Data Partitioning — Tiny Case Study 5

Data Engineer Bob at XYZ Corp has to satisfy both compliance and data scientist users. XYZ Corp receives nightly drops of data from all of its branches. Sometimes, data arrives late, such as within a few days.

Such late data transfers are easy to identify, as the record date that does not match the receive date.

Technologically, it is trivial to partition the data by receive date (as each date can create a new data partition).

Compliance users need a quick way of identifying just the late-transfers: they are OK with having data partitioned by receive date.

Data engineers care when the events happened (they care about the date on the record—not when it was received by the file transfer server). They need a quick way of getting to all the records for a particular date—irrelevant of when they were received.

Bob recognizes that partitioning by receive date, makes data science users unhappy, and partitioning by record date makes compliance users unhappy.

Questions:

1. What are pros/cons of going with either approach?
2. What approach would you pick? Why?
3. How would you confirm that your approach is better?