Open Data Sets

Kaggle

Kaggle has come up with a platform where people can donate open datasets. Data engineers and other community members can have open access to these datasets and can contribute to the open data movement. They have more than 350 datasets in total, with more than 200 as featured datasets. It has a few interesting datasets on the platform that are not present at other places, and it’s a platform to connect with other data enthusiasts.

URL: https://www.kaggle.com/

AWS Open Data Sets Registry

The AWS registry exists to help people discover and share datasets that are available via AWS resources.

URL: https://registry.opendata.aws/

Google

Public Datasets on Google Cloud Platform makes it easy for users to access and analyze data in the cloud. These datasets are freely hosted and accessible using a variety of data warehouse and analytics software, from open source Apache Spark to cutting edge Google technologies like Google BigQuery and Google Cloud Dataflow. From structured genomic or encyclopedic data to unstructured climate data, Public Datasets provide a playground for those new to big data and data analysis and a powerful repository for skilled researchers. You can also integrate with your application to add valuable insights for your users. Whatever your use case, these datasets are freely available on GCP.

URL: https://cloud.google.com/public-datasets/

Azure

Microsoft Azures list of public data sets for data that you can use to prototype and test storage and analytics services and solutions.

URL:
https://docs.microsoft.com/en-us/azure/sql-database/sql-database-public-data-sets#other-statistical-and-scientific-data

IBM

Open data sets, tools, and discussion hosted by IBM Cloud Data Services (CDS). While Analytics Exchange is the primary site for finding and using open data sets with IBM's cloud databases, here

you can find processing tools, sample apps and scripts, and smaller, ad-hoc test data sets. Each directory has it's own specific README.md file describing its contents and purpose.

URL: https://github.com/ibm-watson-data-lab/open-data

US Government Public Data Sets

The home of the U.S. Government's open data. Here you will find data, tools, and resources to conduct research, develop web and mobile applications, design data visualizations, and more.

URL: https://www.data.gov/

Pew Research Data

Pew Research Center is a nonpartisan fact tank that informs the public about the issues, attitudes and trends shaping the world. We conduct public opinion polling, demographic research, content analysis and other data-driven social science research. We do not take policy positions.

URL: http://www.pewresearch.org/

Hadoop Illuminated

'Hadoop illuminated' is the open source book about Apache Hadoop™. It aims to make Hadoop knowledge accessible to a wider audience, not just to the highly technical.

https://hadoopilluminated.com/hadoop_illuminated/Public_Bigdata_Sets.html

ESS-DIVE

The U.S. Department of Energy's (DOE) Environmental Systems Science Data Infrastructure for a Virtual Ecosystem (ESS-DIVE) is a new data archive for Earth and environmental science data. ESS-DIVE is funded by the Data Management program within the Climate and Environmental Science Division under the DOE's Office of Biological and Environmental Research program (BER), and is maintained by the Lawrence Berkeley National Laboratory.

https://ess-dive.lbl.gov/