Рассматриваем два типа пользователя:

1. Администратор – пользователь, который настраивает Hadoop кластер и устанавливает фреймворк
2. Обычный пользователь – пользователь, который будет использовать фреймворк для проведения расчетов

Администратор должен настроить Hadoop кластер (в случае, если его нет) и установить фреймворк. Для этого можно скачать репозиторий отсюда <https://github.com/emtirikov/hadoop_docker_swarm>. Входные данные для использования предложенной установки:

1. IP адреса всех хостов, где планируется установить Hadoop кластер. Все хосты должны иметь сетевой доступ друг к другу по всем портам
2. OC Ubuntu 20.04/centos 7/ или выше. Установка на более старые версии повлечет проблемы с совместимостью (придется многие библиотеки собирать вручную и разбираться с совместимостью)
3. Установленный Python версии 3.8 или выше.
4. Установленный Docker на все хосты (проверить возможность, думаю, можно добавить в установщик). Также должен быть установлен питоновский пакет docker
5. Должен быть установлен openssl версии 1.1 или выше
6. Установленный ansible либо на постороннем хосте, который также имеет доступ ко всем хостам, либо на одном из хостов, где предполагается установка Hadoop
7. Следовать инструкции на сайте github для конфигурирования и установки.
8. Также проверить установку matlab и обозначить для нее требования
9. После установки Hadoop кластера надо установить фреймфорк (подумать, как это надо делать, думаю, скорей всего через pip и также как его записать)

Кастомизация установки, выбор другого компонента. Для того, чтобы выбрать для установки другой компонент (например другая база данных или другая распределенная файловая система) нужно написать свой ansible ноутбук, который будет реализовывать эту установку. Ограничения при реализации данного ноутбука:

1. Установка должна проходить в контейнеры, которые были подняты в ноутбуке startup\_namenode\_datanode\_dockers.yml
2. Установачный компонент должен быть совместим с другими компонентами данной установки

Рассмотрим обычного пользователя. Обычный пользователь, это исследователь, который строит пайплайн для решения конкретной задачи и анализирует результаты.

Первый вопрос – нужно ли скачивать данные или нет. Если нет, то переходить к следующему шагу, если да то нужно воспользовать отдельно модулем upload\_data. Данный компонент не предполагает встраивание в пайплайн, так как это длительное и одноразовое действие. Для загрузки можно воспользоваться одним из готовых компонентов данного модуля или подготовить свой. Свой компонент должен наследоваться от класса UploadData и реализовывать три метода данного класса get\_list\_of\_objects, upload\_to\_ds и upload\_to\_db. Первый метод позволяет получить список субъектов, которые есть в данном источнике. На вход он принимает путь до источника данных, а возвращает список доступных субъектов. Второй метод отвечает за загрузку фМРТ данных из внешней системы в распределенную файловую систему. На вход она принимает список субъектов для загрузки и путь сохранения в распределенной файловой системе. Функция не предполагает возврат значений. Третья функция отвечает за загрузку метаданных в базу данных. На вход она принимает:

1. Словарь, где ключ это субъект, а значение это словарь со всеми метаданными
2. Название создаваемой таблицы (пока ограничение на то, что бд может быть реляционной или колоночной)

Функция таже не возвращает какое либо значение

Далее идет построение основного пайплайна для расчета. Далее описание добавления компонента для каждого модуля. Если это не предполагается этого делать, то можно выбрать из существующих и запустить расчет.

Для добавления компонентов к модулю загрузки данных есть два основных пути: добавление модуля для загрузки из распределенной файловой системы и модуля для загрузки из базы данных. Для загрузки из базы данных нужно наследоваться от класса Database, где нужно определить один метод load\_data. Данный метод должен принимать на вход название таблицы и список субъектов, для которых нужно загрузить данные, на выход должен возвращаться json, где ключом является субъект, а значением словарь с данными.

Для загрузки фМРТ изображений можно пойти двумя путями: можифицировать класс загрузки данных и класс чтения данных.