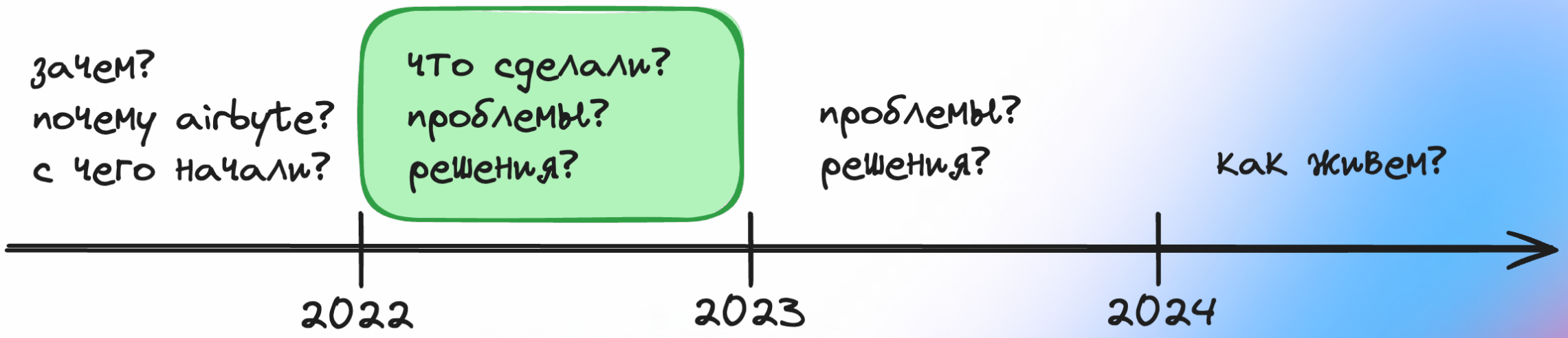
BACK TO end-2021.Что получили?

СБЕР ЗДОРОВЬЕ



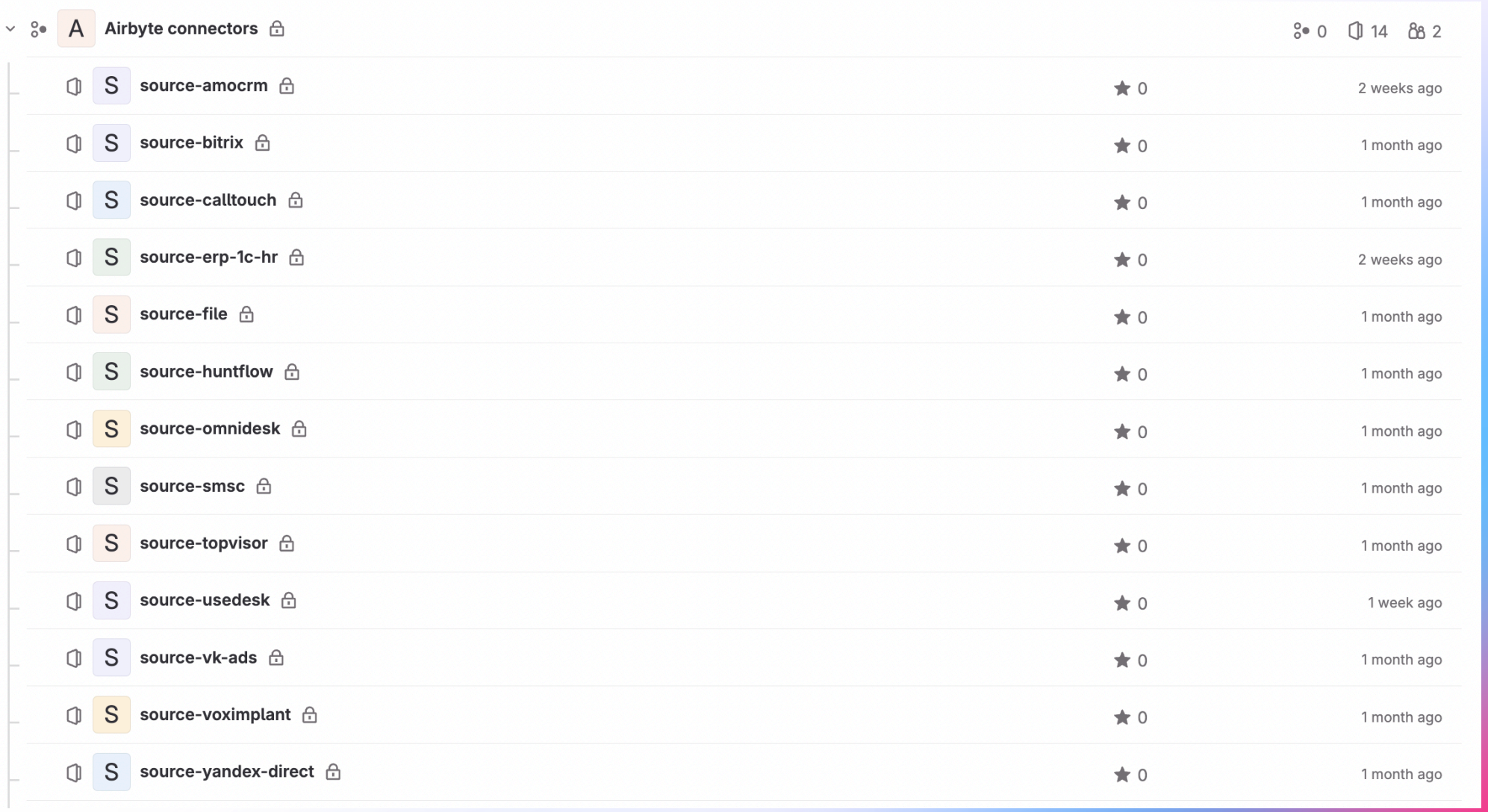
План

СБЕР ЗДОРОВЬЕ



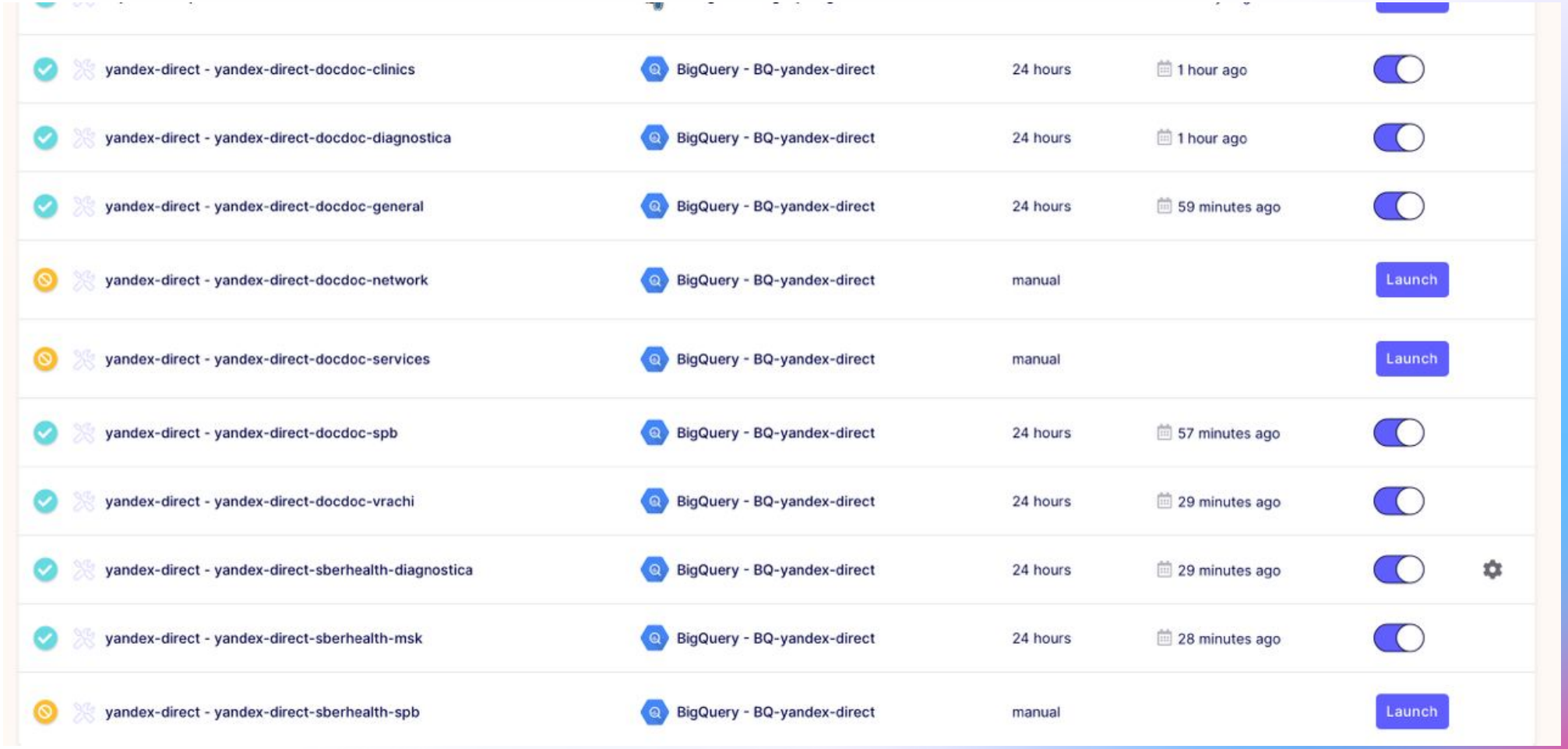
BACK TO begin-2022

СБЕР ЗДОРОВЬЕ



BACK TO begin-2022

СБЕР ЗДОРОВЬЕ

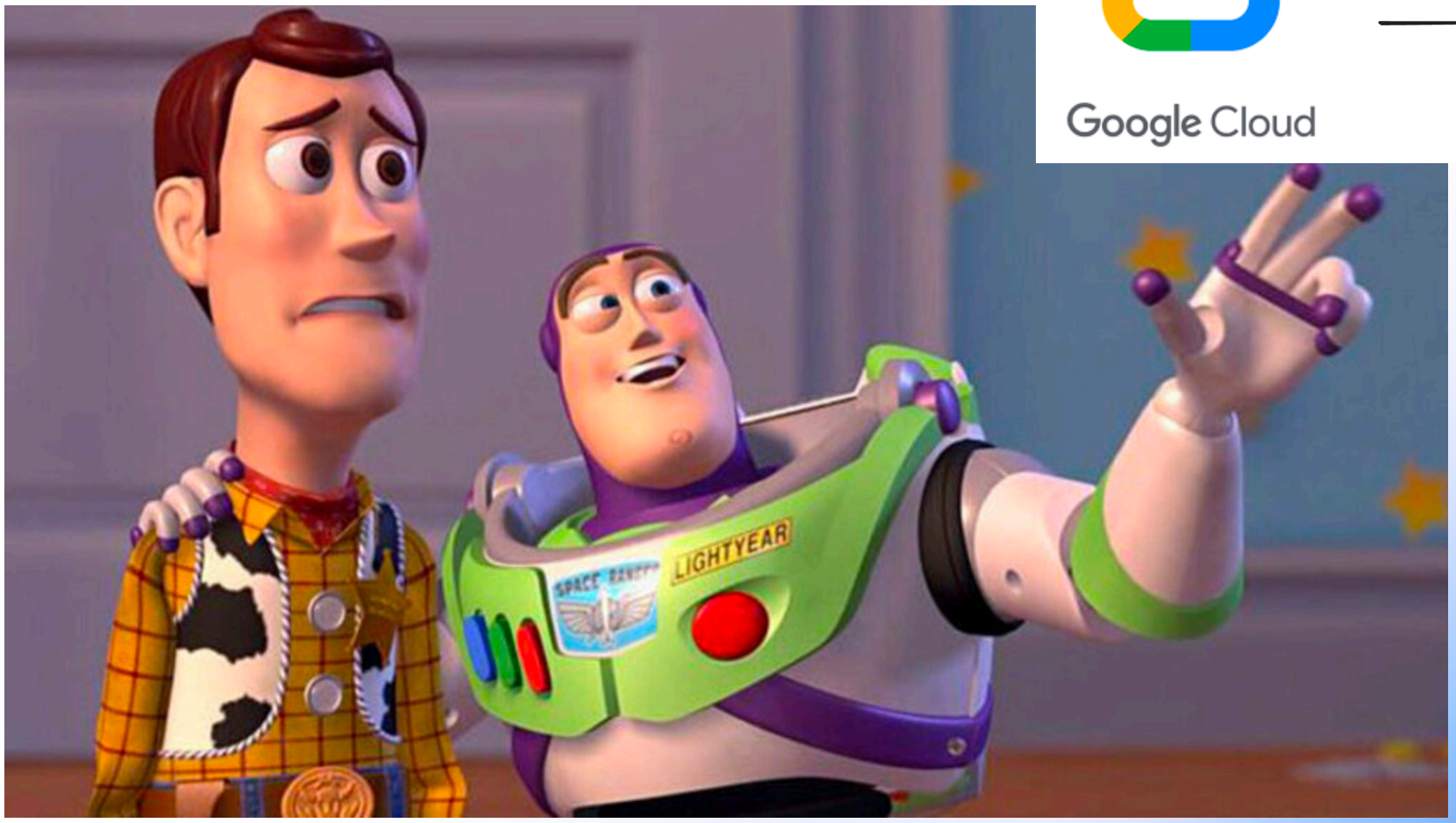


yandex-direct - yandex-direct-docdoc-clinics	BigQuery - BQ-yandex-direct	24 hours	1 hour ago	
yandex-direct - yandex-direct-docdoc-diagnostica	BigQuery - BQ-yandex-direct	24 hours	1 hour ago	
yandex-direct - yandex-direct-docdoc-general	BigQuery - BQ-yandex-direct	24 hours	59 minutes ago	
yandex-direct - yandex-direct-docdoc-network	BigQuery - BQ-yandex-direct	manual		Launch
yandex-direct - yandex-direct-docdoc-services	BigQuery - BQ-yandex-direct	manual		Launch
yandex-direct - yandex-direct-docdoc-spb	BigQuery - BQ-yandex-direct	24 hours	57 minutes ago	
yandex-direct - yandex-direct-docdoc-vrachi	BigQuery - BQ-yandex-direct	24 hours	29 minutes ago	
yandex-direct - yandex-direct-sberhealth-diagnostica	BigQuery - BQ-yandex-direct	24 hours	29 minutes ago	
yandex-direct - yandex-direct-sberhealth-msk	BigQuery - BQ-yandex-direct	24 hours	28 minutes ago	
yandex-direct - yandex-direct-sberhealth-spb	BigQuery - BQ-yandex-direct	manual		Launch

*скрин из прошлого

BACK TO 2022

СБЕР ЗДОРОВЬЕ



Google Cloud

миграция

Yandex Cloud

BACK TO 2022. Миграция

СБЕР ЗДОРОВЬЕ

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

BACK TO 2022

СБЕР ЗДОРОВЬЕ

Airbyte

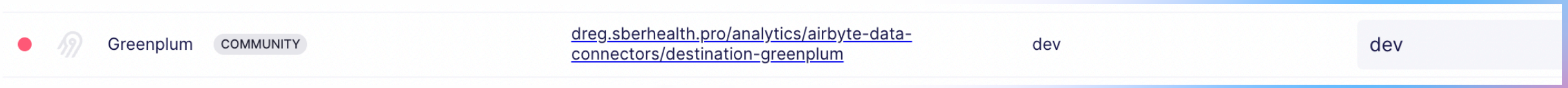
BACK TO 2022

СБЕР ЗДОРОВЬЕ

Проблема: Нет коннектора для Greenplum

Но есть коннектор для Postgres

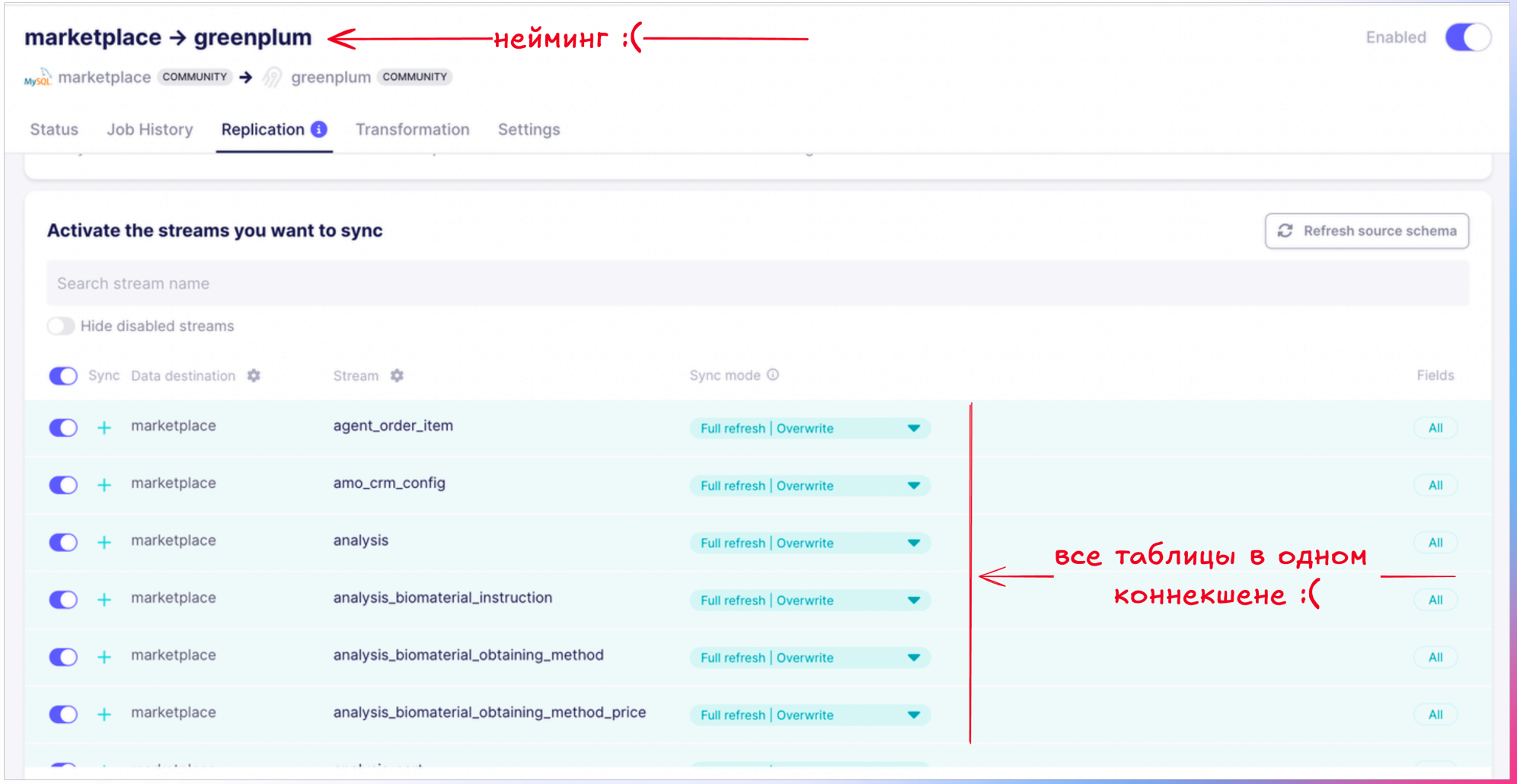
Решение: написать свой, переписав постгресовый :)



BACK TO 2022

СБЕР ЗДОРОВЬЕ

Проблема:



BACK TO 2022

СБЕР ЗДОРОВЬЕ

Решение:

- 1 таблица = 1 коннекшен

, но создавать руками = БОЛЬ 😵‍💫

+ хотели управлять конфигурациями как кодом

BACK TO 2022

СБЕР ЗДОРОВЬЕ

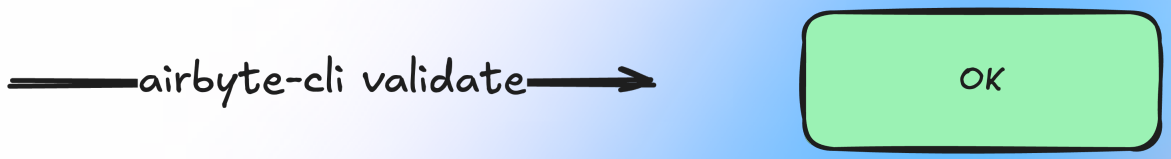
Что делать?

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

BACK TO 2022

СБЕР ЗДОРОВЬЕ

```
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
     ...
```



BACK TO 2022

СБЕР ЗДОРОВЬЕ

```
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
     ...
```

BACK TO 2022

СБЕР ЗДОРОВЬЕ

```
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
     ...
```

BACK TO 2022

СБЕР ЗДОРОВЬЕ

NAME	DESTINATION NAME	CONNECTOR	FREQUENCY	LAST SYNC	ENABLED
marketplace.timezone_list → greenplu...	greenplum	Greenplum	24 hours	2 hours ago	
marketplace.telecheckup_settings → ...	greenplum	Greenplum	24 hours	1 hour ago	
marketplace.subscription_pay_rule → ...	greenplum	Greenplum	24 hours	2 hours ago	
marketplace.specialization → greenpl...	greenplum	Greenplum	7 hours	5 hours ago	
marketplace.snils → greenplum	greenplum	Greenplum	24 hours	1 hour ago	
marketplace.sms_log → greenplum	greenplum	Greenplum	7 hours	4 hours ago	
marketplace.smartblock_smartcard_s...	greenplum	Greenplum	7 hours	4 hours ago	
marketplace.shop → greenplum	greenplum	Greenplum	24 hours	3 hours ago	

BACK TO 2022

СБЕР ЗДОРОВЬЕ

Проблема: Tinyint as bool

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

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

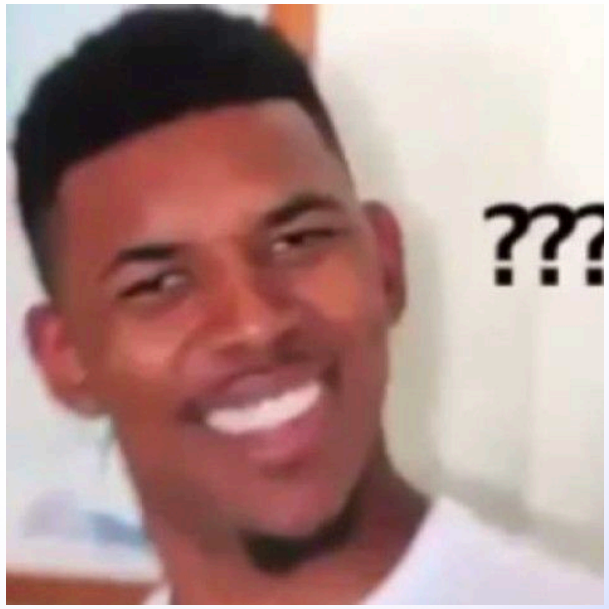
BACK TO 2022

СБЕР ЗДОРОВЬЕ

Проблема: Tinyint as bool

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

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



BACK TO 2022

СБЕР ЗДОРОВЬЕ

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`

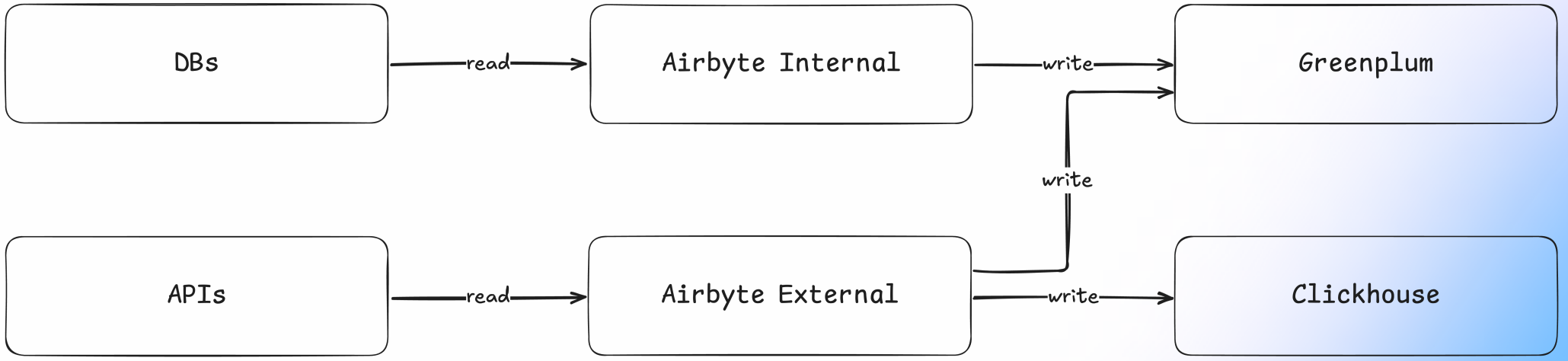
BACK TO 2022

СБЕР ЗДОРОВЬЕ

payment_type
1
2
3
4
5
null

BACK TO end-2022. Что получили?

СБЕР ЗДОРОВЬЕ



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

BACK TO 2022. Какие проблемы оставались?

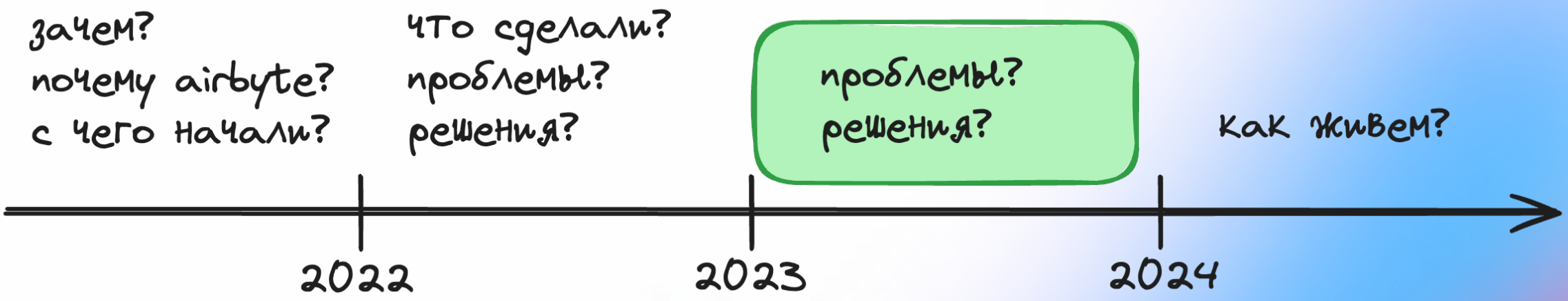
СБЕР ЗДОРОВЬЕ

- Не было расписания по крону
- Не было возможности исключать колонки из синхронизации

—> приходилось обновлять мету в базе :(

План

СБЕР ЗДОРОВЬЕ



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

СБЕР ЗДОРОВЬЕ

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

Решение: использовать CDC

BACK TO 2023

СБЕР ЗДОРОВЬЕ

```
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

СБЕР ЗДОРОВЬЕ

🎉 Support cron strings

Проблема: При ручных запусках расписание ломалось

Решение: перенос запусков на расписание по крону

BACK TO 2023

СБЕР ЗДОРОВЬЕ

🎉 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

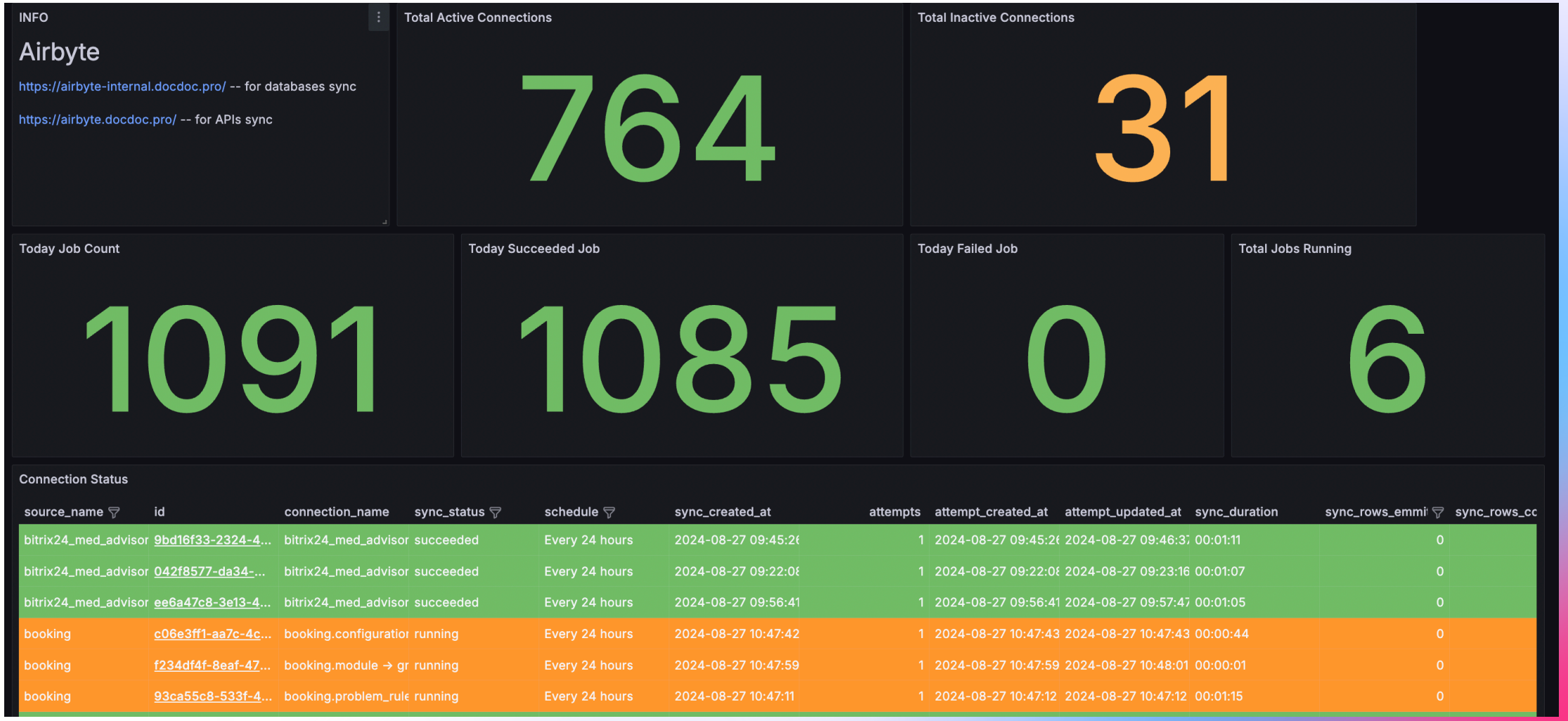
СБЕР ЗДОРОВЬЕ

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

Решение: собираем метрики через sql-exporter

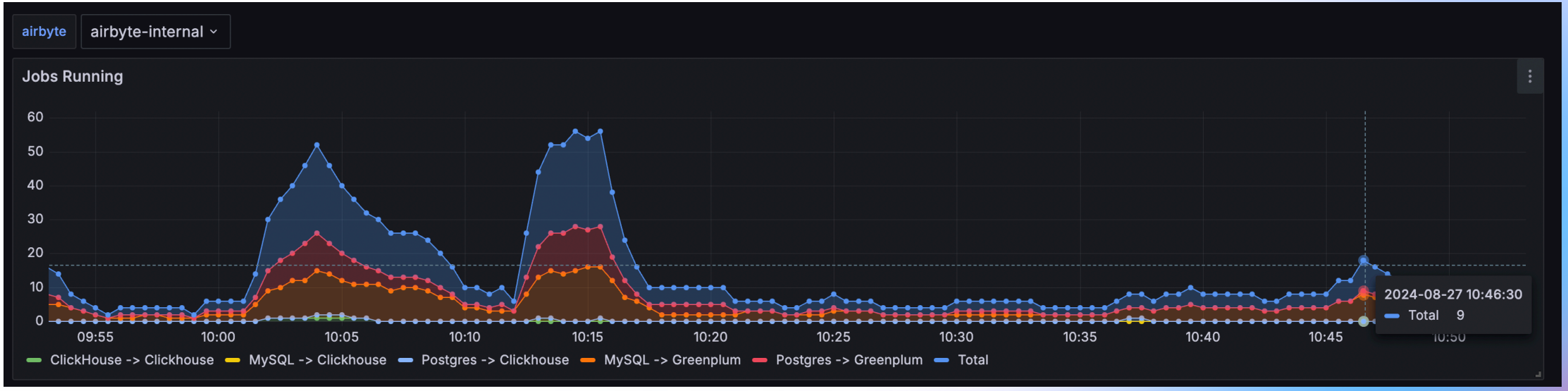
BACK TO 2023

СБЕР ЗДОРОВЬЕ



BACK TO 2023

СБЕР ЗДОРОВЬЕ



BACK TO 2023

СБЕР ЗДОРОВЬЕ

Проблема: каждый запуск требует минимум 3 пода для операций:

- orchestrator
- read
- write

-> большое количество «мертвых» подов от запусков.

Решение: cleanup-policy

BACK TO 2023

СБЕР ЗДОРОВЬЕ

```
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

СБЕР ЗДОРОВЬЕ

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

Решение: разнести статичные сервисы и синки на разные ноды

BACK TO 2023

СБЕР ЗДОРОВЬЕ

```
# static services
webapp:
  nodeSelector:
    type: static
```

✓airbyte-bootloader
✓server
✓temporal

```
# dynamic services
worker:
  nodeSelector:
    type: dynamic
```

✓orchestrator
✓workload-api-server
✓workload-launcher

BACK TO end-2023. Что получили?

СБЕР ЗДОРОВЬЕ

- 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

СБЕР ЗДОРОВЬЕ

Какие проблемы оставались?

- Ролевая модель (только в Cloud)
- Бэкфилл данных

План

СБЕР ЗДОРОВЬЕ

NOW (2024)

СБЕР ЗДОРОВЬЕ

- Проблем пока не было 😎

  Как живем?
- Синки из баз переехали в Debezium

  Почему?
- Целевое решение
- Гибкость в управлении коннекторами

NOW (2024)

СБЕР ЗДОРОВЬЕ

- Проблем пока не было 😎

  Как живем?

- Продолжаем писать коннекторы до API

NOW (2024)

СБЕР ЗДОРОВЬЕ

Что еще сделали?

- Перенесли запуск коннекшенов в Dagster

Почему?

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

Выводы

СБЕР ЗДОРОВЬЕ

🔥 Python CDK для API

👍 Для баз, если:

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

🤔 Если синхронизаций >1000

СБЕР ЗДОРОВЬЕ

Вопросы?