Data Repositories: CKAN & the Azure DataMarket

Alex D. Wade
Director – Scholarly Communication
Microsoft Research
Data Repositories

• Types of data repositories
  • National
  • Institutional
  • Disciplinary
  • Multi-disciplinary
  • Project-Specific

• Discovery of Data Repositories
  • DataBib.org (Purdue U)
  • R3Data.org – Registry of Research Data Repositories (Berlin iSchool)
Data Workflows

Curation Workflows
Takeaways

1. Consider role of data repositories *broadly* along the full lifecycle of data

2. Especially at the birth/ingest and the discovery/reuse phases

3. Awareness of some ways that Azure can help you today!
Potential uses of the Cloud

- Data repositories
  - Hosted repositories (e.g. figshare)
  - Blob storage for local data repositories (e.g. EPrints)
  - Offsite backup and ‘dark archiving’ (e.g. DuraCloud)
- Large scale data analysis
- Cloud services for data curation
- Discovery
CKAN Data Repository Software

Publish & find datasets
Publish datasets via import or through a web interface. Search by keyword or filter by tags. See dataset information at a glance. Full change history lets you easily undo changes or view old versions.

Store & manage data
Store the raw data and metadata. Visualise structured data with interactive tables, graphs and maps. Get statistics and usage metrics for your datasets. Search geospatial data on a map by area.

Engage with users & others
Federate networks with other CKAN nodes. Theme with CSS or integrate with a CMS. Build a community with extensions that allow users to comment on and follow datasets.

Customise & extend
Use the API’s rich programming interface, and benefit from over 60 extensions including link checking, comments, and analytics. CKAN’s Open Source licence allows you to download and run it for free.
CKAN vs. Dikan

- From OKFN
- [http://ckan.org/](http://ckan.org/)

- Open Source
- Existing Repositories
  - [http://ckan.org/instances/](http://ckan.org/instances/) (70)

- On Azure
  - Runs as two Azure VMs (web + db)

- From Nuams
- [http://nucivic.com/dkan](http://nucivic.com/dkan)

- Open Source
- CKAN + Drupal

- On Azure
  - Runs as Azure website
Creating a CKAN repository in Azure

1. Copy two VM Images from the VM Depot
   a) CKAN 2.1 DB
   b) CKAN 2.1 Web

2. Create a VM instance of each
   a) Create 2\textsuperscript{nd} VM instance within same cloud service as first
   b) Link the VMs

• Step by step instructions here:
  • https://github.com/okfn/ckan/wiki/How-to-install-CKAN-2.0-on-Azure-from-VM-
<table>
<thead>
<tr>
<th>NAME</th>
<th>TYPE</th>
<th>STATUS</th>
<th>SUBSCRIPTION</th>
<th>LOCATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Microsoft</td>
<td>Directory</td>
<td>Active</td>
<td>Shared by all Microsoft...</td>
<td>United States</td>
</tr>
</tbody>
</table>
virtual machines

VIRTUAL MACHINE INSTANCES  IMAGES  DISKS

No virtual machines have been created. To get started, click Create a virtual machine.
## Create a Virtual Machine

### Choose an Image

<table>
<thead>
<tr>
<th>Featured</th>
<th>Image Name</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Windows Server 2012 R2 Datacenter</td>
<td>Windows Server 2012 R2 Datacenter</td>
</tr>
<tr>
<td></td>
<td>Windows Server Essentials Experience</td>
<td>Windows Server Essentials Experience</td>
</tr>
<tr>
<td></td>
<td>Windows Server 2012 R2</td>
<td>Windows Server 2012 R2</td>
</tr>
<tr>
<td></td>
<td>Windows Server 2012 Datacenter</td>
<td>Windows Server 2012 Datacenter</td>
</tr>
<tr>
<td></td>
<td>Windows Server 2008 R2 SP1</td>
<td>Windows Server 2008 R2 SP1</td>
</tr>
<tr>
<td></td>
<td>SharePoint Server 2013 Trial</td>
<td>SharePoint Server 2013 Trial</td>
</tr>
<tr>
<td></td>
<td>SQL Server 2014 CTP2 Evaluation</td>
<td>SQL Server 2014 CTP2 Evaluation for Data Warehousing</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Windows Server 2012 R2</td>
</tr>
</tbody>
</table>

### Windows Server 2012 R2 Datacenter

At the heart of the Microsoft Cloud OS vision, Windows Server 2012 R2 brings Microsoft’s experience delivering global-scale cloud services into your infrastructure. It offers enterprise-class performance, flexibility for your applications and excellent economics for your datacenter and hybrid cloud environment.

- **OS FAMILY**: Windows
- **PUBLISHER**: Microsoft Windows Server Group
- **LOCATIONS**: East Asia, Southeast Asia, North Europe, West Europe, Japan, East Japan, West East US, West US

### Pricing Information

Pricing varies based on the subscription you select to provision your virtual machine.
virtual machines

No images have been created. To get started, click Create an image or Browse VM Depot.

CREATE AN IMAGE

BROWSE VM DEPOT
Choose an image

- Alfresco 4.2.c-2 (Ubuntu 12.10)
- Alfresco 4.2.c-3 (Ubuntu 12.10)
- Alfresco 4.2.d-0 (Ubuntu 12.10)
- Alfresco 4.2.d-1 (Ubuntu 12.10)
- Alfresco 4.2.e-0 (Ubuntu 12.10)
- Alfresco 4.2.e-1 (Ubuntu 12.10)
- Alfresco 4.2.e-2 (Ubuntu 12.10)
- Alliance Key Manager 3.0.1

Alfresco powered by BitNami is a pre-configured, ready to run image for running Alfresco on Windows Azure. Alfresco is an Enterprise Content Management (ECM) system.

PUBLISHER
Bitnami

PUBLISH DATE
3/25/2013

PACKAGES
alfresco, java, jdbc, mysql, tomcat

SIZE
30 GB
Choose an image

CKAN 2.1 DB

CKAN is open source, open data portal software used by many governments and organizations around the world to publish their data (incl. data.gov and data.gov.uk).

More

PUBLISHER
OKFN Services

PUBLISH DATE
10/16/2013

PACKAGES
postgisql-9.1, solr
SIZE
29.3 GB
Choose a storage account

Select a storage account to copy the image to

- IMAGE REGION
  - West US

- STORAGE ACCOUNT IN IMAGE REGION
  - Create a new storage account

- NEW STORAGE ACCOUNT NAME

ATTENTION

Microsoft is not responsible for images provided & licensed by the community members and does not screen for security, compatibility or performance. Users should perform their own due diligence before using the community images. Support for images is provided by the community and not by Windows Azure customer support. By using unsupported images, you might forfeit Windows Azure availability SLA.
Choose a storage account

Select a storage account to copy the image to

**IMAGE REGION**
West US

**STORAGE ACCOUNT IN IMAGE REGION**
Create a new storage account

**NEW STORAGE ACCOUNT NAME**
ckanbias

**ATTENTION**
Microsoft is not responsible for images provided & licensed by the community members and does not screen for security, compatibility or performance. Users should perform their own due diligence before using the community images. Support for images is provided by the community and not by Windows Azure customer support. By using unsupported images, you might forfeit Windows Azure availability SLA.

**CKAN 2.1 DB**

CKAN is open source, open data portal software used by many governments and organizations around the world to publish their data (incl. data.gov and data.gov.uk).

**PUBLISHER**
OKFN Services

**PUBLISH DATE**
10/15/2013

**PACKAGES**
postgresq1-9.1, solr

**SIZE**
29.3 GB
<table>
<thead>
<tr>
<th>NAME</th>
<th>SUBSCRIPTION</th>
<th>STATUS</th>
<th>LOCATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>CKAN-2-1-DB</td>
<td>Pay-As-You-Go</td>
<td>Copy in progress (67%)</td>
<td><a href="http://ckandbtest.blob.core.windows.net/ckan">http://ckandbtest.blob.core.windows.net/ckan</a></td>
</tr>
</tbody>
</table>

The image shows the Azure portal interface with a focus on virtual machines. The table displays information about a virtual machine named `CKAN-2-1-DB`, which is in a `Copy in progress`. The table also shows the subscription type as `Pay-As-You-Go`. The location information is provided as a URL. The screenshot includes the Azure portal UI with various options and sections for creating, managing, and deleting virtual machines and other services.
Choose a storage account

Select a storage account to copy the image to

IMAGE REGION

West US

STORAGE ACCOUNT IN IMAGE REGION

ckanbittest

ATTENTION

Microsoft is not responsible for images provided & licensed by the community members and does not screen for security, compatibility or performance. Users should perform their own due diligence before using the community images. Support for images is provided by the community and not by Windows Azure customer support. By using unsupported images, you might forfeit Windows Azure availability SLA.
## virtual machines

<table>
<thead>
<tr>
<th>NAME</th>
<th>SUBSCRIPTION</th>
<th>STATUS</th>
<th>LOCATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>CKAN-2.1-DB</td>
<td>Pay-As-You-Go</td>
<td>* Copy in progress (28 %)</td>
<td><a href="http://ckandblest.blob.core.windows.net/cc">http://ckandblest.blob.core.windows.net/cc</a></td>
</tr>
<tr>
<td>CKAN-2.1-Web</td>
<td>Pay-As-You-Go</td>
<td>* Copy in progress (4 %)</td>
<td><a href="http://ckandblest.blob.core.windows.net/cc">http://ckandblest.blob.core.windows.net/cc</a></td>
</tr>
</tbody>
</table>
## Virtual Machines

<table>
<thead>
<tr>
<th>Name</th>
<th>Subscription</th>
<th>Status</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>CKAN-2.1-DB</td>
<td>Pay-As-You-Go</td>
<td>Pending registration</td>
<td><a href="http://ckandblest.biob.core.windows.net/cc">http://ckandblest.biob.core.windows.net/cc</a></td>
</tr>
<tr>
<td>CKAN-2.1-Web</td>
<td>Pay-As-You-Go</td>
<td>Pending registration</td>
<td><a href="http://ckandblest.biob.core.windows.net/cc">http://ckandblest.biob.core.windows.net/cc</a></td>
</tr>
<tr>
<td>Name</td>
<td>Subscription</td>
<td>Status</td>
<td>Location</td>
</tr>
<tr>
<td>---------------</td>
<td>--------------------</td>
<td>--------------</td>
<td>-----------------------------------------------</td>
</tr>
<tr>
<td>CKAN-2.1-DB</td>
<td>Pay-As-You-Go</td>
<td>✔ Available</td>
<td><a href="http://ckandblest.blob.core.windows.net/cc">http://ckandblest.blob.core.windows.net/cc</a></td>
</tr>
<tr>
<td>CKAN-2.1-Web</td>
<td>Pay-As-You-Go</td>
<td>✔ Available</td>
<td><a href="http://ckandblest.blob.core.windows.net/cc">http://ckandblest.blob.core.windows.net/cc</a></td>
</tr>
</tbody>
</table>
virtual machines

No virtual machines have been created. To get started, click Create a virtual machine.
CREATE A VIRTUAL MACHINE

Choose an Image

- Windows Server 2012 R2 Datacenter
- Windows Server Essentials Experience
  - Windows Server 2012 R2
- Windows Server 2012 Datacenter
- Windows Server 2008 R2 SP1
- SharePoint Server 2013 Trial
- SQL Server 2014 CTP2 Evaluation
  - Windows Server 2012 R2
- SQL Server 2014 CTP2 Evaluation for Data Warehousing
  - Windows Server 2012

Windows Server 2012 R2 Datacenter

At the heart of the Microsoft Cloud OS vision, Windows Server 2012 R2 brings Microsoft's experience delivering global-scale cloud services into your infrastructure. It offers enterprise-class performance, flexibility for your applications and excellent economics for your datacenter and hybrid cloud environment.

OS FAMILY: Windows
PUBLISHER: Microsoft Windows Server Group
LOCATIONS: East Asia, Southeast Asia, North Europe, West Europe, Japan, East Japan, West East US, West US

PRICING INFORMATION
Pricing varies based on the subscription you select to provision your virtual machine.
CKAN is open source, open data portal software used by many governments and organizations around the world to publish their data (incl. data.gov and data.gov.uk) More

OS FAMILY: Linux
SUBSCRIPTION: Pay-As-You-Go
LOCATIONS: West US

PRICING INFORMATION
Pricing varies based on the subscription you select to provision your virtual machine.
Virtual machine configuration

VIRTUAL MACHINE NAME
CKANdb

SIZE
Medium (2 cores, 5.5 GB memory)

NEW USER NAME
azureuser

AUTHENTICATION
☐ UPLOAD COMPATIBLE SSH KEY FOR AUTHENTICATION
☑ PROVIDE A PASSWORD

NEW PASSWORD
CONFIRM

CKAN-2-1-DB

CKAN is open source, open data portal software used by many governments and organizations around the world to publish their data (incl. data.gov and data.gov.uk).

OS FAMILY
Linux

SUBSCRIPTION
Pay-As-You-Go

LOCATIONS
West US

PRICING INFORMATION
Pricing varies based on the subscription you select to provision your virtual machine.
Virtual machine configuration

CLOUD SERVICE
Create a new cloud service

CLOUD SERVICE DNS NAME
CKANdatabase .cloudapp.net

REGION/AFFINITY GROUP/VIRTUAL NETWORK
West US

AVAILABILITY SET
(None)
## Virtual Machine Configuration

### Endpoints

<table>
<thead>
<tr>
<th>NAME</th>
<th>PROTOCOL</th>
<th>PUBLIC PORT</th>
<th>PRIVATE PORT</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSH</td>
<td>TCP</td>
<td>22</td>
<td>22</td>
</tr>
</tbody>
</table>

### Enter or Select a Value

- Remote Desktop
- PowerShell
- SSH
- FTP
- SMTP
- DNS
- HTTP
- POP3
- IMAP
- LDAP
- HTTPS

---

**CKAN-2-1-DB**

CKAN is open source, open data portal software used by many governments and organizations around the world to publish their data (incl. data.gov and data.gov.uk... More)

**OS Family**

- Linux

**Subscription**

- Pay-As-You-Go

**Locations**

- West US

**Pricing Information**

Pricing varies based on the subscription you select to provision your virtual machine.
## Virtual Machines

### Virtual Machine Instances

<table>
<thead>
<tr>
<th>NAME</th>
<th>STATUS</th>
<th>SUBSCRIPTION</th>
<th>LOCATION</th>
<th>DNS NAME</th>
</tr>
</thead>
<tbody>
<tr>
<td>CKAN1tb</td>
<td>Started</td>
<td>Pay-As-You-Go</td>
<td>West US</td>
<td>ckendatabase.cloudepo.net</td>
</tr>
</tbody>
</table>
Store and manage your data using SQL Databases, Blobs, Tables and more.
CREATE A VIRTUAL MACHINE

Choose an Image

ALL

MICROSOFT
WINDOWS SERVER
SHAREPOINT
SQL SERVER
BIZTALK SERVER
UBUNTU
CENTOS
SUSE
ORACLE

FEATURED

Windows Server 2012 R2 Datacenter
Windows Server Essentials Experience
Windows Server 2012 R2
Windows Server 2012 Datacenter
Windows Server 2008 R2 SP1
SharePoint Server 2013 Trial
SQL Server 2014 CTP2 Evaluation
SQL Server 2014 CTP2 Evaluation for Data Warehousing

Windows Server 2012 R2 Datacenter

At the heart of the Microsoft Cloud OS vision, Windows Server 2012 R2 brings Microsoft’s experience delivering global-scale cloud services into your infrastructure. It offers enterprise-class performance, flexibility for your applications and excellent economics for your datacenter and hybrid cloud environment.

OS FAMILY
Windows

PUBLISHER
Microsoft Windows Server Group

LOCATIONS
East Asia,Southeast Asia,North Europe,West Europe,Japan

WEST EAST US

PRICING INFORMATION
Pricing varies based on the subscription you select to provision your virtual machine.
CKAN is open source, open data portal software used by many governments and organizations around the world to publish their data (incl. data.gov and data.gov.uk... More

OS FAMILY: Linux
SUBSCRIPTION: Pay-As-You-Go
LOCATIONS: West US

PRICING INFORMATION
Pricing varies based on the subscription you select to provision your virtual machine.
CREATE A VIRTUAL MACHINE

Virtual machine configuration

VIRTUAL MACHINE NAME
ckanweb

SIZE
Medium (2 cores, 3.5 GB memory)

NEW USER NAME
azureuser

AUTHENTICATION

☑ PROVIDE A PASSWORD

NEW PASSWORD
 CONFIRM

CKAN-2-1-Web

CKAN is open source, open data portal software used by many governments and organizations around the world to publish their data (incl. data.gov and data.gov.uk...)

OS FAMILY
Linux

SUBSCRIPTION
Pay-As-You-Go

LOCATIONS
West US

PRICING INFORMATION
Pricing varies based on the subscription you select to provision your virtual machine.
Virtual machine configuration

CLOUD SERVICE
CKANdatabase

CLOUD SERVICE DNS NAME
CKANdatabase.cloudapp.net

REGION/AFFINITY GROUP/VIRTUAL NETWORK
West US

AVAILABILITY SET
(None)

CKAN-2-1-Web

CKAN is open source, open data portal software used by many governments and organizations around the world to publish their data (incl. data.gov and data.gov.uk... More

OS FAMILY
Linux

SUBSCRIPTION
Pay-As-You-Go

LOCATIONS
West US

PRICING INFORMATION
Pricing varies based on the subscription you select to provision your virtual machine.
## Virtual machine configuration

### ENDPOINTS

<table>
<thead>
<tr>
<th>NAME</th>
<th>PROTOCOL</th>
<th>PUBLIC PORT</th>
<th>PRIVATE PORT</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSH</td>
<td>TCP</td>
<td>22</td>
<td>22</td>
</tr>
</tbody>
</table>

### CKAN-2-1-Web

CKAN is open source, open data portal software used by many governments and organizations around the world to publish their data (incl. data.gov and data.gov.uk...)

**More**

**OS FAMILY**
- Linux

**SUBSCRIPTION**
- Pay-As-You-Go

**LOCATIONS**
- West US

### LEGAL TERMS

By clicking the Submit button, I acknowledge that I am getting this software from OKFN Services and that the legal terms of OKFN Services apply to it. Microsoft does not provide rights for third-party software.

**NOTE**

This is an unsupported image supplied by a third party.
## Virtual Machines

### Virtual Machine Instances

<table>
<thead>
<tr>
<th>Name</th>
<th>Status</th>
<th>Subscription</th>
<th>Location</th>
<th>DNS Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>ckanweb1</td>
<td>Running</td>
<td>Pay-As-You-Go</td>
<td>West US</td>
<td>ckanweb1.cloudapp.net</td>
</tr>
<tr>
<td>ckanweb2</td>
<td>Starting (Provisioning)</td>
<td>Pay-As-You-Go</td>
<td>West US</td>
<td>ckanweb2.cloudapp.net</td>
</tr>
</tbody>
</table>
Your cloud service has been created!

Here are a few options to get you started:

☐ Skip Quick Start the next time I visit

Get the tools

Install a Windows Azure SDK
## Virtual Machines

<table>
<thead>
<tr>
<th>Name</th>
<th>Status</th>
<th>Subscription</th>
<th>Location</th>
<th>DNS Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>CKAN1ib</td>
<td>Running</td>
<td>Pay-As-You-Go</td>
<td>West US</td>
<td>ckandatabase.cloudapp.net</td>
</tr>
<tr>
<td>ckanweb2</td>
<td>Running</td>
<td>Pay-As-You-Go</td>
<td>West US</td>
<td>ckandatabase.cloudapp.net</td>
</tr>
</tbody>
</table>
web endpoint status PREVIEW

You have not configured a web endpoint for monitoring. Configure one to get started.

CONFIGURE WEB ENDPOINT MONITORING

quick glance

View Applicable Add-ons
<table>
<thead>
<tr>
<th>NAME</th>
<th>PROTOCOL</th>
<th>PUBLIC PORT</th>
<th>PRIVATE PORT</th>
<th>LOAD-BALANCED SET NAME</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSH</td>
<td>TCP</td>
<td>64940</td>
<td>22</td>
<td>-</td>
</tr>
</tbody>
</table>
ADD ENDPOINT

Add an endpoint to a virtual machine

Traffic coming to this endpoint will be sent to the virtual machine.

- [ ] ADD A STAND-ALONE ENDPOINT
- [ ] ADD AN ENDPOINT TO AN EXISTING LOAD-BALANCED SET

(None)
Specify the details of the endpoint

NAME
- web

PROTOCOL
- TCP

PUBLIC PORT
- 80

PRIVATE PORT
- 80

CREATE A LOAD-BALANCED SET

ENABLE DIRECT SERVER RETURN
**UPDATE IN PROGRESS**
An update is in progress. You can change the configuration settings after it finishes.

<table>
<thead>
<tr>
<th>NAME</th>
<th>PROTOCOL</th>
<th>PUBLIC PORT</th>
<th>PRIVATE PORT</th>
<th>LOAD-BALANCED SET NA...</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSH</td>
<td>TCP</td>
<td>64940</td>
<td>22</td>
<td></td>
</tr>
<tr>
<td>web</td>
<td>TCP</td>
<td>80</td>
<td>80</td>
<td></td>
</tr>
</tbody>
</table>
### Cloud Services

<table>
<thead>
<tr>
<th>NAME</th>
<th>SERVICE STATUS</th>
<th>PRODUCTION</th>
<th>STAGING</th>
<th>SUBSCRIPTION</th>
<th>LOCATION</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>CKANDatabase</td>
<td>Created</td>
<td>Running</td>
<td>-</td>
<td>Pay-As-You-Go</td>
<td>West US</td>
<td><a href="http://CKANdatabase.cloudapp.net">http://CKANdatabase.cloudapp.net</a></td>
</tr>
</tbody>
</table>
ckandatabase

Your cloud service has been created!

Here are a few options to get you started:

- [ ] Skip Quick Start the next time I visit

Get the tools 🛠️
Install a Windows Azure SDK
Discovery

Azure DataMarket
Windows Azure Marketplace

One-Stop Shop for Premium Data and Applications

Hundreds of Apps, Thousands of Subscriptions, Trillions of Data Points

applications

Idac-Facturas
Idac-Facturas es una herramienta para la gestión y almacenamiento de los archivos de factura...

CloudAnalyzer
CloudAnalyzer is a reporting, analytics and monitoring solution for Azure applications. It's...

PushLocker
PushLocker is a Software as a service solution to Push Notifications on the Windows Phone and Win...

data

Drug Prescribing by GP Practices in England
Covering all general practices in England, the data includes figures on the number of prescription...

Millennium Development Goals Database - United Nations Stat!
The Millennium Development Goals Database presents official data for more than 60 indicators to m...

FinViews Fundamentals
FinViews specialises in solutions for the financial services industry and investors that enable d...

announcements

New Bing Speech Recognition Control and Updated Bing OCR and Translator Controls on Windows Azure Marketplace
10/21/2013

Changes in Brazil local currency charges
10/11/2013

SOC 2 with CSA CCM Attestation Streamlines Security Evaluation for Windows Azure Customers
8/22/2013

Windows Azure Marketplace is available in 50 additional countries and features new, exciting content.
Integration into Azure, Excel, CRM, etc.
discussion?