Samen aanjagen van vernieuwing

Eindrapportage pilot – DCC voor Praktijkgericht onderzoek

Pilotgegevens	
Titel van pilot	Data-intensief onderzoek rondom de berichtgeving over de oorlog in Oekraïne.
Hogeschool	Hogeschool Utrecht
Contactpersoon	Stefan Leijnen
Emailadres	stefan.leijnen@hu.nl
Datum	14 december 2022

1 – Beschrijving van pilot

- *Geef een korte beschrijving van de pilot*
- *Wat was het doel van de pilot?*
- *Welk vraagstuk of probleem heb je aangepakt en hoe (doorlopen stappen/proces)?*

Aan de hand van een casus, een data-intensief onderzoek rondom de berichtgeving over de oorlog in Oekraïne, is onderzocht hoe onderzoekers, docenten en studenten kunnen worden uitgenodigd tot het werken in SURF Research Cloud.

Doel van deze pilot is om SURF Research Cloud te implementeren als platform voor data-analyse in de projecten die voortkomen uit de SPRONG Responsible Applied AI ("RAAIT"). De casus moet op een aansprekende manier duidelijk maken hoe data-analyse in SURF Research Cloud veilig en betrouwbaar mogelijk is.

Concreet resultaat is een serie tutorials waarin de mogelijkheden van data-analyse op SURF Research Cloud uitgelegd en reproduceerbaar gemaakt worden. Daarbij is expliciet het doel om te komen tot een aanpak die inzetbaar is zowel binnen als buiten het themagebied media.

Stappen Datasteward (Tineke van der Meer HU):

1. Opzetten Research Drive omgeving voor de CO door Tineke van der Meer
2. Ondersteuning Fabian Kok met Research Drive omgeving

Stappen Research Engineer (Rins Rutgers HU):

1. Opzetten SURF infrastructuur samen met Martin Brandt (SURF)
2. Opzetten data-verzameling samen met Fabian Kok
3. Ondersteuning Fabian Kok met Research Cloud omgeving

Stappen Datascientist (Fabian Kok HU):

1. Opzetten data-verzameling samen met Rins Rutgers
2. Opzetten werkomgeving in Research Cloud (incl. koppelingen bouwen met Research Drive, SRAM, remote SSH, etc.)
3. 1e kleine model bouwen in Research Cloud op een subset van de data, testen van de infrastructuur
4. Groot model draaien op grote workspace, resultaten borgen in Research Drive
5. Opstellen van documentatie en gedocumenteerde script tutorial
6. Opnemen en editen van video tutorial

2 – Resultaat & impact

- *Wat zijn de resultaten? Welk resultaat heb je bereikt? Wat is er klaar?*
- *Wat heb je geleerd?*
- *Hoe ga je hiermee verder?*
- *Wat is de impact op (het faciliteren) van praktijkgericht onderzoek?*

De opgeleverde resultaten zijn als volgt:

- Een 1-uur durende video tutorial die uitlegt wat de SURF Research Cloud en aanverwante services zijn, wanneer je deze zou willen gebruiken, en hoe je deze op kunt zetten binnen je eigen organisatie met de HU als voorbeeld. De video heeft zowel timestamps voor gebruiksgemak als verschillende links naar belangrijke bronnen die kijkers verder kunnen helpen indien ze geïnteresseerd zijn in de besproken services. Afsluitend wordt de kijker meegenomen door een praktische use-case waarbij een workflow wordt getoond gebruikmakend van de Research Drive, Research Cloud en SRAM.
- Een PowerPoint presentatie met het script van de video tutorial in de comments dat gebruikt kan worden als alleenstaande documentatie of als naslagwerk. Links naar belangrijke en/of interessante bronnen zijn hierin opgenomen om te ondersteunen in het opzetten van een eigen Research Cloud omgeving (en aanverwanten).
- Twee Research Cloud Catalogus items:
    - **File Client**: Een gedocumenteerd Jupyter Notebook script dat dient als tutorial voor het gebruik van BERTopic modeling in de Research Cloud omgeving, waarbij deze gelinkt is aan de Research Drive en de file server. Hierin wordt uitgelegd hoe de data ingelezen kan worden, de techniek toegepast kan worden, en de resultaten naar Research Drive geëxporteerd kunnen worden. Een file client wordt na het opstarten automatisch gekoppeld aan de file server.
    - **File server**: Een linux sshfs waarop de data persistant staat opgeslagen, deze geeft direct toegang tot de data aan alle file clients.

Er is veel nieuwe kennis opgedaan binnen de organisatie, en voornamelijk ook het lectoraat AI van Stefan Leijnen, hoe we de Research Cloud en aanverwante services kunnen opzetten en welke waarde ze leveren. Al meerdere onderzoekers binnen het lectoraat AI van de HU zijn via onze communicatie enthousiast geworden over de mogelijkheden die het platform biedt en zijn al in gesprek met ons team Digitale Onderzoeksomgeving om Research Cloud in gebruik te nemen. Tevens is er een scherper beeld gekomen van zowel de bronnen binnen en buiten de Hogeschool Utrecht (HU) die betrekking hebben tot de SURF Research Cloud als welke individuen betrokken zijn bij het opzetten van de infrastructuur. De kennis van dit netwerk is al van waarde gebleken toen vragen binnenkwamen van geïnteresseerde onderzoekers om ze in de goede richting te kunnen wijzen.

Met de opgedane kennis en opgeleverde resultaten voorzien we een grote impact op de zichtbaarheid van de mogelijkheden die SURF biedt voor onderzoekers, docenten, en zelf studenten, alsmede ook het verlagen van de drempel voor de doelgroepen om aan de slag te gaan met de onderzoek services van SURF.

3 – Diensten

- *Welke SURF-diensten waren onderdeel van de pilot?*
- *Hoe zijn deze diensten ingezet of gebruikt binnen de pilot?*

Er is gebruik gemaakt van de volgende SURF-diensten:

- SURF Research Cloud
- SURF Research Drive
- SRAM

Van SRAM is gebruik gemaakt om nieuwe leden toe te voegen aan de CO en remote SSH op te zetten. De Research Cloud is gebruikt voor het opslaan van de GDELT data die onttrokken is uit Google BigQuery, en het uitvoeren van het NLP project via BERTopic modelling op deze dataset. De Research Drive is gebruikt om de resultaten te borgen en gemakkelijk deelbaar te maken. Allen zijn verder gebruikt voor het maken van een video tutorial over het gebruik – en het opzetten van – de SURF-diensten ter ondersteuning van onderzoek.

4 – Betrokkenen en inzet

- *Wie waren betrokken bij de pilot (rollen/functies)?*
- *Welk lectoraat(en) of kenniscentra waren betrokken bij de pilot?*
- *Welke inzet en expertise is vanuit SURF geleverd?*
- *Urenverantwoording graag aanleveren in de verstrekte Excelfile*

Het kenniscentrum Digital Business en Media van Hogeschool Utrecht is betrokken bij deze pilot. In het bijzonder zijn via de SPRONG Responsible Applied AI de volgende lectoraten betrokken:

- Artificial Intelligence, Stefan Leijnen, lector Artificial Intelligence;
- Kwaliteitsjournalistiek in digitale transitie, Yael de Haan, lector Kwaliteitsjournalistiek in Digitale Transitie;
- Betekenisvol Digitaal Innoveren, Johan Versendaal, lector Betekenisvol Digitaal Innoveren.

Naast onderzoekers, docenten en studenten zijn verder de volgende personen betrokken:

Daan Kolkman, programmamanager Digital Business en Media
Tineke van der Meer, datasteward Kenniscentrum Digital Business en Media en Digitale Onderzoeksomgeving.
Rins Rutgers, research engineer Digitale Onderzoeksomgeving
Margreet Riphagen, product owner Digitale Onderzoeksomgeving
Fabian Kok, docent onderzoeker Lectoraat AI

5 - Lessons learned

- *Zijn er lessons learned?*
- *Zijn er aandachtspunten of verbeterpunten?*

Met opmerkingen [DK1]: Ik zou hier ook graag een "doorkijk" richting de toekomst willen opnemen. Zie hieronder voor een voorzet, wellicht kunnen jullie daaraan schaven zodat het inhoudelijk ook klopt?

Punten mbt data opslag:

- Het was niet mogelijk om vanuit een workspace data aan te passen/toe te voegen aan de gelinkte fileserver waar de GDELT data opstond.
- Gemaakte storage lijkt niet verbonden te kunnen worden aan verschillende workspaces, waar hardware-technisch wel begrip voor is, maar wel eventueel interessant zou kunnen zijn in collaborative workspaces.

Punten mbt Research Cloud:

- Fabian heeft zichzelf access rights moeten geven om packages te kunnen installeren als user. In Anaconda (of miniconda dat geïnstalleerd is) wil je niet iets installeren op sudo niveau, maar op user niveau, zodanig heeft hij zichzelf via chmod write rechten toe moeten geven binnen de opt/miniconda folder.
- Toegang tot de Research Cloud via researchcloud.surf.nl gaf aparte oauth2 errors, met wat doorklikken kon er alsnog ingelogd worden.
- Inloggen op een workspace via JupyterHub gaf vaak een redirect naar een pagina die aangaf dat de omgeving opgezet werd, en dat de pagina automatisch zou refreshen wanneer dit klaar zou zijn. Dit gebeurde enkel niet, en elke keer dat deze pagina opkomt geeft een snelle f5 wel meteen toegang tot de omgeving.
- Applicaties die geen JupyterHub omgeving hebben (bijv een simpele Ubuntu 20.0 workspace) hebben wel een gele access knop en als je daarop klikt kom je wel op een JupyterHub log-in scherm terecht, enkel werken de inloggegevens niet, wat nogal verwarrend kan werken voor nieuwe users.
- Fabian gebruikte remote ssh om via VSCode te programmeren in de Research Cloud, enkel werkte de automatische prompts niet van VSCode om bijvoorbeeld Python te installeren voor het runnen van notebooks (omdat de Python kernels op de een of andere manier nog niet erkend werden). Fabian moest hier zelf even op Googlen en wat manuele installs doen. Niet heel veel werk, maar zou een drempel kunnen vormen voor sommigen die net wat minder technisch onderlegd zijn.
- Het is ietwat onduidelijk wanneer je een workspace aanmaakt en daar aangemaakte HPC storage aan wilt linken dat deze niet opduikt onder het tabje Datasets maar 2 kopjes later onder Opties.

Naast de verbeterpunten die hierboven uiteengezet zijn, zien we aanvullende mogelijkheden om de dienstverlening van SURF bekender en bereikbaarder te maken voor (praktijkgericht) onderzoek. Hierbij kan onder andere gedacht worden aan het (door)ontwikkelen van de video tutorials door verdere uitdieping op Natural Lanugage Processing gebied, of het uitbreiden naar andere AI-tereinen zoals computer vision.

6 - Kennisdeling en disseminatie

- *Hoe zijn de resultaten van het pilot beschikbaar gesteld binnen de eigen hogeschool?*
- *In welke vorm is de opgedane kennis beschikbaar gesteld aan andere hogescholen?*

De resultaten zijn gedeeld op (interne) kanalen; de video tutorial is geplaatst op het YouTube kanaal van de HU en de resultaten worden gedeeld via het platform DCC met andere hogescholen, en de resultaten zijn zodanig ook opgesteld dat deze inzetbaar zijn binnen andere organisaties. Verder worden de resultaten ook gedeeld op de RAAIT social media (zoals LinkedIn) en meenemen op het RAAIT jaarevent in februari 2023.) De code is geplaatst op de GitHub repository van de HU.

Met opmerkingen [DK2]: Ik zou de resulterende tutorials graag ook delen op de RAAIT social media (e.g. LinkedIn) en meenemen op het RAAIT jaarevent in februari 2023.

Aankomende jaren zal er gekeken worden hoe de betrokkenen zich nog nauwer kunnen aansluiten bij het SPRONG AI programma en waar het research datamanagement kan worden versterkt door samenwerking met onderzoekers, research engineers en datastewards van andere hogescholen.

Ook is de huidige groep van betrokkenen geïnteresseerd in het aansluiten bij hackathons georganiseerd door SURF. Hierbij kunnen specifieke wensen en behoeften vanuit het onderzoeksdomein verder doorontwikkeld worden. Ook kunnen er trainingen verzorgd worden en kunnen er bezoeken gebracht worden aan andere hogescholen om best practices te delen. Hier zijn wel aanvullende middelen voor nodig. Mocht hier behoefte aan zijn, dan staat de huidige groep van betrokkenen daarvoor open.