미니 프로젝트_hadoop_pig 데이터 전처리

상태	완료
소유자	㉦ 재혁 최
날짜	@2023년 7월 25일 → 2023년 7월 25일
요약	kaggle에서 가져온 BTC_USD.csv를 pig로 전처리 데이터는 날짜(chararray), 시가(float), 고가(float), 저가(float), 종가(float), 조정후 종가(float), 거래량(long) 으로 구성되어 있음. 목표 전처리 : 1. 연도별 가격 평균 종가 2. 월별 가격 평균 종가 3. 연도별 가격 변화율 = (연말 종가 - 연초 시가)/연초 시가 * 100 4. 월별 가격 변화율 = (월말 종가 - 월초 시가)/월초 시가 * 100 5. 하루 평균 변동성 = (high - low)/low * 100 6. 특정 거래량(거래량 : 하루동안 거래된 코인 수)별 가격 변화율 (ex 볼륨 최소, 최대 뽑아보고, 구간별 나눠서 패턴을 찾을 수 있게) 7. [시간남으면 pig 문법들 사용해보기]

프로젝트 과정 및 결과

로컬에서 cmd로 BTC_USD.csv를 하둡으로 보내고, 하둡에서 pig(root/Data/demos) 로 보낸 후,

쓸 수 있게 BTC_USD.csv를 LOAD해서 릴레이션 'btc' 생성

```
btc = LOAD 'BTC_USD.csv' USING PigStorage(',') AS
(date:chararray,open:float,high:float,low:float,close:float,adj_close:float,volume:long);
```

```
grunt> btc = LOAD 'BTC_USD.csv' USING PigStorage(',') AS (date:chararray,open:float,high:float
,low:float,close:float,adj_close:float,volume:long);
2023-07-25 08:26:33,377 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered War
ning IMPLICIT_CAST_TO_LONG 2 time(s).
grunt> lm_btc = LIMIT btc 10;
2023-07-25 08:26:42,047 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered War
ning IMPLICIT_CAST_TO_LONG 2 time(s).
grunt> dump lm_btc;
```

```
(Date,,,,,,)
(2014-09-17,465.864,468.174,452.422,457.334,457.334,21056800)
(2014-09-18,456.86,456.86,413.104,424.44,424.44,34483200)
(2014-09-19,424.103,427.835,384.532,394.796,394.796,37919700)
(2014-09-20,394.673,423.296,389.883,408.904,408.904,36863600)
(2014-09-21,408.085,412.426,393.181,398.821,398.821,26580100)
(2014-09-22,399.1,406.916,397.13,402.152,402.152,24127600)
(2014-09-23,402.092,441.557,396.197,435.791,435.791,45099500)
(2014-09-24,435.751,436.112,421.132,423.205,423.205,30627700)
(2014-09-25,423.156,423.52,409.468,411.574,411.574,26814400)
```

(기본으로 활용할 X)

X = FOREACH btc GENERATE date,open,high,low,close,volume; (adj_close 컬럼은 무의미해서 제거)

```
grunt> X = FOREACH btc GENERATE date,open,high,low,close,volume;
2023-07-25 08:28:39,505 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered War
ning IMPLICIT_CAST_TO_LONG 2 time(s).
grunt> lm_X=LIMIT X 10;
```

```
(Date,,,,,)
(2014-09-17,465.864,468.174,452.422,457.334,21056800)
(2014-09-18,456.86,456.86,413.104,424.44,34483200)
(2014-09-19,424.103,427.835,384.532,394.796,37919700)
(2014-09-20,394.673,423.296,389.883,408.904,36863600)
(2014-09-21,408.085,412.426,393.181,398.821,26580100)
(2014-09-22,399.1,406.916,397.13,402.152,24127600)
(2014-09-23,402.092,441.557,396.197,435.791,45099500)
(2014-09-24,435.751,436.112,421.132,423.205,30627700)
(2014-09-25,423.156,423.52,409.468,411.574,26814400)
```

(헤더제거) RANK 연산자로 행에 순서 번호 할당하고, 필터 적용해서 첫번째 행 제거해야 함.

ranked_btc = RANK X;
filtered_btc = FILTER ranked_btc BY $0 > 1; -- rank는 결과의 첫 번째 컬럼($0)에 위치

(**수정된 X)
X = FOREACH filtered_btc GENERATE $1, $2, $3, $4, $5, $6; -- date, open, high, low, close, volume의 열 위치

```
grunt> ranked_btc = RANK X;
2023-07-25 08:32:39,970 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered War
ning IMPLICIT_CAST_TO_LONG 2 time(s).
grunt> filtered_btc = FILTER ranked_btc BY $0 > 1;
2023-07-25 08:32:59,458 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered War
ning IMPLICIT_CAST_TO_LONG 3 time(s).
grunt> X = FOREACH filtered_btc GENERATE $1,$2,$3,$4,$5,$6;
2023-07-25 08:33:29,519 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered War
ning IMPLICIT_CAST_TO_LONG 3 time(s).
grunt> lmlm_X = LIMIT X 10;
2023-07-25 08:33:59,783 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered War
ning IMPLICIT_CAST_TO_LONG 3 time(s).
grunt> dump lmlm_X;
2023-07-25 08:34:06,370 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered War
ning IMPLICIT_CAST_TO_LONG 1 time(s).
```


```
grunt> ranked_btc = RANK X;
2023-07-25 08:32:39,970 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered War
ning IMPLICIT_CAST_TO_LONG 2 time(s).
grunt> filtered_btc = FILTER ranked_btc BY $0 > 1;
2023-07-25 08:32:59,458 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered War
ning IMPLICIT_CAST_TO_LONG 3 time(s).
grunt> X = FOREACH filtered_btc GENERATE $1,$2,$3,$4,$5,$6;
2023-07-25 08:33:29,519 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered War
ning IMPLICIT_CAST_TO_LONG 3 time(s).
grunt> lmlm_X = LIMIT X 10;
2023-07-25 08:33:59,783 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered War
ning IMPLICIT_CAST_TO_LONG 3 time(s).
grunt> dump lmlm_X;
2023-07-25 08:34:06,370 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered War
ning IMPLICIT_CAST_TO_LONG 1 time(s).
```


1. 연도별 평균 종가

날짜에서 연도를 추출

btc_year = FOREACH X GENERATE SUBSTRING($0,0,4) AS year,open,close; // 종가까지 포함 + (수정 - 시가까지 넣어서 3번 부터 전처리 해볼 변화율 계산하기 위함)

lm_btc_year = LIMIT btc_year 10;

dump lm_btc_year; // 맨 위 헤더부분인(Date)도 나오니까 첫번째 row를 빼줘야 함.

```
grunt> btc_year = FOREACH X GENERATE SUBSTRING($0,0,4) AS year,open,close;
2023-07-25 08:36:57,245 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered War
ning IMPLICIT_CAST_TO_LONG 3 time(s).
grunt> lm_btc_year = LIMIT btc_year 10;
2023-07-25 08:37:11,352 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered War
ning IMPLICIT_CAST_TO_LONG 3 time(s).
grunt> dump lm_btc_year;
```

```
2023-07-25 08:37:32,804 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - T
otal input files to process : 1
2023-07-25 08:37:32,804 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRed
Util - Total input paths to process : 1
(2014,465.864,457.334)
(2014,456.86,424.44)
(2014,424.103,394.796)
(2014,394.673,408.904)
(2014,408.085,398.821)
(2014,399.1,402.152)
(2014,402.092,435.791)
(2014,435.751,423.205)
(2014,423.156,411.574)
(2014,411.429,404.425)
```

연도별로 그룹화

grp_year = GROUP btc_year BY year;

연도별 평균 종가 - 계산 완성

average_price_by_year = FOREACH grp_year GENERATE group,
AVG(btc_year.close);

dump average_price_by_year;

```
grunt> grp_year = GROUP btc_year BY year;
2023-07-25 08:40:43,256 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered War
ning IMPLICIT_CAST_TO_LONG 3 time(s).
grunt> average_price_by_year = FOREACH grp_year GENERATE group,AVG(btc_year.close);
2023-07-25 08:41:17,901 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered War
ning IMPLICIT_CAST_TO_LONG 3 time(s).
grunt> dump average_price_by_year;
```

```
2023-07-25 08:41:52,704 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - T
otal input files to process : 1
2023-07-25 08:41:52,704 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRed
Util - Total input paths to process : 1
(2014,363.6930847167969)
(2015,272.453380595168)
(2016,568.4924068763607)
(2017,4006.0336286988977)
(2018,7572.298946516481)
(2019,7395.246281704837)
(2020,11116.378092447916)
(2021,47436.932020547945)
(2022,28197.754098886988)
(2023,25941.86935546875)
```

2. 월별 평균 종가

날짜와 연도와 월을 따로 추출하고,

btc_month = FOREACH X GENERATE SUBSTRING($0,0,4) AS year,
SUBSTRING($0,5,7) AS month, open, close;

연도와 월별로 그룹화
grp_month = GROUP btc_month BY (year,month);

average_price_by_month = FOREACH grp_month GENERATE group,
AVG(btc_month.close);
dump average_price_by_month;