1.

from datetime import datetime, timedelta

from airflow.operators.oracle\_operator import OracleOperator

from airflow.operators.sensors import SqlSensor

default\_args={

    'start\_date':datetime(2022,1,4),

}

with DAG( dag\_id='quiz2\_1',default\_args=default\_args,schedule\_interval='@monthly',catchup=False) as dag :

    t1 = SqlSensor(

        task\_id='check\_data\_arrived',

        conn\_id="oracle\_conn",

        sql="SELECT COUNT(\*) ceyhandb.pipeline;",

        poke\_interval=600,

        timeout=1800

)

sqoop import # bu komutu hangi operatörde kullanacğımı bilemedim

--connect "jdbc:oracle:........"

--username "root "

--password "cloudera"

--hive-import

--query "select \* from ceyhandb.pipeline where pdate=SELECT TO\_DATE(current\_date - 1) AS yesterday\_date From Dual"

--hive-table "......"

--hive-partition-key "pdate "

--target-dir /user/sqoop\_import/"....."

--split-by id

2-)

from airflow import DAG

from datetime import datetime, timedelta

from airflow.operators.oracle\_operator import OracleOperator

from airflow.operators.sensors import SqlSensor

default\_args={

    'start\_date':datetime(2020,5,20),

}

with DAG( dag\_id='quiz2\_1',default\_args=default\_args,schedule\_interval='@daily',catchup=False) as dag :

    t1 = SqlSensor(

        task\_id='check\_data\_arrived',

        conn\_id="oracle\_conn",

        sql="SELECT COUNT(\*) main.a1;",

        poke\_interval=600,

        timeout=1800

)

sqoop export --connect \

jdbc:jdbc:oracle:................ \

 --table elvan1.a1 \

 --username root --password cloudera \

 --export-dir /user/hive/warehouse/main.db/a1\

 --input-fields-terminated-by ','

3.  Local filesystemimizde(lfs) /a/b/final  konumuna her gün düzenli dosya akışı olmaktadır. Biz de bu dosyaları başka bir sisteme/makinaya - local filesysteme göndereceğiz.

a.  İlgili dagi yazın.(başka bir locale sahip olmadığınız çalışmayabilir operatör ve parametrelerin doğru olması yeterlidir.)

b.  Eğer /a/b/final  konumunda data yoksa Linux admin ekibine mail atsın (Linux\_admin@dilisim.com) (Mail komutunuz çalışmayabilir operatör ve parametrelerin doğru olması yeterlidir.)

from airflow import DAG

from datetime import datetime,timedelta

from airflow.contrib.operators.sftp\_operator import SFTPOperator

from airflow.contrib.hooks.ssh\_hook import SSHHook

from airflow.contrib.sensors.file\_sensor import FileSensor

from airflow.operators.email\_operator import EmailOperator

default\_args = {

    'start\_date':datetime(2020,5,20),

    "email": ['Linux\_admin@dilisim.com'],

    "email\_on\_failure": False,

    "email\_on\_retry": False,

}

with DAG(dag\_id='quiz2\_3',default\_args=default\_args,schedule\_interval='@daily',catchup=False) as dag:

    transfer\_data\_local\_to\_local = SFTPOperator(

    task\_id="local\_to\_local",

    ssh\_conn\_id="ssh\_default",

    local\_filepath="/a/b/final",

    remote\_filepath="/a/b/final",

    operation="put",

    create\_intermediate\_dirs=True

    )

    check\_file\_exists = FileSensor(

        task\_id='check\_file\_exists',

        filepath='/a/b/final',

        fs\_conn\_id='fs\_default',

        poke\_interval=5,

        timeout=25,

    )

    send\_email = EmailOperator(

        task\_id='send\_email',

        to='Linux\_admin@dilisim.com',

        subject='Dosya mevcut degil',

        html\_content=""" <h1> DOSYA MEVCUT DEGIL </h1> """,

        trigger\_rule='one\_failed'

    )

    transfer\_data\_local\_to\_local>>check\_file\_exists>>send\_email

4-)Günlük çalışan bir dagimiz var.

a.  Tüm dagi clear ettiğimde aynı anda en fazla 2 gün çalışsın istiyorum.

b.  Dagin günlük değil de her cumartesi saat aksam 5te çalışsın istiyorum.

a-) tüm dagler için aynı anda en fazla 2 dag run edilsin istiyorsam airflow.cfg dosyasındaki max\_acitve\_runs\_per\_dag paramatresi varsayılan olarak 16 bunu 2 yapmam gerekir diğer türlü sadece tek bir dag için istiyorsam bunu dagi tanımlarken max\_active\_runs=2 demem lazım.

catchup paramatresinin de true olması lazım

b-)Schedule\_interval = 00 17 \* \* 6 yapmam gerekir

5-)Başka bir dagde çalışan sqoop işinin hangi sisteme gidip hangi tabloyu aldığını görmek istiyorum. Ama dag’in kodunu okuyamıyorum , okusamda zaten parametrik (dümdüz hardcoded yazılmamış yani) ve elimde sadece tree ekranım var.

a.  İlgili db-tablo adına nasıl erişebilirim?

b.  Çalışan sqoop sorgusuna nasıl erişebilirim?

Sqoop işini yapan taskin log veya rendered kısmına bakılabilir

6-)Elinizde 5 adet task var. A1,a2,a3,a4,a5 . A5 ilk çalışmalı. Daha sonra a3 ve a4 birbirinden bağımsız çalışmalı. A4 bittiği an a2 ve a1 birbirinden bağımsız çalışmalı.

a.  Tek satırda

b.  Her bağlamayı farklı bir satırda.

from airflow  import DAG

from datetime import datetime, timedelta

from airflow.operators.dummy\_operator import DummyOperator

default\_args = {

    "start\_date": datetime(2022, 1, 8)

}

with DAG( dag\_id='quiz2\_6',default\_args=default\_args,schedule\_interval='@daily',catchup=False) as dag :

    a5=DummyOperator(task\_id='a5')

    a4=DummyOperator(task\_id='a4')

    a3=DummyOperator(task\_id='a3')

    a2=DummyOperator(task\_id='a2')

    a1=DummyOperator(task\_id='a1')

    #a5>>[a4,a3]

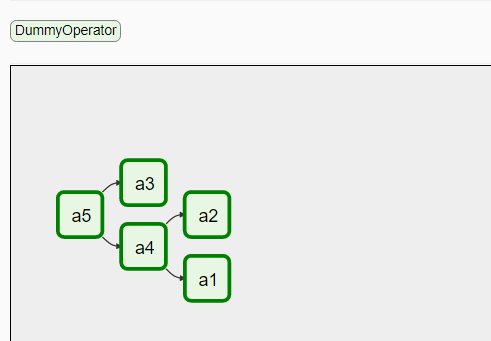
    #a4>>[a2,a1]

    a3.set\_upstream(a5)

    a4.set\_upstream(a5)

    a2.set\_upstream(a4)

    a1.set\_upstream(a4)



7.  Her gece çalışan sqoop data export işiniz sqoop connectionlarında oluşan bir sıkıntıdan dolayı retry durumuna düşüyor tablo aktarılmadan. Ancak tablo bir kere karşı sqoop dbde oluşturulduğu için sonraki denemelerinizde table already exist hatası alıyorsunuz?

a.  Çözüm önerileriniz nelerdir? Min 2 yorum yeterli.

Eğer daha önce oluşturduğumuz bir tabloyu tekrardan oluşturmaya çalışıyorsak bu hatayı alırız tabloyu oluştururken if not exists ifadesini kullanabiliriz

Tabloyu oluşturma query sini tek bir defa çalıştırabiliriz

Tabloyu silip tekrar oluşturabiliriz

Cretae recreate paramatreleri true olarak verilebilir

8.  Sisteminizden başka bir oracle dbye (y db diyelim) tablo aktaracaksınız. Datanız günlük olarak x sisteminden sizin sisteminize geliyor ve tablonuzu önce truncate edip sonra 300m veri yüklüyor ancak ne zaman geldiğini bilmiyorsunuz. Y dbsine tabloyu gönderecek olan iş her gün düzenli olarak çalışacak.  Ancak işin 4. Günde hata aldığını gördünüz çünkü runtime’da tablonuz boş (data sisteme gelmiş ancak tabloya yüklenmemiş. 300 m data 1 saniyede yüklenecek değil ya).

a.  Bunun önüne nasıl geçersiniz? 2 farklı fikir yeterli.

SqlSensor operatörünü kullanabilirim 300m datanın ne zaman yüklenceğini tahmin ediyorsam poke ve timeout değerlerini ona göre ayarlarım

Sql sensorde query olarak Select count(\*) from tablename bu sayede çıktının 0 olmadığını anlarım

Sqlsensor taskinden sonra y tablosuna datayı gönderirim

9-)

distcp\_dlsm = BashOperator(dag=dag, task\_id=’distcp\_dlsm’,

                             bash\_command=’hadoop–config ‘ + distcp\_conf\_dir +

                                          ‘distcp -Dmapreduce.job.hdfs-servers.token-renewal.exclude=dlsm ‘

                         ‘-D ipc.client.fallback-to-simple-auth-allowed=true -skipcrccheck -update ‘

                                          ‘/data/domain/dilisim/dilisim.db/abcd\_ldm/pdate={{ ds\_nodash }} ‘

                                          ‘hdfs://data/domain/dilisim/input/abcd\_bigdata/pdate={{ds\_nodash }}’)

Usage: hadoop [--config confdir] [--loglevel loglevel] [COMMAND] [GENERIC\_OPTIONS] [COMMAND\_OPTIONS]

Tüm gerekli yapılandırma tamamlandıktan sonra, dosyaları tüm makinelerdeki HADOOP\_CONF\_DIR dizinine dağıtın, varsayılan olarak olarak ${HADOOP\_HOME}/conf .

**Distcp**

**Dosya ve dizinleri yinelemli kopyalamak için kümeler arası kopyalama için kullanılır**

**-D özellik vermek için**

2-  default\_args = {

    ‘depends\_on\_past’: False,

    ‘email’: operation\_mail\_list,

    ‘email\_on\_failure’: True,

    ‘email\_on\_retry’: True,

    ‘retries’: 3,

    ‘retry\_delay’: timedelta(minutes=10),

}

dag de buluınan tasklerde aksi belirtilmedikçe yukarıdaki paramatreler geçerli olacak

depends\_on\_past’: False, kendinden önceki dag rundaki task ile ilgili

diyelimki günlük çalışan dagim var 5.günde dagdeki task x fail oldu bu dag 6.gün de de çalışacak eğer bu paramatre true olsaydı 6.gündeki task x de çalışmıcaktı, false olduğu için kendinden önceki dag runlar daki taskın durumuna bakmadan çalışır

emai: paramatresi de bir task fail veya retry durumuna düştüğünde hangi maile alert gidecek

email\_on\_failure: True eğer task fail durumuna düşerse emailde belirtilen mail adresine alert gönderilcek

email\_on\_retry: task retry durumuna düştüğünde emailde belirtilen adrese alert gönderilcek

retries: bir task fail duruman düşmeden önce kaç defa tekrar run edilcek için denencek

retry\_delay: kaç dakika aralıklarla denencek

3-  tableops = HiveOperator(

    task\_id=’tableops’,

    dag=dag,

    hive\_cli\_conn\_id=’dlsm\_hiveserver,

    hql=’dlsm.hql’

)

Hive Operatorleri hql yazmamızı sağlar

Task\_id= her taşkın idsi olmalı ve uniqe olmalı

Dag=hangi dagin taskı

Hive\_cli\_conn\_id = hive için connection id si

hql=’dlsm.hql’ dlsm.hql dosyasındaki hql query yi çalıştırır

 dilisimhrtable = BashOperator(

task\_id=’dilisimhrtable’,

             dag=dag, bash\_command=’/usr/bin/impala-shell –impalad dilisim001:21000 -k –ssl –delimited –output\_delimiter=”,”  --quiet -q  ‘alter table  dilisim.hr drop if exists;’

BashOperator :Bash komutları yazmamızı sağlar

Bash\_command = bash komutu yazdığımız paramatre

/usr/bin/impala-shell –impalad dilisim001:21000 -k –ssl impalayı noninteractive modda çalıştırır interactive modda ve non interactive modda çalıştırmanın farkı

İntteractive modda impalaya bağllandıktan sonra query i yazarsın çıkışı da kendin sağlarsın

Noninteractive modda bağlantıyı ve queryi tek seferde çalıştırırsın bu modda çıkış yapmana gerek yok

--quiet :çıktıda info mesajlarını vermez

-q: impala query si yazmamızı sağlar

alter table  dilisim.hr drop if exists : dilişim dbsinde hr tablosu varsa onu siler