**Use Case 4**

**Bookstore chain**

**Use Case**

A bookstore chain called “Your Best Readings” has realized that book sales have not increased as expected. The company wants to answer quickly the following queries in order to take decisions

i. What is the most popular book?

ii. What are the books sold together?

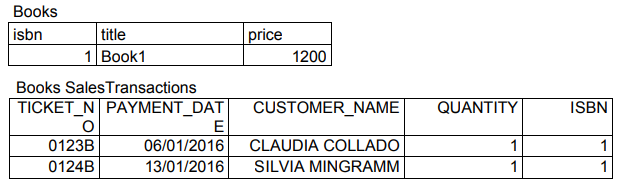
iii. What is the least sold book?

iv. How can we increase book sales?

Can we raise the average ticket by selling products in combo?

The bookstores have massive sales transactions, so would need to manage a high volume of data from the sales tickets and if they store all this information they could answer these important questions. However, the bookshops are distributed across the country and the CEO has decided to create a distributed repository. He is planning to store tweets, blogs or WhatsApp messages to recommend the most appropriate books to customers and increase sales.

Suppose you have two data sources with the following information:



What solution would you bring to the company?

The student should submit its proposal and justification of the following elements:

1. Typical architecture according to the type of information system

2. Kind of database or repository to implement and its design, ETL process if any, etc.

3. Answers to the queries required

4. How does it support scalability

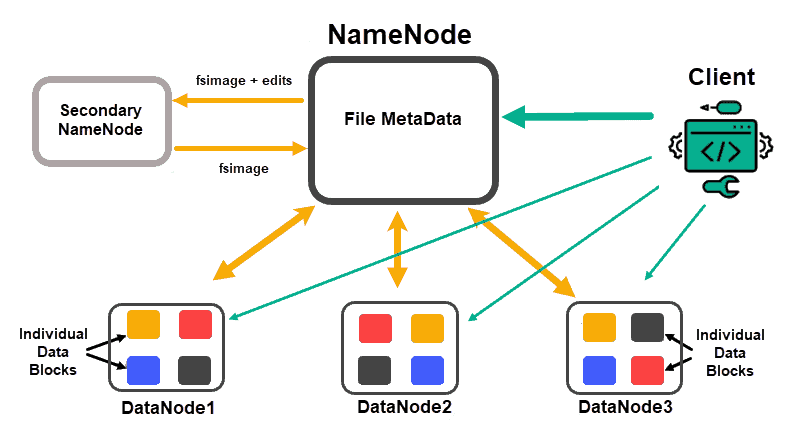
5. How does it support maintainability

6. How does it support security and reliability

**Solution**

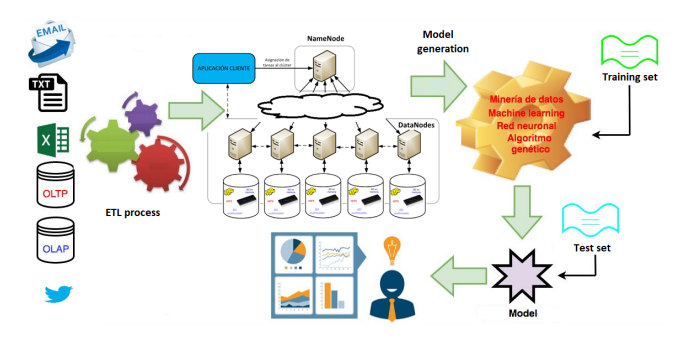
**1. Typical architecture according to the type of information system**

Since we need the power of analytical system. Multiple data sources of structured and unstructured data, and must apply data mining techniques, thus the best architecture suited for use case is HDFS (Hadoop distributed file system)



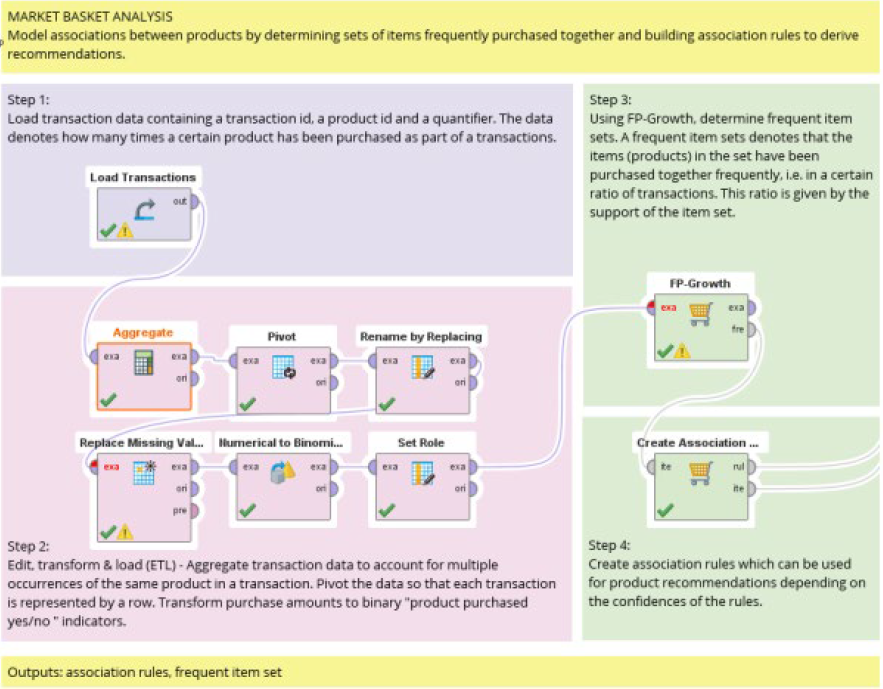
An HDFS in a distributed cluster across the country where we can store ticket information and use a Hadoop tool like mahout or weka, R, or any other tool to process data and execute a market basket analysis.

**2. Kind of database or repository to implement and its design, ETL process if any, etc.**



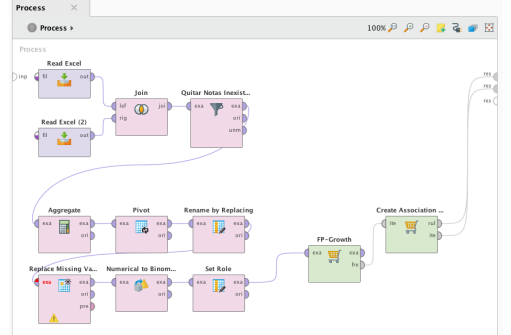
The Datasources can be joined together and uploaded to HDFS with the operator Read Excel, then it was necessary to use the Join operator to have the relationship between IDPROD and PRODUCT, then a filter was applied to eliminate those rows with notes empty, subsequently the ETL process was followed.

Further for Market based analysis to identify the association between products, we can do using FP growth pattern mining technique.

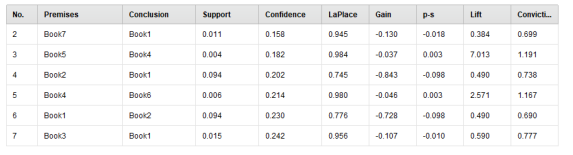


**3. Answers to the queries required**

Here are some deduction from RapidMiner process using Associations and FPGrowth.



Here are some associations from FP growth pattern matching technique:



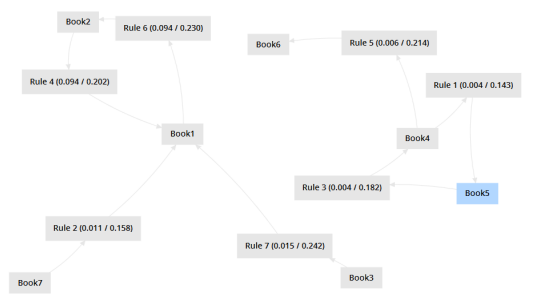
Here the sample answers to the queries:

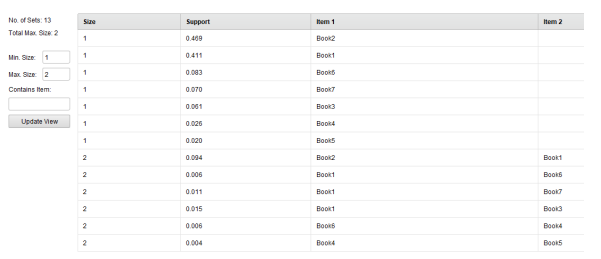
i. What is the most popular book? **Book1**

ii. What are the books sold together? **Book1 with Book7 and Book5 with Book4**

iii. What is the least sold book? **Book3**

iv. How can we increase book sales? **Making a marketing campaign promoting best sellers.**

****



**4. How does it support scalability**

We can achieve scalability by adding more nodes to the cluster.

**5. How does it support maintainability**

Maintainability is guaranteed with HDFS architecture.

**6. How does it support security and reliability**

Security and Reliability is guaranteed with HDFS architecture.