Open data using Cloud infrastructure

An initiative to host, share and use open data using cloud services

Vani Mandava
Microsoft Research,
Redmond, WA, USA
vanim@microsoft.com

Monojit Choudhury
Microsoft Research,
Bangalore, India
monojitc@microsoft.com

ABSTRACT

In this paper, we discuss Microsoft Research Open Data, a new data repository in the cloud dedicated to facilitating collaboration across the global research community. The repository provides a single, convenient location for research datasets. In a single, convenient, cloud-hosted location, research datasets across many domains such as computer science, social science, biology, genomics and others, representing many years of data curation efforts by researchers. The datasets are accompanied by meaningful research assets such as meta data and publications. The data can seamlessly be copied to a data user's cloud subscription on powerful data science virtual machines that accelerate research reproducibility and further research outcomes using the data.

KEYWORDS

cloud, open data, data science, virtual machine, Azure, fourth paradigm, findable, accessible, interoperable, re-usable, models

1 Background

In this section, we explore some of the reasons why there is a need for a cloud based open data repository

1.1 Fourth Paradigm of Scientific Discovery

Jim Gray's fourth paradigm of discovery based on data-intensive science has come true and we observe that almost all research projects have a data component to them. According to Jim Gray in a talk[1] he gave in 2007 "I want to point out that almost everything about science is changing because of the impact of information technology.

Experimental, theoretical, and computational science are all being affected by the data deluge, and a fourth, "data-intensive" science paradigm is emerging. The goal is to have a world in which all of the science literature is online, all of the science data is online, and they interoperate with each other. Lots of new tools are needed to make this happen"

This data deluge also demonstrates a clear need for curated and meaningful datasets in the research community, not only in computer science but also in interdisciplinary and domain sciences. Microsoft Research Open Data, in a single, convenient, cloud-hosted location, offers datasets representing many years of data curation and research efforts by Microsoft that were used in published research studies.

1.2 Reproducibility of Research

Reproducibility of research[2] continues to be a difficult problem leading many to question the validity and utility of investing in research. Publishing datasets related to a research outcome is an integral part of reproducible research. While there are some datasets (PII, sensitive, etc.) that may never publish, there are hundreds of datasets that researchers have already published and are motivated to continue to do so. However, these datasets are in fragmented and hard to find locations.

1.3 Cloud based infrastructure

With data growing at an exponential rate, perceived to be over 150 ZB of data available by 2025, it is now recognized that we need to prioritize bringing processing to data versus relying on data movement through Internet bandwidth that is growing at a much slower pace. We believe that there is real utility in providing an option to bring the processing to the data. After extensive experience from 2013 through 2017 helping the academic community take their research workloads to the cloud[3], it was clear

CoDS-COMAD '19, January 3–5, 2019, Kolkata, India

ACM ISBN 978-1-4503-6207-8/19/01.
https://doi.org/10.1145/3297001.3297038

to the team building the repository that the cloud was the solution for seamless and efficient data access and reusability.

Hence, the Microsoft data repository was built in the cloud connected directly to the ability to copy data over to a cloud based virtual machine with a variety of data science tools.

The goal is to provide a simple platform to Microsoft researchers and collaborators to share datasets and related research technologies and tools. Microsoft Research Open Data[4] is designed to simplify access to these datasets, facilitate collaboration between researchers using cloud-based resources and enable reproducibility of research. We will continue to shape and grow this repository and add features based on feedback from the community.

2 Open Data Repository features

2.1 Experiences: Browse

Users can access the repository via a browse experience and an authenticated experience. The site is built using the Angular UI framework that renders it beautifully across browsers, platforms and devices. A single dataset has detailed description and meta data about the research associated with the dataset.

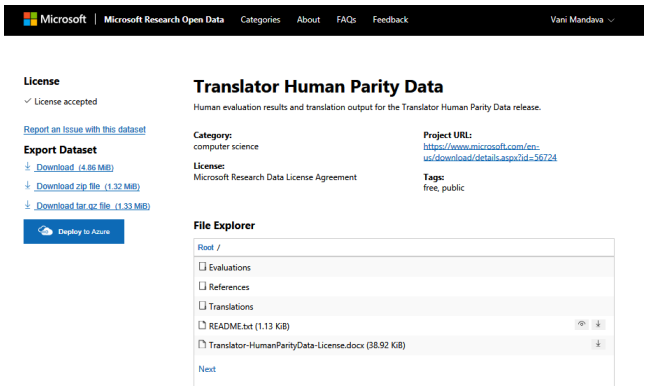

Figure 1: Dataset in Open Data Repository

Datasets in Microsoft Research Open Data are categorized by their primary research area, as shown in Figure 4. You can find links to research projects or publications with the dataset. You can browse available datasets and download them or copy them directly to an Azure subscription through an automated workflow. To the extent possible, the repository meets the highest standards for data sharing to ensure that datasets are findable, accessible, interoperable and reusable; the entire corpus does not contain personally identifiable information.

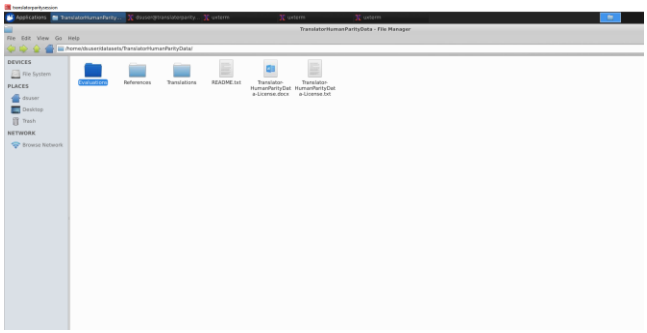

Figure 2: Dataset copied to Data Science Virtual machine

2.2 Experiences: Data Access and Reuse

If the user is interested in using a dataset, they can log in using a standard authentication provider and after accepting the license terms, either copy the dataset over to a Linux Data science virtual machine or download it directly. Figure 2 shows the dataset copied to a Linux DSVM and figure 3 shows the variety of tools available to the data scientist on the Linux DSVM, that can then be used to work with the data. Azure's Data Science VM enables out of the box data science development using popular Python and R development tools such as Juypter notebooks, Microsoft ML Server, Visual Studio code IDEs et al, and popular open source deep learning frameworks.

The data download feature generates a temporary shared access signature token that is then shared with the authenticated user to get access to the blob storage location of the dataset.

For larger datasets, it is expected that the reusability will be based on cloud based virtual machine access as discussed as it will be impractical to download the data over internet bandwidth.

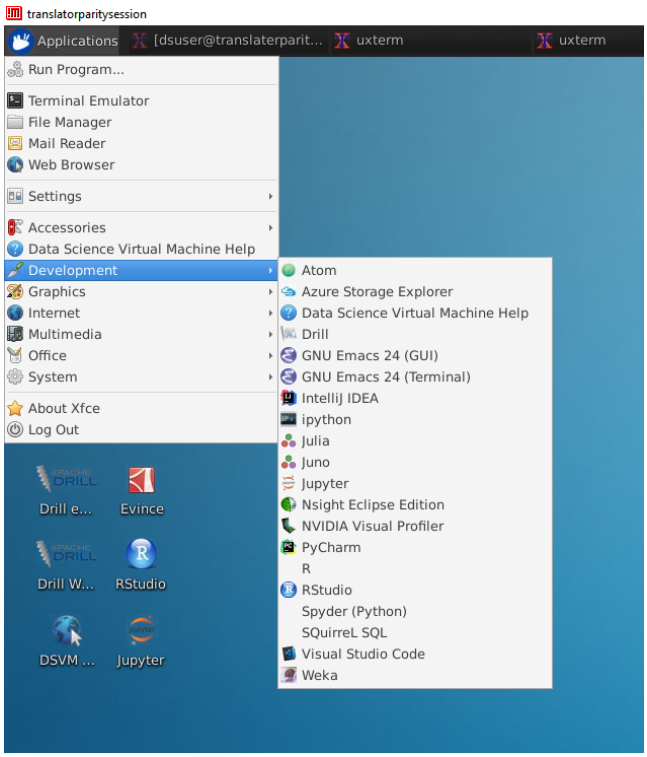

Figure 3 – IDEs available on Linux data science virtual machine

The rise of data science has resulted in an evolving need on the human infrastructure side for a data manager[5] role to help onboard and collect datasets. The repository also includes a dataset nominate feature that the data manager can use to queue new dataset requests for consideration to add to the data repository. A semi-automated process is kicked off with an email based workflow that allows the data repository administrator to approve the dataset. The dataset then gets copied over to Azure blob storage and gets assigned a GUID that is then the unique identifier for the dataset in the repository

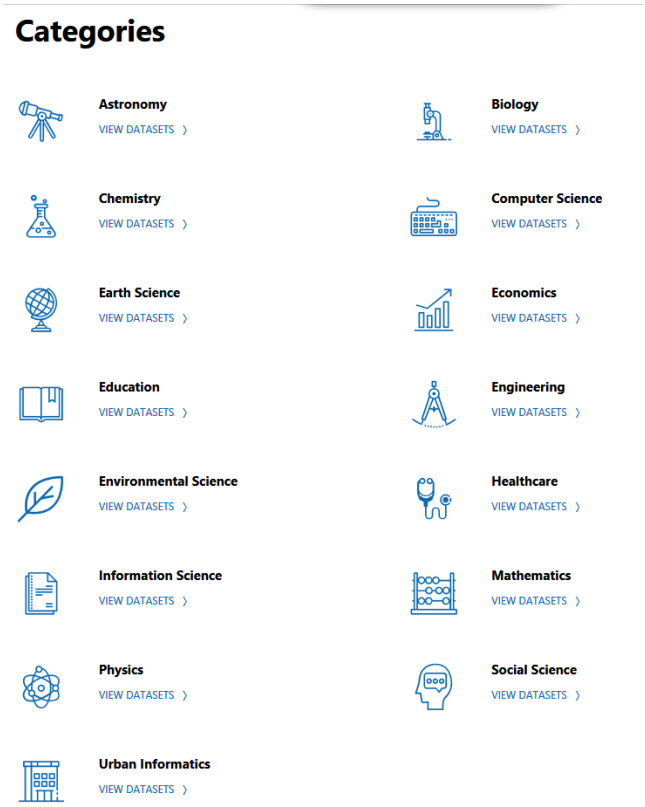

Figure 4: Dataset categories in the Open Data Repository

ACKNOWLEDGMENTS

The authors would like to acknowledge the researchers who contributed datasets to the repository, and the Azure Data Science Virtual machine team in Microsoft.

REFERENCES

[1] Jim Gray , Alex Szalay, 2007, eScience – A Transformed Scientific Method http://research.microsoft.com/en-us/um/people/gray/talks/NRC-CSTB_eScience.ppt

[2] National Academy of Sciences, Reproducibility of Research: Issues and Proposed Remedies https://phys.org/news/2018-03-scientists-issues-remedies.html

[3] Vani Mandava, 2018, National Academies, Opportunities from the Integration of Simulation Science and Data Science, https://www.nap.edu/read/25199/chapter/3#13

[4] KDNuggets, 2018 https://www.kdnuggets.com/2018/06/microsoft-research-open-data.html

[5] K Kervin, RB Cook, WK Michener, The Backstage Work of Data Sharing DOI: https://doi.acm.org/10.1145/2660398.2660406