**NY News**

**Program :** Data Engineer

**Difficulty :** 9/10

**Description :**

The objective of this project is to use the developer portal of the American newspaper [NY Times](https://developer.nytimes.com/apis), which offers several APIs to explore, to create its own API

The newspaper also offers a [dashboard](https://www.nytimes.com/interactive/2021/us/covid-cases.html) on the situation of Covid, the data is updated daily on [github](https://github.com/nytimes/covid-19-data/tree/master/live). It would be interesting to process this data too.

| **Step** | **Description** | **Goal** | **Modules / Masterclass / Templates** | **Conditions of validation** |
| --- | --- | --- | --- | --- |
| **1** | Collecting data | You need to use the :   * Article Search * Books * Times Wire API   but it is recommended to test all the APIs and choose one with a purpose.  Here, we will make an API that retrieves articles, and sends the user to the Times article. Then there will be a Books section where we will list information about bestsellers (we can retrieve meta-data for these books via webscraping) but also about where they were purchased. Then, with the Times Wire  Finally, but as a bonus, we would like to get this data updated on the Covid to keep the user up to date with the latest information. | You will need to use the requests library or you can use the Postman tool.  Webscraping (Selenium, Beautiful Soup) | Explanatory file of the treatment and the different data accessible (doc / pdf)  An example of collected data. |
| **2** | Data modeling | There are several types of data. The aim will be to use different databases depending on the need. (ElasticSearch,SQL)  Here, there is a possibility to process data in real time via the Times Wire API, it will also be necessary to perform the ingestion step. | 142 - SQL  Elasticsearch  143 - MongoDB  Kafka/Spark Streaming | A relational database  UML Diagram  A file who creates and queries the SQL database .  Same files for a Elastic/Mongo/ DataBase |
| **3** | Data consumption | There is no Machine Learning to be done on this type of data, it is more about data for reporting but feel free to find a problem to model. For reporting, if you use ElasticSearch, you can use Kibana or you can use a Dash app  Then, you will need to create an API of this and you will use Docker to deploy | ElasticSearch  DE121  FastAPI/Flask  Docker | ML Notebooks  API FastAPI  Docker |
|
| **4** | Automation of flows | The aim here is to link the various stages together and to provide for an update of the Dashboard/API | Airflow | Python file for Airflow |
| **5** | Defense | Demonstrate their application and explain the reasoning behind their project. | X | Defense Documentation |

**Bibliographie :**