Adding Third-Party Libraries to Data Flow Applications

Reference Link: [Adding Third-Party Libraries to Data Flow Applications (oracle.com)](https://docs.oracle.com/en-us/iaas/data-flow/using/third-party-libraries.htm#dependency-archive-structure)

Your PySpark applications might need custom dependencies in the form of Python wheels or virtual environments. Data Flow allows you to provide a ZIP archive in addition to your application. It is installed on all Spark nodes before launching the application. If you construct it properly, the Python libraries will be added to your runtime, and the JAR files will be added to the Spark classpath.

## Build a Dependency Archive Using the Data Flow Dependency Packager

1. Download docker.
2. Download the packager tool image using the command:

**docker pull phx.ocir.io/oracle/dataflow/dependency-packager:latest**

1. For Python dependencies, create a requirements.txt file.

**Trade\_detection\_df requirements.txt:**

-i <https://pypi.org/simple>

certifi==2020.4.5.1

cffi==1.14.0

cryptography==2.8

oci==2.14.3

pycparser==2.20

pyopenssl==19.1.0

python-dateutil==2.8.1

pytz==2020.1

six==1.15.0

**Trade\_processing\_df requirements.txt:**

-i <https://pypi.org/simple>

certifi==2020.4.5.1

cffi==1.14.0

oci==2.14.3

pyopenssl==19.1.0

python-dateutil==2.8.1

distlib==0.3.4

numpy==1.19.5

virtualenv==20.14.1

wheel==0.37.1

zipp==3.6.0

1. To create the archive, you can use docker or podman. Use the docker container to create the archive:

* Use this command if using macOS or Linux:

**docker run --rm -v $(pwd):/opt/dataflow --pull always -it phx.ocir.io/oracle/dataflow/dependency-packager:latest -p 3.6**

* If using the Windows command prompt as the Administrator, use this command:

**docker run --rm -v %CD%:/opt/dataflow --pull always -it phx.ocir.io/oracle/dataflow/dependency-packager:latest -p 3.6**

* If using Windows PowerShell as the Administrator, use this command:

**docker run --rm -v ${PWD}:/opt/dataflow --pull always -it phx.ocir.io/oracle/dataflow/dependency-packager:latest -p 3.6**

These commands create a file called archive.zip.

1. (Optional) You can add static content. You might want to include other content in your archive. For example, you might want to deploy a data file, an ML model file, or an executable your Spark program calls at runtime. Check the reference link for more information.
2. Upload archive.zip to your object store.

## Validate an Archive.zip File Using the Data Flow Dependency Packager.

You can use the Data Flow Dependency Packager to validate an archive.zip file locally, before uploading the file to Object Storage.

Navigate to the directory containing the archive.zip file, and run the following command:

**docker run --rm -v $(pwd):/opt/dataflow --pull always -it phx.ocir.io/oracle/dataflow/dependency-packager:latest --validate archive.zip**