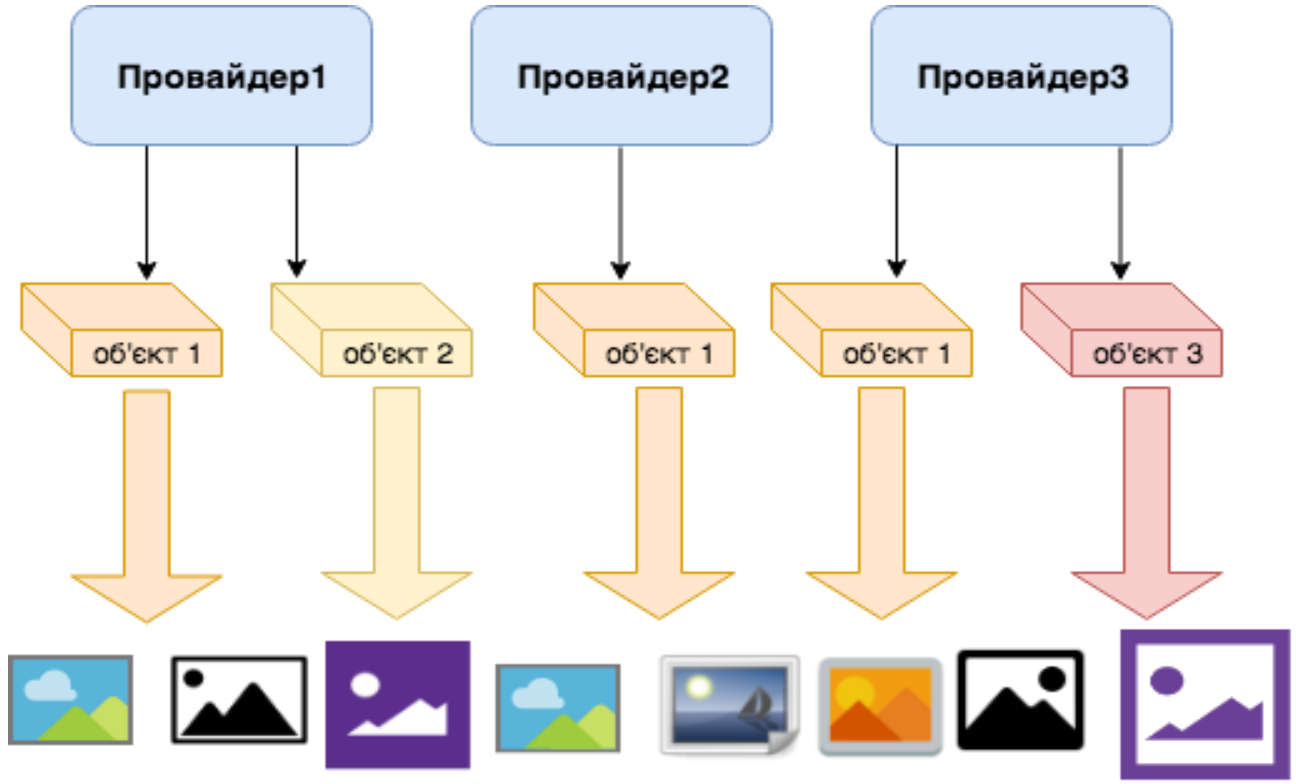
ТМ-71МП, ТЕРЯЄВ Г. О.

КЕРІВНИК: ВЕРЛАНЬ А.А.

ПОШУК ТА ВИДАЛЕННЯ ДУБЛІКАТІВ ЗОБРАЖЕНЬ

ПРОБЛЕМА ДУБЛІКАТІВ ЗОБРАЖЕНЬ



Концептуальна схема агрегації даних

ПРОБЛЕМА ТА МЕТА

- **Проблема**

Дублікований графічний контент в системах, що є агрегаторами товарів, послуг, тощо. Дублікати зображень формують негативне враження про якість контенту системи.

- **Мета**

Розробити програмний продукт, що знаходить дублікати серед зображень різного формату та розміру, здатен масштабуватись в залежності від кількості вхідних даних.

ДУБЛІКАТИ ЗОБРАЖЕНЬ

Які зображення називаються дублікатами?

- Ідентичні зображення
- Обрізані зображення
- Зображення зі зміненим рівнем яскравості чи контрасту
- Зжаті зображення
- «Схожі» зображення

ПРИКЛАД «СХОЖИХ ЗОБРАЖЕНЬ»

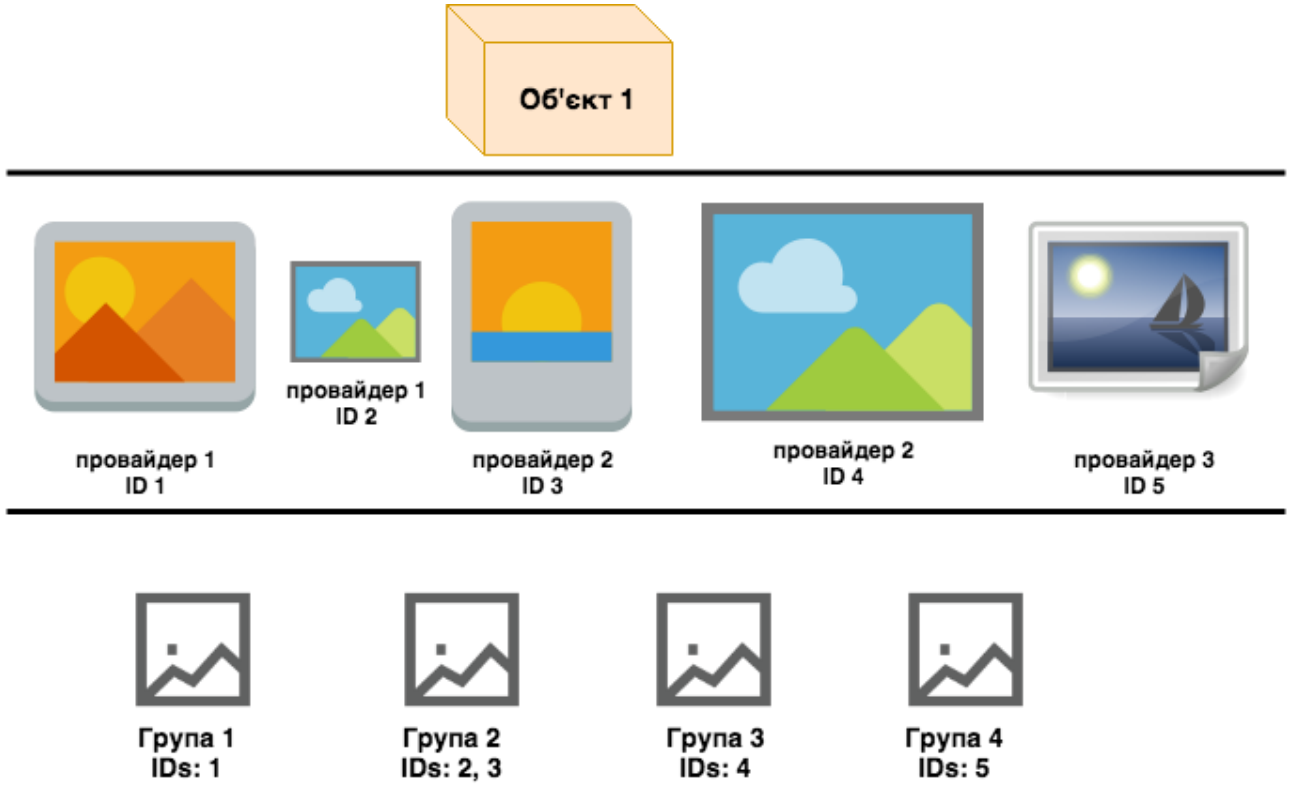
Фото однієї і тієї ж будівлі зроблені в різний час

ВХІДНІ ДАНІ


```
{'provider': 'ean',
 'provider_id': 53462,
 'internal_id': 34552,
 'images': [{'default': True,
   'description': 'Featured Image',
   'url': 'https://i.travelapi.com/hotels/3000000/2460000/2459500/2459478/6975b635_b.jpg'},
  {'default': False,
   'description': 'Lobby Sitting Area',
   'url': 'https://i.travelapi.com/hotels/3000000/2460000/2459500/2459478/009f72b7_b.jpg'},
  {'default': False,
   'description': 'Guestroom',
   'url': 'https://i.travelapi.com/hotels/3000000/2460000/2459500/2459478/2459478_45_b.jpg'},
  {'default': False,
   'description': 'Guestroom',
   'url': 'https://i.travelapi.com/hotels/3000000/2460000/2459500/2459478/2459478_65_b.jpg'},
```


Вхідні дані в форматі JSON

КОНЦЕПТУАЛЬНА СХЕМА

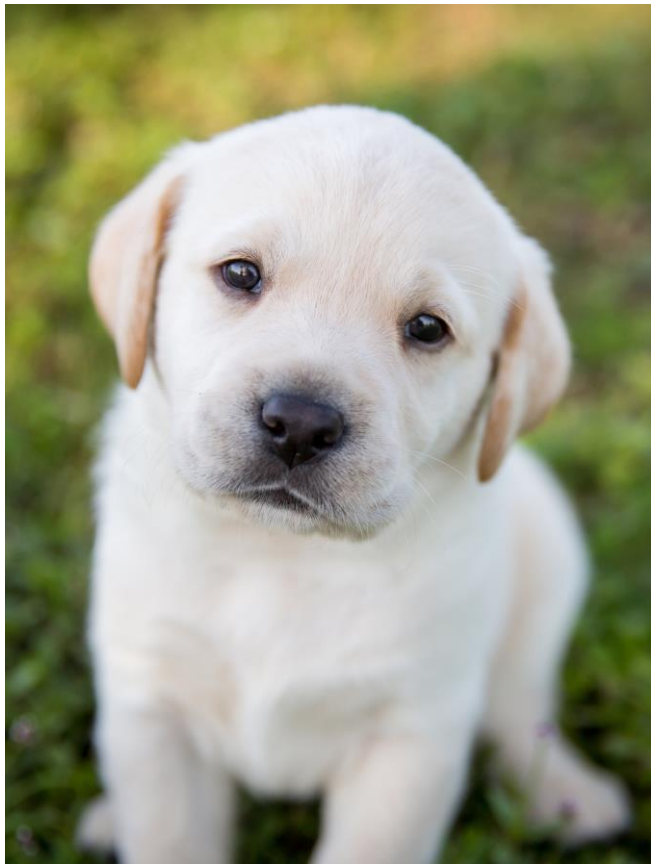


Концептуальна схема об’єктів з якими оперує система

ПОРІВНЯННЯ ЗОБРАЖЕНЬ

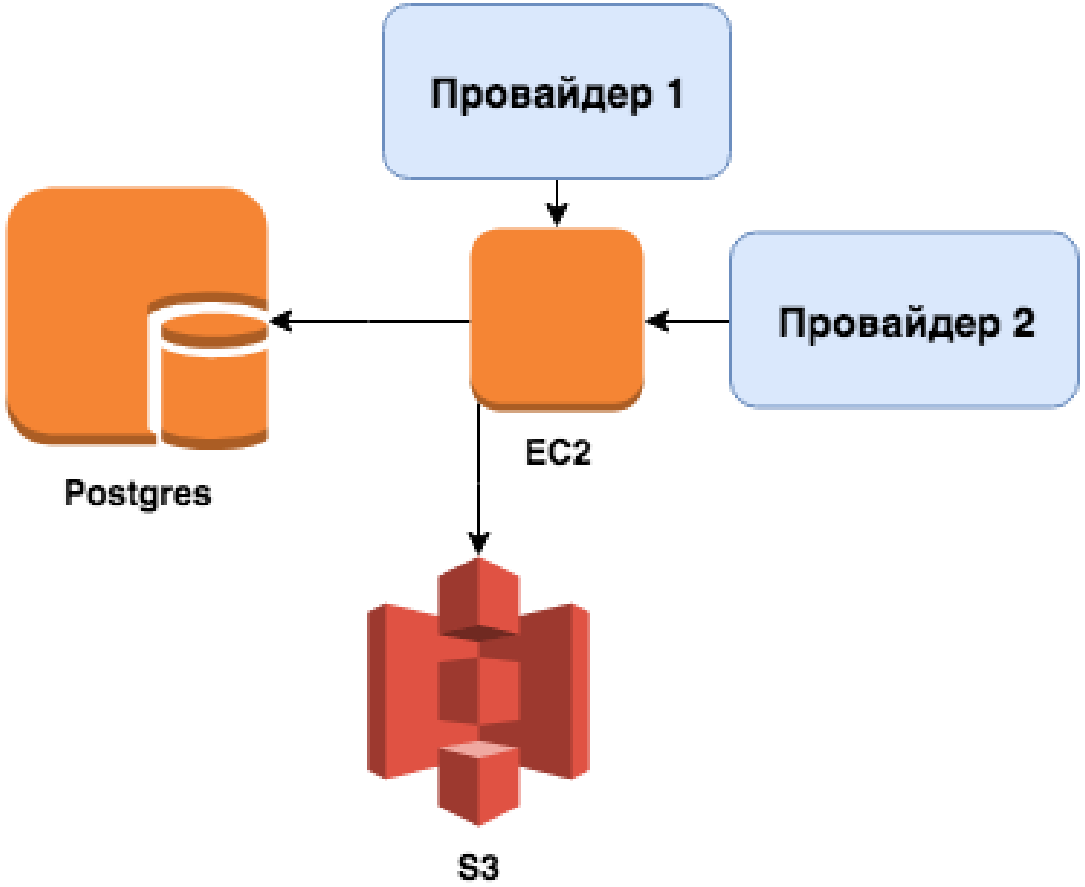
- **Представлення зображень: перцептивне хешування**
- **Порівняння хешів: Гемінгова відстань**
- **Бінарна класифікація: логістична регресія**

ПРОБЛЕМИ AVG HASHING



AVG HASH Цих зображень свідчить про те, що вони однакові

АРХІТЕКТУРА. ЗБІР ДАНИХ



АРХІТЕКТУРА

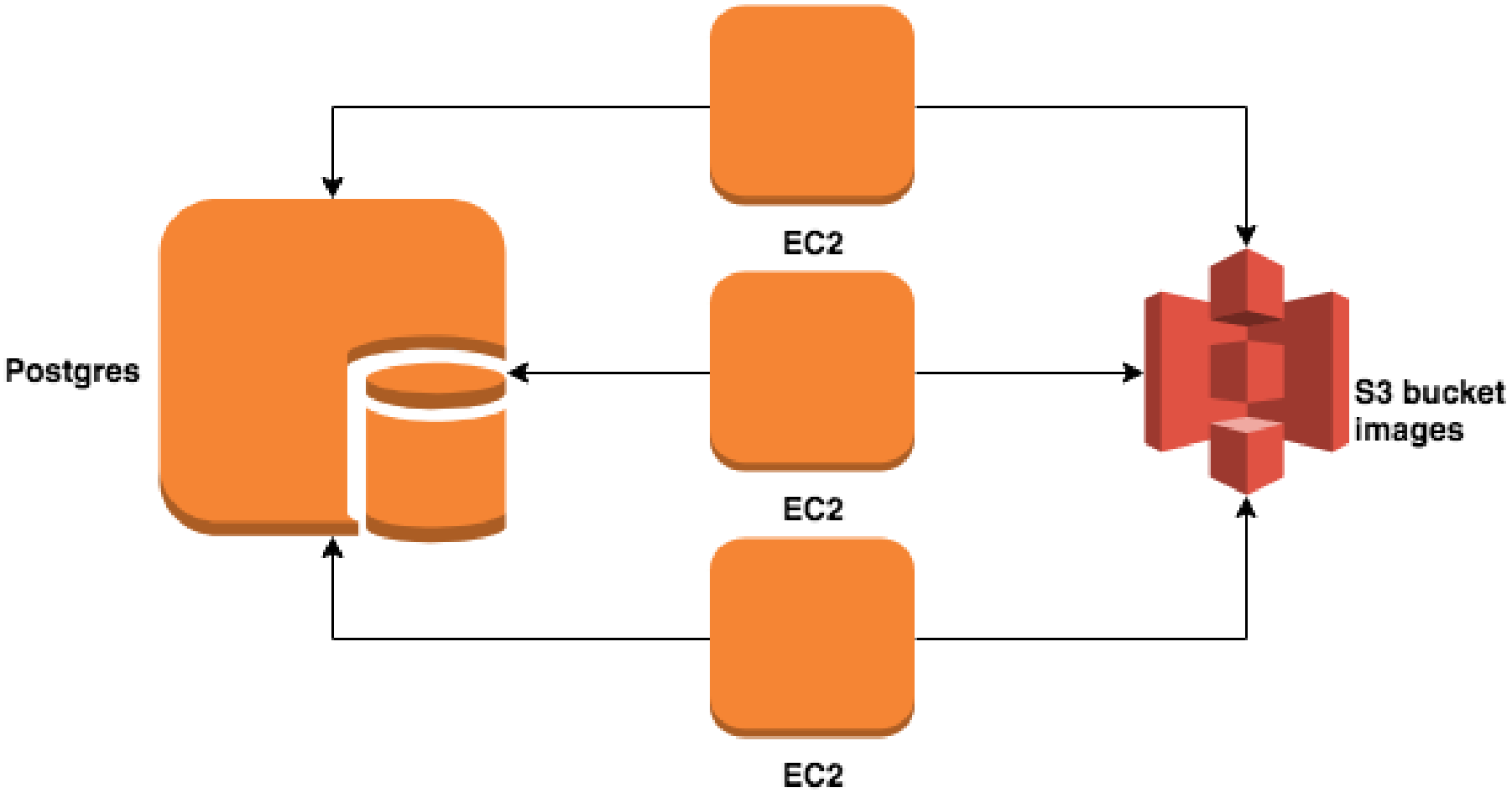
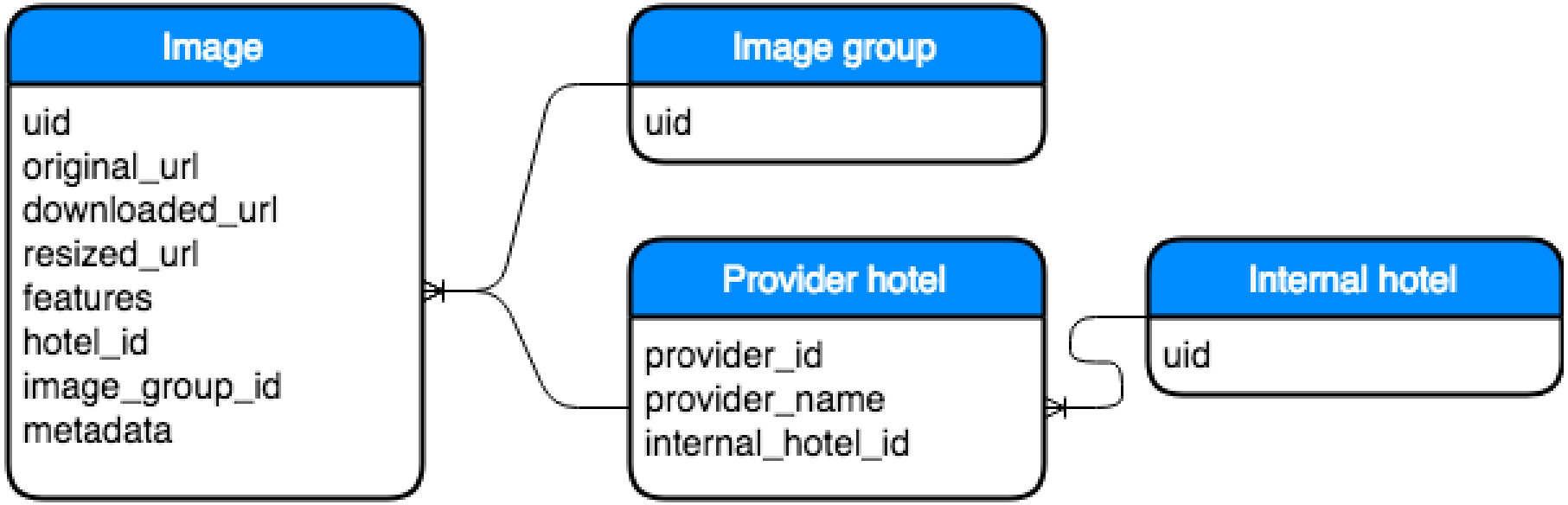


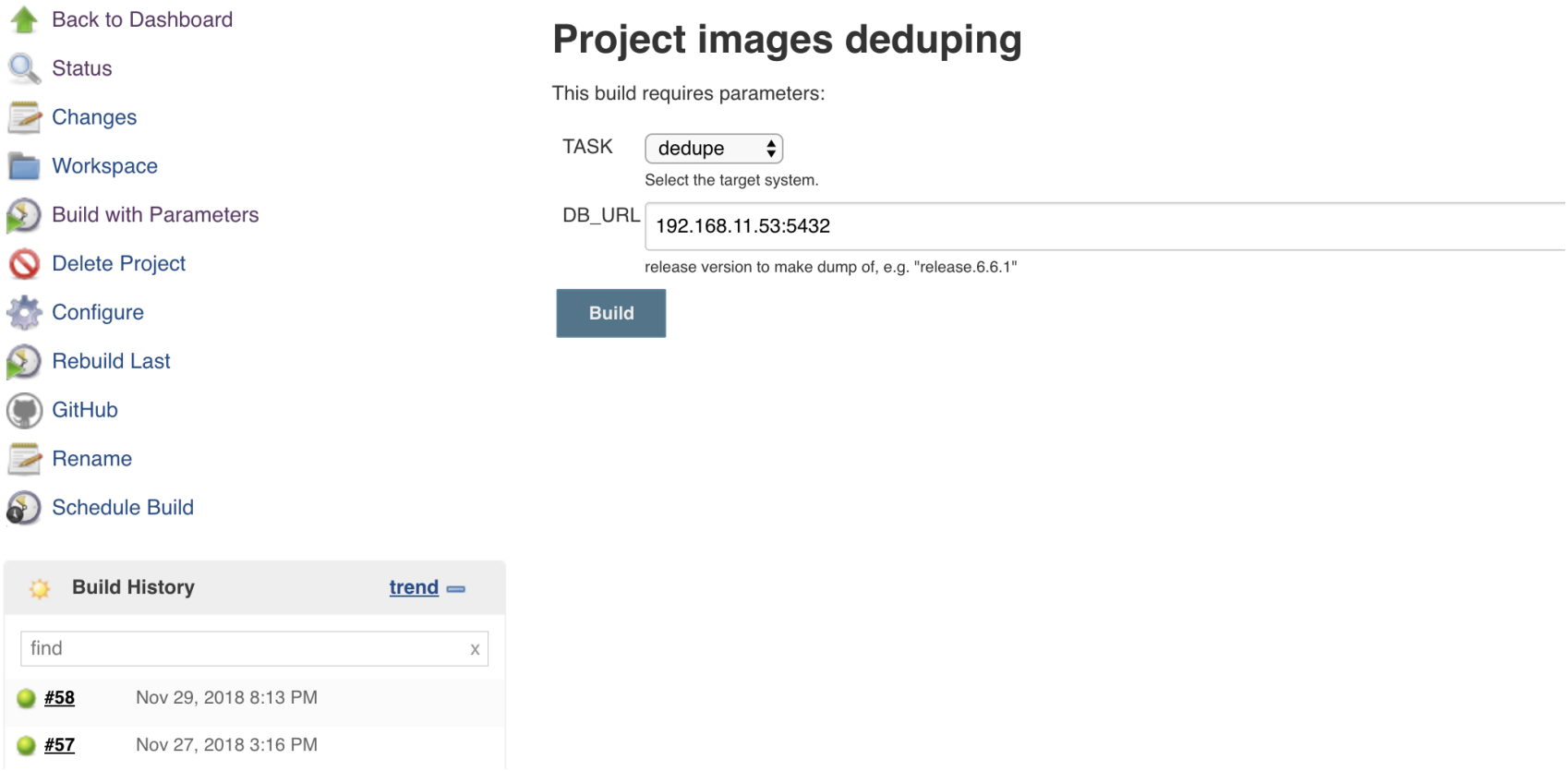
СХЕМА БАЗИ ДАНИХ



КОНФІГУРАЦІЯ ПРОГРАМИ

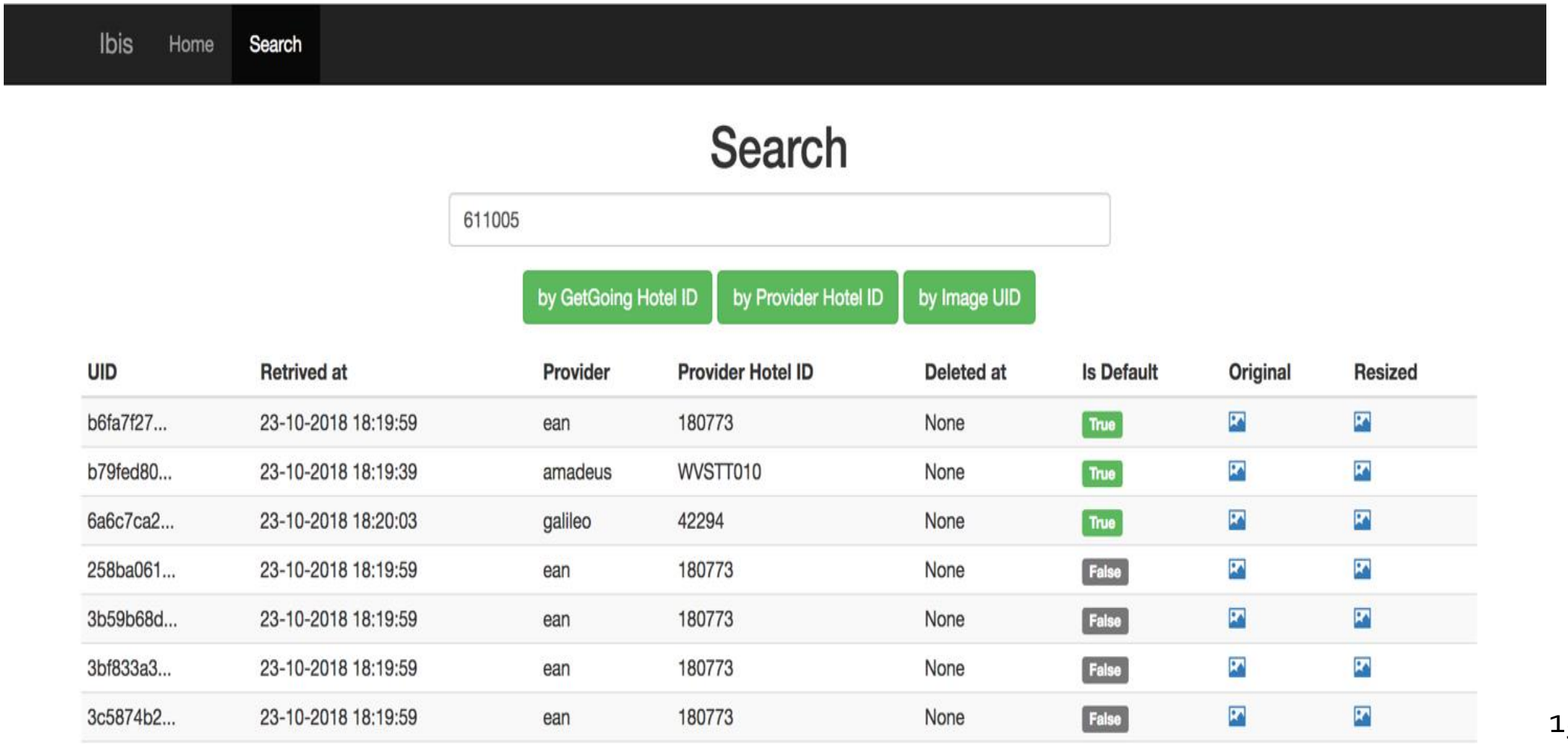
Вікно конфігурації запуску програми в системі Jenkins

ЗАПУСК ПРОГРАМИ



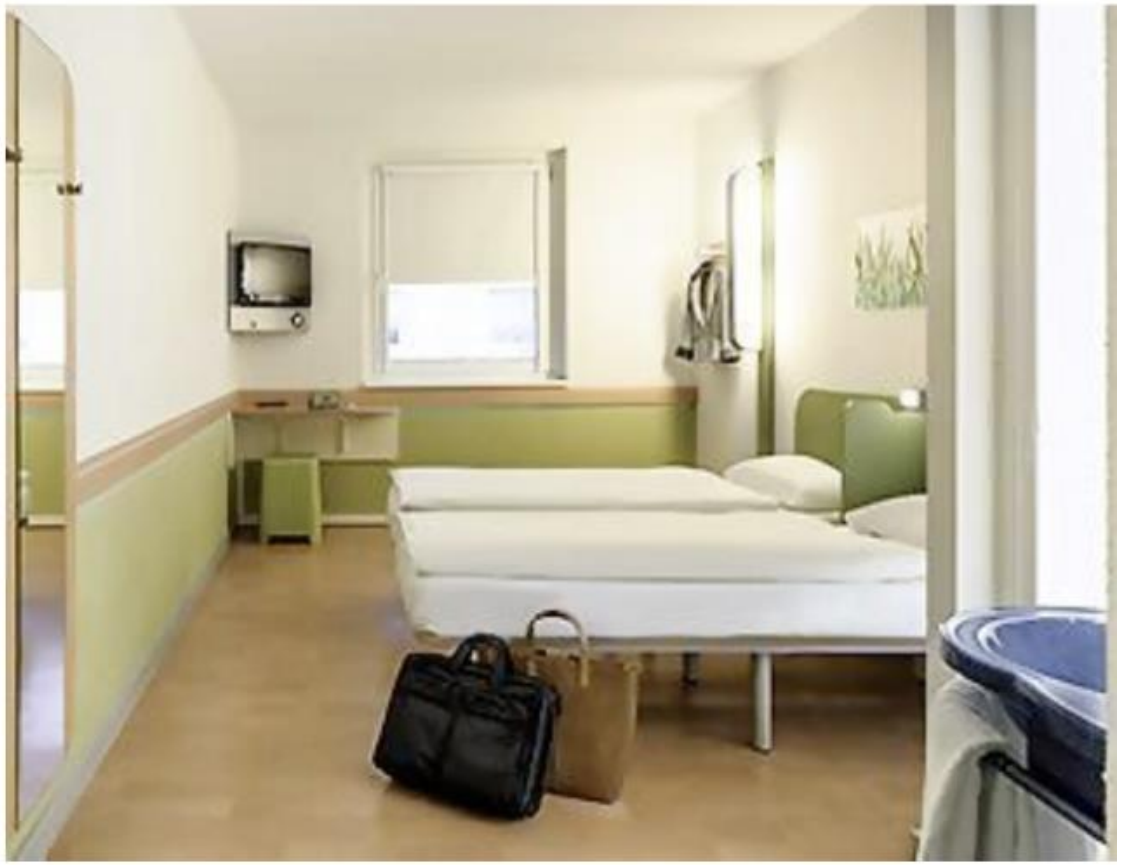
Запуск програми в системі Jenkins

ВЕБ ДОДАТОК. ПОШУК ЗА ІДЕНТИФІКАТОРОМ ГОТЕЛЮ



UID	Retrived at	Provider	Provider Hotel ID	Deleted at	Is Default	Original	Resized
b6fa7f27...	23-10-2018 18:19:59	ean	180773	None	True		
b79fed80...	23-10-2018 18:19:39	amadeus	WVSTT010	None	True		
6a6c7ca2...	23-10-2018 18:20:03	galileo	42294	None	True		
258ba061...	23-10-2018 18:19:59	ean	180773	None	False		
3b59b68d...	23-10-2018 18:19:59	ean	180773	None	False		
3bf833a3...	23-10-2018 18:19:59	ean	180773	None	False		
3c5874b2...	23-10-2018 18:19:59	ean	180773	None	False		

РЕЗУЛЬТАТИ



Приклад успішно класифікованих дублікатів зображень

РЕЗУЛЬТАТИ

Приклад успішно класифікованих дублікатів зображень (87%)

ВИСНОВКИ

- Найкращий результат досягається комбінуванням різних алгоритмів перцептивного хешування
- Процес пошуку дублікатів зображень потребує багато ресурсів (30 EC2 c3.large, 6m images, 24h)
- Результат роботи створеного програмного продукту зберігається в БД
- Результати на тестовій вибірці (1000 зображень 100 готелів) = 87% успіху