FAST: Near Real-time Data Analytics for the Cloud

Yu Hua	Hong Jiang	Dan Feng	Lei Tian
Huazhong Univ. of Sci. and Tech.	Univ. of Nebraska-Lincoln	HUST	UNL
Wuhan, China	Lincoln, NE, USA	Wuhan, China	Lincoln, NE, USA
csyhua@hust.edu.cn	jiang@cse.unl.edu	dfeng@hust.edu.cn	tian@cse.unl.edu

Abstract

Existing cloud storage systems have largely failed to offer an adequate capability for real-time data analytics. Since the true value of data heavily depends on how efficiently data analytics can be carried out on the data in (near-) real-time, large fractions of data unfortunately end up with their values being lost or significantly reduced due to the staleness of data. To address this problem, we propose a near real-time and cost-effective data analytics methodology, called FAST, in the cloud. FAST explores and exploits the semantic correlation property within and among datasets via correlation-aware hashing and manageable flat-structured addressing to significantly reduce the processing latency, while incurring acceptably small loss of accuracy. FAST is demonstrated to be a useful tool in supporting near real-time processing of real-world cloud applications.

Categories and Subject Descriptors

D.4.2 **[Operating Systems]:** Storage Management

Keywords

Cloud Storage, Data Analytics

1 The FAST Methodology

So far, only a tiny fraction of the data being produced has been explored for their potential values through the use of data analytics tools. Real-time data analytics are very important in dealing with large-scale datasets. This is also non-trivial to cloud systems, although they contain high processing capability (hundreds of thousands of cores) and huge storage capacity (PB-level). The fundamental reason is because the analytics must be subject to hard time deadlines that usually cannot be met by brute force with an abundance of resources alone. Existing approaches often fail to meet the (near) real-time requirements because they need to handle high-dimensional features and rely on high-complexity operations to capture the correlation.

SoCC'13, Oct 01-03, 2013, Santa Clara, CA, USA.

ACM 978-1-4503-2428-1/13/10.

http://dx.doi.org/10.1145/2523616.2525932.

We propose a novel, near real-time methodology for analyzing massive data, called FAST, with a design goal of efficiently processing such data in a real-time manner. FAST is to explore and exploit the correlation property within and among datasets via correlation-aware hashing [2] and flat-structured addressing [4] to significantly reduce the processing latency of parallel queries, while incurring acceptably small loss of accuracy.

It is worth noting that the FAST methodology can be extended to and well suited for multiple data types. Many data types can be represented as vectors based on their multi-dimensional attributes, including metadata (e.g., created time, size, filename, etc) and contents (e.g., chunk fingerprints, image interest points, video frames, etc). FAST extracts key property information of a given data type in the form of multi-dimensional attributes and represents this information in multi-dimensional vectors (i.e., multi-dimensional tuples). Each dimension is one component of the vector. The vector-based representation is fed as input to FAST for the subsequent operations of hash-based summarization, semantic aggregation and flat-structured addressing.

FAST is a generalizable methodology of which some components and aspects are derived from and have been used in existing storage systems, such as Spyglass [3] and SmartStore [1]. However, due to their specific and custom designs, these existing systems, while achieving their original design goals, fail to efficiently support near real-time data analytics. Moreover, by incorporating the FAST methodology, existing systems can obtain better performance improvements. The FAST methodology is demonstrated, by way of a "finding missing children" use case that involves identifying near-identical images in near real-time from massive image datasets, to have a great potential to efficiently support the near real-time analytics for heterogeneous types of data. In fact, the FAST use case outperforms the state-of-the-art schemes by 3-4 orders of magnitude in query latency.

Acknowledgment

This work was supported in part by NSFC 61173043; National Basic Research 973 Program of China 2011CB302301; NSFC 61025008, 61232004; US NSF-IIS-0916859, NSF-CCF-0937993, NSF-CNS-1016609, NSF-CNS-1116606. Authors greatly appreciate anonymous reviewers for constructive comments.

References

[1] Y. Hua, H. Jiang, Y. Zhu, D. Feng, and L. Tian. SmartStore: A New Metadata Organization Paradigm with Semantic-Awareness for Next-Generation File Systems. *Proc. International Conference for High Performance Computing, Networking, Storage and Analysis (SC)*, 2009.

[2] P. Indyk and R. Motwani. Approximate nearest neighbors: towards removing the curse of dimensionality. *Proc. STOC*, pages 604–613, 1998.

[3] A. W. Leung, M. Shao, T. Bisson, S. Pasupathy, and E. L. Miller. Spyglass: Fast, Scalable Metadata Search for Large-Scale Storage Systems. *Proc. FAST*, 2009.

[4] R. Pagh and F. Rodler. Cuckoo hashing. *Proc. ESA*, pages 121–133, 2001.