Appendix A
Survey on Business Intelligence Tools

In the following, tools and systems are provided and described that support the BI tasks *data modeling and understanding* (cf. Chap. 2) and *ETL support* (cf. Chap. 3); *big data and cloud* (cf. Chap. 3); *visualization, visual mining, and reporting* (cf. Chap. 4); *data mining* (cf. Chaps. 5 and 6); *process analysis* (cf. Chaps. 7 and 8); and *text mining* (cf. Chap. 8).

Each tool description will be based on the following schema (cf. Table A.1). The schema features general criteria which can be more detailed for special purposes, e.g., mobile access or dashboard creation. An abundance of commercial and noncommercial tools tailored for specific purposes can be found on the Web. Every decision depends to some extent on personal preferences.

Tool collection: The tool descriptions shall cover all tools used throughout the book, complemented with further tools. Please note that the focus is on open source tools.

A.1 Data Modeling and ETL Support

The open-source tools mentioned in Chaps. 2 and 3 are COMA 3.0 and Protègè for data modeling and schema integration as well as Pentaho Spoon and Talend for supporting the entire ETL process.

In Table A.2, Protègè and COMA 3.0 are described which provide support for the data modeling and schema integration. Protègè, for example, enables the definition of ontologies for data modeling that can also be used in order to resolve, for example, ambiguities for later schema integration. Plug-in PROMPT for Protègè additionally enables the mapping between ontologies. COMA 3.0 supports schema matching. Similar functionality is offered by, for example, Altova Mapforce [21]. Here, two schemas can be loaded and matched manually. The respective

W. Grossmann, S. Rinderle-Ma, *Fundamentals of Business Intelligence*,
Data-Centric Systems and Applications, DOI 10.1007/978-3-662-46531-8

Table A.1 Description schema for selected BI tools

Availability
Link, url
Existing documentation such as white papers
Licensing
Existing evaluations
Technical criteria
Operating system
Supported data formats
Extensibility
User interfaces
Evaluation
Functionality
Algorithms, techniques, visualizations
Data export/import, interfaces
Data preprocessing
Interactivity
Community, e.g., forum, blog

transformation queries or statements can be produced automatically in different ways, e.g., XQuery of Java. Altova MapForce is not open source, but commercial.

Note that both tools, Protègè and COMA 3.0, feature a Web-based version, i.e., WebProtege and COMA+ Web Edition 0.5.

The following Table A.3 provides an overview for Pentaho and Talend. Note that Pentaho is a full-fledged BI suite, i.e., consists of the *Business Analytics Platform*, *Kettle* as ETL tool, the *Report Designer*, and the *Marketplace*, complemented by further functionality provided by, for example, the *Aggregation Designer* [12].

Many open sources and commercial tools for supporting tasks of the extraction-transformation-load (ETL) process are available (for an overview, see, e.g., [10, 11]). To name a few, other open-source solutions are CloverETL [17] and Jitterbit [18].

A.2 Big Data

Tools on big data follow two directions, i.e., (a) supporting the analysis of big data and (b) supporting the extraction and integration of big data. Moreover, it is important to distinguish between tools and techniques. MapReduce, for example, is more a technique than a tool. It supports, for example, (a) by fragmenting an analysis job into smaller jobs for which their results are aggregated afterwards. Different tools offer MapReduce implementations, for example, Apache Hadoop [23] (open source). Also as mentioned in Sect. A.1, Pentaho offers support for big

Table A.2 Tools supporting schema integration: Protègè and COMA 3.0

	Protègè	COMA 3.0
Availability		
Link, url	[19]	[20]
Existing documentation	[22]	[2]
Licensing	MPL	AGPL
Existing evaluations	Several scientific papers, e.g., [3]	Several scientific papers, e.g., [1]
Technical criteria		
Operating system	Protègè Desktop: Linux, Windows, Mac OSX; WebProtege: Web-based	COMA 3.0: Windows, Linux; Coma+ Web Edition 0.5 web-based;
Supported data formats	OWL 2	SQL
Extensibility	Development of plugins based on OSGi	Java-based API
User interfaces		
Evaluation	Graphical GUI	Graphical GUI
Functionality		
Algorithms, techniques, visualizations	Reasoning can be connected	Support of different matching strategies
Data export/import, interfaces	Data export/import: RDF, XML, OWL	XSD, OWL
Data preprocessing		
Interactivity		Mapping candidates can be adapted by users
Community, e.g., forum, blog	Protègè and COMA 3.0 are both supported by a variety of documentations and fora	

data analytics via so-called Hadoop Shims [12]. Table A.4 compares two tools for big data analytics, i.e., Pentaho and H2O.

For the extraction and integration of big data (b), challenges volume, variety, and velocity are vital, as well. Let us first comment on challenge volume. As discussed for big data analytics, volume has been tackled among other approaches by suggesting NoSQL databases such as key value stores or graph databases. Chapter 3 mentions sonesDB which is a graph database. As sonesDB is nice for illustration, but seems to be no longer supported, in this section, we will introduce OrientDB as graph database. For challenge variety, Chap. 3 discussed BaseX as XML database. Finally, tackling velocity, Apache Storm supports the analysis of streaming data, i.e., data that is continuously injected into the analytical database and hence addresses the big data challenge velocity. Table A.5 summarized and compares the representatives mentioned before, i.e., OrientDB, BaseX, and Apache Storm.

Table A.3 ETL tools: Pentaho and Talend

	Pentaho	Talend
Availability		
Link, url	[12]	[13]
Existing documentation such as white papers	[14]	[15]
Licensing	GPLv2, LGPL, Apache, depending on the version	Basic: open source, extended functionality: commercial
Existing evaluations	[6]	Four software tests (in German) [16]
Technical criteria		
Operating system	Linux, Windows, Mac OSX	Linux, Windows, Mac OSX
Supported data formats	Variety, e.g., XML, SQL, text, csv	Variety of data formats, e.g., XML, SQL, text, csv, and standards, e.g., BPMN
Limitations		
Extensibility	Java-based API; Pentaho Marketplace stimulates testing and exchange of developed plug ins	Several extensions possible based on, e.g., java-based API, Web services
User interfaces		
Evaluation	ETL processes can be designed and traced in a graphical way	
Functionality		
Algorithms, techniques, visualizations	With Pentaho business analytics platform and report designer, various analyses and reports/visualizations can be created; in particular, aggregation designer supports OLAP analysis	Reporting and dashboards are not supported in the open source version
Data export/import, interfaces	Various interfaces to many of the existing tools/systems, e.g., databases, excel, XML	Talend specifically powers the connection to NoSQL sources such as Hive, MongoDB, Apache
Data preprocessing	Pentaho report designer enables reporting on data through the integration process, hence fosters data understanding; data can be preprocessed through Kettle	In the open source solutions, Talend open studio for data quality offers data profiling as well as graphical charts on the data in order to foster data understanding
Interactivity	Strongly supported throughout all phases of the BI process	
Community, e.g., forum, blog	Pentaho and Talend are both supported by a variety of documentations and fora	

Table A.4 Big data analytics tools: Pentaho and H2O

	Pentaho	H2O
Availability		
Link, url	[12]	[24]
Existing documentation such as white papers	[14]	[27]
Licensing	GPLv2, LGPL, Apache 2.0, depending on the version	Basic: Apache 2.0
Existing evaluations	[6]	Benchmarks [28]
Technical criteria		
Operating system	Linux, Windows, Mac OSX	Java-based platform, Web interface
Supported data formats	Variety, e.g., XML, SQL, text, csv	Local sources, Hadoop, EC2, multiple nodes
Extensibility	Java-based API; Pentaho Marketplace stimulates testing and exchange of developed plug ins	APIs to R and JSON
User interfaces		
Evaluation	Graphical UI	Graphical UI
Functionality		
Algorithms, techniques, visualizations	With Pentaho business analytics platform and report designer, various analyses and reports/visualizations can be created; in particular, aggregation designer supports OLAP analysis	Variety of analysis algorithms and techniques, e.g., regression, classification, neural networks
Data export/import, interfaces	Hadoop distributions via abstraction layer (shim); provision of predefined shims, but not for open source distribution	Import of csv, SQL
Data preprocessing	Reporting and transformation functions on different Hadoop clusters, e.g., Hive;	n.a.
Interactivity	Graphical UI	Graphical UI
Community, e.g., forum, blog	All supported by a variety of documentations and fora	

Table A.5 Big data integration tools: OrientDB, BaseX, and Apache Storm

	OrientDB	BaseX	Apache Storm
Availability			
Link, url	[29]	[30]	[25]
Existing documentation such as white papers	Documentation available on [29]	[31]	[26]
Licensing	Apache 2.0	BSD	Apache 2.0
Existing evaluations			
Technical criteria			
Operating system	Linux, Windows, Mac OSX	Linux, Windows, Mac OSX	Java-based framework
Supported data formats	Key value pairs, graphs	XML	Streams of key value pairs
Extensibility	Several APIs, e.g., Java API, SQL	Java-based API	Implementation in java or another language possible
User interfaces			
Evaluation	Graphical UI, Web frontend	Graphical UI	No GUI
Functionality			
Algorithms, techniques, visualizations	Supported query languages: SQL and Gremlin (graph-based)	Tree-based visualization of XML documents; support of XPath and XQuery	Enables the integration of data streams from different sources
Data export/import, interfaces	Import from RDBMS and Neo4J (graph database)	Import: XML, export: XML, HTML, csv	Can be used to feed streaming data into other systems such as Hive
Interactivity	Query language	Query language	n.a.
community, e.g., forum, blog	All supported by a variety of documentations and fora		

A.3 Visualization, Visual Mining, and Reporting

Modeling and layouting process models and instances is described in Sect. 4.2, and several tools are mentioned. As these tools provide much more functionality, an evaluation of their layouting functionality is presented directly in Sect. 4.2.2.

For the visualization of cross-sectional data, Chap. 4 used a number R packages for graphics, in particular, the packages `lattice` and `ggplot2`. The latter is probably one of the most advanced tools for producing statistical and other graphics. A tool for dynamic graphics for data exploration is GGobi [33]. GGobi can be used as stand-alone software or in connection with R in the package `ggobi`.

For dynamic and interactive graphics, the application of HighChart was shown. HighChart is a Javascript library which requires a HTTP server for local visualiza-

Table A.6 Visualization tools: R, HighCharts, Tableau Public

	R-graphics	HighChart	Tableau Public
Availability			
Link, url	[5]	[34]	[35]
Existing documentation such as white papers	On the website, [9]	On the website tutorial and publications	On the website tutorial
Licensing	GPL	Creative commons-NonCommercial	Free
Technical criteria			
Operating system	Linux, Mac OS, Unix, Windows	Javascript,jQuery, HTTP-Server	Windows, Mac OS X
Supported data formats	csv, excel	csv, excel, json, xml	csv, excel
Extensibility	Yes		
User interfaces			
Evaluation	Command line	JavaScript	GUI
Functionality			
Algorithms, techniques, visualizations	Statistical graphics, dynamic graphics for data exploration	Interactive graphics, dashboards	Interactive graphics, dashboards
Data export/import, interfaces	Interface to all DB	Export to JPEG, PNG, pdf, SVG	Web
Data preprocessing	Yes	Yes	Yes
Interactivity	Yes	Yes	Yes
Community, e.g., forum, blog	All supported by a variety of documentations and fora		

tion. For personal use and nonprofit organizations, high chart is freely available. An open-source tool for reporting and infographics is Tableau Public. Tableau Public has an easy-to-use interface and proposes a data visualization after parsing the uploaded data. Afterwards, the user can customize this basic layout in drag-and-drop style. The produced infographic can be published on the Web.

From the commercial products for visualization of cross-sectional data, we want to mention the SAS data mining software and the IBM SPSS Modeler which integrate visualization in the data mining activities. For an overview on R-graphics, HighChart, and Tableau Public see Table A.6.

There are numerous tools for Web-based graphics and infographics. Table A.7 lists ManyEyes, Gapminder, and Piktochart.

ManyEyes is an advanced visualization tool from IBM. The main emphasis is on sharing graphics within the ManyEyes community. Users can create their own graphics in easy steps or modify the graphics of other community members.

Gapminder is based on the Trendanalyzer software developed for the animated presentation of statistics, so-called *motion charts*. These charts show impressively the development of demographic, economic, or environmental facts. Many time

Table A.7 Web based visualizations: ManyEyes, Gapminder, Piktochart

	ManyEyes	Gapminder	Piktochart
Availability			
Link, url	[36]	[37]	[38]
Existing documentation such as white papers	Not much documentation available, introduction see [8]	On website	On website
Licensing	Free, data and visualizations are directly shared, copyright should be cleared	Free	Free
Technical criteria			
Operating system	Web browser		
Supported data formats	csv, spread sheet	Google spreadsheet	csv, spreadsheet
Extensibility			
User interfaces			
Evaluation	Interactive graphical user interface		
Functionality			
Algorithms, techniques, visualizations	Various basic layouts for graphics		
Data export/import, interfaces	Web publishing or download		
Data preprocessing	Limited		
Interactivity	Interactive editing of visualization		
Community, e.g., forum, blog	All supported by a community		

series at the national level as well as from international organizations are available on the site. The Trendanalyzer software is now available as interactive chart in the Google spreadsheet. This allows users the creation of motions charts with their own data.

Piktochart is an easy-to-use tool for creation of infographics. Numerous templates for infographics are available which can be adapted by the user. The created infographics allow interactive elements and are readable by search engines.

In addition, there are several other tools that enable the creation of infographics, e.g.:

- http://www.hongkiat.com/blog/infographic-tools/
- http://www.coolinfographics.com/tools/
- http://www.fastcodesign.com/3029239/infographic-of-the-day/30-simple-tools-for-data-visualization

A.4 Data Mining

In this book, we used R for data mining applications. Strictly speaking, R is a programming language for statistical computing and statistical graphics. It is a popular data mining tool for scientists, researchers, and students. Consequently, there exists a large community with fora and blogs which helps to learn how to use the numerous packages necessary for data mining. Besides data mining, a rich set of statistical methods for data preparation and graphics is available. R has strong object-oriented programming facilities which allow the extension of the software as soon as one has mastered the R language.

For usage of R as a BI production tool, the package `DBI` offers an interface to relational database systems. For big data, a number of solutions are provided. The package `data.table` is a fast tabulation tool as long as the data fit in the memory, e.g., 100 GB in RAM. For using Hadoop and the MapReduce approach, a number of packages have to be installed. For details, we refer to [4]. For a number of algorithms, there are also parallel implementations available.

From a more practical point of view, big data problems can be handled by sampling data from a database, develop a decision rule for the sample, and deploy the learned rule afterwards in the database. Thus, R can be used as an analysis tool in connection with an analytical sandbox. Alternatively, many a time it may be useful to aggregate the data and analyze the aggregated data.

Basically R is command line oriented, but a number of GUIs exist. For the development, RStudio offers an IDE, for data mining the Rattle GUI can be used, and Revolution Analytics provides a visual studio-based IDE. Further, the RWeka interface facilitates the application of Weka data mining algorithms in within R.

Weka is a Java-based data mining software which offers analysis tools similar to R. It also provides numerous data preprocessing techniques. With respect to data visualization, the facilities are not so comprehensive. The main user interface of Weka is the *Explorer* which provides in several panels access to the different data mining tasks. There exist panels for preprocessing, for variable selection, for visualization, and for different data mining techniques like classification, clustering, or association analysis. Weka supports two other BI tools: the Pentaho Business Analytics Platform uses Weka for data mining and predictive analytics; inside ProM Weka can be used for data mining, for example, in decision point analysis.

As a third open-source data mining software, we want to mention RapidMiner. Due to the fact that it has an easy-to-use interface, it is one of the most popular data mining tools in BI. It captures the entire life cycle of a BI application, allows model management, and is well designed for the collaboration between the business analyst and the data scientist. With respect to analysis capacities, it offers algorithms for data preparation and for analysis. Algorithms from external sources like R or Weka can be included in the analysis. Further, it supports the analysis of data in the memory, in databases, in the cloud, and supports Hadoop. For an overview on R, RapidMiner, and Weka see Table A.8.

Table A.8 Data mining tools: R, RapidMiner, Weka

	R	RapidMiner	Weka
Availability			
Link, url	[5]	[40]	[42]
Existing documentation such as white papers	On the website: manuals, R journal, FAQs	On the website: documentation	[43]
Licensing	GPL	AGPL	GPL
Technical criteria			
Operating system	Linux, (Mac) OSX, Windows	All platforms (Java based)	All platforms (Java based)
Supported data formats	Basically csv, but various other data formats are supported		
Extensibility	Yes	Yes	Yes
User interfaces			
Evaluation	Command line/various GUIs	GUI	Command line, various GUIs
Functionality			
Algorithms, techniques, visualizations	Algorithms for all data mining algorithms, various visualization techniques		
Data export/import, interfaces	Interfaces to all DB systems		
Data preprocessing	Supported by various algorithms		
Interactivity	Depending on the application		
Community, e.g., forum, blog	[32]	[41]	[44]

From the commercial products, the SAS data mining software and the IBM SPSS Modeler are two powerful data mining tools. Both products offer a visual interface and allow applications without programming.

A.5 Process Mining

Table A.9 summarizes details on the process mining tool ProM which is applied in Chap. 7. There is no comparable tool available as open-source solution; hence, only ProM is introduced here. Nonetheless, one can mention Disco [45] as commercial process mining tool which developed from ProM.

Table A.9 Process mining tool: ProM

	ProM
Availability	
Link, url	[39]
Existing documentation such as white papers	[7]
Licensing	
Existing evaluations	ProM 5.2: CPL, ProM 6.2: LGPL, ProM 6.3: LGPL, ProM 6.4: GPL
Technical criteria	
Operating system	All platforms
Supported data formats	Log formats: MXML, XES, and csv; process model formats: PNML, YAWL specification, BPEL, CPN
Extensibility	Development of java-based plug ins
User interfaces	
Evaluation	n.a.
Functionality	
Algorithms, techniques, visualizations	Several algorithms for process discovery, conformance checking, filtering, organizational mining, etc.; visualizations as, e.g., graph-based process models or dotted charts
Data export/import, interfaces	Import: MXML, XES, csv, PNML, etc.; export: process models as graphics, e.g., eps, svg; Petri Nets: pnml; logs: MXML, XES; reports: HTML
Data preprocessing	Filtering
Interactivity	Partly, e.g., mouseover and dragging of social networks
Community, e.g., forum, blog	Supported by fora, developer support, ProM task force

A.6 Text Mining

All the data mining software products reviewed in Sect. A.4 offer text mining facilities for classification and cluster analysis of text data represented as document term matrices. The applicability of these tools essentially depend on the data sources which can be read by the software, the available transformations for preprocessing, the availability of linguistic knowledge for the language under consideration, and the analysis algorithms. For example, the package `tm` can process a number of formats by using plugins. Regarding the linguistic knowledge, part-of-speech tagging and stemming can be done, and WordNet can be accessed as English lexical database. For analysis, a number of advanced statistical models like topic maps or specific classification and cluster algorithms can be used.

An open-source tools which puts more emphasis on natural language processing is *GATE*. GATE stands for General Architecture for Text Engineering and was

developed at the University Sheffield. On the homepage [46], one can find an extensive documentation.

GATE consists of a number of components. A core component of GATE is an information extraction tool which offers modules for tokenization, part-of-speech tagging, sentence splitting, entity identification, semantic tagging, or referencing between entities. A number of plugins offer applications for data mining algorithms or the management of ontologies. Another important component allows indexing and searching of the linguistic and semantic information generated by the applications.

GATE supports analysis of text documents in different languages and in various data formats. The GATE Developer is the main user interface that supports the loading of documents, the definition of a corpus, the annotations of the documents, and the definition of applications.

References

1. Arnold P, Rahm E (2014) Enriching ontology mappings with semantic relations. Data Knowl Eng 93:1–18
2. COMA 3.0 CE, program description (2012) Database chair. University of Leipzig, Leipzig
3. Fridman NN, Tudorache T (2008) Collaborative ontology development on the (semantic) web. In: Symbiotic relationships between semantic web and knowledge engineering. Papers from the 2008 AAAI spring symposium, Technical Report SS-08-07, AAAI
4. Prajapati V (2013) Big data analytics with R and hadoop. http://it-ebooks.info/book/3157/. Accessed 11 Nov 2014
5. R Core Team (2014) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. http://www.R-project.org. Accessed 12 Dec 2014
6. Tuncer O, van den Berg J (2012) Implementing BI concepts with Pentaho, an evaluation. Delft University of Technology, Delft
7. van der Aalst WMP (2011) Process mining: discovery, conformance and enhancement of business processes. Springer, Heidelberg
8. Viegas FB, Wattenberg M, van Ham F, Kriss J, McKeon M (2007) Manyeyes: a site for visualization at internet scale. IEEE Trans Vis Comput Graph 13(6):1121–1128
9. Wickham H (2009) ggplot2: Elegant graphics for data analysis. Springer, New York
10. http://www.databaseanswers.org/modelling_tools.htm. Accessed 4 Dec 2014
11. http://www.etltools.net/free-etl-tools.html. Accessed 4 Dec 2014
12. http://community.pentaho.com/. Accessed 4 Dec 2014
13. http://www.talend.com/products/big-data. Accessed 4 Dec 2014
14. http://wiki.pentaho.com/display/EAI/Latest+Pentaho+Data+Integration+%28aka+Kettle%29+Documentation. Accessed 4 Dec 2014
15. http://www.talendforge.org/tutorials/menu.php. Accessed 4 Dec 2014
16. https://de.talend.com/resources/whitepapers?field_resource_type_tid[]=79. Accessed 4 Dec 2014
17. http://www.cloveretl.com/. Accessed 4 Dec 2014
18. http://www.jitterbit.com/. Accessed 4 Dec 2014
19. http://protege.stanford.edu/. Accessed 4 Dec 2014
20. http://dbs.uni-leipzig.de/Research/coma.html. Accessed 4 Dec 2014
21. http://www.altova.com/mapforce.html. Accessed 4 Dec 2014
22. http://protegewiki.stanford.edu/wiki/ProtegeDesktopUserDocs. Accessed 4 Dec 2014

23. http://hadoop.apache.org/. Accessed 5 Dec 2014
24. http://docs.0xdata.com/. Accessed 5 Dec 2014
25. https://storm.apache.org/. Accessed 5 Dec 2014
26. https://storm.apache.org/documentation/Home.html. Accessed 5 Dec 2014
27. http://docs.0xdata.com/. Accessed 5 Dec 2014
28. http://docs.0xdata.com/benchmarks/benchmarks.html. Accessed 5 Dec 2014
29. http://www.orientechnologies.com/orientdb/. Accessed 5 Dec 2014
30. http://basex.org/. Accessed 5 Dec 2014
31. http://docs.basex.org/wiki/Main_Page. Accessed 5 Dec 2014
32. http://www.inside-r.org/. Accessed 12 Dec 2014
33. http://www.ggobi.org/. Accessed 5 Dec 2014
34. http://www.highcharts.com/. Accessed 12 Dec 2014
35. http://www.tableausoftware.com/public/. Accessed 12 Dec 2014
36. http://www-969.ibm.com/software/analytics/manyeyes/. Accessed 9 Dec 2014
37. http://www.gapminder.org/. Accessed 12 Dec 2014
38. http://piktochart.com/. Accessed 12 Dec 2014
39. http://www.processmining.org/. Accessed 11 Dec 2014
40. https://rapidminer.com/. Accessed 12 Dec 2014
41. http://forum.rapid-i.com/. Accessed 12 Dec 2014
42. http://www.cs.waikato.ac.nz/ml/weka/. Accessed 12 Dec 2014
43. http://www.cs.waikato.ac.nz/ml/weka/book.html. Accessed 12 Dec 2014
44. http://www.cs.waikato.ac.nz/ml/weka/help.html. Accessed 12 Dec 2014
45. http://www.fluxicon.com/disco/. Accessed 19 Dec 2014
46. https://gate.ac.uk/. Accessed 12 Dec 2014