Workshop: “Machine Learning” and “Big Data”: basic concepts and applications Hosted by BPLIM

Detection of Frauds and Transactions in Local Trade: SAF-T with Big Data Technology

Eurico Lopes, Elói Lopes (IPCB)

19/dec/2018

Detection of Frauds and Transactions in Local Trade: SAF-T with Big Data Technology

BANCO DE PORTUGAL
EUROSYSTEM

- **Agenda**
  - Motivation
  - Contribution
  - Fraud Detection: Literature Review
  - Requirements
  - Architecture Proposal
  - Data Modelling
  - Implementation
  - Discussion
  - Conclusion
  - Future Work
  - Contribuition

Detection of Frauds and Transactions in Local Trade: SAF-T with Big Data Technology

BANCO DE PORTUGAL
EUROSYSTEM

Motivation

- Fraud and tax evasion has been something that the tax authorities have faced hard in recent years [1], with the adoption of e-fatura [2] in Portugal was introduced some control of transactions between the consumer and the shopkeeper, getting all recorded in a file named SAF-T [3];
- Prototype that aims to detect fraud in sales, stocks and control products inventories for local businesses using technology in the area of Big Data.

Detection of Frauds and Transactions in Local Trade: SAF-T with Big Data Technology

BANCO DE PORTUGAL
EUROSYSTEM

Contribution

- Collect SAFT-T XML files
- Storage using a DW
- Infrastructure using Big Data Technology:
  - HDFS
  - Hive
  - YARN/MapReduce
  - NoSQL under Cloudera
  - Pentaho Report Designer

Detection of Frauds and Transactions in Local Trade: SAF-T with Big Data Technology

BANCO DE PORTUGAL
EUROSYSTEM

Literature Review

- The methodology used are based on pattern search at digital library: ACM.org and IEEE.org;
- The most widely used approach to detect suspicious spending behavior relies on fraud rules because they can be (1) easily implemented, (2) efficiently evaluated, and (3) interpreted by the payment processor, the merchant and the customer to explain why a transaction was rejected [4].

Detection of Frauds and Transactions in Local Trade: SAF-T with Big Data Technology

Literature Review

- Other area is Health Care claims data and identify multiple research problems that can be solved using the existing big data analytics solutions. Healthcare data can be broadly categorized into four groups [5]:
  - **Clinical data** (patient health records, medical images, lab and surgery reports, etc.);
  - **Patient behavior data** (collected through monitors and wearable devices);
  - **Pharmaceutical research data** (clinical trial reports, high throughput screening results);
  - and **Health insurance data** (data collected and stored for several years by various health insurance agencies).

Detection of Frauds and Transactions in Local Trade: SAF-T with Big Data Technology

BANCO DE PORTUGAL
EUROSYSTEM

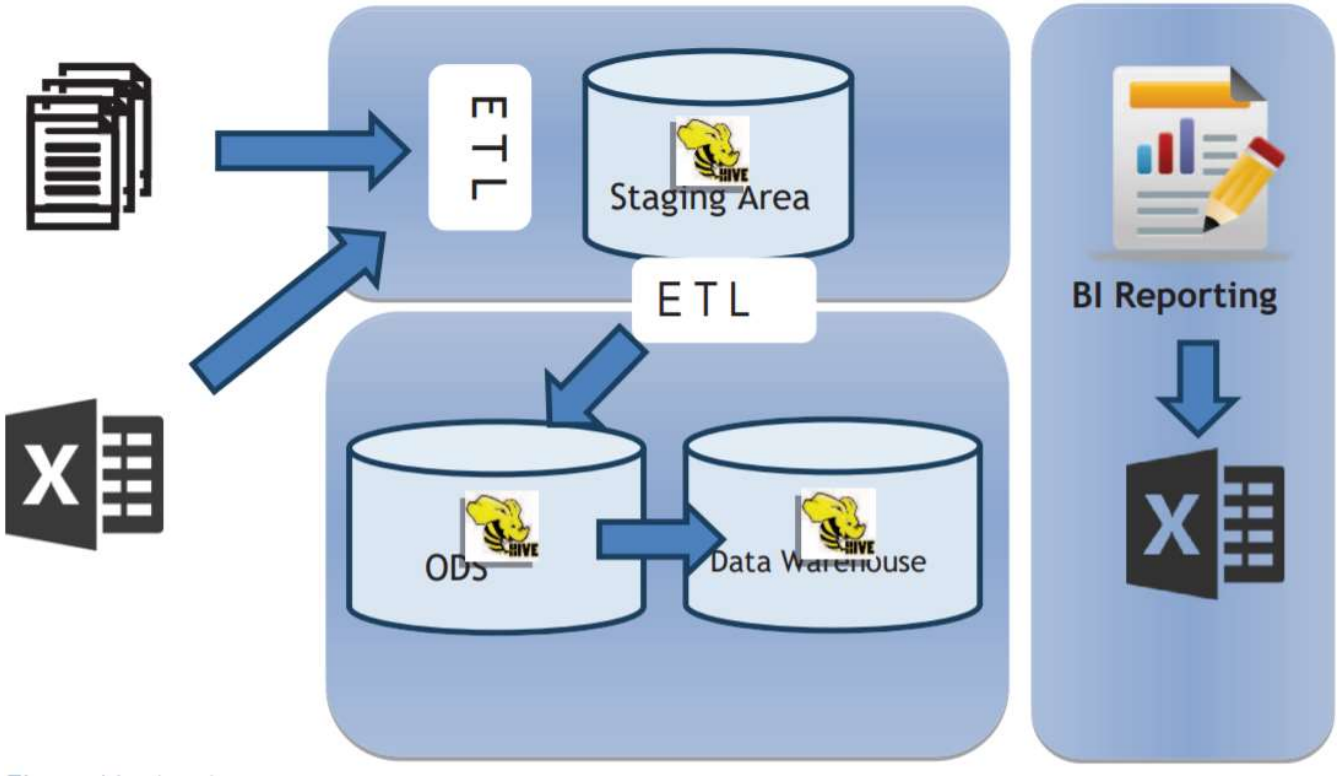
Requirements

- Report Analysis to detect certain types of frauds in:
  - Sales, stocks and control of inventories of products for local trades.
    - For example, the cross-checking of data between customer (merchant) and retailer to confirm the consistency between what is declared by both, in audits if the goods (products) that the merchant acquired, corresponds to what was billed, as well as detects if all the products that are sold were billed, and that there was no tax evasion.
  - Big Data tools from the Hadoop ecosystem.
  - The source of the data are the SAF-T files generated by the cash registers or accounting software that are installed in shops.

Detection of Frauds and Transactions in Local Trade: SAF-T with Big Data Technology

BANCO DE PORTUGAL
EUROSYSTEM

Architecture



Detection of Frauds and Transactions in Local Trade: SAF-T with Big Data Technology

BANCO DE PORTUGAL
EUROSYSTEM

Architecture

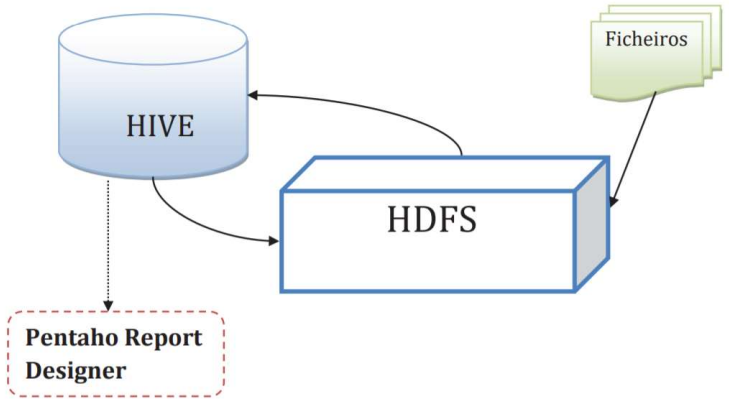
- Centralized data repository using a Data Warehouse:
  - supported by Hive, with a star data model;
  - data sources being the SAF-T.
- In order to analyze and visualize the data, it was used the Pentaho Report Designer.

Detection of Frauds and Transactions in Local Trade: SAF-T with Big Data Technology

BANCO DE PORTUGAL
EUROSYSTEM

Architecture

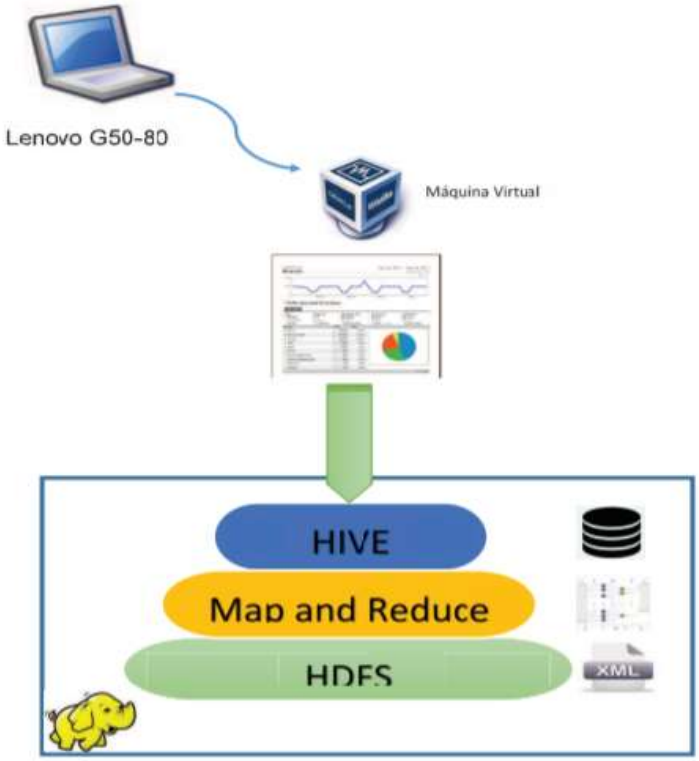
- The architecture design with Hadoop uses HDFS, Map Reduce and Hive.
- Files (XML) are loaded into the lower layer of Hadoop, the HDFS, in this layer the data generated by the sales and stocks are stored and are available for use by the remaining Hadoop components, Hive use the files that are in HDFS and load them in their tables, in this process Hive internally uses Map Reduce, after loading the data, a JDBC connection is opened to communicate with the reporting tool where the reports will be presented.



Detection of Frauds and Transactions in Local Trade: SAF-T with Big Data Technology

Technological Architecture

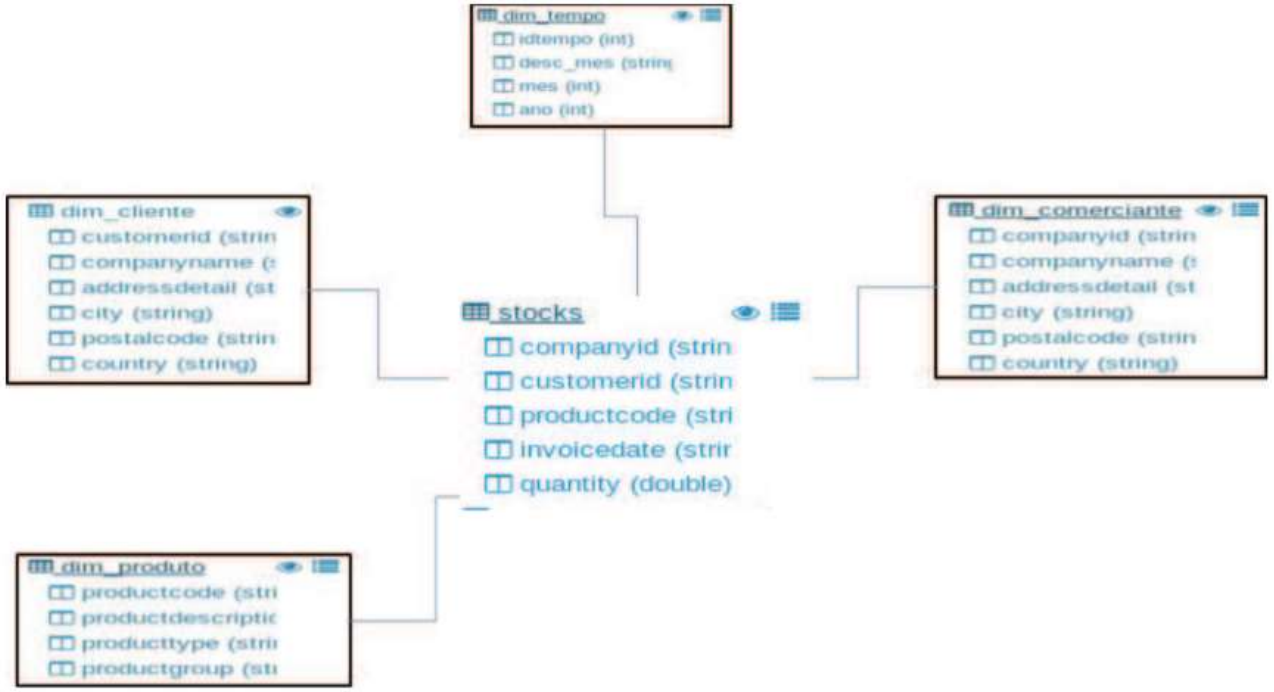
- Hardware is a Lenovo G50-80 equipped with an Intel® Core i7-5500U processor with a 2.4GHz dual-core speed, with 16GB of DDR3L RAM 1600MHz, with 1TB of internal storage with a speed of 5400 rpm.
- This hardware runs a virtual machine with CentOS release 6.5 (Final) operating system with 2 CPUs, 11GB of RAM. In the virtual machine Hadoop 2.5.0 is installed with a single node, as well as the remaining components of its ecosystem.
- Pentaho Report Designer 3.7 was installed for the construction and visualization of reports.



Detection of Frauds and Transactions in Local Trade: SAF-T with Big Data Technology

Data Modelling

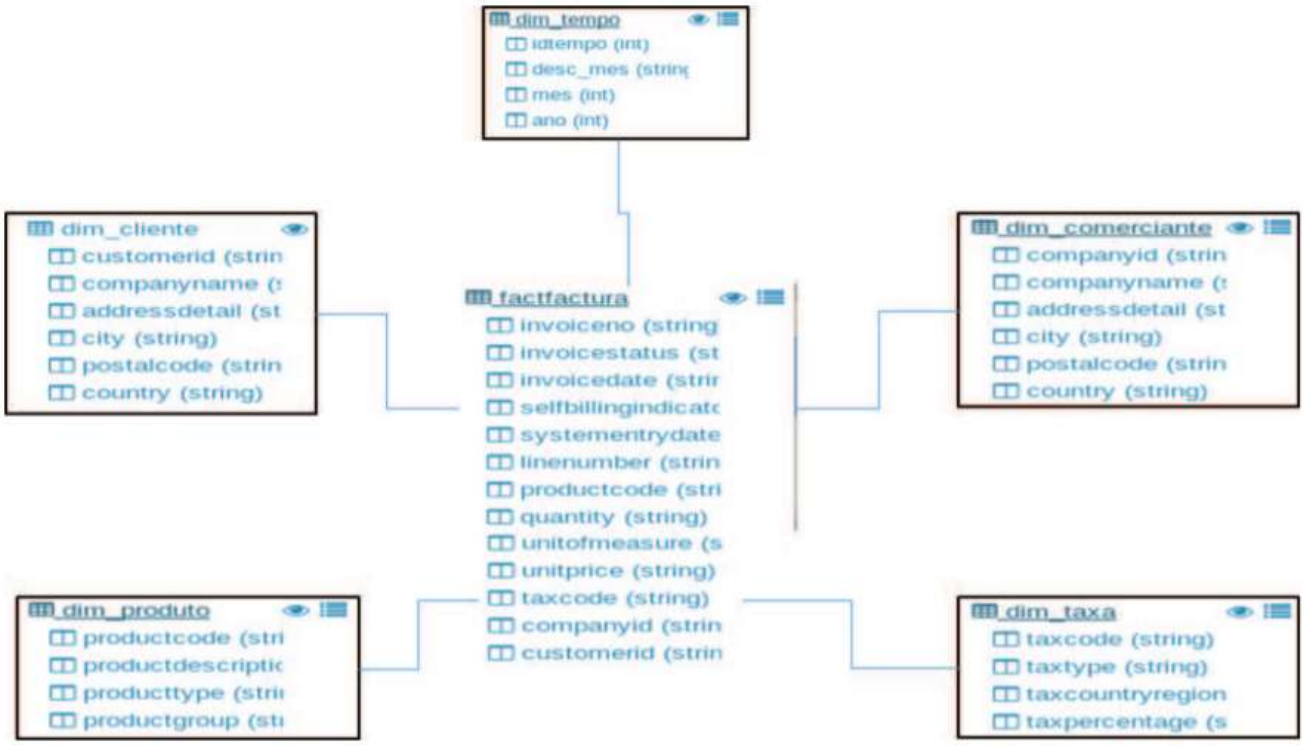
- **Stocks**



Detection of Frauds and Transactions in Local Trade: SAF-T with Big Data Technology

Data Modelling

- **Invoices**



Detection of Frauds and Transactions in Local Trade: SAF-T with Big Data Technology

Data Source Structure

- **SAF-T XML files**

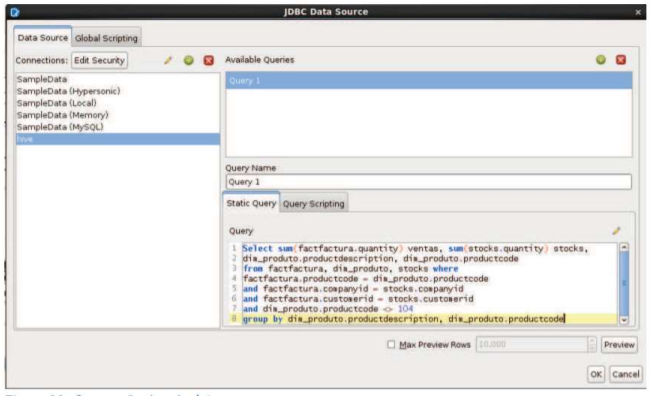
```
- <Product>
    <ProductType>P</ProductType>
    <ProductCode>25060618776084507</ProductCode>
    <ProductGroup>Família</ProductGroup>
    <ProductDescription>OUTROS PRODUTOS</ProductDescription>
    <ProductNumberCode>25060618776084507</ProductNumberCode>
  </Product>
- <Product>
    <ProductType>P</ProductType>
    <ProductCode>25060618776084509</ProductCode>
    <ProductGroup>Família</ProductGroup>
    <ProductDescription>OUTROS PRODUTOS</ProductDescription>
    <ProductNumberCode>25060618776084509</ProductNumberCode>
  </Product>
- <Product>
    <ProductType>P</ProductType>
    <ProductCode>25060618776084809</ProductCode>
    <ProductGroup>Família</ProductGroup>
    <ProductDescription>MINI CROISSANT</ProductDescription>
    <ProductNumberCode>8413760005901</ProductNumberCode>
  </Product>
- <Product>
    <ProductType>P</ProductType>
```

Detection of Frauds and Transactions in Local Trade: SAF-T with Big Data Technology

BANCO DE PORTUGAL
EUROSYSTEM

Implementation

- **ETL Process**
  - ETL process was developed in the Hadoop ecosystem, namely using the HiveQL
  - The open-source Hive-XML-SerDe project [7] was used to load the SAF-PT files to the external tables located in the Hive, in the Staging Area layer of the defined architecture.
  - Then all the Dimensions of the Fact Tables are loaded (For detail see page 39 to 43 [8])
- **Report**
  - Report Designer from Pentaho, which used the
    - Report Wizard available in the Pentaho suite



Detection of Frauds and Transactions in Local Trade: SAF-T with Big Data Technology

BANCO DE PORTUGAL
EUROSYSTEM

Discussion

- All the dimensions and consequently the fact tables where loaded successfully and report where produced without errors
- A report has also been built which allows an entity such as the tax authority, who in possession of the information of a trader and retailer, crosses the sales of both and can find differences between what the trader bought from the retailer and subsequently what has been sold.

Detection of Frauds and Transactions in Local Trade: SAF-T with Big Data Technology

BANCO DE PORTUGAL
EUROSYSTEM

Discussion

- The Hadoop components used in this project are a credible and significant alternative to be used as ETL tools, having the advantage of being open source and being able to develop functions tailored to the needs of the project, with Hadoop having a capacity to process and process data that current ETL tools can not achieve without much costlier than the one presented in the project.

Detection of Frauds and Transactions in Local Trade: SAF-T with Big Data Technology

BANCO DE PORTUGAL
EUROSYSTEM

Conclusion

- The initiative of this project originated in the great importance that Big Data is having in the short and medium term in the business strategy and business intelligence in companies and the willingness of the author to get involved in this new technology that will play a central role in the world of information technologies.
- Big Data also brought to this project new tools with an ability to process data. Among the many advantages associated with Big Data are the possibility of detecting fraud, predicting based on a large cluster of data, namely the SAF-T files that both tax authorities and marketers use, to record the movements of your business activity.

Detection of Frauds and Transactions in Local Trade: SAF-T with Big Data Technology

BANCO DE PORTUGAL
EUROSYSTEM

Conclusion

- Difficulties to implement the project:
  - Configuration and installation of the entire where the project was to be developed.
  - ETL process from XML files that are highly expandable files and that have a defined structure that takes to be respected, collecting the information for the temporary tables and later for the Data Warehouse were one of the most difficult points to overcome.

Detection of Frauds and Transactions in Local Trade: SAF-T with Big Data Technology

BANCO DE PORTUGAL
EUROSYSTEM

Future Work

- At the development environment level it would be very useful to have an infrastructure in node and have multiple nodes so that in case of failure it is possible to have the information on another node, in the case described it was not imperative due to the
- Volume of tested data small, but if the number of users and volume of data grows, this improvement would have to be adopted.
- Finally, SAF-T files of significant size, in the order of Giga's or Terabytes, could be included so that data could be processed with files of this size and such files are only generated by large companies.
- Add more tools do detect pattern of use that define profiles under a BI technology.

Detection of Frauds and Transactions in Local Trade: SAF-T with Big Data Technology

BANCO DE PORTUGAL
EUROSYSTEM

Papers

- Luis Rolo, Alexandre Fonte, Eurico Lopes (21015) Sistema de gestao e auditoria fiscal na nuvem, *Atas da 15 Conferencia da Associacao Portuguesa de Sistemas de Informacao*, 2 e 3 outubro 2015, ICTE, Lisboa
- Eurico Lopes et al, 2018. Feelings Detection System - A Proposal. Proceedings of *Fifth International Workshop on Cultures of Participation in the Digital Age - CoPDA 2018* Castiglione della Pescaia, Italy, May 29, 2018, CEUR-WS.org, online, Vol. 2101: 53-61, urn:nbn:de:0074-2101-9 Eurico Lopes, et al., 2018. Pitch

Detection of Frauds and Transactions in Local Trade: SAF-T with Big Data Technology

BANCO DE PORTUGAL
EUROSYSTEM

- References

[1] - Combate à Fraude e Evasão Fiscais e Aduaneiras [online] https://www.portugal.gov.pt/media/3322199/20150129-mf-plano-combate-fraude-fiscal-2015_2017.pdf [acedido a 03 december 2018]

[2] – Sobre o E-Fatura [online] http://info.portaldasfinancas.gov.pt/pt/faturas/pages/sobre_efatura.aspx pdf [acedido a 03 december 2018]

[3] - SAF-T PT (Standard Audit File for Tax purposes) - Versão Portuguesa –

[Online] http://info.portaldasfinancas.gov.pt/pt/apoio_contribuinte/SAFT_PT/Paginas/news-saf-t-pt.aspx

[acedido a 03 december 2018]

[4] - Davy Preuveneers, Bavo Goosens, and Wouter Joosen. 2017. Enhanced fraud detection as a service supporting merchant-specific runtime customization. In Proceedings of the *Symposium on Applied Computing* (SAC '17). ACM, New York, NY, USA, 72-76. DOI: https://doi.org/10.1145/3019612.3019886

[5] - Varun Chandola, Sreenivas R. Sukumar, and Jack C. Schryver. 2013. Knowledge discovery from massive healthcare claims data. In Proceedings of the *19th ACM SIGKDD international conference on Knowledge discovery and data mining* (KDD '13), Rayid Ghani, Ted E. Senator, Paul Bradley, Rajesh Parekh, and Jingrui He (Eds.). ACM, New York, NY, USA, 1312-1320. DOI: https://doi.org/10.1145/2487575.2488205

[6] - Saleh Al-Furiah and Lamia Al-Braheem. 2009. Comprehensive study on methods of fraud prevention in credit card e-payment system. In Proceedings of the *11th International Conference on Information Integration and Web-based Applications & Services* (iiWAS '09). ACM, New York, NY, USA, 592-598. DOI=http://dx.doi.org/10.1145/1806338.1806450

[7] - XML Serializer/Deserializer for Apache Hive [online] https://github.com/dvasilen/Hive-XML-SerDe/wiki/XML-datasources[

acedido a 11 de Março de 2016]

[8] Dissertação Mestrado Elói Lopes (2016) [online] https://repositorio.ipcb.pt/bitstream/10400.11/5321/1/Disserta%C3%A7%C3%A3o%20El%C3%B3i%20Lopes_A.pdf