Workshop №2

Airflow и MinIO (S3)

Яндекс Практикум @moneretin Михаил Неретин

План воркшопа

1. Что такое Data Lake на примере MinIO (S3)?
2. Напишем (разберем) пайплайн в Airflow

Что такое Data Lake (DL)?

Что такое Data Lake (DL)?

Концепция DL — хранить сырые данные, доступные для анализа

Что такое Data Lake (DL)?

Концепция DL — хранить сырые данные, доступные для анализа

DL — это хранилище сырых данных любого формата (JSON, CSV, PNG)

Что такое Data Lake (DL)?

Концепция DL — хранить сырые данные, доступные для анализа

DL — это хранилище сырых данных любого формата (JSON, CSV, PNG)

Зачем?

Что такое Data Lake (DL)?

Концепция DL — хранить сырые данные, доступные для анализа

DL — это хранилище сырых данных любого формата (JSON, CSV, PNG)

Зачем?

- анализ сырых данных

Что такое Data Lake (DL)?

Концепция DL — хранить сырые данные, доступные для анализа

DL — это хранилище сырых данных любого формата (JSON, CSV, PNG)

Зачем?

- анализ сырых данных
- возможность хранения любых типов данных

DWH vs DL

DWH vs DL

DWH	DL
Схема на запись	Схема на чтение

DWH vs DL

DWH	DL
Схема на запись	Схема на чтение
Структурированные данные	Любые типы данных

DWH vs DL

DWH	DL
Схема на запись	Схема на чтение
Структурированные данные	Любые типы данных
Стоимость высокая	Стоимость низкая

DWH vs DL

DWH	DL
Схема на запись	Схема на чтение
Структурированные данные	Любые типы данных
Стоимость высокая	Стоимость низкая
Доступность данных высокая	Доступность данных низкая

Что под капотом у DL?

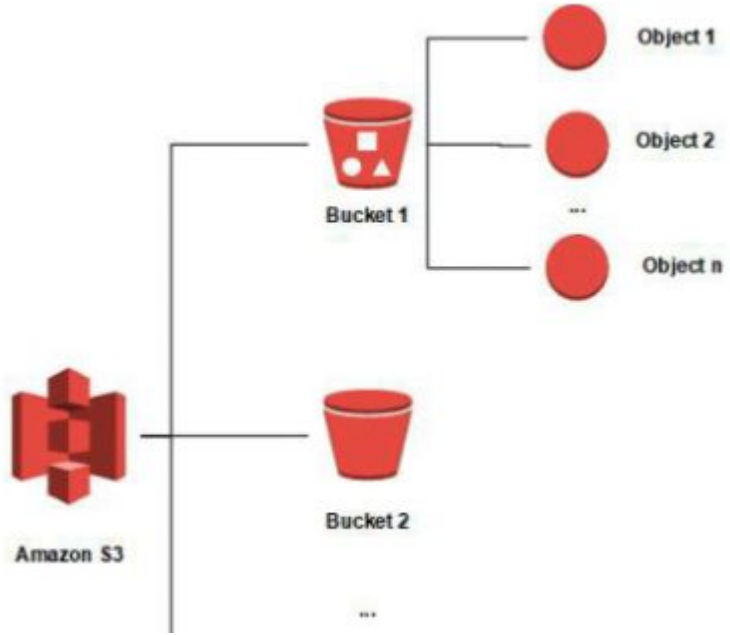
Что под капотом у DL?

- Облачные решения
    - Amazon S3
    - Google Cloud Storage
    - Azure Data Lake Storage

Что под капотом у DL?

- Облачные решения
  - Amazon S3
  - Google Cloud Storage
  - Azure Data Lake Storage
- Open-source решения
  - Apache Hadoop
  - MinIO
  - Ceph
  - HyperStore

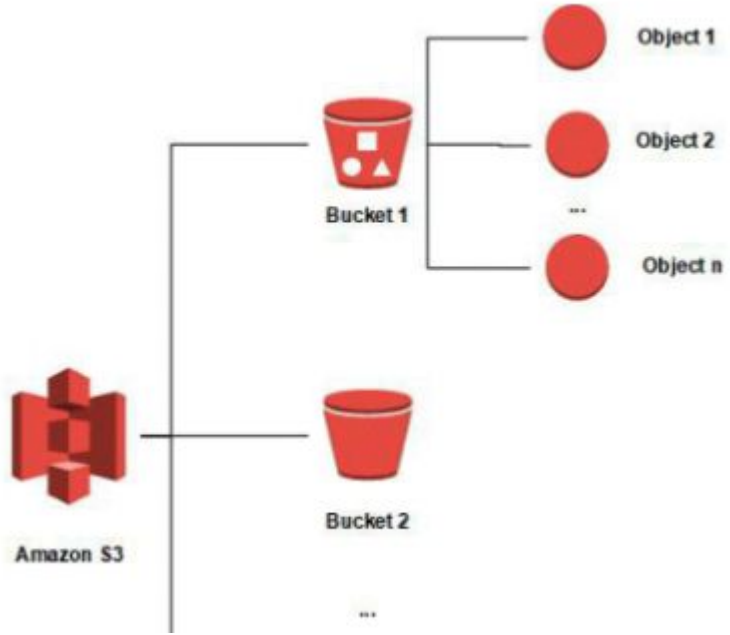
Amazon S3 (Simple Storage Service)



Amazon S3 (Simple Storage Service)

Основные сущности:

- bucket
- object

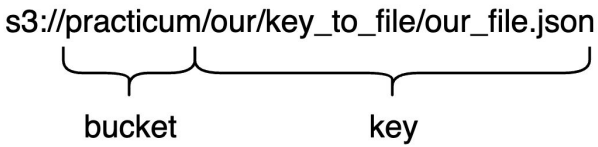


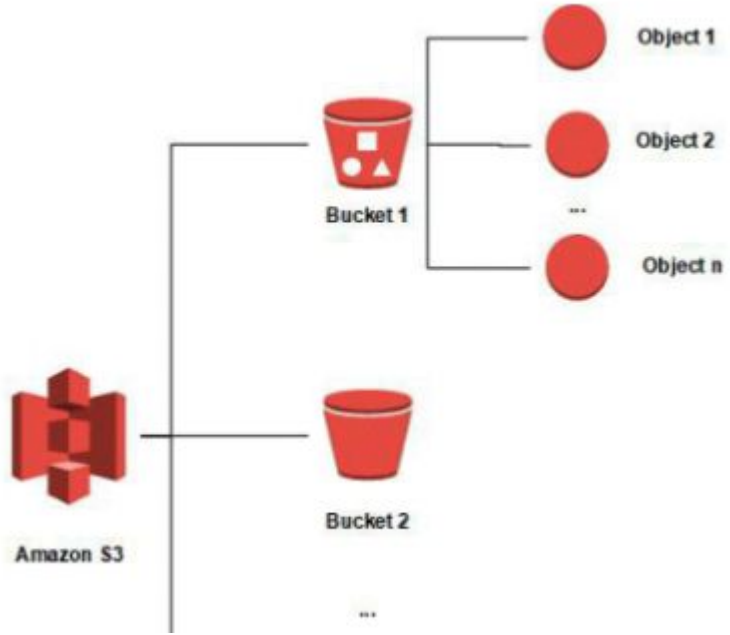
Amazon S3 (Simple Storage Service)

Основные сущности:

- bucket
- object

Путь состоит из bucket + object key (key)



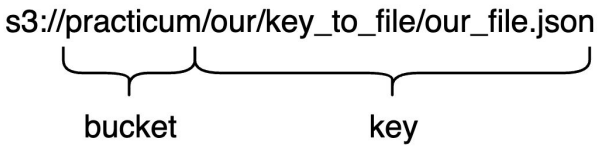


Amazon S3 (Simple Storage Service)

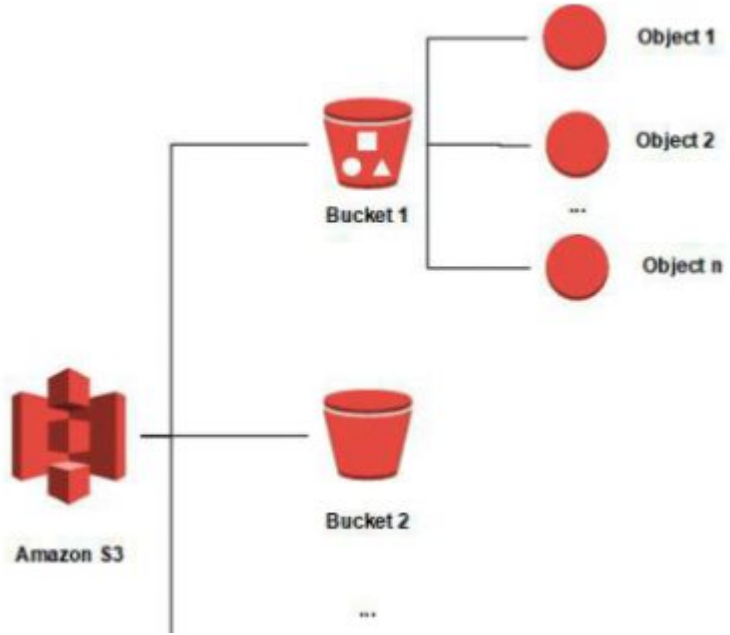
Основные сущности:

- bucket
- object

Путь состоит из bucket + object key (key)



Работает по HTTP протоколу



MinIO (S3 compatible storage)

Яндекс Практикум @moneretin Михаил Неретин

MinIO (S3 compatible storage)

- Open-source решение

MinIO (S3 compatible storage)

- Open-source решение
- S3 совместимое API

MinIO (S3 compatible storage)

- Open-source решение
- S3 совместимое API
- Не совсем честный S3 (метод хранения данных реализован поверх файловой системы) → не стоит хранить кучу мелких файлов

Практика

Ссылка на репу

Практика

Задача: Пришли Data Scientists и попросили предоставить данные о валютной паре USD/RUB.

Практика

Задача: Пришли Data Scientists и попросили предоставить данные о валютной паре USD/RUB.

Исходные данные:

Практика

Задача: Пришли Data Scientists и попросили предоставить данные о валютной паре USD/RUB.

Исходные данные:

1. У нас DL на базе MinIO

Практика

Задача: Пришли Data Scientists и попросили предоставить данные о валютной паре USD/RUB.

Исходные данные:

1. У нас DL на базе MinIO
2. Соседняя команда грузит к нам сырые данные

Практика

Задача: Пришли Data Scientists и попросили предоставить данные о валютной паре USD/RUB.

Исходные данные:

1. У нас DL на базе MinIO
2. Соседняя команда грузит к нам сырые данные
3. У нас есть открытое API откуда нам надо забирать данные

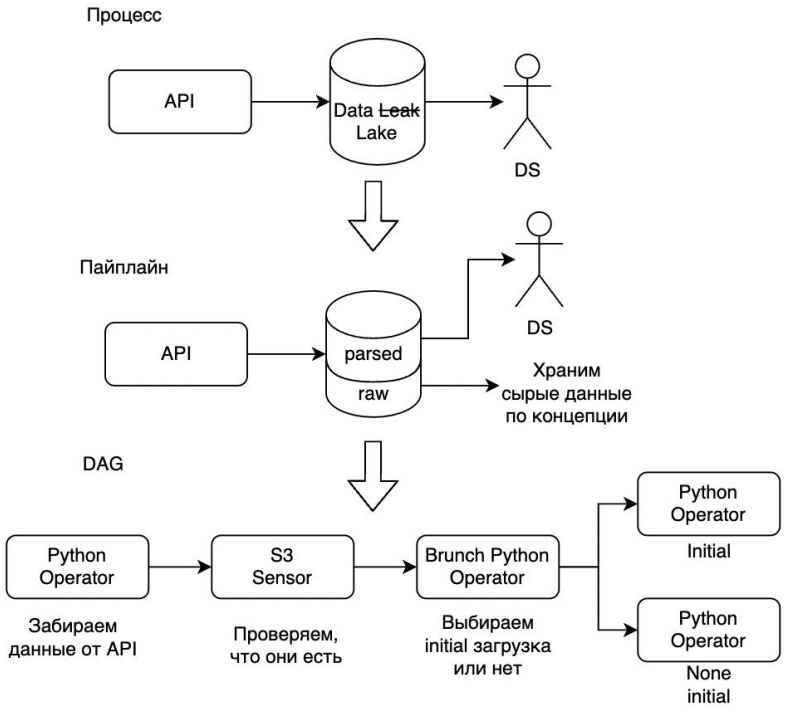
Практика

Задача: Пришли Data Scientists и попросили предоставить данные о валютной паре USD/RUB.

Исходные данные:

1. У нас DL на базе MinIO
2. Соседняя команда грузит к нам сырые данные
3. У нас есть открытое API откуда нам надо забирать данные
4. Будем обрабатывать сырые данные и предоставлять DS обработанные данные

Практика


Процесс
API
Data ~~Leak~~ Lake
DS
Пайплайн
API
parsed
raw
DS
Храним сырые данные по концепции
DAG
Python Operator
Забираем данные от API
S3 Sensor
Проверяем, что они есть
Brunch Python Operator
Выбираем initial загрузка или нет
Python Operator
Initial
Python Operator
None initial

Домашнее задание (по желанию)

1. Нужно запускать пайплайн каждый день в 7 утра (ежедневная прогрузка)
2. Доработать хранение файлов (партиционирование + название файла)

Что почитать

- https://docs.min.io/docs/distributed-minio-quickstart-guide
- https://docs.min.io/docs/minio-erasure-code-quickstart-guide
- https://ivan-shamaev.ru/data-engineering-etl-pipeline-data-warehouse-datalake