CIS 4930 Introduction to Hadoop and Big Data

Assignment 6

Akshat Shah

U10348398

a. Using Pig for ETL Processing

i. Introduction

In this project, we use pig to explore, correct, and reorder data from two different ad networks. We use the local mode option of pig to work with the smaller samples of data from the two ad networks before processing the complete set of data in the hadoop file system in MapReduce mode. We learn how to work in grunt shell and use various commands. We are analyzing the data and using Pig to filter invalid records, reorder fields, correct inconsistencies, and write corrected data to the hadoop file system.

ii. Data Format in the Original input file

- Data from ad_data1.txt

Index	Field	Data Type	Description	Example
0	keyword	chararray	Keyword that triggered ad	tablet
1	campaign_id	chararray	Uniquely identifies the ad	A3
2	date	chararray	Date of ad display	05/29/2013
3	time	chararray	Time of ad display	15:49:21
4	display_site	chararray	Domain where ad shown	www.example.com
5	was_clicked	int	Whether ad was clicked	1
6	cpc	int	Cost per click, in cents	106
7	country	chararray	Name of country in which ad ran	USA
8	placement	chararray	Where on page was ad displayed	TOP

- Data from ad_data2.txt

Index	Field	Data Type	Description	Example
0	campaign_id	chararray	Uniquely identifies the ad	A3
1	date	chararray	Date of ad display	05/29/2013
2	time	chararray	Time of ad display	15:49:21
3	display_site	chararray	Domain where ad shown	www.example.com
4	placement	chararray	Where on page was ad displayed	TOP
5	was_clicked	int	Whether ad was clicked	Y
6	cpc	int	Cost per click, in cents	106
7	keyword	chararray	Keyword that triggered ad	tablet

iii. Data Load Method in Pig

To load input data in this lab, instead of using the Grunt Shell we write a Pig Script to load and manipulate data. The data load function uses the LOAD function to mention the file and the AS function to mention the data types.

The function used in the file first_etl.pig is:

```
data = LOAD '/dualcore/ad_data1.txt'
                AS (
                        keyword: chararray,
                        campaign_id: chararray,
                        date: chararray,
                        time: chararray,
                        display_site: chararray,
                        was_clicked: int,
                        cpc: int,
                        country: chararray,
                        placement: chararray
                );
```

The function used in second_etl file is

```
data = LOAD '/dualcore/ad_data2.txt' USING PigStorage(',')
                AS (
                        campaign_id: chararray,
                        date: chararray,
                        time: chararray,
                        display_site: chararray,
                        placement: chararray,
                        was_clicked: int,
                        cpc: int,
                        keyword: chararray
                );
```

Here the LOAD function mentions the address of the sample to load and the AS function mentions the data type for each field.

We use the local function first to test a sample data from the whole file we are supposed to test to make sure the functionality is working fine without having to test the whole data on pig which takes longer. This is done by using the head functionality in hadoop which obtains a subset of the data.

iv. Data Processing Procedure

- For first_etl:

  First of all we the LOAD function to load the data and use the AS function to manipulate the data type for each field.

  Then use the FILTER functionality to filter out the data that is not from USA

Then we use the FOREACH….GENERATE statement to store the fields in a different order from the one we received them in. The country field is removed because we are only showing records from the USA. We use the UPPER and TRIM functions to change the keyword field to uppercase.

Then we use the STORE functionality to store the data to ad_data2 directory.

Index	Field	Description
0	campaign_id	Uniquely identifies the ad
1	date	Date of ad display
2	time	Time of ad display
3	keyword	Keyword that triggered ad

4	display_site	Domain where ad shown
5	placement	Where on page was ad displayed
6	was_clicked	Whether ad was clicked
7	cpc	Cost per click, in cents

- For second_etl:

  First of all we the LOAD function to load the data and use the AS function to manipulate the data type for each field.

  Then use the DISTINCT functionality to make sure all the script returns is unique.

  Then we use the FOREACH….GENERATE statement to store the fields in a different order from the one we received them in. We use the UPPER and TRIM functions to change the keyword field to uppercase and we use the REPLACE functionality to correct the date from MM-DD-YYYY to MM/DD/YYYY.

  Then we use the STORE functionality to store the data to ad_data2 directory.

v. Data Output and Results

- For first_etl.pig

The output only had results from USA but did not show the country firlsa. The fields are in correct order as given by the FOREACH…GENERATE function. The keywords were all uppercase.

- For second_etl.pig

  The ouput only had unique records. All the keywords were in uppercase and the fields were in correct order as mentioned in FOREACH...GENERATE function and the dates had MM/DD/YYYY format.

b. Analyzing Ad Campaign Data with Pig

i. Introduction

In this project, we use pig to analyze the data and optimize advertising for Dualcore. The project uses pig scripts to optimize various functionality and help Dualcore save money. We do this by finding low cost sites and high cost keywords for both ad networks and this way Dualcore can make an optimal marketing plan.

ii. Data Format in the Original input file

- Data from ad_data1.txt

Index	Field	Data Type	Description	Example
0	keyword	chararray	Keyword that triggered ad	tablet
1	campaign_id	chararray	Uniquely identifies the ad	A3
2	date	chararray	Date of ad display	05/29/2013
3	time	chararray	Time of ad display	15:49:21
4	display_site	chararray	Domain where ad shown	www.example.com
5	was_clicked	int	Whether ad was clicked	1
6	cpc	int	Cost per click, in cents	106
7	country	chararray	Name of country in which ad ran	USA
8	placement	chararray	Where on page was ad displayed	TOP

- Data from ad_data2.txt

Index	Field	Data Type	Description	Example
0	campaign_id	chararray	Uniquely identifies the ad	A3
1	date	chararray	Date of ad display	05/29/2013
2	time	chararray	Time of ad display	15:49:21
3	display_site	chararray	Domain where ad shown	www.example.com
4	placement	chararray	Where on page was ad displayed	TOP
5	was_clicked	int	Whether ad was clicked	Y
6	cpc	int	Cost per click, in cents	106
7	keyword	chararray	Keyword that triggered ad	tablet

iii. Data Load Method in Pig

To load input data in this lab, instead of using the Grunt Shell we write a Pig Script to load and manipulate data. The data load function uses the LOAD function to mention the file and the AS function to mention the data types.

The function used in the file low_cost_sites.pig is:

```
data = LOAD '/dualcore/ad_data[1-2]/part*'
                    AS (
                        campaign_id:chararray,
                        date:chararray,
                        time:chararray,
                        keyword:chararray,
                        display_site:chararray,
                        placement:chararray,
                        was_clicked:int,
                        cpc:int
                    );
```

The function used in high_cost_keywords.pig:

```
data = LOAD '/dualcore/ad_data[1-2]/part*'
```

```
AS (
        campaign_id:chararray,
            date:chararray,
            time:chararray,
            keyword:chararray,
            display_site:chararray,
            placement:chararray,
            was_clicked:int,
            cpc:int
);
```

Here the LOAD function mentions the address of the sample to load and the AS function mentions the data type for each field. We use a file glob pattern to load both directories.

We use the local function first to test a sample data from the whole file we are supposed to test to make sure the functionality is working fine without having to test the whole data on pig which takes longer. This is done by using the head functionality in hadoop which obtains a subset of the data.

iv. Data Processing Procedure

- For low_cost_sites.pig:

  First of all we the LOAD function to load the data and use the AS function to manipulate the data type for each field. We use file glob functionality to load both files.

  Then use the FILTER functionality to filter records where was_clicked has a value of 1.

  Then we use the GROUP functionality to group the data that was clicked ones by display sites.

Then we use the FOREACH….GENERATE statement to find the total cost for the grouped sites by summing the cost per click value for each site and create a new display relation that contains only the sites and the total cost of clicks on that site.

Then we sort the totals by ORDER..BY..ASC functionality,

Then we use the DUMP functionality to display the top 3 cheapest sites.

- For high_cost_keywords.pig:

  Since this script is only a variation of the code we wrote earlier we create a copy of low_cost_keywords.pig

  First of all we the LOAD function to load the data and use the AS function to manipulate the data type for each field. We use file glob functionality to load both files.

  Then use the FILTER functionality to filter records where was_clicked has a value of 1.

  Then we use the GROUP functionality to group the data that was clicked once by keywords.

  Then we use the FOREACH….GENERATE statement to find the total cost for the grouped keywords by summing the cost per click value for each site and create a new display relation that contains only the keywords and the total cost of clicks on that keyword.

  Then we sort the totals by ORDER..BY..DESC functionality,

  Then we use the DUMP functionality to display the top 5 cheapest keywords.

v. Data Output and Results

- For low_cost_sites.pig:

  The output only had 3 results which displayed the top 3 cheapest

sites in ascending order.

The sites were:

(bassoonenthusiast.example.com,1246)

(grillingtips.example.com,4800)

(footwear.example.com,4898)

- For high_cost_keywords.pig:

  The ouput only had 5 results which were in descending order by their total cost for each keywords.

  The results were:

  (PRESENT,165606)

  (TABLET,106509)

  (DUALCORE,95124)

  (BARGAIN,67913)

  (MOBILE,56348)

c. Bonus Lab #1, #2, and #3.

vi. Introduction

In the bonus labs, we use pig to analyze the data and optimize advertising for Dualcore. The project uses pig scripts to optimize various functionalities and help Dualcore save money. We do this by finding the number of ad clicks and estimate the maximum cost of the next Ad campaign. We also compare the click through rates for all the sites and output the lowest ones. We do this for both ad networks and this way Dualcore can make an optimal marketing plan.

vii. Data Format in the Original input file

- Data from ad_data1.txt

Index	Field	Data Type	Description	Example
0	keyword	chararray	Keyword that triggered ad	tablet
1	campaign_id	chararray	Uniquely identifies the ad	A3
2	date	chararray	Date of ad display	05/29/2013
3	time	chararray	Time of ad display	15:49:21
4	display_site	chararray	Domain where ad shown	www.example.com
5	was_clicked	int	Whether ad was clicked	1
6	cpc	int	Cost per click, in cents	106
7	country	chararray	Name of country in which ad ran	USA
8	placement	chararray	Where on page was ad displayed	TOP

- Data from ad_data2.txt

Index	Field	Data Type	Description	Example
0	campaign_id	chararray	Uniquely identifies the ad	A3
1	date	chararray	Date of ad display	05/29/2013
2	time	chararray	Time of ad display	15:49:21
3	display_site	chararray	Domain where ad shown	www.example.com
4	placement	chararray	Where on page was ad displayed	TOP
5	was_clicked	int	Whether ad was clicked	Y
6	cpc	int	Cost per click, in cents	106
7	keyword	chararray	Keyword that triggered ad	tablet

viii. Data Load Method in Pig

To load input data in this lab, instead of using the Grunt Shell we write a Pig Script to load and manipulate data. The data load function uses the LOAD function to mention the file and the AS function to mention the data types.

The function used in the file total_click_count.pig is:

```
data = LOAD '/dualcore/ad_data[1-2]/part*'    AS (
                                        campaign_id:chararray,
                                            date:chararray,
                                            time:chararray,
                                            keyword:chararray,
                                            display_site:chararray,
```

```
                placement:chararray,
                was_clicked:int,
                cpc:int
        );
```

The function used in project_next_campaign_cost.pig:

```
    data = LOAD '/dualcore/ad_data[1-2]/part*'    AS (
                campaign_id:chararray,
                date:chararray,
                time:chararray,
                keyword:chararray,
                display_site:chararray,
                placement:chararray,
                was_clicked:int,
                cpc:int
        );
```

The function used in lowest_ctr_by_site.pig:

```
    data = LOAD '/dualcore/ad_data[1-2]/part*'    AS (
                campaign_id:chararray,
                date:chararray,
                time:chararray,
                keyword:chararray,
```

```
                                    display_site:chararray,

                                    placement:chararray,

                                    was_clicked:int,

                                    cpc:int

                            );
```

Here the LOAD function mentions the address of the sample to load and the AS function mentions the data type for each field. We use a file glob pattern to load both directories.

ix. Data Processing Procedure

- For total_click_count.pig:

  First of all we the LOAD function to load the data and use the AS function to manipulate the data type for each field. We use file glob functionality to load both files.

  Then use the FILTER functionality to filter records where was_clicked has a value of 1.

  Then we use the GROUP functionality to group all the ads that were clicked.

  Then we use the FOREACH….GENERATE statement to find the total count for the grouped sites by using the COUNT functionality to count the number of clicks.

  Then we use the DUMP functionality to display the total.

- For next_campaign_cost.pig:

  First of all we the LOAD function to load the data and use the AS function to manipulate the data type for each field. We use file glob functionality to load both files.

Then use the FILTER functionality to filter records where was_clicked is not null.

Then we use the GROUP functionality to group the data that was clicked.

Then we use the FOREACH….GENERATE statement to find the total cost by generating the mac and multiplying the cpc by 50000

Then we use the DUMP functionality to display the total.

- For lowest_ctr_by_site.pig:

  First of all we the LOAD function to load the data and use the AS function to manipulate the data type for each field. We use file glob functionality to load both files.

  We group data by display_site.

  We create a FOREACH loop to loop through every single record. Then use the FILTER functionality to filter records where was_clicked is 1. Then we use the COUNT functionality to count the number of times an instance was clicked. Then we count the total number of data.

  The click through rate is determined by multiplying the number of times clicked by 100 and dividing it by the total.

  Then we use the ORDER functionality to sort the data in ascending orders.

  Then we dump the top 3 values.

  For the extra work, we output the top three keywords with highest ctr.

  In the FOREACH loop we group the data by keyword instead of display_site.

  Then we use the ORDERED functionality to sort the data in descending order.

  Then we dump the top 3 values.

x. Data Output and Results

- For total_click_count.pig:

  The output only displayed the total click count.

  The total was 18243.

- For next_campaign_cost.pig:

  The output only displayed the maximum total campaign cost.

  The total was 8000000.

- For lowest_ctr_by_site.pig:

  This script displays the lowest ctr by site.

  The lowest ctrs by site are:

  (bassoonenthusiast.example.com,1.000741289844329)

  (grillingtips.example.com,1.7343173431734318)

  (butterworld.example.com,1.90032269630692)

  The lowest ctrs by keywords are:

  (PRESENT,6.449976753032672)

  (BARGAIN,3.7029166445306276)

  (BYTEWEASEL,3.5706739911261356)