2024-06-13

오전

1. producer의 네트워크를 nat->어댑터에 브릿지로 변경
2. sudo vim /etc/hosts를 들어가서 ip설정
3. scp datanode1:/home/hadoop/hadoop/etc/hadoop/ */home/hadoop/hadoop/etc/hadoop/*
   known.hosts 에 등록이 된다. yes를 한애들은 저파일에 저장이 되는데, 그런데 ip는 같은데 내용이 다르 거부를 일으킨다.
   ssh-keygen -f "/home/hadoop/.ssh/known_hosts" -R "datanode1"
   회사에서는 저렇게 해야한다
   rm known_hosts
   ssh datanode1
   다시가서 datanode1의 know_hosts에 저장하고
   exit가서 다시 producer로 간다음
   scp datanode1:/home/hadoop/hadoop/etc/hadoop/ /home/hadoop/hadoop/etc/hadoop/
   다시 실행하면된다
4. hdfs dfs -mkdir /jotaesik 폴더만들어보기
5. hdfs dfs -mkdir /mort
6. producer에서 jupyter-notebook실행하기
7. ip:port에 들어갈때 ip는 내 ip를 쳐야한다
8. nohup jupyter notebook --ip=0.0.0.0 &
   터미널을 꺼도 백그라운드에서 돌아간다. 하지만 프로세서 메모리를 잡아먹는다
   kill -9 pid
   9.hdfs dfs -put ./*.csv /이름**
   hdfs dfs -put
   /home/hadoop/hadoop_test1/demo/src/main/java/com/example/tpss_bcycl_od_statnhm_202001.csv /mort

free -h로 남은 사용량을 알수있다.

producer를 8192mb로 바꾼다.
캐시 - 책상과 책꽂이의 그 중간 가장 비싼 메모리
1차캐시,2차캐시,3차캐시
스왑 - 지금 당장 안필요한걸 갖다놓는행위

file_path = '/encore/tpss_bcycl_od_statnhm_202001.csv'
import pandas as pd

with hdfs.open_input_file(file_path) as f:
#table = csv.read_csv(f, encoding='cp949')
df = pd.read_csv(f, encoding='cp949')

2024.06.13 ipynb 파일참조

32기가 노트북이상을사자\

producer에 다운받기
wget https://dlcdn.apache.org/hive/hive-4.0.0/apache-hive-4.0.0-bin.tar.gz

압축풀기

tar -zxvf 압축풀기

폴더명을 hive로 바꾸기

mv apache-hive-4.0.0-bin hive

hadoop의 hive 폴더 데이터엔지니어링폴더

파일질라를 통해서

/home/hadoop/hive/lib -다운받을 jar pip파일들
/home/hadoop/hive/conf - 환경변수파일 xml

vim ~/.bashrc
export HIVE_HOME=/home/hadoop/hive
export PATH=$PATH$:HADOOP_HOME/sbin:
$HADOOP_HOME/bin : /home/hadoop/hadoop/lib/native$:HIVE_HOME/bin

그리고 source ~/.bashrc
vim hive-site.xml에 hiveid hivepw가 있다.
순서가 되게 중요하다!!!
문제가 생기면 echo $PATH로 확인해보자

CREATE EXTERNAL TABLE stock(time STRING, contract_price STING, prev_diff STRING, sell STRING, trading STRING, volume STRING) STORED AS PARQUET LOCATION '/encore_test/stock.parquet';

hive

hue apache 쿼리를 보내면 결과를 보여준다.
zeppline apache 쿼리 날리는곳

heidi에서도 접속이된다 secondnode

hive에서 쿼리를 보내면
바로 mapreduce를 연결시켜서 일을한다.

hdfs는 배치성으로 많이쓴다 실시간은 약하므로, 그래서 에어플로우로종종

hive가 접속이 안되므로 버전을3으로한다

hive 3.x 기준

wget https://dlcdn.apache.org/hive/hive-3.1.3/apache-hive-3.1.3-bin.tar.gz
tar xvfz apache-hive-3.1.3-bin.tar.gz

이름 변경

hive3으로 바꿈

sudo apt install openjdk-8-jdk -y
hive/conf/hive-env.sh 밑에
JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64/ 추가

그리고 다시 lib안의 jar파일과 xml파일은 conf에
vim .bashrc에서
hive_home을 hive3로 바꿔기
그리고 source .bashrc

테스트해보기

```
CREATE EXTERNAL TABLE IF NOT EXISTS bicycle (
a STRING,
b STRING,
c STRING,
d STRING,
e STRING,
f STRING,
q STRING,
o INT,
r INT,
q1 INT
)

ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
LINES TERMINATED BY '\n'
```

STORED AS TEXTFILE
LOCATION '/encore';

ALTER TABLE bicycle SET TBLPROPERTIES('serialization.encoding'='euc-kr');

```
hive> select * from bicycle limit 5;
OK
기준_날짜       구분코드        기준_시간       시작_대여소ID   시작_대여소명   종료_대여소ID   종료_대여소명   NULL   N
ULL     NULL
20200101        0       0000    ST-443  성산2동_041_2   ST-82   성산2동_041_3   1       2       450
20200101        0       0000    ST-56   여의동_005_10   ST-1321 영등포동_036_1  1       10      1460
20200101        0       0000    ST-1546 조원동_004_1    ST-703  신사동_023_1    1       5       700
20200101        1       0000    ST-1701 당산2동_065_1   ST-1701 당산2동_065_1   1       28      70670
Time taken: 1.822 seconds, Fetched: 5 row(s)
```

오후

다시 hive4로 하기

xml 은 conf로
jar은 lib으로

cmd 3개켜서하기기
hive --service metastore
hive --service hiveserver2
hive

!connect jdbc:hive2://localhost:10000/default;auth=noSasl hive password
org.apache.hive.jdbc.HiveDriver

안되는 와중 다양한 시도중 왜 mv hive4를 hive로 바꾸면 되는것인가?
버전은 왜 17이면 안되는건가 11로 바꿔야한느건가?

자바 11설치
hive로 들어가면 3은 hive였지만 4는 beeline이 되어야한다.

```
beeline> !connect jdbc:hive2://localhost:10000/default;auth=noSasl hive password org.apache.hive.jdbc.HiveDriver
Connecting to jdbc:hive2://localhost:10000/default;auth=noSasl
Connected to: Apache Hive (version 4.0.0)
Driver: Hive JDBC (version 4.0.0)
Transaction isolation: TRANSACTION_REPEATABLE_READ
0: jdbc:hive2://localhost:10000/default>
```

아니다 트래픽이 많아서 그런건가? 자바 버전 17도 가능하다 그런데 connet가안된다

하둡의 관리자가 필요하다 ambari 모니터링 관리 하둡에코시스템

데이터모으기

ssh hadoop@secondnode
/home/hadoop/encore_data/

amazon

http://jmcauley.ucsd.edu/data/amazon/