New Data Lake/Tutorial 2 ...

Tutorial 2: Introduction to Citi Bike New York and Setup

FINISHED

This tutorial was built for BDCS-CE version 17.4.1 as part of the New Data Lake User Journey: here (https://github.com/oracle/learning-library/tree/master/workshops/journey2-new-data-lake). Questions and feedback about the tutorial: david.bayard@oracle.com (mailto:david.bayard@oracle.com)

Contents

As a reminder, the documentation for BDCS-CE can be found here (https://docs.oracle.com/cloud/latest/big-data-compute-cloud/index.html)

Took 0 sec. Last updated by anonymous at November 15 2017, 1:49:43 PM.

About Citi Bike New York City

READY

This user journey uses bike ride data available from the New York City bike share program known as Citi Bike NYC. Citi Bike consists of a fleet of bikes and a network of docking stations. Bikes can be unlocked from one station and returned to any other. Details about Citi Bike can be found here: https ://www.citibikenyc.com/ (https://www.citibikenyc.com/).

We will use Citi Bike bike trip data to illustrate some of the capabilities of BDCS-CE and its sister cloud services throughout this journey.

Parts of the journey are inspired by this analysis: http ://toddwschneider.com/posts/a-tale-of-twenty-two-million-citi-bikes-analyzing-the-nyc-bike-share-system/ (http://toddwschneider.com/posts/a-tale-of-twenty-two-million-citi-bikes-analyzing-the-nyc-bike-share-system/)

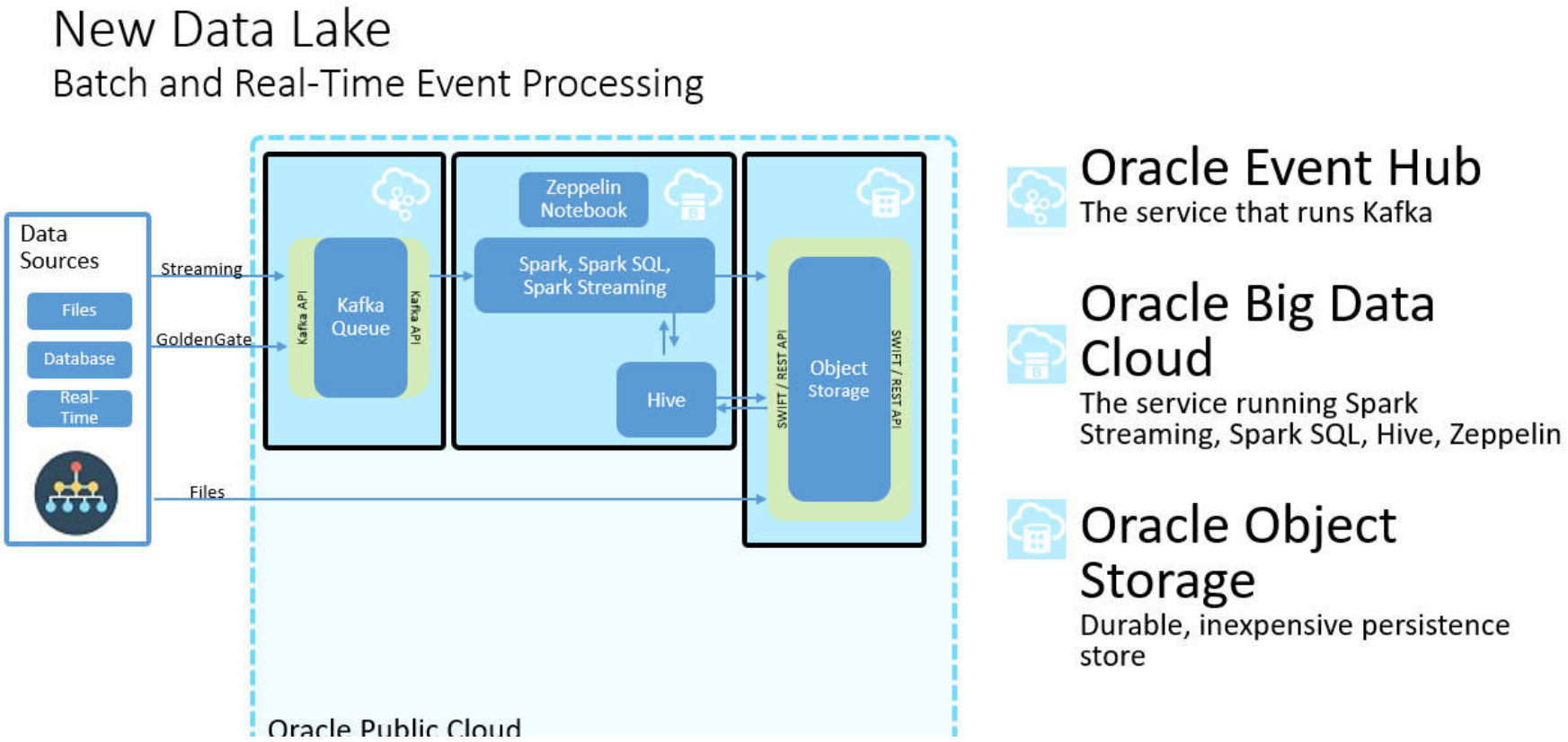
Downloading Citi Bike data and Storing in the Object Store

READY

The first step will be to download a month of data. We will get our data from https: //www.citibikenyc.com/system-data (https://www.citibikenyc.com/system-data). We will grab data for December 2016.

Once the data is downloaded, we will unzip it. Finally, we will store the unzipped data into a container in Object Store. The container does not need to exist– the code will create it if needed. By default, the container will be named "journeyC". We will put our data into a "citibike" directory within the "journeyC" container.

Here is a logical diagram of the architecture we will use in these tutorials:



In this step, we are illustrating how to move files into the Object Storage. We will store the original data file into a directory within our container called “citibike/raw”. We will create a modified version of the data file (that has the header row removed) and store that into a directory called “citibike/modified”.

Shell commands to download data and copy data to Object Storage (takes 1-3 minutes)

FINISHED

```
%sh

# the very first time you run a shell interpreter with zeppelin 0.70, the first few characters of output seem to get jumbled.
# since this paragraph is typically the first thing run, here is a silly workaround
echo "                                        "
sleep 5
echo "                                        "
# end silly workaround


CONTAINER=journeyC
DIRECTORY=citibike
FILENAME=201612-citibike-tripdata

echo "Object Storage Container Name        :" $CONTAINER
echo "Directory Name        :" $DIRECTORY
echo "Data Set name (remove .zip or .csv)  :" $FILENAME
echo "-----------------------------------------------------------------"

test -e $DIRECTORY || mkdir $DIRECTORY
cd $DIRECTORY
rm $FILENAME*

echo "Downloading $FILENAME.zip.  This may take a few minutes."
# https://www.citibikenyc.com/system-data links us to https://s3.amazonaws.com/tripdata/
wget -nv https://s3.amazonaws.com/tripdata/$FILENAME.zip
echo "Extracting the csv from the zip file"
unzip $FILENAME.zip
#head -3 $FILENAME.csv
echo "Creating a new version of the file without header information named _nh.csv"
sed '1d' $FILENAME.csv > $FILENAME.nh.csv
ls -l
```

```
Object Storage Container Name        : journeyC
Directory Name        : citibike
Data Set name (remove .zip or .csv)  : 201612-citibike-tripdata
---------------------------------------------------------------
rm: cannot remove `201612-citibike-tripdata*': No such file or directory
Downloading 201612-citibike-tripdata.zip.  This may take a few minutes.
2017-11-15 18:50:08 URL:https://s3.amazonaws.com/tripdata/201612-citibike-tripdata.zip [27546951/27546951] -> "201612-citibike-tripdata.
zip" [1]
Extracting the csv from the zip file
Archive:  201612-citibike-tripdata.zip
  inflating: 201612-citibike-tripdata.csv
Creating a new version of the file without header information named _nh.csv
total 293824
-rw-rw-r-- 1 zeppelin zeppelin 136661429 Jan 20  2017 201612-citibike-tripdata.csv
-rw-rw-r-- 1 zeppelin zeppelin 136661199 Nov 15 18:50 201612-citibike-tripdata.nh.csv
-rw-rw-r-- 1 zeppelin zeppelin  27546951 Jan 23  2017 201612-citibike-tripdata.zip
Storing both versions of the csv files to Object Storage.  This may take a few minutes.
List the directory. directory should be empty or missing
ls: `swift://journeyC.default/citibike': No such file or directory
Make the raw directory in Object Store
Copy First File to Object Store. May take a minute
```

```
Make the modified directory in Object Store
Copy Second File to Object Store.  May take a minute
Validate by listing the 2 csv files that got copied to Object Store (you should see 2 .csv files)
Found 1 items
-rw-rw-rw-   1  136661429 2017-11-15 18:50 swift://journeyC.default/citibike/raw/201612-citibike-tripdata.csv
Found 1 items
-rw-rw-rw-   1  136661199 2017-11-15 18:50 swift://journeyC.default/citibike/modified/201612-citibike-tripdata.nh.csv
done
```

Took 1 min 0 sec. Last updated by anonymous at November 15 2017, 1:50:57 PM. (outdated)

Additional setup for BDCS-CE for the Journey

READY

The next paragraph contains some commands to help setup the BDCS-CE environment with some additional tools that will be used later. In particular, it will install the “swift” command-line for interacting with the Object Store. The “swift” command line is written in Python, so we’ll first do some work on updating various Python components.

Run the following paragraph to make the needed changes.

Additional setup for BDCS-CE for the Journey

FINISHED

```
%sh
echo "Please run this paragraph to do some additional setup for our journey"

sudo whoami
#the sudo whoami output should say root.  If not, your bootstrap.sh script did not work.  The easiest fix is to start over and recreate
    instructions.
#Or you can follow the Reference document on manually running bootstrap.

# If you get a SIGTERM error before this finishes, it is another sign your bootstrap.sh script did not work.

echo "running yum to install python-setuptools. may take 5 to 10 minutes to setup yum cache initially."
echo "running easy_install"
sudo easy_install --upgrade --index-url https://pypi.python.org/simple/ pip
echo "after easy_install"
echo "use pip to upgrade setuptools"
sudo pip install --upgrade setuptools
echo "after upgrade setuptools"
echo "use pip to install python-swiftclient"
```

```
Please run this paragraph to do some additional setup for our journey
root
running yum to install python-setuptools. may take 5 to 10 minutes to setup yum cache initially.
running easy_install
Searching for pip
Reading https://pypi.python.org/simple/pip/
Best match: pip 9.0.1
Processing pip-9.0.1-py2.6.egg
pip 9.0.1 is already the active version in easy-install.pth
Installing pip script to /usr/bin
Installing pip2.6 script to /usr/bin
Installing pip2 script to /usr/bin
Using /usr/lib/python2.6/site-packages/pip-9.0.1-py2.6.egg
Processing dependencies for pip
Finished processing dependencies for pip
after easy_install
use pip to upgrade setuptools
DEPRECATION: Python 2.6 is no longer supported by the Python core team, please upgrade your Python. A future version of pip will drop su
pport for Python 2.6
/usr/lib/python2.6/site-packages/pip-9.0.1-py2.6.egg/pip/_vendor/requests/packages/urllib3/util/ssl_.py:318: SNIMissingWarning: An HTTPS
request has been made, but the SNI (Subject Name Indication) extension to TLS is not available on this platform. This may cause the serv
er to present an incorrect TLS certificate, which can cause validation failures. You can upgrade to a newer version of Python to solve t
his. For more information, see https://urllib3.readthedocs.io/en/latest/security.html#snimissingwarning.
  SNIMissingWarning
/usr/lib/python2.6/site-packages/pip-9.0.1-py2.6.egg/pip/_vendor/requests/packages/urllib3/util/ssl_.py:122: InsecurePlatformWarning: A
true SSLContext object is not available. This prevents urllib3 from configuring SSL appropriately and may cause certain SSL connections
to fail. You can upgrade to a newer version of Python to solve this. For more information, see https://urllib3.readthedocs.io/en/latest/
security.html#insecureplatformwarning.
  InsecurePlatformWarning
Collecting setuptools
  Downloading setuptools-36.7.2-py2.py3-none-any.whl (482kB)
```

```
Installing collected packages: setuptools
  Found existing installation: setuptools 0.6rc11
    DEPRECATION: Uninstalling a distutils installed project (setuptools) has been deprecated and will be removed in a future version. Th
is is due to the fact that uninstalling a distutils project will only partially uninstall the project.
    Uninstalling setuptools-0.6rc11:
      Successfully uninstalled setuptools-0.6rc11
Successfully installed setuptools-36.7.2
/usr/lib/python2.6/site-packages/pip-9.0.1-py2.6.egg/pip/_vendor/requests/packages/urllib3/util/ssl_.py:122: InsecurePlatformWarning: A
true SSLContext object is not available. This prevents urllib3 from configuring SSL appropriately and may cause certain SSL connections
to fail. You can upgrade to a newer version of Python to solve this. For more information, see https://urllib3.readthedocs.io/en/latest/
security.html#insecureplatformwarning.
  InsecurePlatformWarning
after upgrade setuptools
use pip to install python-swiftclient
DEPRECATION: Python 2.6 is no longer supported by the Python core team, please upgrade your Python. A future version of pip will drop su
pport for Python 2.6
Collecting python-swiftclient==2.7.0
/usr/lib/python2.6/site-packages/pip-9.0.1-py2.6.egg/pip/_vendor/requests/packages/urllib3/util/ssl_.py:318: SNIMissingWarning: An HTTPS
request has been made, but the SNI (Subject Name Indication) extension to TLS is not available on this platform. This may cause the serv
er to present an incorrect TLS certificate, which can cause validation failures. You can upgrade to a newer version of Python to solve t
his. For more information, see https://urllib3.readthedocs.io/en/latest/security.html#snimissingwarning.
  SNIMissingWarning
/usr/lib/python2.6/site-packages/pip-9.0.1-py2.6.egg/pip/_vendor/requests/packages/urllib3/util/ssl_.py:122: InsecurePlatformWarning: A
true SSLContext object is not available. This prevents urllib3 from configuring SSL appropriately and may cause certain SSL connections
to fail. You can upgrade to a newer version of Python to solve this. For more information, see https://urllib3.readthedocs.io/en/latest/
security.html#insecureplatformwarning.
  InsecurePlatformWarning
  Downloading python_swiftclient-2.7.0-py2.py3-none-any.whl (58kB)
Collecting futures>=2.1.3 (from python-swiftclient==2.7.0)
  Downloading futures-3.1.1-py2-none-any.whl
Requirement already satisfied: six>=1.5.2 in /usr/lib/python2.6/site-packages (from python-swiftclient==2.7.0)
Requirement already satisfied: requests>=1.1 in /usr/lib/python2.6/site-packages (from python-swiftclient==2.7.0)
Installing collected packages: futures, python-swiftclient
Successfully installed futures-3.1.1 python-swiftclient-2.7.0
after swiftclient
testing swiftclient
..
..
Usage: swift [--version] [--help] [--os-help] [--snet] [--verbose]
             [--debug] [--info] [--quiet] [--auth <auth_url>]
             [--auth-version <auth_version>] [--user <username>]
```

```
            [--key <api_key>] [--retries <num_retries>]
            [--os-username <auth-user-name>] [--os-password <auth-password>]
            [--os-user-id <auth-user-id>]
            [--os-user-domain-id <auth-user-domain-id>]
            [--os-user-domain-name <auth-user-domain-name>]
            [--os-tenant-id <auth-tenant-id>]
            [--os-tenant-name <auth-tenant-name>]
            [--os-project-id <auth-project-id>]
            [--os-project-name <auth-project-name>]
            [--os-project-domain-id <auth-project-domain-id>]
            [--os-project-domain-name <auth-project-domain-name>]
            [--os-auth-url <auth-url>] [--os-auth-token <auth-token>]
            [--os-storage-url <storage-url>] [--os-region-name <region-name>]
            [--os-service-type <service-type>]
            [--os-endpoint-type <endpoint-type>]
            [--os-cacert <ca-certificate>] [--insecure]
            [--no-ssl-compression]
            <subcommand> [--help] [<subcommand options>]
Command-line interface to the OpenStack Swift API.
Positional arguments:
  <subcommand>
    delete               Delete a container or objects within a container.
    download             Download objects from containers.
    list                 Lists the containers for the account or the objects
                         for a container.
    post                 Updates meta information for the account, container,
                         or object; creates containers if not present.
    stat                 Displays information for the account, container,
                         or object.
    upload               Uploads files or directories to the given container.
    capabilities         List cluster capabilities.
    tempurl              Create a temporary URL.
    auth                 Display auth related environment variables.
Examples:
  swift download --help
  swift -A https://auth.api.rackspacecloud.com/v1.0 -U user -K api_key stat -v
  swift --os-auth-url https://api.example.com/v2.0 --os-tenant-name tenant \
      --os-username user --os-password password list
  swift --os-auth-url https://api.example.com/v3 --auth-version 3\
      --os-project-name project1 --os-project-domain-name domain1 \
      --os-username user --os-user-domain-name domain1 \
```

```
      --os-password password list
  swift --os-auth-url https://api.example.com/v3 --auth-version 3\
      --os-project-id 0123456789abcdef0123456789abcdef \
      --os-user-id abcdef0123456789abcdef0123456789 \
      --os-password password list
  swift --os-auth-token 6ee5eb33efad4e45ab46806eac010566 \
      --os-storage-url https://10.1.5.2:8080/v1/AUTH_ced809b6a4baea7aeab61a \
      list
  swift list --lh
Options:
  --version             show program's version number and exit
  -h, --help            show this help message and exit
  --os-help             Show OpenStack authentication options.
  -s, --snet            Use SERVICENET internal network.
  -v, --verbose         Print more info.
  --debug               Show the curl commands and results of all http queries
                        regardless of result status.
  --info                Show the curl commands and results of all http queries
                        which return an error.
  -q, --quiet           Suppress status output.
  -A AUTH, --auth=AUTH  URL for obtaining an auth token.
  -V AUTH_VERSION, --auth-version=AUTH_VERSION
                        Specify a version for authentication. Defaults to 1.0.
  -U USER, --user=USER  User name for obtaining an auth token.
  -K KEY, --key=KEY     Key for obtaining an auth token.
  -R RETRIES, --retries=RETRIES
                        The number of times to retry a failed connection.
  --insecure            Allow swiftclient to access servers without having to
                        verify the SSL certificate. Defaults to
                        env[SWIFTCLIENT_INSECURE] (set to 'true' to enable).
```

Took 9 sec. Last updated by anonymous at November 15 2017, 2:22:29 PM. (outdated)

Next Steps

READY

So far, we have downloaded a Citi Bike data zip and created csv files. Then we stored the data into an Object Storage container.

In the next part of the journey, we will set up a Hive table using the data.

Change Log

FINISHED

November 15, 2017 - Tweaked easy_install to use https address for index-url
October 6, 2017 - Added small workaround for Zeppelin shell output jumbling during first execution
September 7, 2017 - Confirmed it works with 17.3.5-20
August 23, 2017 - A few minor tweaks
August 12, 2017 - Confirmed it works with 17.3.3-20
August 11, 2017 - Journey version2 changes.
July 31, 2017 - Added comments about 17.3.1-20 core-site.xml swift auth url issue that affects some users.
July 28, 2017 - Confirmed that it works with 17.3.1-20.

Took 0 sec. Last updated by anonymous at November 15 2017, 2:23:32 PM.

```
%md
```

READY