andy.petrella

Setup Instructions

Distributed Data Science Pipeline

Environment

For the training, we'll be using as the simplest environment to run most of the pipeline.

This environment will be available from a single docker image.

Prerequisites

Install a docker environment, the installation you'll have to go through is described on the docker website.

RESOURCES (WINDOWS / MAC OS X)

Docker will have to run as a service within a Virtual Machine. However, the pipeline will require a few resources (he, we're talking about many computations here).

So, please, assign quite some resources to this VM, here is the recommended setup:

- 8G memory
- 4 CPUs/Cores

Pull image

The first step is to download locally the docker image put by Data Fellas on dockerhub.

To do so, run this command in a shell

```sh
docker pull datafellas/oreilly-pipeline:1.0
```

Verify the image is available locally,

```sh
docker images
```

Check that you see a line referring `datafellas/oreilly-pipeline`.

Run

When the image is available locally we can start it using this command:

```sh
docker run --rm -it -m 8g -p 19000:9000 -p 14040:4040 -p 14041:14041 -p 14042:4042 -p 14043:14043 datafellas/oreilly-pipeline:1.0 bash
```

In essence, this command will start a container with

- `--rm` means that the container will be killed when exiting
- `-it` allows interative session
- `-m 8g` gives 8G memory to the container
- `-p $lp:$cp` exports container's port `$cp` on the local (host) port `$lp`
- `bash` puts you in a shell

Setup

When the container has been started, we're ready to start and prepare the different services

- Cassandra
- Kafka
- Spark Notebook

Three scripts in the `pipeline` have to be invoked sequentially

```sh
source var.sh
source start.sh
source create.sh
```

Access

From now on, you should have the docker image locally, a container running with all services started... so you're ready to start the Spark Notebook interface which will drive the whole pipeline end to end.

This interface is available at http://localhost:19000/notebooks/pipeline which leads you directly in the *right* notebooks folder for the training, `pipeline`.

> IMPORTANT
>
> On Windows / Mac OS X, you'll run the container in a VM, hence localhost won't work. So you'll have to use the IP of the VM instead.