How to maximize the value of big data with the open source SpagoBI suite through a comprehensive approach

Monica Franceschini
SpagoBI Competency Center
(Engineering Group)
Corso Stati Uniti 23/C
35127 Padova - Italy
+39 (0)49 8283411
monica.franceschini@eng.it

ABSTRACT

This paper describes the approach adopted by SpagoBI suite (www.spagobi.org) to manage large volumes of heterogeneous structured and unstructured data, to perform real-time Business Intelligence on Big Data streaming and to give meaning to data through the semantic analysis. SpagoBI supplies meaningful data insights through the main concept of persistable and schedulable datasets, and using tools such as self-service BI, ad-hoc reporting, interactive dashboards and explorative analysis.

1. A COMPREHENSIVE APPROACH TO BUILD INFORMATION AND GIVE IT A VISUAL REPRESENTATION

Managing and analyzing Big Data is one of the most interesting challenges that new technologies have to face nowadays. Big Data does not only refer to different kinds of sources. It also means managing structured and unstructured data, real-time data streams, as well as semantic ontologies applied to data.

Specifically, approaching Big Data requires the definition of new ways of operating, according the "*building of information*" and the "*visualization*" points of view.

Information building involves tasks allowing data extraction from different repositories or sources, as well as data aggregation capabilities for different analytical purposes. On the other hand, visualization stands for the process of giving data a visual representation.

1.1 Building of Information

SpagoBI suite faces the Big Data challenge through three main scenarios: Big Data Storages, Data in motion, Semantic information.

1.1.1 Big Data Storages

SpagoBI suite provides access to different types of data sources.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Articles from this volume were invited to present their results at The 39th International Conference on Very Large Data Bases, August 26th - 30th 2013, Riva del Garda, Trento, Italy.
Proceedings of the VLDB Endowment, Vol. 6, No. 11
Copyright 2013 VLDB Endowment 2150-8097/13/09... $ 10.00.

In detail, it can extract data from various platforms, ranging from analytical databases and appliances to NoSQL databases of different types (document-oriented databases, column families, graphs, multi-models), as well as from Hadoop ecosystem datasets, accessing Hive or Hbase, even Impala for better performances.

In order to build information, SpagoBI suite uses an independent component called "dataset", which enables developers to extract information from the above-mentioned data sources. There are different kinds of datasets. However, a dataset can be defined as a query performed on data that can be written using a specific query language (depending on the data source) or designed through a free enquiry engine.

Furthermore, a dataset can be scheduled, allowing developers to provide information at predefined time intervals, or it can be persisted. By persisting a dataset, the information extracted through the dataset can be stored in a physical storage (e.g. a more performing database or a file system) or onto an in-memory support for further low-latency analytical queries.

1.1.2 Data in Motion

SpagoBI can also analyse data on the fly, thanks to an adapter that can be embedded in the Stream Processing Engine or in the CEP (Complex Event Processing Engine), which notifies SpagoBI Server that new information is available. This information is kept by the dataset component, which can be persisted onto the in-memory support. This technology follows the "push based" paradigm. According to this, SpagoBI acts as a subscriber to the stream processor.

1.1.3 Semantic Information

SpagoBI suite provides specific tools to analyze huge amounts of data coming from heterogeneous sources and to easily create relationships among them. A widely adopted solution consists in classifying data through domain ontologies and semantic annotations. Afterwards this additional information can be stored in NoSql graph databases for future analysis. Specifically, SpagoBI retrieves this data and gives it a graphical representation through its network analysis engine, which acts as a filter on Big Data.

1.2 Visualization

SpagoBI dataset is used by different kinds of documents (e.g. reports, charts, interactive cockpits, maps, network analysis)

thanks to its wide range of analytical engines. In other words, datasets feed any type of analytics with the specific data that need to be analyzed and visually represented.

Specifically, cockpits and interactive dashboards allow end users to interact with several documents that are represented together in a single view. In this case, each document can be used as a driver for the others and the dashboard itself supplies the information integration at a glance.

SpagoBI suite also offers an ad-hoc reporting engine that allows users to create their own analytical documents including tables and charts, with a few "drag and drop" actions. This tools is a valuable support for data scientists who need to refine their analysis in order to give their datasets an instant graphical meaning.

1.3 Building of Information & Visualization

An agile and interactive way to perform insights over data and to refine one's own queries on Big Data can be obtained through the Self-service BI feature provided by SpagoBI suite. In fact, SpagoBI provides a Query by Example engine (QbE), which allows power users to create their own queries in a graphical and smart way over their datasets. Several filter conditions and aggregations can be applied on the datasets, so that results can instantly been used by the ad-hoc reporting engine for further graphical analysis. Newly created datasets can be saved and used by the other SpagoBI engines, as well as scheduled and persisted.

2. SHORT AUTHOR'S BIOGRAPHY

Monica Franceschini is SpagoBI Architect and Big Data specialist at the SpagoBI Competency Center of Engineering Group. She delivers consulting services and training courses on SpagoBI. She also contributes to the development of the suite and to the realization of customers' projects.

As Big Data specialist, she is currently developing Big Data solutions for SpagoBI analytics. With more than ten years of experience as an IT solution architect, she has developed many projects for the industry and public administration, based on Java technologies. Monica is teacher of big data technologies at Engineering Group's ICT Training School.