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

Ayat Fekry
akmf3@cl.cam.ac.uk
Computer Laboratory, University of Cambridge

The success of big data solutions principally rely on the timely extraction of valuable insights from data. This will continue to become more challenging due to the growths in data volume without corresponding increase in velocity. We advocate that storage systems of the future should include functionality to detect and harness fundamental data characteristics such as similarity and correlation. This has the potential to optimize storage space, reduce amount of processing needed for further information extraction, and save I/O and network communications.

We propose storage self-cleaning as a service that performs intrinsic similarity and correlation analysis with minimal overhead to achieve a set of goals. 1) Optimize storage space by performing deduplication on the dataset level in contrast to the conventional block level. 2) In addition to serving data, storage would provide correlation information. This will speed up data analytics, as correlation is a crucial element in most analytics algorithms [1]. 3) Save I/O, CPU, and network consumption due to the optimized volume and enhance velocity.

Why storage?

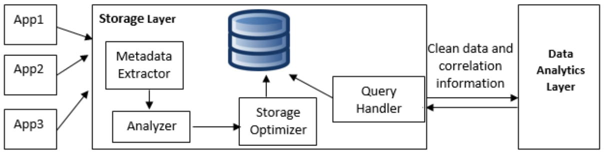
Traditionally, data similarity and correlation analysis is done in the Data Analytics Layer (DAL). This implies slower velocity and increase in I/O and communication costs due to data movement. On the contrary, the storage layer provides a centralized and proximate place for data. Furthermore, this analysis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

A Motivating Example

In today's connected world, we have several smart city applications that use various sensors. Consider traffic, air quality, weather, and energy consumption sensors that report their periodic readings in a big data storage system. Healthy sensor data has high similarity and relatively fixed correlation patterns. For example, with the same temporal and spatial context, air quality readings negatively correlate with traffic congestion, and energy consumption readings correlate with temperature (positively in hot countries and negatively in cold ones). A System with these characteristics would benefit from the proposed self-cleaning service by deduplicating similar data and detecting anomalous data. While this is a use case, most datasets are self-similar and would benefit from the proposed service, as they tend to have embedded correlation patterns.

Key Idea

Leverage lightweight correlation and similarity analysis at the storage level to clean redundant, anomalous data and accelerate extraction of insights in the DAL.



Advantages

Reduce the "Garbage in garbage out" problem and facilitate big data cleaning which is the most time and effort-consuming task in big data analysis [2]. The proposed self-cleaning service imposes minimal changes to the application layer and provides a unified way of data cleaning at the storage level. Ultimately, this allows faster and richer insights at the DAL to arise, for example forecasting, recommending, and what-if analysis built on top of clean data and correlation information.

Challenges and open questions

1) Overhead: this relies on the complexity of correlation and similarity algorithms and selecting convenient time to execute them so that storage performance is minimally affected. Recent correlation and anomaly detection algorithms have shown lightweight ones that do not need complex machine learning [3]. 2) Hardware support: Carrying out correlation at line-rate on large datasets is spatially and computationally expensive process. ASIC hardware support is necessary to support the software design; however, the ASIC needs to be generally programmable and adhere to power and silicon area budgets. This requires work at the hardware level. 3) Similarity/correlation definition: similarity can be defined as exact duplicates or similar semantics depending on data domain; correlation and similarity can be expressed either using rules as in Intelliclean [4] or modeled via machine learning algorithms as in Tamr and twitter long term anomaly detection [5, 6]. Rules might be tedious to define and generalize yet have less overhead, and while ML automates similarity/correlation extraction, it prerequisites feature engineering which metadata could assist in. 5) Data Lineage: has high capture overhead but can accelerate data cleaning tasks. Analytics tasks were significantly accelerated using lineage in SEeSAW [7].

REFRENCES

[1] Chen, M., Mao, S. and Liu, Y., 2014. Big data: A survey. Mobile Networks and Applications, 19(2), pp.171-209.

[2]http://www.iqworkforce.com/how-data-scientists-spend-their-time/

[3] Barbhuiya, S., Papazachos, Z.C., Kilpatrick, P. and Nikolopoulos, D.S., 2015. A Lightweight Tool for Anomaly Detection in Cloud Data Centres. In CLOSER (pp. 343-351).

[4] Lee, M.L., Ling, T.W. and Low, W.L., 2000, August. IntelliClean: a knowledge-based intelligent data cleaner. In Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 290-294). ACM.

[5] Stonebraker, M., Bruckner, D., Ilyas, I.F., Beskales, G., Cherniack, M., Zdonik, S.B., Pagan, A. and Xu, S., 2013, January. Data Curation at Scale: The Data Tamer System. In CIDR.

[6] Vallis, O., Hochenbaum, J. and Kejariwal, A., 2014, June. A Novel Technique for Long-Term Anomaly Detection in the Cloud. In HotCloud.

[7] Kannan, K., Bhattacharya, S., Raj, K., Murugan, M. and Voigt, D., 2016, June. SEeSAW-Similarity Exploiting Storage for Accelerating Analytics Workflows. In 8th USENIX Workshop on Hot Topics in Storage and File Systems (HotStorage 16). USENIX Association.