# ETL

KARPOV.COURSES

# БЛОК ETL

1. Введение в ETL
2. Знакомство с Airflow
3. Сложные пайплайны, часть 1
4. Сложные пайплайны, часть 2
5. Разработка своих плагинов
6. Установка и настройка Airflow

# ВВЕДЕНИЕ В ETL

**KARPOV.COURSES**

# ЛЕКЦИЯ «ВВЕДЕНИЕ В ETL»

1. Что такое ETL
2. Как правильно готовить ETL
3. Обзор планировщиков
4. Почему Airflow

# ЧТО ТАКОЕ ETL

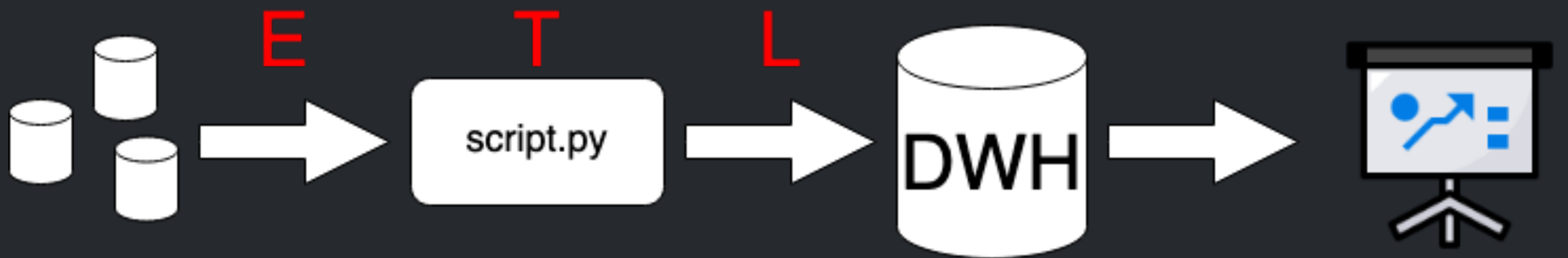
# ETL – ЭТО

процесс миграции данных из одного хранилища в другое

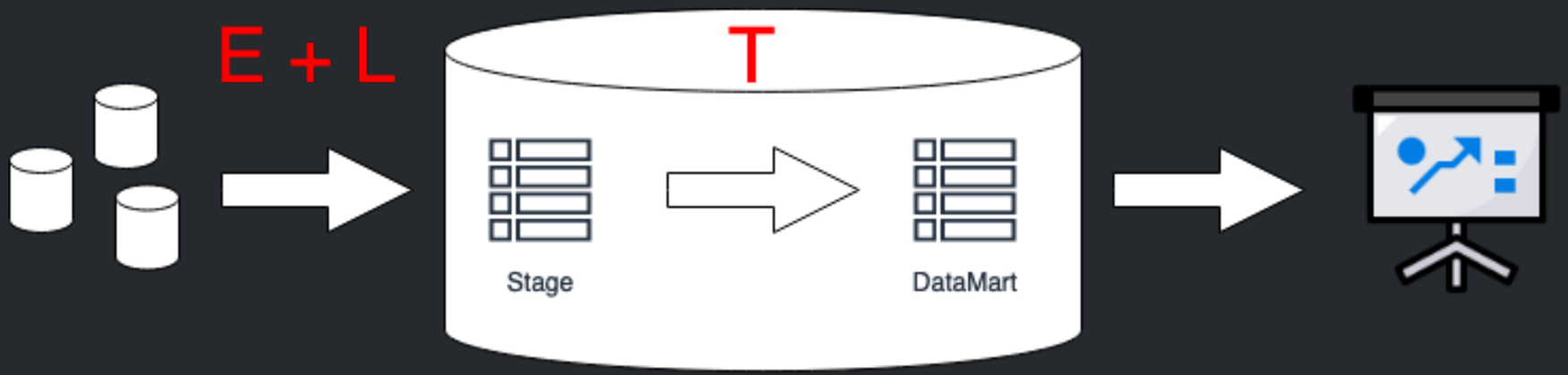
# ETL

- Extract
- Transform
- Load

# ПРИМЕР ETL



# ПРИМЕР ELT



# КАК ПРАВИЛЬНО ГОТОВИТЬ ETL

# Принципы построения ETL

— Чистый код

# Принципы построения ETL

- Чистый код
- Простота

# Принципы построения ETL

- Чистый код
- Простота
- Единообразие

# Принципы построения ETL

- Чистый код
- Простота
- Единообразие
- Время выполнения пайплайна

# Принципы построения ETL

- Чистый код
- Простота
- Единообразие
- Время выполнения пайплайна
- Меньше сетевого трафика

# Принципы построения ETL

- Чистый код
- Простота
- Единообразие
- Время выполнения пайплайна
- Меньше сетевого трафика
- Работа с репликой

# Принципы построения ETL

- Чистый код
- Простота
- Единообразие
- Время выполнения пайплайна
- Меньше сетевого трафика
- Работа с репликой
- Оптимизация забора данных

# Принципы построения ETL

- Чистый код
- Простота
- Единообразие
- Время выполнения пайплайна
- Меньше сетевого трафика
- Работа с репликой
- Оптимизация забора данных
- Способы загрузки данных (SCD)

# Принципы построения ETL

- Чистый код
- Простота
- Единообразие
- Время выполнения пайплайна
- Меньше сетевого трафика
- Работа с репликой
- Оптимизация забора данных
- Способы загрузки данных (SCD)
- Партиционирование

# Принципы построения ETL

- Чистый код
- Простота
- Единообразие
- Время выполнения пайплайна
- Меньше сетевого трафика
- Работа с репликой
- Оптимизация забора данных
- Способы загрузки данных (SCD)
- Партиционирование
- Инкрементальный пересчёт витрин

# Принципы построения ETL

- Чистый код
- Простота
- Единообразие
- Время выполнения пайплайна
- Меньше сетевого трафика
- Работа с репликой
- Оптимизация забора данных
- Способы загрузки данных (SCD)
- Партиционирование
- Инкрементальный пересчёт витрин
- Загрузка всего без ограничений

# Принципы построения ETL

- Чистый код
- Простота
- Единообразие
- Время выполнения пайплайна
- Меньше сетевого трафика
- Работа с репликой
- Оптимизация забора данных
- Способы загрузки данных (SCD)
- Партиционирование
- Инкрементальный пересчёт витрин
- Загрузка всего без ограничений
- Избавляться от неактуального

# Принципы построения ETL

- Чистый код
- Простота
- Единообразие
- Время выполнения пайплайна
- Меньше сетевого трафика
- Работа с репликой
- Оптимизация забора данных
- Способы загрузки данных (SCD)
- Партиционирование
- Инкрементальный пересчёт витрин
- Загрузка всего без ограничений
- Избавляться от неактуального
- Идемподентность

# Принципы построения ETL

- Чистый код
- Простота
- Единообразие
- Время выполнения пайплайна
- Меньше сетевого трафика
- Работа с репликой
- Оптимизация забора данных
- Способы загрузки данных (SCD)
- Партиционирование
- Инкрементальный пересчёт витрин
- Загрузка всего без ограничений
- Избавляться от неактуального
- Идемподентность
- Аудиторский след

# Будьте готовы

# Будьте готовы

- Отсутствие целостности

# Будьте готовы

- Отсутствие целостности
- Сетевые проблемы

# Будьте готовы

- Отсутствие целостности
- Сетевые проблемы
- Незапланированные изменения

# Будьте готовы

- Отсутствие целостности
- Сетевые проблемы
- Незапланированные изменения
- Пайплайны задерживаются

# Будьте готовы

- Отсутствие целостности
- Сетевые проблемы
- Незапланированные изменения
- Пайплайны задерживаются
- Данные в разных системах противоречивы

# ОБЗОР ПЛАНИРОВЩИКОВ

# CRON

**Плюсы**

— Максимально простой

**Минусы**

— Максимально простой

# JENKINS / GITLAB CI

— Инструмент для CI/CD

GitLab

# НАПИСАТЬ СВОЙ

# HH.RU (2021-10-03)

| | |
|---|---|
| — Airflow | 685 |
| — SSIS | 296 |
| — Oracle Data Integrator | 252 |
| — NiFi | 138 |
| — Oozie | 67 |
| — Talend | 60 |
| — Informatica PowerCenter | 55 |
| — Luigi | 50 |
| — SAS DIS | 44 |

# Платные шедулеры

- Дорогие
- Нет доступа к коду
- Но есть поддержка
- Визуальный редактор

# MS SQL SERVER INTEGRATION SYSTEM

- Интегрирован в MS SQL Server
- Транзакционность из коробки
- Визуальный редактор
- https://www.microsoft.com/en-us/sql-server/sql-server-2019

# ORACLE DATA INTEGRATOR

- Входит в экосистему Oracle
- Визуальный редактор
- https://www.oracle.com/middleware/technologies/data-integrator.html

# INFORMATICA POWER CENTER

- Визуальный редактор
- Встроенные утилиты контроля качества данных
- https://www.informatica.com/products/data-integration/powercenter.html

# SAS DATA INTEGRATION STUDIO

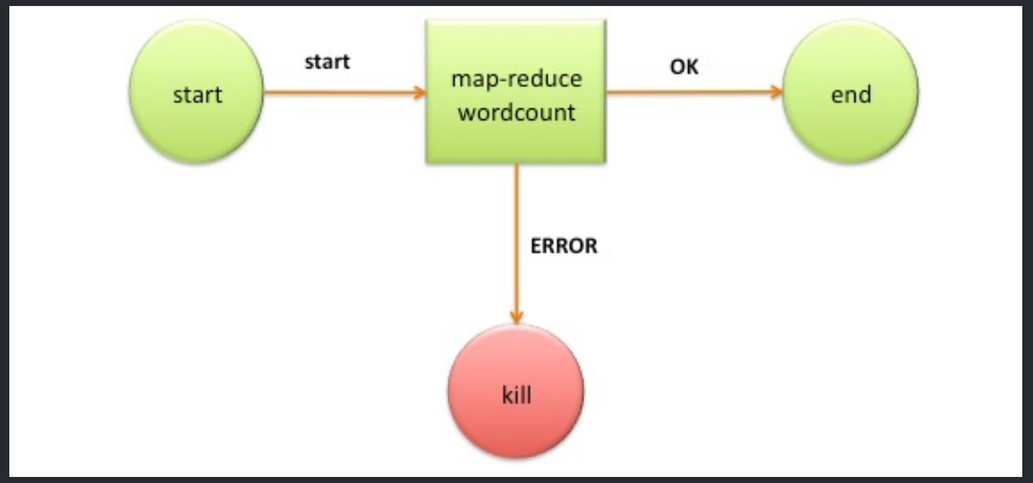
- Бесплатное обучение
- Визуальное построение пайплайнов
- Язык SAS Base
- SAS Access – плагины для источников

# Open Source

- Бесплатные
- Можно посмотреть в код
- Можно контрибьютить

# APACHE OOZIE

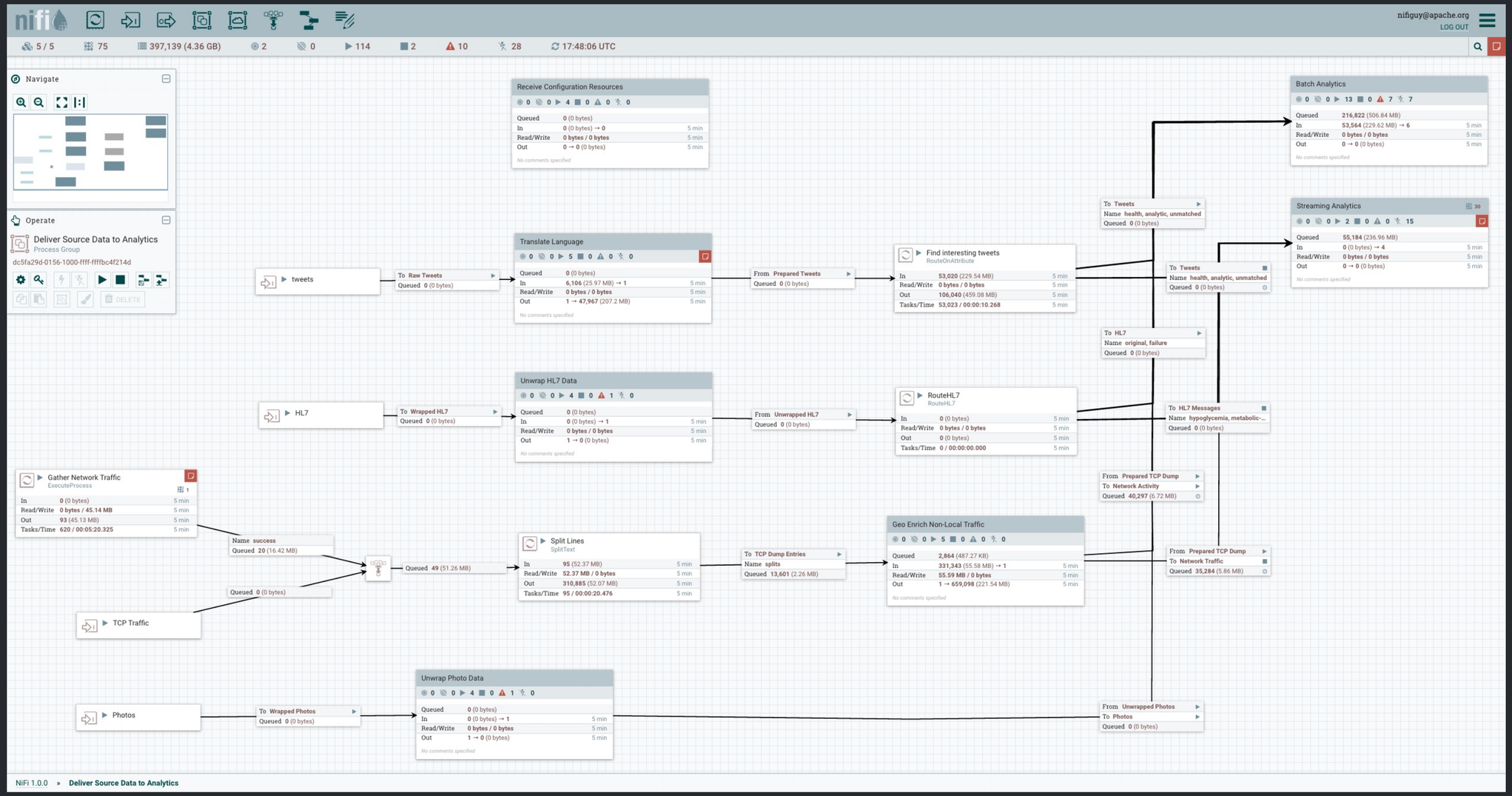

- Планировщик для Hadoop
- Пайплайны на xml
- https://oozie.apache.org/



# APACHE NIFI


- Open Source
- Работает внутри Hadoop
- В том числе с внешними источниками
- Потоковая и батчевая обработка
- https://nifi.apache.org/

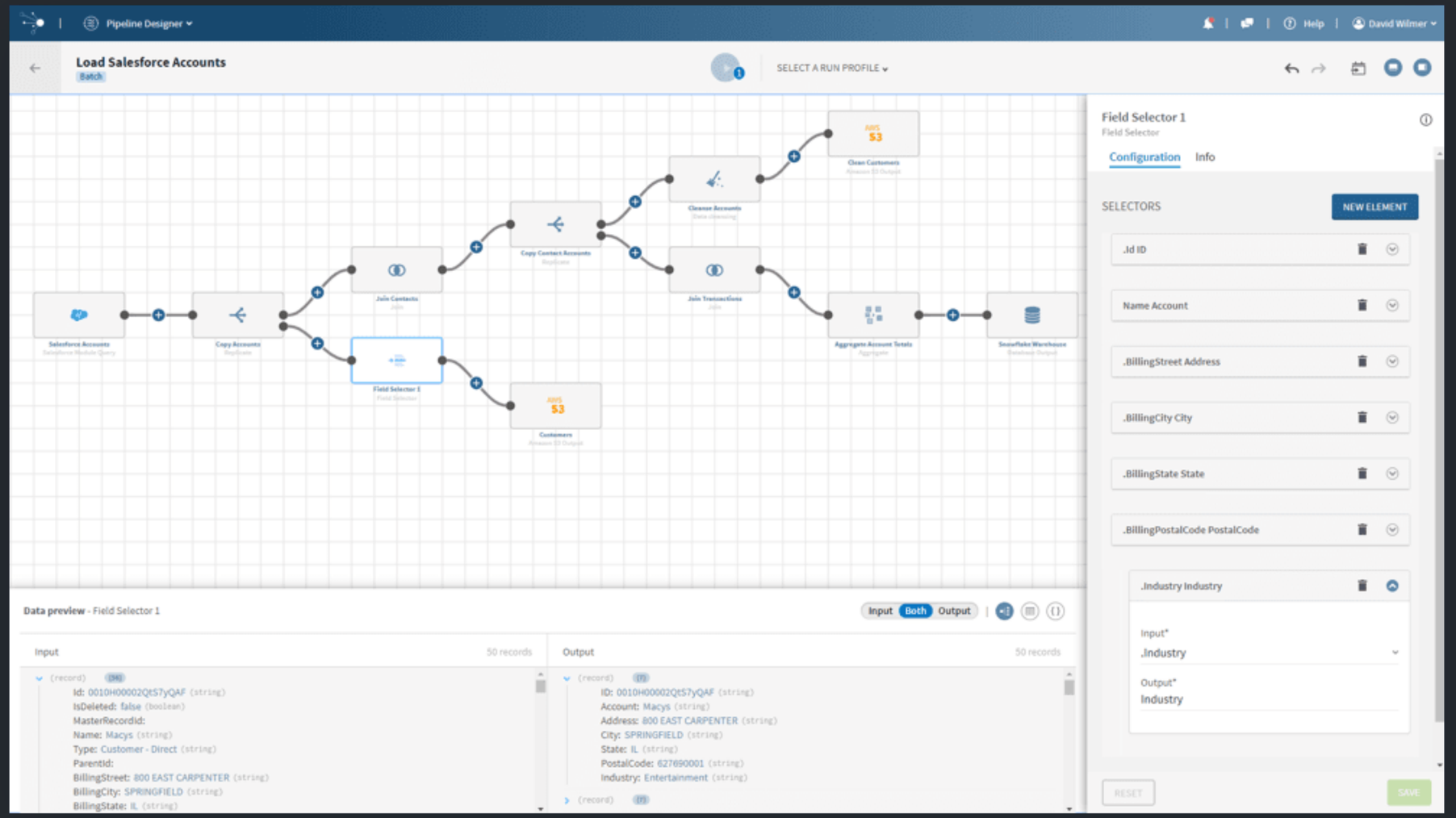
# APACHE NIFI


nifiguy@apache.org
LOG OUT
5 / 5
75
397,139 (4.36 GB)
2
0
114
2
10
28
17:48:06 UTC
Navigate
Operate
Deliver Source Data to Analytics
Process Group
dc5fa29d-0156-1000-ffff-ffffbc4f214d
DELETE
Receive Configuration Resources
Batch Analytics
Streaming Analytics
Translate Language
tweets
To Raw Tweets
From Prepared Tweets
Find interesting tweets
RouteOnAttribute
To Tweets
Name health, analytic, unmatched
To HL7
Name original, failure
Unwrap HL7 Data
HL7
To Wrapped HL7
From Unwrapped HL7
RouteHL7
To HL7 Messages
Name hypoglycemia, metabolic-...
Gather Network Traffic
ExecuteProcess
From Prepared TCP Dump
To Network Activity
Queued 40,297 (6.72 MB)
Name success
Queued 20 (16.42 MB)
Queued 49 (51.26 MB)
Queued 0 (0 bytes)
Split Lines
SplitText
To TCP Dump Entries
Name splits
Queued 13,601 (2.26 MB)
Geo Enrich Non-Local Traffic
Queued 2,864 (487.27 KB)
To Network Traffic
Queued 35,284 (5.86 MB)
TCP Traffic
Unwrap Photo Data
Photos
To Wrapped Photos
From Unwrapped Photos
To Photos
NiFi 1.0.0 » Deliver Source Data to Analytics

# TALEND


- Open Source
- Визуальный редактор
- Потоки между задачами
- Требует знания Java
- https://www.talend.com/

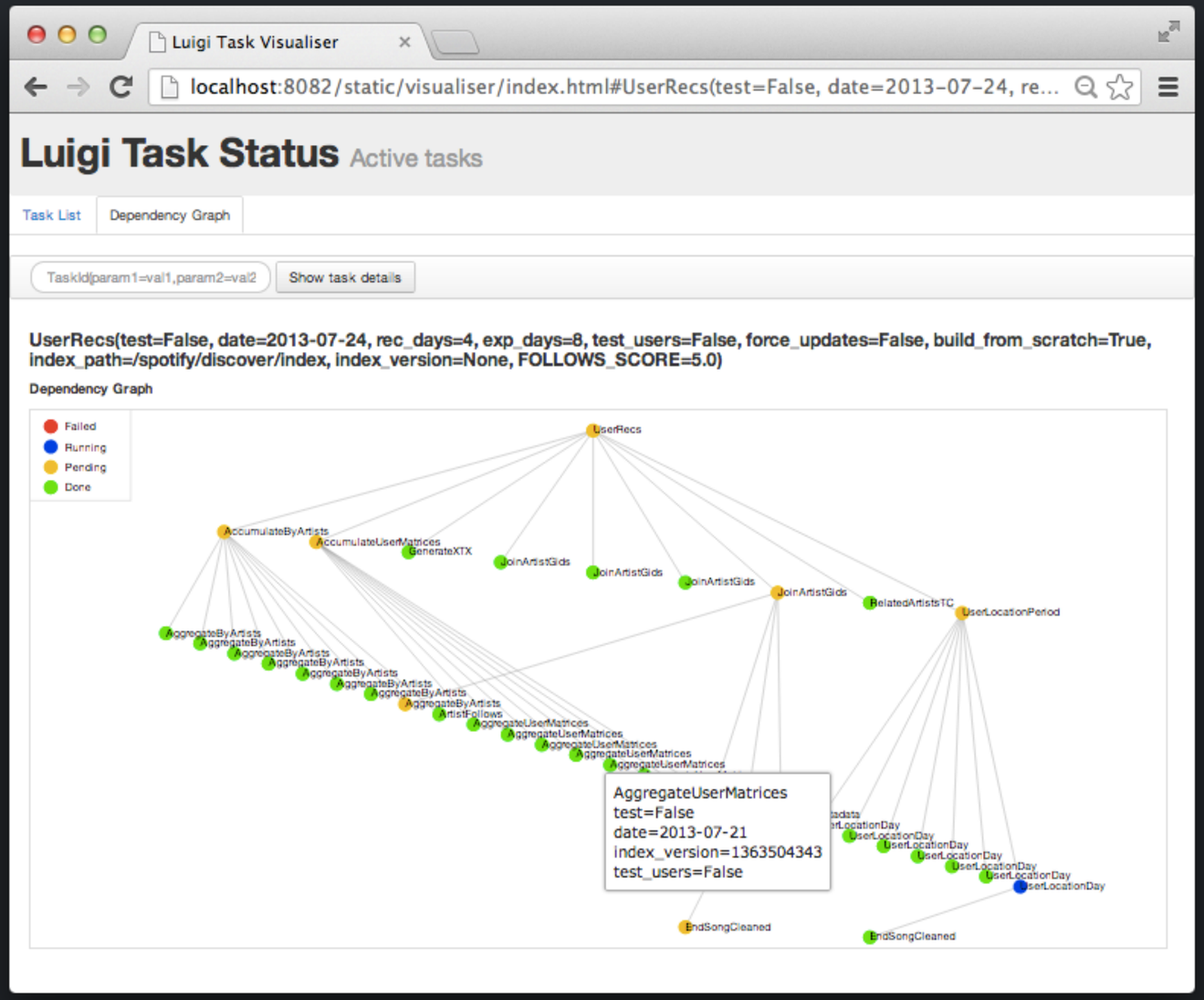
# TALEND


Pipeline Designer
Help
David Wilmer
Load Salesforce Accounts
Batch
SELECT A RUN PROFILE
Field Selector 1
Field Selector
Configuration
Info
SELECTORS
NEW ELEMENT
.Id ID
Name Account
.BillingStreet Address
.BillingCity City
.BillingState State
.BillingPostalCode PostalCode
.Industry Industry
Input*
.Industry
Output*
Industry
RESET
SAVE
Data preview - Field Selector 1
Input
Both
Output
Input
50 records
Output
50 records
Id: 0010H00002QtS7yQAF (string)
IsDeleted: false (boolean)
MasterRecordId:
Name: Macys (string)
Type: Customer - Direct (string)
ParentId:
BillingStreet: 800 EAST CARPENTER (string)
BillingCity: SPRINGFIELD (string)
BillingState: IL (string)
ID: 0010H00002QtS7yQAF (string)
Account: Macys (string)
Address: 800 EAST CARPENTER (string)
City: SPRINGFIELD (string)
State: IL (string)
PostalCode: 627690001 (string)
Industry: Entertainment (string)

# LUIGI


- Open Source
- Код на Питоне
- Работает поверх артефактов
- Нет встроенного планировщика
- Почти не развивается
- https://github.com/spotify/luigi

# LUIGI


Luigi Task Visualiser
localhost:8082/static/visualiser/index.html#UserRecs(test=False, date=2013-07-24, re...
Luigi Task Status Active tasks
Task List
Dependency Graph
TaskId(param1=val1,param2=val2
Show task details
UserRecs(test=False, date=2013-07-24, rec_days=4, exp_days=8, test_users=False, force_updates=False, build_from_scratch=True, index_path=/spotify/discover/index, index_version=None, FOLLOWS_SCORE=5.0)
Dependency Graph
Failed
Running
Pending
Done
UserRecs
AccumulateByArtists
AccumulateUserMatrices
GenerateXTX
JoinArtistGids
JoinArtistGids
JoinArtistGids
JoinArtistGids
RelatedArtistsTC
UserLocationPeriod
AggregateByArtists
ArtistFollows
AggregateUserMatrices
AggregateUserMatrices
AggregateUserMatrices
AggregateUserMatrices
AggregateUserMatrices
AggregateUserMatrices
UserLocationDay
EndSongCleaned
EndSongCleaned
AggregateUserMatrices
test=False
date=2013-07-21
index_version=1363504343
test_users=False

# ПОЧЕМУ AIRFLOW

# Преимущества Airflow

- Open-source

# Преимущества Airflow

- Open-source
- Отличная документация

# Преимущества Airflow

- Open-source
- Отличная документация
- Простой код на Питоне

# Преимущества Airflow

- Open-source
- Отличная документация
- Простой код на Питоне
- Удобный веб-интерфейс

# Преимущества Airflow

- Open-source
- Отличная документация
- Простой код на Питоне
- Удобный веб-интерфейс
- Алертинг и мониторинг

# Преимущества Airflow

- Open-source
- Отличная документация
- Простой код на Питоне
- Удобный веб-интерфейс
- Алертинг и мониторинг
- Интеграция с основными источниками

# Преимущества Airflow

- Open-source
- Отличная документация
- Простой код на Питоне
- Удобный веб-интерфейс
- Алертинг и мониторинг
- Интеграция с основными источниками
- Кастомизация

# Преимущества Airflow

- Open-source
- Отличная документация
- Простой код на Питоне
- Удобный веб-интерфейс
- Алертинг и мониторинг
- Интеграция с основными источниками
- Кастомизация
- Масштабирование

# Преимущества Airflow

- Open-source
- Отличная документация
- Простой код на Питоне
- Удобный веб-интерфейс
- Алертинг и мониторинг
- Интеграция с основными источниками
- Кастомизация
- Масштабирование
- Большое комьюнити

# СПАСИБО

**ДИНА САФИНА**