Upgrades — Tiny Case Study 8

A tech startup has been custom building and configuring their Hadoop clusters for over a decade. They have 100T of data stored in HDFS, with triple replication, that they run Hive queries on.

With the recent economic downturn, the decision was made to move to AWS.

The lead engineer (the one who has been most involved with keeping Hadoop running) is pushing to allocate N EC2 nodes on AWS and configure HDFS and Hadoop on those: essentially replicating the in-house environment, except using AWS hardware.

Another engineer is suggesting to store all data in S3, and spin up EMR clusters whenever there is a need to query data.

Questions:

1. What are the pros/cons of either approach?
2. Why did Hadoop couple compute and storage into the same system?
3. Why is everyone decoupling compute from storage these days?