



PlanetData

Network of Excellence

FP7 – 257641

D6.1 Training Curriculum

Coordinator: Simeona Cruz Pellkvist (STI2)

With contributions from: Lyndon Nixon (STI2), Oscar Corcho (UPM), Jean Paul Calbimonte (UPM), Ying Zhang (CWI), Chris Bizer (FUB), Graham Hench (STI2)

1st Quality reviewer: Zoltán Miklós (EPFL)

2nd Quality reviewer: Thomas Bauereiß (UIBK)

Deliverable nature:	Report (R)
Dissemination level: (Confidentiality)	Public (PU)
Contractual delivery date:	31-03-2011
Actual delivery date:	31-03-2011
Version:	1.0
Total number of pages:	16
Keywords:	curriculum, training

Abstract

The PlanetData Network of Excellence intends to ensure that the technologies and tools for large scale data management will be transferred via training to technology adopters and students, and that this training will be delivered through multiple channels. In this deliverable we present the result of the training curriculum that has been designed by the expert institutions of the Network of Excellence, which focuses on the relevant fields of large-scale data management. This field consists of Semantic Technology, Database Technology, Linked Data and Data Streams. The consortium expects that the curriculum will be used as a standard guidance for the graduate students and professionals that will be involved in large-scale data management. The proposed curriculum is expected to be modified and adjusted to the developing research situation in the future.

[End of abstract]

Executive summary

In responding to the PlanetData Network of Excellence objectives and fulfilling the impact activity, the consortium has designed and prepared a set of curricula. The purpose of designing this curriculum is to encourage and educate future researchers, organisations, technology vendors and collaborators, through a specific targeted training.

This deliverable focuses on how the proposed curriculum has been designed by the PlanetData Network of Excellence. In the first part of this deliverable we describe who the PlanetData NoE curriculum is targeted at. The targeted people will be the university students and professionals. Ways of delivering the curriculum is described in 4 ways. They are Self-training, Distance Learning, Webinars and On-Site Training. We also describe what should be the prerequisite qualification for the targeted people.

The second part of this deliverable contains the proposed curriculum, which is the main part of the deliverable. The curriculum consists of four topics that are relevant to large-scale data management. These four topics are Semantic Technology, which includes the Semantic Web, Database Technology, and Linked Data and Data Streams. The proposed curriculum can be modified, mixed and matched in order to better suit the trainees' previous level of knowledge and intended goals of the training.

In the end we conclude that since the project has just started when this deliverable is written, therefore we designed the curriculum in such a way that it can be modified and adjusted to the current situation later when it is implemented.

Document Information

IST Project Number	FP7 - 257641	Acronym	PlanetData
Full Title	PlanetData		
Project URL	http://www.planet-data.eu/		
Document URL			
EU Project Officer	Leonhard Maqua		

Deliverable	Number	D6.1	Title	Training Curriculum
Work Package	Number	WP6	Title	Training

Date of Delivery	Contractual	M6	Actual	M6
Status	version 1.0		final <input type="checkbox"/>	
Nature	prototype <input type="checkbox"/> report <input checked="" type="checkbox"/> dissemination <input type="checkbox"/>			
Dissemination level	public <input checked="" type="checkbox"/> consortium <input type="checkbox"/>			

Authors (Partner)	STI2			
Responsible Author	Name	Simeona Cruz Pellkvist	E-mail	simeona.pellkvist@sti2.org
	Partner	STI2	Phone	+43- 1 23 64 002

Abstract (for dissemination)	
Keywords	curriculum, training

Version Log			
Issue Date	Rev. No.	Author	Change
14-01-2011	0.1	Simeona Cruz Pellkvist	Initial Draft ToC & Semantic Technology curriculum structure
19-01-2011	0.2	Oscar Corcho & Jean Paul Calbimonte	Data Stream curriculum structure
22-01-2011	0.3	Ying Zhang	Databases technology curriculum structure
24-01-2011	0.4	Chris Bizer	Link Data curriculum structure
27-01-2011	0.8	Simeona Cruz Pellkvist	Content input
31-01-2011	0.9	Lyndon Nixon	Addition and modification
03-03-2011	1.0	Simeona Cruz Pellkvist	Completing rest of the items
10-03-2011	1.5	Simeona Cruz Pellkvist	Refinement after QA Review
28-03-2011	1.6	Simeona Cruz Pellkvist	Further refinement prior to final EC submission

Table of Contents

Executive summary.....	3
Document Information.....	4
Table of Contents.....	5
Abbreviations.....	6
1 Introduction.....	7
1.1 Purpose.....	7
1.2 Target Level of knowledge.....	7
1.3 Prerequisite qualifications.....	8
1.4 Delivery Methods.....	8
2 Structure of the curriculum.....	9
2.1 Semantic Technology.....	9
2.2 Databases technology.....	11
2.3 Linked Data.....	12
2.4 Data Streams.....	13
3 Conclusion.....	15
References.....	16

Abbreviations

AS3AP – ANSI SQL Standard Scalable and Portable

DBMSs – Database Management System standard

DL – Distance Learning

DoW – Description of Work

EU – European Union

IRS-III – Internet Reasoning Service-III

IT – Information Technology

KR – Knowledge Representation

OWL – Ontology Web Language

OWL-S – OWL for Services/ OWL-based Web Service Ontology (formerly DAML-S)

PhD – Doctor of Philosophy

RDF/S – Resource Description Framework / Schema

RDBMS - Relational Database Management System

SAWSDL – Semantic Annotations for Web Services Description Language

SPARQL – SPARQL Protocol and RDF Query Language

SQL – Structured Query Language

TPC-* - Transaction Processing Performance Council

WP – Work Package

WSM – Web Service Module

WSMO – Web Service Modeling Ontology

WSMT – Web Service Modeling Toolkit

WSMX – Web Service Modeling eXecution environment

XML - Extensible Markup Language

1 Introduction

The PlanetData Network of Excellence aims to enable maintainable large-scale access to structured data that has been exposed by organisations within the European community. To support the future production and consumption of large scales of data, there should be training and education programs for the interested organisations in the industry and the academic community in large-scale data management and the underlying approaches and technologies.

In order to maintain and sustain the results of the PlanetData network, we aim to make possible training for both the academic and industrial communities. The training should be based on a curriculum that has been designed according to the PlanetData network to cover all necessary approaches and technologies, which apply to large-scale data management. This deliverable will present the initial foreseen content of the PlanetData curriculum and discuss possible implementation of the curriculum in academic and professional training.

1.1 Purpose

In order to support the objectives of the PlanetData project, we need to ensure that there is access to appropriate guidance in the technologies and approaches, which will underlie future large-scale data management tasks. The PlanetData consortium proposes as a first step a curriculum. A curriculum is important for both trainers and learners in order to achieve more understanding of what is going to be expected to have been covered by the end of the training so that the learner achieves the necessary level of competency in the appropriate technologies and approaches in order to be able to consider themselves skilled in large scale data management. Therefore we have to specify in advance what we are trying to achieve and how we are going about it. The given curriculum is set up to offer a better grasp of the principals of large-scale data management. The achievement in our case is to cover the latest state of the art in the different fields related to data management but also the unifying methodologies developed by the PlanetData Network of Excellence. In short the curriculum is a benchmark of the training. The curriculum acts as a guideline for creating concrete training and education offers in the course of, and following, the project.

1.2 Target Level of knowledge

This section describes the target of the PlanetData project, in terms of the expected candidates for training and education based on the PlanetData curriculum. The curriculum should fit according to the learner's level of knowledge in large-scale data management. In this deliverable, the PlanetData curriculum will only focus both general categories of learner: "professionals" and "graduates" - as our target of the curriculum. Here we describe the typical foreseen candidate for the curriculum.

- **Graduates**

The PlanetData Graduate curriculum is targeted towards university graduate-level students, i.e. at Masters and PhD level in the EU context (as defined in the Bologna process). This target has already achieved a core broad understanding of Computing Science topics, such as programming skills, software development approaches, databases and computing systems (e.g. networks), and now intend to deepen their knowledge in a particular topic, and in this case particularly dig deeper into the complexities of large-scale data management. The participants will be taught through regular seminars on curriculum topics as well as participation in dedicated events such as the PlanetData summer schools where there will be keynote presentations and tutored sessions run by large-scale data management experts such as PlanetData members.

- **Professionals**

The PlanetData professional candidate is a learner who needs a broader and deeper understanding of how to deal and manage large-scales of data in the context of their professional activities. At first the participants should undertake an exam in order to identify, which level the participant, would belong to. During the training the participants are also expected to complete reading assignments, work through hands-on exercises, and devote some time after the course hours to familiarize themselves with the concepts presented. In professional training, it would be expected that the concrete training material would focus more on providing real business application examples and relating the learnt topics to those examples, as well as additional hands-on work with tools and technologies. There will be a final exam on the last training day.

The successful participants in passing the exam will receive a credential proof of each level of expertise. The PlanetData professional curriculum could be divided further into several levels. The levels depend on the usual levels of competency recognised by the context in which a specific training offer is applied. This would be based on the purpose and qualifications of each participant. As an example, the Semsphere Company¹, which offers training in semantic technology, recognises two levels (specialist and professional) and uses an online exam to determine a candidate's proficiency. We expect a similar approach can be used in PlanetData training.

1.3 Prerequisite qualifications

As a basic prerequisite for both level of candidate are require having a basic level of computer literacy. . A background in IT development and engineering is an advantage but not necessarily a requirement. Most important is that candidates have a keen interest in learning about the foundations of knowledge systems and the Semantic Technology, Linked Data, Data Streams and Databases, and that they have an intention to produce, consume and manage large-scale data. As already mentioned, the precise pre-requirements will depend upon the candidate's goals and upon the training context.

1.4 Delivery Methods

To get a better understanding of each topic of the training curriculum, there are different methods of delivering the course material. This subsection will outline how the curriculum will be delivered to the candidate mentioned in part 1.2. The delivery method will depend on the candidate's context and preference, in which they could choose between online training and offline training.

- **Self-training**

Using this delivery method, the candidate trains in and learns the course material him- or herself. The course material could be self-obtained (purchased directly) or acquired online, e.g. via an online portal made available by the training provider. The materials could be used in printed or online form. The self-learning consists of two parts: course material and the certification examination. It is up to the student how he or she learns the materials and can take the certification examination. In this context there is usually no direct trainer support and onsite tutorial. It is only the official course material, which the candidate learns by themselves, and the taking of the certification examination.

- **Distance learning**

Distance learning is the most convenient delivery method compared to other alternatives. This method is flexible since the learner can just take an exam or do a full-qualified training. The exam can be taken directly, since the learner has already done self-learning.

The distance learning (DL) is an extensive online course lasting for a couple of months with enhanced support material and possibilities for interaction with the trainer and other trainees respectively.

The DL features could consist of live webinars, guided tutorials, the official course material and access to recorded live webinars. The distance learning can also include the certification examination. The DL offer is more expensive than the Self-Learning Package

- **Webinars**

It is an online lecture with some interactive elements. This online training includes interactive presentation, lecture, workshop or seminar that is transmitted over the web. The participant can directly interact in giving, receiving and discussing material of the course in real time. It will be costly compare to other online delivery methods.

- **On-site training**

On-site training is a very effective and high-level interaction between the trainer and the learner using this method. It is a delivery method where the trainer will come to the training site in order to train several participants simultaneously. It is the best way to learn and get in contact with trainers for discussions and direct questions.

¹ <https://www.semsphere.com/index.php>

2 Structure of the curriculum

In this section, the PlanetData network of excellence proposes different topical curricula that cover large-scale data management. These various topics will include databases, data and Web mining, knowledge representation, Web reasoning, stream processing and Linked Open Data. PlanetData will integrate various disciplines that can contribute to enhance large-scale data management across the four objectives. These four objectives are integration in research, in data provisioning, in data management and in impact creation. As already mentioned earlier, this proposed curriculum will provide a basis for the development of graduate or professional training courses, to ensure an organized guidance for their content focusing on the necessary skills and technologies for large-scale data management. The proposed curricula can be modified, mixed and matched in order to better suit the trainees' previous level of knowledge and intended goals of the training. The proposed curriculum could be later distributed as several modules to form a shared structure of training development. Each structure will consist of different modules (for different levels of competency), from "Introductory level" to "Proficient level". The topics were chosen according to the insight of the PlanetData network into the necessary underlying technologies which will be part of future large scale data management approaches:

- Semantic technologies, including semantic web services
- Database technologies
- Linked Data
- Data streams

Within each, we present as a curriculum the sub-topics which are relevant for future large-scale data managers, i.e. persons who will need to be able to use computer systems and software to produce, consume, and manage large scales of data. These sub-topics are ordered in terms of candidate proficiency to learn them, i.e. we start with introductory sub-topics within the high level topic and progress to more advanced topics which need already higher proficiency in the topic, as gained through earlier sub-topics. As a result, the more proficient a candidate is in the respective topic, the 'later' they can begin within the presented curriculum.

2.1 Semantic Technology

The Semantic Technology curriculum has been created following the research done by STI International and STI Innsbruck for the Austrian national project SARID (Service Austria: Research and Industry Dissemination).[1]

1. Semantic Web Foundations
 - a. A Short History of Knowledge Systems
 - b. Birth of the Semantic Web
 - c. Semantic Applications
 - d. Core Concepts
 - e. RDF/S
 - f. OWL
 - g. Rules
 - h. SPARQL
2. Ontologies and the Semantic Web
 - a. Ontologies: a brief history
 - b. Ontology development process
 - c. Hands-on: use an example and create a requirements document
 - d. Knowledge elicitation

-
- e. Hands-on: formulate competency questions, carry out interviews, and make first draft of ontology
 - f. Ontology creation (tools)
 - g. Ontology design
 - h. Hands-on: create ontology in the chosen editor
3. Application Development
 - a. Semantic Web Application Framework
 - b. Development frameworks
 - c. Development methodologies
 - d. Creating the Semantic Web
 - e. Storing the Semantic Web
 - f. Creating Semantic Web clients
 - g. Querying the Semantic Web
 - h. An application example, e.g. a semantic web portal
4. KR and Reasoning on the Semantic Web
 - a. Core concepts in reasoning and logic
 - b. Description Logic-based Knowledge Representation
 - c. RDFS and Taxonomic reasoning
 - d. OWL semantics
 - e. Hands on session: seeing inferences using an ontology editor and a reasoner
 - f. Reasoners for the Semantic Web
 - g. Logic Programming
 - h. Semantic Web and Logic Programming
5. Ontology Lifecycle
 - a. Ontology lifecycle
 - b. Collaboratively developing an ontology
 - c. Finding ontologies
 - d. Ontology modularisation
 - e. Re-using ontologies
 - f. Ontology evaluation
 - g. Ontology refinement
 - h. Ontology evolution and versioning
6. Semantic Web Services
 - a. From Web Services to Semantic Web Services
 - b. Adding Semantics to existing services: SAWSDL
 - c. OWL-S
 - d. The WSM stack
 - e. Semantic Web Service application deployment
 - f. Hands-on: creating a WSMO service with WSMT
 - g. Hands-on: discovering and executing WSMO services with WSMX
-

- h. Hands-on: creating SWS with IRS-III and WSMO Studio
7. Semantic Web Services in Depth
- a. SWS matching
 - b. SWS mediation
 - c. SWS orchestration
 - d. SWS choreography
 - e. SWS co-ordination
 - f. Trust and agreement between services
 - g. Capturing business rules
 - h. Capturing business processes

2.2 Databases technology

This training curriculum for database technology contains topics that are often used in, e.g. master database courses, summer schools/seminars for Ph.D. students. The items "brief history" and "basic database techniques" are intended to contain general topics in database management systems. Everyone who wants to learn about how DBMSs work, should know that these are the topics that play important roles in DBMSs, what these topics are about and what basic techniques are used in the scope of these topics.

The contents and the structure of the list below have been adjusted for PlanetData, to take into account the topics relevant to PlanetData deliverables. The items "Performance benchmarking", "Column-store architectures", "XML query processing" and "Advanced database techniques" are directly related with the topics dealt within WP1 in DoW of PlanetData.

This training curriculum for database technology contains topics dealt with in well-known DBMS books and courses for bachelor and master students. The following book[9] has been consulted for some of the curriculum.

As this list is a preliminary list, we expect that its contents and structure will be refined and adjusted more in later versions to better meet the needs of audience of PlanetData.

1. A brief history of DBMS
 - a. Data models
 - b. Query languages
 - c. Kinds of DBMS: relational, object-oriented, semi-structured, XML, Multi-dimensional, ...
 - d. Distributed architectures: distributed, federated, multi-DBMSs, P2P DBMSs
 - e. Major commercial/open-source DBMSs
2. Performance Benchmarking
 - a. Components: hardware platform, data structures, algebraic optimizer, SQL parser
 - b. Measures: throughput, response time, availability; speedup, scaleup, sizeup.
 - c. Relational benchmarks: Wisconsin, AS3AP, TPC-*, ...
 - d. XML benchmarks: XMark, XPathMark, XBenck, ...
3. Basic database techniques²
 - a. Relational data model and query language
 - i. The relational data model
 - ii. Data storage

² <http://homepages.cwi.nl/~manegold/teaching/DBtech/>

- iii. Relational algebra
 - iv. The RDBMS language SQL
 - b. Query execution
 - i. The query compiler
 - ii. Query execution algorithms: join, sort, hash based, etc...
 - iii. Index structures
 - iv. Buffer management
 - v. Transaction management
 - c. Column-store architectures (MonetDB)
 - i. Database structures
 - ii. Execution paradigm
 - iii. Query optimiser
 - iv. DBMS architecture
 - v. Database cracking
 - vi. Database recycling
 - d. XML query processing
 - i. Introduction to XML, XPath, XQuery
 - ii. Relational XQuery processing (with MonetDB/XQuery)
 - iii. Handling XQuery Updates
 - iv. Other XQuery processing approaches
4. Advanced database techniques³
- a. Parallel and Distributed Database
 - i. Introduction
 - ii. Architectures
 - iii. Query processing techniques
 - iv. Data storage
 - v. Data replication
 - vi. Transaction management
 - b. Spatial and Geographic Databases
 - i. Sequoia benchmark for earth science (ES) databases
 - ii. Data model
 - iii. Spatial data access
 - iv. Spatial indexes

2.3 Linked Data

The Linked Data curriculum has been generated from a book following[2].

Introduction

- a. The Data Deluge

³ <http://homepages.cwi.nl/~manegold/teaching/adt/>

- b. The Rationale for Linked Data
 - c. Intended Audience
 - d. Introducing *Big Lynx Productions*
1. Principles of Linked Data
 - a. The Principles in a Nutshell
 - b. Naming Things with URIs
 - c. Making URIs Defererencable
 - d. Providing Useful RDF Information
 - e. Including Links to other Things
 2. The Web of Data
 - a. Bootstrapping the Web of Data
 - b. Topology of the Web of Data
 3. Linked Data Design Considerations
 - a. Using URIs as Names for Things
 - b. Describing Things with RDF
 - c. Publishing Data about Data
 - d. Choosing and Using Vocabularies
 - e. Making Links with RDF
 4. Recipes for Publishing Linked Data
 - a. Linked Data Publishing Patterns
 - b. The Recipes
 - c. Additional Approaches to Publishing Linked Data
 - d. Testing and Debugging Linked Data
 - e. Linked Data Publishing Checklist
 5. Consuming Linked Data
 - a. Deployed Linked Data Applications
 - b. Architecture of Linked Data Applications
 - c. Effort Distribution between Publishers, Consumers and Third

2.4 Data Streams

Data streams curriculum has taken ideas from several books,[3][4][5] papers[7], presentations[6] and courses[8] as a basis. It has been adapted for the PlanetData project needs in large-scale data processing. However, we have not generated any material yet and we have not exercised or used this curriculum in any other courses yet.

1. Motivation
 - a. Comparison with Relational DB storage
2. Streaming data models
 - a. Unbounded streams
 - b. Tuples, Windows
 - c. Timestamps

- d. K-constraints
- 3. Query Languages
 - a. Relational operators
 - b. Window operators, temporal operators
 - c. Aggregators
 - d. Joins
- 4. Semantic streaming data
 - a. RDF Stream data models
 - b. SPARQL extensions for RDF Streams
 - c. Reasoning with Streams
 - d. Complex event processing
 - e. Linked Streaming Data
- 5. Query processing
 - a. Continuous queries
 - b. Window evaluation
 - c. Aggregates evaluation, approximative queries
 - d. Static optimization
 - e. Query optimization, statistics
 - f. Load shedding
 - g. Sampling

3 Conclusion

This deliverable has introduced the different PlanetData curricula that can be applied for the development of training offers, both graduate and professional. This curriculum has been proposed so that the future trainers and learners will have a standard guidance to what topics need to be learnt as a benchmark for considering oneself adequately skilled as a large-scale data manager. Since this is just the beginning of the project, we cannot foresee yet how the project results will be by the end of the project. Therefore the curriculum in this deliverable could be modified and adjusted to the current situation later, e.g. to address training in the use of software and platforms provided by the PlanetData partners. The developed initial curriculum presented here focuses on covering the related underlying topics of large-scale data management which the PlanetData network partners have identified as core to enabling future large-scale data management solutions, and hence represent expected core skills of any large-scale data manager.

In the rest of the project, we will explore how the curricula here may form part of concrete training offers. In the context of graduate training, we will look at how this may be integrated into Masters' and PhD courses, e.g. within the Erasmus Mundi program, or PhD summer school. In the context of professional training, a specific certified training course in Large Scale Data Management could be established. STI International's professional training company Semsphere would be a potential driver for this. Online training in line with this curriculum can be supported using the REASE platform hosted by STI International or the Video Lectures platform expected to form the core of a PlanetData training infrastructure, see deliverable D6.2 for more details on this. Finally, as PlanetData research and platform development proceeds, we can update and expand this curriculum accordingly.

References

- [1] SARIT D1.1 Training Material Plan, Emilia Cimpian and Elena Simperl, December 1, 2008
- [2] Tom Heath and Christian Bizer (2011) *Linked Data: Evolving the Web into a Global Data Space* (1st edition). *Synthesis Lectures on the Semantic Web: Theory and Technology*, 1:1, 1-136. Morgan & Claypool.
- [3] *Data streams: models and algorithms*, Aggarwal, C.C. 2007, Springer-Verlag New York Inc.
- [4] *Data Stream Management: Processing High-Speed Data Streams* M. Garofalakis, J. Gehrke, and R. Rastogi, Springer, 2007.
- [5] *Stream Data Management - Advances in Database Systems* Nauman Chaudhry, Kevin Shaw, Mahdi Abdelguerfi, Springer 2010.
- [6] ICDE Tutorial slides material: Data stream query processing by Nick Koudas, University of Toronto and Divesh Srivastava, AT&T Labs-Research. Tutorial presented at IEEE International Conference on Data Engineering (ICDE), 1145, 2005
- [7] Brian Babcock, Shivnath Babu, Mayur Datar, Rajeev Motwani, and Jennifer Widom. 2002. Models and issues in data stream systems. In *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems (PODS '02)*. ACM, New York, NY, USA.
- [8] *Streaming for Dummies*. Zdonik, S. and Sibley, P. and Rasin, A. and Sweetser, V. and Montgomery, P. and Turner, J. and Wicks, J. and Zgolinski, A. and Snyder, D. and Humphrey, 2004, Citeseer.
- [9] Widom, Jennifer, H. Garcia-Molina and J. D. Ullman (2002) *Database Systems: The Complete Book*, New Jersey: Prentice-Hall. ISBN 978-0130319951.