UNIVERSITY OF CAMBRIDGE

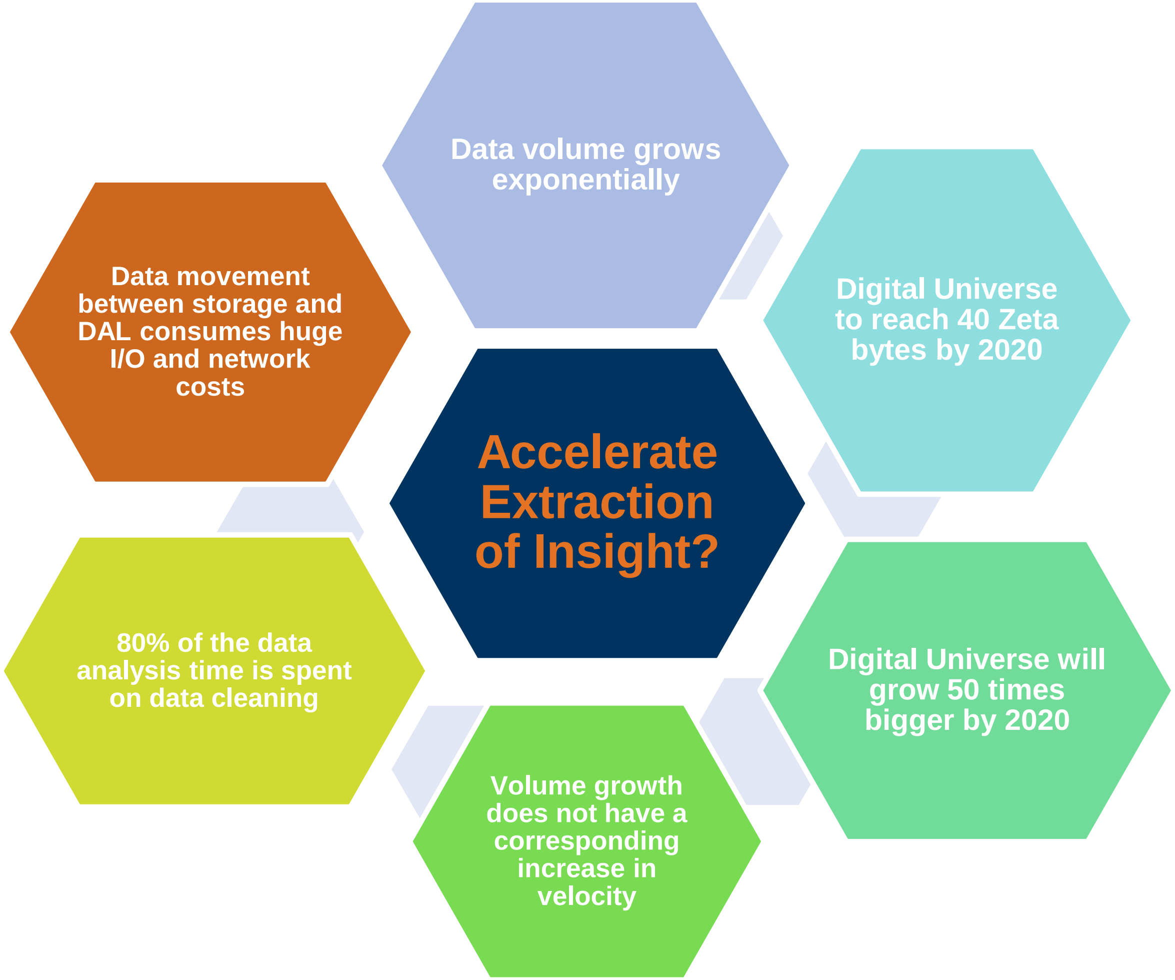
# Big data gets bigger: what about data cleaning as a storage service?

**Ayat Fekry**

**akmf3@cl.cam.ac.uk**

*Computer Laboratory, University of Cambridge*

# What is the Problem?



Accelerate Extraction of Insight?

Data volume grows exponentially

Digital Universe to reach 40 Zeta bytes by 2020

Digital Universe will grow 50 times bigger by 2020

Volume growth does not have a corresponding increase in velocity

80% of the data analysis time is spent on data cleaning

Data movement between storage and DAL consumes huge I/O and network costs

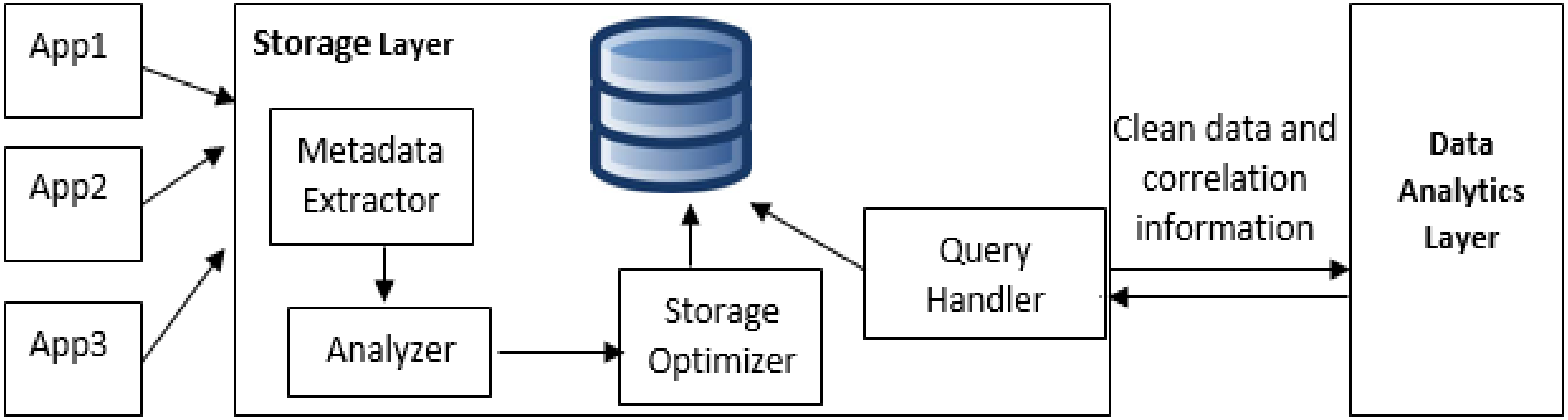
# How to solve?

Storage self-cleaning as a service that performs basic similarity and correlation analysis with minimal overhead to optimize storage space.

# Why Storage Layer?

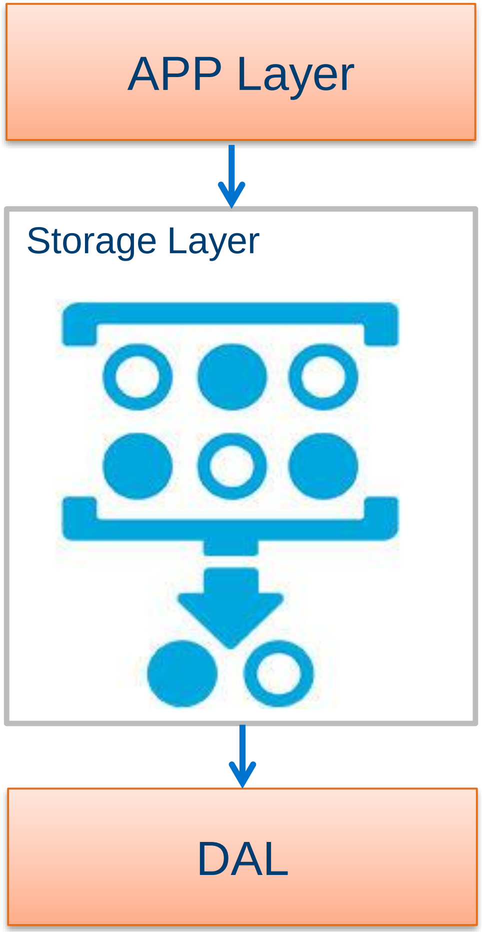
- Traditionally, data similarity and correlation analysis is done in the Data Analytics Layer (DAL).
- This implies slower velocity and increase in I/O and communication costs due to data movement.
- On the contrary, the storage layer provides a centralized and proximate place for data.
- This analysis would benefit the storage layer in terms of space optimization.

# Key Idea


App1
App2
App3
Storage Layer
Metadata Extractor
Analyzer
Storage Optimizer
Query Handler
Clean data and correlation information
Data Analytics Layer

# What are the benefits?

- Optimize storage space: deduplication on the dataset level in contrast to the conventional block level.
- Speed up data analytics: Storage provides correlation information to DAL.
- Save I/O, CPU, and network consumption due to the optimized volume.
- Reduce the "Garbage in garbage out" problem.



APP Layer

Storage Layer

DAL

# Challenges and Open questions

- Overhead: algorithm complexity?  When to execute?
- Similarity definition: exact duplicates or similar semantics? How to generalize? Rules vs ML? … Drools, Spark, and Alluxio
- Lineage support: how to leverage and generalize?
- Hardware support: ASIC support?

# Thank You!

## Questions?