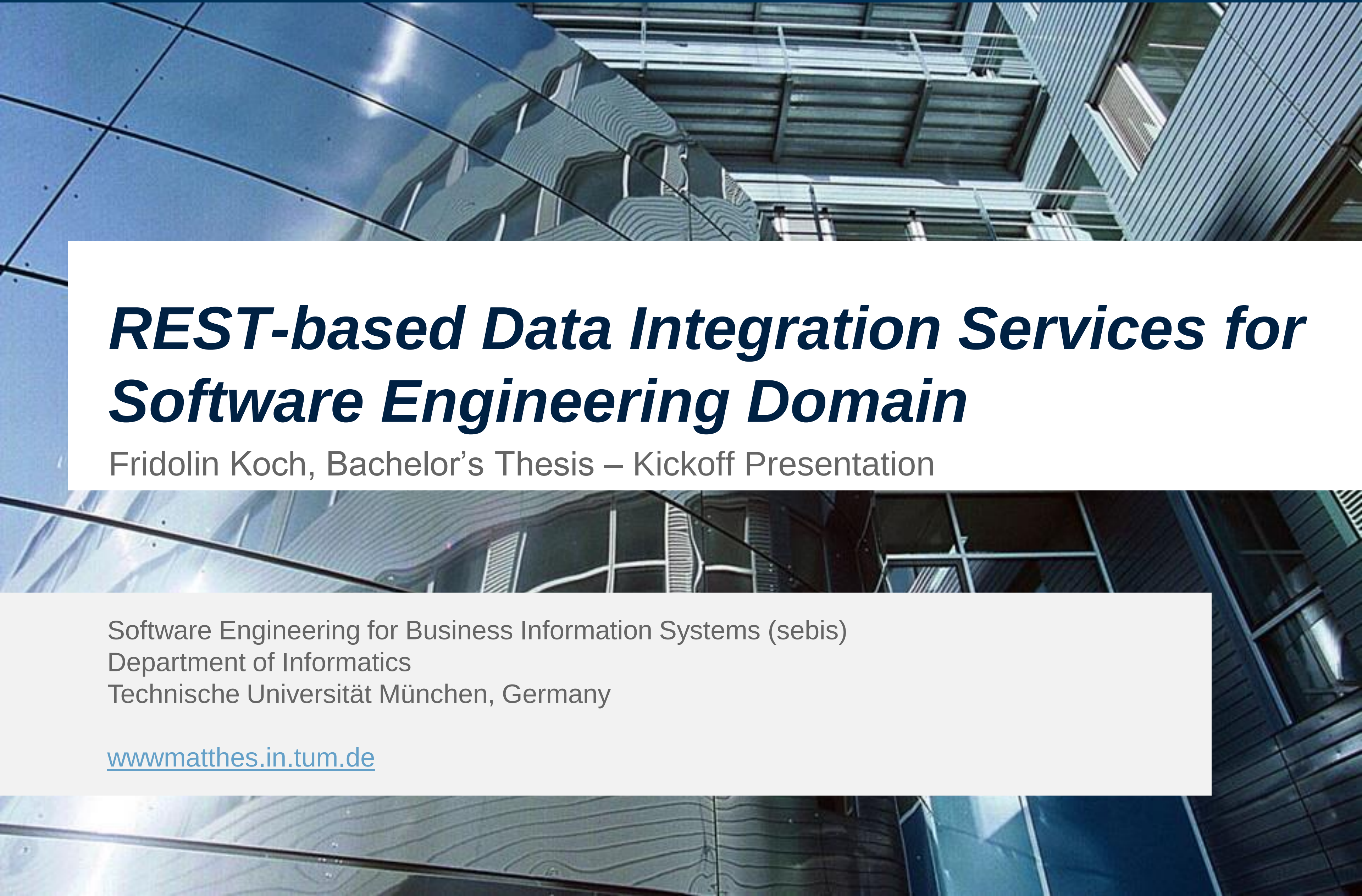
Technische Universität München

sebis



REST-based Data Integration Services for Software Engineering Domain

Fridolin Koch, Bachelor’s Thesis – Kickoff Presentation

Software Engineering for Business Information Systems (sebis)
Department of Informatics
Technische Universität München, Germany

wwwmatthes.in.tum.de

Outline

1. **Motivation**
   - Problem statement
   - Existing ETL solutions
2. **Research Questions**
3. **Solution Approach**
   - UI Prototype
   - Framework Workflow
   - Technology Stack
   - Current Architecture
4. **Next Steps**
5. **Timeline**

Problem Statement

- Existing barrier in the adoption of knowledge management systems in software engineering domain
  - Many different software architecture life cycle tools produce data in different formats (Enterprise Architect, Excel, Jira, etc.)
  - Repeatedly integrating this data into such a system can be a challenging and tedious task
- In general the task of data integration is addressed by **E**xtract-**T**ransform-**L**oad-Tool (ETL-Tool)
  - Wide range of commercial and open source ETL-Tool available
  - But: Mostly tailored to generic use cases → Difficult to embedded in existing domain specific tools
- **Potential Solution**: Analyze popular ETL-Tools and create an easily extendable framework

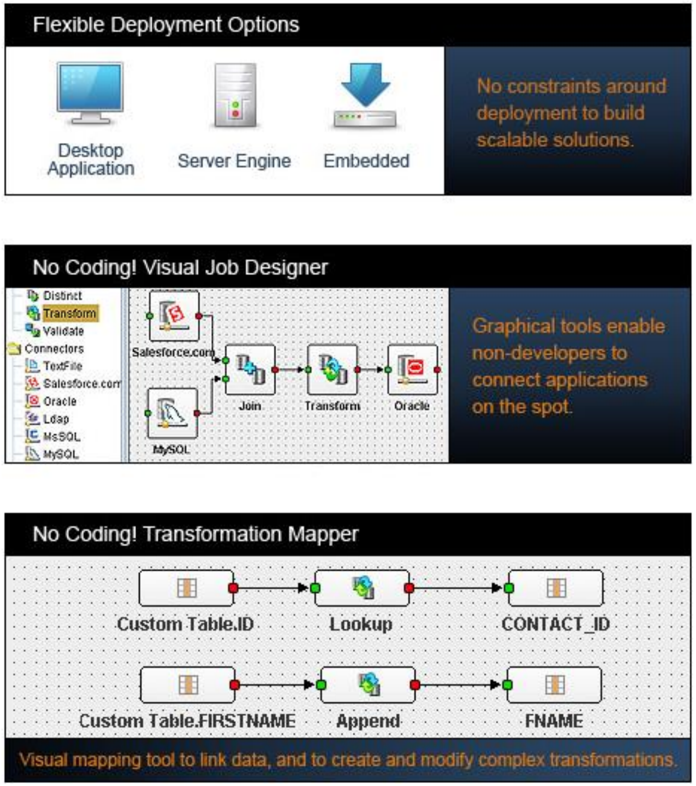
Existing ETL-Tools: Overview

Tool	OpenSource	Technology	Mode	Domain
Apatar	Yes	Java	Standalone	Generic
CloverETL	Core only	Java	Standalone, Embedded	Generic
Talend Open Studio for Data Integration	Yes	Java	Designer / Script-Generator	Generic
Pentaho	Yes, but less functionality	Java	Standalone	Generic
RhinoETL	Yes	C# .net	Framework	Generic
UnifiedViews	Yes	Java	Standalone	Linked-Data (RDF)

Existing ETL-Tools: Apatar

apatar
connecting data

- Java-Based
- Open-Source
- Visual job designer
- Generic usage domain



Flexible Deployment Options

No constraints around deployment to build scalable solutions.

No Coding! Visual Job Designer

Graphical tools enable non-developers to connect applications on the spot.

No Coding! Transformation Mapper

Visual mapping tool to link data, and to create and modify complex transformations.

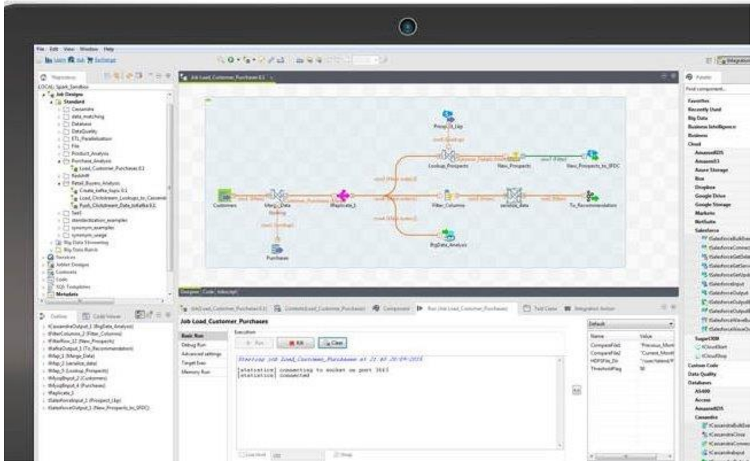
Source: www.apatar.com

Existing ETL-Tools: Clover-ETL

- Java-Based ETL-Tool
- Open-Source (Core only)
- Visual job designer (Community ad Commercial Edition)
- Standalone and embedded
- Generic usage domain
- Custom Domain-Specific-Language to define business logic (“CTL”)
- Clusterable
- Many data connectors

Source: http://www.cloveretl.com/

Existing ETL-Tools:
Talend Open Studio for Data Integration



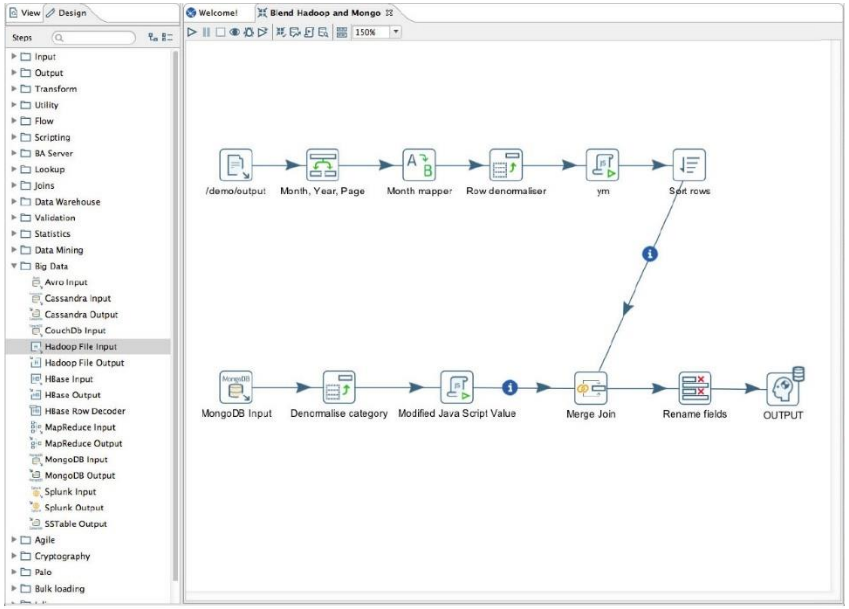
```
[statistics] connecting to socket on port [illegible]
[statistics] connected
```

- Java-Based ETL-Tool
- Open-Source
- Code generator for data transformation scripts (Java)
- Based on Eclipse
- +900 connector and components
- Generic usage domain

Source: https://www.talend.com/download/talend-open-studio

Existing ETL-Tools: Pentaho

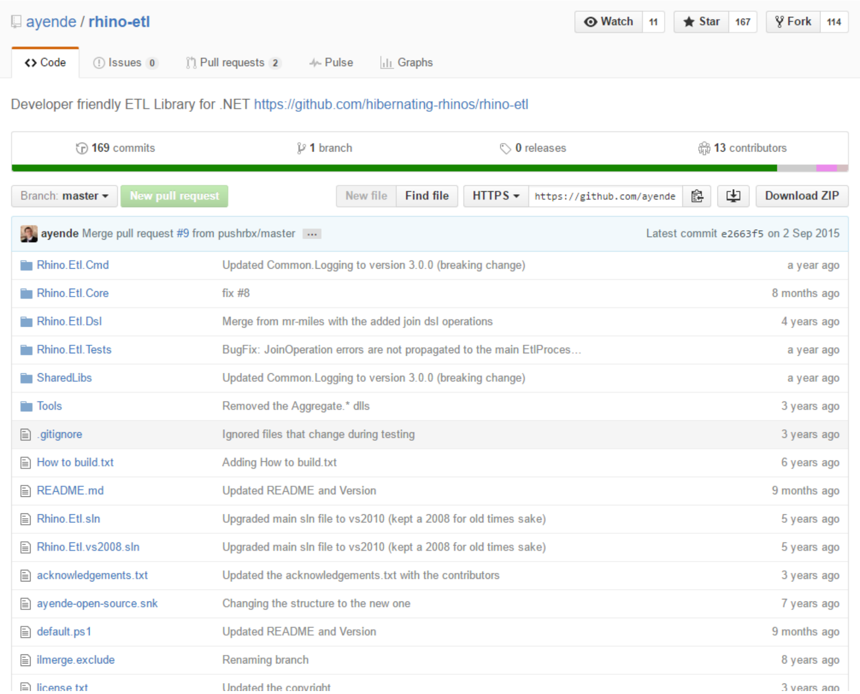
- Java-Based
- Community and Enterprise edition
- Visual Designer
- Rich library of pre-built ETL components
- Generic usage domain



Source: http://www.pentaho.com/product/data-integration

Existing ETL-Tools: Rhino ETL

hibernating rhinos



Developer friendly ETL Library for .NET https://github.com/hibernating-rhinos/rhino-etl

ayende Merge pull request #9 from pushrbx/master … Latest commit e2663f5 on 2 Sep 2015

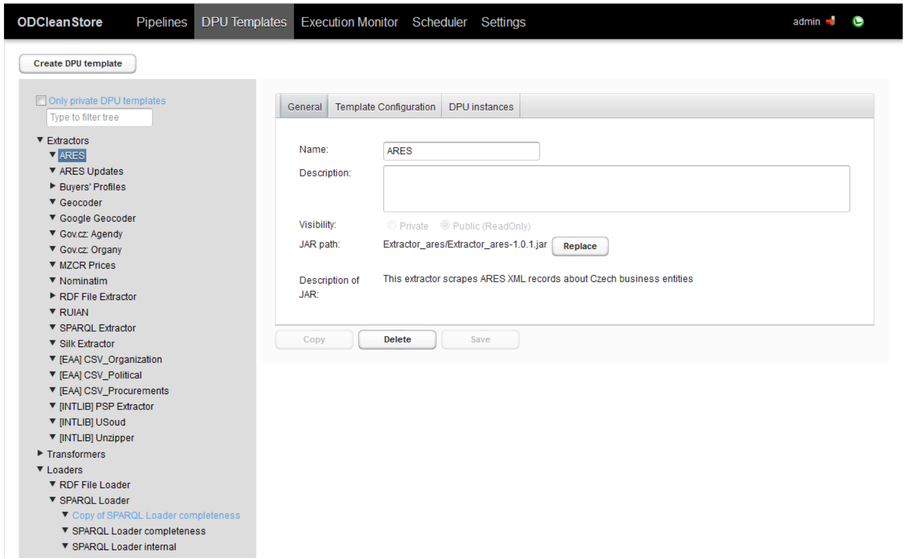
Rhino.Etl.Cmd	Updated Common.Logging to version 3.0.0 (breaking change)	a year ago
Rhino.Etl.Core	fix #8	8 months ago
Rhino.Etl.Dsl	Merge from mr-miles with the added join dsl operations	4 years ago
Rhino.Etl.Tests	BugFix: JoinOperation errors are not propagated to the main EtlProces…	a year ago
SharedLibs	Updated Common.Logging to version 3.0.0 (breaking change)	a year ago
Tools	Removed the Aggregate.* dlls	3 years ago
.gitignore	Ignored files that change during testing	3 years ago
How to build.txt	Adding How to build.txt	6 years ago
README.md	Updated README and Version	9 months ago
Rhino.Etl.sln	Upgraded main sln file to vs2010 (kept a 2008 for old times sake)	5 years ago
Rhino.Etl.vs2008.sln	Upgraded main sln file to vs2010 (kept a 2008 for old times sake)	5 years ago
acknowledgements.txt	Updated the acknowledgements.txt with the contributors	3 years ago
ayende-open-source.snk	Changing the structure to the new one	7 years ago
default.ps1	Updated README and Version	9 months ago
ilmerge.exclude	Renaming branch	8 years ago
license.txt	Updated the copyright	3 years ago

- C# .Net Framework
- Open source
- Hello-World application available on GitHub
- Pure framework no additional connectors

Source: https://hibernatingrhinos.com/oss/rhino-etl

Existing ETL-Tools: UnifiedViews

UnifiedViews



Name: ARES

Description:

Visibility: ○ Private ● Public (ReadOnly)

JAR path: Extractor_ares/Extractor_ares-1.0.1.jar

Description of JAR: This extractor scrapes ARES XML records about Czech business entities

- Java based
- Open Source
- Specialized on RDF-Data (Linked data)
- Visual Designer to build Job (Web-Based)
- Extendable through plugins
- Developed at *Charles University, Prague*

Source: http://unifiedviews.eu/

Existing ETL-Tools: Conclusion

- Almost all tool have a generic use case domain, but are manly advertised for Business Intelligence and Big Data Integration / Analysis
  - Tools have thousands of adapters, transformers and settings → High entry barrier
  - Heavy duty tools for “Big Data” → Higher configuration and maintenance effort
- SyncPipes is lightweight an quick to integrate into your infrastructure
  - TypeScript / JavaScript provides an ecosystem that is easily extensible
  - ~260.000 Packages available through *npm* to speed up development
  - RESTful API assures easy integration into existing system architecture
  - Rule of thumb: Create new adapters and the corresponding Workflow within a day
  - Docker-Support out of the Box (“Zero configuration”)

Research Questions

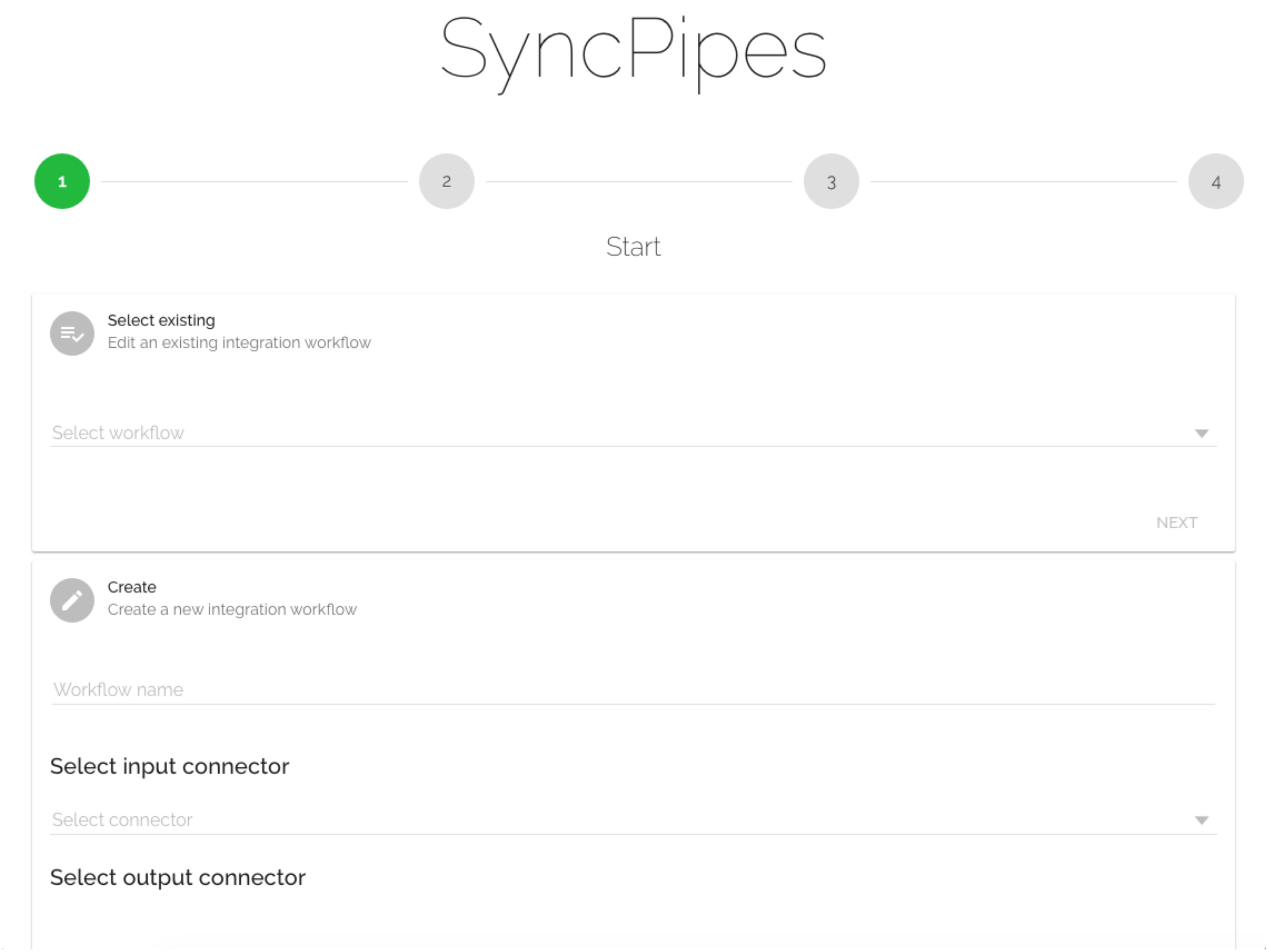
Research Objective: Create a REST-based Data Integration Framework to enable developers to implement adapters for ETL-Workflows easily. Facilitate the End-User to visualize the source and target system’s domain model in the conjunction with creating new Data Integration Workflows.

Research Questions

Q1: “What are the key features that must be supported by data integration framework?”

Q2: “How does the framework's architecture support its extensibility with new adapters?”

UI Prototype (I)



SyncPipes

Start

Select existing

Edit an existing integration workflow

Select workflow ▾

Create

Create a new integration workflow

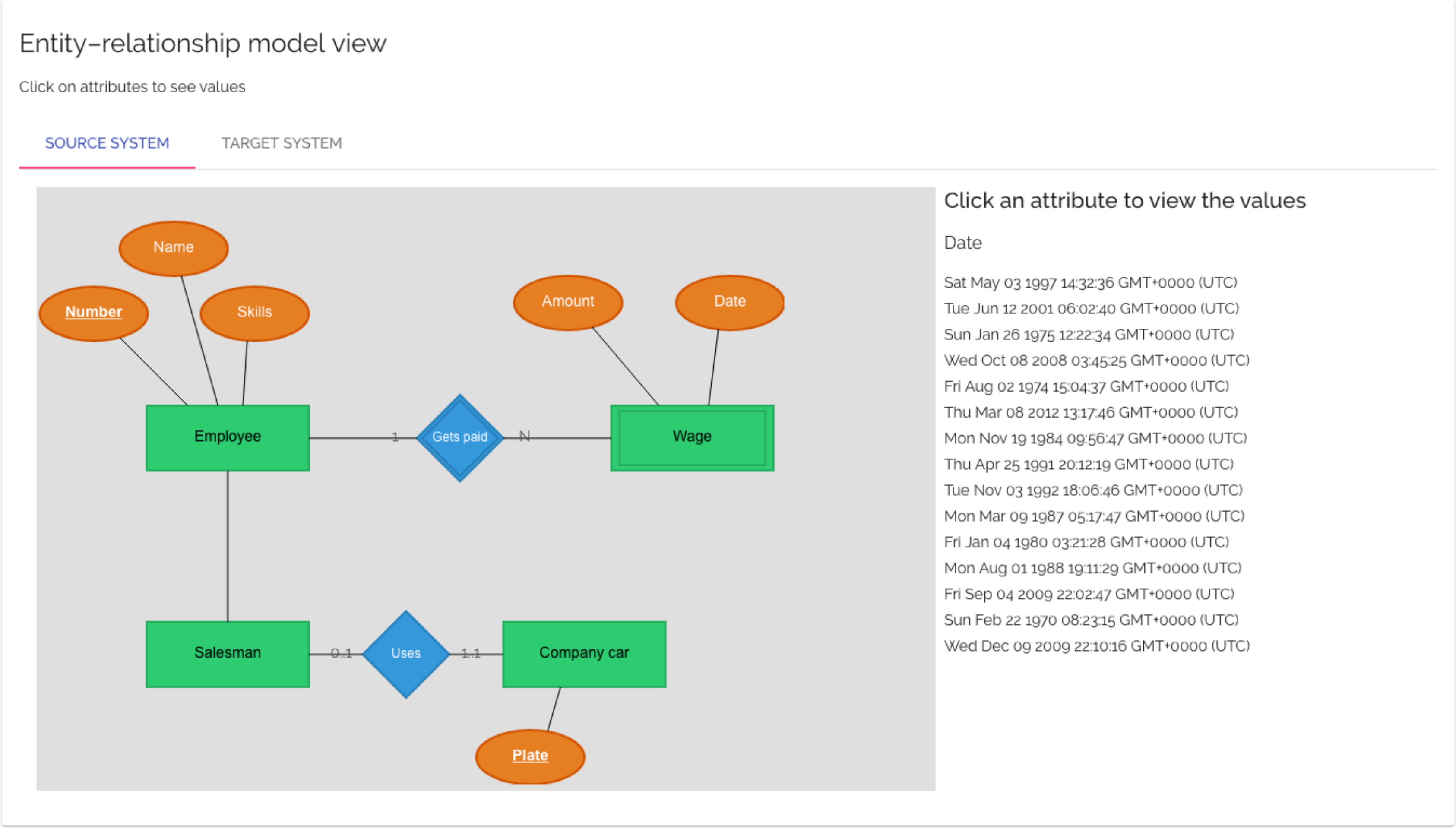
Workflow name

Select input connector

Select connector ▾

Select output connector

UI Prototype (II)



Mapping

Entity–relationship model view

Click on attributes to see values

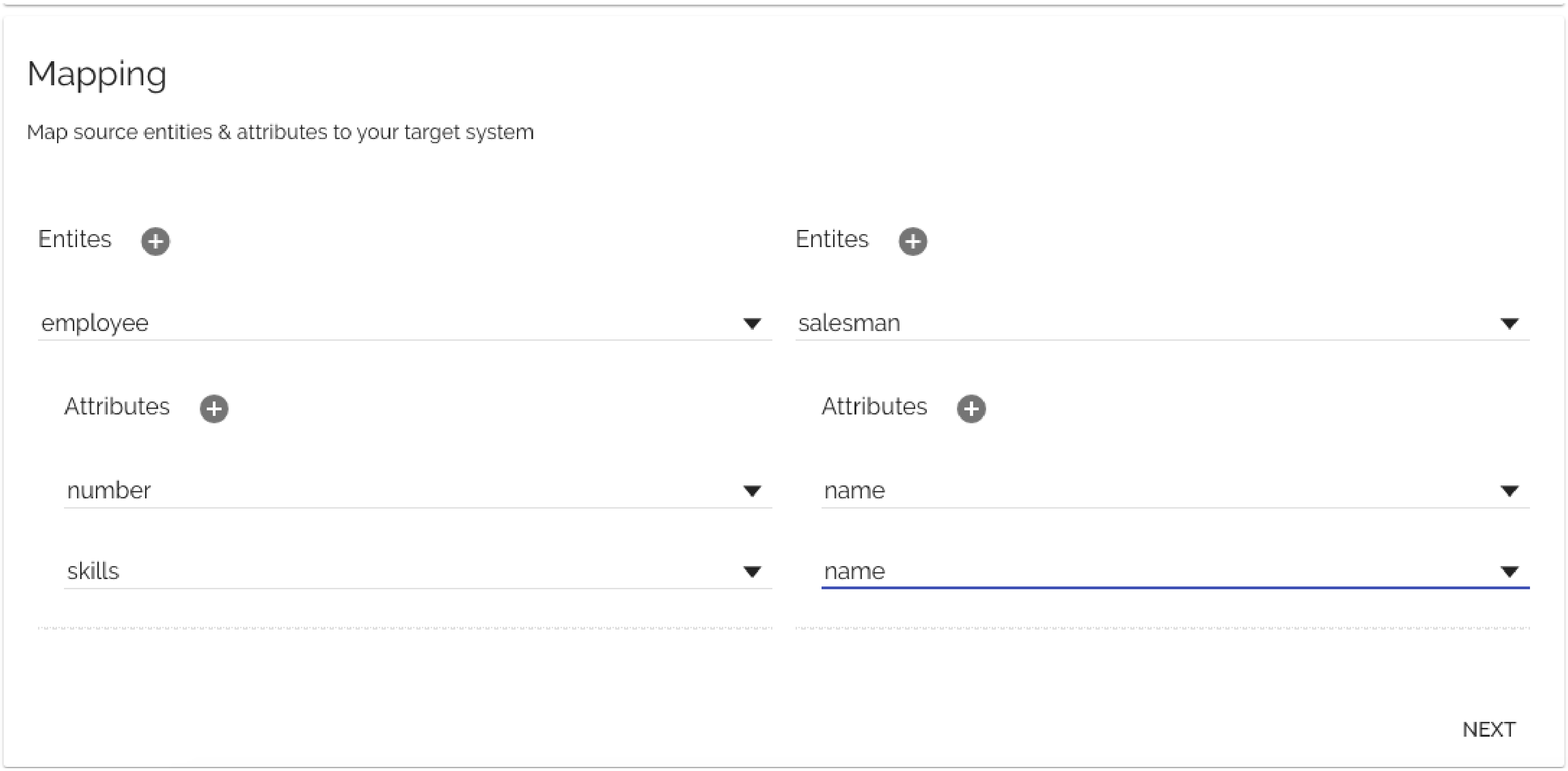
SOURCE SYSTEM | TARGET SYSTEM

Click an attribute to view the values

Date

Sat May 03 1997 14:32:36 GMT+0000 (UTC)
Tue Jun 12 2001 06:02:40 GMT+0000 (UTC)
Sun Jan 26 1975 12:22:34 GMT+0000 (UTC)
Wed Oct 08 2008 03:45:25 GMT+0000 (UTC)
Fri Aug 02 1974 15:04:37 GMT+0000 (UTC)
Thu Mar 08 2012 13:17:46 GMT+0000 (UTC)
Mon Nov 19 1984 09:56:47 GMT+0000 (UTC)
Thu Apr 25 1991 20:12:19 GMT+0000 (UTC)
Tue Nov 03 1992 18:06:46 GMT+0000 (UTC)
Mon Mar 09 1987 05:17:47 GMT+0000 (UTC)
Fri Jan 04 1980 03:21:28 GMT+0000 (UTC)
Mon Aug 01 1988 19:11:29 GMT+0000 (UTC)
Fri Sep 04 2009 22:02:47 GMT+0000 (UTC)
Sun Feb 22 1970 08:23:15 GMT+0000 (UTC)
Wed Dec 09 2009 22:10:16 GMT+0000 (UTC)

UI Prototype (III)



Mapping

Map source entities & attributes to your target system

Entites ⊕

employee ▼

Attributes ⊕

number ▼

skills ▼

Entites ⊕

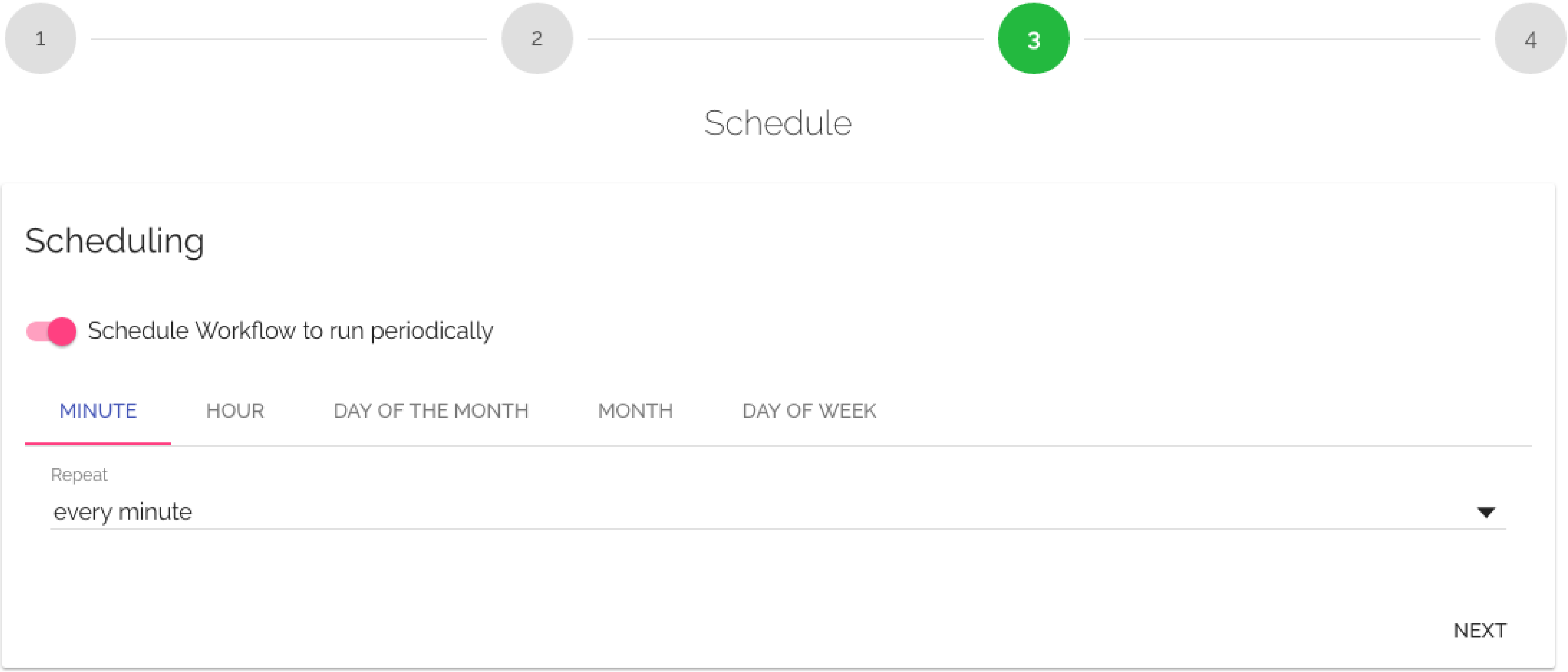
salesman ▼

Attributes ⊕

name ▼

name ▼

UI Prototype (IV)

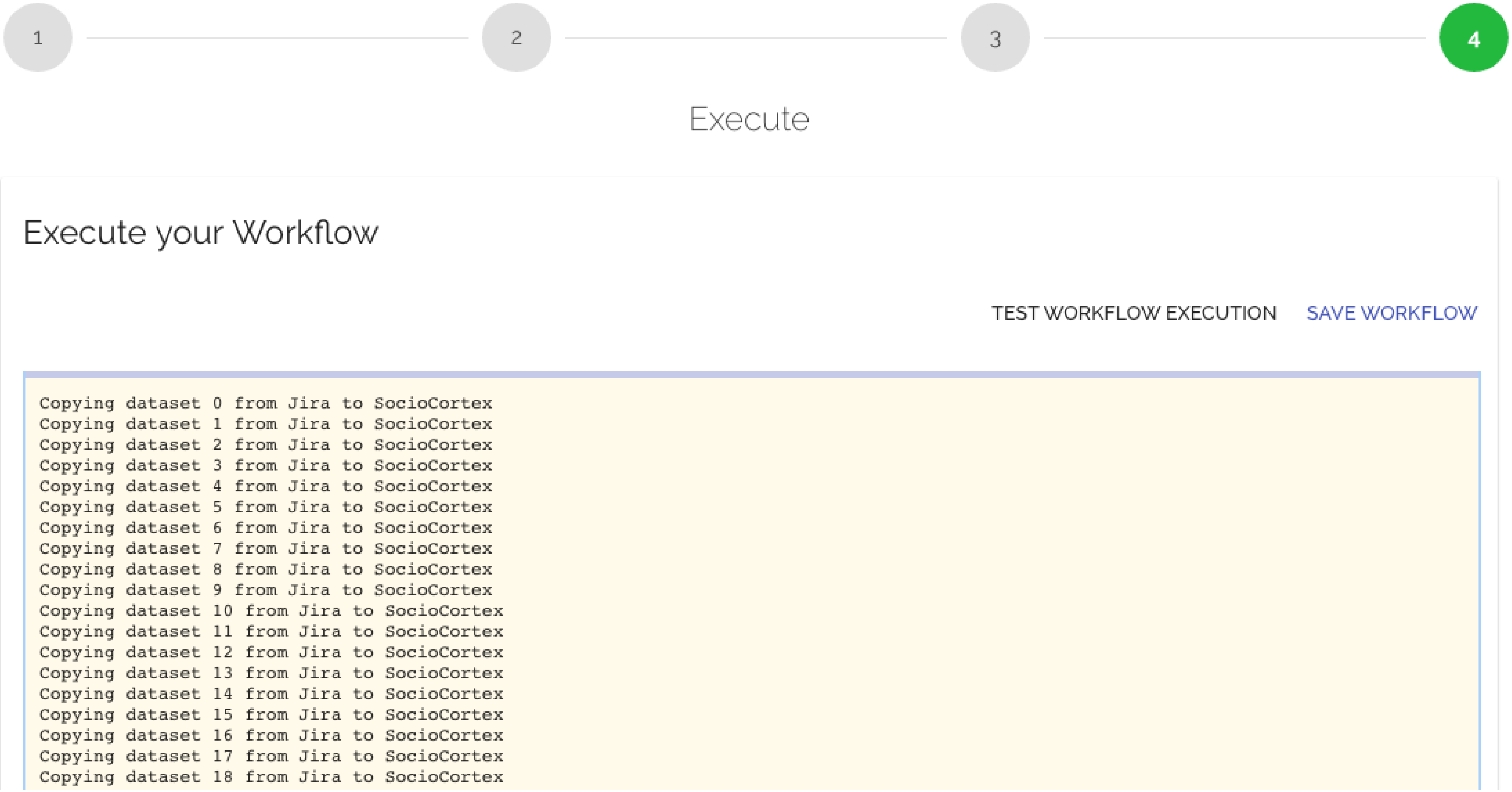


Scheduling

Schedule Workflow to run periodically

Repeat: every minute

UI Prototype (V)

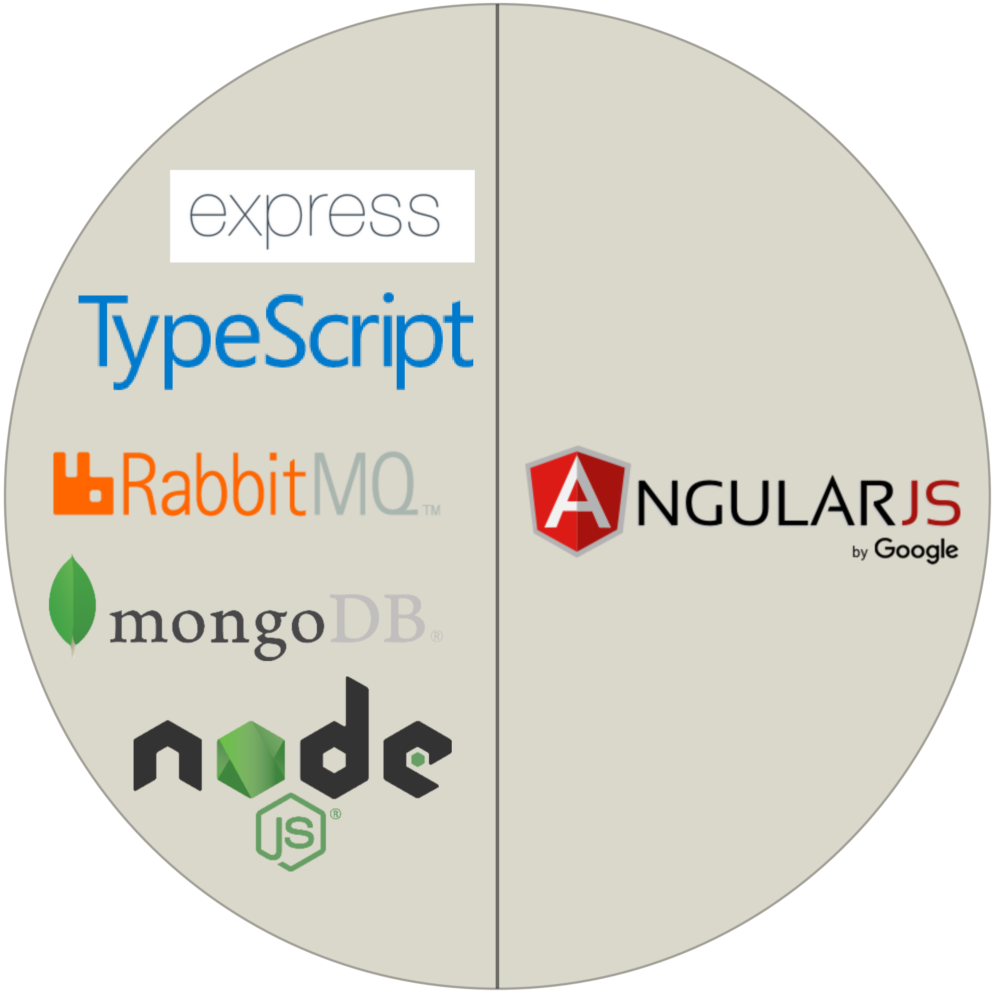


Execute

Execute your Workflow

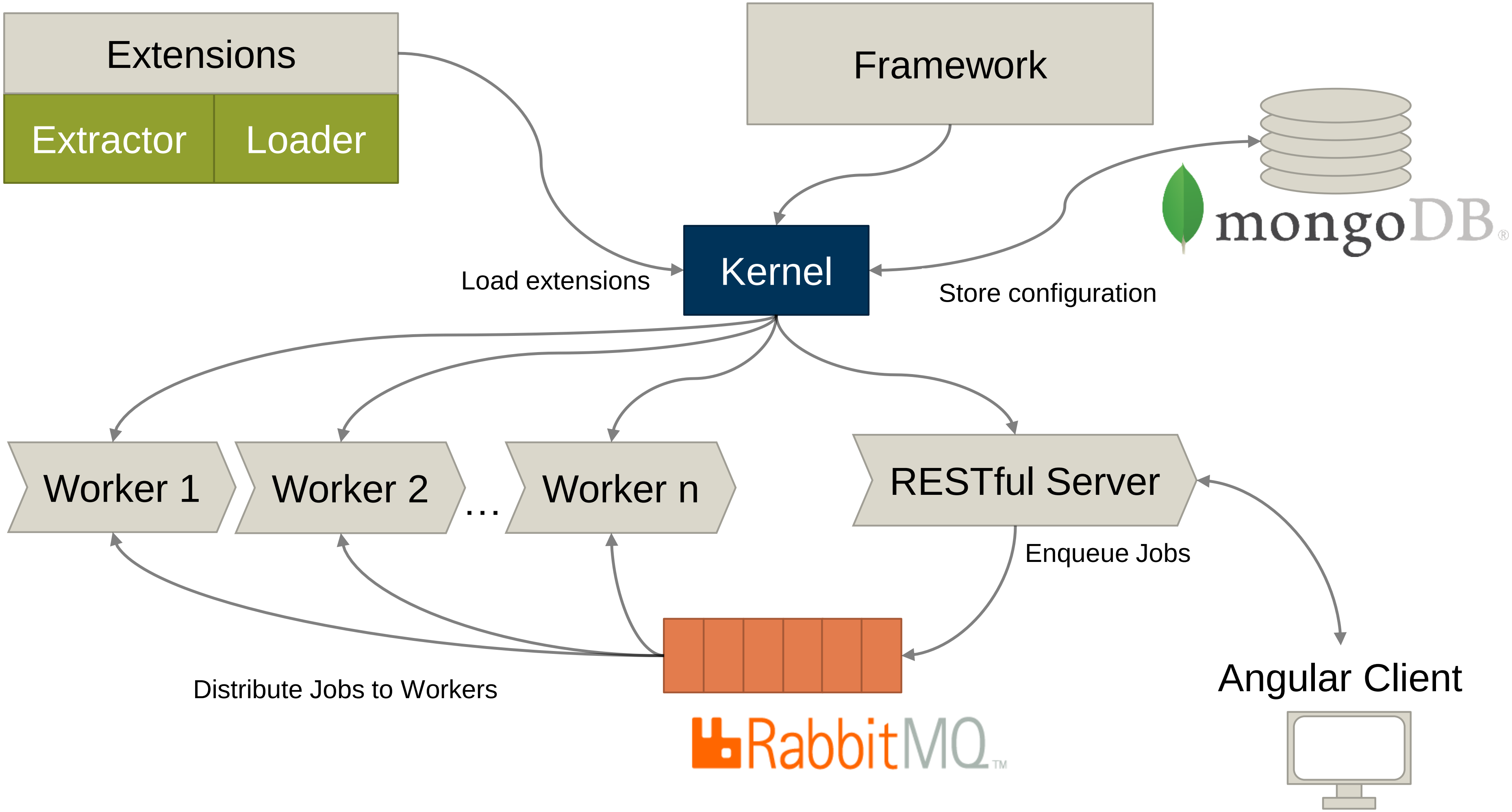
```
Copying dataset 0 from Jira to SocioCortex
Copying dataset 1 from Jira to SocioCortex
Copying dataset 2 from Jira to SocioCortex
Copying dataset 3 from Jira to SocioCortex
Copying dataset 4 from Jira to SocioCortex
Copying dataset 5 from Jira to SocioCortex
Copying dataset 6 from Jira to SocioCortex
Copying dataset 7 from Jira to SocioCortex
Copying dataset 8 from Jira to SocioCortex
Copying dataset 9 from Jira to SocioCortex
Copying dataset 10 from Jira to SocioCortex
Copying dataset 11 from Jira to SocioCortex
Copying dataset 12 from Jira to SocioCortex
Copying dataset 13 from Jira to SocioCortex
Copying dataset 14 from Jira to SocioCortex
Copying dataset 15 from Jira to SocioCortex
Copying dataset 16 from Jira to SocioCortex
Copying dataset 17 from Jira to SocioCortex
Copying dataset 18 from Jira to SocioCortex
```

Technology Stack



MongoDB **E**xpress.js **A**ngular.js **N**ode.js

Current Application Architecture



Extensions

Extractor

Loader

Framework

Kernel

Load extensions

Store configuration

Worker 1

Worker 2

…

Worker n

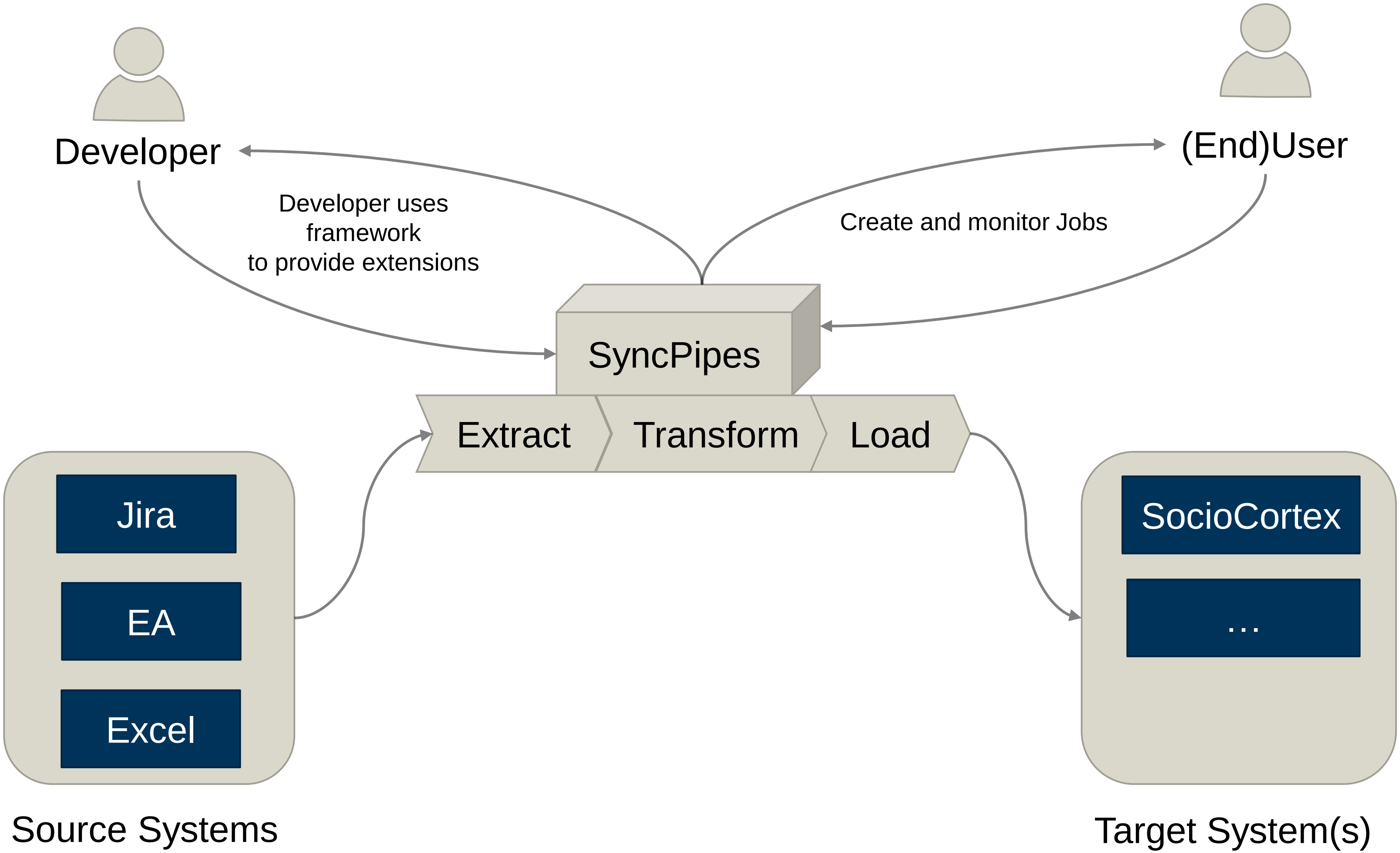
RESTful Server

Enqueue Jobs

Distribute Jobs to Workers

Angular Client

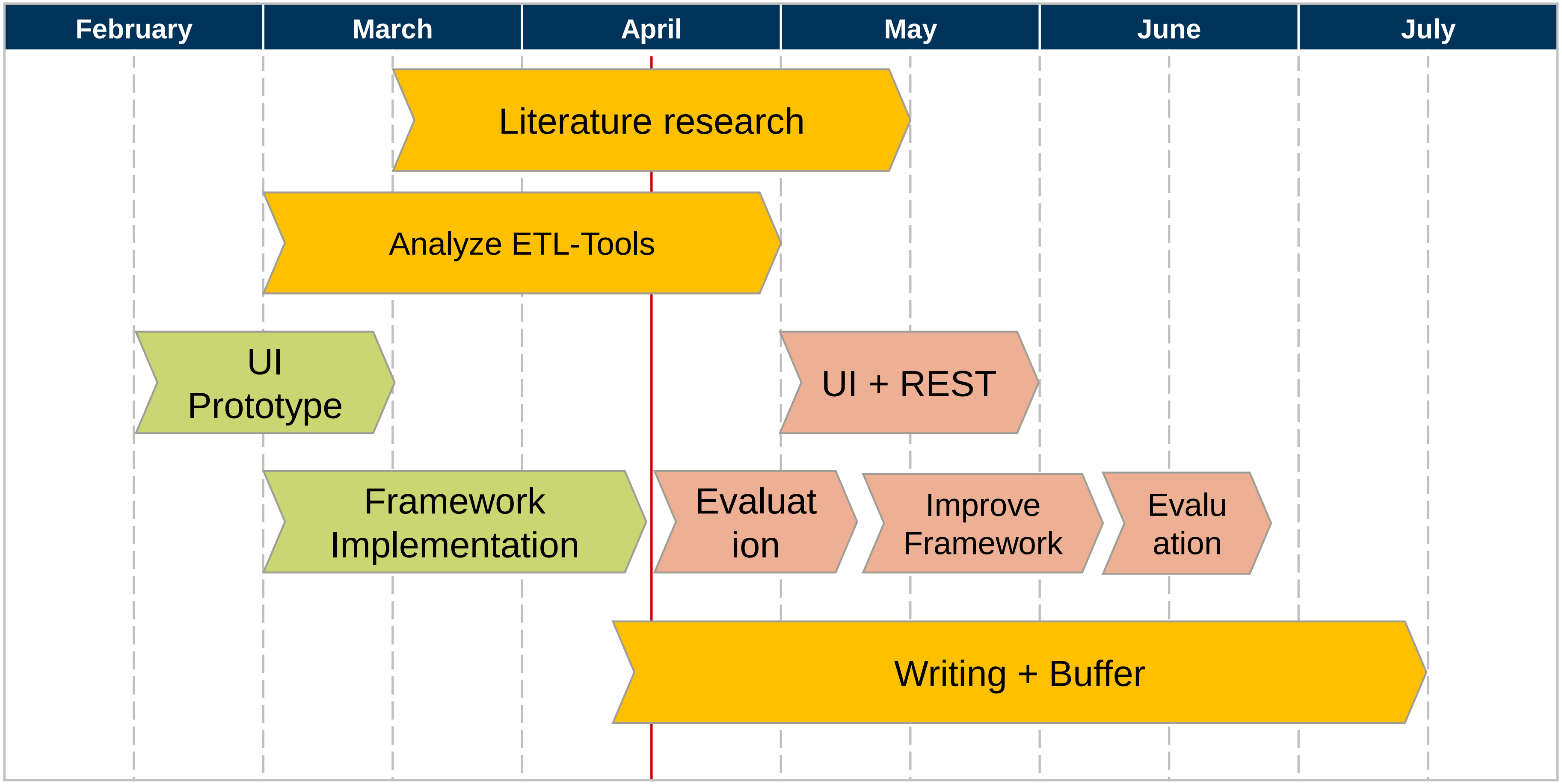
Workflow


Developer
(End)User
Developer uses framework to provide extensions
Create and monitor Jobs
SyncPipes
Extract
Transform
Load
Jira
EA
Excel
Source Systems
SocioCortex
…
Target System(s)

Next Steps

- Prototype evaluation
  1. Present prototypical implementation to 2-3 developers which a familiar with the target domain (e.g. researchers at the SEBIS chair)
  2. Ask developers to implement extractor and loader extensions
  3. Gather feedback through interviews
  4. Improve prototype based on the provided feedback
  5. Ask developers to implement similar adapters again
  6. Gather feedback
- Improve UI / Frontend to work with RESTful backend
- Write thesis

Timeline


February
March
April
May
June
July
Literature research
Analyze ETL-Tools
UI Prototype
UI + REST
Framework Implementation
Evaluat ion
Improve Framework
Evalu ation
Writing + Buffer

Thank you for your attention.

Fridolin Koch

Technische Universität München
Department of Informatics
Chair of Software Engineering for
Business Information Systems

Boltzmannstraße 3
85748 Garching bei München

frido.koch@tum.de
wwwmatthes.in.tum.de