CUSTOMER

Hadoop Connector Configuration Guide

Typographic Conventions

Type Style	Description
Example	Words or characters quoted from the screen. These include field names, screen titles, pushbuttons labels, menu names, menu paths, and menu options. Textual cross-references to other documents.
Example	Emphasized words or expressions.
`EXAMPLE`	Technical names of system objects. These include report names, program names, transaction codes, table names, and key concepts of a programming language when they are surrounded by body text, for example, SELECT and INCLUDE.
`Example`	Output on the screen. This includes file and directory names and their paths, messages, names of variables and parameters, source text, and names of installation, upgrade and database tools.
`Example`	Exact user entry. These are words or characters that you enter in the system exactly as they appear in the documentation.
`<Example>`	Variable user entry. Angle brackets indicate that you replace these words and characters with appropriate entries to make entries in the system.
`EXAMPLE`	Keys on the keyboard, for example, `F2` or `ENTER`.

Document History

Version	Date	Release	Change
1.0	<2015-11-11>	NW 740 SP13 and NW 750 SP01	Initial Delivery
1.1	<2018-04-05>	NW 740 SP13 and NW 750 SP01 and later releases	Edited creating Hadoop connection

Table of Contents

Hadoop Connector Configuration Guide

Table of Contents

1 Introduction

1.1 About this Document

Use

You use this document to perform the Hadoop Connector configuration.

1.2 Overview Hadoop

Apache™ Hadoop® is an open source software project that enables distributed processing of large data sets across clusters of commodity servers. It is designed to scale up from a single server to thousands of machines, with a very high degree of fault tolerance. Rather than relying on high-end hardware, the resiliency of these clusters comes from the software's ability to detect and handle failures at the application layer.

Hadoop File System (HDFS) indicates the main difference between Hadoop and a database. While databases store data in the form of transparent tables, Hadoop operates with flat files in a directory structure. The HDFS stores files under specific directories. Therefore the HDFS has to know the name of the file and the directory path where it is located. Here's a comparison of the main Uniform Resource Identifier (URI)/path components of the ILM Store and the Hadoop File System:

ILM Store	Hadoop File System
Collection	Directory
Resource	File

Resources within the ILM Store are stored in specific collections that create resource URIs. These resource URIs are also used in the Hadoop File System. This facilitates:

- easier orientation of files in a Hadoop File System for an administrator
- flexible selection of resources when moving or cleaning files

Before you continue, you should decide on how you want to structure your storage.

2 File System Layout

The layout of the file system can be custom-modified by implementing the BAdI (BADI_ILM_STOR_PROCESS) in enhancement spot ENHS_ILM_STOR_PROCESS. The file path (path & file name) is queried from the method IF_ILM_STOR_ES_PROCESS~GET_STORE_TARGET_INFO. By implementing this BAdI, you are completely free to define a path of your choice: You may, however, make use of the properties available in the ILM Store as shown below.

2.1 Default Path Assembly

By default, the path to the file is assembled according to the following scheme:
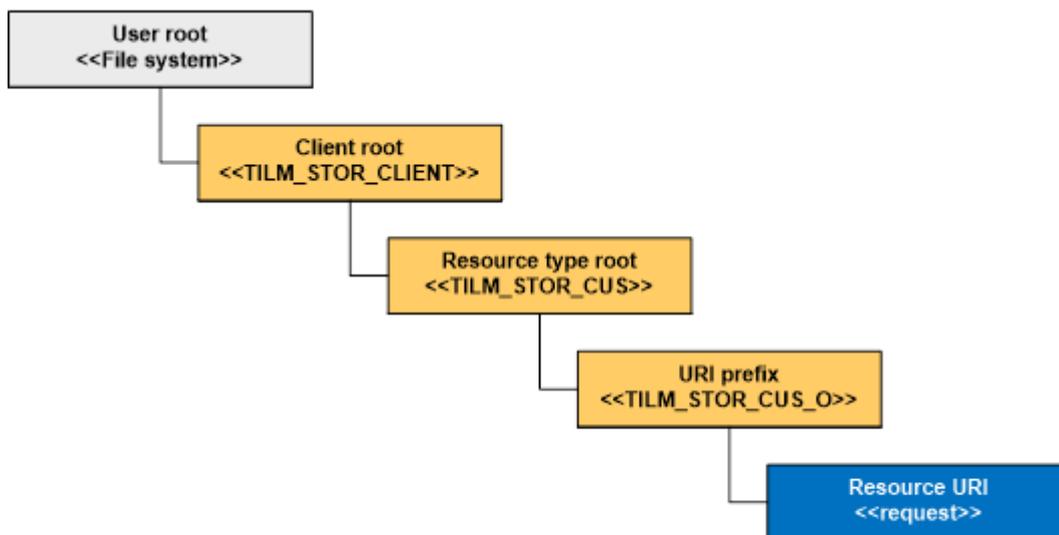

2.1.1 <User Root>

[Mandatory]

This is used to store the home directory path of the target Hadoop user. This path must be specified within an RFC destination used to connect to the Hadoop File System. For more information see, RFC Destination Maintenance.

2.1.2   <Client Root>

[Optional]

The client root path is defined by a logical path name in the client's Customizing. For more information, refer to the topic Create Clients and Origins (Transaction ILM_STOR_ADM_CUST) in the Installation and Configuration Guide for the ILM Store. This is the common root of all file system-based storage locations of a client.
:
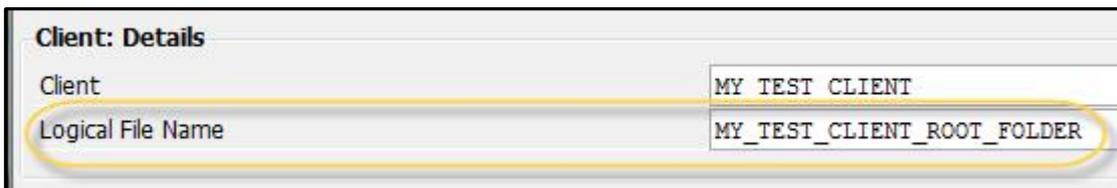
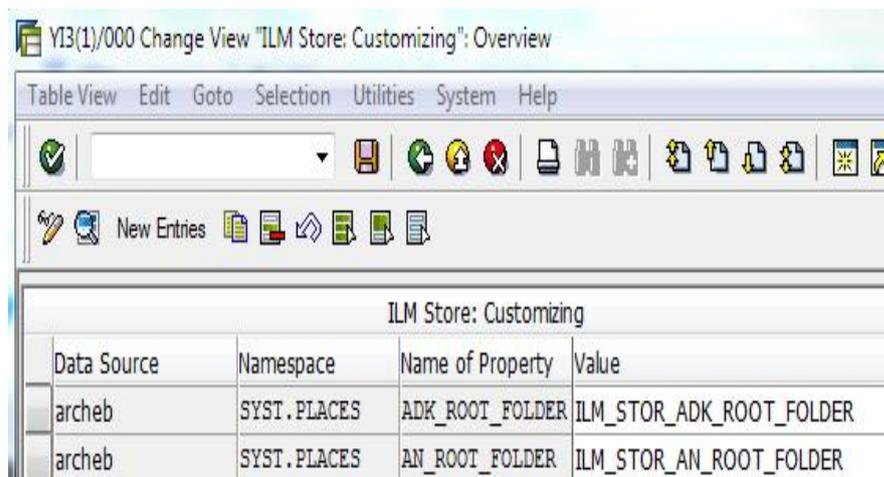| Client: Details	
Client	MY TEST CLIENT
Logical File Name	MY_TEST_CLIENT_ROOT_FOLDER

## 2.1.3   <Resource type root>

[Optional]

The resource type root path is defined by a logical path specified in ILM Store's Operational Customizing under namespace (SYST.PLACES) and properties (AN_ROOT_FOLDER & ADK_ROOT_FOLDER). It differentiates the resources by their type: ADK files and Anonymous files (Archive Link).

YI3(1)/000 Change View "ILM Store: Customizing": Overview

Table View Edit Goto Selection Utilities System Help

New Entries

ILM Store: Customizing

Data Source	Namespace	Name of Property	Value
archeb	SYST.PLACES	ADK_ROOT_FOLDER	ILM_STOR_ADK_ROOT_FOLDER
archeb	SYST.PLACES	AN_ROOT_FOLDER	ILM_STOR_AN_ROOT_FOLDER

The physical paths are defined in transaction FILE:
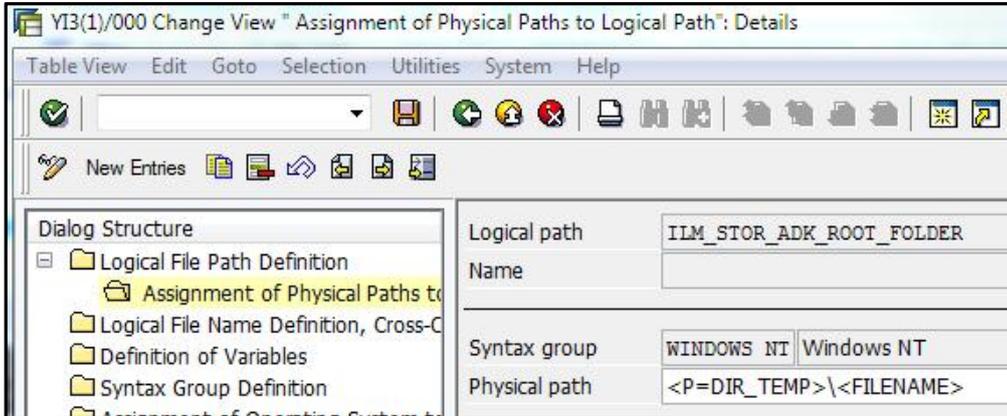
2.1.4   <URI prefix

[Optional]

The URI prefix path is defined by a path prefix belonging to the current origin under which the store operation is performed. It is customizable in the Operative origin setup of the client's Customizing. . For more information, refer to the topic Create Clients and Origins in the Installation and Configuration Guide for the ILM Store.
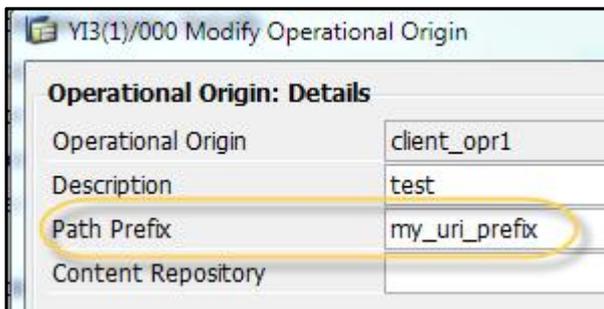
2.1.5   <Resource URI>

This is the resource URI as specified in the request to store a file and must not be changed.

2.2   Topology of ILM Store

Topology enables you distinguish between the system landscapes in the ILM Store. For example, if you have the test landscape TST and the productive landscape PRD and you want to use one single ILM Store server to handle both of them. To get a real time experience, you can use a copy of the productive data in the test environment. You have to be careful not to mix the data. The Store's topology is built on the above mentioned client: A client groups origins that belong together semantically into one virtual group. The most obvious distinction is the root path for file systems. Using our example, the topology could look like this:
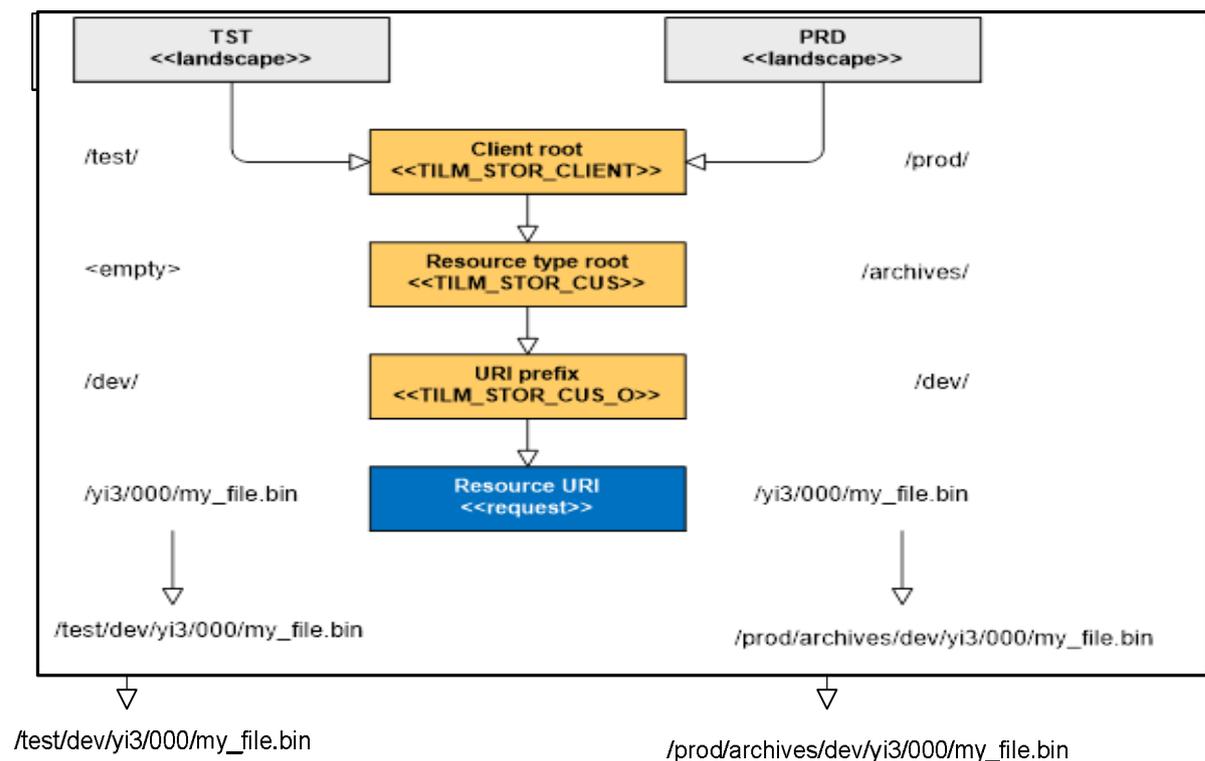
Finally, this path fragment is created under the home directory of the user who is connected to Hadoop. You can use a separate user, but even if you use the standard HDFS user, your data will stay separated.

3 Prerequisites

- Hadoop integration is available for the ILM Store as of:
  - o SAP NetWeaver 7.40 SPS 13
  - o SAP NetWeaver 7.50 SPS 01
- Storage and Retention Service should be enabled.
- The ILM Store has been configured and tested to run with a default configuration.

Further information available online on the Information Lifecycle Management page.

Note: This documentation does not cover installation support for a Hadoop cluster. It is expected that a live Hadoop cluster is available and reachable from the system landscape. Install and enable HttpFS or WebHDFS service on your Hadoop Cluster.

HttpFS and WebHDFS are HTTP services covered by the Apache™ Hadoop® project. They provide HTTP REST API that supports all HDFS File System operations. The ILM Store's Hadoop Connector supports connection to both of these services. For more details about the services, refer to official documentation of the Apache™ Hadoop® project.

4 Hadoop Connector Configuration

This process consists of four main steps:

1. Hadoop connection enablement in ILM Store
   - Create connection to Hadoop
   - Registration of storage medium operators
2. Maintenance of Hadoop preferences in ILM Store
   - Customizing of Hadoop connection
   - Customizing of further access
3. RFC destination maintenance.
4. Customizing of Hadoop origin-dependent connection

4.1 Hadoop Connection Enablement in ILM Store

The connection parameters are maintained in the ILM Store to establish the communication with the Hadoop instance. We use the terms "connection type" and "connection ID" as follows:

- The connection type describes a storage medium
- The connection ID points to a set of properties

4.1.1 Create Connection

You need to access the maintenance view of Customizing table TILM_STOR_CUS. All changes to Customizing depend on the ILM Store origin value and are specified accordingly. We will use the sample origin archeb.

Perform the following steps:

1. Create a new Connection ID and have it assigned to DBCON.TILM_STOR_BLOB.

Namespace: DB   Property: DBCON.TILM_STOR_BLOB   Value: HADOOP_CONNECTION_ID
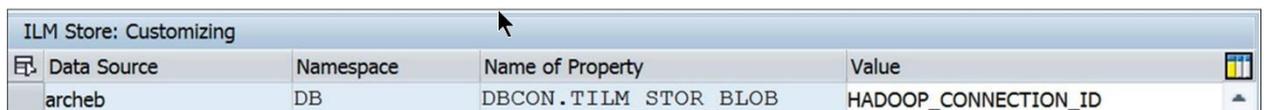
ILM Store: Customizing				
Data Source	Namespace	Name of Property	Value	
archeb	DB	DBCON.TILM_STOR_BLOB	HADOOP_CONNECTION_ID	

2. Update this connection ID to point to the connection type.

Namespace: DB   Property: CONSYS.HADOOP_CONNECTION_ID   Value: HADOOP
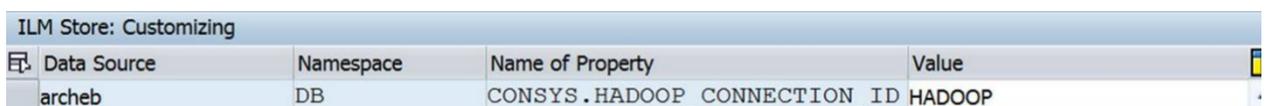
ILM Store: Customizing				
Data Source	Namespace	Name of Property	Value	
archeb	DB	CONSYS.HADOOP_CONNECTION_ID	HADOOP	

4.1.2 Registration of Storage Medium Operators

The Storage Medium Operators are responsible for handling all of the communication with a storage medium. They are registered in the class factory.

You need to access the maintenance view of the class factory registery table TILMSTOR_CF.

Maintain the configuration for routing of the interface to the storage medium operators. (Ensure your settings are in sync with the naming convention used in TILM_STOR_CUS. The value of the Parameter field must correspond to value of the property DBCON.TILM_STOR_BLOB, in other words it must be the connection type).

Mandatory values:

Interface: IF_ILM_STOR_DB_OPERATOR
API Version: Default
Implementing Class: CL_ILM_STOR_DB_OPERATOR_HDP
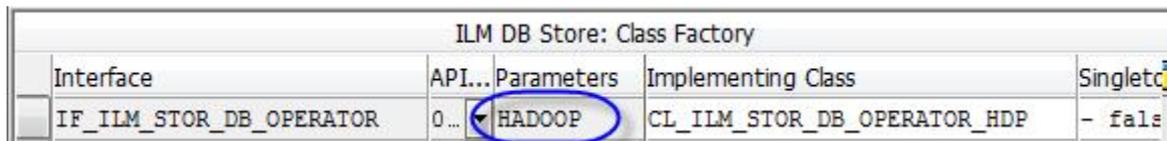Singleton: False

For example:

Interface: IF_ILM_STOR_DB_OPERATOR
API Version: Default
Parameter: HADOOP (customer-defined)
Implementing Class: CL_ILM_STOR_DB_OPERATOR_HDP
Singleton: False

4.2 Maintenance of Hadoop Preferences in ILM Store

The connection parameters, authentication, and access methods should be maintained to ensure successful operation and data transfer between ILM Store and Hadoop.

4.2.1 Hadoop Connection Customizing

You need to access the maintenance view of the Hadoop origin independent, connection customizing table TILM_STOR_HDP_C.

The connection ID name should be the same as the name maintained in TILM_STOR_CUS.

The Customizing table contains the following properties:

·   Max Data Size - Maximum amount of data, in Bytes, transferrable by an HTTP request. The value depends on HTTP service setup in Hadoop.

·   Authentication Method – Authentication method used for user access to Hadoop. The value depends on the setup of the HTTP service in Hadoop. Pseudo and no authentication are currently the only available options as of Hadoop release NW740 SP13.

Example:

Connection ID: HADOOP_CONNECTION_ID
Max Data Size [B]: 2.147.483.648
Authentication Method: PSEUDO_AUTHENTICATION

ILM Store: Connection Customizing for Hadoop Connector		
Connection ID	Max Data Size [B]	Authentication Method
HADOOP_CONNECTION_ID	2.147.483.648	PSEUDO_AUTHENTICATION

4.2.2   Access Customizing

The no authentication method requires no further maintenance, pseudo authentication requires additional configuration steps to complete. You need to access the maintenance view of the Customizing table TILM_STOR_HDP_PU. The table defines the ORIGIN (data source)-to-Hadoop user mapping that is supposed to perform the connection and data transfer. The Hadoop user must exist in the Hadoop system.

Example:

Connection ID: HADOOP_CONNECTION_ID
Data Source (origin): archeb
User Name: guest

ILM Store: Origin to User Mapping for Pseudo Authentication		
Connection ID	Data Source	User Name
HADOOP_CONNECTION_ID	archeb	guest

4.3   RFC Destination Maintenance

You need to access RFC destination maintenance transaction SM59 and create a connection to the HADOOP instance.

The RFC destination should contain the target host where the Hadoop HTTP Service is initialized and running, as well as the service (port) number.

The path prefix should refer to the home directory, which refers to the user to be invoked for the connection. It has the following syntax:

/webhdfs/v1/<hdp_usr_home_path>

·   /webhdfs/v1 – This is the mandatory path prefix defined by the HttpFS and WebHDFS APIs.

·   <hdp_usr_home_path> -This is the home directory path of the target Hadoop user. Each Hadoop user has their own home directory, which, in UNIX-type systems, for example, is typically '/user/<user_name>'. The user is the owner of the directory and has higher access privileges.

4.4 Hadoop Origin-Dependent Connection Customizing

You need to access the maintenance view of the origin dependent-connection Customizing table TILM_STOR_HDP_CO. The connection ID name should be same as the database connection name maintained in TILM_STOR_CUS. You need to set up or maintain the relevant RFC destination, you have configured for the communication with the HADOOP cluster.

5   Security

There are three main security tasks:

- Ensure that data is only accessible to its owners
- Ensure confidentiality of data during HTTP communication
- Ensure integrity of the data

5.1   Authentication

The Hadoop connector offers two authentication methods:

- NO_AUTHENTICATION
- PSEUDO_AUTHENTICATION

Since Hadoop is an open source software, it can implement other authentication protocols. To enable this kind of freedom on the SAP side as well, customers can implement their own authentication methods. These methods must be registered in the table TILM_STOR_HDP_A and a corresponding implementation of the authentication BAdI must be created in the enhancement spot ILM_STOR_HDP_CONNECTIVITY. The ILM Store's Hadoop Connector also supports the standard HTTP Basic authentication. User credentials are maintained within the corresponding RFC destination.

5.2   Apache KNOX

The Apache Knox Gateway is a REST API Gateway for interacting with Hadoop clusters. The Knox Gateway provides a single access point for all REST interactions with Hadoop clusters.

It supports HTTP Basic authentication. HTTP Basic authentication can be maintained within an RFC destination. The authentication method NO_AUTHENTICATION is used for this purpose.

5.3   HttpFS over HTTPS

HTTPS can be enabled on the level of RFC destination. Transaction SM59 enables you to set up whether an SSL certificate shall be used. The prerequisite is a correct setup of security certificate in transaction STRUST.

6　References

References	Link
ILM Store Help Portal	Link
ILM Store Guide	Link

7 Appendix

7.1 Terms for Included Open Source Software

This SAP software contains also the third party open source software products listed below. Please note that for these third party products the following special terms and conditions shall apply.

1. This software was developed using ANTLR.

2. gSOAP

Part of the software embedded in this product is gSOAP software. Portions created by gSOAP are Copyright (C) 2001-2004 Robert A. van Engelen, Genivia inc. All Rights Reserved.

THE SOFTWARE IN THIS PRODUCT WAS IN PART PROVIDED BY GENIVIA INC AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE AUTHOR BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

3. SAP License Agreement for STLport SAP License Agreement for STLPort between SAP Aktiengesellschaft Systems, Applications, Products in Data Processing Neurottstrasse 16 69190 Walldorf, Germany (hereinafter: SAP) and you (hereinafter: Customer)

a) Subject Matter of the Agreement

A) SAP grants Customer a non-exclusive, non-transferrable, royalty-free license to use the STLport.org C++ library (STLport) and its documentation without fee.

B) By downloading, using, or copying STLport or any portion thereof Customer agrees to abide by the intellectual property laws, and to all of the terms and conditions of this Agreement.

C) The Customer may distribute binaries compiled with STLport (whether original or modified) without any royalties or restrictions.

D) Customer shall maintain the following copyright and permissions notices on STLport sources and its documentation unchanged:

Copyright 2001 SAP AG

E) The Customer may distribute original or modified STLport sources, provided that:

o The conditions indicated in the above permissions notice are met;

o The following copyright notices are retained when present, and conditions provided in accompanying permission notices are met:

Coypright 1994 Hewlett-Packard

Company

Copyright 1996,97 Silicon Graphics

Computer Systems Inc.

Copyright 1997 Moscow Center for

SPARC Technology.

Copyright 1999,2000 Boris Fomitchev

Copyright 2001 SAP AG

Permission to use, copy, modify, distribute and sell this software and its documentation for any purposes is hereby granted without fee, provided that the above copyright notice appear in all copies and that both that copyright notice and this permission notice appear in supporting documentation. Hewlett-Packard Company makes no representations about the suitability of this software for any purpose.

It is provided "as is" without express or implied warranty.

Permission to use, copy, modify, distribute and sell this software and its documentation for any purpose is hereby granted without fee, provided that the above copyright notice appear in all copies and that both that copyright notice and this permission notice appear in supporting documentation. Silicon Graphics makes no representations about the suitability of this software for any purpose. It is provided "as is" without express or implied warranty.

Permission to use, copy, modify, distribute and sell this software and its documentation for any purposes is hereby granted without fee, provided that the above copyright notice appear in all copies and that both that copyright notice and this permission notice appear in supporting documentation. Moscow Center for SPARC makes no representations about the suitability of this software for any purpose. It is provided "as is" without express or implied warranty.

Boris Fomitchev makes no representations about the suitability of this software for any purpose. This material is provided "as is", with absolutely no warranty expressed or implied.

Any use is at your own risk. Permission to use or copy this software for any purpose is hereby granted without fee, provided the above notices are retained on all copies.

Permission to modify the code and to distribute modified code is granted, provided the above notices are retained, and a notice that the code was modified is included with the above copyright notice.

Permission to use, copy, modify, distribute and sell this software and its documentation for any purposes is hereby granted without fee, provided that the above copyright notice appear in all copies and that both that copyright notice and this permission notice appear in supporting documentation. SAP makes no representations about the suitability of this software for any purpose. It is provided with a limited warranty and liability as set forth in the License Agreement distributed with this copy.

SAP offers this liability and warranty obligations only towards its customers and only referring to its modifications.

b) Support and Maintenance SAP does not provide software maintenance for the STLport. Software maintenance of the STLport therefore shall be not included.

All other services shall be charged according to the rates for services quoted in the SAP List of Prices and Conditions and shall be subject to a separate contract.

c) Exclusion of warranty

As the STLport is transferred to the Customer on a loan basis and free of charge, SAP cannot guarantee that the STLport is error-free, without material defects or suitable for a specific application under third-party rights. Technical data, sales brochures, advertising text and quality descriptions produced by SAP do not indicate any assurance of particular attributes.

d) Limited Liability

A) Irrespective of the legal reasons, SAP shall only be liable for damage, including unauthorized operation, if this (i) can be compensated under the Product Liability Act or (ii) if caused due to gross negligence or intent by SAP or (iii) if based on the failure of a guaranteed attribute.

B) If SAP is liable for gross negligence or intent caused by employees who are neither agents or managerial employees of SAP, the total liability for such damage and a maximum limit on the scope of any such damage shall

depend on the extent to which its occurrence ought to have anticipated by SAP when concluding the contract, due to the circumstances known to it at that point in time representing a typical

transfer of the software.

C) In the case of Art. 4.2 above, SAP shall not be liable for indirect damage, consequential damage caused by a defect or lost profit.

D) SAP and the Customer agree that the typical foreseeable extent of damage shall under no circumstances exceed EUR 5,000.

E) The Customer shall take adequate measures for the protection of data and programs, in particular by making backup copies at the minimum intervals recommended by SAP. SAP shall not be liable for the loss of data and its recovery, notwithstanding the other limitations of the present Art. 4 if this loss could have been avoided by observing this obligation.

F) The exclusion or the limitation of claims in accordance with the present Art. 4 includes claims against employees or agents of SAP.

4. Adobe Document Services Adobe, the Adobe logo, Acrobat, PostScript, and Reader are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States and / or other countries. For information on Third Party software delivered with Adobe document services and Adobe LiveCycle Designer, see SAP Note 854621.

www.sap.com/contactsap