Оркестриране обработката на големи данни с Apache Airflow

OpenFest 2021

Кирил Митов

Кирил Митов

Chief Technical Office, BeMe.ai

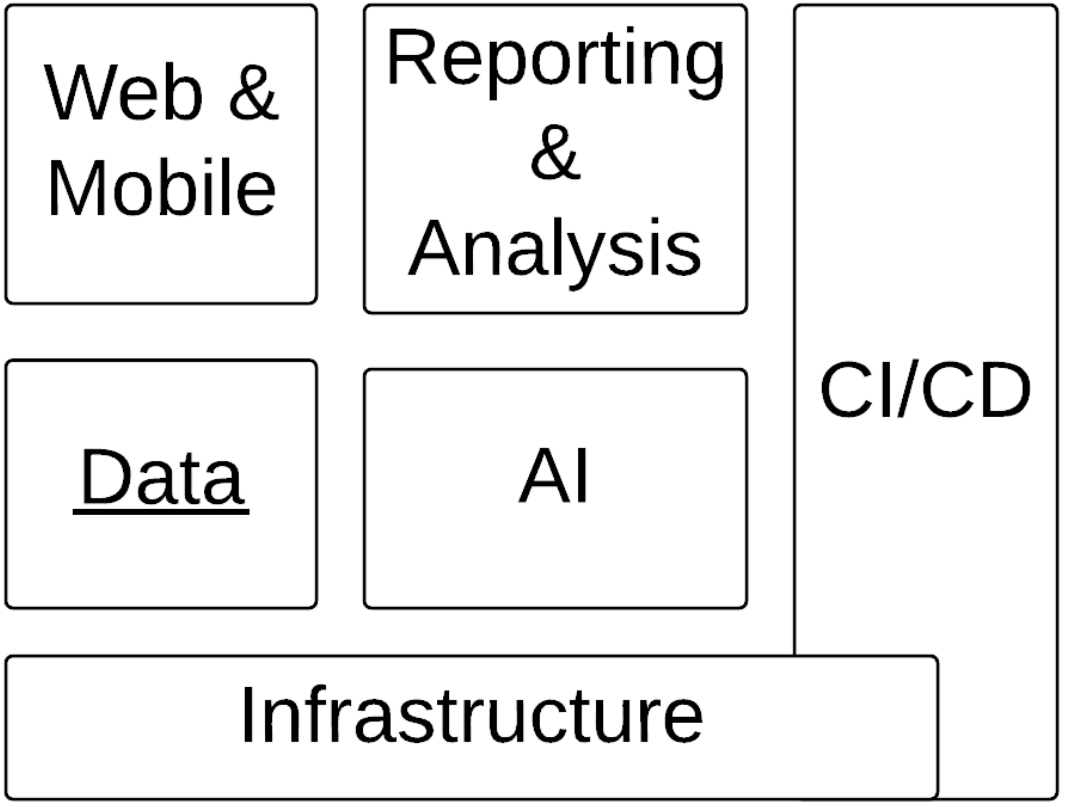
thebravoman (github, twitter)
kmitov.com
linkedin.com/in/kirilmitov/

FLLCasts, Robopartans, TUES

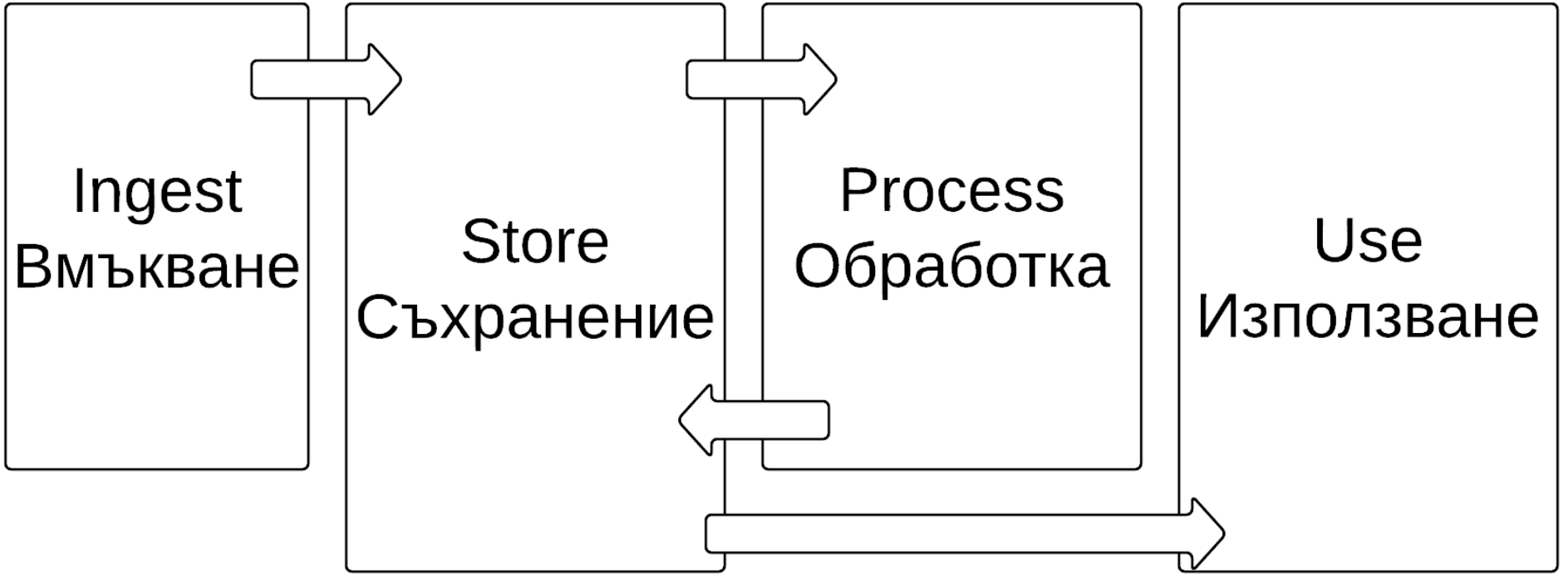
Apache Airflow

Rails, 3D, Google Closure Compiler

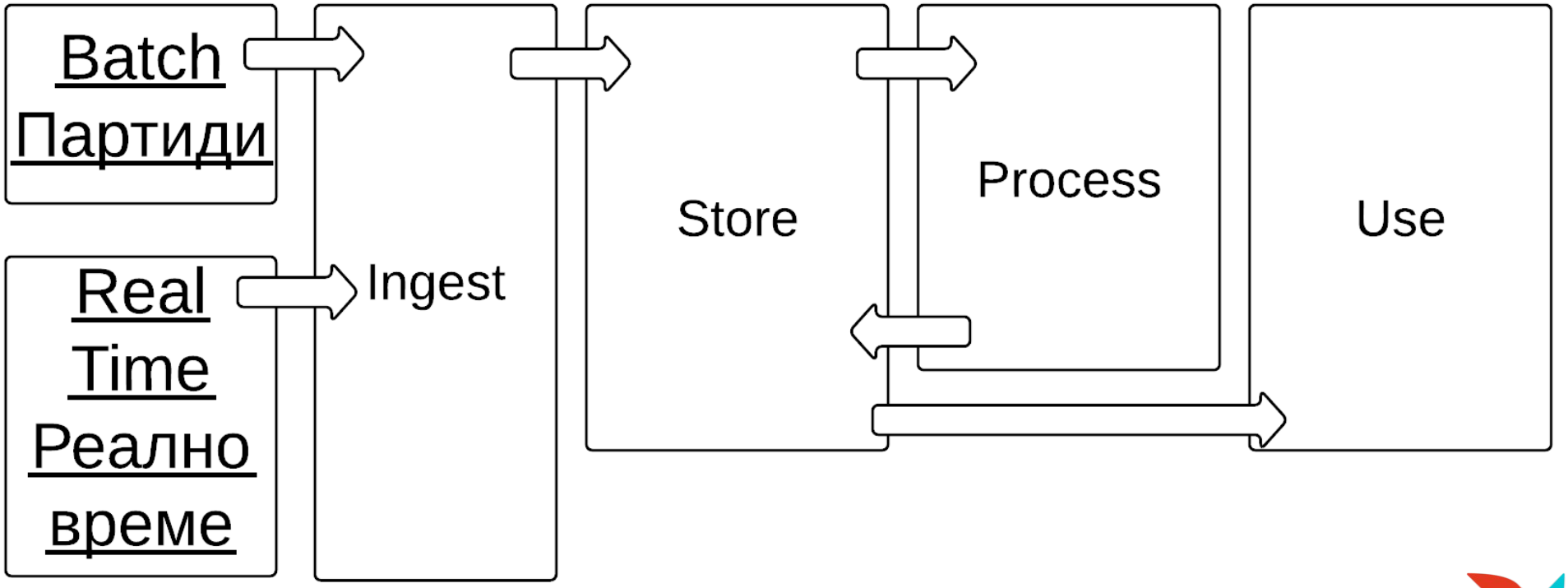
Стекът


Web &
Mobile
Reporting
&
Analysis
CI/CD
Data
AI
Infrastructure

Платформа за данни



“Партиди” и “в реално време”



Data{X}

Data{Base} – $0.5/GB

Data{Warehouse} – $0.2/GB

Data{Lake} – $0.02/GB

Големи данни (BigData)

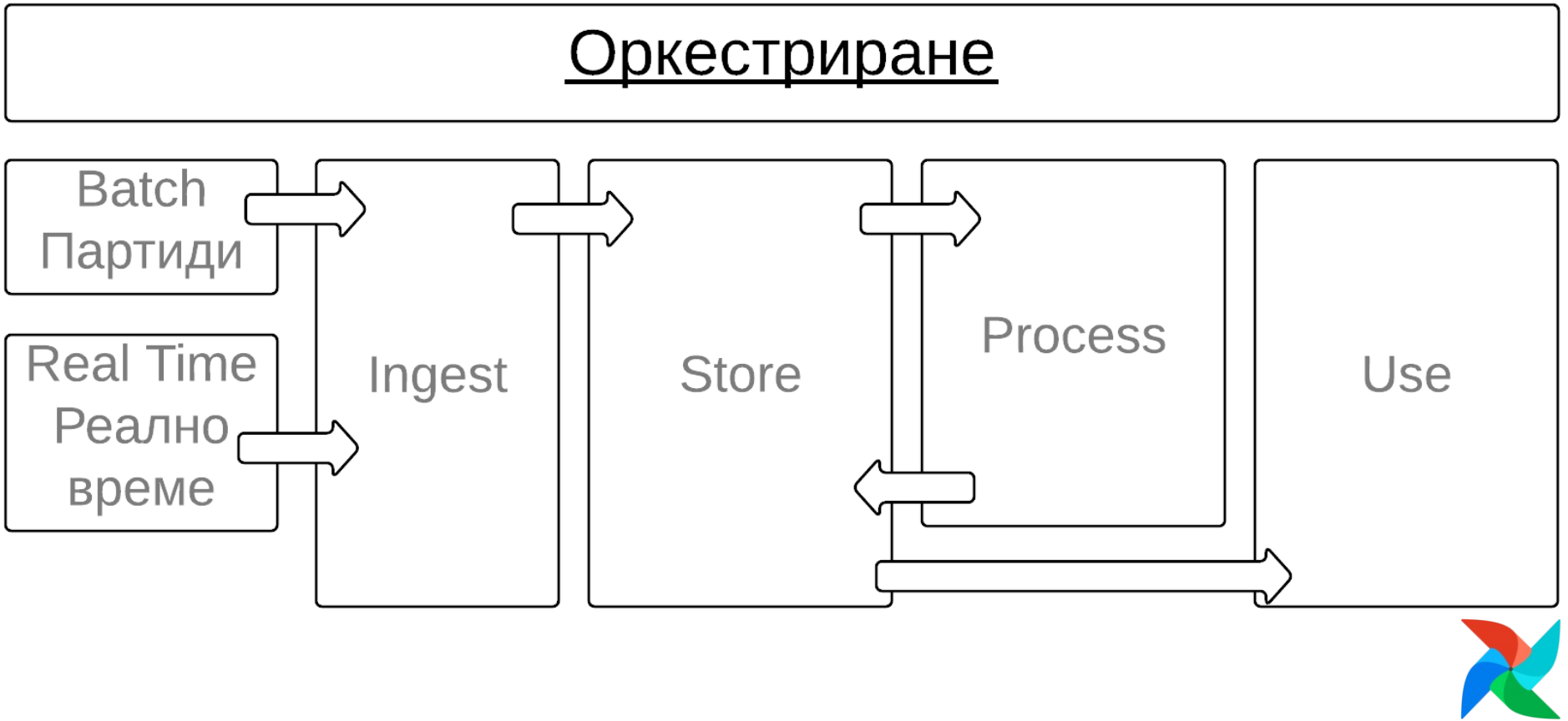
Трудно е да ги преместиш

Разнообразни схеми

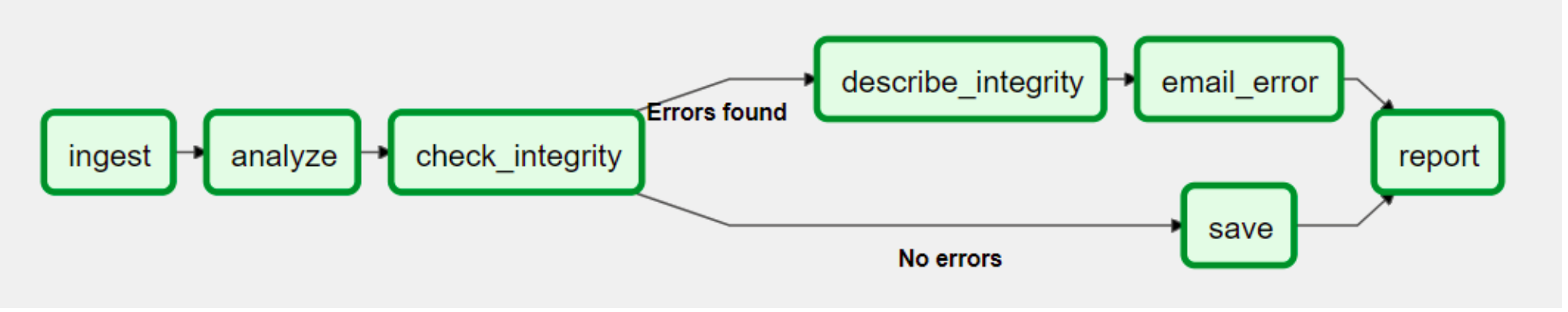
Не знаеш къде е ценното

Кой е клиентът?

Оркестриране



Directed Acyclic Graph



Apache Airflow

```
task1 = BashOperator(…)
task2 = PythonOperator(…)
task3_1 = {Custom}Operator(...)
task3_2 = DummyOperator(...)
task1 >> task2 >> [task3_1, task3_2]
```

Разклонения и условия

```
task1 >> [task2, task2_1]
```

```
[task1, task1_2] >> task3
```

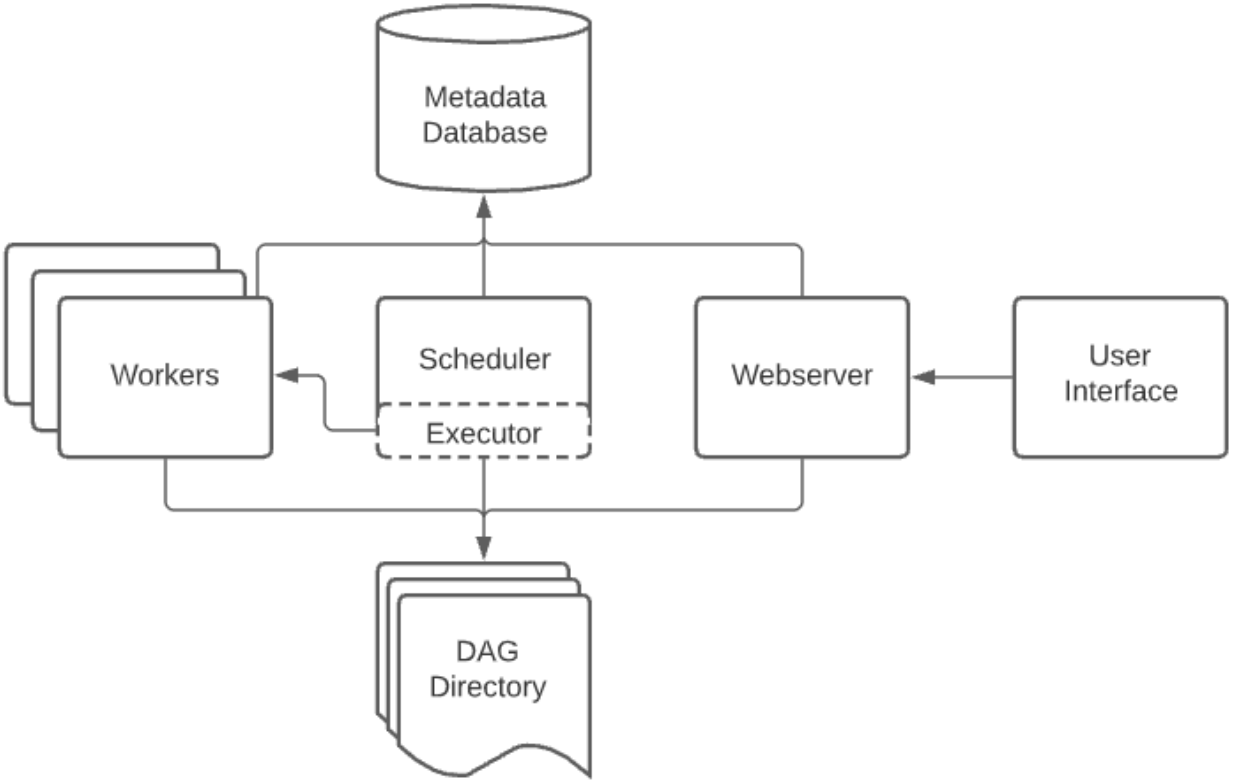
Комуникация между задачи

```
value = task.xcom_pull(task_ids='push_task')
```

Периоди и време

```
airflow dags backfill beme \
    --start-date 2020-08-14 \
    --end-date 2021-08-14
```

Архитектура



Docker & K8S

```
tX = DockerOperator(…)


t1 >> t2
t1 >> t3
t3 >> t4
```

Облаци

- AWS
- Azure
- Google Cloud Platform

3 в 1

Деца аутисти

Големи данни

Apache Airflow