Dark data - information that organizations store but never use - represents a growing crisis across industries in our increasingly digital world. While this challenge affects industries ranging from finance to manufacturing, we chose to focus our investigation on healthcare, where inefficient data management has both significant financial impact and direct consequences for patient care.

The scale of the problem in healthcare alone is staggering: every 10 minutes, hospitals worldwide generate approximately 44 petabytes of medical data [calculated from healthcare's current 10 zettabytes of data (Coughlin, 2018)]. What's most concerning is that 97% of this data becomes "dark data" - information that is stored but never accessed or utilized again (Arcadia, n.d.).

The current approach to healthcare data storage is fundamentally flawed in three critical ways:

1. Cost Inefficiency: Despite the common assumption that storage is inexpensive, hospitals currently maintain over 60% of their data in high-cost, on-premise storage tiers (Verified Market Research, n.d.). This misallocation of resources has significant financial implications - analysis of current cloud storage pricing shows that moving just 1 petabyte of data from hot to cold storage could save $264,000 annually [calculated from AWS S3 pricing tiers (Amazon Web Services, n.d.)]. More broadly, optimizing merely 1% of healthcare data storage could yield industry-wide savings of $26.41 billion per year.
2. Outdated Classification Methods: Current storage systems rely on simplistic, rule-based classification methods, primarily using "days since last access" as the main criterion for storage tier assignment. This outdated approach fails to account for the complex patterns of data usage in modern healthcare settings.
3. Lack of Predictive Capabilities: Existing storage systems operate reactively rather than proactively. They cannot predict future access patterns or adapt to changing data usage requirements, leading to scenarios where critical patient data may be buried in cold storage while rarely accessed files consume expensive high-speed storage resources.

The implications of this problem extend beyond mere cost considerations. When critical medical data is misclassified or improperly stored, it can impact healthcare delivery speed and quality. Furthermore, maintaining unnecessary data in readily accessible storage creates potential security vulnerabilities, as excess exposed data increases the attack surface for potential data breaches.

While our solution focuses on healthcare as a proof of concept, the principles and approaches we develop could be adapted to address similar dark data challenges across other industries. The healthcare sector serves as an ideal initial focus due to its clear metrics, significant potential for cost savings, and direct impact on human welfare. This situation represents a growing crisis that requires a fundamental rethinking of how organizations approach data storage and management, starting with healthcare but potentially extending to all industries grappling with the dark data challenge.