Cross-Industry Preservation Architectures on Oracle

Robert Sharpe, Tessella

London PASIG, April 2011

Tessella
Technology & Consulting

Contents

- Tessella cross-industry archiving history
- Which industries?
- SDB4 Architecture:
  - Flexibility
  - Scalability
- ENSURE
- Conclusions:
  - Flexibility
  - Scalability

Tessella
Technology & Consulting

History

- Tessella have been working in digital archiving for a decade:
  - Mostly memory institutions
  - But some engagements with pharmaceutical etc.
- Out of this Safety Deposit Box (SDB) grown:
  - 12 customers
  - Now on version 4
  - Product roadmap
  - Support team
  - SDB Users Group

Tessella
Technology & Consulting

SDB4 Solutions Worldwide

The National Archives
UK National Archives

wellcome library / wellcome collection
Wellcome Trust

ARKISTOLAITOS ARKIVVERKET
Finnish National Archives

RAHVUSARHIIV
Estonian National Archives

Gemeente Rotterdam
Gemeentearchief
Rotterdam City Archive

nationaal archief
Dutch National Archives

FAMILYSEARCH
FamilySearch

Malaysian Archives

Science & Technology Facilities Council
STFC

Schweizerisches Bundesarchiv
Swiss Federal Archives

FEDERAL CHANCELLERY AUSTRIA
Austrian Government

Tessella
Technology & Consulting

Why do digital archiving?

- Avoid damage to organisation:
  - Need to comply with regulatory requirements
  - Defend legal claims / patent infringements etc.
  - Reputation: Need to be seen to treat information with respect
  - Cost of maintaining existing systems prohibitive
- Gain benefits:
  - Need to reuse information
- Applies to everyone but in particular:
  - Pharmaceutical
  - Health care
  - Financial
  - Aerospace
  - Nuclear
  - Oil/gas

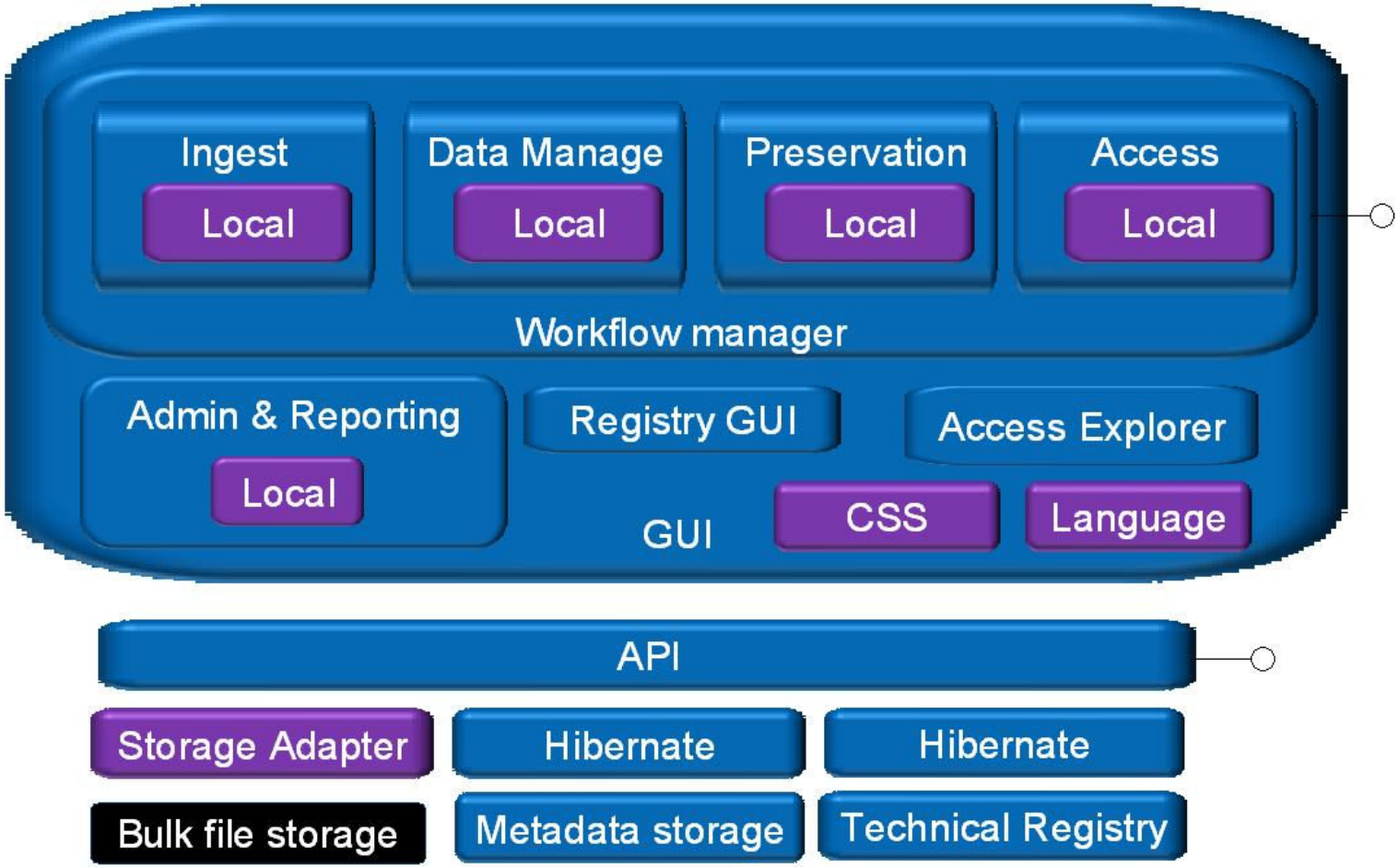
Tessella
Technology & Consulting

Demands of “other” domains

- Everything in archives / libraries etc.:
  - All of OAIS etc.
- Flexibility:
  - Take stuff from different sources in many different formats
  - Structured (data) as well as unstructured (documents)
    - Often in highly specialised formats / bespoke databases etc.
  - Privacy very important
- Scalability:
  - Hundreds of thousands of employees:
  - Process huge volumes, preferably at short notice
- Need cost/benefit analysis

Tessella
Technology & Consulting

Safety Deposit Box



Tessella
Technology & Consulting

SDB4: Cross-industry flexibility

- Choose ingest source:
  - EDRMS
  - Workflow systems
  - Web sites (e.g., via Heritrix)
  - Flat files & catalogue (easy-to-use create SIP tool)
- Choose descriptive metadata schema:
  - DON'T convert
  - Support heterogeneous schemas
  - Still allow view / edit / fielded search
  - Plus synchronisation with external catalogues (e.g., via OAI-PMH)

Tessella
Technology & Consulting

SDB4: Cross-industry flexibility

- Choose functionality (via workflow system):
  - Can add new steps
  - Can create new workflows
- Choose security (multiple tenancy):
  - Single administered instances
  - Multiple organisations / departments
- Choose storage system and AIP structure:
  - Use existing or add new storage adaptor
- Choose database engine:
  - Oracle, mySQL, SQL Server, …
- Choose reporting options

Tessella
Technology & Consulting

SDB4: Cross-industry flexibility

- Choose characterisation functionality:
  - Format identification tool
  - Format validation tool (per format)
  - Property extraction tool (per format)
  - Embedded object extraction tool (per format)
  - Logical characterisation tool
- Choose preservation functionality:
  - What's at risk?
  - Migration pathway / tool
  - Validation criteria

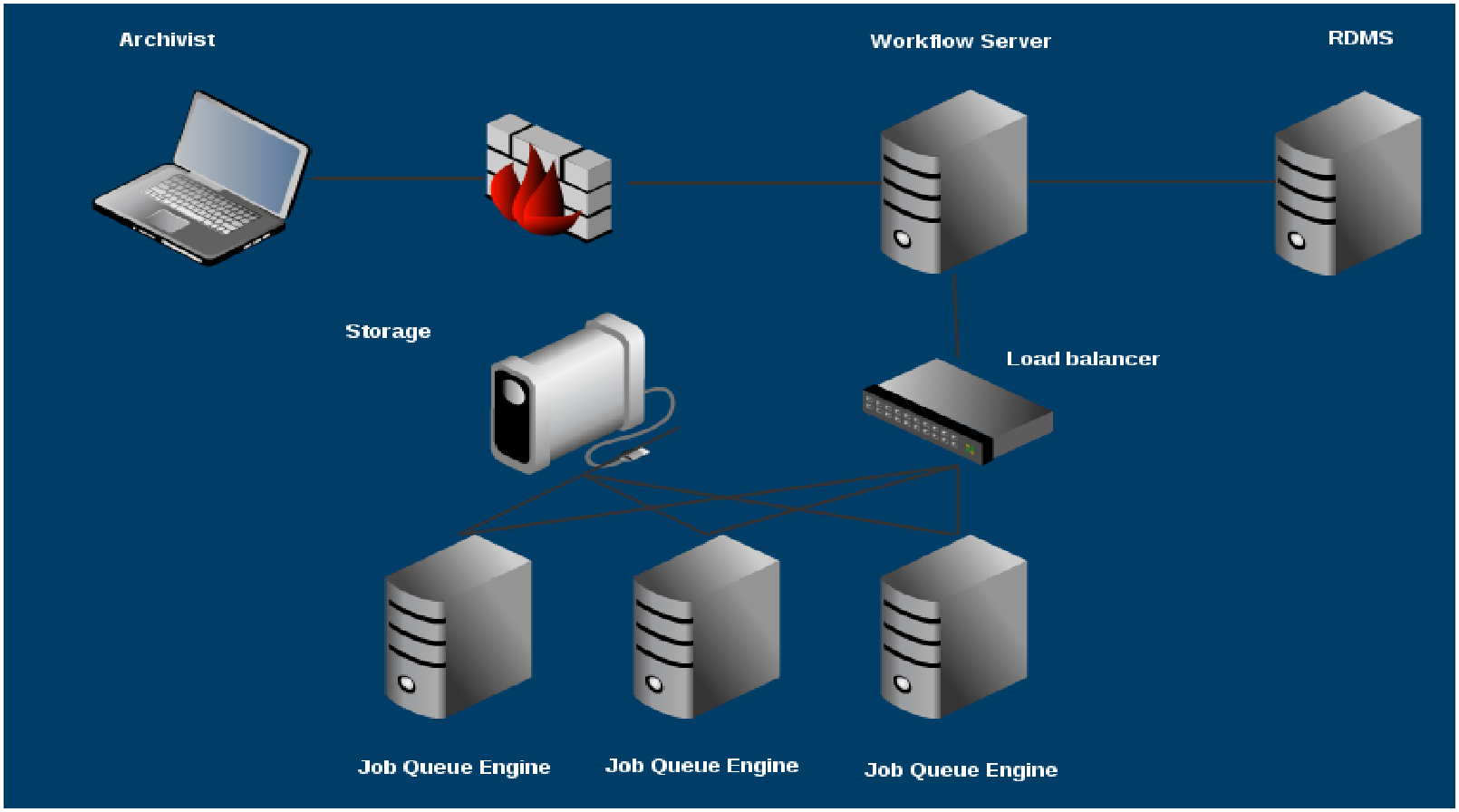
Tessella
Technology & Consulting

SDB4: Cross-industry scalability

- Lots of long-running jobs:
  - Farm out to multiple servers
  - Utilise job queuing system (control threads per server)
- How fast can we ingest?:
  - Used Oracle hardware and database in a test suite
  - Test data:
    - Thousands of 1GB SIPs (100 c.10MB files each)
    - Mix of formats (PDF, TIFF, JPEG)
  - Workflow:
    - Copy from source
    - Fixity check
    - Integrity checks
    - Characterise
    - Store content
    - Store metadata

Tessella
Technology & Consulting

SDB4: Cross-industry scalability



Tessella
Technology & Consulting

SDB4: Cross-industry scalability

- Tuned system parameters:
  - Built performance model
- Achieved 2TB/day per server (SunFire X4140):
  - BUT local server almost idle.
  - Held up by speed of reading content from source
  - Network also close to saturation
  - Hence, adding more job queue servers didn’t help

Tessella
Technology & Consulting

SDB4: Cross-industry scalability

- Working with FamilySearch:
  - 4.4GB test SIPs (c. 10MB JPEG2000 files)
  - Similar workflow
  - Need c. 20 TB/day
- Initially similar barrier, so:
  - Updated storage array (ingest queue) to 168 disks in parallel
  - Don’t move content more than needed
  - Added second job queue server
- Now achieved:
  - > 20 TB/day
- Tuning work will continue

FAMILYSEARCH
WHERE GENERATIONS MEET

Tessella
Technology & Consulting

The future

- EU FP7 Project:
  - Started 1-FEB-2011, 3 years
- Apply digital preservation to
  - Health Care
  - Clinical Trials
  - Financial Data
- Flexibility:
  - Preservation lifecycle management
  - Content-aware long term data protection
  - Privacy
- Scalability
- Evaluating cost and value

ENSURE

Tessella
Technology & Consulting

Conclusions

- Cross-organisation needs vary within memory institutions
- Cross-industry demands are also varied
- Generally all domains demand:
  - Flexibility
  - Scalability
- SDB4 is designed to meet these demands
- Research underway to demonstrate digital preservation in “new” domains

Tessella
Technology & Consulting