Report on Data Infrastructure Set-up

Contents

1 Preface

This document is the Report on Data Infrastructure Set Up by the IDMAPS project at Newcastle University.[1] It has been made available under a *Creative Commons Attribution-Share Alike 3.0 License* to the wider Higher Education community in the expectation that our experiences will prove useful to other institutions undertaking similar activities.[2]

Any references to third-party companies, products or services in this document are purely for informational purposes, and do not constitute any kind of endorsement by the IDMAPS Project or Newcastle University.

2 Introduction

2.1 Report

This report provides details of how the data infrastructure which has been set up by the IDMAPS project addresses the needs identified in the Requirements Analysis document.

Each requirement is addressed individually, and is referenced back to the corresponding need in the Requirements Analysis. For some Optional or Desirable requirements, the project is not currently using functionality provided by the infrastructure. Where this is the case, this is noted as "Unused".

The use of Talend as a data Extraction, Transform and Load (ETL) tool underpins a number of the technical aspects of the data architecture.

3 Meeting the Requirements Analysis

3.1 General Principles: Simplicity

Priority	***R.A. Ref.***	***Description***	***Met?***
Essential	4.1	See details below	☑

A key aim was simplicity in the institutional data infrastructure. This requirement was met in several ways.

Talend provides a graphical IDE for the development of data feeds, which significantly simplified the creation of data flows. This provides an easy to understand visual display of the intricacies of complex data flows, and enables their swift creation. It also makes it incredibly easy to re-purpose existing feeds to accommodate changes in the required inputs and outputs for the feed, such as a change of format.

A crucial part of the project is to ensure that future data feeds are adequately documented and that appropriate governance is put in place for data provision. The procedural requirements of the data infrastructure have therefore been simplified as far as possible, by:

[1] *Institutional Data Management for Personalisation and Syndication* (IDMAPS) is a JISC-funded Institutional Innovation project which aims to improve the quality and reliability of institutional data flows. For more information, please visit the project website at http://research.ncl.ac.uk/idmaps.
[2] http://creativecommons.org/licenses/by-sa/3.0/.

1. The Institutional Data Feed Service (IDFS) has a clear remit, supported by the senior management of the department.
2. There is a simple process for requesting a data feed, which involves contacting the IDFS team directly by e-mail to discuss requirements.
3. The Data Integration Template, which must accompany each data feed request, has been kept as simple as possible. It uses sections to clearly stage the process, and is accompanied by a visual guide. The IDFS will assist Application Developers in completing the Template.

3.2 Technical Requirements

Priority	*R.A. Ref.*	*Description*	*Met?*
3.2.1 Field validation			
Essential	5.1.1	Field validation against schemas and data types is supported by Talend.	☑
3.2.2 Input accuracy screening			
Optional	5.1.2	Input accuracy screening is supportable through Talend data quality plug-ins.	☑ Unused
3.2.3 Feedback mechanism to master systems			
Essential	5.1.3	Talend provides the ability to log errors and use this information to change the data flow process, allowing feedback to to master systems.	☑
3.2.4 Abnormal status reporting			
Essential	5.2.1	Talend supports both e-mail alerts and a data dashboard, allowing abnormal status reporting.	☑
3.2.5 Error notification system			
Essential	5.2.2	Talend supports both e-mail alerts and a data dashboard, providing an error notification system.	☑
3.2.6 Data analysis tools			
Desirable	5.3.1	Data Analysis is supported in Talend Open Profiler.	☑ Unused
3.2.7 Data load reporting			
Essential	5.3.2	Talend provides statistics on data loads through its Administration tool.	☑
3.2.8 Statistical reporting of field quality			
Optional	5.3.3	Although this is not explicitly supported by Talend, as part of the data flow creation process, it is possible to ensure that data flows provide this information.	☑
3.2.9 Final destination systems			
Essential	5.3.4	This is inherently supported: Talend controls the data flow process, from retrieving the original data to pushing it to the recipient systems, forming an unbroken “chain” independent of any other systems.	☑

3.2.10 Daily dashboard

Optional	5.3.5	Supported, though at present unused.	☑ Unused

3.2.11 Auditability and interrogability of end systems

Optional	5.3.6	Supported, though at present unused. There are security considerations inherent in allowing one system seamless access to many others in this manner.	☑ Unused

3.2.12 Grace periods

Optional	5.4.1	Not explicitly supported by Talend itself, but flows can be created to ensure this functionality.	☑

3.2.13 Mixed update frequency environment

Essential	5.4.2	Talend supports a mixed update frequency environment through a number of technologies, including scheduling jobs itself, SOA, and passing jobs to a cron scheduler.	☑

3.2.14 Future-proofing

Desirable	5.4.3	All three technologies (MOM, SOA and Web services) are supported by Talend.	☑

3.2.15 Data update types

Desirable	5.4.4	This is supported by the creation of specific data flows.	☑

3.2.16 Redundancy and resilience

Essential	5.4.5	As Talend pushes data to end systems, if the latter do not receive it, they will continue to operate on pre-existing data. In addition, the jobs created by Talend are standard Java/Perl code, which can be taken and run from any system.	☑

3.2.17 Data auditing

Essential	6.1	Auditing is supported through the procedures put in place as part of the project, including the use of forms to capture information regarding data flows, and the support of the Senior Management Team in the department and wider University.	☑

3.2.18 Legal obligations

Essential	6.2	Legal obligations, as data auditing, are supported through the use of appropriate forms and policies.	☑

4 Installing Talend

4.1 Hardware

Talend does not require much in the way of hardware. Our installation is running on a virtual machine with Centos 5.3 OS, 4GB of RAM and 20GB of disc space. This seems more than adequate for the time being however more disc space may be required in the future.

4.2 The Installation, Setup and Testing Process

Installing Talend is as simple as unzipping a file into the appropriate directory. It was very easy and it just worked without problems.

The documentation[3] is relatively good, for open source software, and certainly adequate. However, only the basics are documented, meaning that there is a steep learning curve. Talend do offer (paid for) courses which go into a lot more depth and detail. In addition to this, there is an active forum[4] with a community eager to help each other out and answer questions.

There is also good support from the developers. During testing, a small number of bugs were discovered. These were reported and have been resolved, as well as three enhancement requests.

Prototype testing against requirements was used to check that Talend was fit for purpose; using data feed inputs with known results.

[3] www.talend.com/resources/documentation.php
[4] www.talendforge.org/forum