File characterization

After analyzing layers and images, we conducted a deeper analysis on the files that are stored in containers.Specifically, we characterize files in terms of size and type. Based on this characterization, we create three-level classification hierarchy as shown in Figure fig:file-type-hierarchy. At the highest level, we created two categories: Common used file types and non-common used file types based on the total file size and file count for each type. Totally, we got around 1,500 types after analyzing our whole dataset. We found that only 133 file types that take up more than 7 GB individually and occupy the most of capacity (98.4%, with 166.8 TB) totally. We put these 133 file types into common used file type group and the remaining files into non-common used file types. Our further classification expands on the 98.4% common used file types.

At the second level of the hierarchy, we clustered common used file types based on the major file format, usage, or platform involved by each file type. We identified common used file types relevant to EOF (executable, object code, and libraries), source code, scripts, documents, archival, images, databases, and others.

At the third level, we present the specific redundant file types which take a large percentage of redundant files or storage space.

figure* [width=1]graphs/graph-types-hierarchy Taxonomy of file types. fig:file-type-hierarchy