Wk2Post1

Considering HDFS, how would you structure the data directories to ensure efficient data management and retrieval? Would you keep all data in one directory or segregate it, and on what basis?

I wasn’t sure how to answer this question, so I looked it up and found this answer on Quora:  
<https://www.quora.com/What-is-the-best-directory-structure-to-store-different-types-of-logs-in-HDFS-over-time/answer/Eric-Sammer>

Eric Sammer provides this layout and descriptions:

/data : Contains canonical, raw data sets ingested from other systems. Read only to users

/user/<username> : Home directories / scratch pads for users

/etl : Contains ETL process queue directories

/tmp : Sticky-bit set scratch for tools and users (no guarantee on longevity).

How organizations choose to structure their data directories is going to specific to their needs, but this is a good place to start and will likely fit the needs of most organizations.

The article goes into much more detail and was well worth the read.