Escuela de Datos Vivos

Curso Data Engineer: Creando un pipeline de datos

Módulo E - Clase 9

Agenda

- Airflow Xcom
- Streaming process
- Ejercitación

Airflow - Xcom

Airflow - xcom

XCom significa “cross-communication” y permite intercambiar mensajes o **pequeñas cantidades** de datos entre tareas.

List XComs

Search

Actions

Record Count: 14

	Key	Value	Timestamp	Dag Id	Task Id	Run Id	Map Index
	return_value	23/04/30 10:24:02 INFO ShutdownHookManager: Deleting directory /tmp/spark-34a8ac70-1aab-4ed5-ad2a-eafda7c107b9	2023-04-30, 13:24:02	ingest-transform	transform	scheduled__2023-04-29T00:00:00+00:00	
	return_value	rm: `/ingest/*.*': No such file or directory	2023-04-30, 13:23:09	ingest-transform	ingest	scheduled__2023-04-29T00:00:00+00:00	
	model_accuracy	5.9778372006977385	2023-04-30, 12:58:06	model_trining	training_model_C	scheduled__2023-04-29T00:00:00+00:00	
	model_accuracy	4.348391678519117	2023-04-30, 12:58:00	model_trining	training_model_B	scheduled__2023-04-29T00:00:00+00:00	

Airflow

	Key	Value	Timestamp	Dag Id	Task Id	Run Id	Map Index
	return_value	23/04/30 10:24:02 INFO ShutdownHookManager: Deleting directory /tmp/spark-34a8ac70-1aab-4ed5-ad2a-eafda7c107b9	2023-04-30, 13:24:02	ingest-transform	transform	scheduled__2023-04-29T00:00:00+00:00	

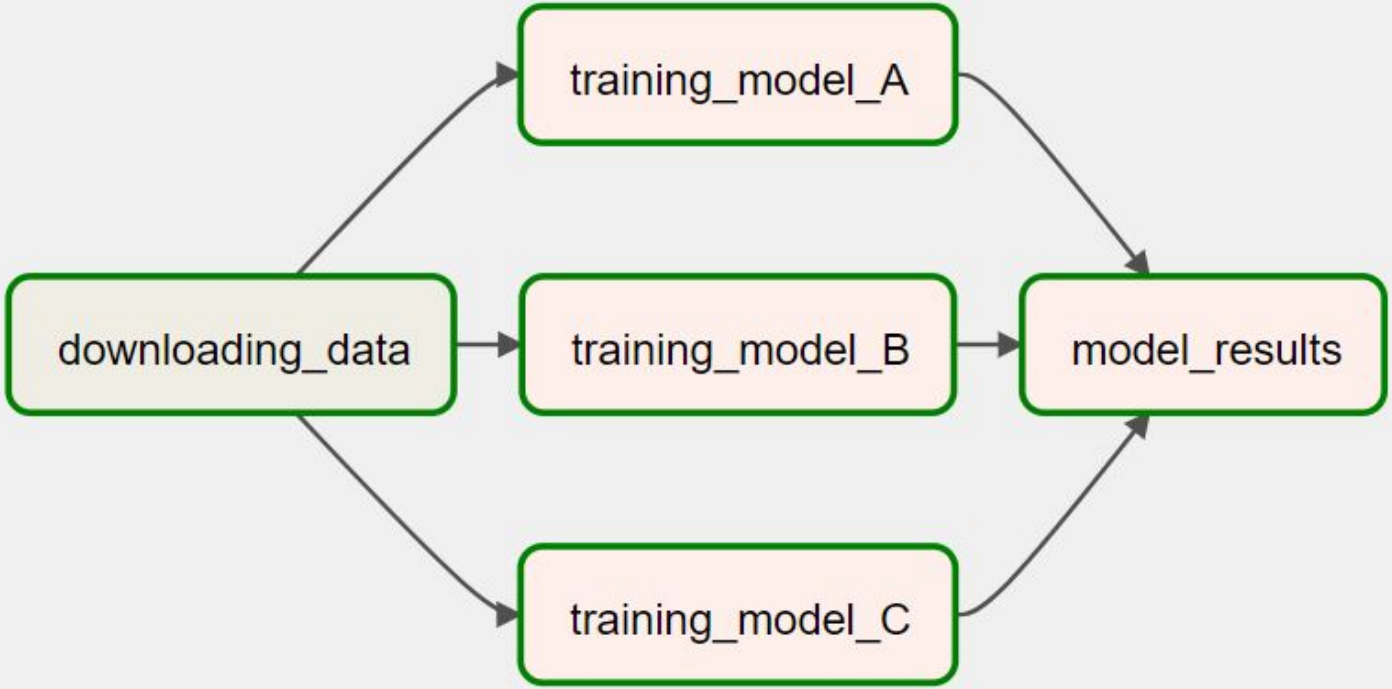
- **Key** es el identificador del XCom.
- **Value** es el valor del XCom.
- **Timestamp** es cuando el XCom es creado.

Airflow

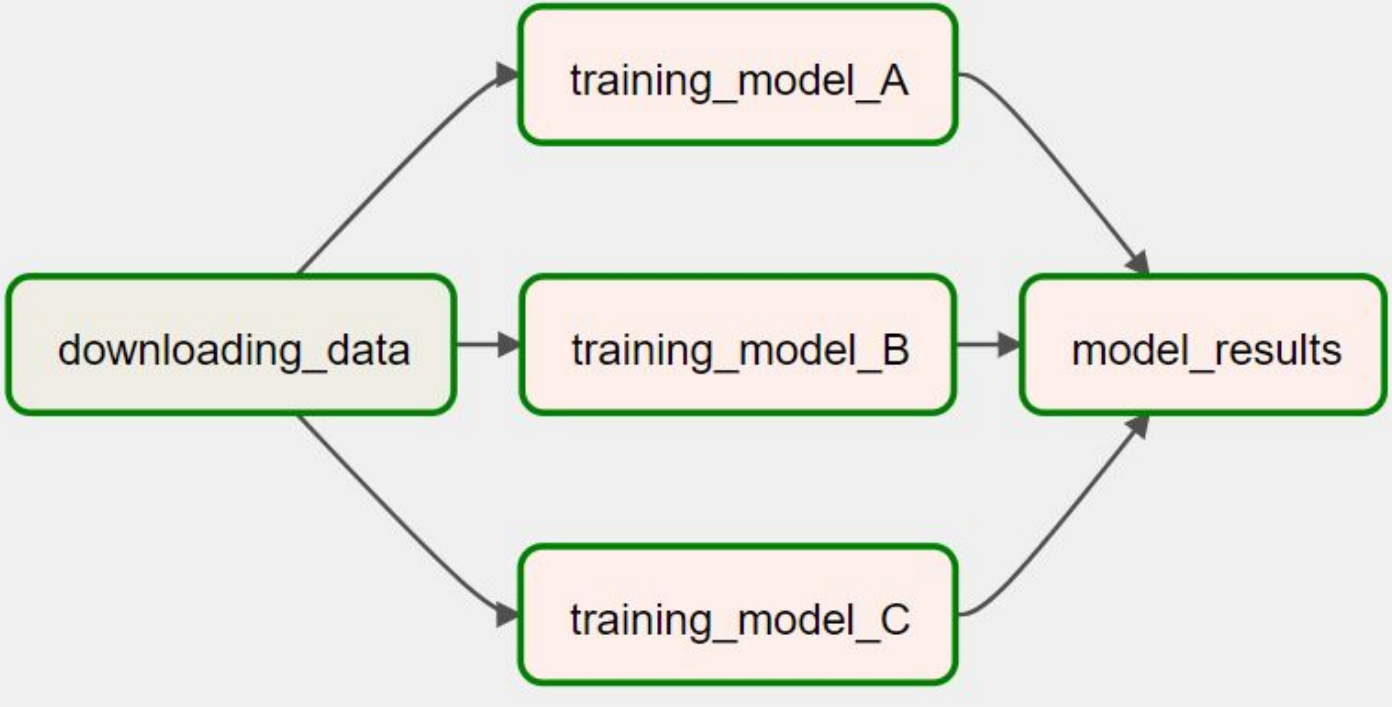
	Key	Value	Timestamp	Dag Id	Task Id	Run Id	Map Index
	return_value	23/04/30 10:24:02 INFO ShutdownHookManager: Deleting directory /tmp/spark-34a8ac70-1aab-4ed5-ad2a-eafda7c107b9	2023-04-30, 13:24:02	ingest-transform	transform	scheduled__2023-04-29T00:00:00+00:00	

- **Dag id** del dag donde el XCom fue creado.
- **Task id** de la tarea donde el XCom fue creado.
- **Run id** corresponde a la fecha de ejecución del DagRun donde se generó el XCom.

Airflow - xcom

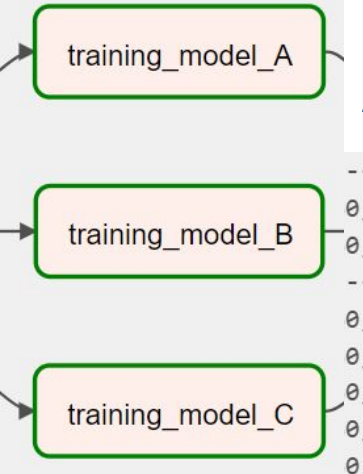


Airflow - model_training_1



Airflow - model_training_1

```python
from airflow import DAG
from airflow.operators.bash import BashOperator
from airflow.operators.python import PythonOperator
from random import uniform
from datetime import datetime

default_args = {
'start_date': datetime(2023, 1, 1)
}

def _training_model(ti):
    accuracy = uniform(0.1, 10.0)
    print(f'model\'s accuracy: {accuracy}')
    return accuracy


def _model_results(ti):
    print(f'model results')


with DAG('model_training_1', schedule_interval='@daily', default_args=default_args, catchup=False) as dag:
    downloading_data = BashOperator(
        task_id='downloading_data',
        bash_command='sleep 20'
        )

```



Airflow - model_training_1

```
--------------------------------------------------------------------------------
0, 15:30:46 UTC] {taskinstance.py:1357} INFO - Starting attempt 2 of 2
0, 15:30:46 UTC] {taskinstance.py:1358} INFO -
--------------------------------------------------------------------------------
0, 15:30:46 UTC] {taskinstance.py:1377} INFO - Executing <Task(PythonOperator): training_model_A
0, 15:30:46 UTC] {standard_task_runner.py:52} INFO - Started process 2853 to run task
0, 15:30:46 UTC] {standard_task_runner.py:79} INFO - Running: ['airflow', 'tasks', 'run', 'model
0, 15:30:46 UTC] {standard_task_runner.py:80} INFO - Job 137: Subtask training_model_A
0, 15:30:47 UTC] {task_command.py:369} INFO - Running <TaskInstance: model_training_1.training_m
[2023-04-30, 15:30:47 UTC] {taskinstance.py:1569} INFO - Exporting the following env vars:
AIRFLOW_CTX_DAG_OWNER=airflow
AIRFLOW_CTX_DAG_ID=model_training_1
AIRFLOW_CTX_TASK_ID=training_model_A
AIRFLOW_CTX_EXECUTION_DATE=2023-04-29T00:00:00+00:00
AIRFLOW_CTX_TRY_NUMBER=2
AIRFLOW_CTX_DAG_RUN_ID=scheduled__2023-04-29T00:00:00+00:00
[2023-04-30, 15:30:47 UTC] {logging_mixin.py:115} INFO - model's accuracy: 5.891220799941742
[2023-04-30, 15:30:47 UTC] {python.py:173} INFO - Done. Returned value was: 5.891220799941742
[2023-04-30, 15:30:47 UTC] {taskinstance.py:1395} INFO - Marking task as SUCCESS. dag_id=model_training_1,
[2023-04-30, 15:30:47 UTC] {local_task_job.py:156} INFO - Task exited with return code 0
[2023-04-30, 15:30:47 UTC] {local_task_job.py:273} INFO - 0 downstream tasks scheduled from follow-on sche
```

Airflow - model_training_1

List XComs

Search

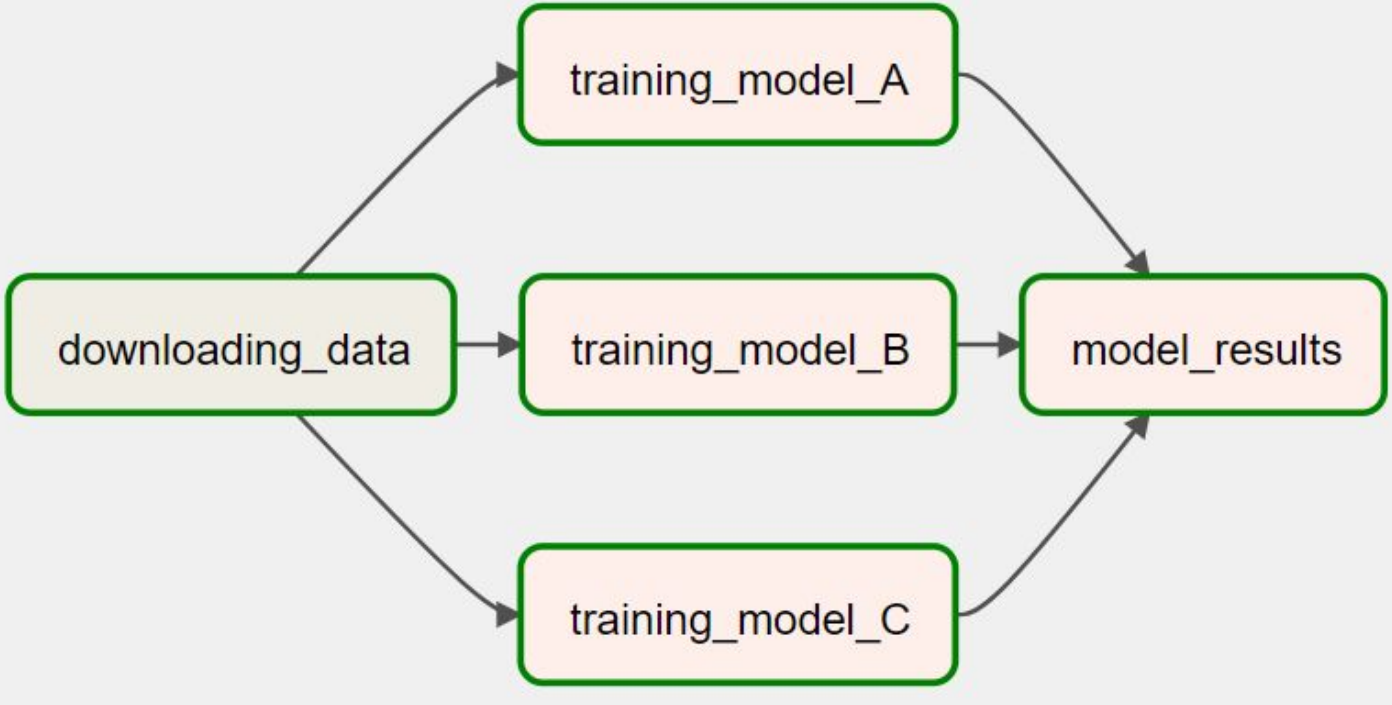
Actions

Record Count: 4

	Key	Value	Timestamp	Dag Id	Task Id	Run Id	Map Index
	return_value	1.6117023999412345	2023-04-30, 19:14:34	model_training_1	training_model_C	scheduled__2023-04-29T00:00:00+00:00	
	return_value	9.32002554177192	2023-04-30, 19:14:31	model_training_1	training_model_B	scheduled__2023-04-29T00:00:00+00:00	
	return_value	0.6933815602995332	2023-04-30, 19:14:29	model_training_1	training_model_A	scheduled__2023-04-29T00:00:00+00:00	
	return_value		2023-04-30, 19:14:27	model_training_1	downloading_data	scheduled__2023-04-29T00:00:00+00:00	

Nos queda un return_value vacío ???

Airflow - model_training_2



Airflow - model_training_2

```python
from airflow import DAG
from airflow.operators.bash import BashOperator
from airflow.operators.python import PythonOperator
from random import uniform
from datetime import datetime

default_args = {
'start_date': datetime(2023, 1, 1)
}

def _training_model(ti):
    accuracy = uniform(0.1, 10.0)
    print(f'model\'s accuracy: {accuracy}')
    return accuracy

def _model_results(ti):
    print(f'model results')

with DAG('model_training_2', schedule_interval='@daily', default_args=default_args, catchup=False) as dag:
    downloading_data = BashOperator(
        task_id='downloading_data',
        bash_command='sleep 20',
        do_xcom_push=False
        )

```

Airflow - model_training_2

List XComs

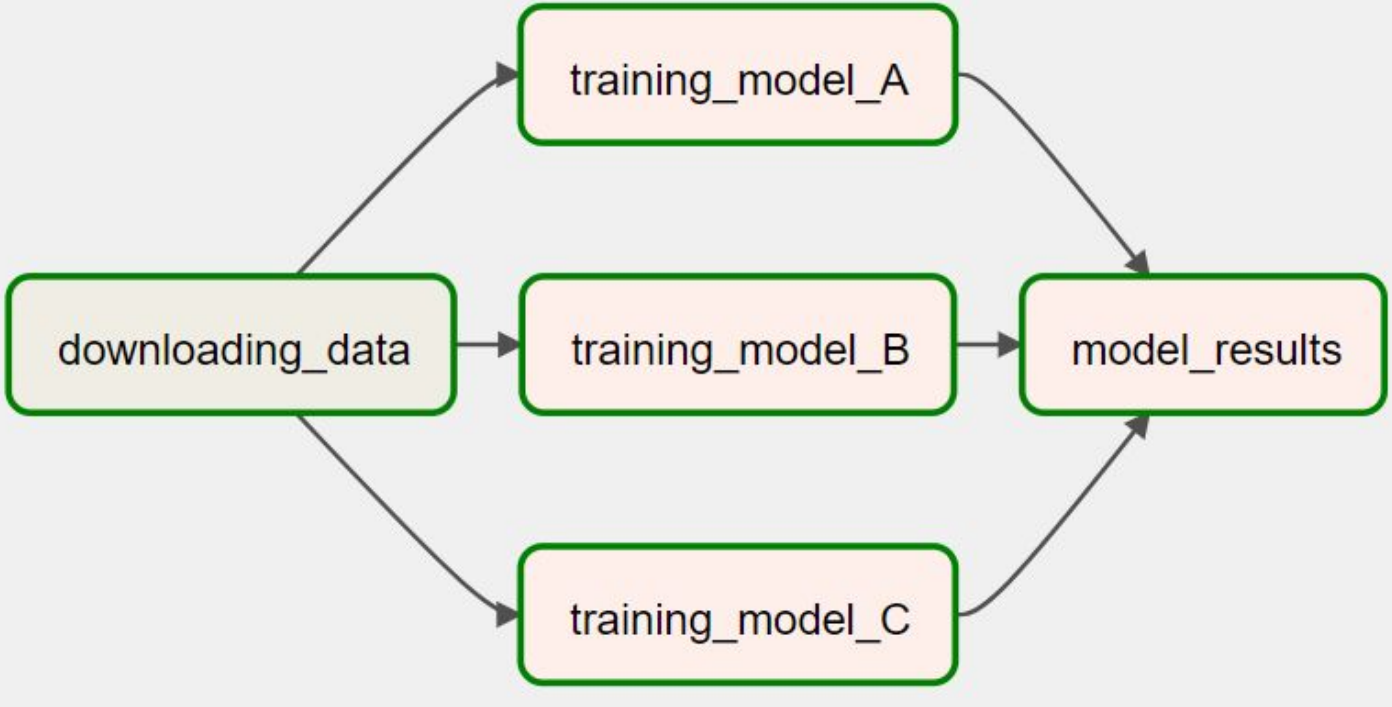
Search

Actions

Record Count: 3

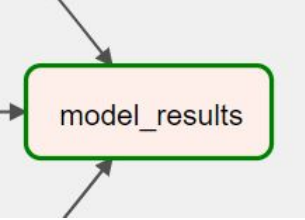
	Key	Value	Timestamp	Dag Id	Task Id	Run Id	Map Index
	return_value	1.1863582650140008	2023-04-30, 19:26:34	model_training_2	training_model_C	scheduled__2023-04-29T00:00:00+00:00	
	return_value	8.885779668764442	2023-04-30, 19:26:32	model_training_2	training_model_B	scheduled__2023-04-29T00:00:00+00:00	
	return_value	8.876635890055711	2023-04-30, 19:26:30	model_training_2	training_model_A	scheduled__2023-04-29T00:00:00+00:00	

Airflow - model_training_3



Airflow - model_training_3

```python
from airflow import DAG
from airflow.operators.bash import BashOperator
from airflow.operators.python import PythonOperator
from random import uniform
from datetime import datetime

default_args = {
'start_date': datetime(2023, 1, 1)
}

def _training_model(ti):
    accuracy = uniform(0.1, 10.0)
    print(f'model\'s accuracy: {accuracy}')
    ti.xcom_push(key='model_accuracy', value=accuracy)

def _model_results(ti):
    fetched_accuracies = ti.xcom_pull(key='model_accuracy', task_ids=['training_model_A', 'training_model_B', 'training_model_C'])
    fetched_accuracies.sort()
    print(f'Best accuracy: {fetched_accuracies[-1]}')

with DAG('model_training_3', schedule_interval='@daily', default_args=default_args, catchup=False) as dag:
    downloading_data = BashOperator(
        task_id='downloading_data',
        bash_command='sleep 20',
        do_xcom_push=False
        )
```



Airflow - model_training_3

```
--------------------------------------------------------------------------------
[2023-04-30, 16:30:46 UTC] {taskinstance.py:1357} INFO - Starting attempt 2 of 2
[2023-04-30, 16:30:46 UTC] {taskinstance.py:1358} INFO -
--------------------------------------------------------------------------------
[2023-04-30, 16:30:46 UTC] {taskinstance.py:1377} INFO - Executing <Task(PythonOperator): model_results> on 2023-04-29 00:0
[2023-04-30, 16:30:46 UTC] {standard_task_runner.py:52} INFO - Started process 3812 to run task
[2023-04-30, 16:30:46 UTC] {standard_task_runner.py:79} INFO - Running: ['airflow', 'tasks', 'run', 'model_training_3', 'mo
[2023-04-30, 16:30:46 UTC] {standard_task_runner.py:80} INFO - Job 170: Subtask model_results
[2023-04-30, 16:30:47 UTC] {task_command.py:369} INFO - Running <TaskInstance: model_training_3.model_results scheduled__20
[2023-04-30, 16:30:47 UTC] {taskinstance.py:1569} INFO - Exporting the following env vars:
AIRFLOW_CTX_DAG_OWNER=airflow
AIRFLOW_CTX_DAG_ID=model_training_3
AIRFLOW_CTX_TASK_ID=model_results
AIRFLOW_CTX_EXECUTION_DATE=2023-04-29T00:00:00+00:00
AIRFLOW_CTX_TRY_NUMBER=2
AIRFLOW_CTX_DAG_RUN_ID=scheduled__2023-04-29T00:00:00+00:00
[2023-04-30, 16:30:47 UTC] {logging_mixin.py:115} INFO - Best accuracy: 7.817875987181118
[2023-04-30, 16:30:47 UTC] {python.py:173} INFO - Done. Returned value was: None
[2023-04-30, 16:30:47 UTC] {taskinstance.py:1395} INFO - Marking task as SUCCESS. dag_id=model_training_3, task_id=model_re
[2023-04-30, 16:30:47 UTC] {local_task_job.py:156} INFO - Task exited with return code 0
[2023-04-30, 16:30:47 UTC] {local_task_job.py:273} INFO - 0 downstream tasks scheduled from follow-on schedule check
```

Airflow - model_training_3

List XComs

Search

Actions

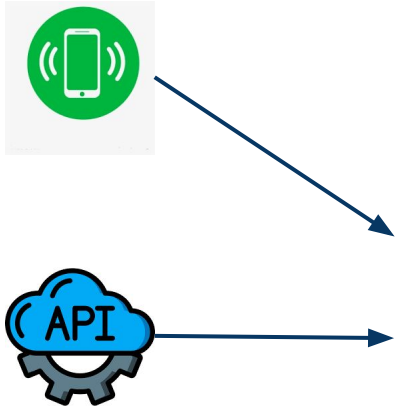
Record Count: 3

	Key	Value	Timestamp	Dag Id	Task Id	Run Id	Map Index
	model_accuracy	3.6228667994855086	2023-04-30, 19:30:44	model_training_3	training_model_C	scheduled__2023-04-29T00:00:00+00:00	
	model_accuracy	7.817875987181118	2023-04-30, 19:30:42	model_training_3	training_model_B	scheduled__2023-04-29T00:00:00+00:00	
	model_accuracy	3.9822105689063956	2023-04-30, 19:30:40	model_training_3	training_model_A	scheduled__2023-04-29T00:00:00+00:00	

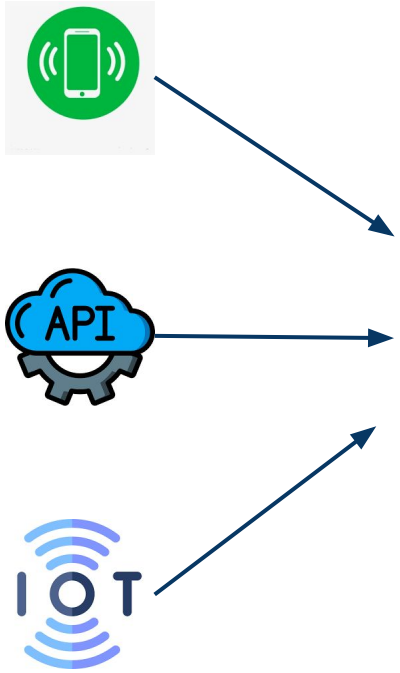
Streaming processing

STREAMING PROCESSING

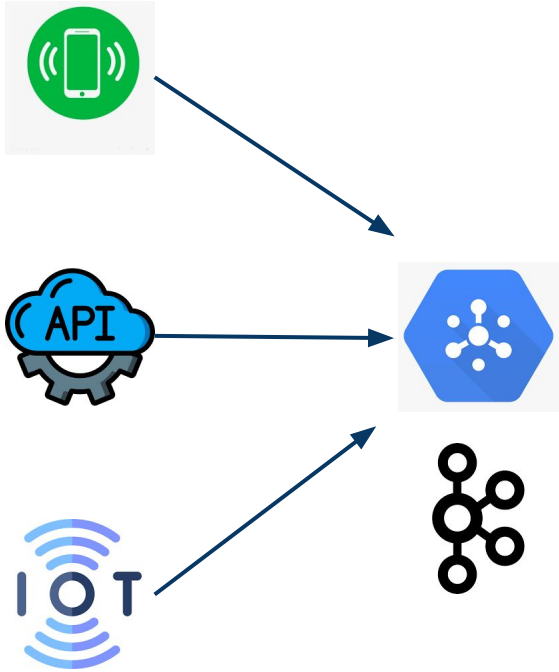
STREAMING PROCESSING



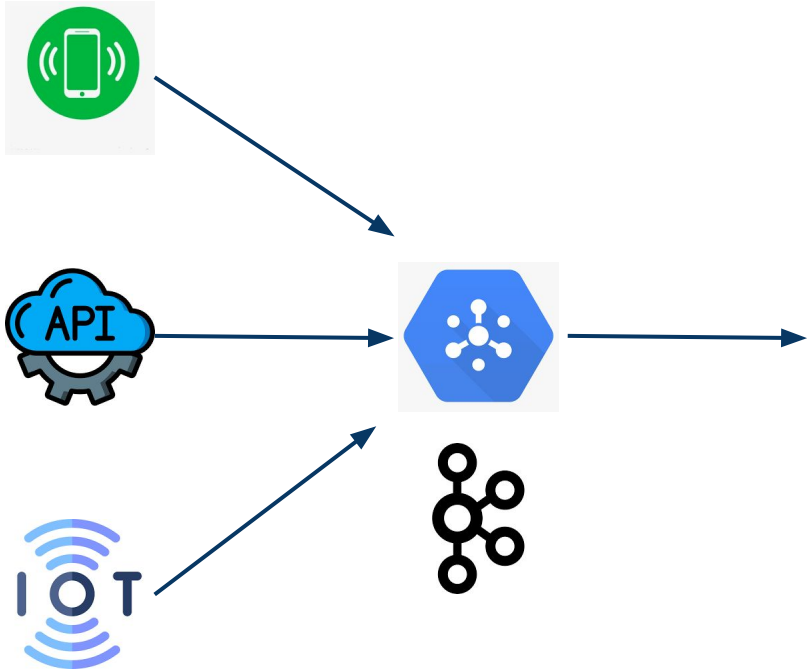
STREAMING PROCESSING



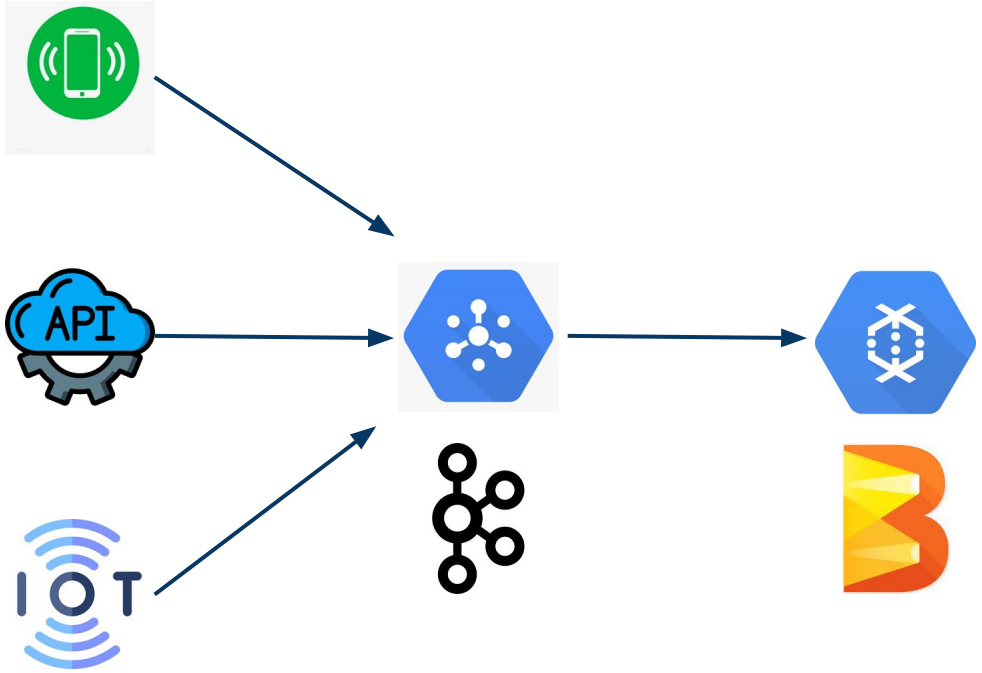
STREAMING PROCESSING



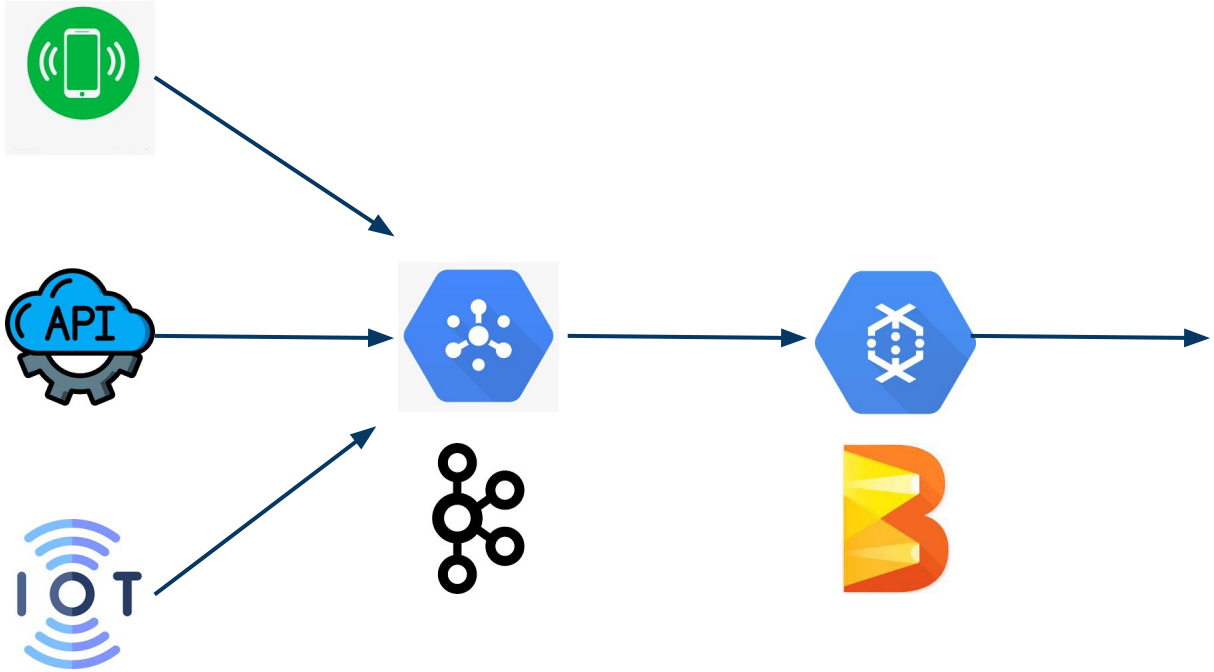
STREAMING PROCESSING



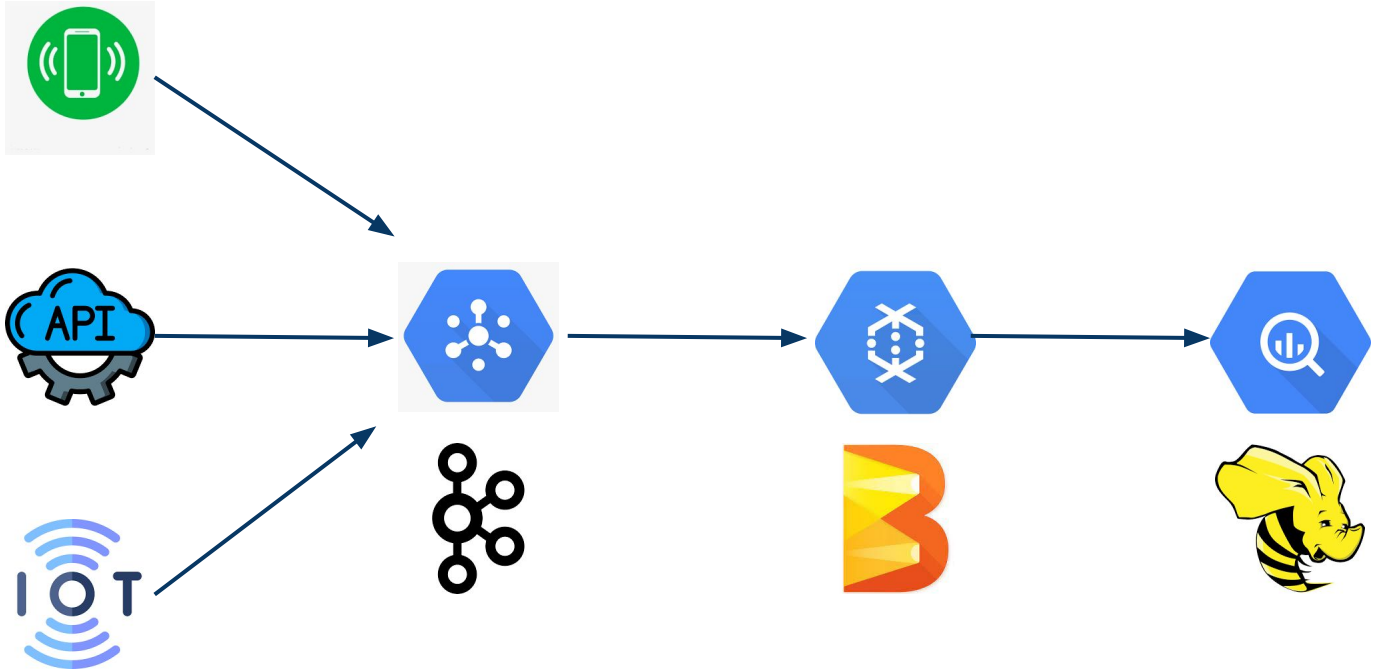
STREAMING PROCESSING



STREAMING PROCESSING



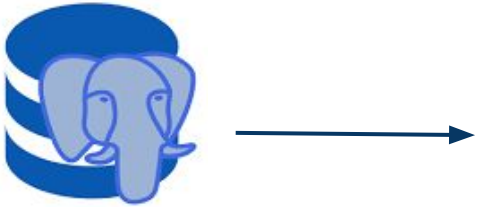
STREAMING PROCESSING



Ejercicio

Ejercicio

Ejercicio



Ejercicio

Ejercicio

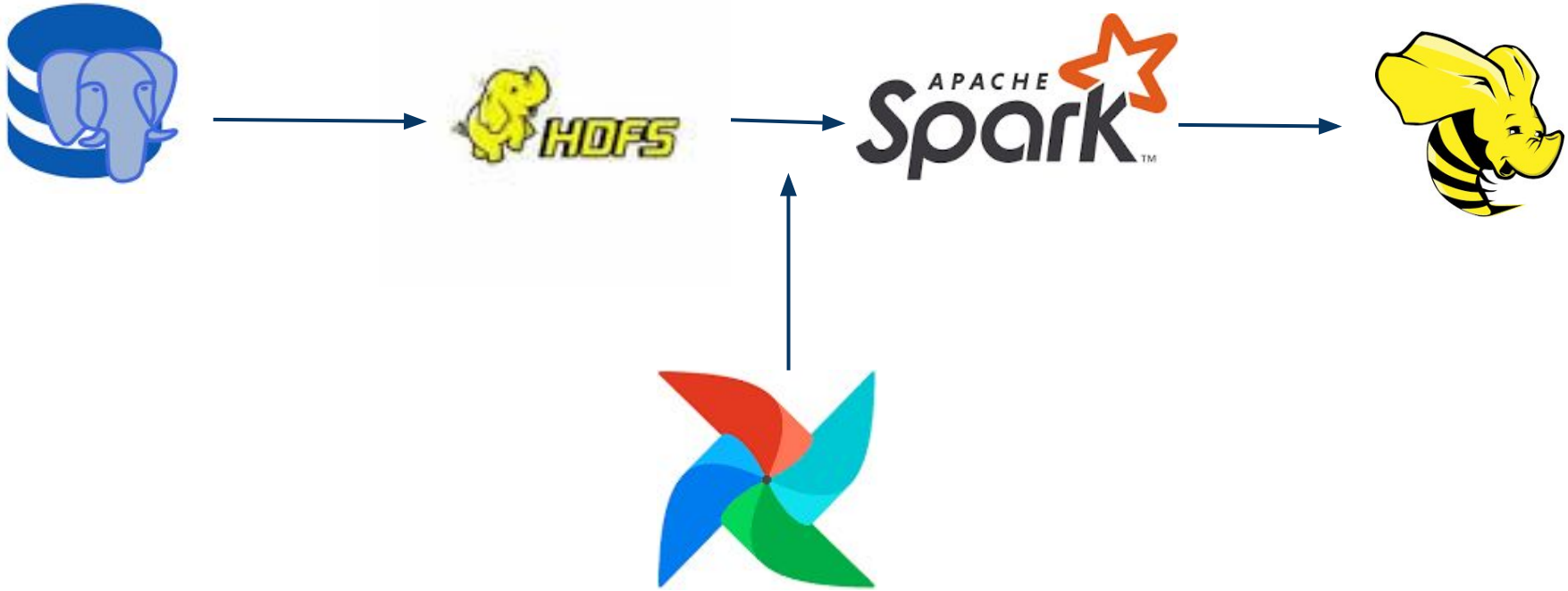
Ejercicio

Ejercicio

Ejercicio

Ejercicio

Ejercicio



Preguntas ?

Gracias