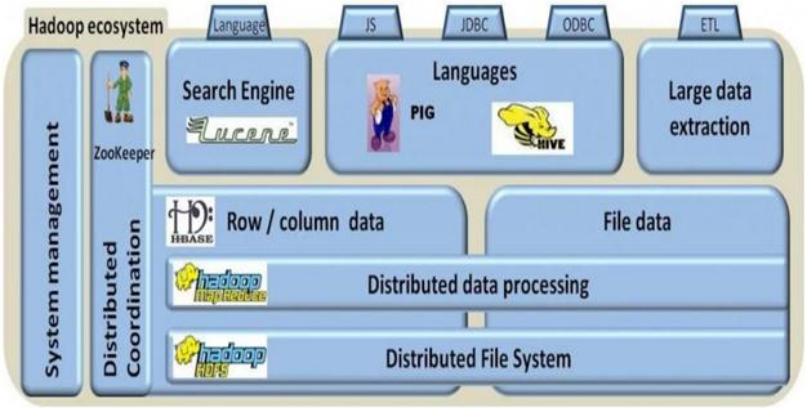
Project Report

(Big data and Hadoop)

By
Supriya Mandal

Index

- POC on Analyzing Book - Crossing Data
- Approach for Problem 1
- Approach for Problem 2
- Approach for Problem 3



POC on Analyzing Book - Crossing Data

Data set Description

The Book-Crossing dataset consists of 3 tables.

BX-Users:

This file contains the list of the users, their age and where they are collected. If that data is unavailable for any field then it is filled with NULL.

BX-Books:

It gives us the details about the book such as Book-Title, Book-Author, Year-Of-Publication, Publisher, Image-URL and ISBN. Here ISBN will act as a unique code for a book. Invalid ISBNs have already been removed from the dataset. URLs linking to cover images are also given, appearing in three different flavors (`Image-URL-S`, `Image-URL-M`, `Image-URL-L`) i.e. small, medium, large. These URLs point to the Amazon web site.

BX-Book-Ratings:

It contains the book rating information. Ratings are either explicitly expressed on a scale from 1-10 (higher values denoting higher appreciation) or implicitly expressed by 0.

POC on Analyzing Book - Crossing Data

Problem Statement

Find out the frequency of books published each year. (Hint: Use Boooks.csv file for this)

Find out in which year maximum number of books were published

Find out how many book were published based on ranking in the year 2002. (Hint: Use Book.csv and Book-Ratings.csv)

Approach for Problem 1

Problem Statement

Find out the frequency of books published each year

Approach

As BX-Books.csv file contains ISBN number and Year of publication is used for the calculation of the frequency of books published each year.

Data has been processed using Map-Reduce in JAVA.
Code has been written in Eclipse editor on windows and a jar file has been created for Map-Reduce Job.

Jar file: booksfreq.jar

Execute job

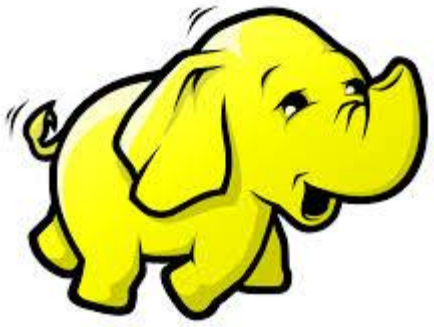
Map-Reduce code was executed using "**hadoop jar"** command

Source code reference

BooksPOC.zip file contains eclipse project with all files.

Output file

books_freq



Problem 1: output

Year	# of Books Published
1376	1
1378	1
1806	1
1897	1
1900	3
1901	3
1902	2
1904	1
1906	1
1908	1
1909	2
1910	1
1911	19
1917	1
1919	1
1920	32
1921	1
1922	2
1923	11
1924	2
1925	2
1926	1
1928	2
1929	4
1930	6
1931	3
1932	3
1933	3
1934	1
1935	3
1936	7
1937	3
1938	5
1939	9
1940	33
1941	8
1942	12

Year	# of Books Published
1943	6
1944	4
1945	6
1946	12
1947	10
1948	7
1949	9
1950	31
1951	39
1952	32
1953	59
1954	51
1955	66
1956	69
1957	71
1958	68
1959	95
1960	115
1961	124
1962	110
1963	117
1964	134
1965	155
1966	162
1967	143
1968	185
1969	288
1970	388
1971	446
1972	637
1973	739
1974	833
1975	1028
1976	1388
1977	1682
1978	1918
1979	2006

Year	# of Books Published
1980	2450
1981	3038
1982	3895
1983	4181
1984	4615
1985	4954
1986	5455
1987	6074
1988	6949
1989	7325
1990	8063
1991	8695
1992	9147
1993	9775
1994	10830
1995	12373
1996	12792
1997	13646
1998	14427
1999	15960
2000	15907
2001	16042
2002	16388
2003	13484
2004	5570
2005	36
2006	3
2010	2
2011	2
2012	1
2020	1
2021	1
2024	1
2026	1
2030	5
2038	1
2050	2

Approach for Problem 2

Problem Statement

Find out in which year maximum number of books were published

Approach

Input files:
As books_freq, the output file from the JAVA based Map-Reduce job file contains Year of publication and number of books published is used to find the year in which maximum number of books were published.
Data has been processed using Hadoop PIG.

Commands to load data into PIG
Copy all 3 files for books crossing into the following directory
/home/cloudera/poc/reference/Module4/input/

Create input directory in the below path /user/cloudera/
hadoop fs -mkfir/user/cloudera/input/

Load data into HDFS
hadoop fs -put /home/cloudera/poc/BX-CSV-Dump/* /user/cloudera/input/

Get result file from HDFS to local file system
hadoop fs -get /user/cloudera/input/poc2/part-r-00000 /home/cloudera/poc/prob2/max_published_year

PIG Script file
max_published_year .pig contains eclipse project with all files.

Approach for Problem 2: cont.

PIG Script

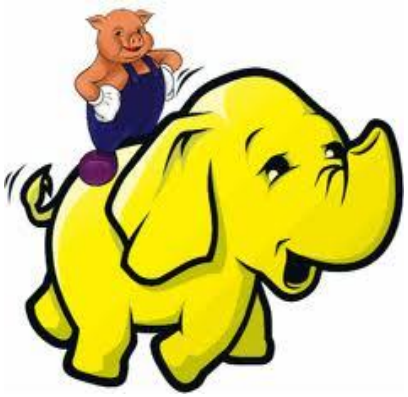
```
books_freq = LOAD 'input/books_freq' using PigStorage('\t') AS ( year:long, freq:long);

a = ORDER books_freq BY freq desc;
c = LIMIT a 1;
dump c;
store c into '/input/poc2' using PigStorage(',');
```

Output

Maximum Number of books published in the year 2002

Year	# of Books Published
2002	16,388



Approach for Problem 3

Problem Statement

Find out how many book were published based on ranking in the year 2002.

Approach

Input files:
Book.csv and Book-Ratings.csv are used to find the the number of book were published based on ranking in the year 2002.

Data has been processed using Hadoop PIG.
Commands to load data into PIG
Copy all 3 files for books crossing into the following directory
/home/cloudera/poc/reference/Module4/input/

Create input directory in the below path /user/cloudera/
hadoop fs -mkfir/user/cloudera/input/

Load data into HDFS
hadoop fs -put /home/cloudera/poc/BX-CSV-Dump/* /user/cloudera/input/

Get result file from HDFS to local file system
hadoop fs -get /user/cloudera/input/poc2/part-r-00000 /home/cloudera/poc/prob2/max_published_year

PIG Script file
booksRanking.pig contains the PIG scripts for the problem 2.

Approach for Problem 3: cont.

PIG Script

```
book_ratings = LOAD 'input/BX-Book-Ratings.csv' using PigStorage(';') AS ( UserID:chararray,
ISBN:chararray, BookRating:chararray);

ratings = FOREACH book_ratings GENERATE TRIM(REPLACE($1, '"', ' ')),
(INT)TRIM(REPLACE($2, '"', ' '));

ratings_f = FILTER ratings by $0 != 'ISBN';

bx_books = LOAD 'input/BX-Books.csv' using PigStorage(';') AS ( ISBN:chararray, BookTitle:chararray,
Book_Author:chararray, Year_Of_Publication:chararray, Publisher:chararray, Image_URL_S:chararray,
Image_URL_M:chararray, Image_URL_L:chararray);

books = FOREACH bx_books GENERATE TRIM(REPLACE($0, '"', ' ')), (INT)TRIM(REPLACE($3,
'"', ' '));

books_2002 = FILTER books by $1 == 2002;

books_2002_rating = join books_2002 by $0, ratings_f by $0;

rating_groups = group books_2002_rating by $3;

books_ranking_count = foreach rating_groups generate group, COUNT( books_2002_rating.$1);
dump books_ranking_count;

STORE books_ranking_count INTO  'output' USING PigStorage(',');
STORE books_ranking_count INTO  'input/ranking_count' USING PigStorage(',');
```

Approach for Problem 3: cont.

Output

Number of books were published based on ranking in the year 2002

Rating	# of Books Published in 2002
10	6189
9	6502
8	9761
7	6569
6	3147
5	3568
4	853
3	531
2	257
1	142
0	53124

Thank You