목차

<빅데이터 분석 시스템 설계 및 개발 - 환율 차익 계산>

1. 개요

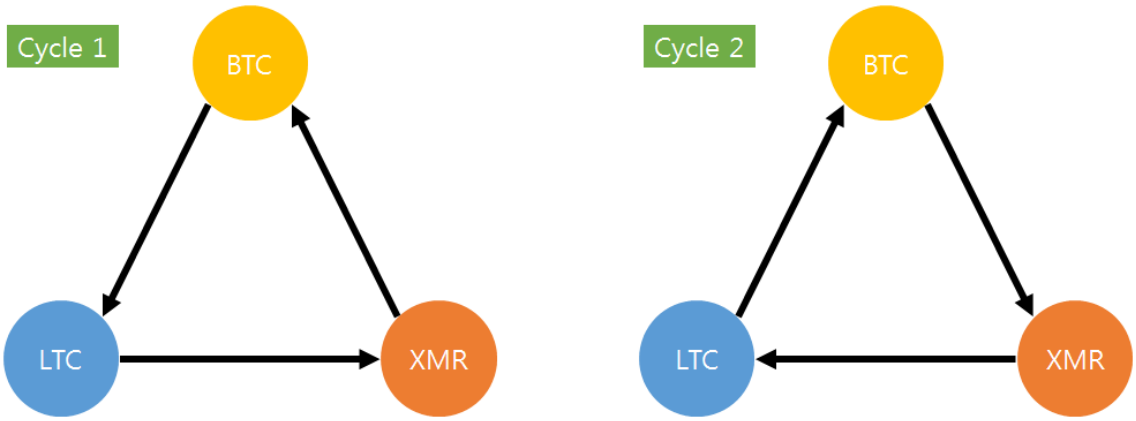
Hadoop 기반의 빅데이터 처리 기술을 활용하여 대용량의 데이터를 수집, 분석, 활용 할 수 있는 시스템을 개발하라. 필요에 따라 Hadoop, HDFS, MapReduce, Spark, Pig, Hive, Kafka, Flume, Sqoop, RDBMS, NoSQL 등 다 양한 도구나 서비스를 활용가능하다.

2. 문제 정의

(1) 프로젝트의 명확한 주제와 목표

- **Triangular Arbitrage (삼각 차익 거래) :** Triangular arbitrage is the result of a discrepancy between three foreign currencies that occurs when the currency's exchange rates do not exactly match up

(출처 : https://www.investopedia.com/terms/t/triangulararbitrage.asp)



- 크롤링을 통한 실시간 환율 정보를 이용해 삼각 차익 거래 방식에 적용한다.

(2) 주제 선정 계기

- 현재 명지대학교 주식투자동아리 'MIRS' 활동 중

- 여러 투자 방법에 대해 공부하다가 '삼각 차익 거래'라는 개념을 알게 되었고, 큰 리스크 없이 수익을 낼 수 있는 구조에 관심이 생겨 이번 프로젝트의 주제로 선정하게 되었다.

3. 시스템 아키텍쳐

(1) 수집 데이터

데이터 수집기간 : 2019-11-24 21:59:27 ~ 2019-12-11 10:03:31

데이터 수집주기 : 컴퓨터가 켜져 있는 한, 16초 마다 수집

전체 데이터 크기 : 890,439 row / 72.4 MB

USD/KRW	1,180.70	1,182.70	1,178.00	1,182.45	3.61	0.31%	2019-12-02 04:50:44.609995
USD/JPY	109.51	109.52	109.4	109.68	0.01	0.01%	2019-12-02 04:50:44.609995
JPY/KRW	10.7807	10.7999	10.7535	10.8011	0.0344	0.32%	2019-12-02 04:50:44.609995
EUR/USD	1.1015	1.1019	1.0982	1.1028	0.001	0.09%	2019-12-02 04:50:44.609995
AUD/USD	0.6764	0.6766	0.6754	0.6781	-0.0003	-0.04%	2019-12-02 04:50:44.609995
BTC/USD	7,363.80	7,363.90	7,288.60	7,613.70	-180	-2.39%	2019-12-02 04:50:44.609995
ETH/USD	150.85	150.86	146.74	153.55	-0.57	-0.38%	2019-12-02 04:50:44.609995
EUR/KRW	1,300.54	1,303.22	1,295.52	1,302.66	5.16	0.40%	2019-12-02 04:50:44.609995
GBP/USD	1.2925	1.293	1.288	1.2948	0.0016	0.13%	2019-12-02 04:50:44.609995
EUR/JPY	120.63	120.68	120.41	120.78	0.12	0.10%	2019-12-02 04:50:44.609995
USD/CHF	0.9998	1.0004	0.9981	1.0024	0.0018	0.18%	2019-12-02 04:50:44.609995
USD/CAD	1.3271	1.3276	1.327	1.3315	-0.0003	-0.03%	2019-12-02 04:50:44.609995
GBP/JPY	141.55	141.59	141.13	141.74	0.19	0.13%	2019-12-02 04:50:44.609995
NZD/USD	0.642	0.6424	0.6409	0.6439	0.0005	0.08%	2019-12-02 04:50:44.609995
AUD/JPY	74.06	74.11	73.92	74.28	-0.03	-0.03%	2019-12-02 04:50:44.609995
USD/CNY	7.0308	7.0341	7.019	7.0355	-0.0022	-0.03%	2019-12-02 04:50:44.609995
USD/RUB	64.3195	64.3195	64.02	64.3414	0.246	0.38%	2019-12-02 04:50:44.609995
USD/BRL	4.2364	4.2369	4.1835	4.2458	0.0466	1.11%	2019-12-02 04:50:44.609995
USD/TRY	5.746	5.748	5.736	5.767	-0.0035	-0.06%	2019-12-02 04:50:44.609995
XAU/USD	1,466.18	1,466.78	1,453.13	1,466.59	0.3	0.02%	2019-12-02 04:50:44.609995
USD/KRW	1,180.70	1,182.70	1,178.00	1,182.45	3.61	0.31%	2019-12-02 04:50:58.724190
USD/JPY	109.51	109.52	109.4	109.68	0.01	0.01%	2019-12-02 04:50:58.724190
JPY/KRW	10.7807	10.7999	10.7535	10.8011	0.0344	0.32%	2019-12-02 04:50:58.724190
EUR/USD	1.1015	1.1019	1.0982	1.1028	0.001	0.09%	2019-12-02 04:50:58.724190
AUD/USD	0.6764	0.6766	0.6754	0.6781	-0.0003	-0.04%	2019-12-02 04:50:58.724190
BTC/USD	7,363.80	7,363.90	7,288.60	7,613.70	-180	-2.39%	2019-12-02 04:50:58.724190
ETH/USD	150.85	150.86	146.74	153.55	-0.57	-0.38%	2019-12-02 04:50:58.724190
EUR/KRW	1,300.54	1,303.22	1,295.52	1,302.66	5.16	0.40%	2019-12-02 04:50:58.724190
GBP/USD	1.2925	1.293	1.288	1.2948	0.0016	0.13%	2019-12-02 04:50:58.724190
EUR/JPY	120.63	120.68	120.41	120.78	0.12	0.10%	2019-12-02 04:50:58.724190
USD/CHF	0.9998	1.0004	0.9981	1.0024	0.0018	0.18%	2019-12-02 04:50:58.724190
USD/CAD	1.3271	1.3276	1.327	1.3315	-0.0003	-0.03%	2019-12-02 04:50:58.724190
GBP/JPY	141.55	141.59	141.13	141.74	0.19	0.13%	2019-12-02 04:50:58.724190
NZD/USD	0.642	0.6424	0.6409	0.6439	0.0005	0.08%	2019-12-02 04:50:58.724190
AUD/JPY	74.06	74.11	73.92	74.28	-0.03	-0.03%	2019-12-02 04:50:58.724190
USD/CNY	7.0308	7.0341	7.019	7.0355	-0.0022	-0.03%	2019-12-02 04:50:58.724190
USD/RUB	64.3195	64.3195	64.02	64.3414	0.246	0.38%	2019-12-02 04:50:58.724190
USD/BRL	4.2364	4.2369	4.1835	4.2458	0.0466	1.11%	2019-12-02 04:50:58.724190
USD/TRY	5.746	5.748	5.736	5.767	-0.0035	-0.06%	2019-12-02 04:50:58.724190
XAU/USD	1,466.18	1,466.78	1,453.13	1,466.59	0.3	0.02%	2019-12-02 04:50:58.724190
USD/KRW	1,180.70	1,182.70	1,178.00	1,182.45	3.61	0.31%	2019-12-02 04:51:12.859979
USD/JPY	109.51	109.52	109.4	109.68	0.01	0.01%	2019-12-02 04:51:12.859979

(2) 전체적인 구성

- virtualbox상에 hadoop 가상환경 구축
- hadoop 가상 환경위에서 watch 명령어를 이용해 크롤링 소스코드를 10초 주기마다 반복
- subprocss 모듈을 사용하여 한번 파일이 실행될 때마다, 생성 후 update된 csv파일을 자동으로 hdfs에 upload
- hive를 이용한 분석

4. 데이터 수집 방법

(1) 소스 코드

```
import subprocess
import requests
from bs4 import BeautifulSoup
import pandas as pd
from datetime import datetime

# 접속설정
req =  "https://kr.investing.com/currencies/live-currency-cross-rates"

# investiong.com 특성상 ban 당하므로 헤더 정보 추가
header={'User-Agent':'Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36
(KHTML, like Gecko) Chrome/41.0.2227.0 Safari/537.36'}
result = requests.get(req, headers=header)
html = result.text

# BeautifulSoup 설정
bs = BeautifulSoup(html, "html.parser")

# 배열 선언
column_list = ['CURRENT_NAME','BID', 'ASK', 'LOW', 'HIGH', 'PC',
'PCP','RECORD_TIME']
currency_list=[]
bid_list = []
ask_list = []
low_list = []
high_list = []
pc_list = []
pcp_list = []
record_time = []
now = datetime.now()


# 데이터 가져오기
for a in range(20):
    columns = bs.findAll("div", "topBox") # div 중에 class 이름이 topbox 인것을 모두
찾아라.
```

```
    title = columns[a].findAll('a')# columns 중에 a 태그를 찾아라.
    currency_name = title[0].text

    data = bs.findAll("div", "contentBox")
    content = data[a].findAll('div')
    bid = content[2].text # 홀수번호는 글자포함, 짝수번호는 데이터만
    ask = content[5].text

    content1 = data[a].findAll('span')
    low = content1[0].find('i').text
    high = content1[1].find('i').text
    pc = content1[2].text
    pcp = content1[3].text

    currency_list.append(currency_name)
    bid_list.append(bid)
    ask_list.append(ask)
    low_list.append(low)
    high_list.append(high)
    pc_list.append(pc)
    pcp_list.append(pcp)
    record_time.append(now)

df_1 = pd.DataFrame(currency_list)
df_2 = pd.DataFrame(bid_list)
df_3 = pd.DataFrame(ask_list)
df_4 = pd.DataFrame(low_list)
df_5 = pd.DataFrame(high_list)
df_6 = pd.DataFrame(pc_list)
df_7 = pd.DataFrame(pcp_list)
df_8 = pd.DataFrame(record_time)

result = pd.concat([df_1,df_2, df_3, df_4, df_5, df_6, df_7, df_8], axis=1)
result.columns = [column_list]

# csv 로 저장
result.to_csv("real_time_exchange_rate.csv", mode = 'a', header=False)

subprocess.call('hadoop fs -copyFromLocal * big_data', shell = True)
subprocess.call('hadoop fs -put -f * big_data', shell = True)
```

(2) 수집 방법 / 전략

- **beautifulSoup 모듈을 사용하여 크롤링**

- **pandas 모듈**을 사용하여 각각의 데이터가 들어있는 list들을 하나의 DataFrame으로 생성

- **Linux 명령어 'watch' :** 'watch' 명령어는 원하는 명령어의 결과를 원하는 시간(초) 주기로 리프레시 하여 결과를 보여주는 명령어이다.

  ex) watch -n 10 python file.py -> file.py를 10초의 interval로 반복 실행

- **파이썬 모듈 subprocess 사용** : 파일이 실행되면 csv파일이 생성이 되는데 이 파일을 어떻게 하면 자동으로 hadoop에 upload할 수 있을까 고민하다가 이 방법을 선택하게 되었다. 파일을 처음 실행 시, -copyFromLocal이 포함되어 있는 문장을 수행하고 그 다음에 실행할때는 해당 문장을 주석처리 후, 아래의 -put -f가 포함되어 있는 문장을 수행하여 덮어쓰기 한다.

- **ssh maria_dev@localhost -p 2222로 접속** 하여 'data'라는 폴더를 생성한다. 해당 폴더에 .py파일을 옮겨 놓은 후, **watch 명령어**를 실행한다. 이 과정에서 **request, pandas와 같은 라이브러리들을 추가로 yum install** 해야 한다.

- **서버시간을 확인해봐야 하는데**, 만약 서버시간과 시간이 맞지 않는다면 **sudo rdate -s time.bora.net**을 이용해 **타임서버를 변경**해주면 된다. 실시간 환율정보이므로 시간이 굉장히 중요하기 때문에 프로그램 수행 전 꼭 확인해야 한다.

5. 데이터 분석 방법

(1) Hive 이용

- 테이블 생성 및 데이터 로드

```
DROP TABLE exchange_rate;
CREATE EXTERNAL TABLE IF NOT EXISTS exchange_rate(
  idx INT,
  currency_name STRING,
  bid double,
```

```
    ask double,
    low double,
    high double,
    pc double,
    pcp STRING,
    time timestamp
)
ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
LOCATION '/user/maria_dev/big_data/real_time_exchange_rate'
tblproperties("skip.header.line.count"="1");
```

(2) 데이터 분석 (2019-11-24 21:59:27 ~ 2019-12-11 10:03:31기간 동안의 평균값 계산)

QUERY :

```
SELECT currency_name, avg(regexp_replace(bid,',','')) as bid_avg,
        avg(regexp_replace(ask,',','')) as ask_avg
FROM exchange_rate2
GROUP BY currency_name
```

-> 매수, 매도 비용의 부분이 천의 단위로 구분이 되어있는 ','로 인해 평균값이 null로 나오는 현상이 나타나 regexp_replace를 사용하여 ','를 공백으로 변경함.

6. 데이터 분석 결과

currency_name	bid_avg	ask_avg
AUD/JPY	74.18413961312926	74.22375305381163
AUD/USD	0.6795854643757874	0.6797830301576575
BTC/USD	7473.522262805938	7473.799321558844
ETH/USD	150.61892042132584	150.6415874083875
EUR/JPY	120.47236309022979	120.52550114147014
EUR/KRW	1305.1245103126396	1307.022548119716
EUR/USD	1.1035618719213953	1.1038523248829284
GBP/JPY	141.68447578997433	141.73918819336754
GBP/USD	1.2978616844882789	1.2982171636833733
JPY/KRW	10.831382511913972	10.846318886620816
NZD/USD	0.646874196003033	0.6472069550244579
USD/BRL	4.216018426047194	4.216605579720304
USD/CAD	1.326464829988551	1.3268096888143068
USD/CHF	0.995273840362314	0.9956756081545142
USD/CNY	7.035499793346036	7.037619610714662
USD/JPY	109.1673072209291	109.18855556889639
USD/KRW	1182.6038121899576	1184.0748569756554
USD/RUB	64.07078621650216	64.07078621650216
USD/TRY	5.7569463158324234	5.759924501584472
XAU/USD	1463.4549900278623	1463.862454002932

<분석 결과 활용 1>

-> 구매시점에 평균 ask보다 가격이 낮다면 구매하고, 판매시점에 평균 bid보다 높다면 판매한다.

<분석 결과 활용 2>

*** 돈을 환전하는데 드는 수수료는 없다고 가정한다.**

a. EUR -> KRW

(1) 현물 거래 시 : EUR/KRW = 1307.023원

(2) 삼각 차익 거래

EUR/USD = 1.104 , USD/JPY = 109.189 , JPY/KRW = 10.846

$$\frac{EUR}{\cancel{USD}} \times \frac{\cancel{USD}}{\cancel{JPY}} \times \frac{\cancel{JPY}}{KRW}$$

=> 1.104 * 109.189 * 10.846 = 1307.427

=> 적정가가 현물가 보다 높기 때문에 구매하지 않는다.

b. ETH -> KRW

(1) 현물 거래 시 : ETH/KRW = 169,500원

(2) 삼각 차익 거래

ETH/USD = 150.642 , USD/EUR = 0.906 , EUR/KRW = 1307.023

$$\frac{ETH}{\cancel{USD}} \times \frac{\cancel{USD}}{\cancel{EUR}} \times \frac{\cancel{EUR}}{KRW}$$

=> 150.642 * 0.906 * 1307.023 = 157,603원

=> 적정가가 현물가 보다 낮기 때문에 구매한다.

=> 이 경우, 삼각 차익 거래 방식을 이용했을 때 1 ETC당 약 11897원의 수익을 낼 수 있다.

7. 추가적인 확장 가능성

- 더 많은 종류의 환율정보를 실시간으로 수집하여 삼각차익거래 뿐 아니라 다양한 투자 방법에 적용할 수 있다.

- 서버위에서 수행할 시에, 더 연속적인 데이터 확보가 가능하다.