DATA
WAREHOUSING
08.12 | EB 7315

THE IMPACT OF DATA TEMPERATURE ON THE DATA WAREHOUSE

Leveraging Data Temperature with Hybrid Technologies

TERADATA

TABLE OF CONTENTS

DATA TEMPERATURE BASICS

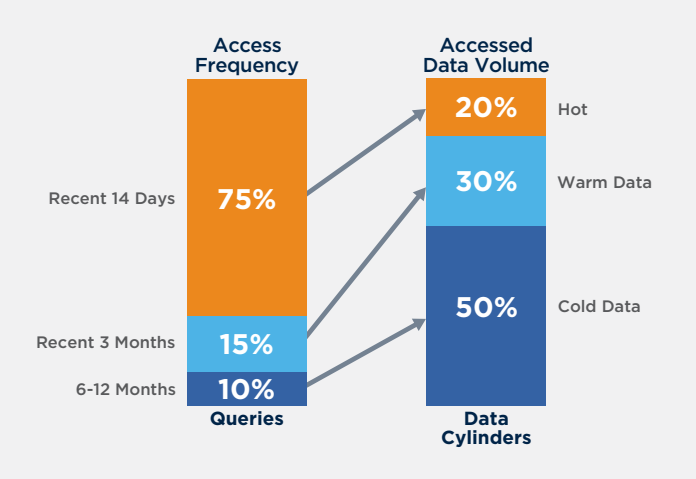
The Enterprise Data Warehouse (EDW) serves as the central repository for information and the users that require increasing access. One of the least understood, and certainly the least managed facets of an EDW, is how often users tap into the information it holds. The frequency at which specific data is accessed, described as its "temperature", affects the performance of the data warehouse. Consequently, managing data by its temperature can open up opportunities to provide value across the enterprise with faster analytics.

Multi-Temperature Data (MTD) refers to different elements of data being subject to different frequencies of access. Teradata has found that data warehouses have both hot and cold data. Hot data is frequently accessed data, reflecting tactically driven behavior and typical short-term need for high-priority business-driven decision support. Cold data is historical information usually seen in data mining, year-over year analysis, and deep analytic activities. Ideally, the hot data should be stored on high performance solid state drives (SSD), while cold data would reside on cost effective high capacity hard disk (HDD) drives.

The concept of data temperature refers to the frequency with which certain data is accessed within a system. It is measured on a cylinder level by the metric collection process. The most frequently accessed data is referred to as 'Hot', the moderately accessed data is referred to as 'Warm', while the occasionally accessed data is referred to as 'Cold'. For some classes of data this parameter is static (or fixed), for others it is dynamic and changes during the system workload cycles depending on the business dynamics.

It is important to separate the concepts of data access (queries) from temperature for a moment. Access frequency is the amount of I/O requests to a set of data blocks, aka cylinders. Temperature is a measure of the number of cylinders (gigabytes) that have high, medium, or low access frequency.

The below example displays how a high percentage of the queries (75% in the example) access a small percentage of cylinders (20% in the example).



DATA TEMPERATURE AND HYBRID TECHNOLOGIES

In Teradata hybrid technology both the solid-state drives (SSDs) and hard disk drives (HDDs) are fully utilized with Teradata Virtual Storage, which automatically migrates data between the two types of storage to achieve optimum performance. This solution tracks data use and intelligently moves it to the appropriate disk with "hot" data on the faster SSDs and the less used "cold" data on the slower HDDs. The result is a platform that makes smarter decisions at hyper speeds.

It's important to note that data temperature changes over time, so to work well, a hybrid system has to be able to respond to changes. Optimum placement is maintained by the data migration provided by the Teradata Virtual Storage (TVS) product. TVS considers many factors such as data access frequency, and data type to automatically and continuously optimize the data placement onto the appropriate storage device: SSD or the HDDs to match the data temperature.

Teradata experts use temperature demographic data to size and configure the hybrid storage, i.e. what percentage is hot/cold data and how much SSD space is needed for hot data.

Optimizing hot and cold data allows processing capacity to be focused in the data warehouse and additional analysis to be undertaken in even shorter time periods. This results in more granular and precise insight and more focused processing power for incremental business improvements to the enterprise. The ultimate technical advantages include better management of workloads and optimizing user activity while ensuring that the full processing capacity of the EDW is exploited.

Managing data by temperature ensures that the precious and expensive storage capacity and processing capability are used most efficiently , earning more from the data warehouse investment and achieving incremental business value through more effective use of the daily EDW processing. Leveraging MTD can also improve the end-user experience since by making the business queries respond faster, the users can take full advantage of new business and technology improvements while optimizing their current operations and activities. In short, it delivers value for the entire organization.

ADVANTAGES OF HYBRID TECHNOLOGY IN THE EDW

Efficient Performance

Until now, there has been a trade-off between cost and performance with storage subsystems. The only way to gain IO performance was by adding more processors and HDDs to a system, often resulting in excess storage capacity. Alternatively boosting performance with just SSDs merely added higher cost to the total capacity ratio. Teradata's hybrid data storage platforms leverage the best attributes of SSD and HDD technologies, perfectly blending the costs and performance. Data is stored on the appropriate drive depending on business requirements and it can automatically be migrated as the requirements change.

If a customer regularly accesses 25% of data for example, the hybrid system may store the 25% on the fastest device and the rest of the data will migrate toward the lower performance device with a better cost per gigabyte.

Teradata's hybrid storage solution can deliver a significant performance improvement without increasing data center footprints, breaking the budget or sacrificing energy efficiency.

THE IMPACT OF DATA TEMPERATURE ON THE DATA WAREHOUSE

Cost Effective Utilization of Storage

Identifying cold or dormant data, then managing it, can result in capacity savings. Upgrades are usually related to expected increases in required processing power, often through new applications being added or existing ones being expanded. The storage performance needed to support that increased processing power would in previous systems have resulted in more storage capacity then required. Storage consists of two aspects: IO performance and capacity. By blending small amounts of SSD for high performance with the capacity of HDDs, its possible to achieve the best of both technologies with hybrid storage. This capability depends on software, such as TVS, that ensures the hot data is always in the SSDs, all done automatically. Moving cold or dormant data to HDDs allows it to cost effectively stay in the data warehouse when needed for a broad analysis over longer time periods.

Analyzing the use of data may show an over-provisioning in capacity, which could result in an added expense. Often this over-capacity is a result of leaving unused capacity in the system as a technique for increasing the effective performance for the remaining data capacity. Hybrid storage eliminates the need for this over capacity approach to performance. In addition, analyzing data usage patterns will give a clearer perspective of the storage that's required for growth and allow better, more effective financial planning.

Enhance the Value of the Data Warehouse

Teradata's hybrid storage empowers organizations to stay ahead of their competitors by giving leaders better information faster. It also enables them to grow the use of their data warehouse across an expanding spectrum of applications and functions by applying increased performance to integrated data.

A data warehouse based on hybrid storage will enhance the value of the business. Here are some examples of business benefits:

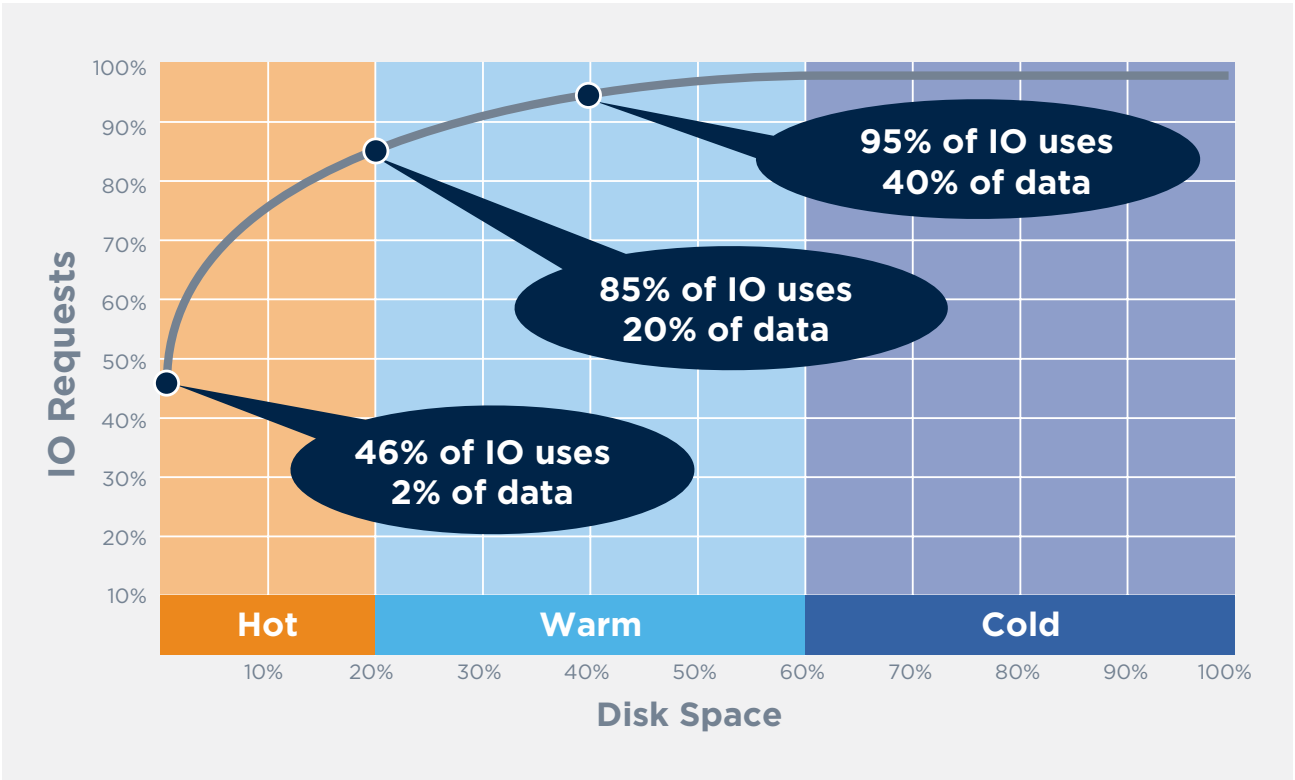
- Lower total cost of ownership (TCO)
- Greater return on investment (ROI)
- New business prospects through enhanced EDW performance
- Increased conformity to service level agreements

In addition, hybrid technology facilitates savings in data center space and energy. For instance, the Teradata Active Enterprise Data Warehouse 6690 delivers up to a 75% lower energy cost over a traditional hard disk drive based data warehouse along with a 75% smaller data center footprint. This dramatically increased performance per square foot and per kilowatt can yield greatly improved total cost of operation.

TAKING THE TEMPERATURE OF YOUR DATA

The assessment of Data Temperature in a Teradata data warehouse requires the use of two tools: the IO Count and Data Temperature analysis tools. The IO Count tool is a software utility that counts the number of I/O's performed on a 'cylinder' or 'zone' of data. It collects the information on how frequently each cylinder is accessed and if the access is a spool (temporary database data) or non-spool (table) data. This usage data is collected over a period of time that would represent the typical workload patterns on the system, typically 1 to 2 weeks. The tool does not impact system performance. Once this collection window has completed the resulting I/O counts per cylinder are saved. This information is then fed into the Data Temperature analysis tool to analyze the temperature demographics of the system's data usage. This analysis provides Teradata experts with the information on how the customer accesses their data so the appropriate system and configuration may be determined. The Data Temperature tool establishes the appropriate amount of SSDs for the configuration to enable optimization of storage to node ratios for a hybrid storage solutions.

The chart below graphically represents the output of the Data Temperature tool which is the measured I/O activity for each portion of the system data space. The values in this chart are examples of data usage that a typical data warehouse might exhibit. Note that nearly half (46%) of the total I/O requests had used a very small amount of the total data space (2%). This "super hot" data would typically reside in the main memory of the Teradata system for optimum performance. Even more significant is that 85% of query I/O workload had used just 20% of available data space. This hot data would ideally reside on the SSD in a hybrid storage based system to enable the majority of queries to take advantage of the speed of SSD. Areas of the data space are colored coded in this chart to correspond to the temperature of the data regions. While the warm and cold colored areas depict a typical distribution of warm and cold data, the key red section shows the hot data that will be kept in the fast SSD devices.

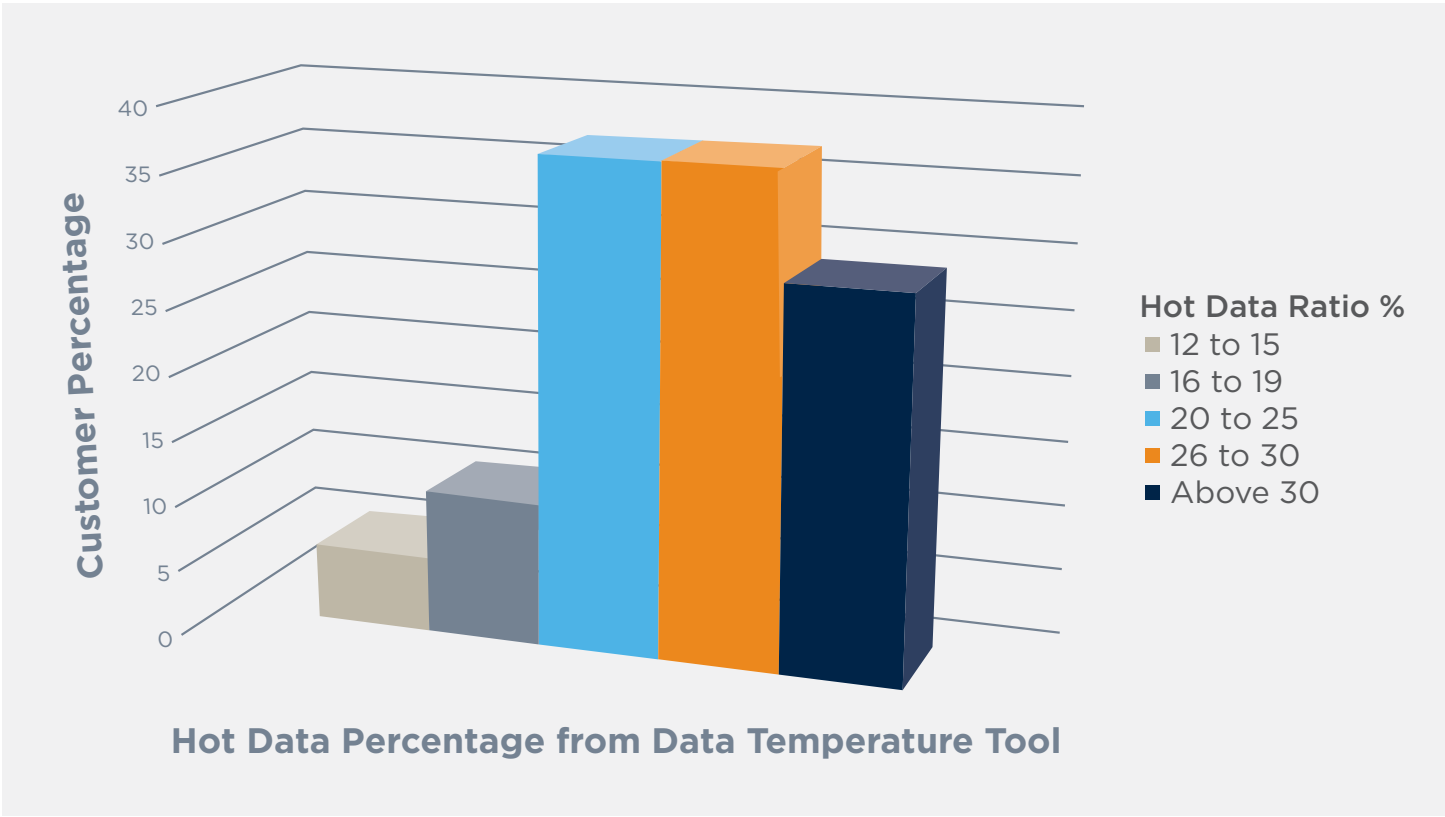


DATA TEMPERATURE IN ACTION

Since early 2011 Teradata has analyzed the data temperature demographics of a large number of EDW systems at companies worldwide in order to determine the temperature demographics of their data. The results are used to accurately estimate the total size of hot data capacity that is required to support business workloads and system usage patterns. The hot data size is most conveniently expressed as a percentage of total data space. A hybrid storage based EDW would best support this hot data capacity with the SSD portion of the hybrid storage sized to hold this hot data, that is, the percentage of SSD would equal the portion of total data that is rated as hot.

The chart below shows the hot data/SSD % results from a representative sample of hundreds of Teradata EDW systems across a spectrum of primary industries. Key observations from this data are:

- 70% of the systems have a hot data ratio between 20% and 30% of total data space, and, in fact, the average for all the systems was 28%. Both of these facts lead to a conclusion that a basic guideline for optimum hybrid storage use would be to accommodate approximately 25% of the total data space in faster SSD storage.
- A significant portion – over one-fourth - of the systems have a rather large portion of total data rated as hot with values of over 30% measured. These systems use a larger portion of their EDW data in normal operations so the performance benefits of hybrid storage will be leveraged across a broader range of data.
- A minority of the systems measured had less than 20% for their hot data ratio. The workloads on these systems would be able to very efficiently benefit from hybrid storage since only a relatively small portion of total data would need to be accommodated in fast, SSD storage.



INDUSTRY SPECIFIC DATA TEMPERATURE TRENDS

While the observations on the whole sample population discussed above provide general guidelines on data temperature in the EDW, insights into data usage by individual industries are even more valuable. By analyzing the sample population from the perspective of the five major industry segments represented in the sample, even deeper insights are evident.

Manufacturing

Manufacturers need to have the ability to continuously analyze large amounts of current process and demand chain data and determine long term trends that drive immediate decisions on the business and its operations, This often requires analysis of longer periods of historical records which in turn tends to "heat up" larger amounts of the EDW data with queries for dashboards and tactical reporting. This is reflected in the fact that over two-thirds of the manufacturing companies in the sample had hot data ratios at or above the sample-wide average of 28%. Also, these decisions are often based on the most current data from operations – even up to the minute – so loading and the associated extraction/transformation activity is a major portion of their regular workload at a high, 24X7 pace. This freshly loaded data is always rated as hot by the TVS software in hybrid storage, so this will tend to increase the total hot data ratio.

High volume and CPG manufacturers take advantage of applications that allows their sales teams to review sales online in dashboards and in complex summary information. Sales teams would obviously consider much of their data "hot" due to the need for short response times in the many short queries used to build these dashboards. The "cold data" in these systems is typically the historical records for their process and operations data which represents only a small number of years.

Communications

In the communications industry segment, like in manufacturing above, over two-thirds of the companies in the sample have hot data ratios at or above the sample-wide average of 28%. The system usage, for both the wireline and wireless providers here, depends on loading large amounts of daily call, billing, and network information to be used in business analytics for analysis such as churn prevention, next "best offer" to customers, and other customer relationship management tasks. In addition, analysis and decisions are typically based on relatively long time periods of customer activity history in the three month range or longer. Both of these factors will result in larger amounts of hot data in relation to the overall data table space, since communications companies tend not to hold history data much longer than two to three years for in-depth analysis.

Retail

Unlike the Manufacturing and Communications industry companies discussed above, this industry reported relatively small ratios of hot data with almost three-quarters of the retailers having hot data ratios at or below the average of 28%. Not only do these companies keep large amounts of colder history data in the EDW (one in the sample kept 11 years' worth), their business analysis is focused primarily on the current period (quarter, week) compared to the like period one year earlier. The portion of data accessed for these workloads is relatively small in relation to the entire data warehouse. Also, very low hot data ratios (less than 20%) are concentrated in the "mass merchandiser" store and food segments of the retail space.

Loading of daily sales and inventory data is a nightly batch process since next day reports satisfy the business needs. This reduces the amount of immediate, hot data focused queries thereby reducing the hot data size.

Financial

Similar to the retail industry results described above, the companies in this segment reported relatively low ratios of hot data with almost three-quarters having hot data ratios at or below the average of 28%. This trend holds valid across all the major elements of this industry including banks, insurance, and equities traders. Financial firms manage extremely hot data like ATM transactions and, securities transactions for event based offers. They tend to do a lot of intense analysis on a smaller amount of data, usually the current month out of a two to 5 year history. For instance, one company in the sample supports 350 different applications that run on their systems daily. At the same time, regulatory demands require that they maintain up to seven years' worth of detailed historical records. Huge volumes of historical data, while providing valuable business perspective, would be too costly to maintain in a performance-oriented environment but perfect for lower cost HDDs. A hybrid solution offers the needed performance for the hot data yet provides a cost-effective solution for storing the regulatory data.

Travel/Transportation

The companies in this industry segment, consisting of airline, shipping, and delivery service firms, reported hot data ratios that centered on the 28% average hot data percentage for the entire sample set. The system usage across these companies typically consists of heavy analysis of up to 6 months of system wide operational and customer transaction data from 2 to 3 years of history data. Workloads typically consist of customer relationship management analysis, transportation system reporting (such as delivery status), and loading of daily business data. Other key workloads produced by management reporting with BI tools such as Cognos and Microstrategy.

CONCLUSION

The data warehouse, by the very nature of its workloads and data environment, exhibits a multi-temperature data distribution in the access patterns of the data space. The majority of accesses to the system data occurs to a small portion of the total data - some is hot and much is not. Based on actual data temperature measurements, Teradata has shown that this multi-temperature data distribution holds true across all industries in a relatively narrow range of hot to cold data ratios, Each industry tends to have some unique patterns of usage that have an impact on the actual ratios they exhibit.

The Teradata hybrid storage technology takes advantage of the data temperature demographic of a system with Teradata Virtual Storage constantly monitoring data usage and then automatically migrate the often used data to the very fastest storage medium – SSD. A hybrid technology based system is configured to match the data temperature demographics of each solution through the use of data temperature monitoring tools. The result is that the Teradata Active EDW with hybrid technology provides higher and more efficient performance to enable enhanced business value for the data warehouse in the vast majority of industries.