Milestone 5: Data Warehousing

1. PaperBook, EBook sale and Smartphone Penetration.

1) Hypothesis

Despite the increasing prevalence of smartphones and electronic devices, people's preference for reading paper books remains strong, suggesting that traditional reading materials may offer unique benefits or experiences not found in digital alternatives.

2) External data sources

- PaperBook: https://wordsrated.com/print-book-sales-statistics/
- Ebook: https://wordsrated.com/ebooks-sales-statistics/
- Smartphone:https://www.statista.com/statistics/201183/forecast-of-smartphone-penetration-in-the-us/

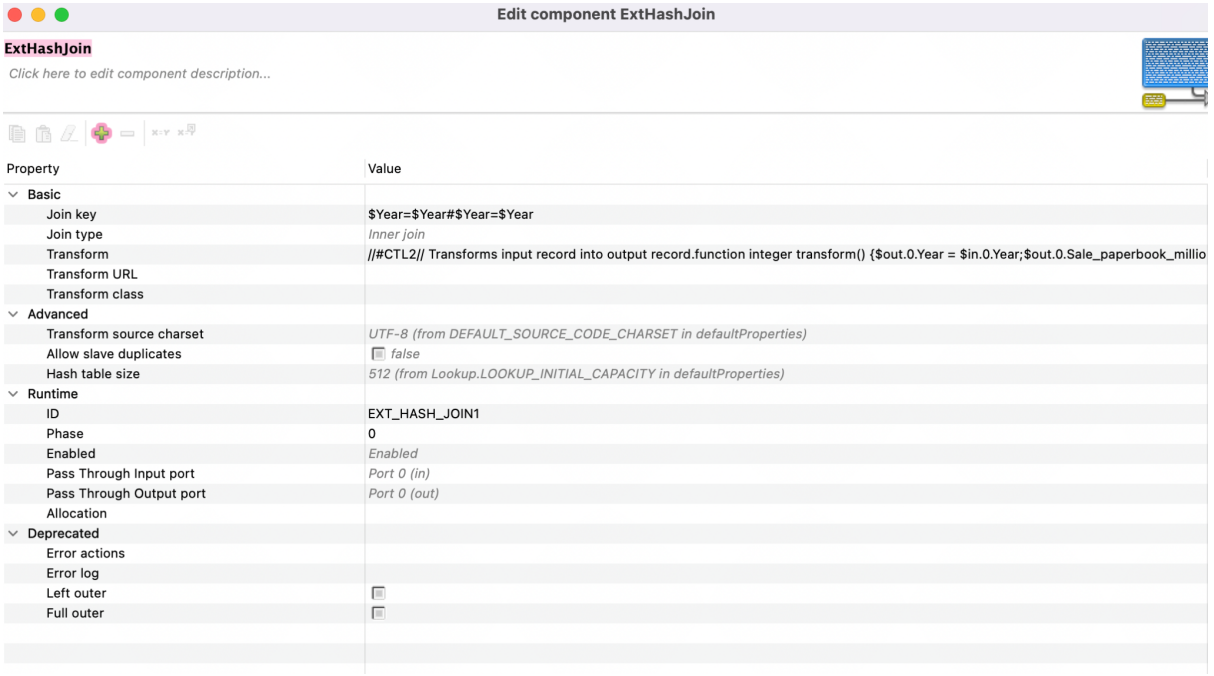
3) ETL workflows

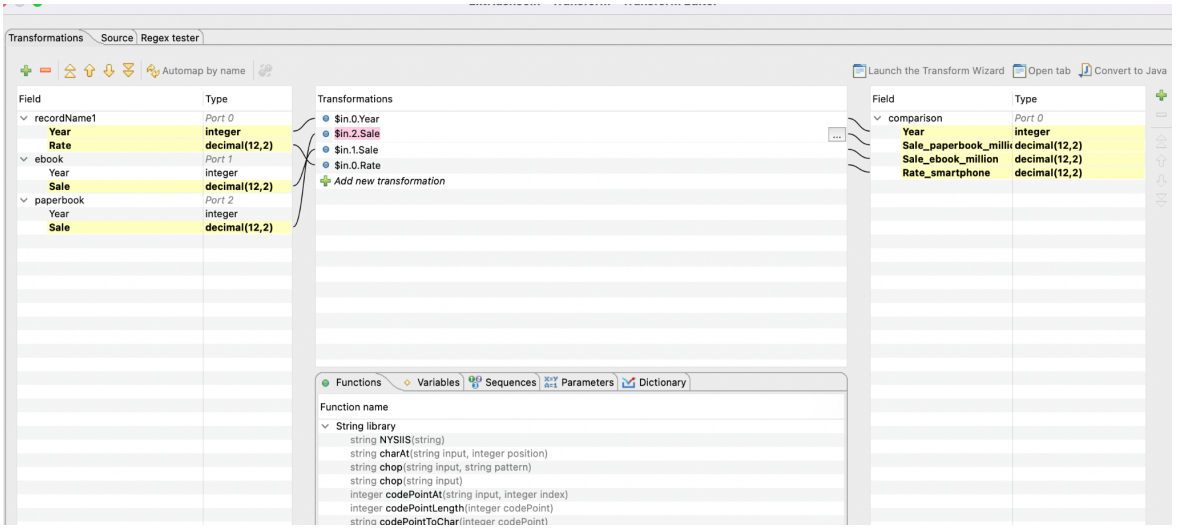
- SpreadsheetDataReader:

  We used the SpreadsheetDataReader to read three external datasets: paper-book sale, e-book sale and smartphone penetration rate.

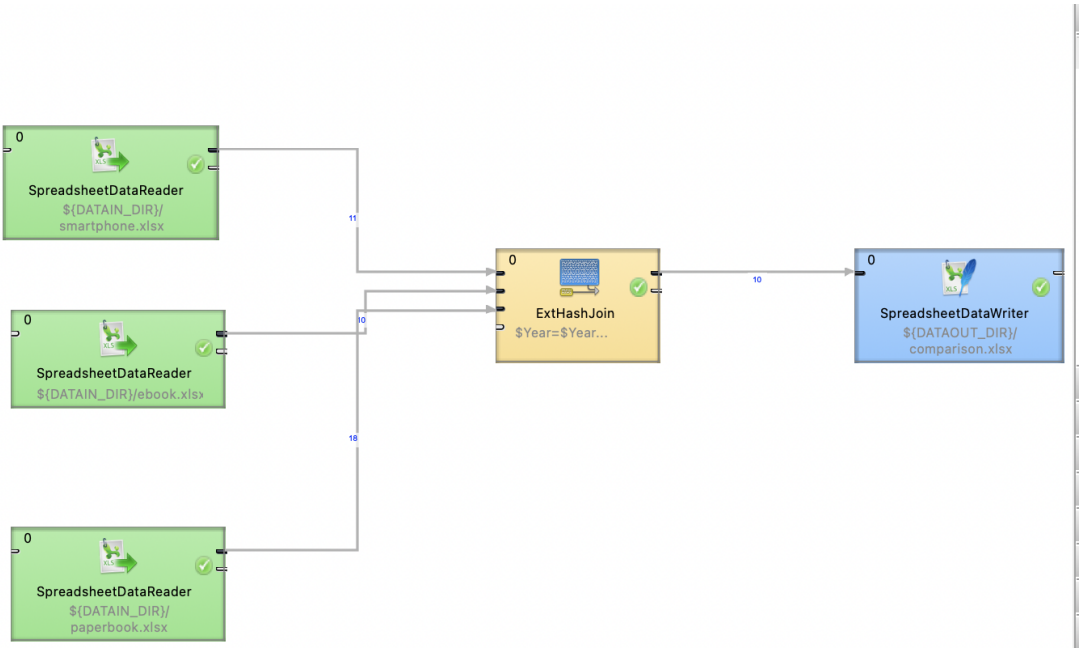
- ExHashJoin

  ExHashJoin performs an inner join on our three tables, merging them based on the shared "Year" attribute (Join key).

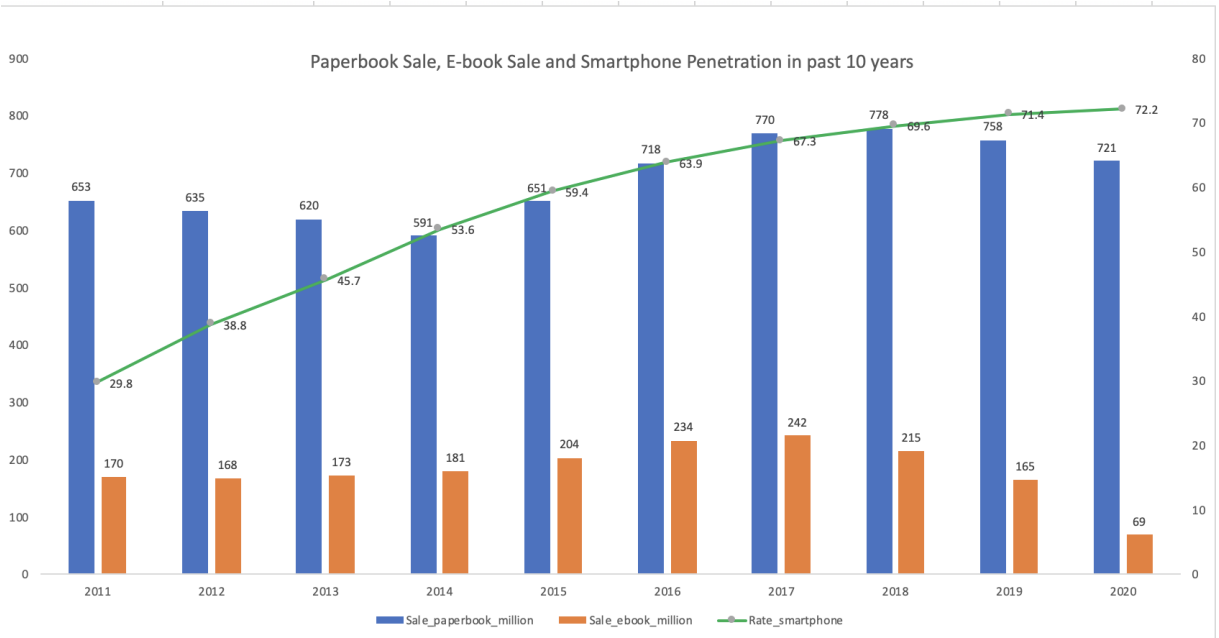

Edit component ExtHashJoin
ExtHashJoin
Click here to edit component description...
Property
Value
Basic
Join key
$Year=$Year#$Year=$Year
Join type
Inner join
Transform
//#CTL2// Transforms input record into output record.function integer transform() {$out.0.Year = $in.0.Year;$out.0.Sale_paperbook_millio
Transform URL
Transform class
Advanced
Transform source charset
UTF-8 (from DEFAULT_SOURCE_CODE_CHARSET in defaultProperties)
Allow slave duplicates
false
Hash table size
512 (from Lookup.LOOKUP_INITIAL_CAPACITY in defaultProperties)
Runtime
ID
EXT_HASH_JOIN1
Phase
0
Enabled
Enabled
Pass Through Input port
Port 0 (in)
Pass Through Output port
Port 0 (out)
Allocation
Deprecated
Error actions
Error log
Left outer
Full outer


Transformations
Source
Regex tester
Automap by name
Launch the Transform Wizard
Open tab
Convert to Java
Field
Type
recordName1
Port 0
Year
integer
Rate
decimal(12,2)
ebook
Port 1
Year
integer
Sale
decimal(12,2)
paperbook
Port 2
Year
integer
Sale
decimal(12,2)
Transformations
$in.0.Year
$in.2.Sale
$in.1.Sale
$in.0.Rate
Add new transformation
comparison
Port 0
Year
integer
Sale_paperbook_milli
decimal(12,2)
Sale_ebook_million
decimal(12,2)
Rate_smartphone
decimal(12,2)
Functions
Variables
Sequences
Parameters
Dictionary
Function name
String library
string NYSIIS(string)
string charAt(string input, integer position)
string chop(string input, string pattern)
string chop(string input)
integer codePointAt(string input, integer index)
integer codePointLength(integer codePoint)
string codePointToChar(integer codePoint)

- SpreadsheetDataWriter: was employed to generate the annual sales figures for both paperbacks and ebooks, as depicted in the flow-chart below.



4) Analysis and Results



The chart highlights the increasing penetration of smartphones in recent years, which indicates a growing digital device market. This trend suggests that consumers are increasingly adopting digital technologies, including e-books, which could have an impact on the future sales of paper-books.

Additionally, it is evident that the sales of paper-books experienced a slight decline from 2011 to 2014. However, the overall trend shows an increase in sales, and paper-books continue to dominate over ebooks in terms of sales volume. This indicates that there is still a preference for paper-books among consumers.

5) **Conclusion**

The findings derived from the graph substantiate our hypothesis, providing empirical evidence that supports the proposed theoretical framework.

With the advancement of economy and technology, people are using smartphones more and more. However, the enthusiasm for reading has not changed significantly. Despite the emergence of various electronic products, it seems that people still prefer reading paper books, compared with ebooks.

2. Economy and number of published books relationship

In this chapter, we want to analyze the relationship between the economy and the number of published books. The economy of a country is reflected by a variety of factors that provide insight into the health and performance of its overall economic system. Some key indicators of an economy's performance and health include GDP, Employment Rate, Inflation Rate, GDI, etc. We choose two factors to analyze: GDP and GNP.

- Gross Domestic Product (GDP): This is the total value of all goods and services produced within a country's borders during a specific period of time, usually a year. GDP is often used as a measure of a country's economic performance and growth.
- GNP stands for Gross National Product, which measures the total economic output of a country's residents and businesses, regardless of their location. It is similar to GDP, but it includes the value of goods and services produced by a country's residents and businesses, even if they are located outside the country's borders.

1) **Hypothesis**

During economic downturns, the publishing industry may face challenges such as reduced book sales, lower budgets for marketing and promotions, and increased competition for fewer book deals. This can result in fewer

books being published overall. On the other hand, during periods of economic growth, publishers may have more resources to invest in new authors and titles, leading to an increase in the number of books being published. Therefore, we suppose that during the period of economic increment, the number of published books will increase. On the other hand, during economic downturns, the number of published books will decrease.

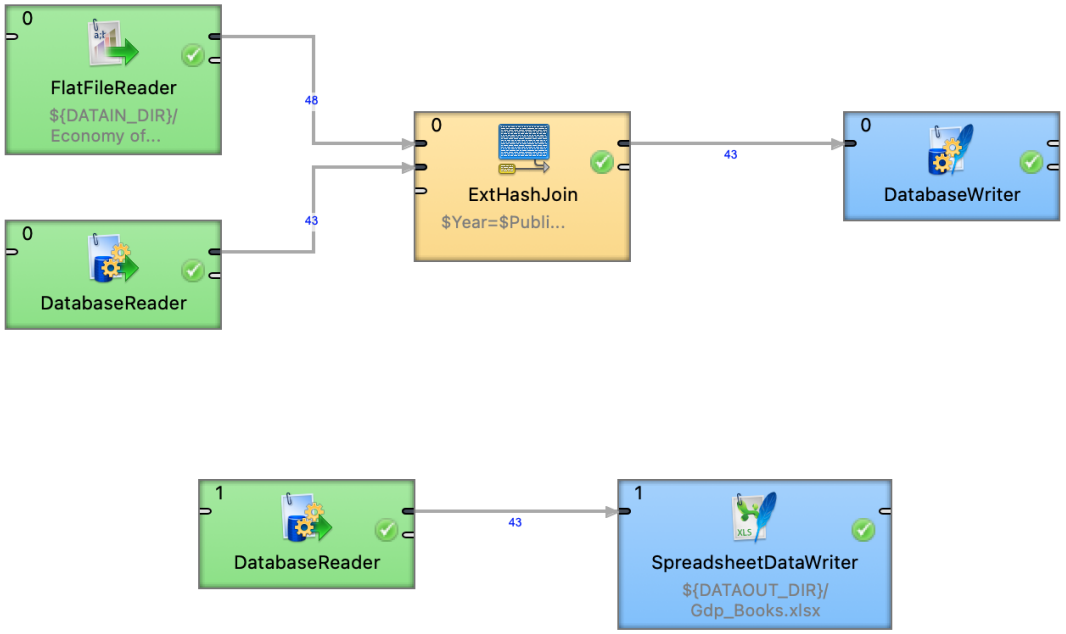
2) External data source

- GDP: https://www.kaggle.com/datasets/rajkumarpandey02/economy-of-the-united-states
- GNP:

  https://www.macrotrends.net/countries/USA/united-states/gnp-gross-national-product

In the GDP data source, we choose the column year and the column GPD(in billion $). In the GNP data source, we choose the column year and the column GNP(in billion $).We combine the year-GDP with the internal data year-number of published books and the year-GNP with the internal data year-number of published books.

3) ETL workflows



➔ **GDP-PublishedBookCounts**

- Data Source:

    A. Download the data (as CSV): GDP - Year https://www.kaggle.com/datasets/rajkumarpandey02/economy-of-the-united-state

    B. DataBaseReader: BookCount - Year

    Select PublishedDate, count(*)

    from BookDotNext_0.BookInfo

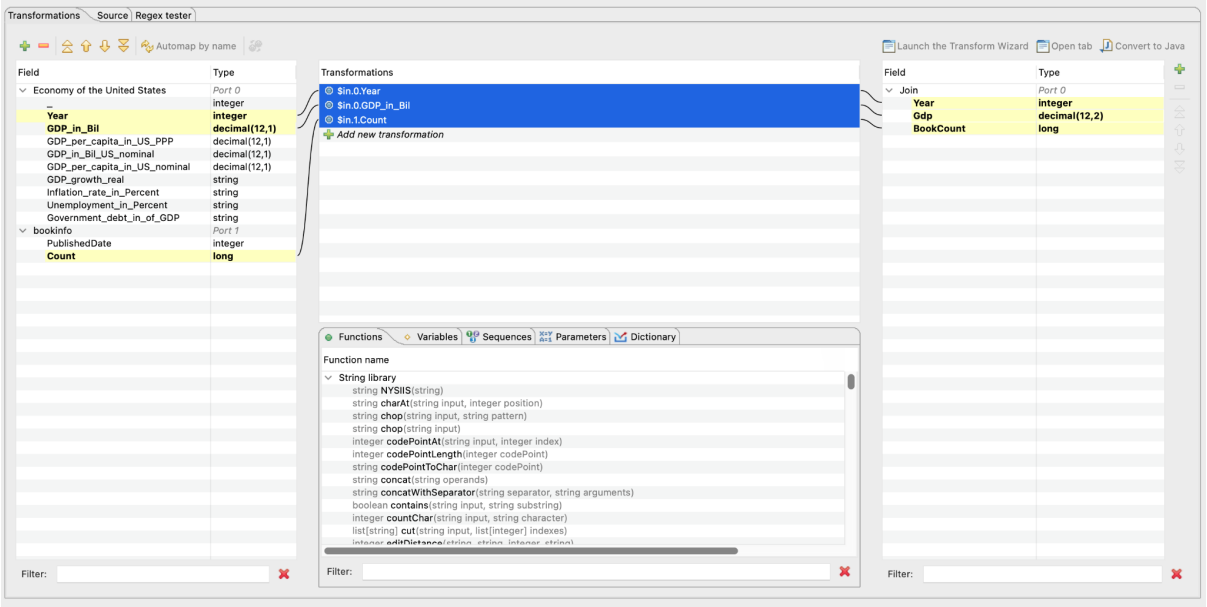
    where PublishedDate>=1980 and PublishedDate<=2022

    Group by PublishedDate

    Order by PublishedDate

- HashJoin:

    The hash joining the economic-year table and the book-publishing-year table based on the year as hashing the join key. The hash table for the economic-year table and the hash table for the book-publishing-year table can then be joined using the hash join process. The resulting joined table will contain the economic data and book-publishing data for each year.



- DataBaseWriter:

    A. Create The GDP-Year-BookCount Table (GDPBOOKS)

CREATE TABLE GDPBOOKS(GDPBOOKSID INT AUTO_INCREMENT,

YEAR INT, GDP DECIMAL , COUNT INT,

CONSTRAINT pk_GDPBOOKS_GDPBOOKSID PRIMARY KEY (GDPBOOKSID));

B. Write the hash joined result to database:

- DataBaseReader:

  Read data from GDPBOOKS.

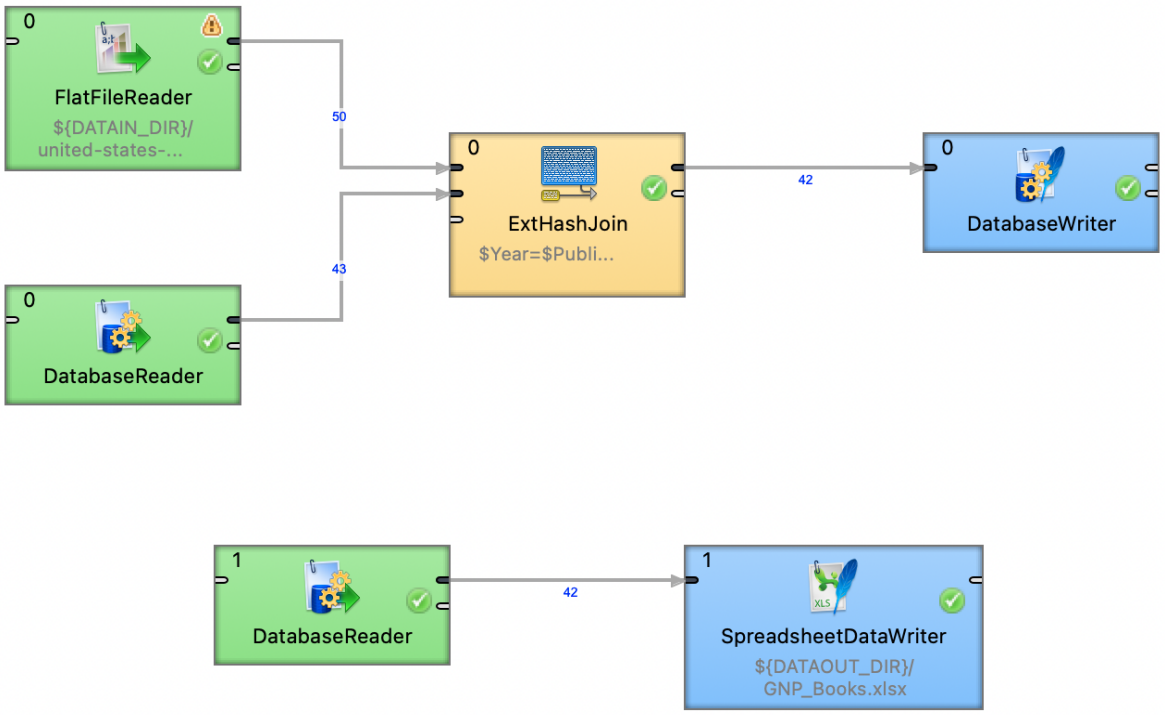
  SELECT GNPBOOKS.YEAR, GNPBOOKS.GNP, GNPBOOKS.COUNT FROM GNPBOOKS

- SpreadSheetWriter:

  Map and export data to Spreadsheet File

  Generate Chart in Spreadsheet

GNP-PublishedBookCounts



- Data Source:

    A. Download the data (as CSV): GNP - Year https://www.macrotrends.net/countries/USA/united-states/gnp-gross-national-product

    B. DataBaseReader: BookCount - Year

    Select PublishedDate, count(*)

    from BookDotNext_0.BookInfo

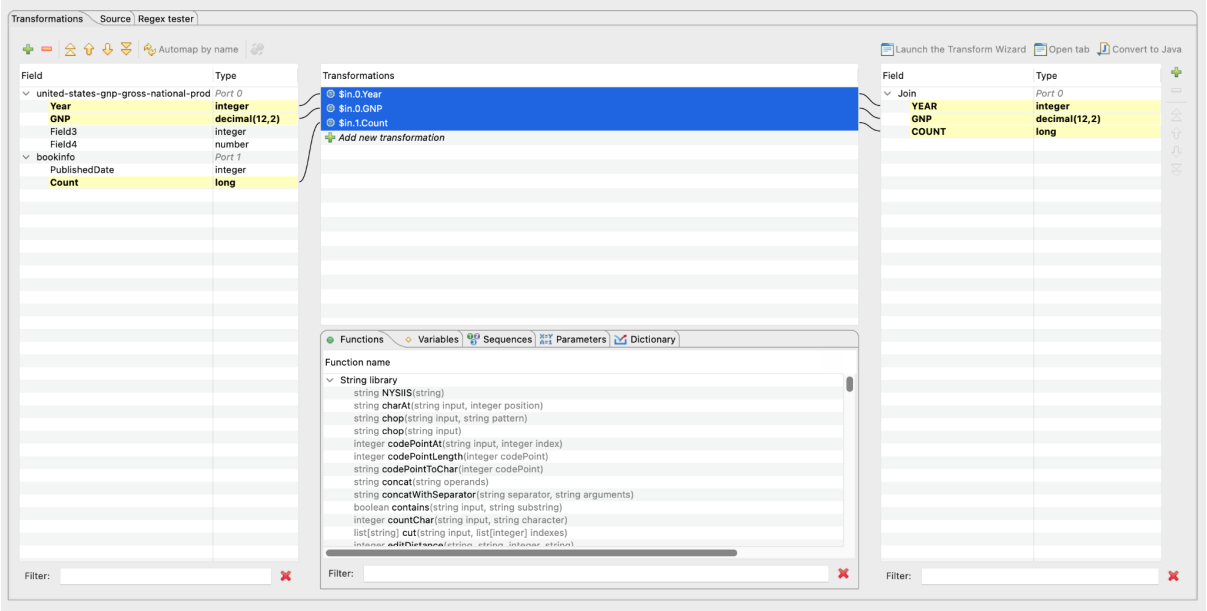
    where PublishedDate>=1980 and PublishedDate<=2022

    Group by PublishedDate

    Order by PublishedDate

- HashJoin:

    The hash joining the economic-year table and the book-publishing-year table based on the year as hashing the join key. The hash table for the economic-year table and the hash table for the book-publishing-year table can then be joined using the hash join process. The resulting joined table will contain the economic data and book-publishing data for each year.



- DataBaseWriter:

    A. Create The GNP-Year-BookCount Table (GNPBOOKS)

CREATE TABLE GNPBOOKS(GNPBOOKSID INT AUTO_INCREMENT,

YEAR INT, GNP DECIMAL , COUNT INT,

CONSTRAINT pk_GNPBOOKS_GNPBOOKSID PRIMARY KEY (GNPBOOKSID));

B. Write the hash joined result to database:

- DataBaseReader:

  Read data from GNPBOOKS.

  SELECT GNPBOOKS.YEAR, GNPBOOKS.GNP, GNPBOOKS.COUNT FROM GNPBOOKS

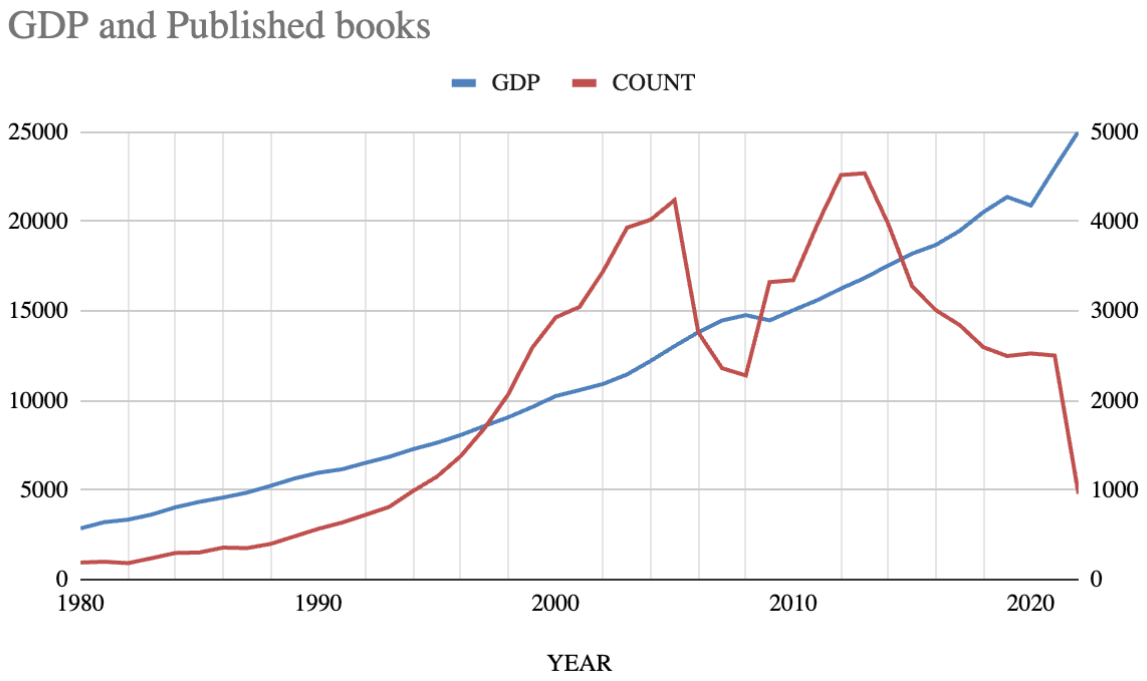
- SpreadSheetWriter:

  Map and export data to Spreadsheet File

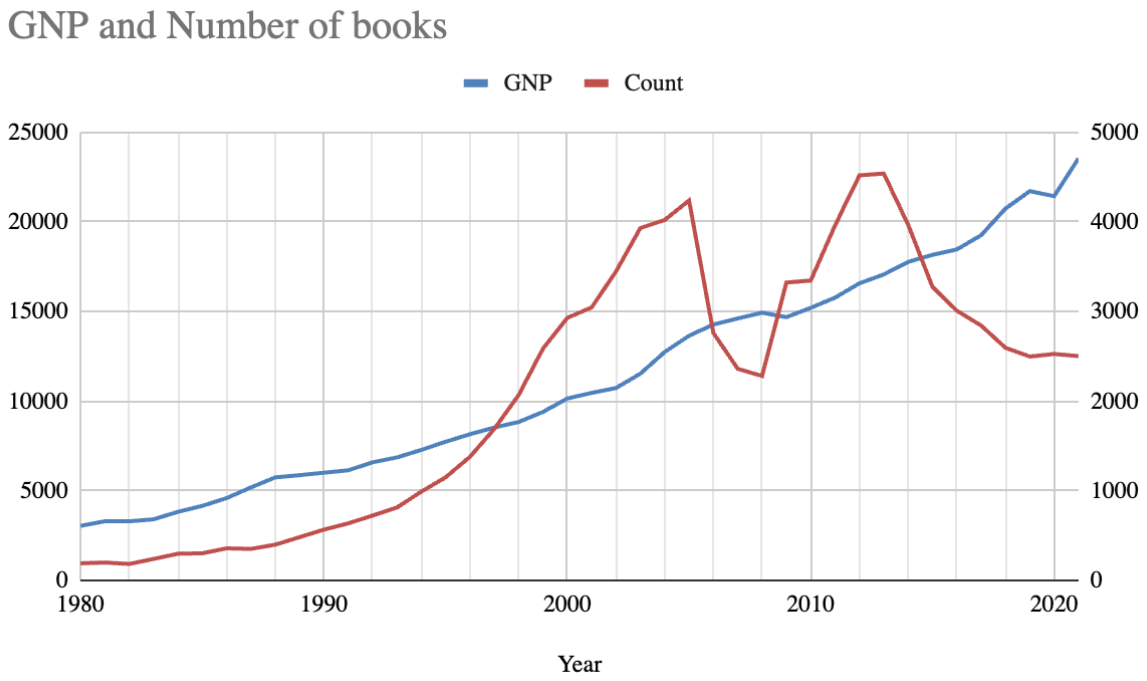
  Generate Chart in Spreadsheet

4) Results & Conclusion

GDP - number of published books



GNP - number of published books



There is a positive correlation between economic indicators and the number of printed books published, it means that as the economic indicators increase, the number of books published also increases. However, the gradual decline in the number of books published after 2012 suggests that there may be other factors at play.

To investigate this phenomenon further, we need to introduce new variables that may be affecting the number of books published. These variables could include changes in reading habits due to the rise of smartphones, digital media, the availability of alternative forms of entertainment, or changes in the publishing industry itself. By studying these variables, we can gain a deeper understanding of the factors that affect the number of printed books published and how they have changed over time. In the next chapter, we launch an analysis of the number of published books and technology development to drive deeper for more information.

3. Technologies and number of published books relationship

The year 2000 marked the beginning of a new era of technological development, often referred to as the "Information Age." Here are some of the major technological developments that have emerged since 2000: smartphones, social media, cloud computing, artificial intelligence, virtual and augmented reality and etc. In this chapter, we choose the smartphone as the representation of technology development. The introduction of smartphones in the early 2000s revolutionized the way we communicate and access information and led to the creation of new industries, such as mobile app development, which might have a great influence on the paper book market.

1) Hypothesis

Technology has had a significant impact on the production and consumption of paper books. While paper books have been the primary format for publishing and reading for centuries, technological advancements have led to the rise of new formats like e-books, audiobooks, and interactive books. Since technologies like smartphones, e-readers, and tablets have made it easier for readers to access books in digital formats, this has led to a decline in paper book sales. While paper books still hold a significant market share, the convenience and portability of digital formats have made them an attractive alternative for many readers.

Therefore, we suppose that during the period of technology blooming, the number of published books will decrease.

2) External data source

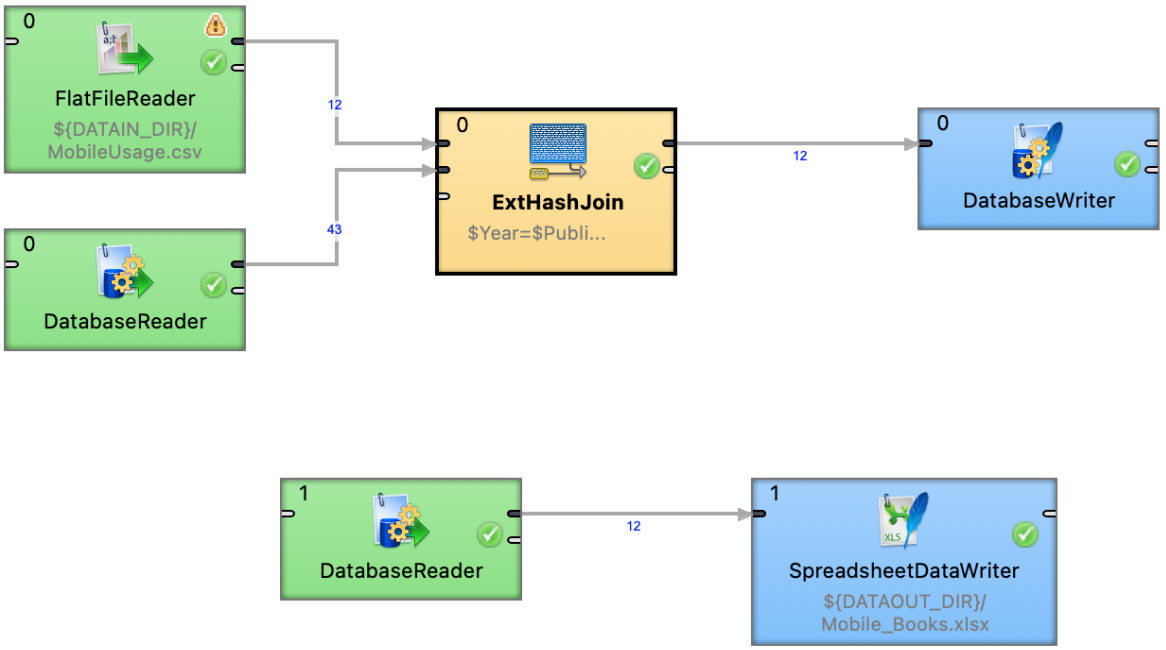
Year - MobilePenetrationRate

https://www.statista.com/statistics/201183/forecast-of-smartphone-penetration-in-the-us/

In the smartphone data source, we choose the column year and the column MobilePenetrationRate.

We combine the year-MobilePenetrationRate with the internal data year-number of published books.

3) ETL workflows



(MobilePenetrationRate vs PaperBooksCount) /yr

- Data Source:

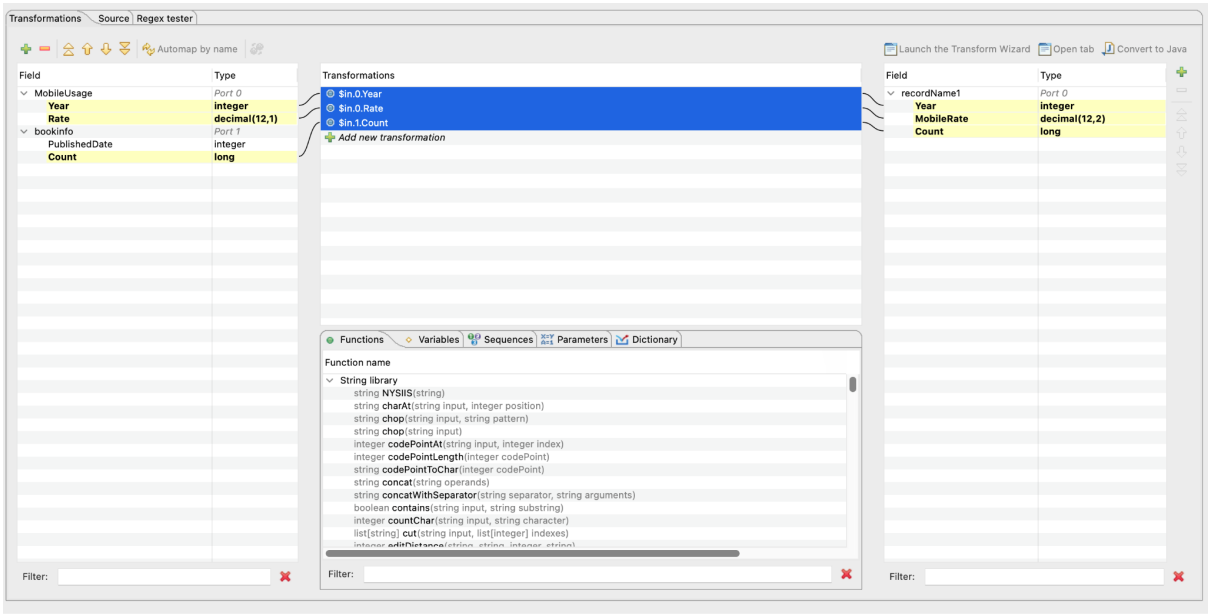
A. Download the data (as CSV): MobilePenetrationRate - Year
https://www.statista.com/statistics/201183/forecast-of-smartphone-penetration-in-the-us/

B. DataBaseReader: BookCount - Year

```
Select PublishedDate, count(*)

from BookDotNext_0.BookInfo

where PublishedDate>=1980 and PublishedDate<=2022

Group by PublishedDate

Order by PublishedDate
```

- HashJoin:

  The hash joining the economic-year table and the book-publishing-year table based on the year as hashing the join key. The hash table for the economic-year table and the hash table for the book-publishing-year table can then be joined using the hash join process. The resulting joined table will contain the economic data and book-publishing data for each year.



- DataBaseWriter:

    A. Create The MobilePenetrationRate-Year-BookCount Table (MobileBOOKS)

    CREATE TABLE MOBILEBOOKS(MOBILEBOOKSID INT AUTO_INCREMENT,

    YEAR INT, MOBILERate DECIMAL , COUNT INT,

    CONSTRAINT pk_MOBILEBOOKS_MOBILEBOOKSID PRIMARY KEY (MOBILEBOOKSID))

    B. Write the hash joined result to database

- DataBaseReader:

    Read data from MOBILEBOOKS.

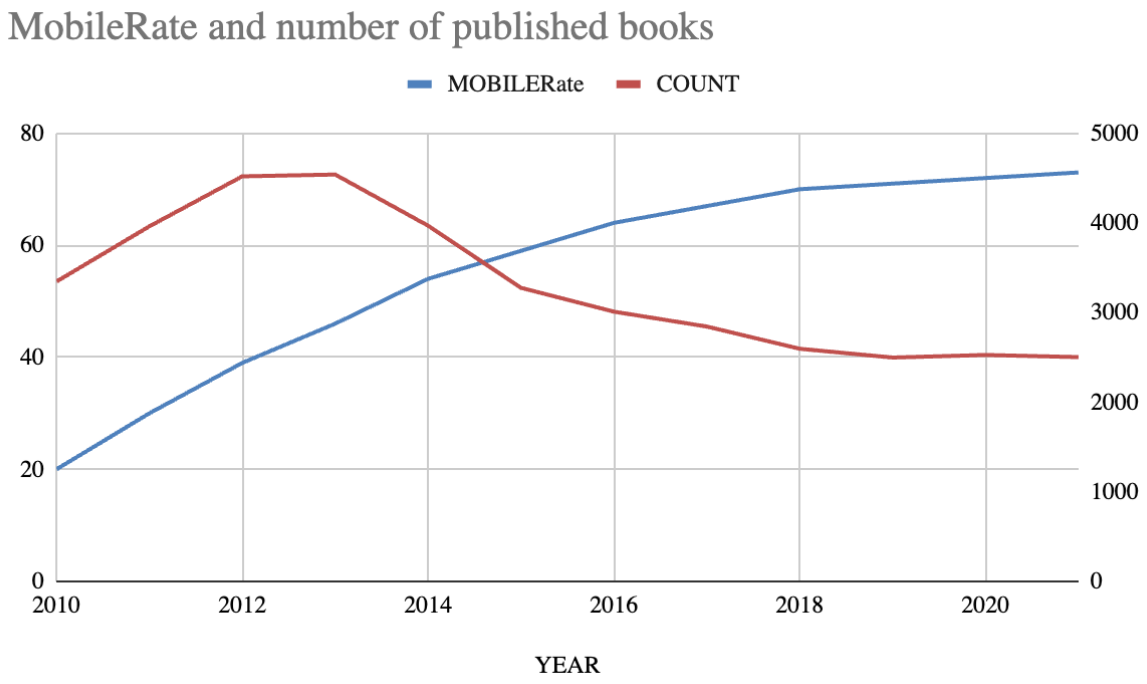
    SELECT MOBILEBOOKS.YEAR, MOBILEBOOKS.MOBILERate, MOBILEBOOKS.COUNT FROM MOBILEBOOKS

- SpreadSheetWriter:

    Map and export data to Spreadsheet File

    Generate Chart in Spreadsheet

4) Results & Conclusion



The chart shows a comparison between smartphone penetration and the number of published paper books over time, specifically from 2010 to 2022.

From 2010 to 2013, as smartphone penetration grew, the number of published paper books also increased. One possible explanation for this trend is that increased access to digital publishing platforms and e-books, facilitated by the proliferation of smartphones, may have actually stimulated the demand for paper books among some readers. In other words, some readers who were initially drawn to e-books and the digital content may have eventually sought out the physical experience of reading a traditional paper book.

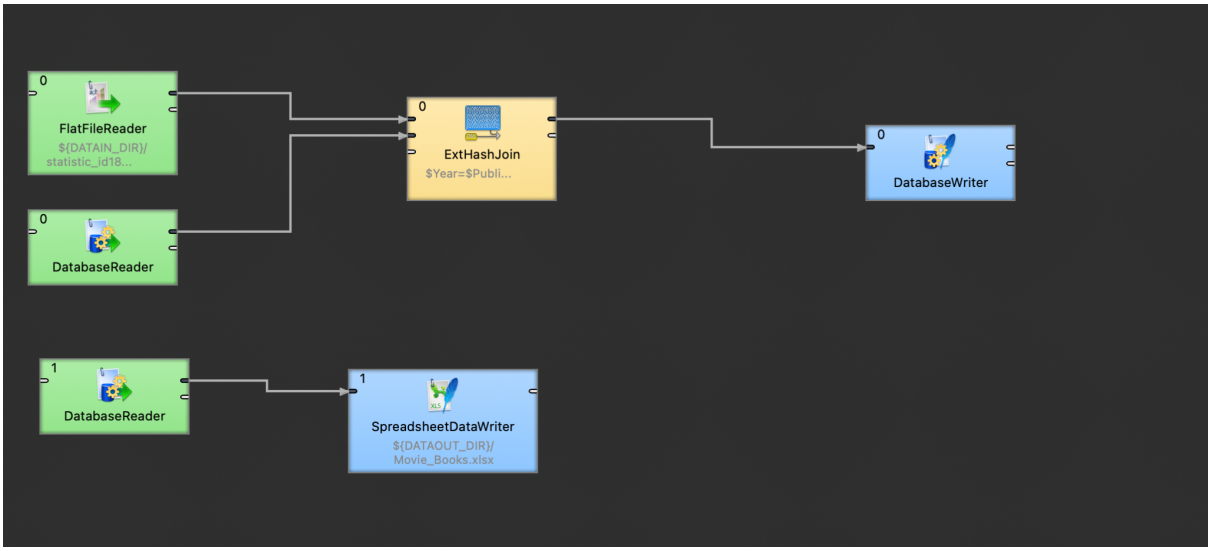
However, from 2014 to 2022, the number of published paper books decreased as smartphone penetration continued to grow. This inverse relationship suggests that as smartphone penetration increased, the number of published paper books decreased. One possible explanation for this trend is that as more people began to consume content on their smartphones, there may have been less demand for traditional print books. This may have been exacerbated by the COVID-19 pandemic, which led to a surge in e-book sales as people sought out more convenient and contactless ways to access reading material.

4. Movie vs Book

1) Hypothesis:

Our initial hypothesis was that there could be a correlation between the number of movies released each year and the number of published books per year. The more movies are released means that people are starting to watch more movies and that they spend more time watching movies in their free time than other activities such as reading books which will impact the book market by decreasing the number of published books. Thus, as the number of movies released increased, the number of books published decreased, and vice versa.

2) ETL WorkFlow



3) Data Set

- External Data : [Movie released in United States and Canada]

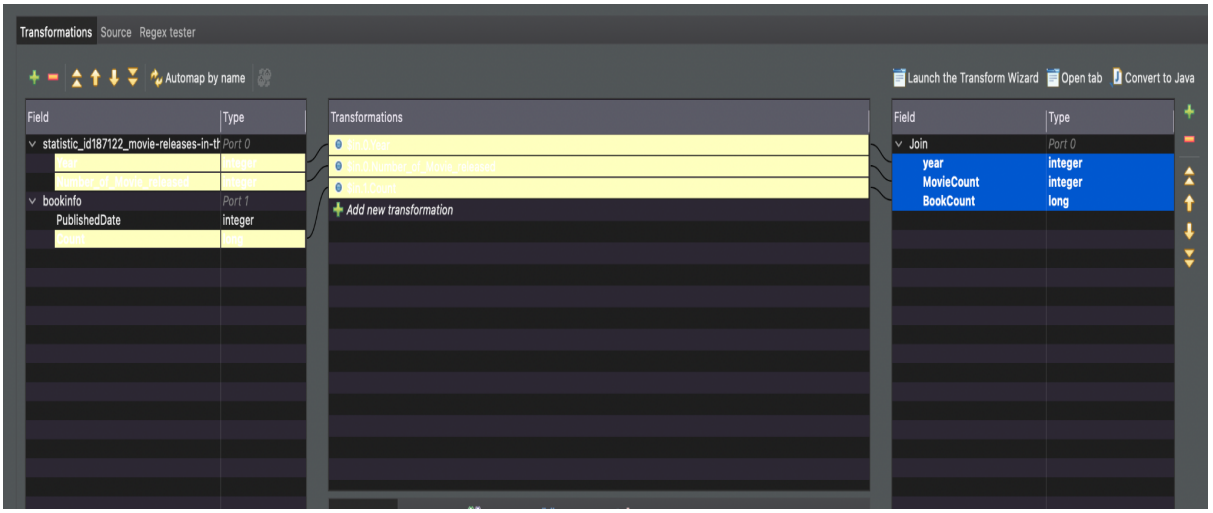
- We used the SpreadsheetDataReader to read Movie data
- Internal Data : Book.Next Database

```sql
Select PublishedDate, Count(*) as Count
from BookDotNext_0.BookInfo
where PublishedDate>=2000 and PublishedDate<=2022
Group by PublishedDate
Order by PublishedDate
```

Extract Data From Book.next database.

- **HASH JOIN**

The hash joining the Movie-year table and the book-publishing-year table based on the year as hashing the join key. The hash table for the Movie-year table and the hash table for the book-publishing-year table can then be joined using the hash join process. The resulting joined table will contain the Movie data and book-publishing data for each year.



- **DataBaseWriter**

Create Table MovieBook to store HashJoined Data

```
Use BookDotNext_0;
CREATE TABLE MOVIEBOOK(MOVIEBOOKID INT AUTO_INCREMENT,
YEAR INT, MOVIECOUNT INT , COUNT INT,
CONSTRAINT pk_MOVIEBOOK_MOVIEBOOKID PRIMARY KEY (MOVIEBOOKID));
```

Create Table MovieBook to store output Data

- **DataBaseReader**: Read data from MOVIEBOOK TABLE.

```
SELECT MOVIEBOOK.YEAR, MOVIEBOOK.MOVIECOUNT, MOVIEBOOK.COUNT FROM MOVIEBOOK
```

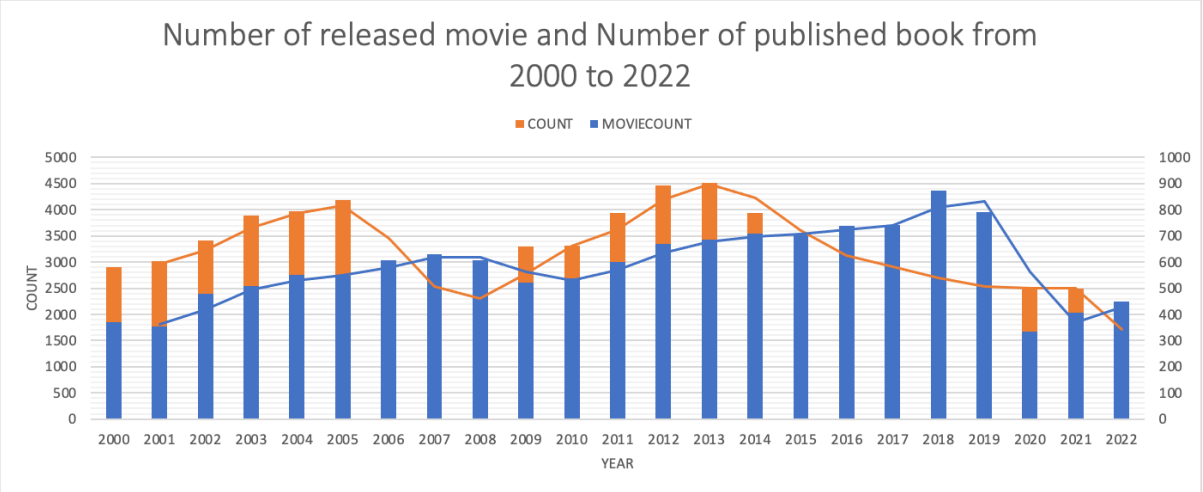
Select From MovieBook

- **SpreadSheetWriter:**

  Map and export data to Spreadsheet File

  Generate Chart in Spreadsheet

4) Graph



5) Analysis:

The number of movies released each year shows some fluctuation from 2000 to 2018. From 2019 because of the pandemic, the movie market crashed, and they couldn't release new movies during that time which lowered the MOVIECOUNT. Unlike MOVIECOUNT, the graph of the number of published books often goes up and down. However, it is gradually decreasing from 2013. We thought there is a correlation between them that if one goes up, one goes and and if one goes down, then one goes up. But contrary to our thought, both increase during 2001 to 2005 and 2010-2013, and both only decrease during 2018 to 2019. Also, there are only two periods being opposite to each other, 2006-2009 and 2015 to 2022. Both cases appeared in the graph, therefore we cannot conclude that they are interacting with each other with this data.

6) Conclusion:

Upon analyzing the data, we didn't observe a trend where an increase in the number of movies released was followed by a decrease in the number of published books, and vice versa. Additionally, we acknowledge that various factors influence the number of books published annually, such as changes in reader preferences, technological advancements, and economic conditions. In addition, the number of movies and books released in a given year can vary significantly based on factors

like studio schedules, publishing cycles, and industry trends. Therefore, while there may be a general trend suggesting an inverse relationship between the two, we cannot establish a direct correlation concluding that they are independent to each other.