How HDFS Works

School of Information Studies
Syracuse University

HDFS: Hadoop Distributed File System

- Based on Google's GFS
- Data distributed over physical nodes
- Designed for failover
- Data stored "as is"
- Data split into blocks
- Default replication factor is three

HDFS at Work

Client:
1) Issues command to write data.csv file to HDFS.

Namenode:
2) Splits the file into 64 MB blocks (size can be changed).
3) Writes each block to a separate data node.
4) Replicates each block a number of times (default is three).
5) Keeps track of which nodes contain each block in the file.