面向大数据的存储系统中近似数据去重方法

华宇

华中科技大学

The Properties of Big Data (from IDC reports)

- Average storage capacity per household will grow from **464 Gigabytes** in 2011 to **3.3 Terabytes** in 2016.
- In 2012, **68%** of all data created was used by consumers watching digital TV, interacting with social media or sending camera-phone generated images.

The Properties of Big Data (con.)

- By 2020, as much as **33%** of all data will contain information that might be **valuable** if analyzed
- Nearly **75%** of our digital world has a copy, i.e., only **25%** is unique.
- *The value is difficult to exploit and use due to the data redundancy (besides noise, un-correlation, etc.).*

The importance of deduplication

- Serve as the premier of cost-efficient data analytics
- Considered as one component of analytics
- Essentially become an indexing problem
- Be helpful to obtain the value of data

The value decreases with the time, i.e., time-sensitive

The importance of deduplication

- The applications are classified into:

- Soft-deadline: scientific computing, gene discovery, earth simulation, etc.
- Hard-deadline: weather forecast, surveillance, etc.

- The support of real-time deduplication/index is important to big data.

Methodology Changes

- Many data to be handled in the Era of Big Data
- **Data value/worth significantly decreases with time**
- **Real-time definition!**
- **The analysis methodology should be changed:**

☐ *Pursing real-time performance, even with the cost of decreasing accuracy*

☐ *Not 100% exact-matching, but approximately*

How to implement from system views

- Real-time performance obtained from a ***Storage Ecosystem***
- **Ecosystem** includes:

- Not a single component due to hierarchical architecture
- But:
- Device: processors (e.g.,IEEE Micro conf.), PCM, etc
- Operating system: file systems
- Applications

Redundancy

- Consume system resources, e.g., computation, storage, network bandwidth, energy, etc.
- ***Managed redundancy*** is purposely introduced by the system to support and improve availability, reliability and load balance through data backups.
- ***The unmanaged redundancy*** is a property of the data itself and thus invisible to the system.

Backgrouds: Deduplication

- Deduplication: to delete duplicate copies
- **Forms:**

- *File-level* (e.g., different filenames with the same contents)
- *Chunk-level* (8KB-64KB, *Fixed-length Chunking or Content Defined Chunking*).

Exact-matching dedup. fails

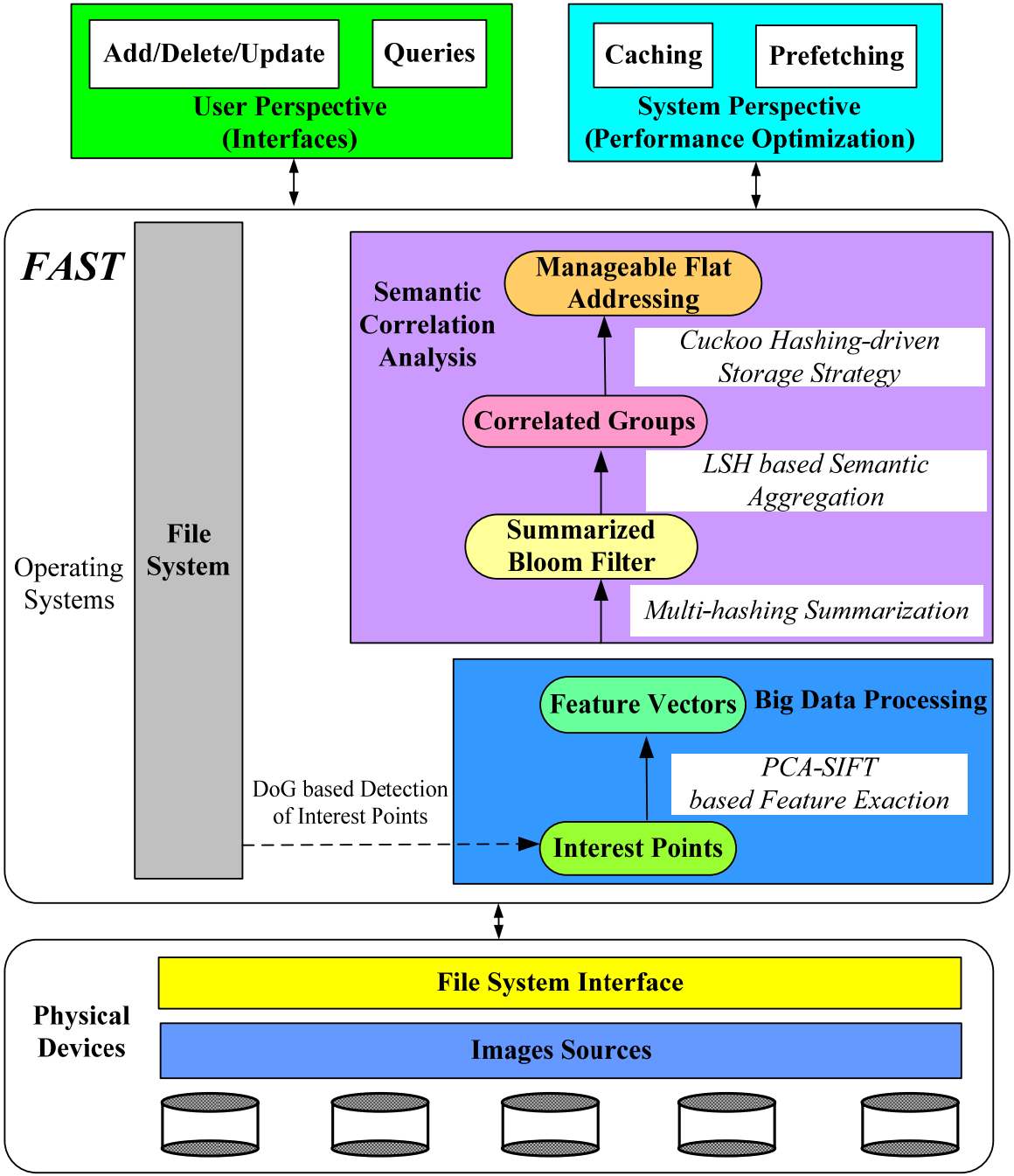
- Performance problems:
  - Handle big data via multi-step operations
  - Long latency, heavy space overhead
- “fail-to-do” problems:
  - Save as….
  - Different angles
  - Continuously taking pictures

An Example

- Finding missing children
- Complementary to video surveillance systems

Yu Hua, Hong Jiang, Dan Feng, "FAST: Near Real-time Searchable Data Analytics for the Cloud", Accepted and to appear in the Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC), November 2014.

The architecture



Semantic hashing

- **Design goals**:

- Fast identify correlated data in O(1)-scale complexity
- Implement “divide and conquer” in groups

- The grouping is probabilistic, but acceptable

Open problems and Conclusions

- Real-time performance is important in the era of big data.
- Storage ecosystem offers comprehensive supports to the methodology.
- Open problems:

✓The definition of near duplicate

➢*Remove the redundancy?*

✓The evaluation of query accuracy

➢*A, M, P?*

✓Related with multimedia research work

➢*The accuracy bound?*

Thanks and Questions