

## Concept and Objectives

**n="false" UnhideWhenUsed="false"**  
**Name="Medium Shading 1 Accent 4" />**  
**<w:LsdException Locked="false" Priority="64"**  
**SemiHidden="false" UnhideWhenUsed="false"**  
**Name="Medium Shading 2 Accent 4" />**  
**<w:LsdException Locked="false" Priority="65"**  
**SemiHidden="false" UnhideWhenUsed="false"**  
**Name="Medium List 1 Accent 4" />**  
**<w:LsdException Locked="false" Priority="66"**  
**SemiHidden="false" UnhideWhenUsed="false"**  
**Name="Medium List 2 Accent 4" />**  
**<w:LsdException Locked="false" Priority="67"**  
**SemiHidden="false" UnhideWhenUsed="false"**  
**Name="Medium Grid 1 Accent 4" />**  
**<w:LsdException Locked="false" Priority="68"**  
**SemiHidden="false" UnhideWhenUsed="false"**  
**Name="Medium Grid 2 Accent 4" />**  
**<w:LsdException Locked="false" Priority="69"**  
**SemiHidden="false" UnhideWhenUsed="false"**  
**Name="Medium Grid 3 Accent 4" />**  
**<w:LsdException Locked="false" Priority="70"**  
**SemiHidden="false" UnhideWhenUsed="false"**  
**Name="Dark List Accent 4" />** **<w:LsdException**  
**Locked="false" Priority="71" SemiHidden="false"**  
**UnhideWhenUsed="false" Name="Colorful**  
**Shading Accent 4" />** **<w:LsdException**

**Locked="false" Priority="72" SemiHidden="false"  
UnhideWhenUsed="false" Name="Colorful List  
Accent 4" /> <w:LsdException Locked="false"  
Priority="73" SemiHidden="false"  
UnhideWhenUsed="false" Name="Colorful Grid  
Accent 4" /> <w:LsdException Locked="false"  
Priority="60" SemiHidden="false"  
UnhideWhenUsed="false" Name="Light Shading  
Accent 5" /> <w:LsdException Locked="false"  
Priority="61" SemiHidden="false"  
UnhideWhenUsed="false" Name="Light List  
Accent 5" /> <w:LsdException Locked="false"  
Priority="62" SemiHidden="false"  
UnhideWhenUsed="false" Name="Light Grid  
Accent 5" /> <w:LsdException Locked="false"  
Priority="63" SemiHidden="false"  
UnhideWhenUsed="false" Name="Medium  
Shading 1 Accent 5" /> <w:LsdException  
Locked="false" Priority="64" SemiHidden="false"  
UnhideWhenUsed="false" Name="Medium  
Shading 2 Accent 5" /> <w:LsdException  
Locked="false" Priority="65" SemiHidden="false"  
UnhideWhenUsed="false" Name="Medium List 1  
Accent 5" /> <w:LsdException Locked="false"  
Priority="66" SemiHidden="false"  
UnhideWhenUsed="false" Name="Medium List 2  
Accent 5" /> <w:LsdException Locked="false"  
Priority="67" SemiHidden="false"**

**UnhideWhenUsed="false" Name="Medium Grid 1  
Accent 5" /> <w:LsdException Locked="false"  
Priority="68" SemiHidden="false"  
UnhideWhenUsed="false" Name="Medium Grid 2  
Accent 5" /> <w:LsdException Locked="false"  
Priority="69" SemiHidden="false"  
UnhideWhenUsed="false" Name="Medium Grid 3  
Accent 5" /> <w:LsdException Locked="false"  
Priority="70" SemiHidden="false"  
UnhideWhenUsed="false" Name="Dark List Accent  
5" /> <w:LsdException Locked="false"  
Priority="71" SemiHidden="false"  
UnhideWhenUsed="false" Name="Colorful  
Shading Accent 5" /> <w:LsdException  
Locked="false" Priority="72" SemiHidden="false"  
UnhideWhenUsed="false" Name="Colorful List  
Accent 5" /> <w:LsdException Locked="false"  
Priority="73" SemiHidden="false"  
UnhideWhenUsed="false" Name="Colorful Grid  
Accent 5" /> <w:LsdException Locked="false"  
Priority="60" SemiHidden="false"  
UnhideWhenUsed="false" Name="Light Shading  
Accent 6" /> <w:LsdException Locked="false"  
Priority="61" SemiHidden="false"  
UnhideWhenUsed="false" Name="Light List  
Accent 6" /> <w:LsdException Locked="false"  
Priority="62" SemiHidden="false"  
UnhideWhenUsed="false" Name="Light Grid**

**Accent 6" /> <w:LsdException Locked="false"  
Priority="63" SemiHidden="false"  
UnhideWhenUsed="false" Name="Medium  
Shading 1 Accent 6" /> <w:LsdException  
Locked="false" Priority="64" SemiHidden="false"  
UnhideWhenUsed="false" Name="Medium  
Shading 2 Accent 6" /> <w:LsdException  
Locked="false" Priority="65" SemiHidden="false"  
UnhideWhenUsed="false" Name="Medium List 1  
Accent 6" /> <w:LsdException Locked="false"  
Priority="66" SemiHidden="false"  
UnhideWhenUsed="false" Name="Medium List 2  
Accent 6" /> <w:LsdException Locked="false"  
Priority="67" SemiHidden="false"  
UnhideWhenUsed="false" Name="Medium Grid 1  
Accent 6" /> <w:LsdException Locked="false"  
Priority="68" SemiHidden="false"  
UnhideWhenUsed="false" Name="Medium Grid 2  
Accent 6" /> <w:LsdException Locked="false"  
Priority="69" SemiHidden="false"  
UnhideWhenUsed="false" Name="Medium Grid 3  
Accent 6" /> <w:LsdException Locked="false"  
Priority="70" SemiHidden="false"  
UnhideWhenUsed="false" Name="Dark List Accent  
6" /> <w:LsdException Locked="false"  
Priority="71" SemiHidden="false"  
UnhideWhenUsed="false" Name="Colorful  
Shading Accent 6" /> <w:LsdException**

```
Locked="false" Priority="72" SemiHidden="false"
UnhideWhenUsed="false" Name="Colorful List
Accent 6" /> <w:LsdException Locked="false"
Priority="73" SemiHidden="false"
UnhideWhenUsed="false" Name="Colorful Grid
Accent 6" /> <w:LsdException Locked="false"
Priority="19" SemiHidden="false"
UnhideWhenUsed="false" QFormat="true"
Name="Subtle Emphasis" /> <w:LsdException
Locked="false" Priority="21" SemiHidden="false"
UnhideWhenUsed="false" QFormat="true"
Name="Intense Emphasis" /> <w:LsdException
Locked="false" Priority="31" SemiHidden="false"
UnhideWhenUsed="false" QFormat="true"
Name="Subtle Reference" /> <w:LsdException
Locked="false" Priority="32" SemiHidden="false"
UnhideWhenUsed="false" QFormat="true"
Name="Intense Reference" /> <w:LsdException
Locked="false" Priority="33" SemiHidden="false"
UnhideWhenUsed="false" QFormat="true"
Name="Book Title" /> <w:LsdException
Locked="false" Priority="37"
Name="Bibliography" /> <w:LsdException
Locked="false" Priority="39" QFormat="true"
Name="TOC Heading" /> </w:LatentStyles> </xml>
```

## Motivation and Main Idea

CUBIST is a visionary approach that leverages Business Intelligence to a new level of precise, meaningful and user-friendly analytics of data. CUBIST follows a best-of-breed approach that combines capabilities of

## Cubist: Objectives

Business Intelligence, Semantic Technologies, and Visual Analytics.

CUBIST in a nutshell: Developing an approach for **semantic and user-friendly Business Intelligence** by

- augmenting Semantic Technologies with BI capabilities, and
- providing conceptually relevant and user friendly visual analytics.

**Business Intelligence** (BI) refers to technologies, applications and practices of gathering, representing and analyzing data to support decision making in enterprises. According to a 2009 Gartner report, BI software is going to use up to 40% of company IT budgets in the next four years. Constantly growing amounts of data, complicated and rapidly changing interactions in the economy and an emerging trend of incorporating unstructured data into analytics bring new challenges to BI:

*Unstructured data.* One of the problems is the fact that trends visible in unstructured data (e.g., in the news) need some time to appear in a data warehouse. It is expected by the Data Warehouse Institute that we will soon witness a significant change of paradigm in Data Warehousing and BI: the proportion of unstructured data in all data is estimated to be over 80%, leading to the fraction of structured data used in Data Warehouses decreasing significantly in favour of unstructured data. Information extraction to provide unstructured data sources to BI brings new challenges regarding the quality of extracted data. Currently BI engines do not handle the meaning of data so detecting contradicting information is difficult.

- *No clear semantics of data.* Another problem is how to manage reports created in the data warehouse. Unnecessary reports are often created because existing results are hard to find. Semantic descriptions do not exist to make results of previous analyses searchable.
- *BI complexity.* Performing BI can require spanning over data from several business processes from various data sources. Expertise from a number of domains may be required to interpret the data. ?Task complexity? and ?lack of expertise? are key quality metrics in decision making. There lacks a reasoning mechanism to facilitate this process.
- *Coarse-grained BI.* A further issue is that users are not provided with tools that present information adapted to their needs. According to a 2008 study by Information Week the complexity of BI tools and their interfaces becomes the biggest barrier for success of these systems.

**Semantic Technologies** (ST) are proposed by CUBIST as key elements in new BI systems to extract and utilise the meaning of data to address these challenges. ST refers to a set of technologies that have grown in popularity over the last decade, due to their potential for automation. ST provide a formal foundation to represent and reason over knowledge. Research within the Semantic Web has standardised a set of knowledge representation languages including the Resource Description Framework (RDF), an abstract model for describing things by making statements about resources and their properties. RDF is supported by logical formalisms and related reasoning techniques and query languages such as SPARQL.

The advantages of using ST can be summarised as:

## Cubist: Objectives

- An open conceptual model for representing and integrating heterogeneous and disparate databases and schemas, independent of the original data representation means or formats
- The ability to declare and apply formal rules and reasoning logic to data in a semantic repository in order to infer implicit facts and discover relationships that would otherwise not be discovered

From an application point of view, ST are well suited for:

- Enhanced information (and meaning) extraction and semantic annotation of documents
- Federating different data and information silos and providing a unified view on them
- Deriving new information from different sources and performing reasoning on them
- Advanced (semantic) search and navigation facilities over the federated information

**Combining ST and BI:** The following table compares classical BI with ST

| Feature                                                  | Classical BI | Semantic Technologies |
|----------------------------------------------------------|--------------|-----------------------|
| Federating data from unstructured and structured sources | bad          | good                  |
| Explicit meaning of data / explicit relations            | bad          | good                  |
| Querying/accessing huge amounts of data                  | good         | average               |
| OLAP functionalities                                     | good         | bad                   |
| Agile DW schema                                          | bad          | good                  |
| Inferring implicit facts                                 | bad          | good                  |
| ETL flexibility                                          | bad          | average               |
| Data reuse                                               | bad          | good                  |
| Use of external datasources / data location independence | bad/average  | good                  |

ST introduce benefits to the BI lifecycle (see Figure 1 for comparison of BI and BI+ST lifecycles):

- *Enhanced Extract Transform Load (ETL)* ? integration of heterogeneous datasources, and improvements over traditional approaches to unstructured data analysis (such as Information Extraction or Text Mining)
- *Flexible data warehouse design* ? by employing the RDF conceptual model instead of traditional ?star schema? approaches
- *Enhanced data cubes* ? by employing RDF triplestores for a data warehouse (information warehouse), implicit facts can be inferred within the database, using sound logical rules
- *Enhanced search and discovery* ? by employing the semantics of the underlying data for more powerful query answering and matchmaking

To realise these benefits CUBIST will address issues relating to traditional ST:

- Inferior performance with respect to data volumes when compared to state-of-the-art enterprise data warehouses
- Lack of Online Analytical Processing (OLAP) extensions to the RDF query language (SPARQL)

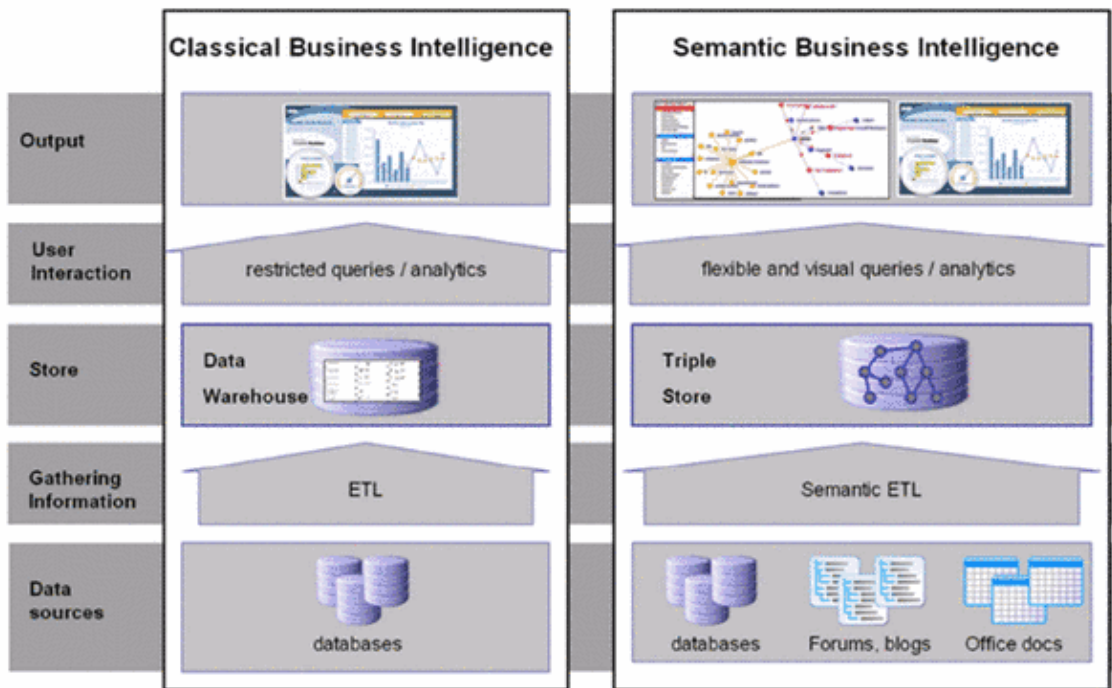


Figure 1: lifecycle of classical and semantic BI

**Analyzing the Information by Means of Interactive Visualizations:** User interaction with data is closely related to its visual representation. Interactive visualization technology displays numerous aspects of multidimensional data using interactive pictures and charts. The colour, size, shape and motion of objects represent the multidimensional aspects of the data. Information visualization has been defined as 'the use of computer-supported, interactive, visual representations of abstract data to amplify cognition'. A good visual representation can amplify user cognition by providing more information, faster, with less cognitive effort. BI platform vendors are currently promoting these technologies as an alternative and enrichment to traditional reporting and online analytical processing capabilities. CUBIST will provide BI users with visualization and interaction techniques enhanced by Semantic Technologies, such as Formal Concept Analysis (FCA).

FCA captures hitherto undiscovered patterns in data. Concepts are formalised as groups of objects associated with groups of attributes. Hierarchical relationships between these groups are formed and visualized and can be used for information retrieval. Information organised in this way has a close correlation to human perception. Thus, FCA has been used as the basis for semantic search engines and as a means of organising information based on its meaning. FCA has also been used to mine data for groups of items and their associated transactions to identify, for example, groups of products that are frequently purchased together.

## Objectives

**Objective 1: Semantic ETL.** Unlike classical BI, CUBIST will provide comprehensive methods to bring unstructured data into analytics. SETL will use lineage information, metadata of data sources, semantic descriptions, and meaning of extracted data to support error detection within data sources and the SETL process. Semi-automated mapping of structured data sources to the RDF model will reduce the complexity of data integration. This objective will be approached by leveraging ETL best practises, orchestrated and wrapped by a semantic layer.

## Cubist: Objectives

*Expected Result:* Semantically enriched lineage information, error detection within extracted data and CUBIST information warehouse and a SETL component that provides SPARQL endpoints for various data sources.

**Objective 2: Semantic Warehouse.** CUBIST will employ an RDF triple store and ontology as the backbone for the information warehouse, to improve performance and reduce the complexity of the integration of heterogeneous data sources. ST will enrich BI by enabling the discovery of new implicit information through logical reasoning. The standard RDF query language SPARQL will be extended by OLAP functionalities for complex aggregates and analysis. The information warehouse will use advanced indexing and materialisation techniques known from state-of-the-art data warehousing to improve the performance of the RDF triple store. A layer within the warehouse will integrate the triple store with the FCA-based visual analytics.

*Expected Result:* OLAP extensions for SPARQL, discovery mechanism for finding implicit information, triple store based information warehouse, integration of triple store with FCA.

**Objective 3: FCA-Based Advanced Visual Analytics.** On top of the Semantic Warehouse, CUBIST will provide ways to visualize and explore hitherto undiscovered BI using FCA.

*Expected Result:* High performance formal concept miner, large scale FCA capability, FCA visualization of BI.